Analysing within- and between-host HIV genetic diversity with `phyloscanner` for detection of transmission and more

Chris Wymant[1,2] on behalf the BEEHIVE and PANGEA Collaborations

[1] Big Data Institute, Li Ka Shing Centre for Health Information and Discovery,
Nuffield Department of Medicine, University of Oxford
[2] Department of Infectious Disease Epidemiology, Imperial College London

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

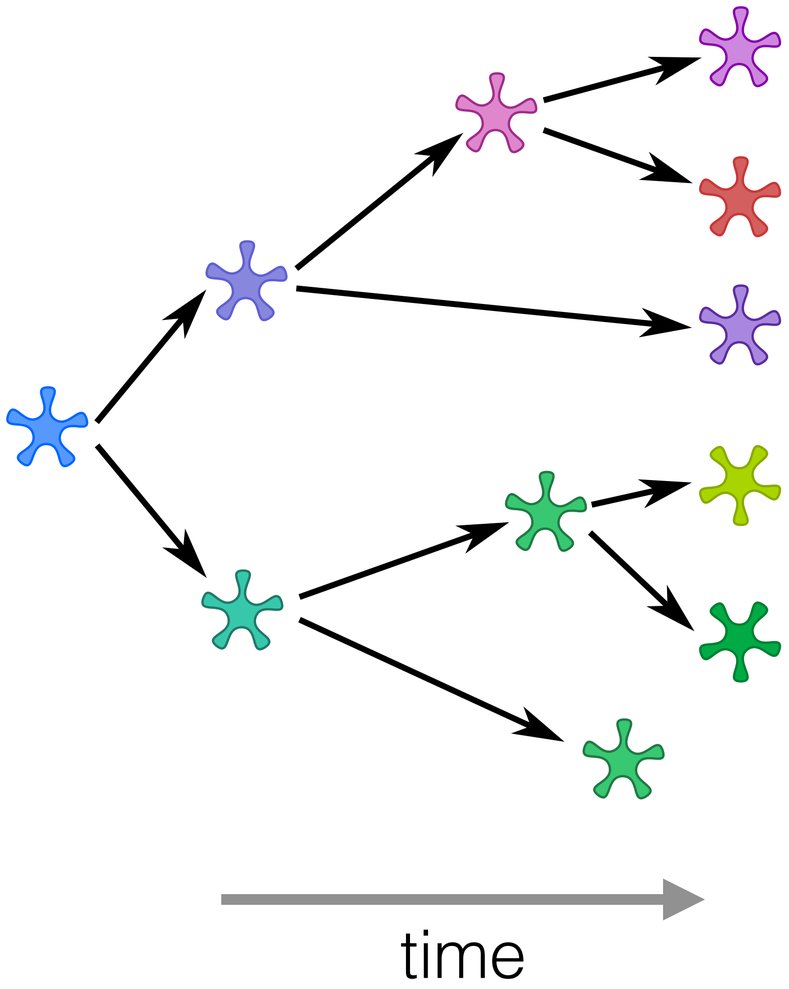
erc

Problem: how can we target our interventions in an epidemic to have a greater impact?

A solution: learn more about the epidemic through analysis of pathogen sequence data.

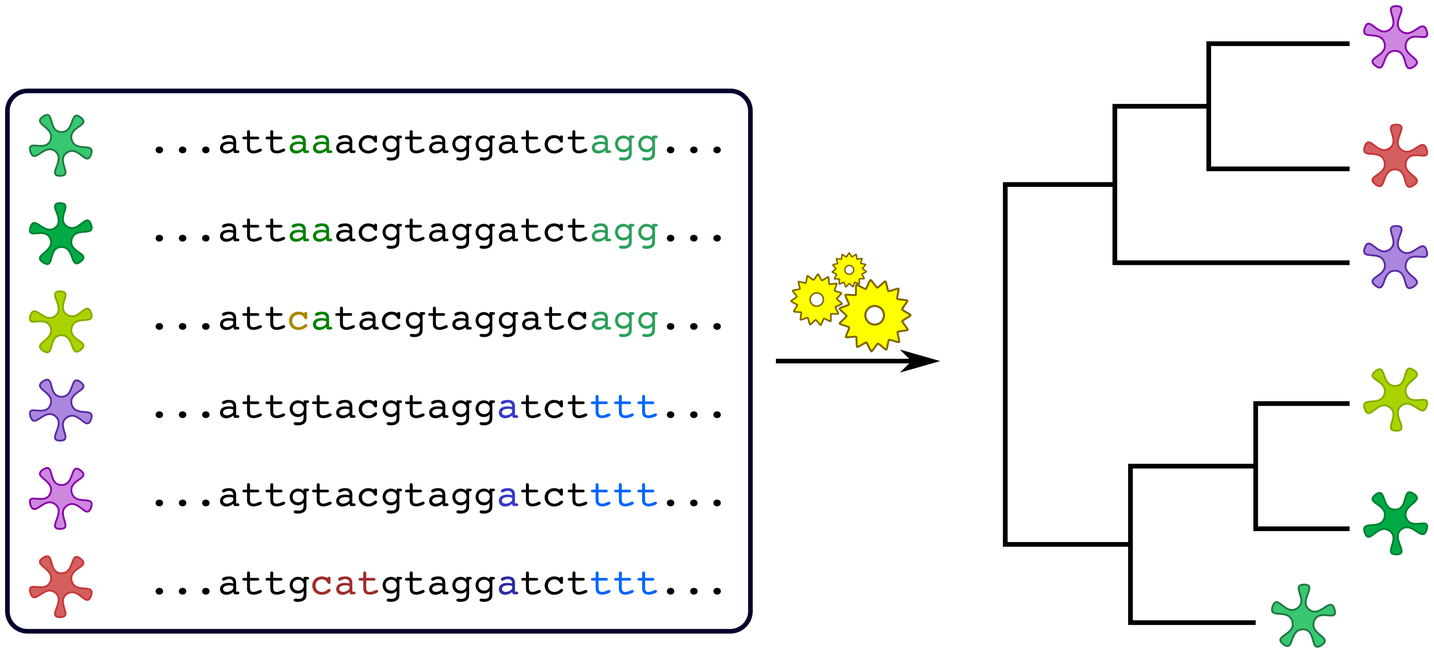
Genetic changes (substitution) accumulate over time

Substitution: replacement of a nucleotide (e.g. a $\rightarrow$ t)



Original slide by
Thibaut Jombart

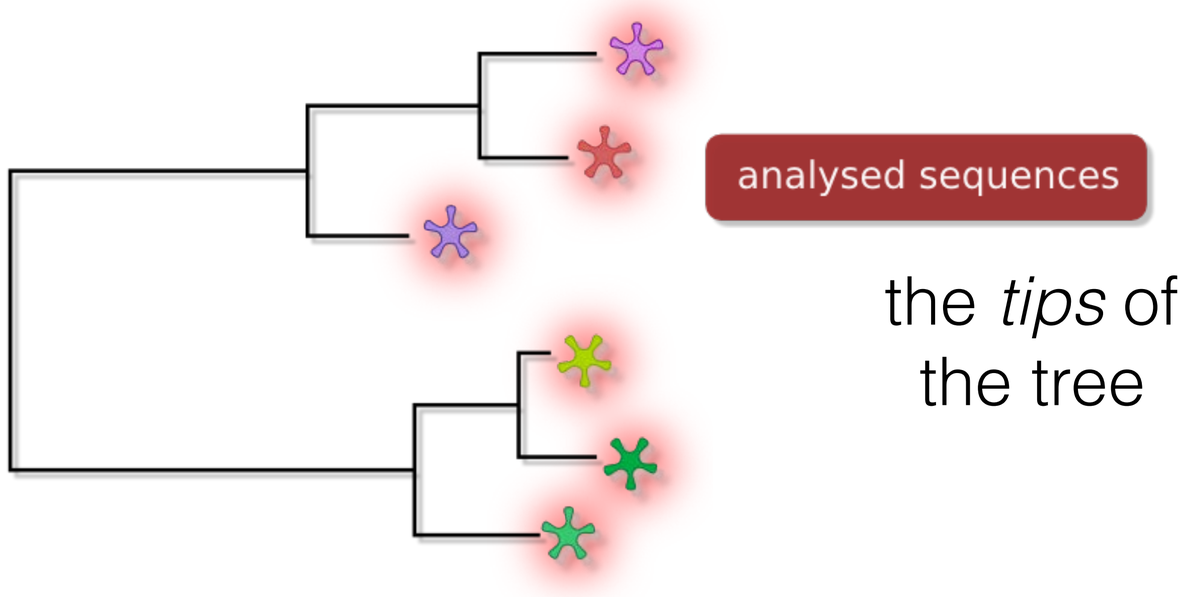
Using substitution patterns to reconstruct the evolutionary history



Phylogenetics aim to reconstruct evolutionary trees (*phylogenies*) from genetic sequence data.

Original slide by Thibaut Jombart

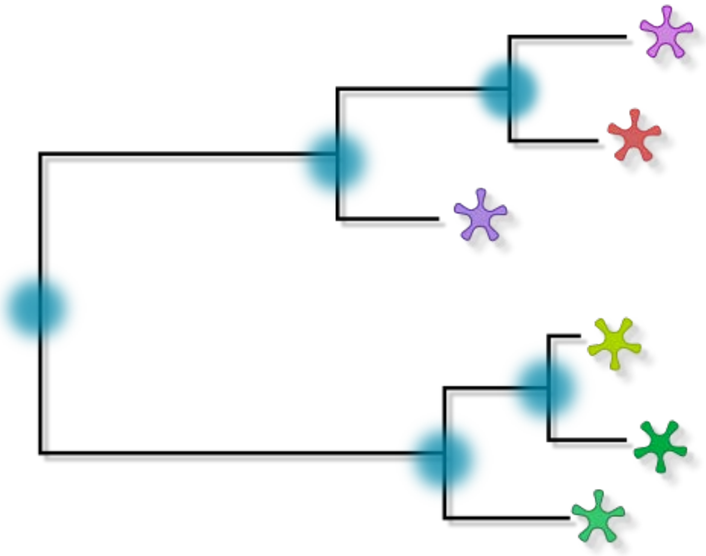
Using trees to represent the evolutionary history



Original slide by
Thibaut Jombart

Using trees to represent the evolutionary history

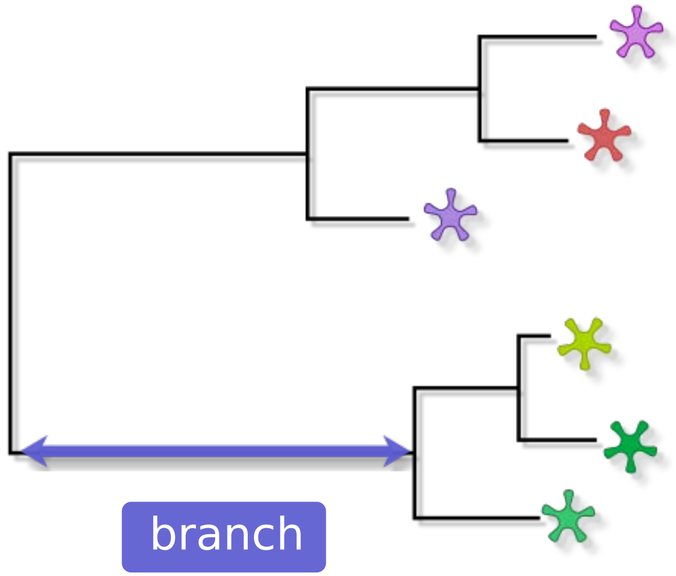
Most Recent Common Ancestors (MRCA)



the *nodes* of the tree

Original slide by Thibaut Jombart

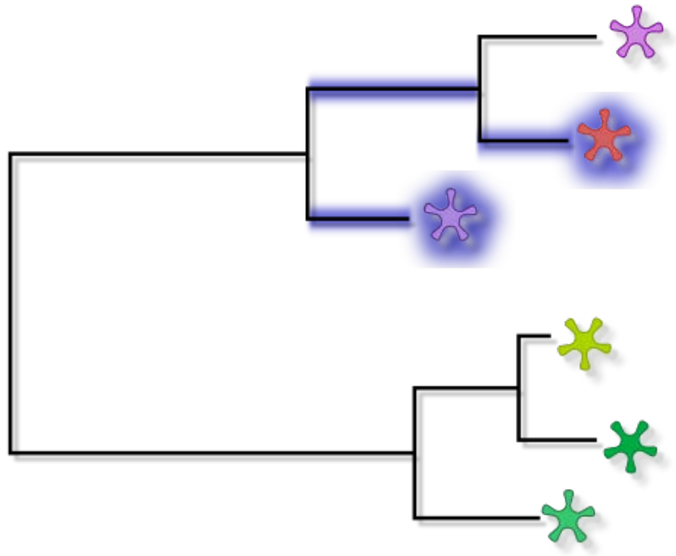
Using trees to represent the evolutionary history



length = amount of evolution (**not time**, as a rule)

Original slide by
Thibaut Jombart

Using trees to represent the evolutionary history



distances between tips

"*patristic*" distance: sum of branch lengths

Original slide by
Thibaut Jombart

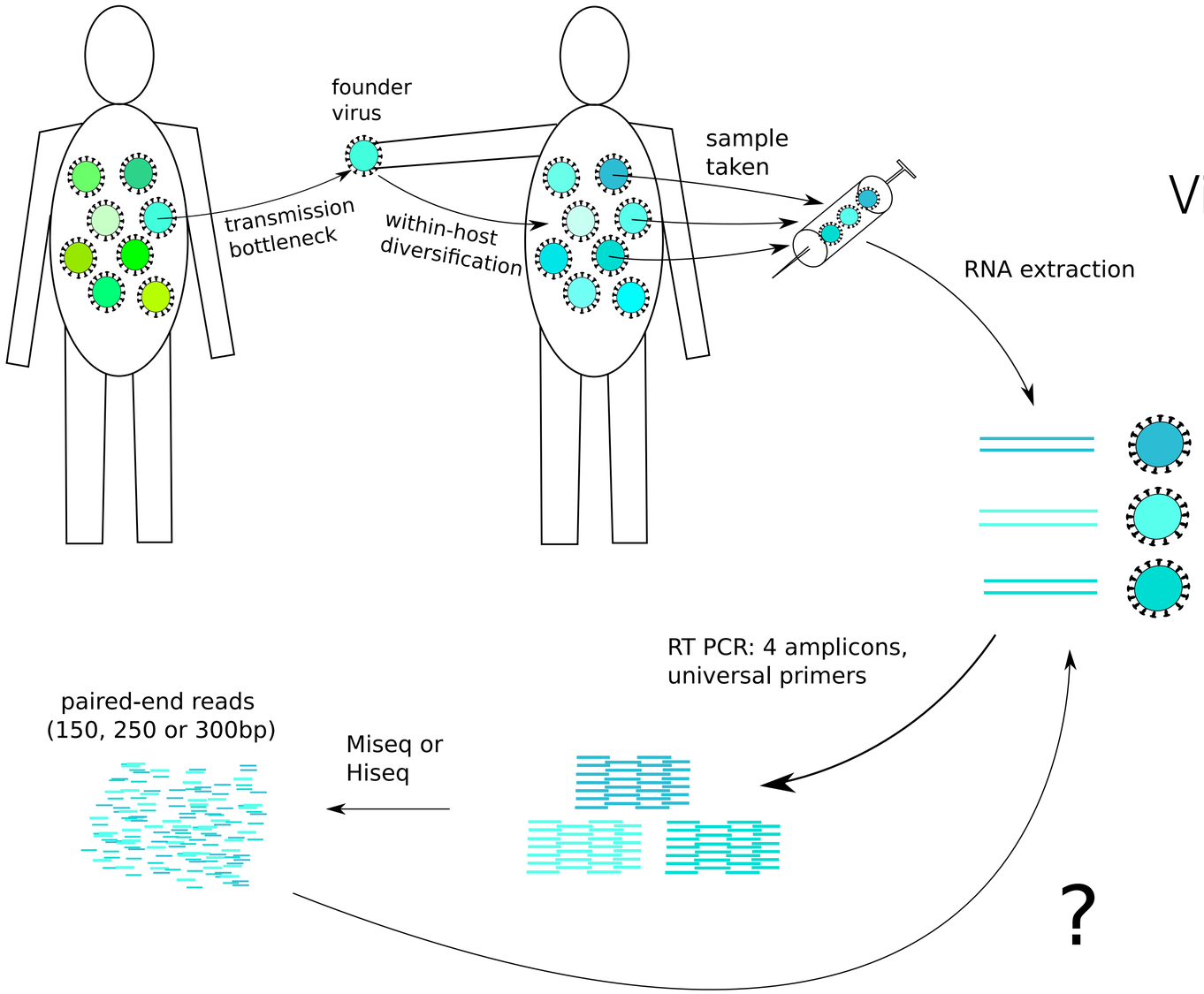
Small distance ⇒ small amount of evolution,
from which we infer proximity in a transmission network.

Problem: to meaningfully interpret small differences between closely related viruses, we want *accurate* genomes.

Problem: to make robust statements, need *many* genomes.

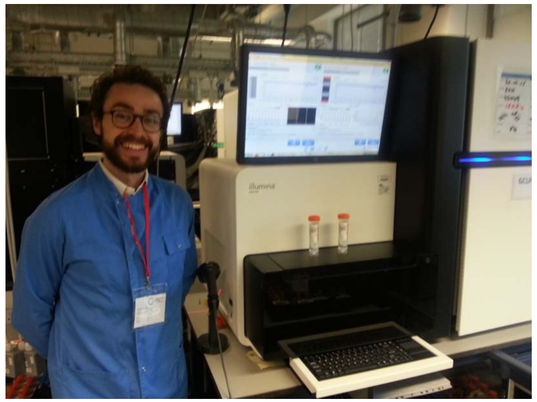
Solution: high-throughput sequencing + accurate, scalable sequence processing.

Our HIV Data



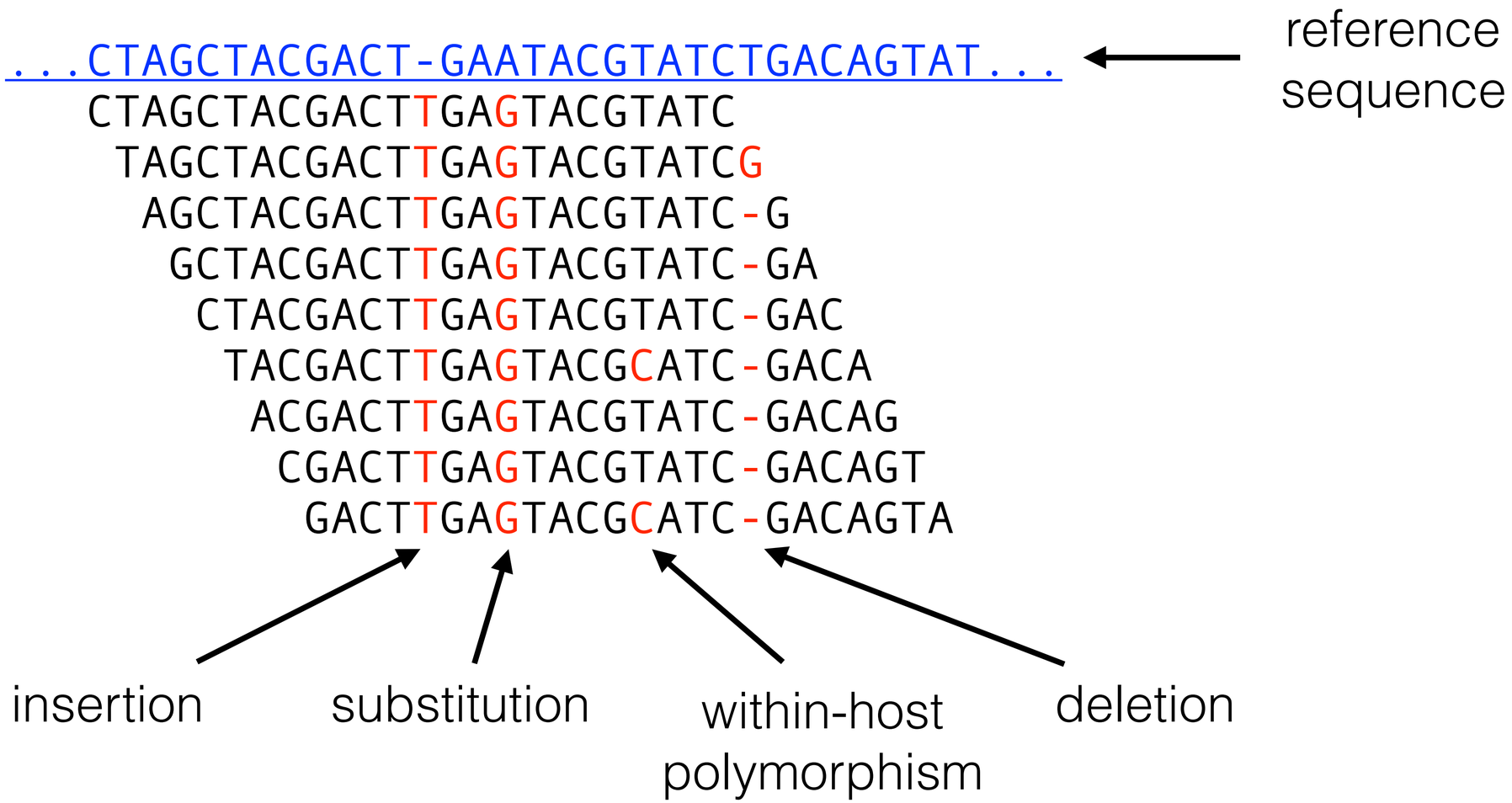
Cornelissen *et al.*
Virus Research 2016

Gall *et al.*
J. Clin. Microbiol.
2012



↑ visiting, not helping

Mapping Reads to a Reference

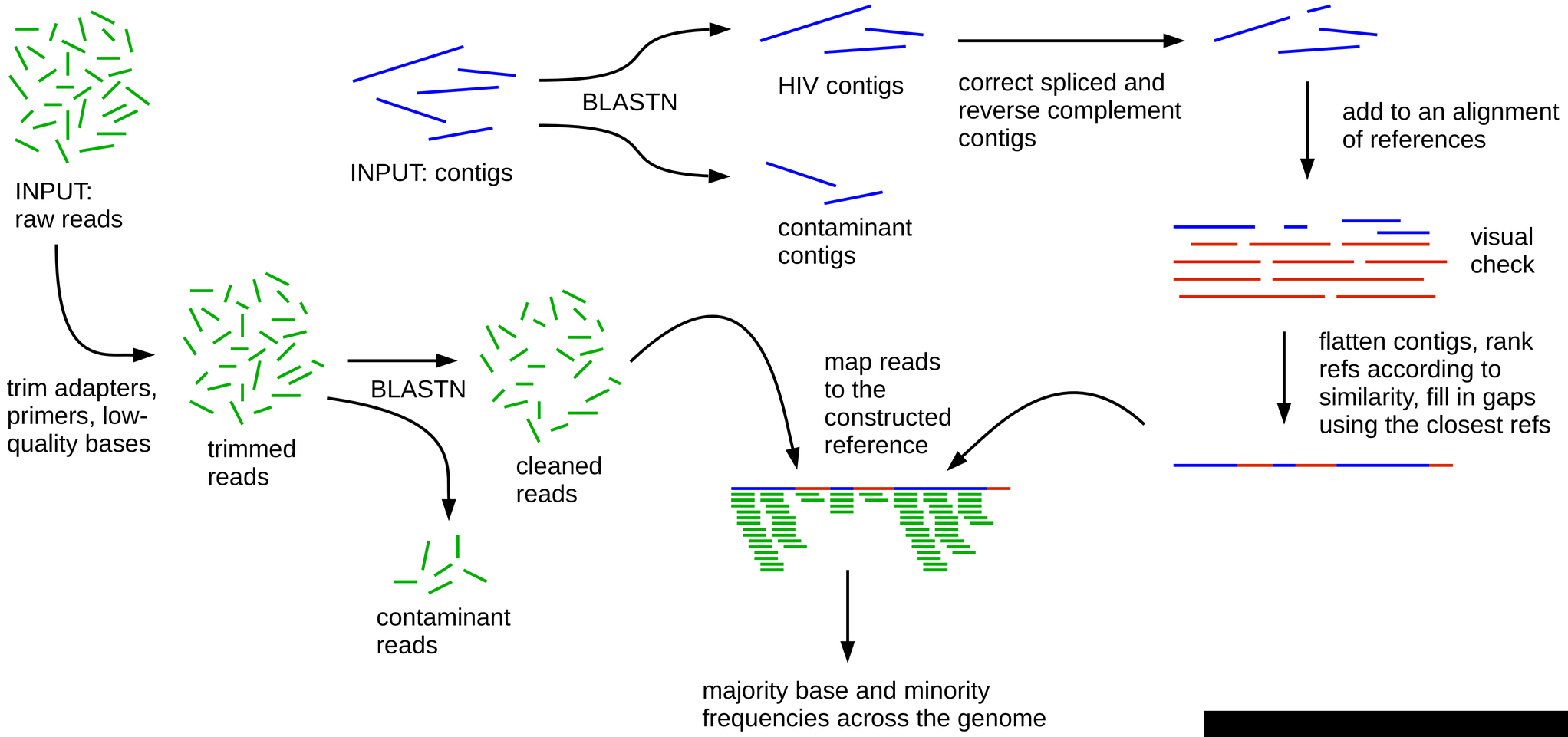


However, the more different a read is from its reference, the more likely it is to be aligned incorrectly or not at all. So **mapping causes biased loss of information**.

e.g. use the same reference for mapping for a group of individuals: bias their viruses to look similar to each other
→ false inference of transmission.

e.g. use old references for mapping new samples
→ false inference of slow evolution

`shiver` - Sequences from HIV Easily Reconstructed



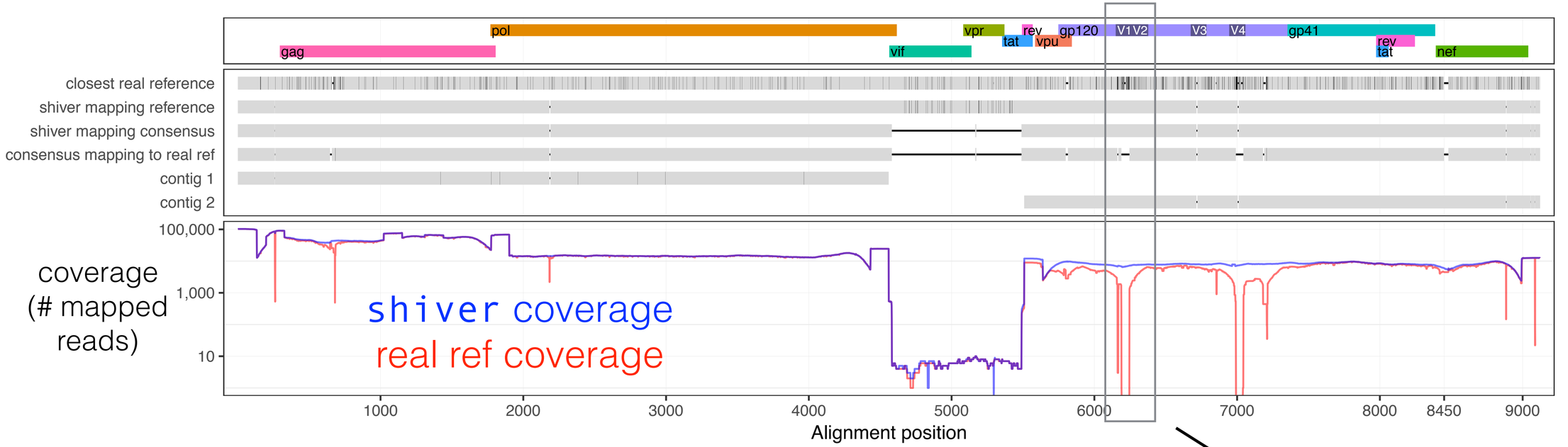
Wymant *et al.*, Virus Evolution 2018

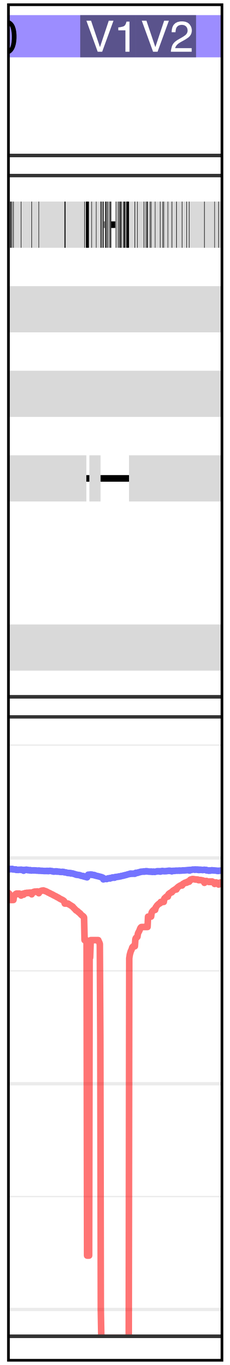
github.com/ChrisHIV/shiver

Also works on HCV, RSV, …

```
$ shiver_align_contigs.sh
$ shiver_map_reads.sh
```

Map to `shiver`'s constructed reference,
also map to the closest of 3259 real references (from database),
compare.





Over 50 BEEHIVE samples + 65 public samples:
median number of bases called differently with higher coverage:
13; with lower coverage, 0.
Recover missing sequence, often in *envelope* (vaccine design!).

Application of `shiver` for 3 HIV projects:

Investigate the viral-genetic basis of virulence

BEEHIVE: N ~ 3,000
PopART: N ~ 1,000 currently, 9,000 eventually → Help interpret the effect of a population-level test-and-treat intervention in Zambia and South Africa
PANGEA: N ~ 13,000 currently, 23,000 eventually

Generate new insights into transmission dynamics in generalized epidemics in Africa

Tanya Golubchik

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 individual

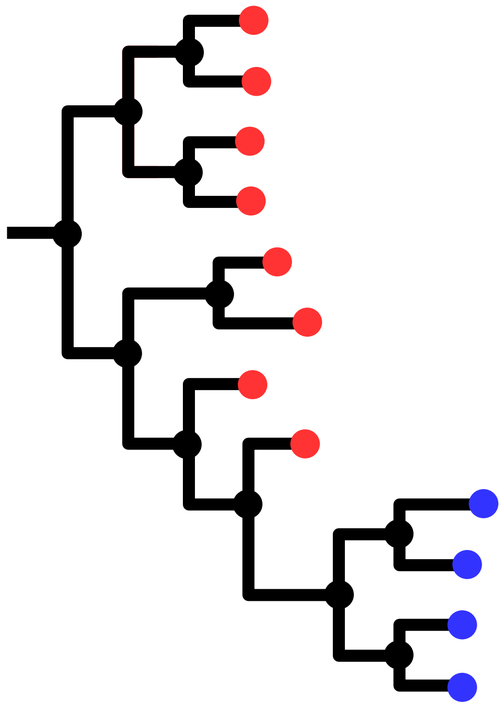
- Make a tree using one pathogen sequence per sampled individual.
- Define clusters: likely to be related by recent transmission.
- Investigate epidemiological correlates.

No information on who is transmitting!

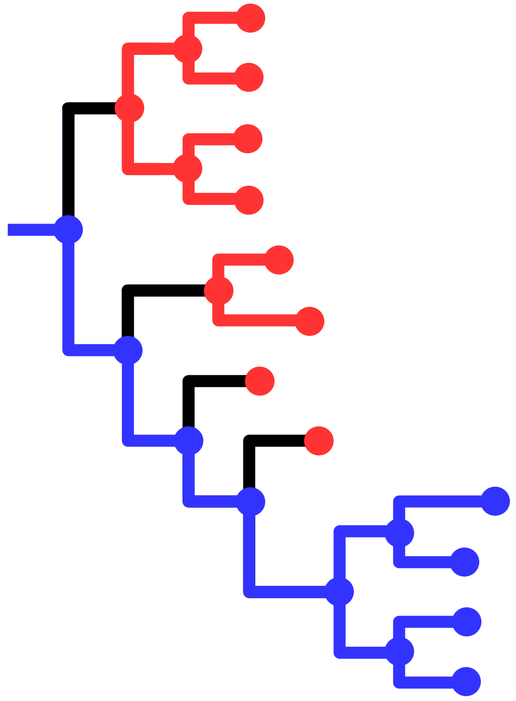
Can do better by also using extra data + transmission model. Conclusions then dependent on both.

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

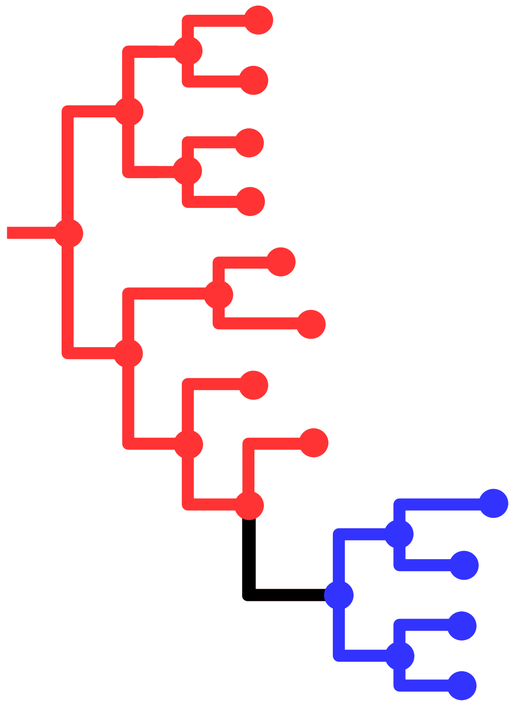
Ancestral State Reconstruction



Known states (red or blue) at the tips.



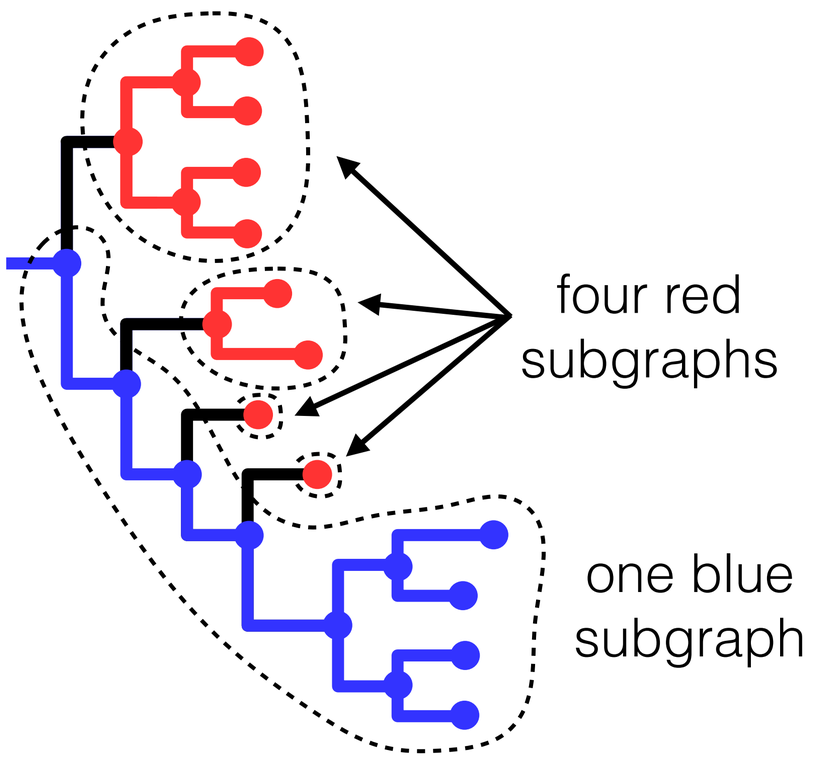
Red evolved from blue: 4+ changes of state needed.

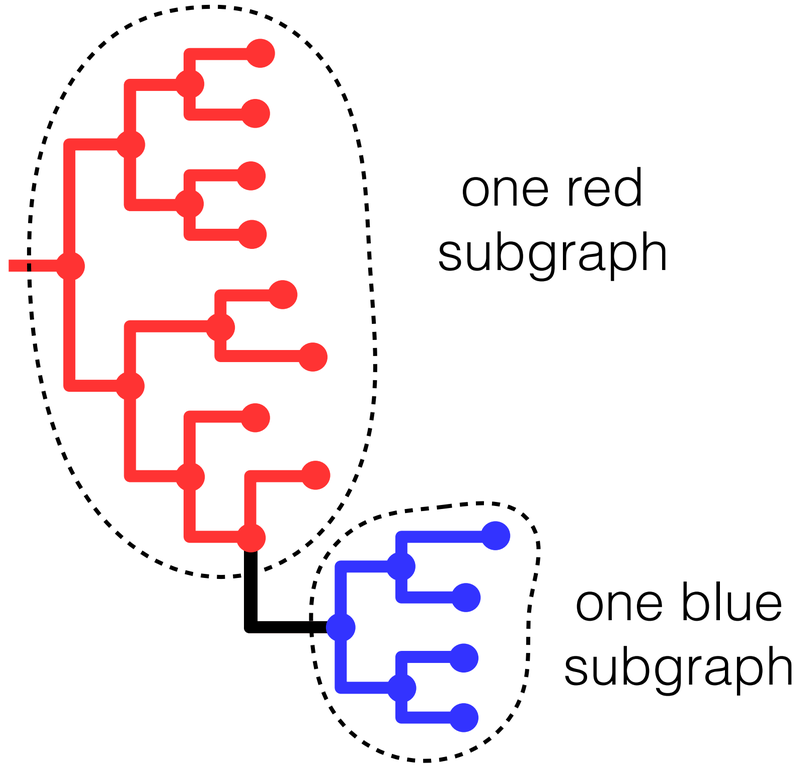


Blue evolved from red: 1 change of state needed.

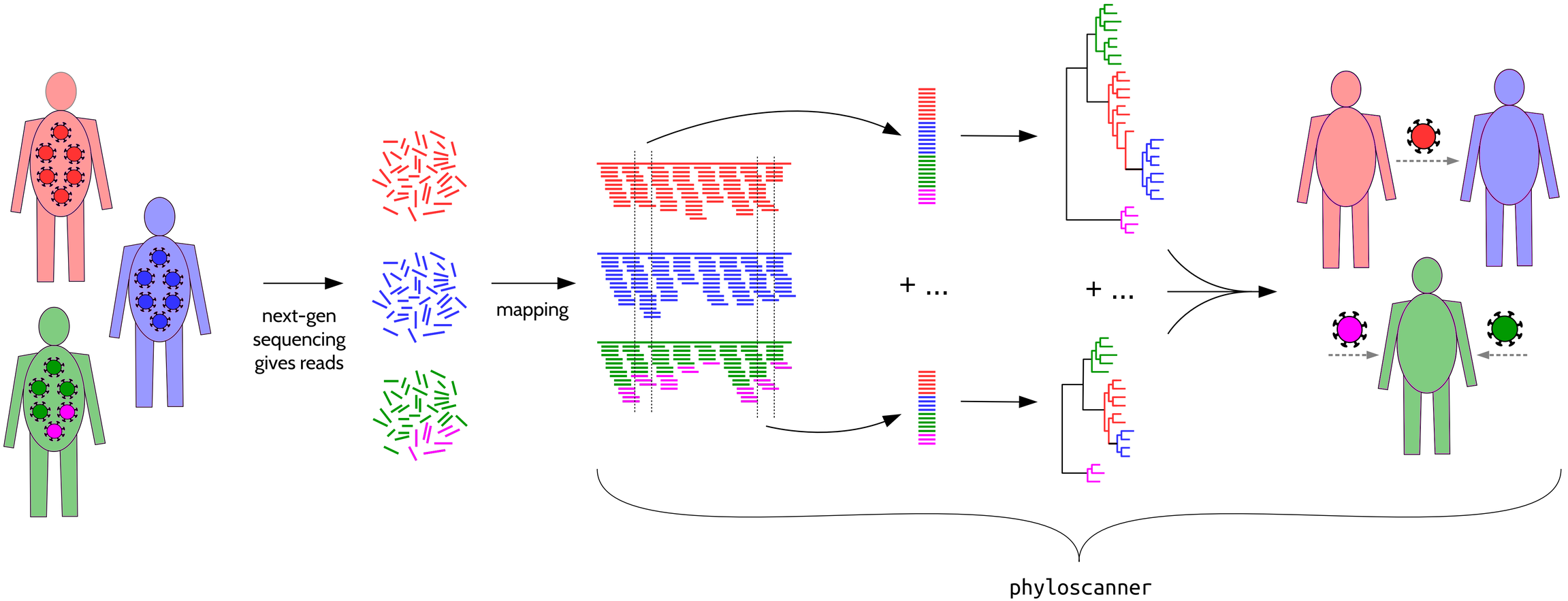
e.g. colour for “which person was this virus in?” - Romero-Severson et al. PNAS 2016

Ancestral State Reconstruction: subgraphs



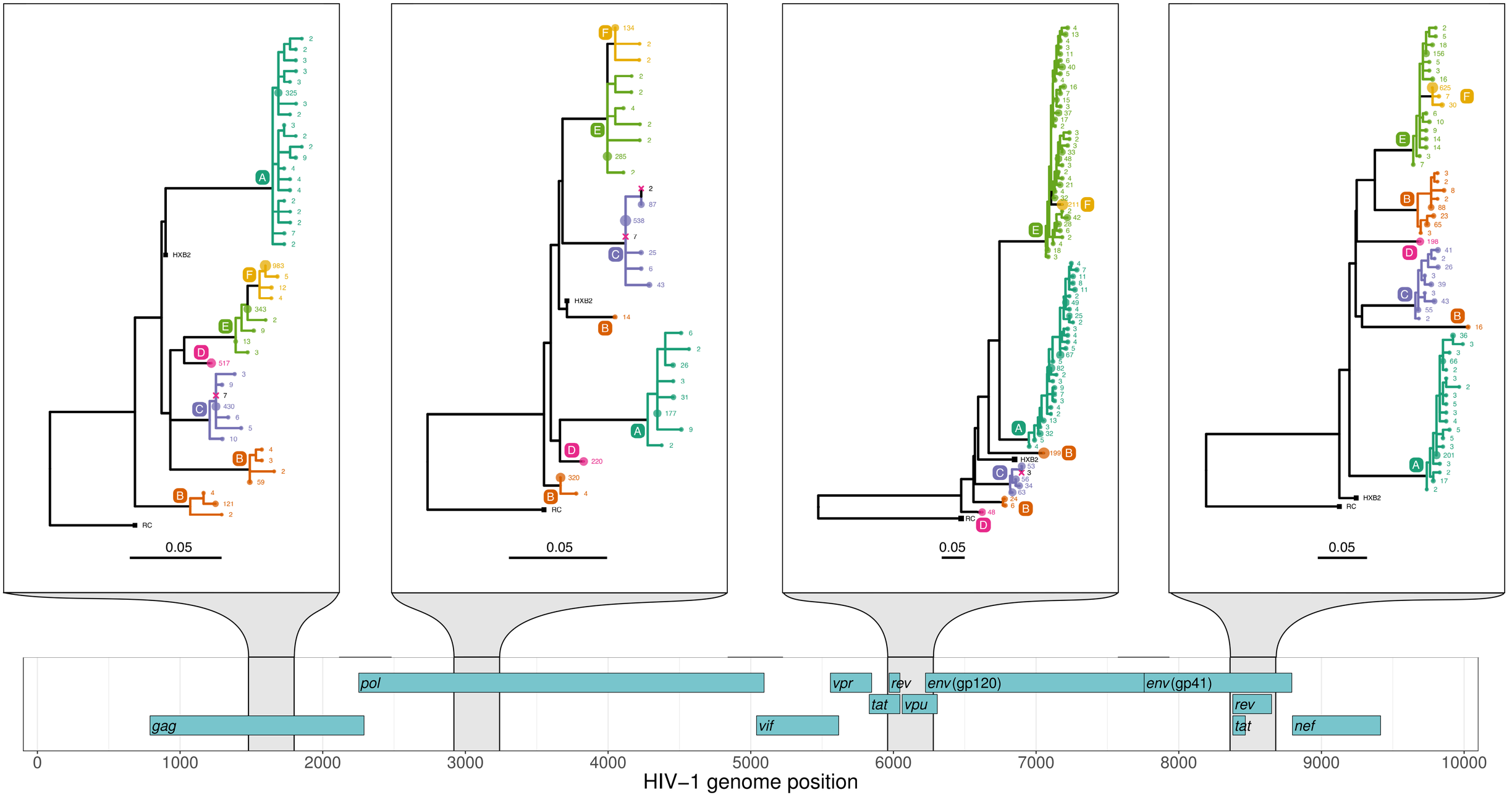


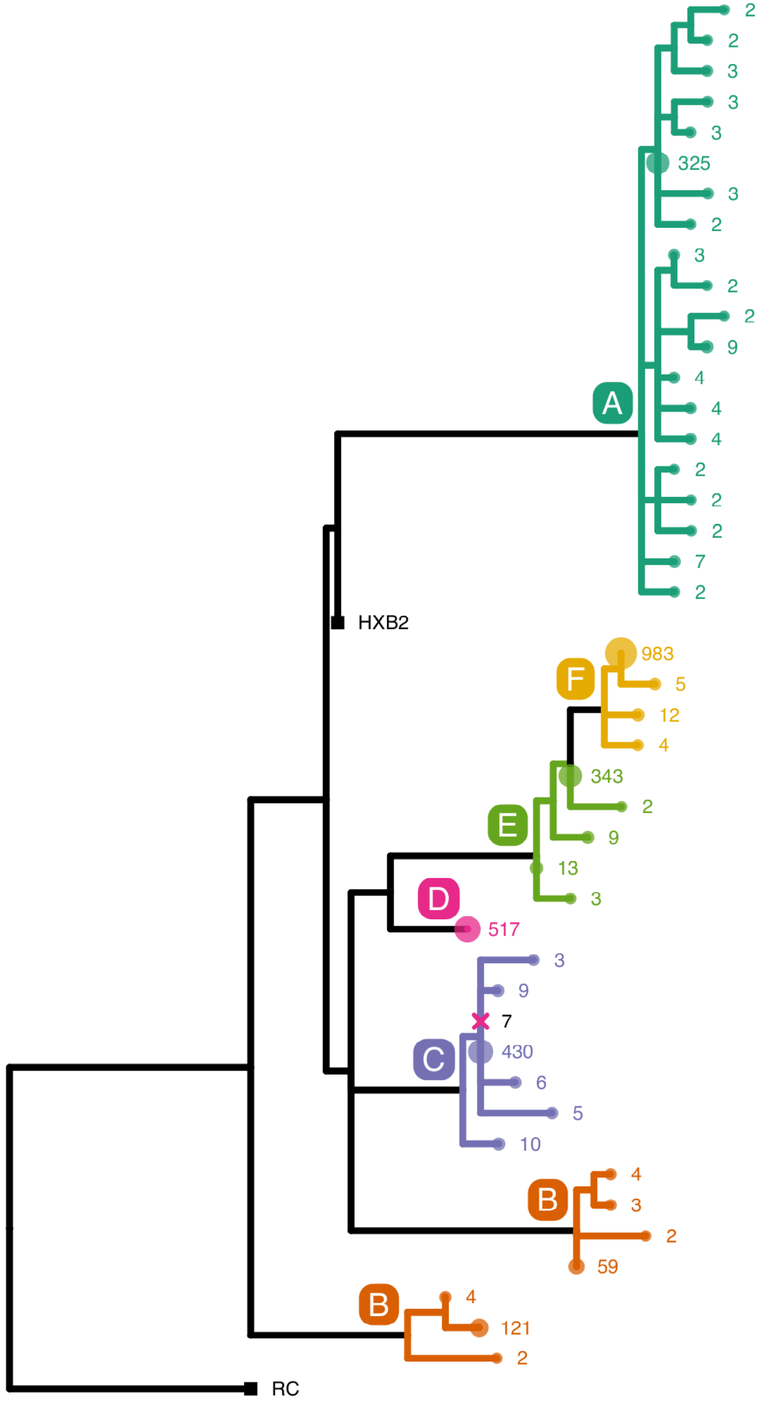
A *subgraph* is connected/continuous region of the phylogeny.
Later, we'll define these regions by the state assigned to them,
i.e. one subgraph = one solid block of colour.



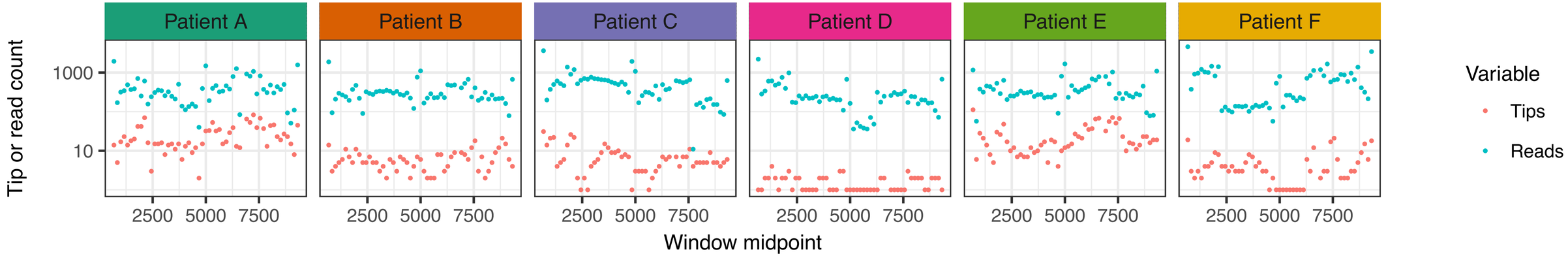
Wymant, Hall *et al.,* MBE 2017

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

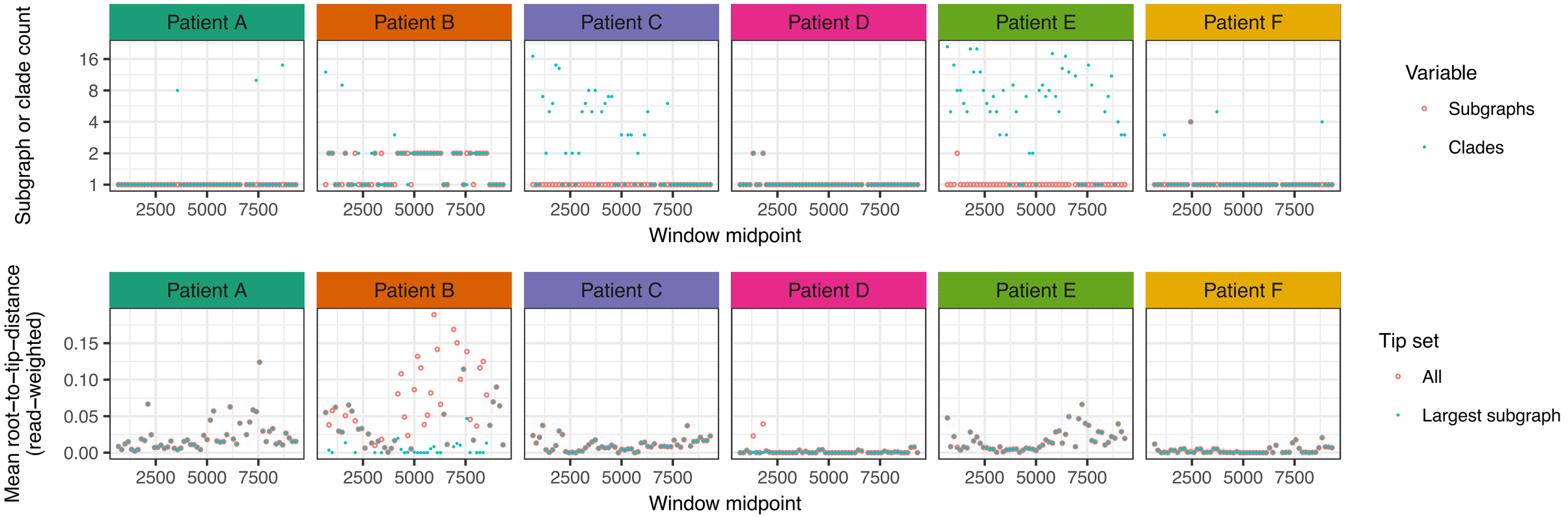

A
2
2
3
3
3
325
3
2
3
2
2
9
4
4
4
2
2
2
7
2
HXB2
F
983
5
12
4
E
343
2
9
13
3
D
517
C
3
9
7
430
6
5
10
B
4
3
2
59
B
4
121
2
RC

Per-patient summary statistics along the genome

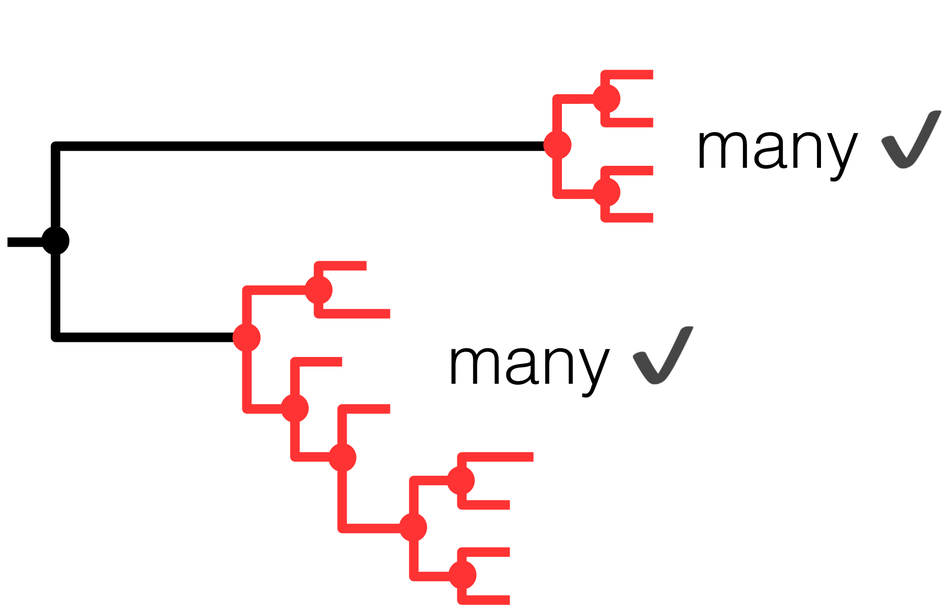


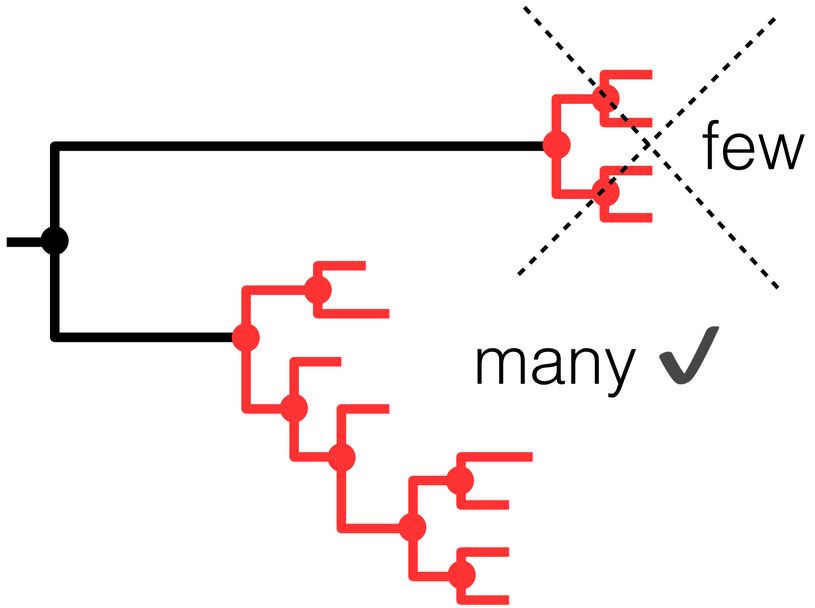
100-1000 reads,
10-100 unique reads.
Diverse!

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

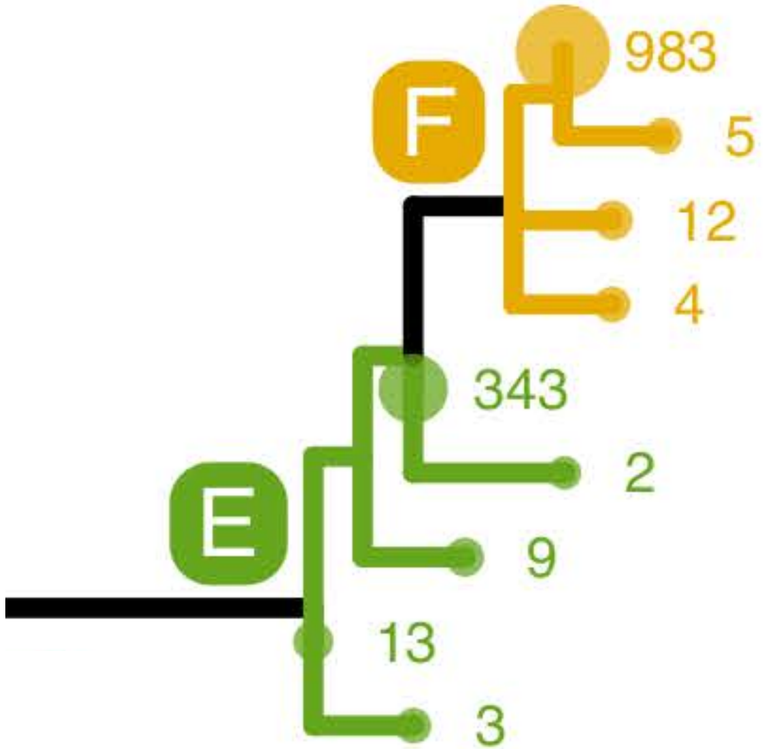

Patient A
Patient B
Patient C
Patient D
Patient E
Patient F
Subgraph or clade count
16
8
4
2
1
2500 5000 7500
Window midpoint
Mean root–to–tip–distance (read–weighted)
0.15
0.10
0.05
0.00
Variable
Subgraphs
Clades
Tip set
All
Largest subgraph

Dually infected individual or contaminated sample?

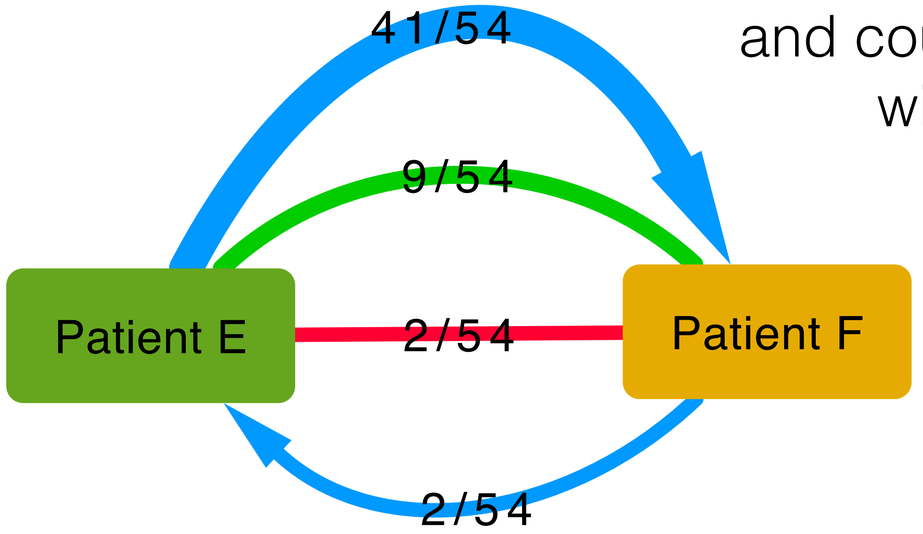


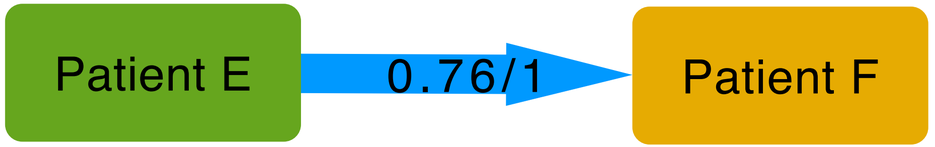


Transmission

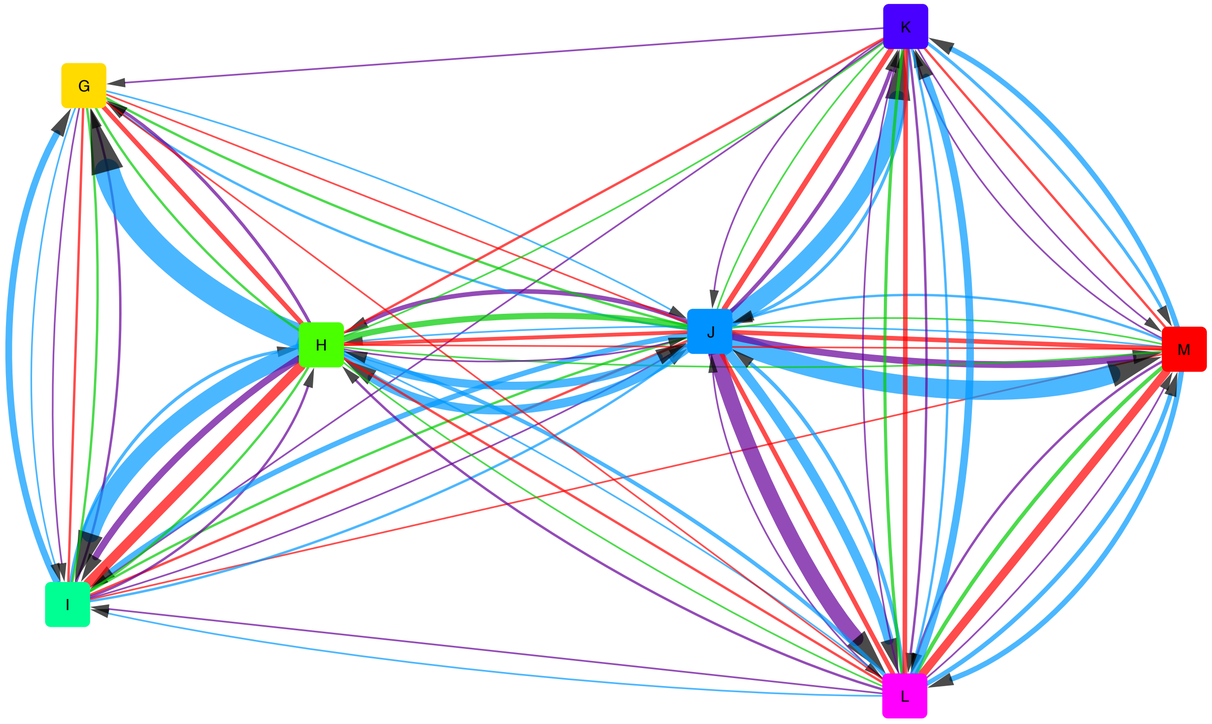


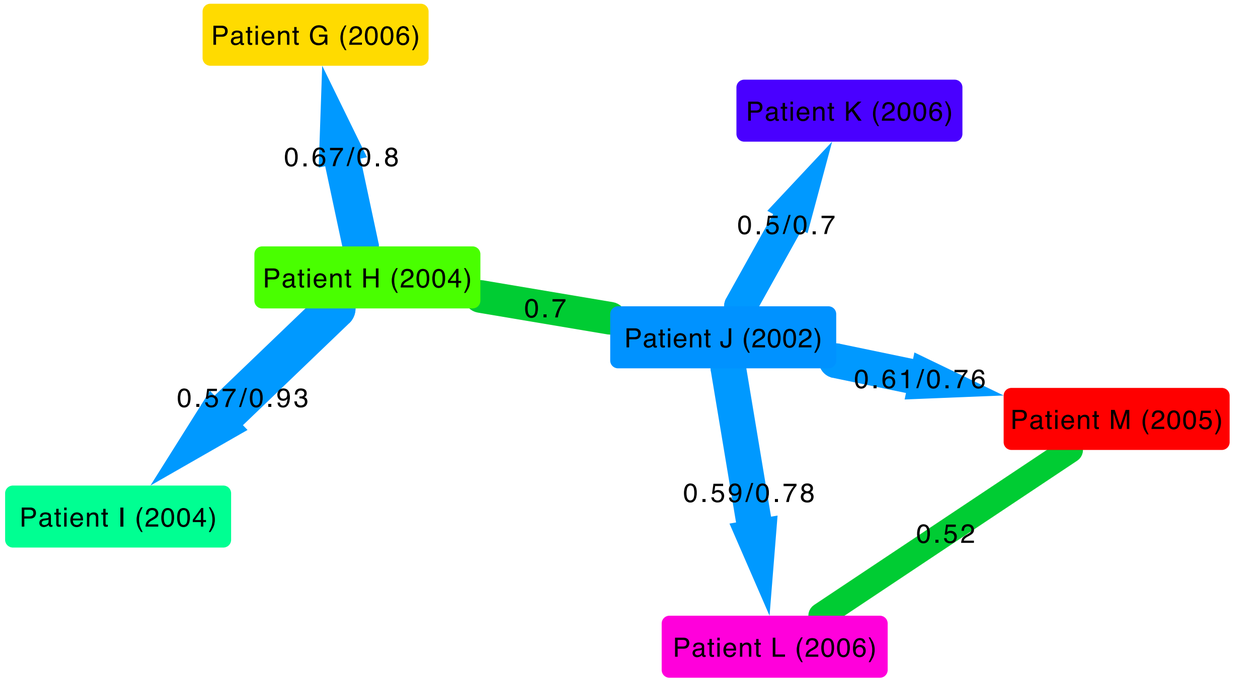
Summary: categorise and count genomic windows

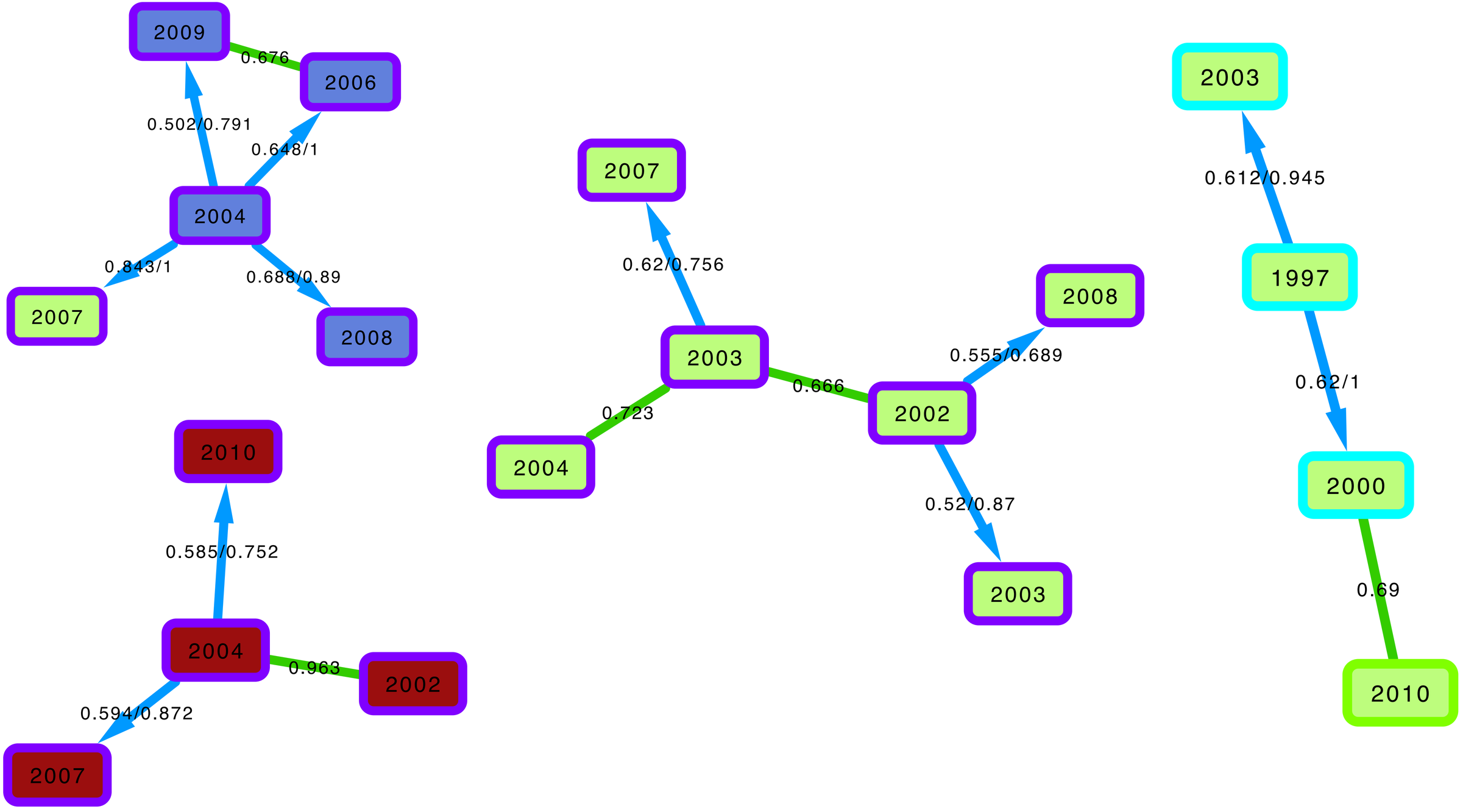




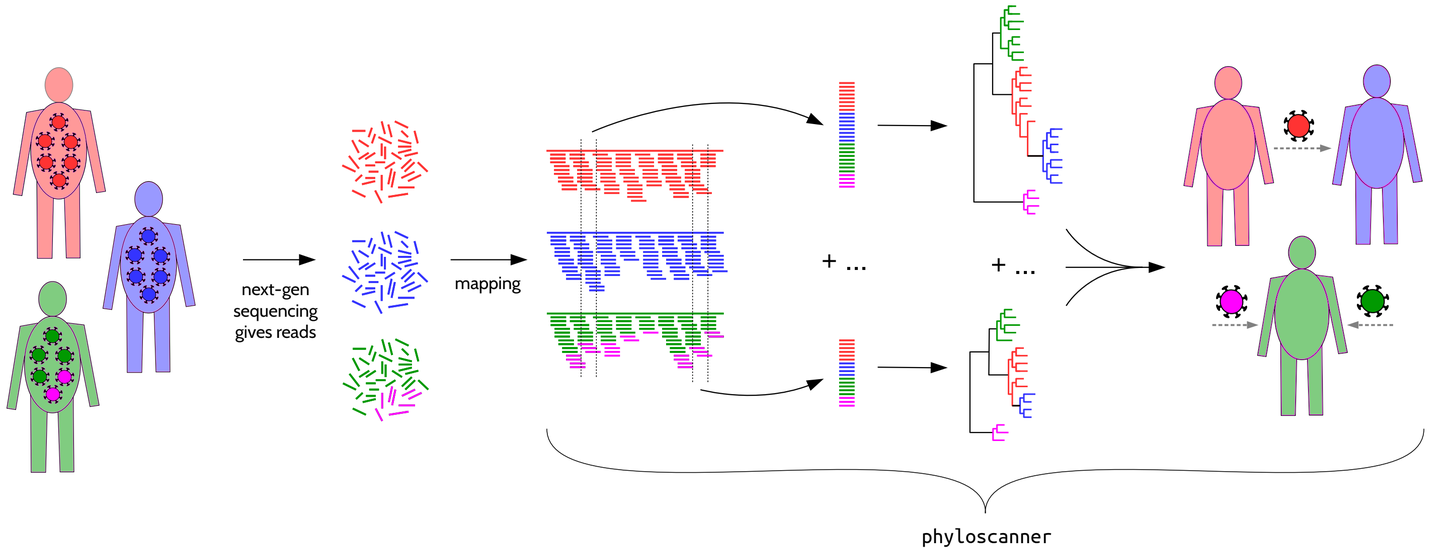
Simpler summary: aggregate categories


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)


2009
0.676
2006
0.502/0.791
0.648/1
2004
0.843/1
0.688/0.89
2007
2008
2010
0.585/0.752
2004
0.963
2002
0.594/0.872
2007
2007
0.62/0.756
2003
0.666
2008
0.555/0.689
0.723
2004
2002
0.52/0.87
2003
2003
0.612/0.945
1997
0.62/1
2000
0.69
2010

Summary



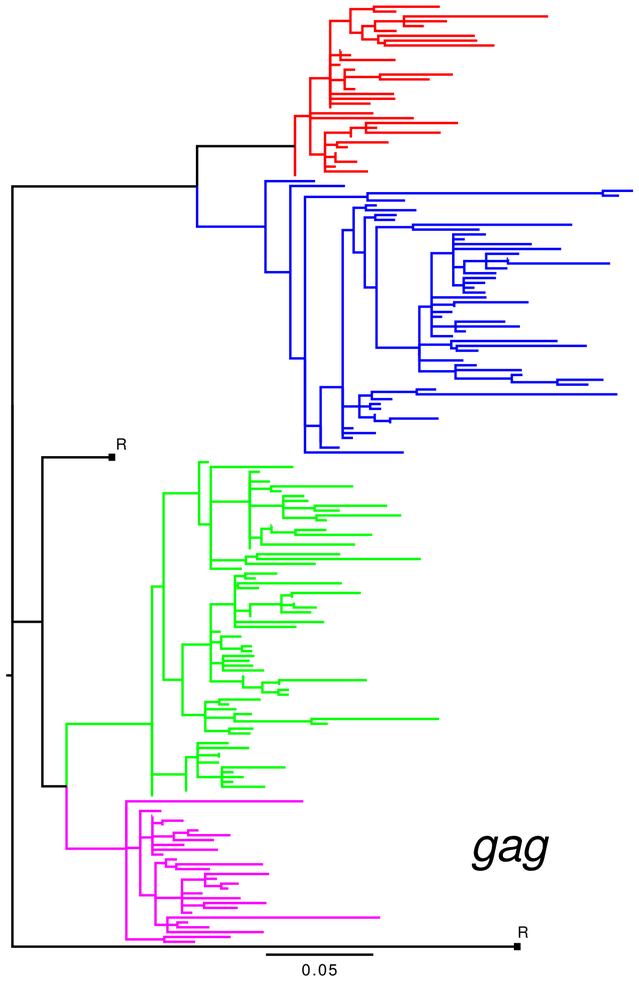
Wymant, Hall *et al.* MBE 2017
GitHub.com/BDI-pathogens/phyloscanner

`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**

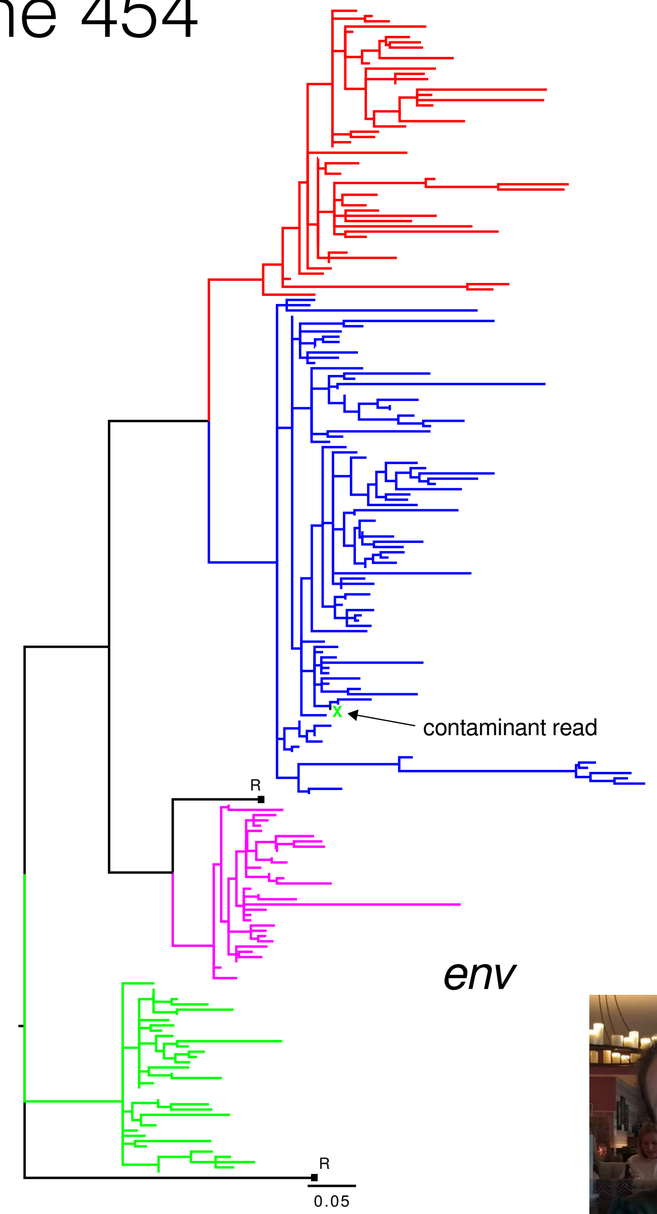
Extra slides: application to other sequencing platforms and pathogens, and using `phyloscanner`

Four more BEEHIVE patients sequenced with Roche 454



gag

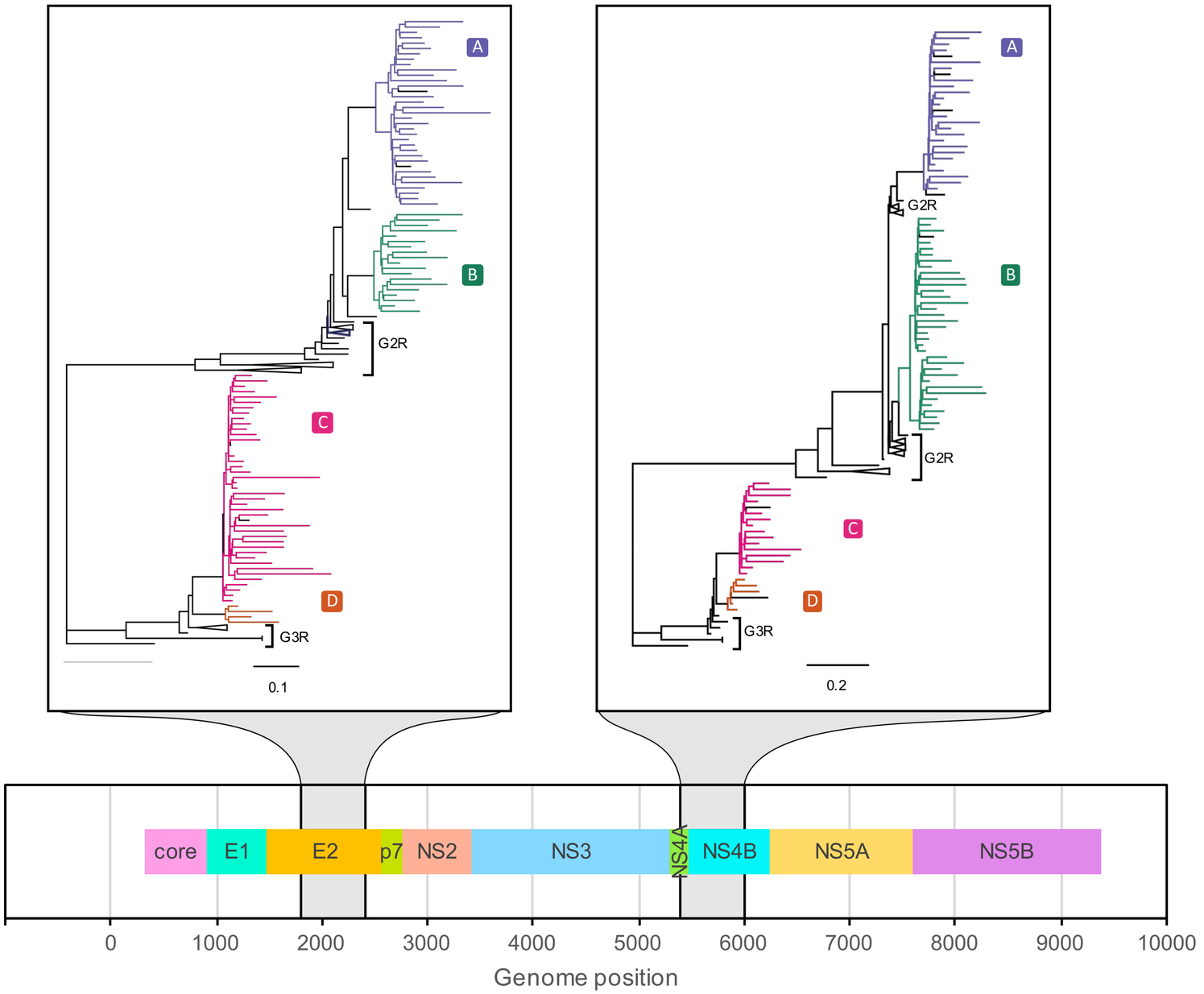
pol



env

Tanya Golubchik

Hepatitis C Virus Sequenced With Oxford Nanopore

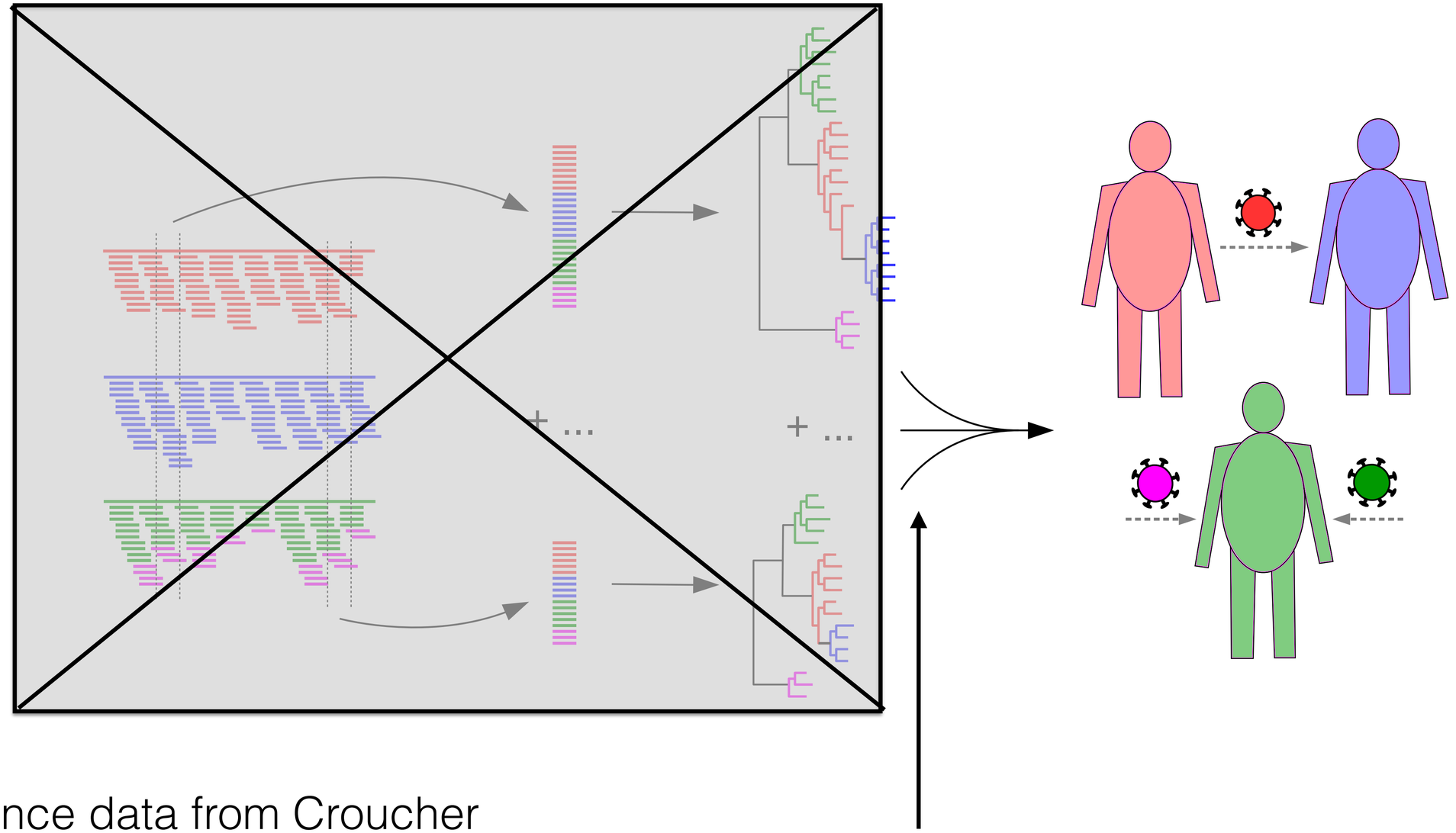


The Stop-HCV Consortium:

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

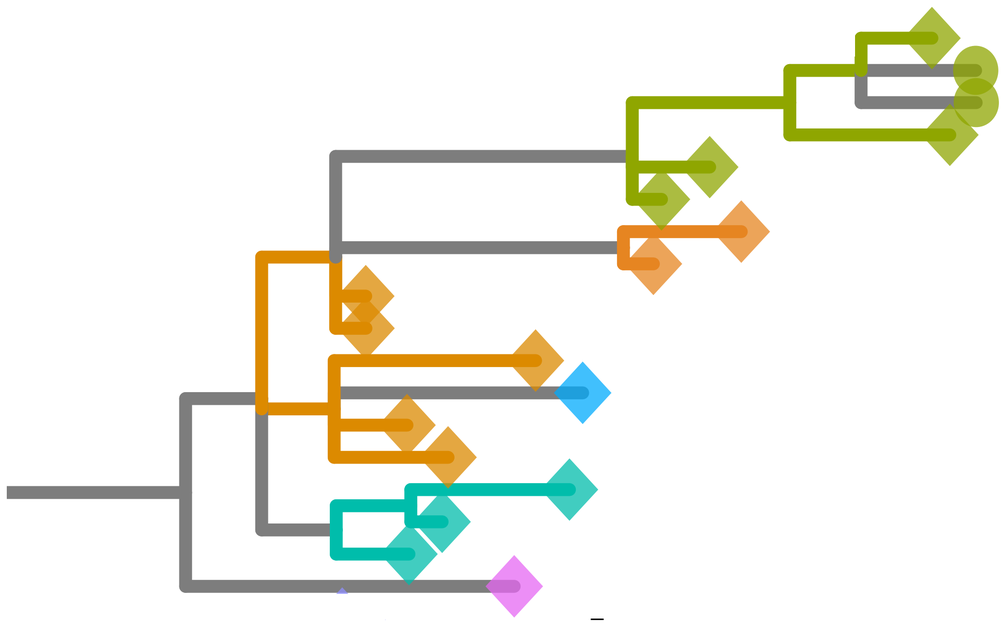


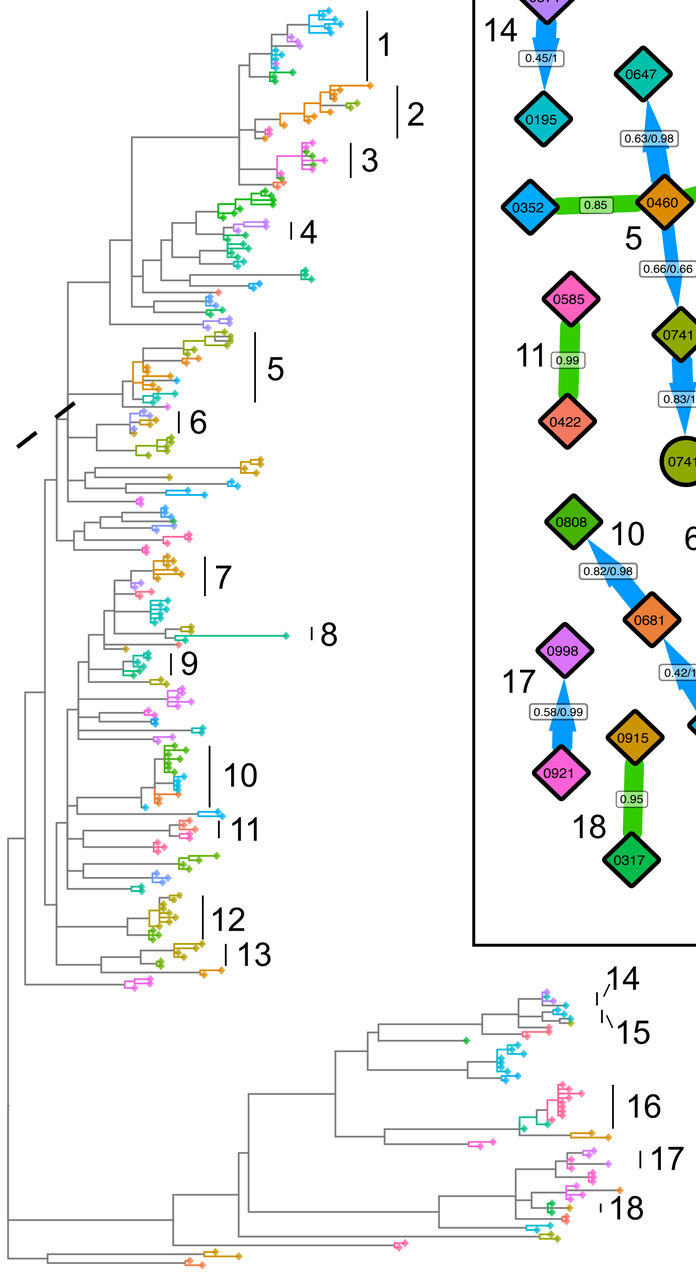
Sequence data from Croucher *et al.* PLoS Bio. 2016: multiple colony picks per carrier of *S. pneumoniae*

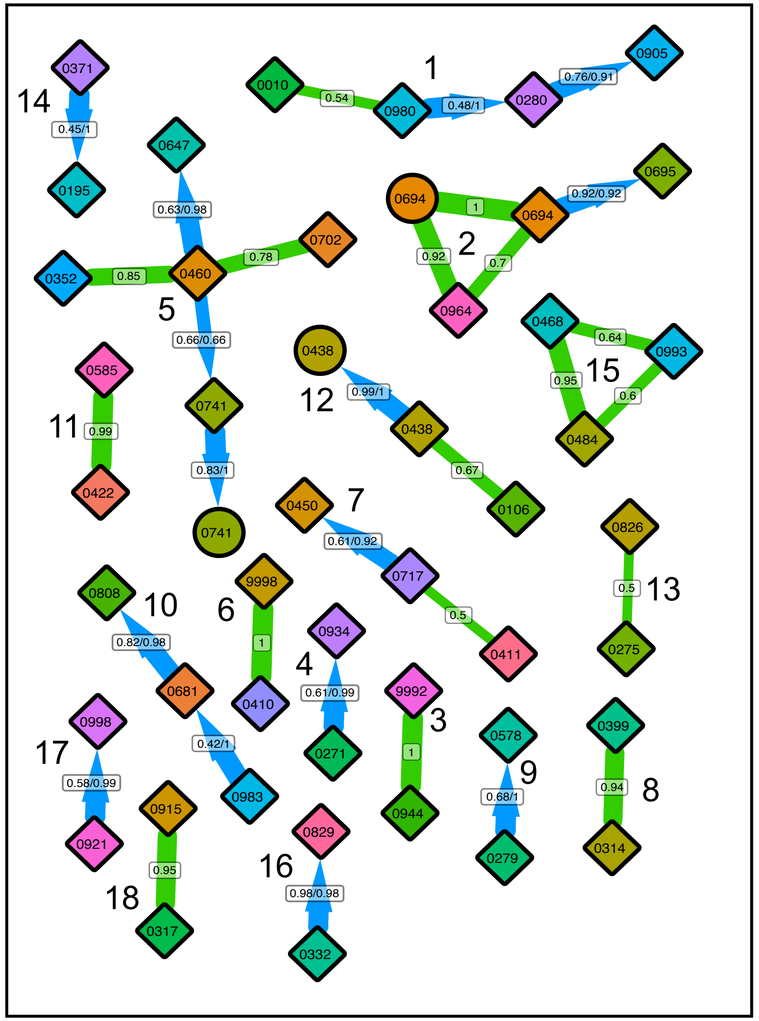
100 posterior trees with `MrBayes`

The Maela Pneumococcal Collaboration:

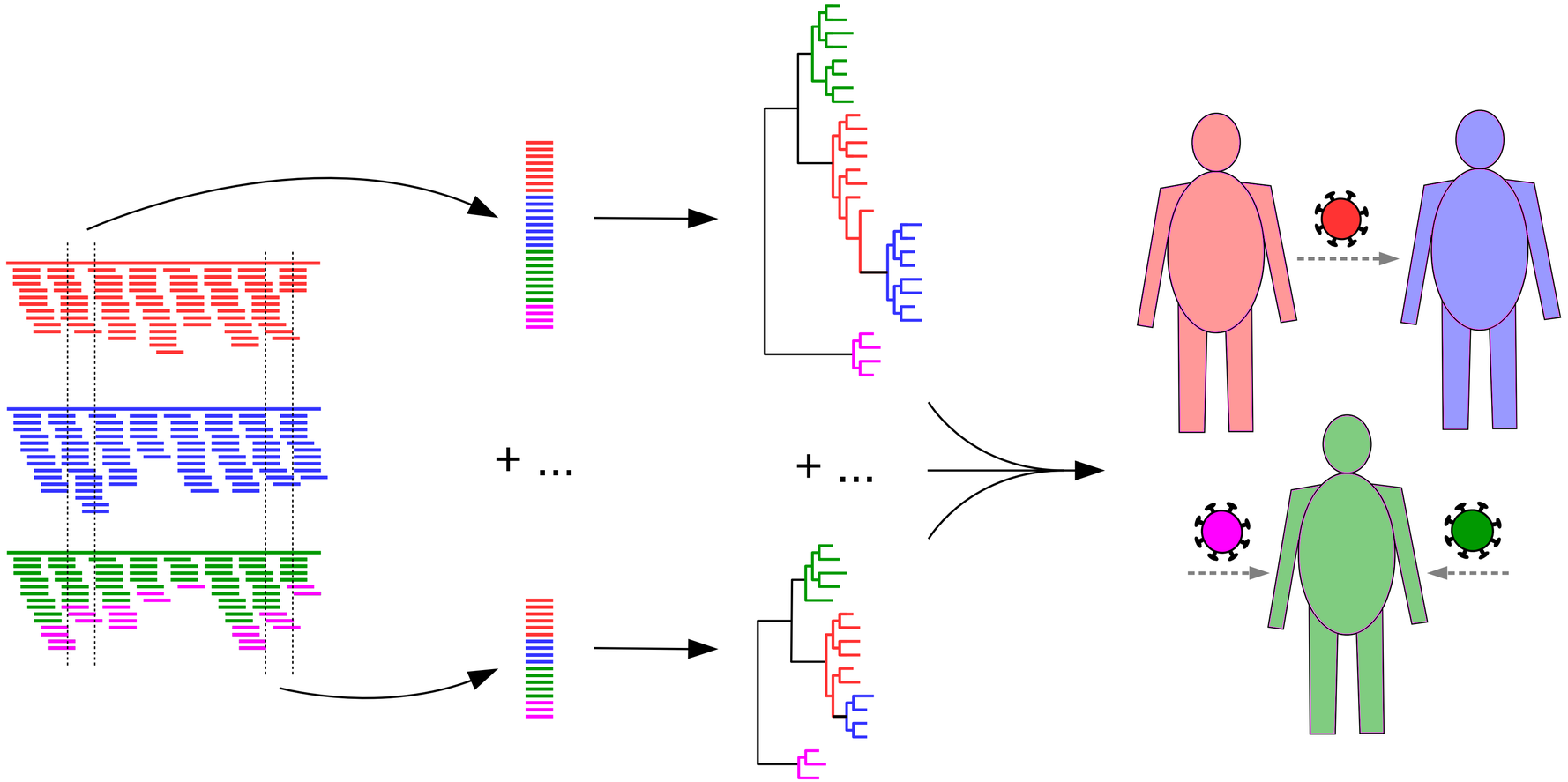
Stephen Bentley
Claire Chewapreecha
Nicholas J. Croucher
Simon Harris
Jukka Corander
David Goldblatt
Julian Parkhill
Francois Nosten
Claudia Turner
Paul Turner







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

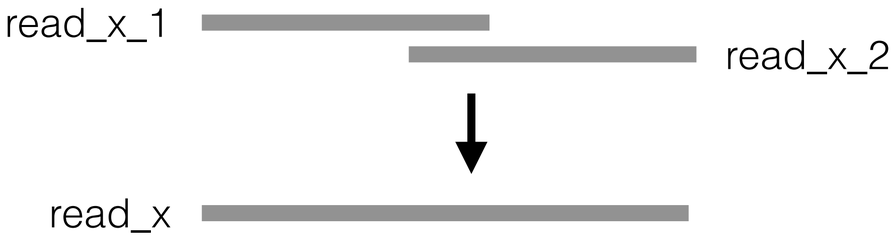


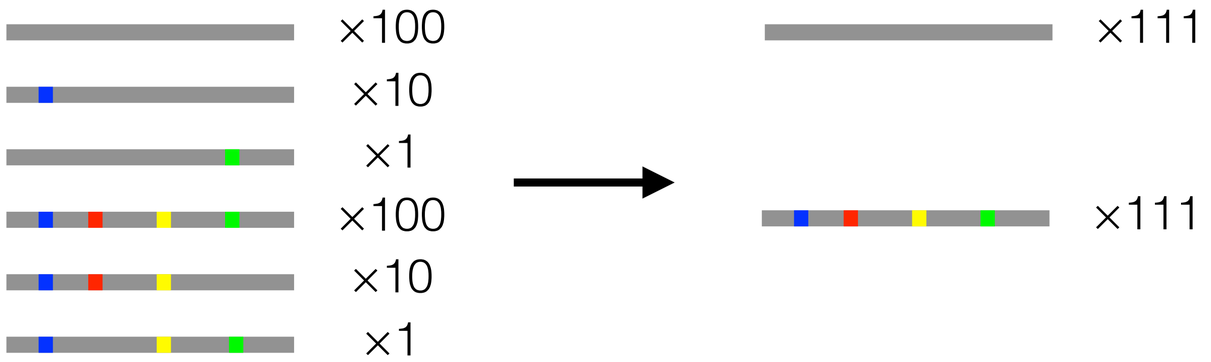

```
$ phyloscanner_make_trees.py  ListOfBamFiles.csv  --windows  1,300,301,600,...

$ phyloscanner_analyse_trees.R  TreeFiles  OutputString  ChoiceOfHostStateReconstruction
```

Options

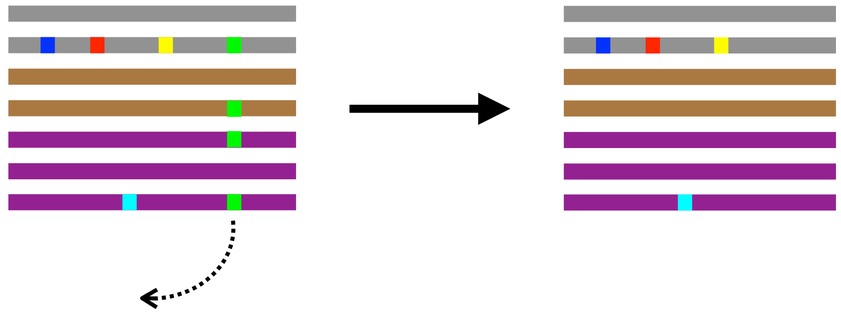
Merge overlapping paired reads into longer reads.





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**).