Phylogenetics Between and Within: Seeing Transmissions and Dual Infections in HIV Deep Sequence Data

Chris Wymant[1,2], Matthew Hall[1,2], François Blanquart[1], Christophe Fraser[1,2], and the BEEHIVE Collaboration

[1] Department of Infectious Disease Epidemiology, Imperial College London, UK

[2] (from 1st July) Oxford Big Data Institute, Nuffield Department of Medicine, University of Oxford, UK

Mathematical and Computation Evolutionary Biology, June 2016: Forecasting Evolution

Notes on the talks:
ChrisWymant on Twitter

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

- Whole viral genome data from >4000 patients.
- Seroconverters: 1 year window between – and + tests, or clinical evidence
- ≥ 1 viral load measurement 6-24 months after + test, pre-ART
- Wealth of clinical data (including later response to ART)

Objectives:

1°: the viral-molecular basis of virulence

2°: molecular epidemiology

3°: dual infections and minority variants

Imperial College London
François Blanquart
Christophe Fraser
Matthew Hall
Frank de Wolf
Chris Wymant
Amsterdam Medical Centre
Margreet Bakker
Ben Berkhout
Marion Cornelissen
Peter Reiss
Wellcome Trust Sanger Institute
Martin Hunt
Astrid Gall
Paul Kellam
Swee Hoe Ong

HIV Monitoring Foundation
Daniela Bezemer
Mariska Hillebregt
Ard van Sighem
Sima Zaheri
Karolinska University Hospital
Jan Albert
Anders Sönnerborg
Oslo University Hospital
Anne-Marte Bakken Kran
Anne Margarita Dyrhol Riise
Antwerp Institute of Tropical Medicine
Katrien Fransen
Guido Vanham
John Hopkins University
M. Kate Grabowski

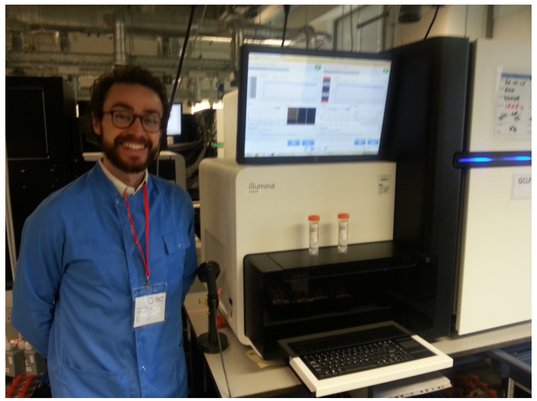
Robert Koch-Institute, Berlin
Barbara Gunsenheimer-Bartmeyer
Claudia Kücherer
University Hospital Zürich
Huldrych Günthard
Roger Kouyos
Laboratory of Immunoregulation
NIAID NIH, Baltimore
Oliver Laeyendecker
Helsinki University Hospital
Kirsi Liitsola
Matti Ristola
Université Paris Sud
Laurence Meyer
University College London
Dan Frampton
Kholoud Porter

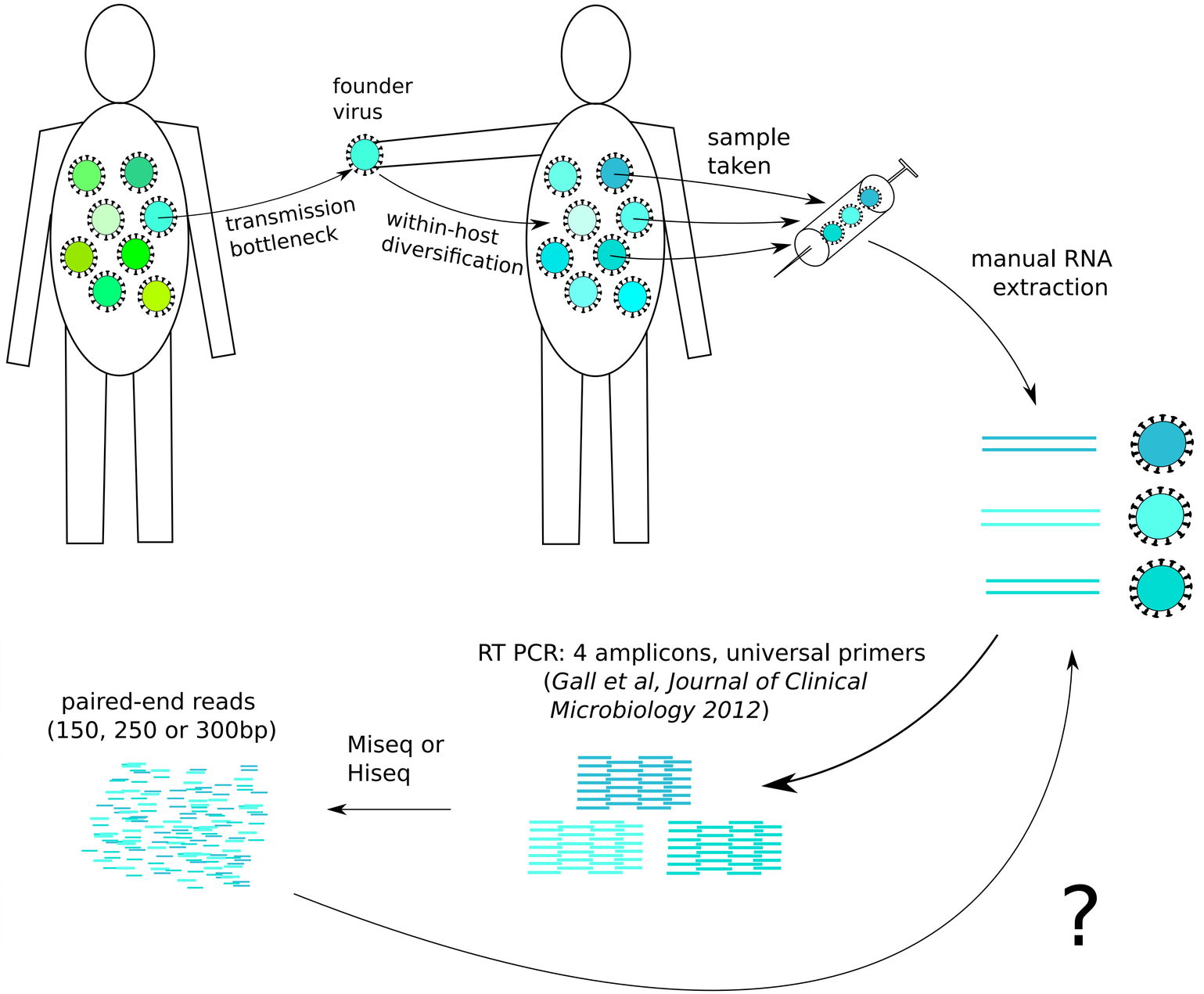
Erasmus Medical Centre, Rotterdam
David van de Vijver
École polytechnique fédérale de Lausanne
Jacques Fellay
Istvan Bartha
Analysis Advisory Group
Sebastian Bonnhoeffer
Gabriel Leventhal
Samuel Alizon
Andrew Rambaut
Oliver Pybus
Gil McVean
. . .and thanks to
Tiziano Gallo Cassarino
Nick Croucher
Katrina Lythgoe
Olli Ratmann

erc

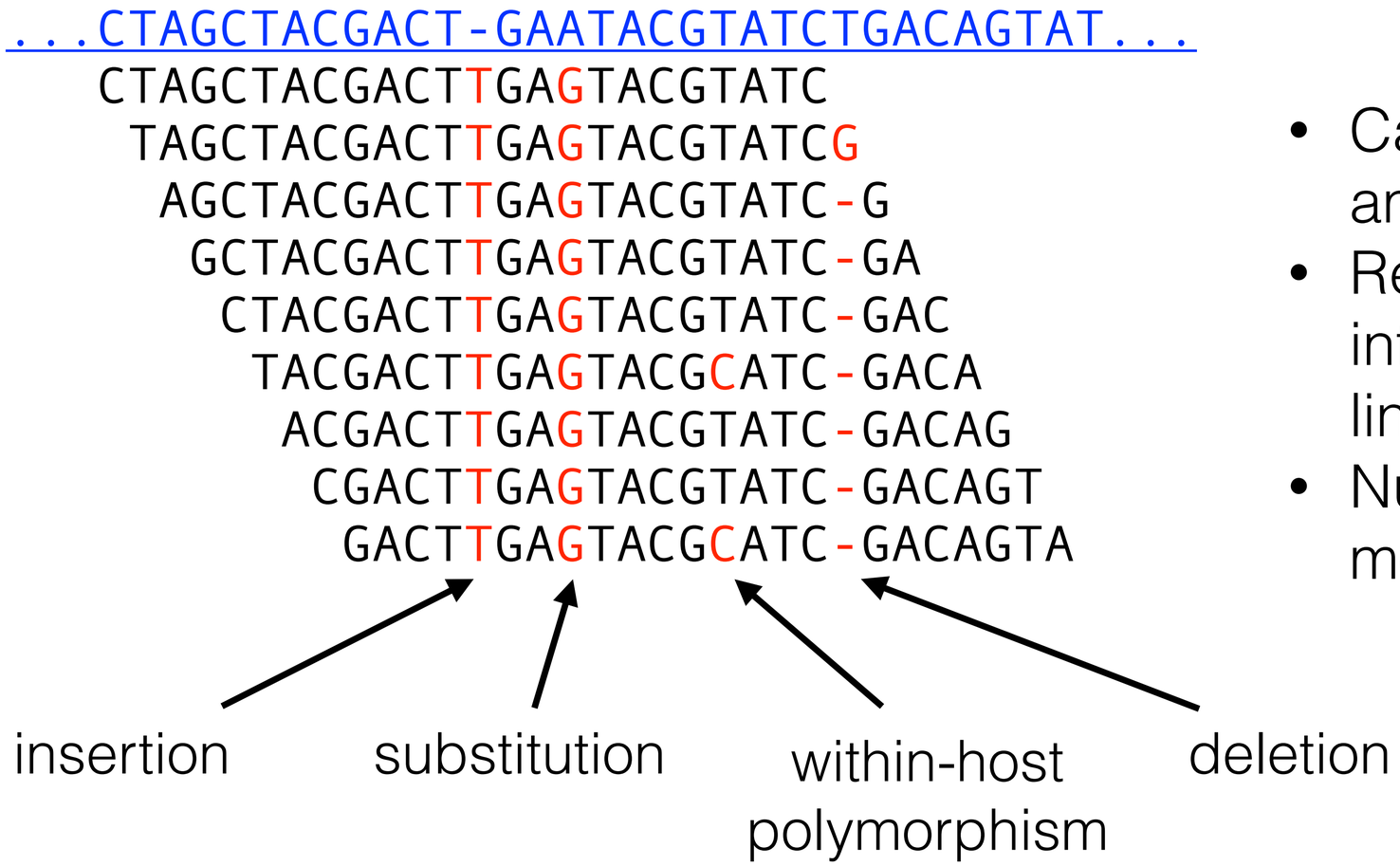
First: reconstructing the viral genotype

(Next: putting it in context.)



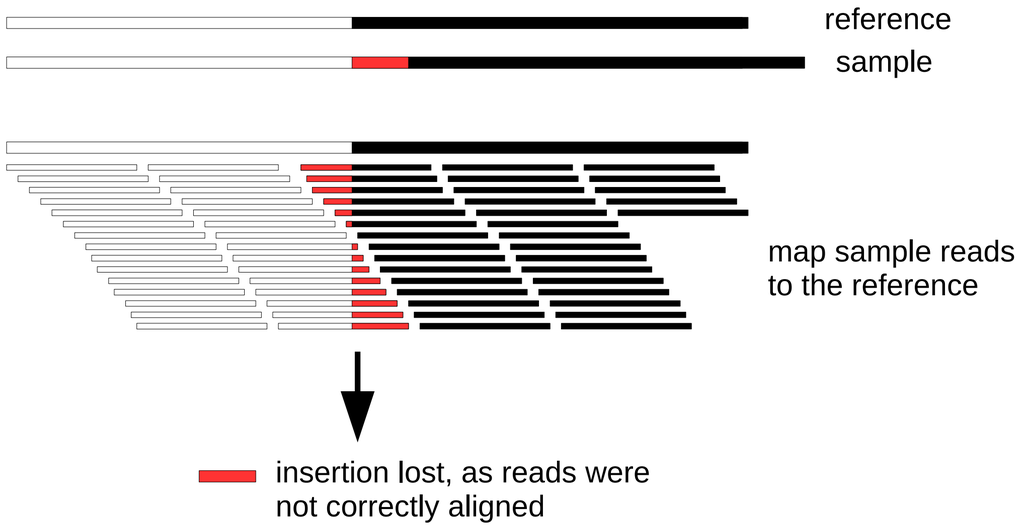

founder
virus
transmission
bottleneck
within-host
diversification
sample
taken
manual RNA
extraction
RT PCR: 4 amplicons, universal primers
(*Gall et al, Journal of Clinical Microbiology 2012*)
Miseq or
Hiseq
paired-end reads
(150, 250 or 300bp)
?

Map Reads to a Known Reference…



- Call the most common base amongst the reads at each position
- Retains within-host diversity information (base frequencies and linkage)
- Number of reads gives some measure of confidence

…but Mapping Induces Bias



The more different a read is from its reference, the more likely it is to be aligned incorrectly or not at all.

↓

Systematic bias to make new sequences resemble known examples. An extremely high mutation rate makes this an acute problem for HIV.

Real data example: what the mapper should have done:

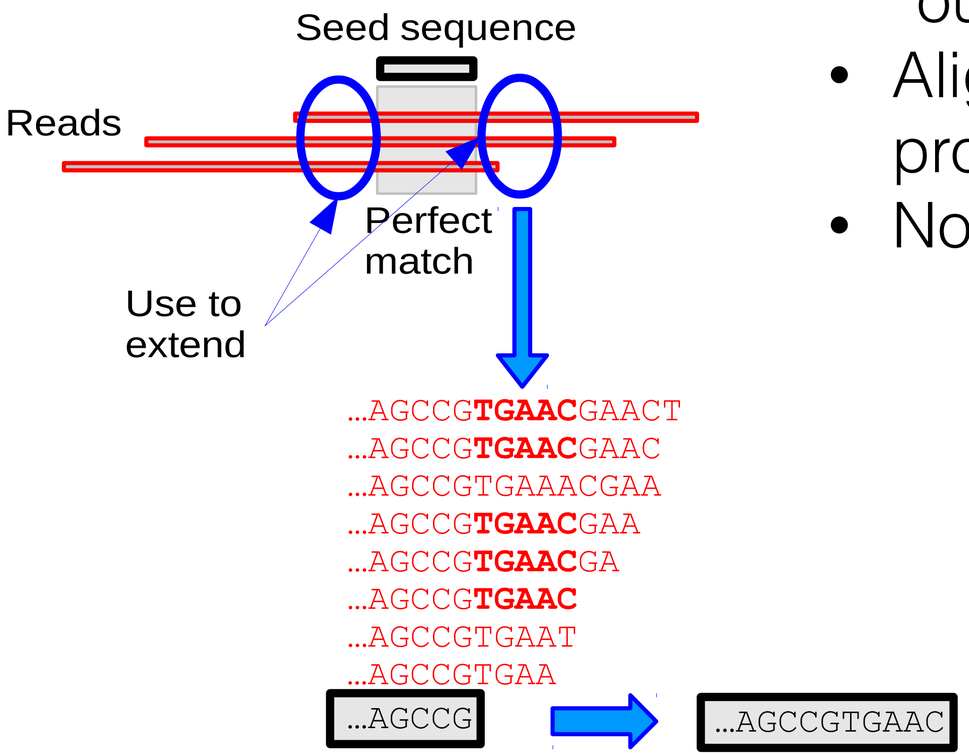
```
reference AAGTGTTAGTGTGGAAGTTTGACAGCAGCCTAGCACTTCATCACAGGGCCCGAGAGCAACATCCGGAGTTTTACAAAGACTGCTGACATC--------------------GAGTTTCCTACAAGGGACTTTCCGCTGGGGACTTTCCAGGGAAGGCGTGGCCTGGGCGGG
read 1    AAGTGTTAATGTGGAAGTTTGACAGCCGCCTAGCATTCCATCACGTAGCCCGAGAGCTGCATCCGGAGTACTACAAAGACTGCTGACATCCTACAAAGACTGCTGACATCGAGCTTCTGCAAAGGGACTTTCCGCTGGGGACTTTCCAGG
read 2                      GACAGCCGCCTAGCATTCCATCACGTAGCCCGAGAGCTGCATCCGGAGTACTACAAAGACTGCTGACATCCTACAAAGACTGCTGACATCGAGCTTCTGCAAAGGGACTTTCCGCTGGGGACTTTCCAGGGGAGGCGTGGCCTGGGCGGG
```

What it did:

```
reference AAGTGTTAGTGTGGAAGTTTGACAGCAGCCTAGCACTTCATCACAGGGCCCGAGAGCAACATCCGGAGTTTTACAAAGACTGCTGACATCGAGTTTCCTACAAGGGACTTTCCGCTGGGGACTTTCCAGGGAAGGCGTGGCCTGGGCGGG
read 1    AAGTGTTAATGTGGAAGTTTGACAGCCGCCTAGCATTCCATCACGTAGCCCGAGAGCTGCATCCGGAGTACTACAAAGACTGCTGACATCCTACAAAGACTGCTGACATCGAGCTTCTGCAAAGGGACTTTCCGCTGGGGACTTTCCAGG
read 2    GACAGCCGCCTAGCATTCCATCACGTAGCCCGAGAGCTGCATCCGGAGTACTACAAAGACTGCTGACATCCTACAAAGACTGCTGACATCGAGCTTCTGCAAAGGGACTTTCCGCTGGGGACTTTCCAGGGGAGGCGTGGCCTGGGCGGG
```

reads clipped

De Novo Assembly

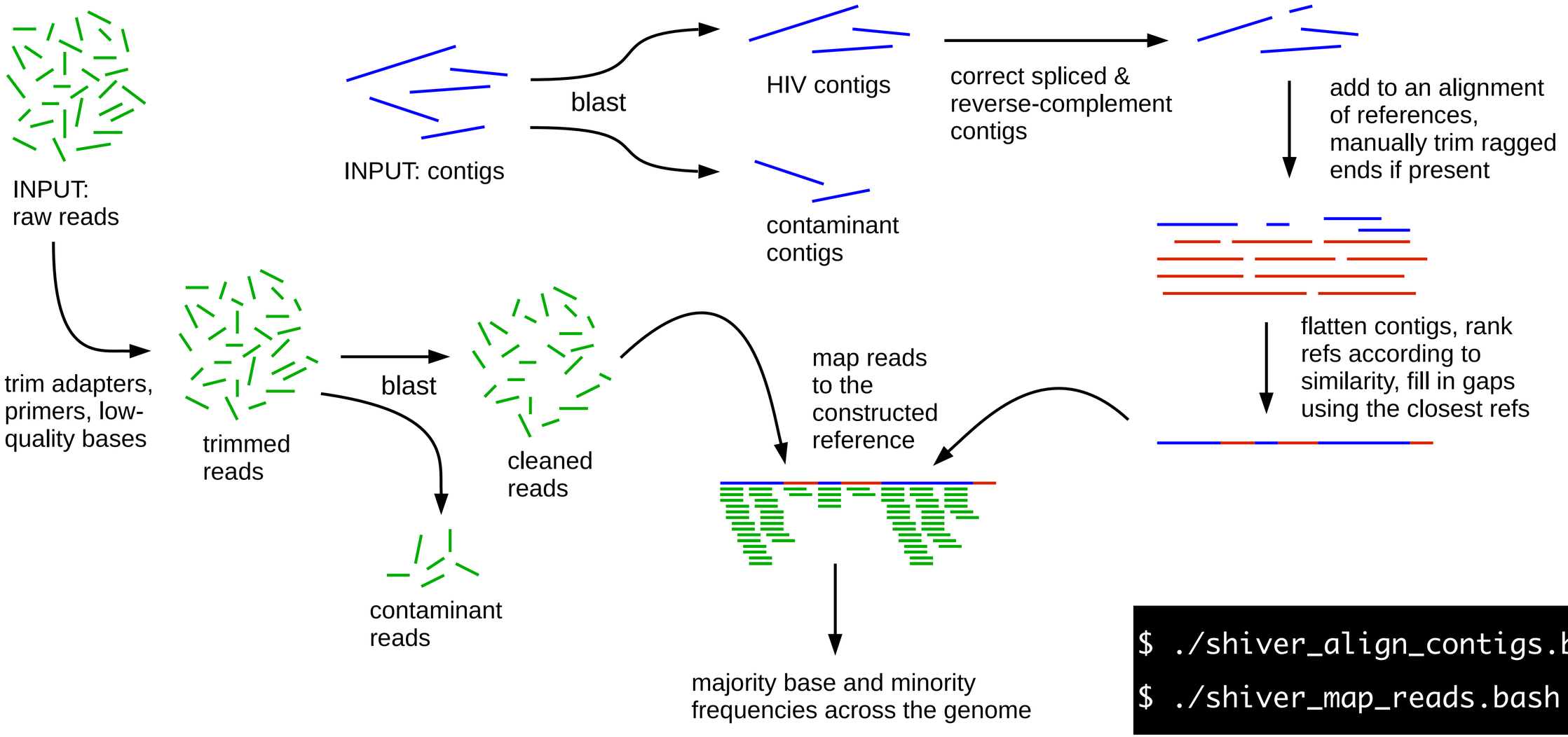


- e.g. IVA (*Hunt et al., Bioinformatics 2015*): on HIV, “outperforms all other virus de novo assemblers”.
- Align the reads to themselves, iteratively extending, producing a number of summary sequences.
- No need for a reference: no bias.

but...

- Diversity & low coverage → no contig
- Diversity & high coverage → two or more overlapping, different contigs. Which one is ‘correct’?
- No idea of base frequencies or linkage
- Occasional assembly errors: splicing separate parts of the genome, reverse complements.

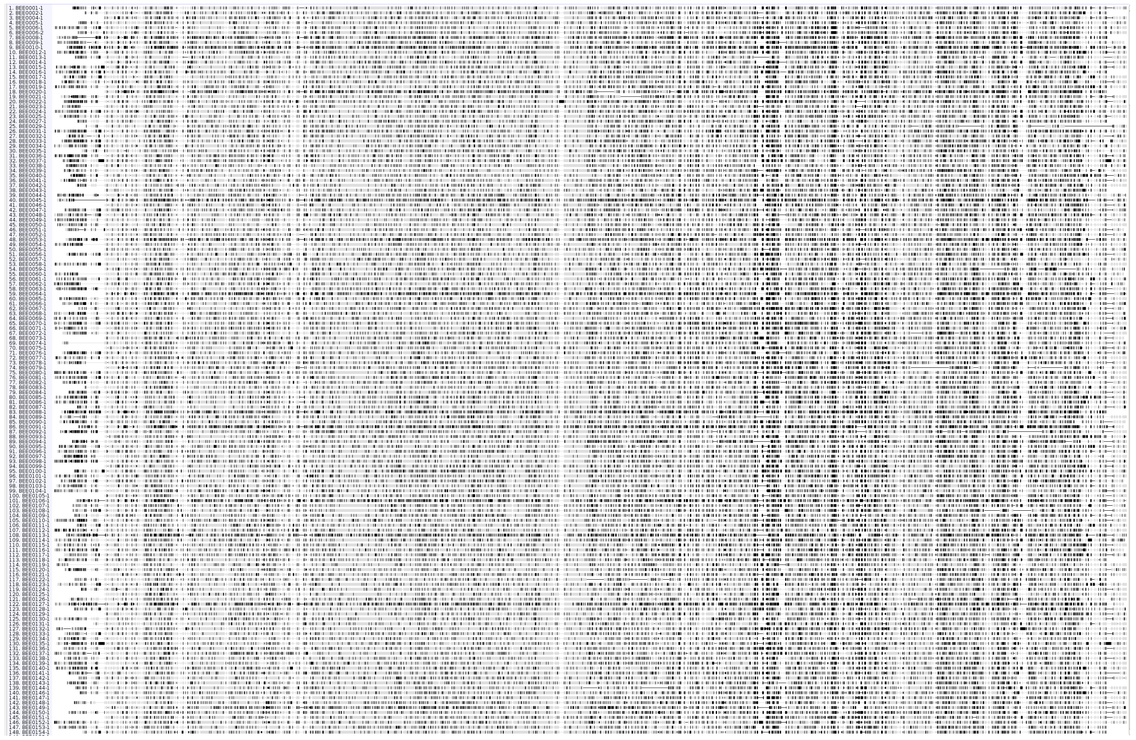
shiver - Sequences from HIV Easily Reconstructed



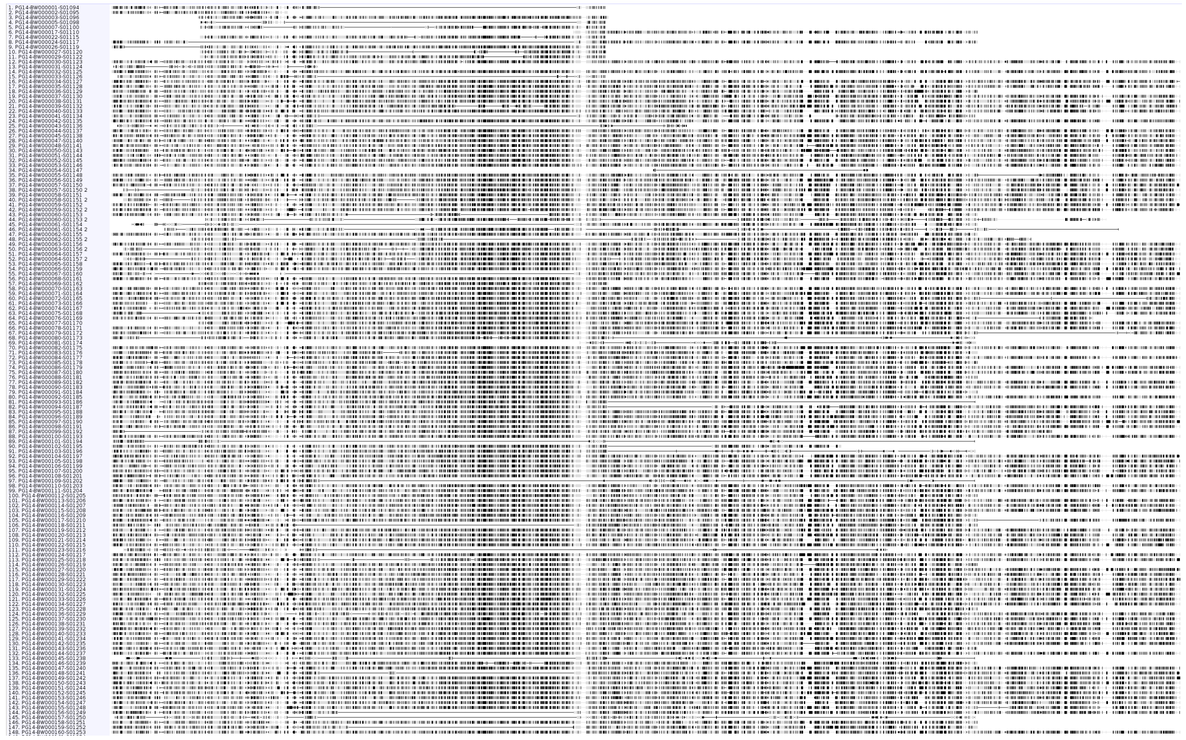

```
$ ./shiver_align_contigs.bash
$ ./shiver_map_reads.bash
```

Used to reconstruct whole* genomes for

BEEHIVE: 1700+ samples



PANGEA: 4300+ samples

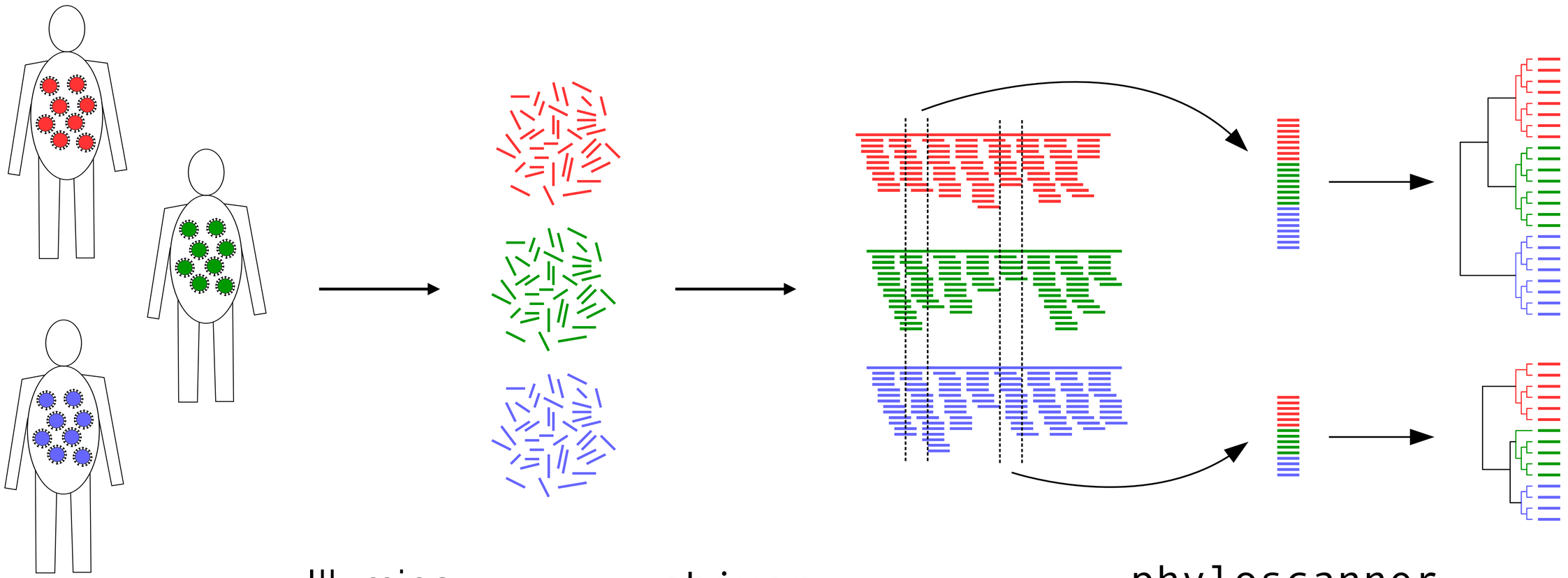


*or not, depending on the sample

A consensus sequence does not show

- **transmission** (or at least not well; important for epidemiology),
- **dual infections** (important clinically, for bioinformatics, for evolution),
- **contamination** (important to distinguish from dual infections),
- **recombination** between strains within-host.

Put within-host diversity in context: examine within- and between-host evolution at once, across the genome.



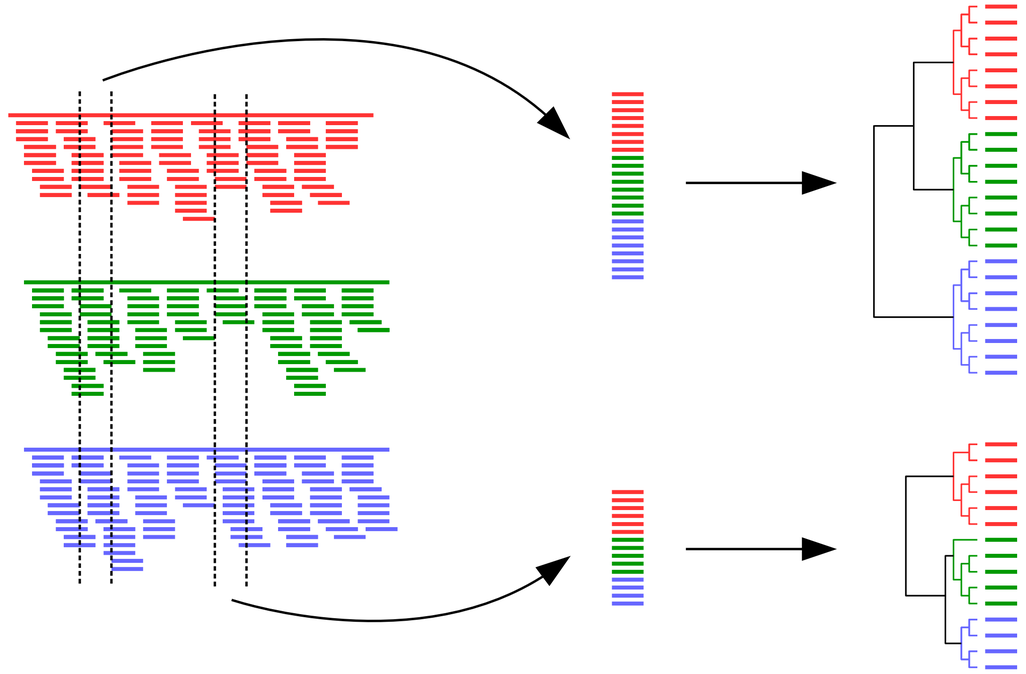
Illumina sequencing

`shiver`

(or your own pipeline)

`phyloscanner`

Easy



```
$ ./phyloscanner.py  MyBamFiles.txt  MyRefFiles.txt  --windows 1,300,200,500,...
```

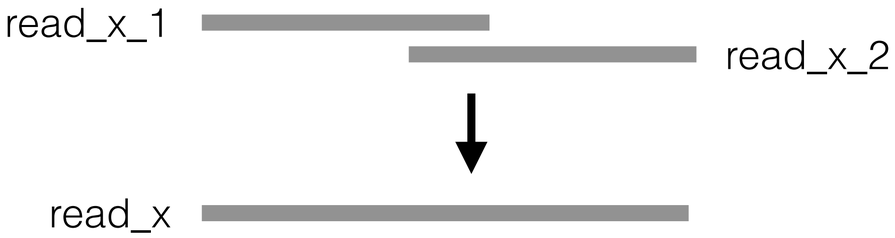
```
red.bam
green.bam
blue.bam
```

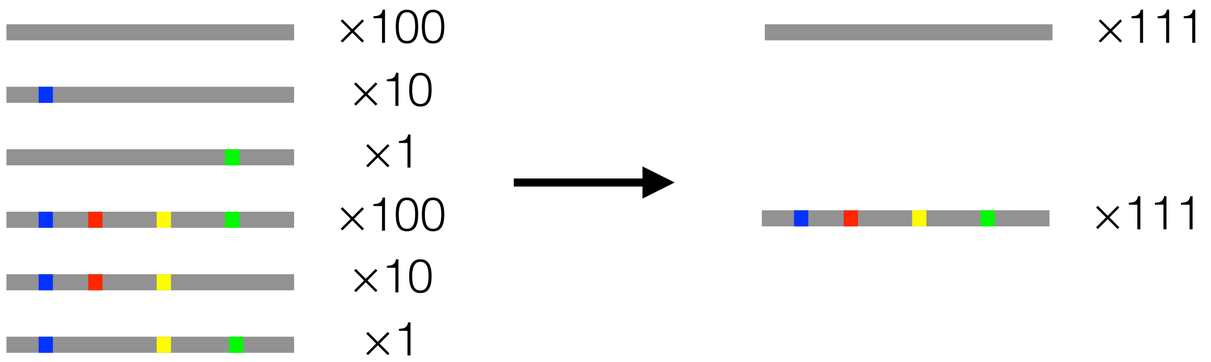
```
redRef.fasta
greenRef.fasta
blueRef.fasta
```

Genomic coordinates of the windows to be analysed; here, 1-300, 200-500, ...

Feature-packed 1

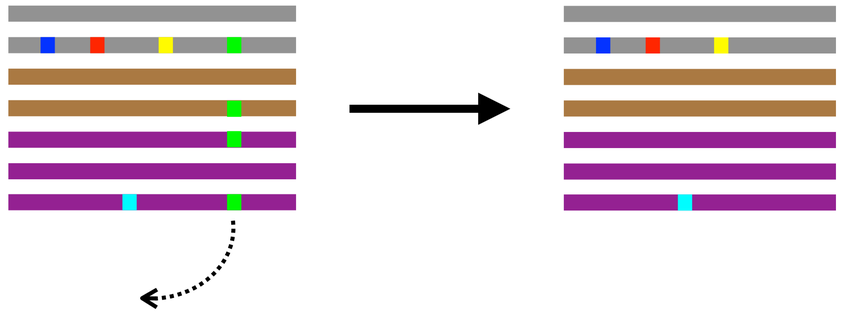
Merge overlapping paired reads into longer reads.





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

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

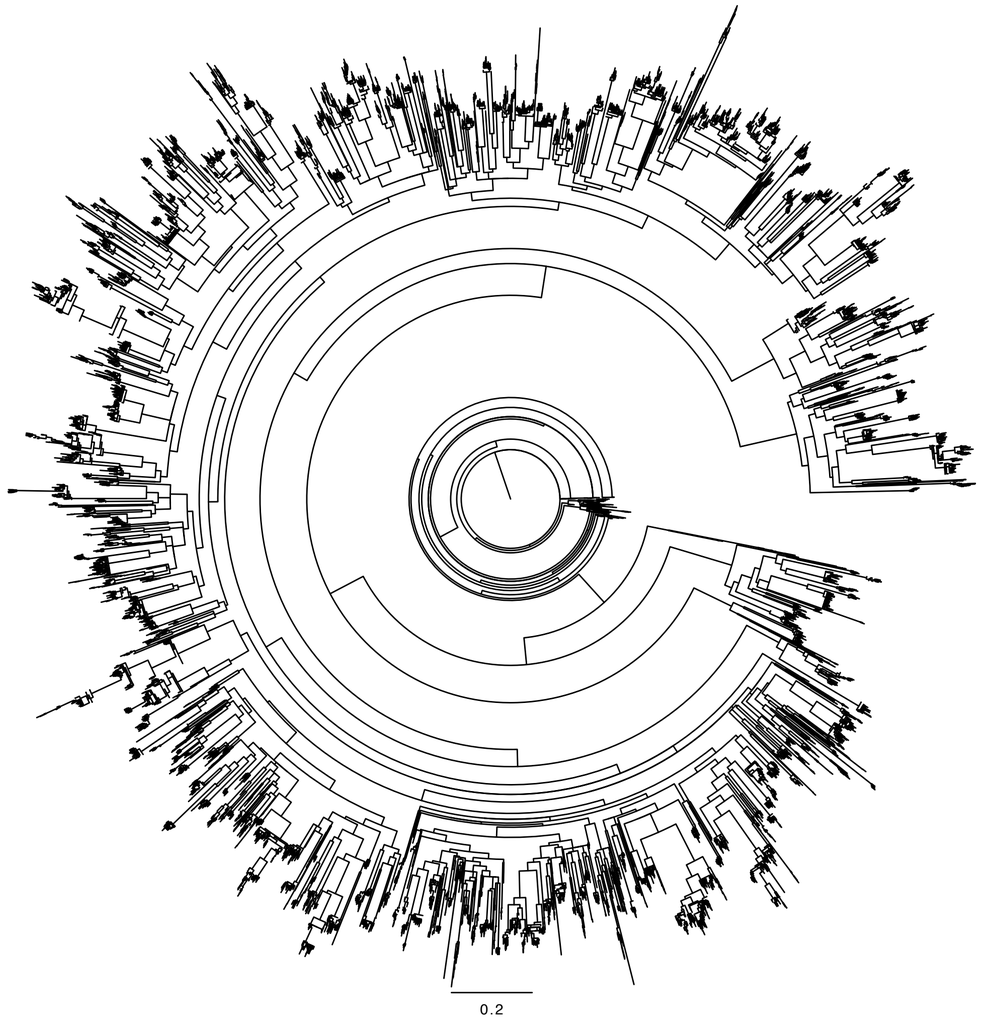


Feature-packed 2

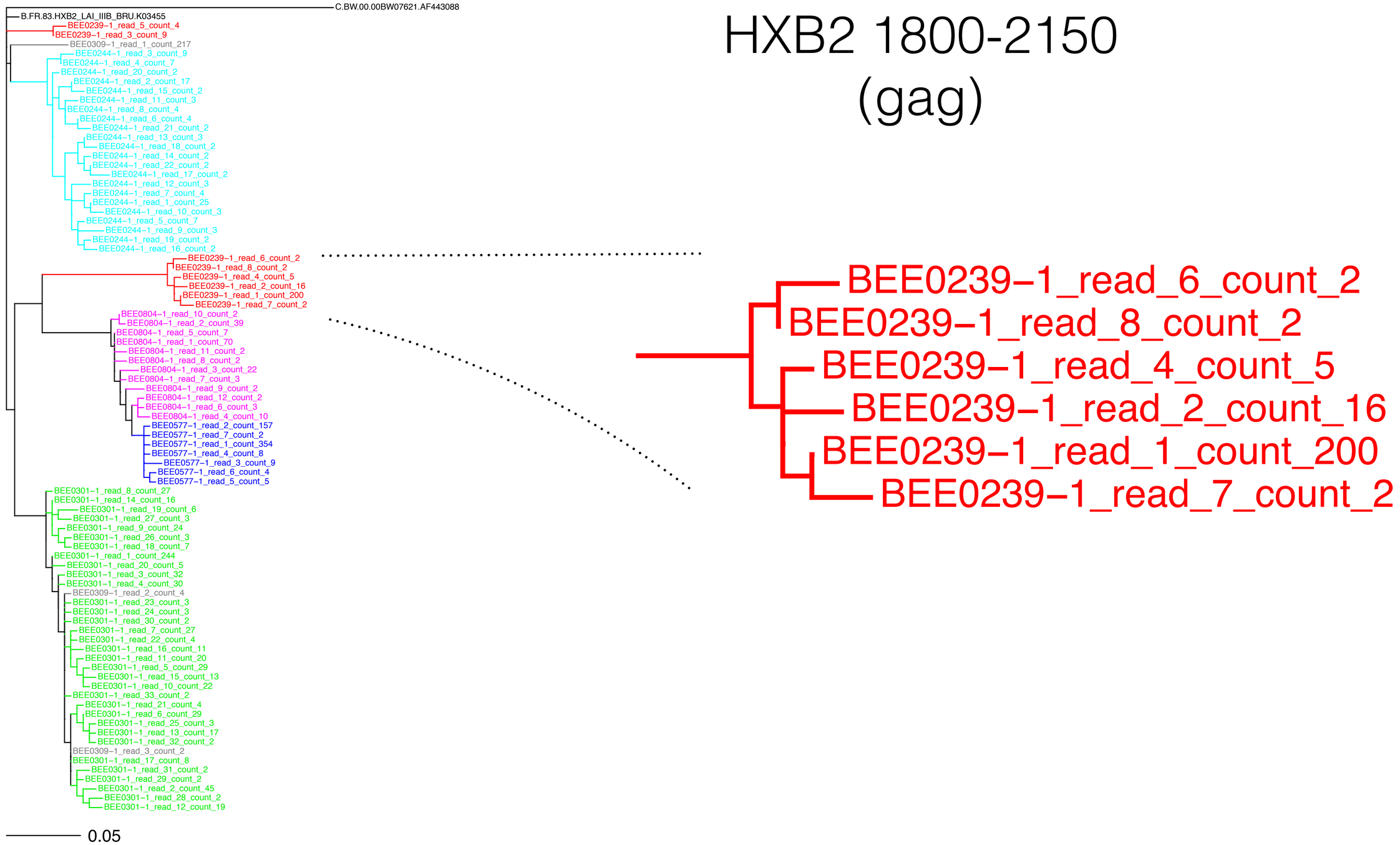
- Include known references with the reads.
- Trim and/or discard low-quality reads.
- Minimum read count.
- Check every read for duplication in any other patient to find contamination. Remove duplicates depending on counts.
- Bootstraps.

Trivially parallelisable: split the whole genome into windows, run each window as a separate job on your cluster.

<1 day for O(1000) patients with coverage O(10,000) (`RAxML` the bottleneck).



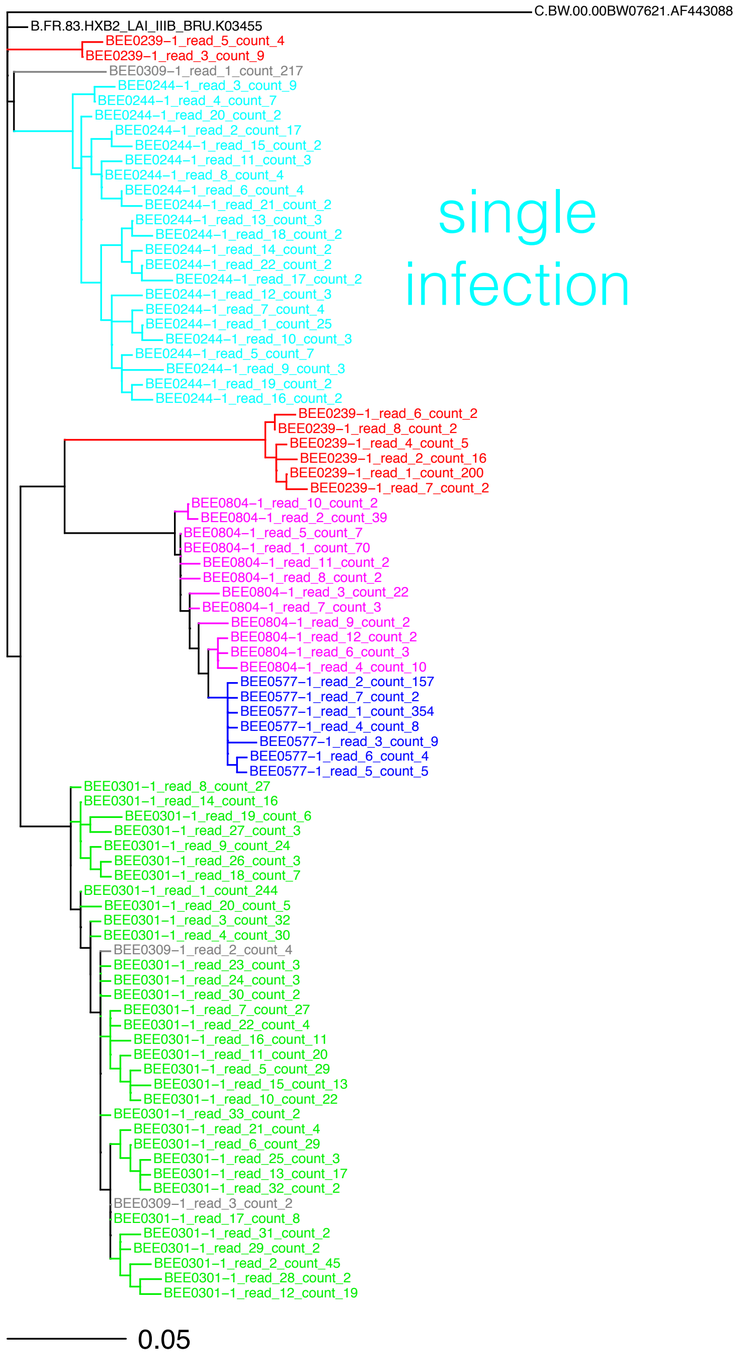
375bp at the end of the gp120 gene. 700 patients, 8574 distinct reads.
(merging threshold = 1bp, minimum read count = 2)



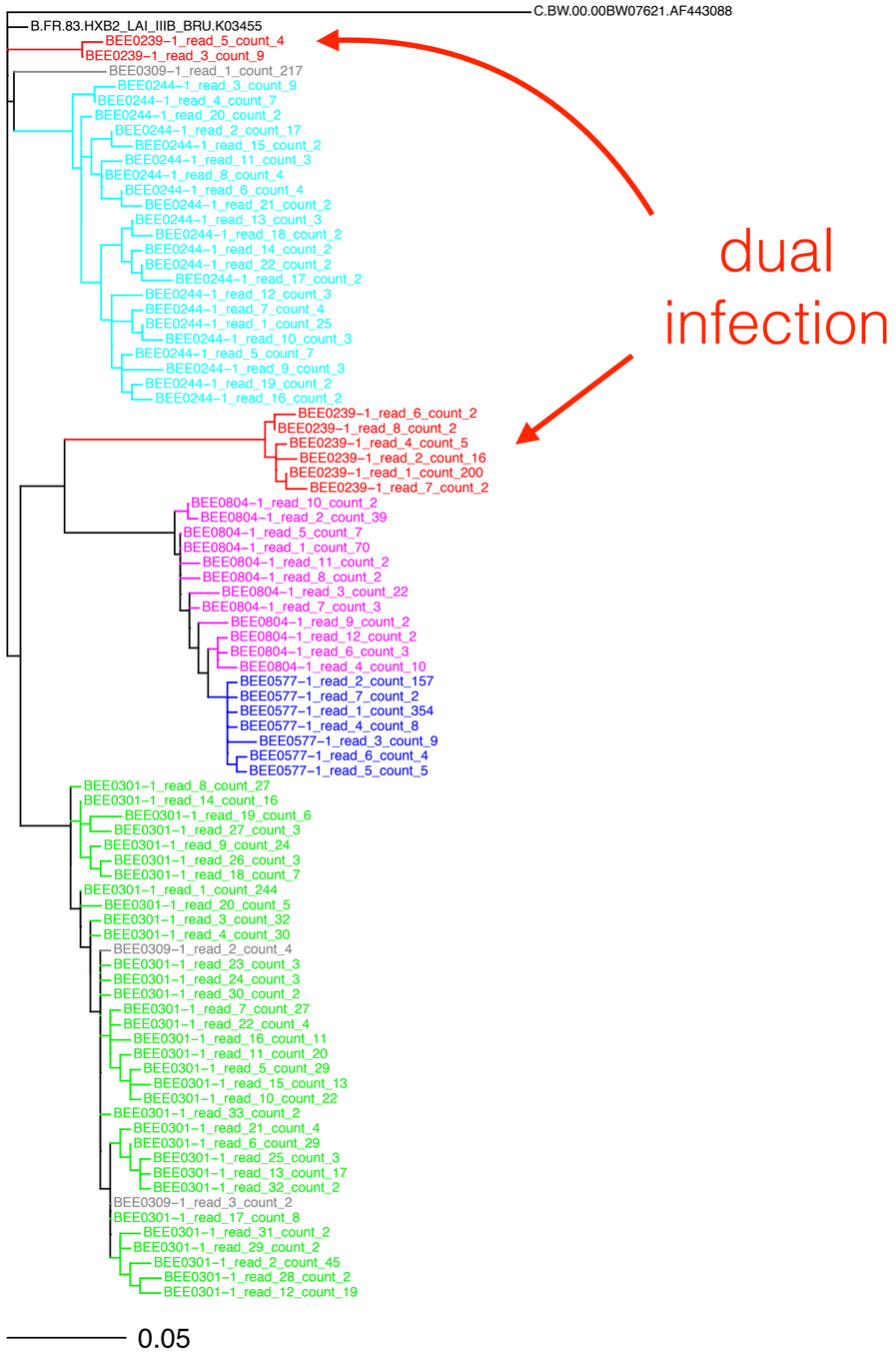
HXB2 1800-2150 (gag)

BEE0239−1_read_6_count_2
BEE0239−1_read_8_count_2
BEE0239−1_read_4_count_5
BEE0239−1_read_2_count_16
BEE0239−1_read_1_count_200
BEE0239−1_read_7_count_2

HXB2 1800-2150 (gag)



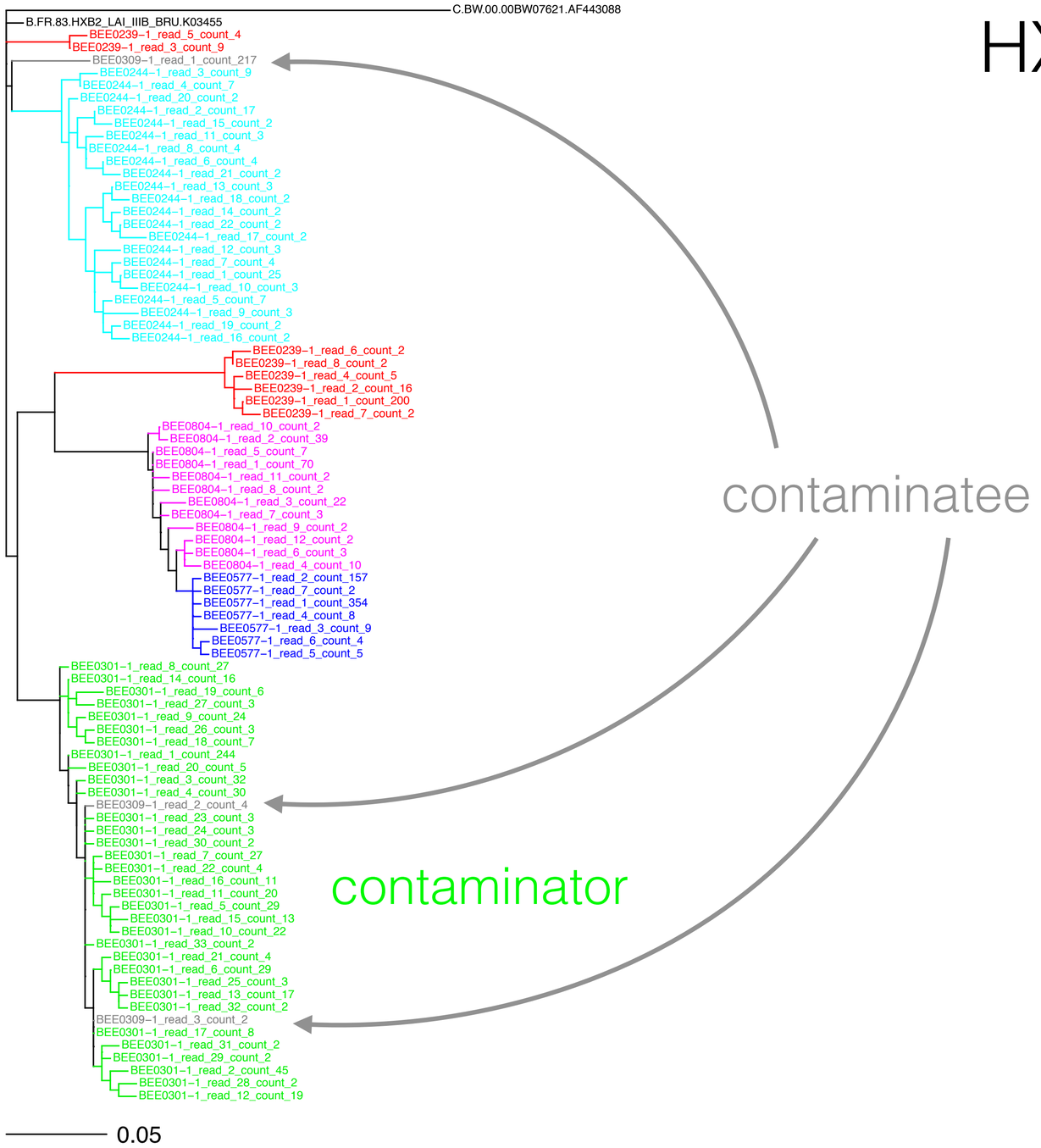
single infection



dual infection

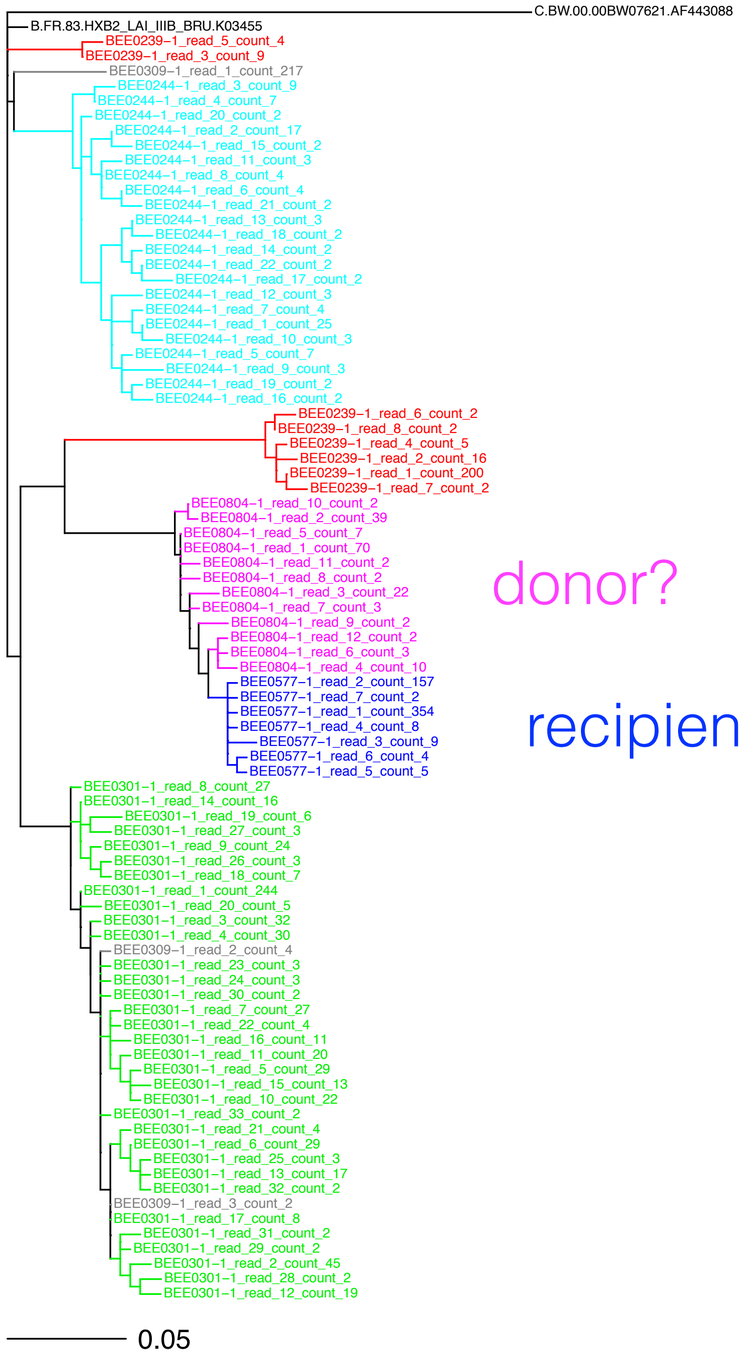
HXB2 1800-2150 (gag)

HXB2 1800-2150 (gag)



contaminatee

contaminator



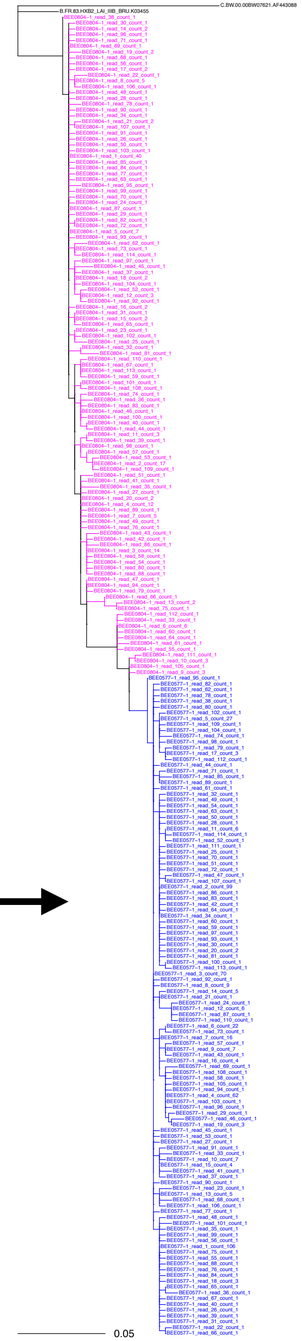
donor?

recipient?

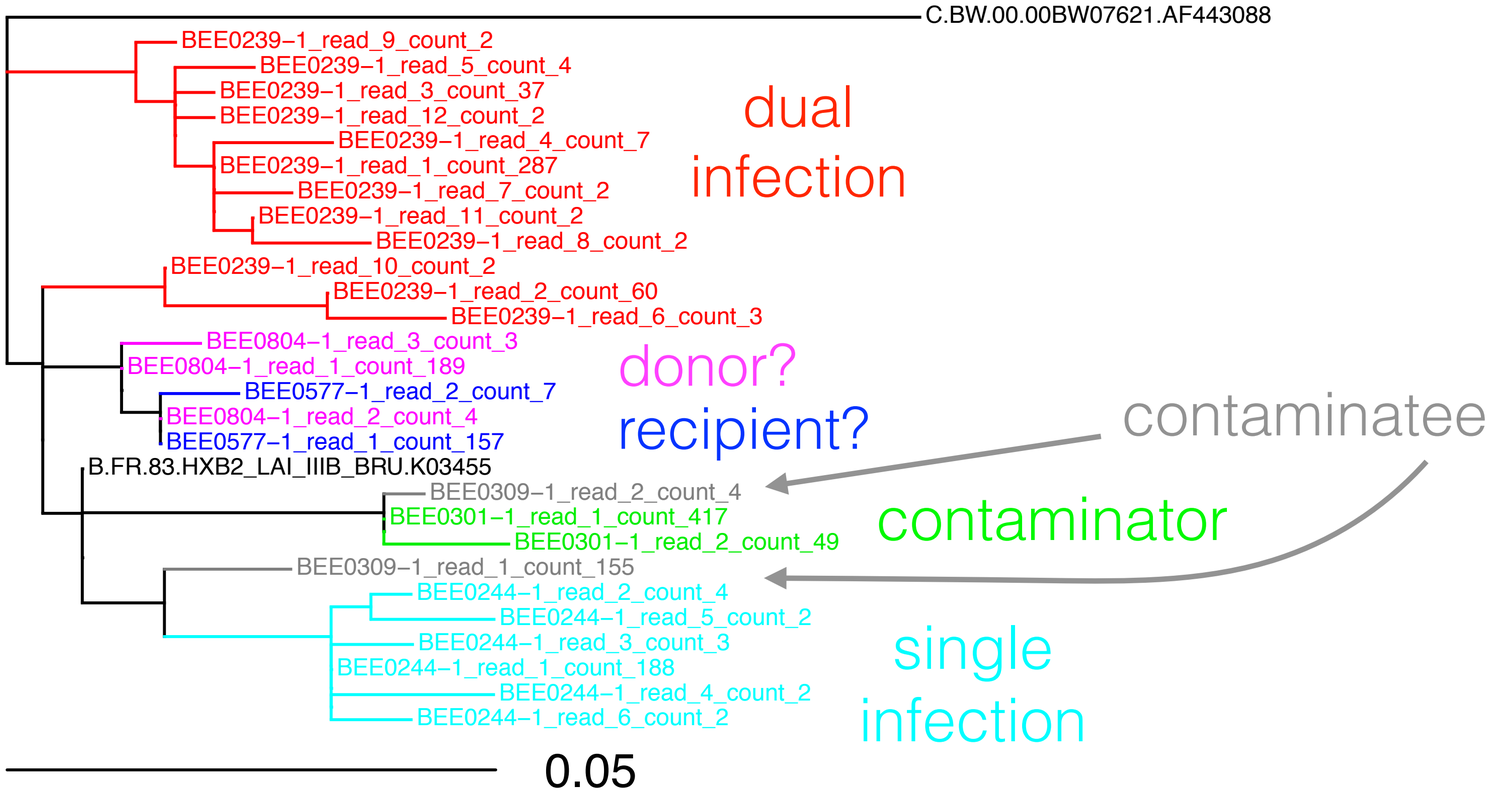
HXB2 1800-2150 (gag)

Left tree: 1bp merging, minimum read count = 2.

Re-run with just the transmission pair, no merging and no minimum read count for a clearer picture



HXB2 2800-3150 (pol)



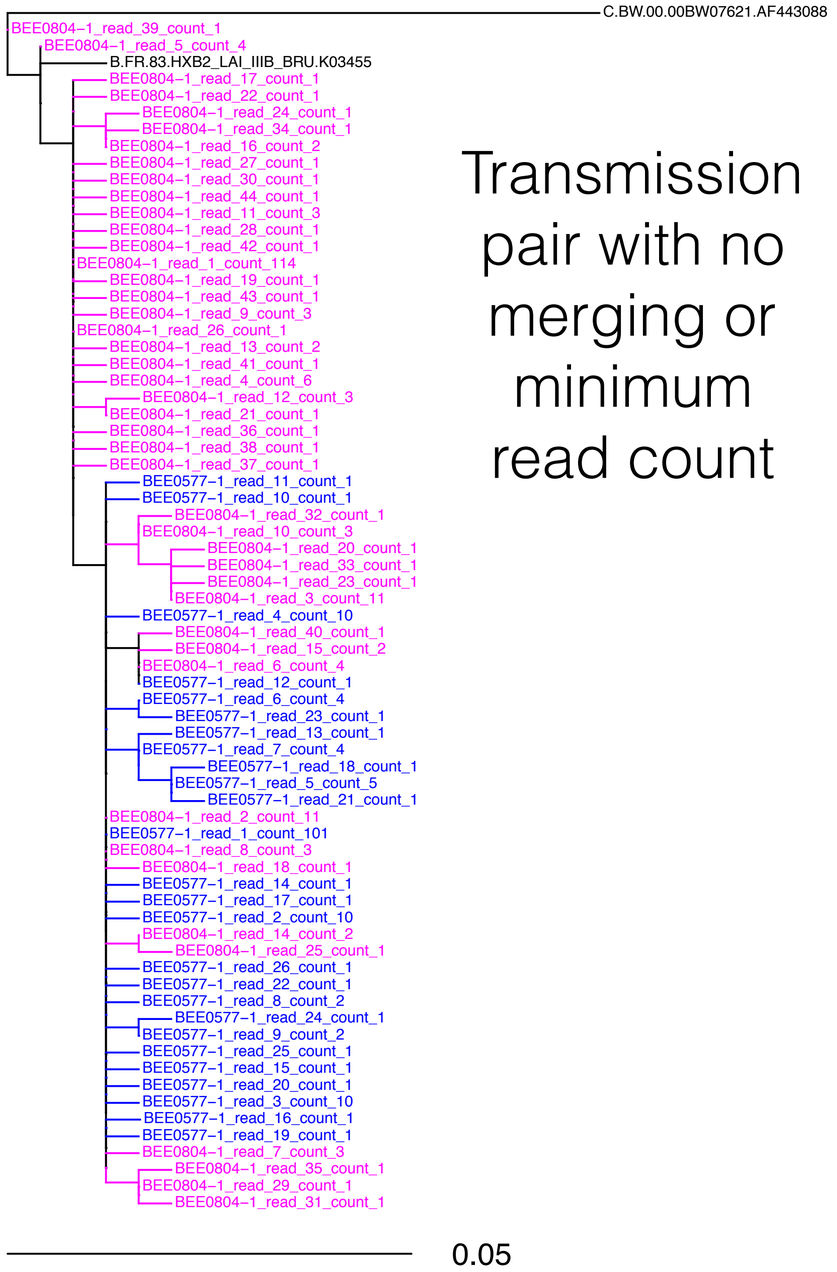
dual infection

donor? recipient?

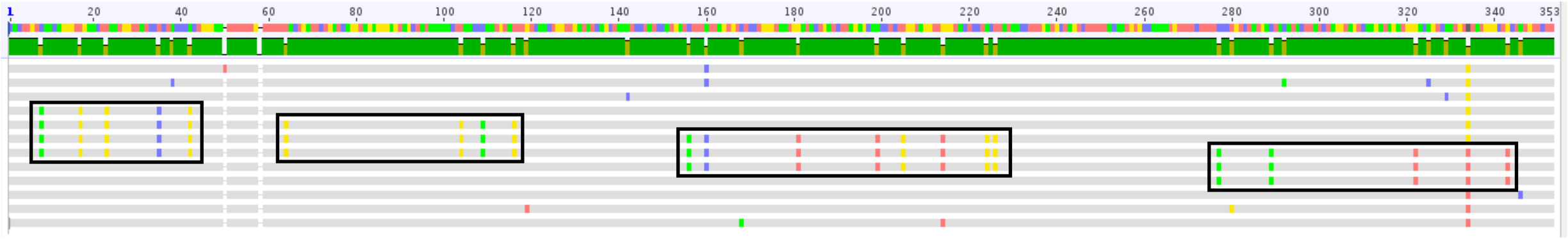
contaminatee

contaminator

single infection

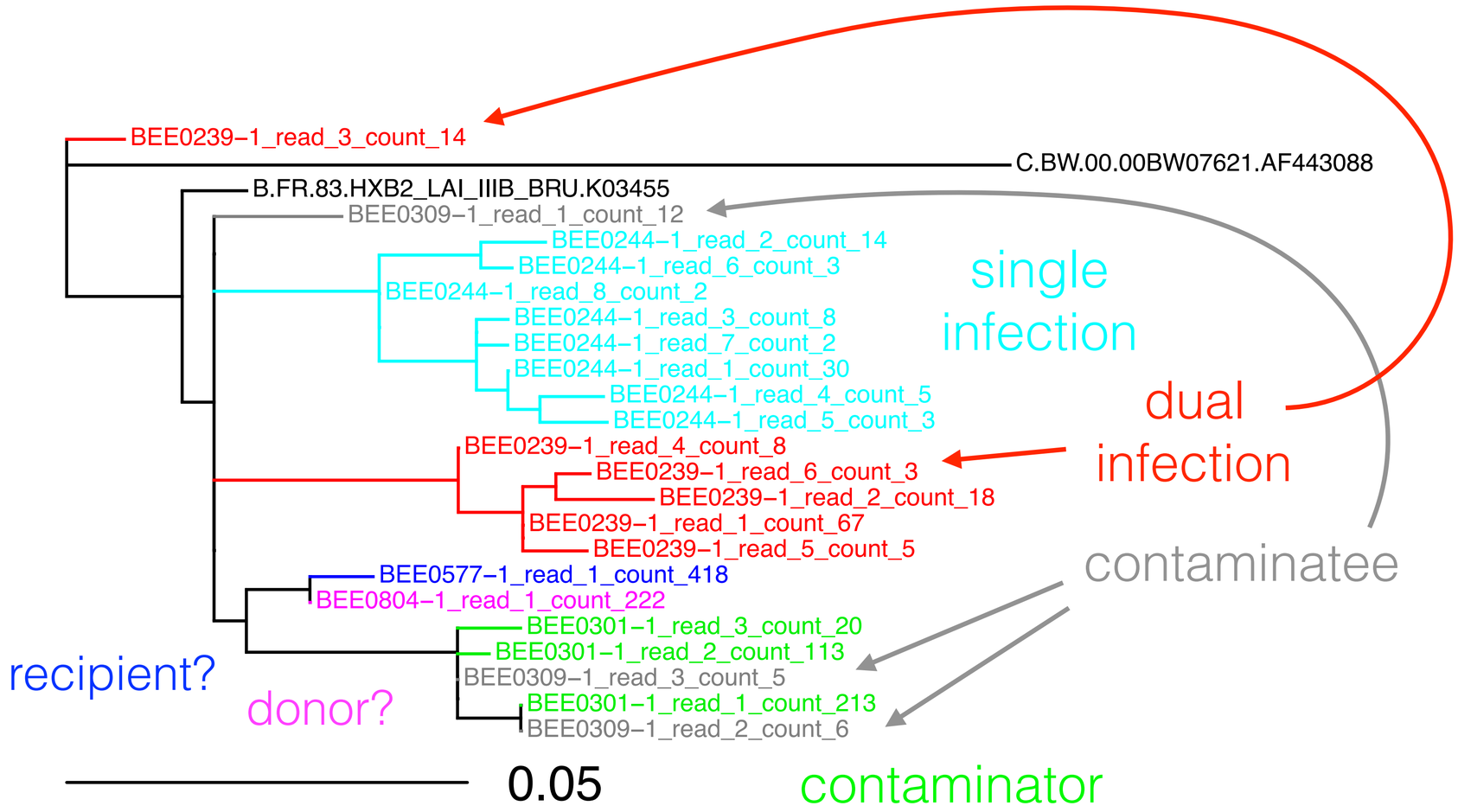


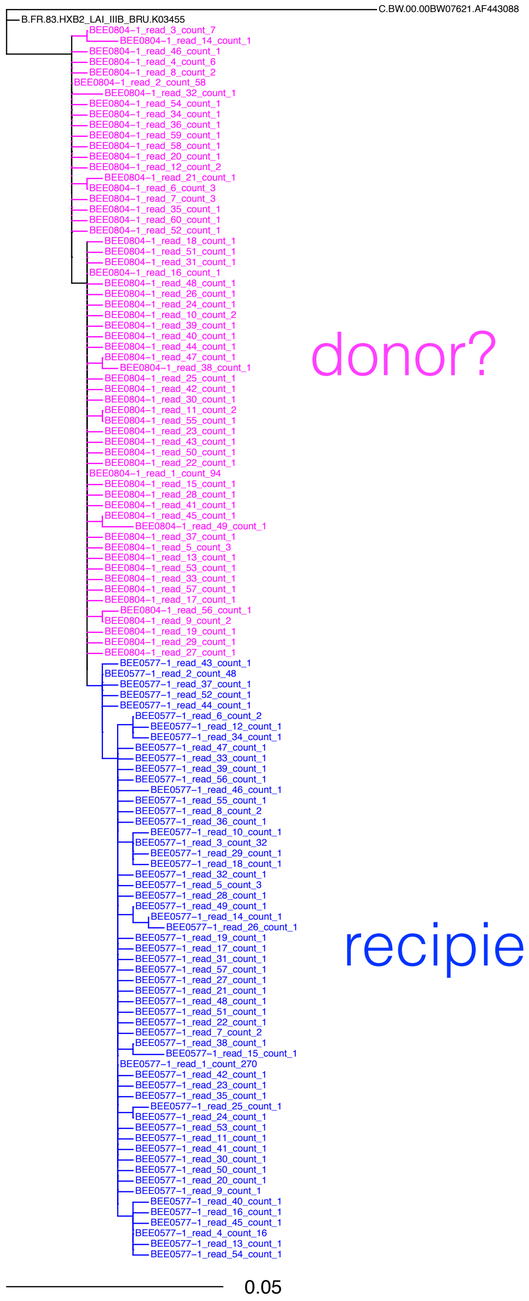
Transmission pair with no merging or minimum read count

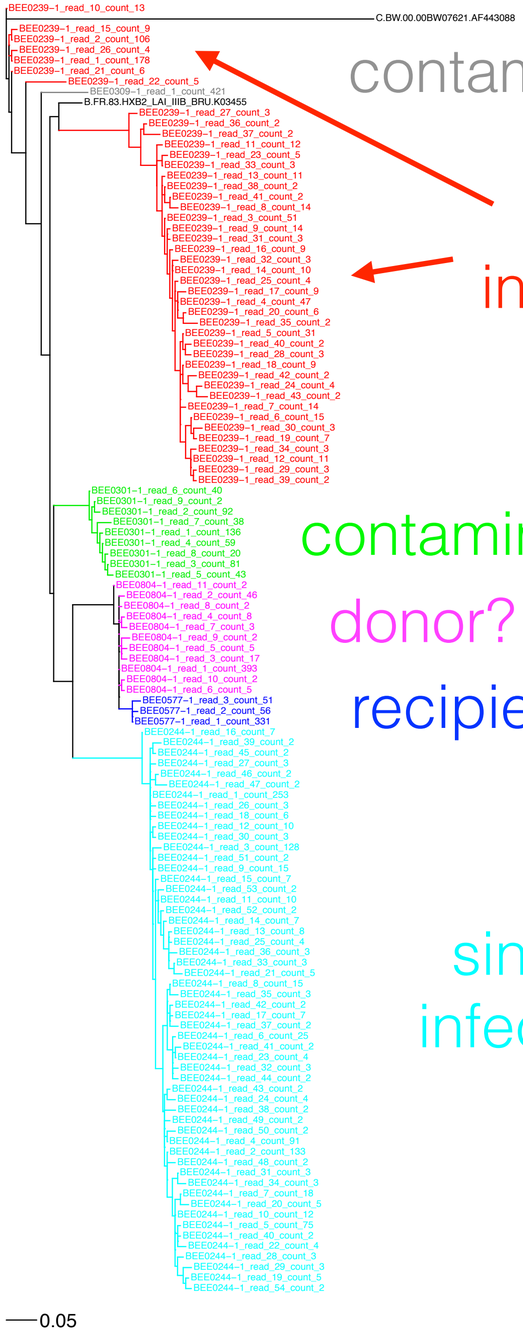


dual infection, point of recombination

HXB2 4800-5150 (pol & vif)







contaminatee (not here)

dual infection

contaminator (not here)

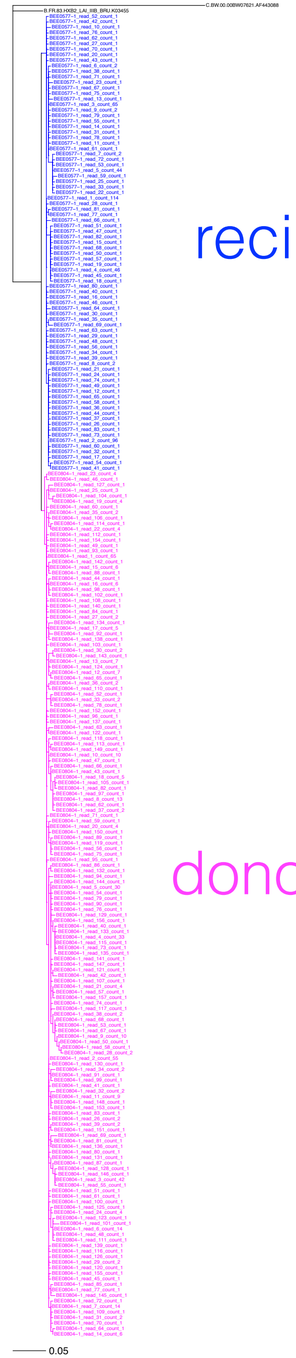
donor?

recipient?

single infection

HXB2 6050-6400 (vpu)

All trees excise the DRMs of *Johnson et al., Top. Anti. Med. 2011* and *Gatanaga et al., J. Biol. Chem. 2002* and the CTL-escape sites of *Carlson et al., J. Virology 2012*

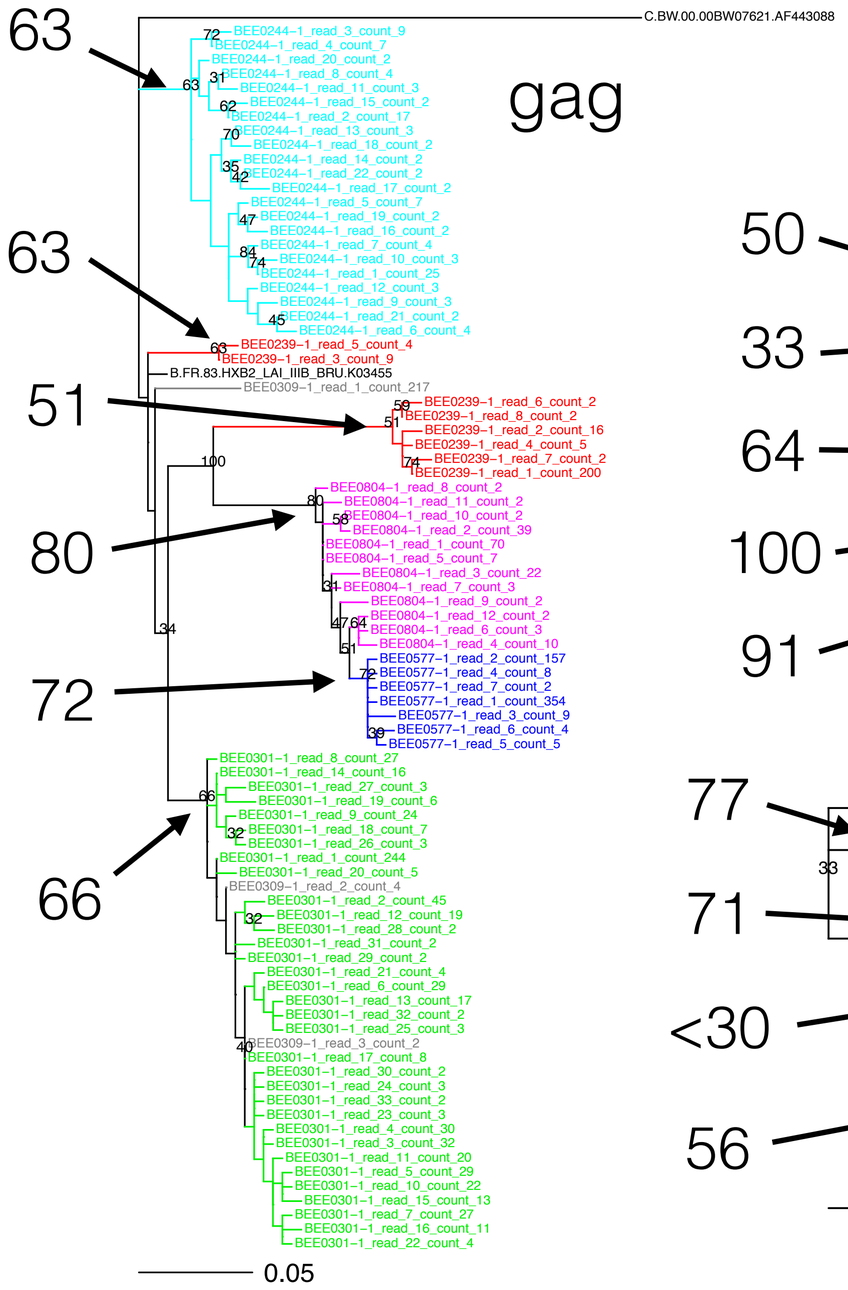


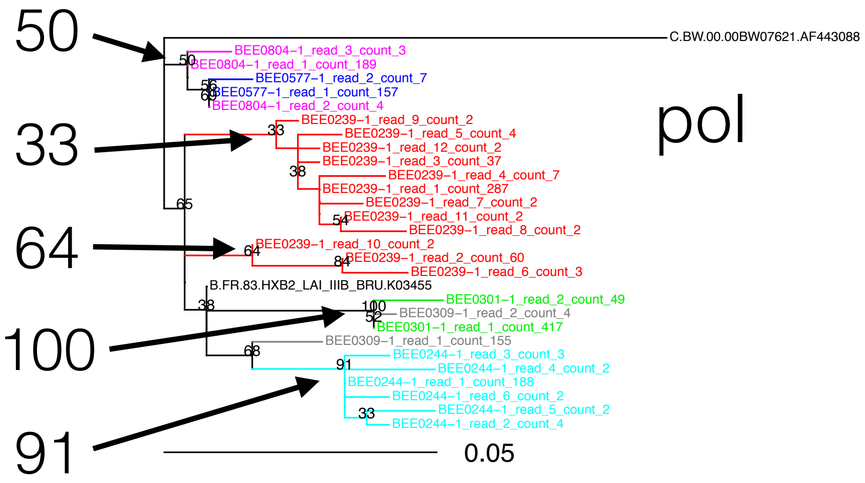
recipient?

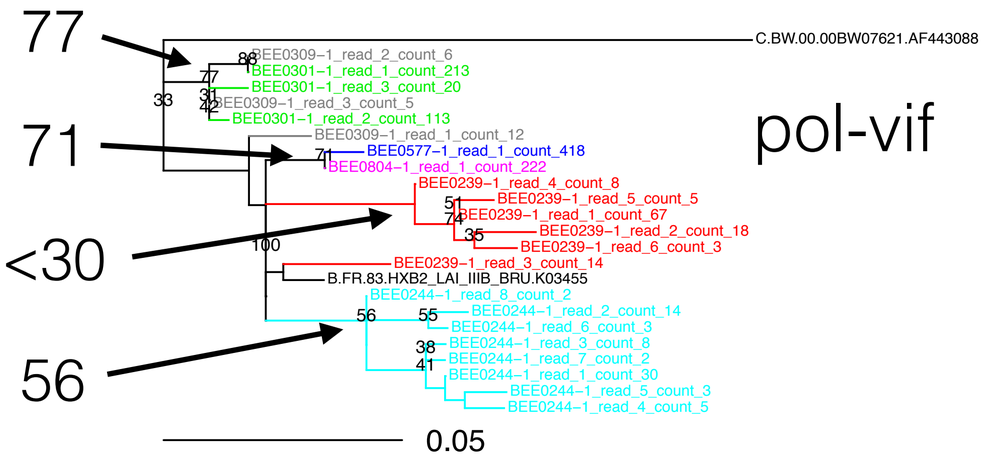
donor?

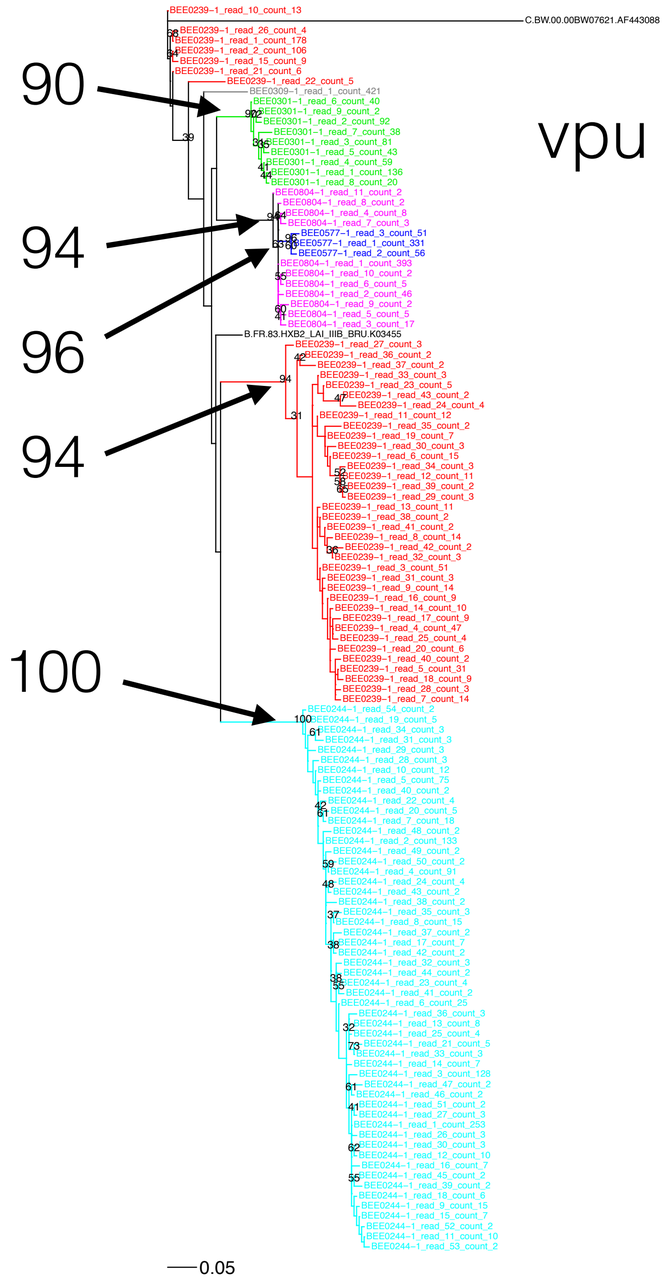
bootstraps (%)

values > 30 shown on nodes

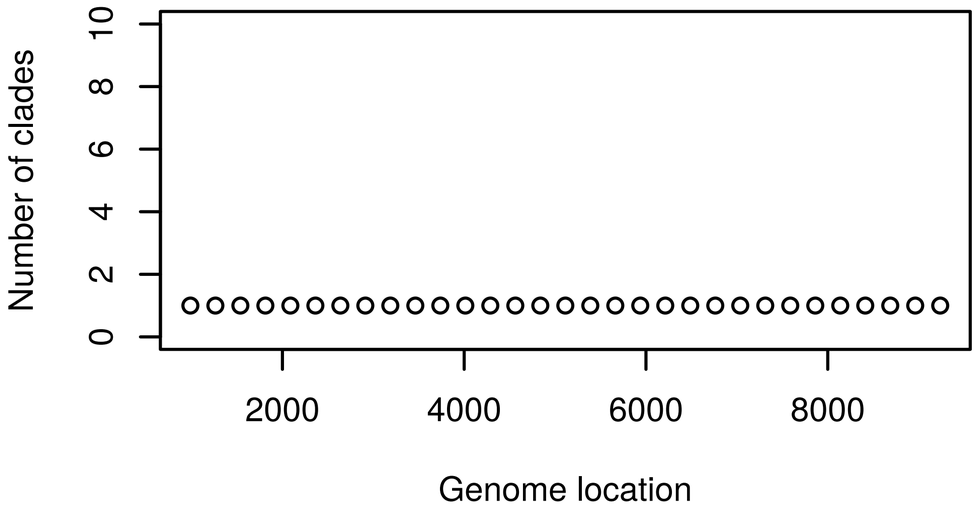






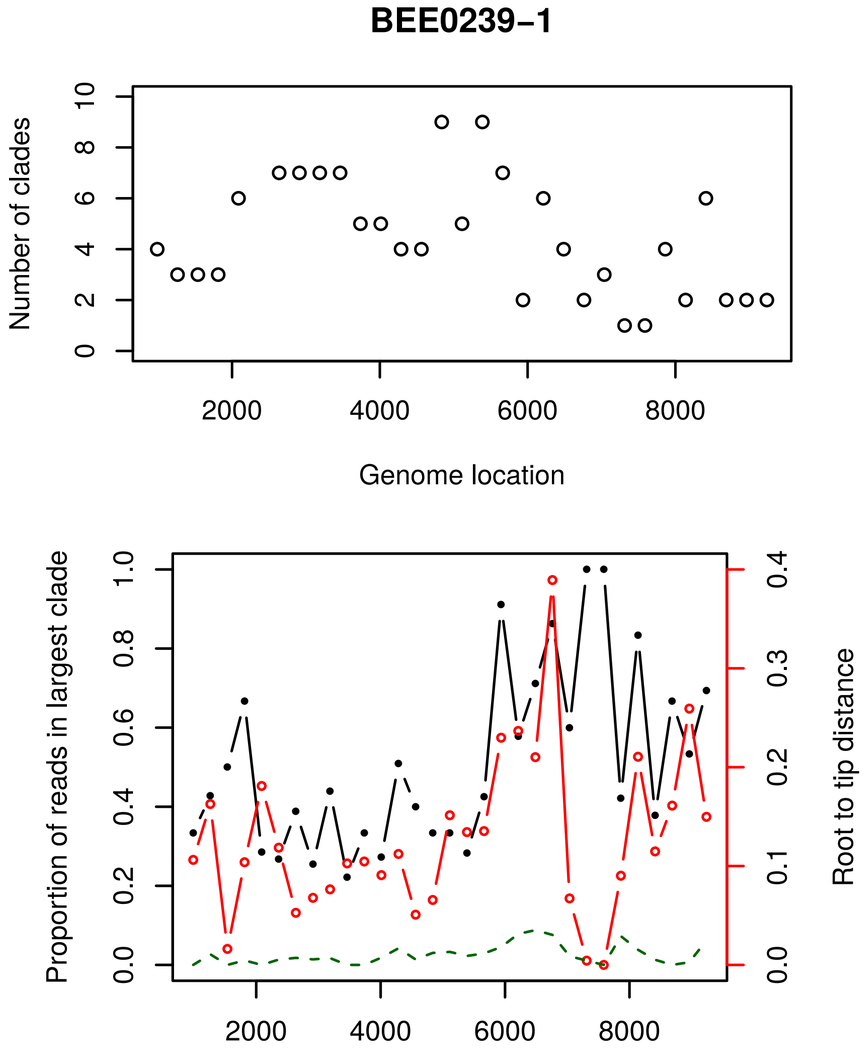


Per-patient summaries over the genome



single infection

clades = 1 throughout

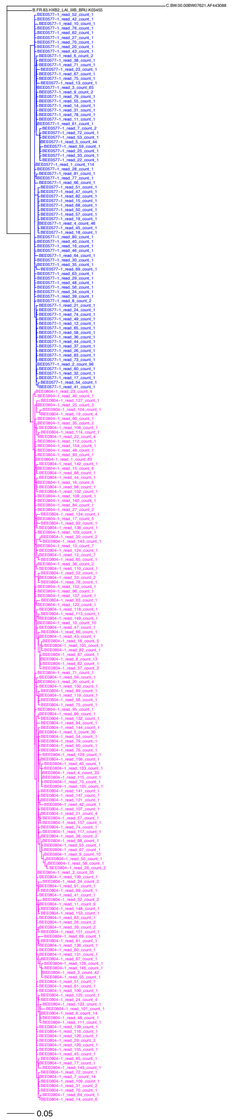


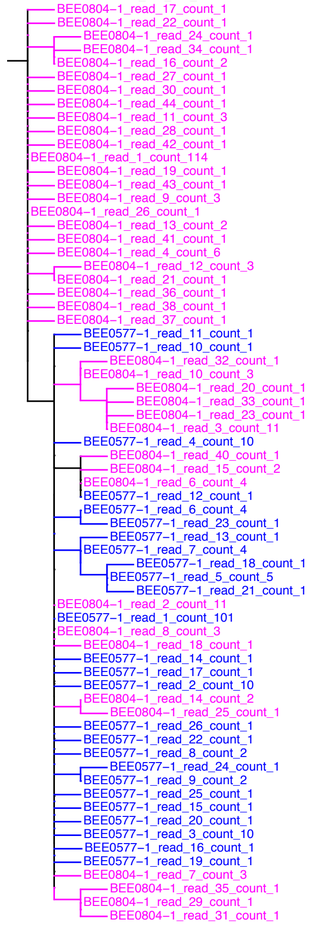
dual infection

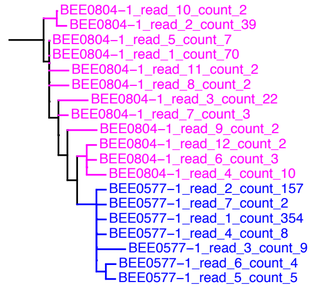
>1 clade, large root-tip, large proportion of reads outside largest clade

Two patients' reads can be...

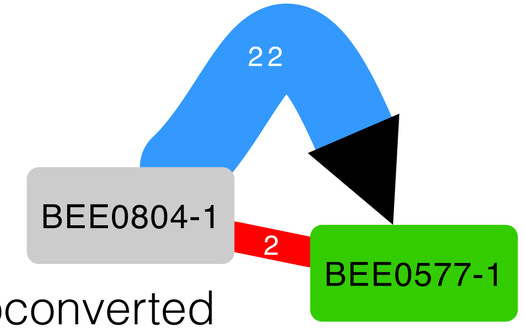
separated, entangled, or clearly directed.







Count these over all windows (31 here):



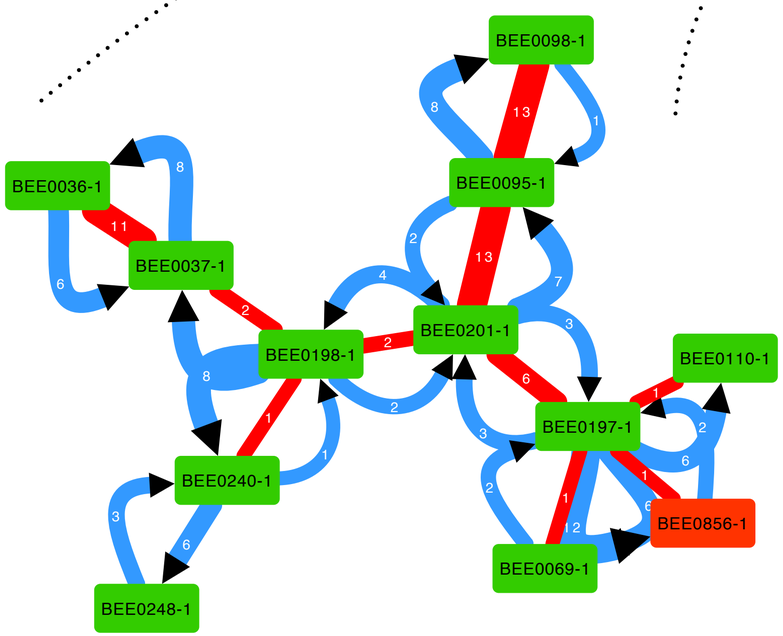
seroconverted 2007, Amsterdam

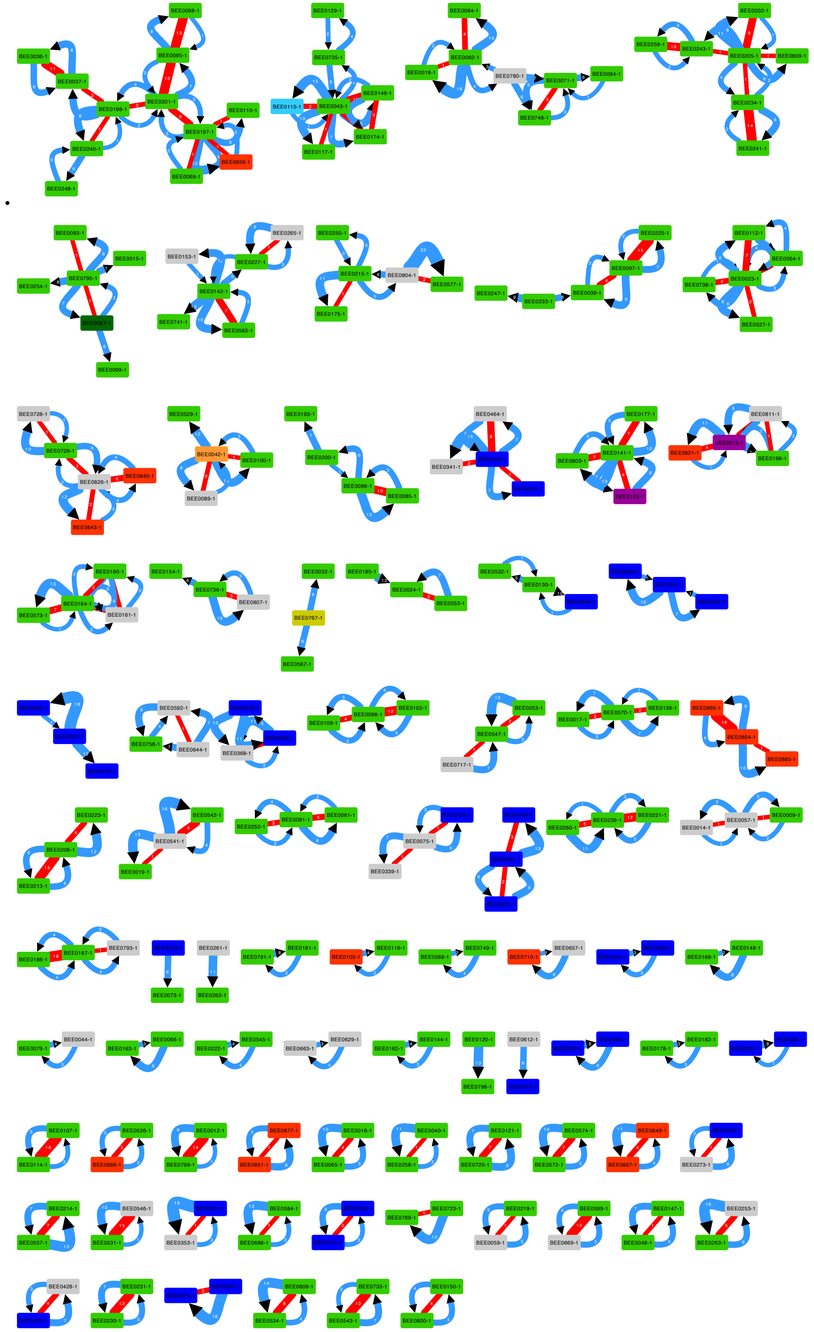
seroconverted 2010, Amsterdam

Matthew Hall

See BEAST2 package *BEASTLIER* (github: *twoseventwo*)

c.f. Ypma et al., *Genetics* 2013;
Didelot et al., *Mol. Biol. Evol.* 2014;
Hall et al. *PLoS Comp. Biol.* 2015:
Kenah et al., arxiv 2016.
Romero-Severson et al. 2016

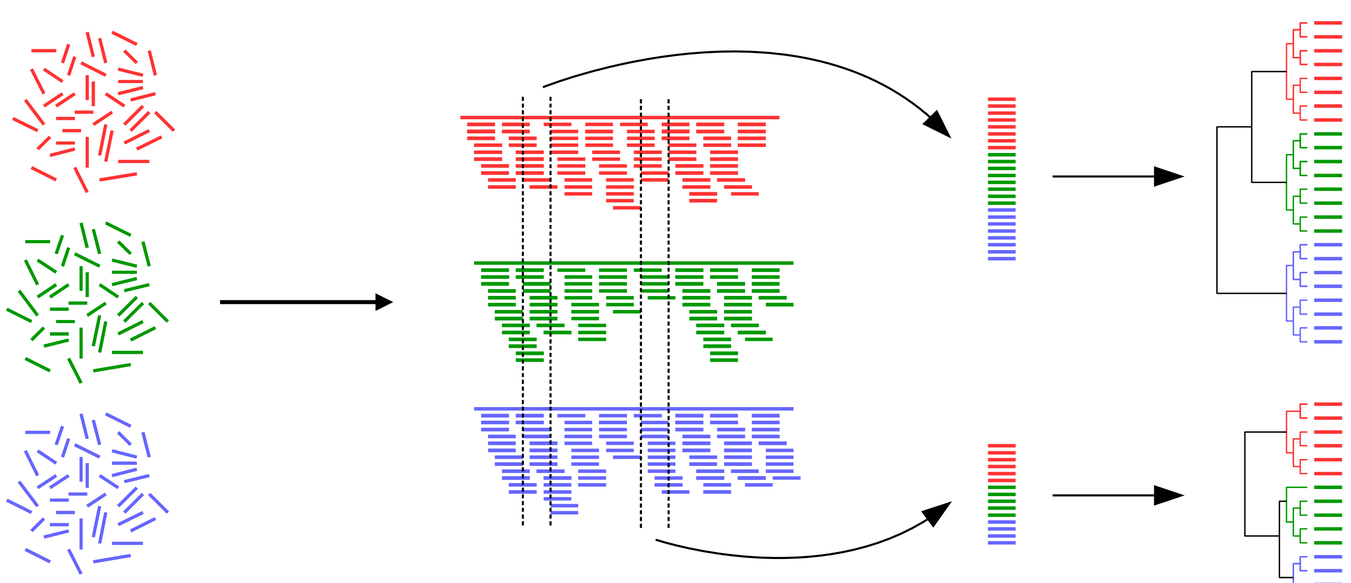




Node Fill Color	LONG_COUNTRY_INF
	Belgium
	Germany
	Morocco
	Netherlands
	Portugal
	Spain
	United Kingdom
	United States
	Unknown

Edge Stroke Color (Unselected)	interaction
	int
	trans

Summary



`shiver` `phyloscanner`

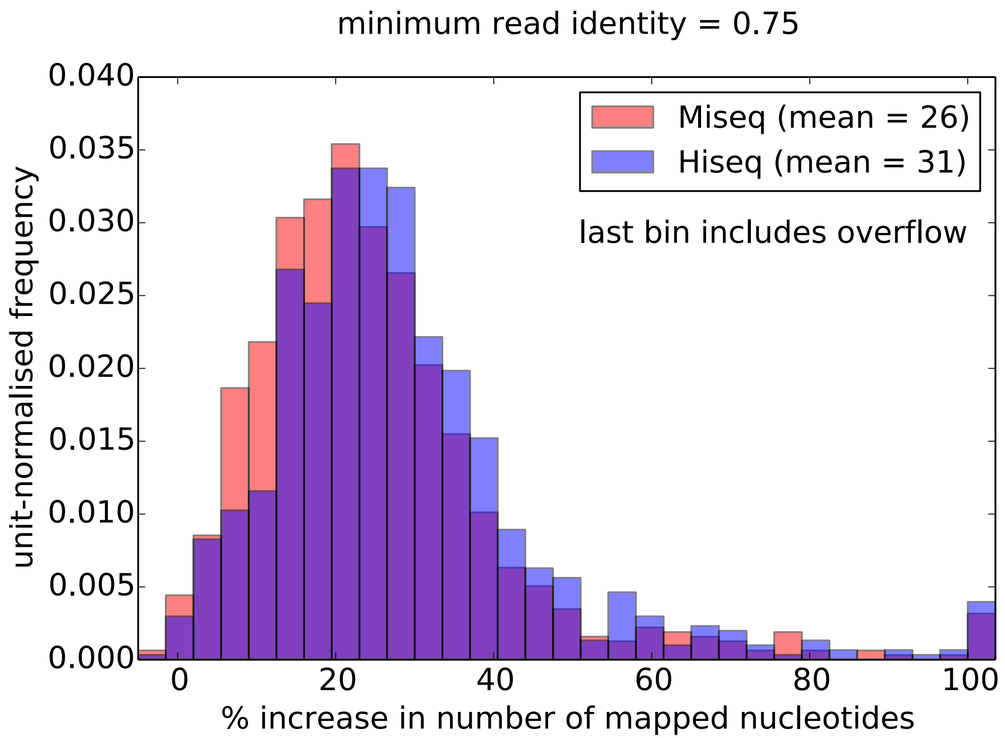
Analysis of phylogenetics between and within, along the genome, shows:

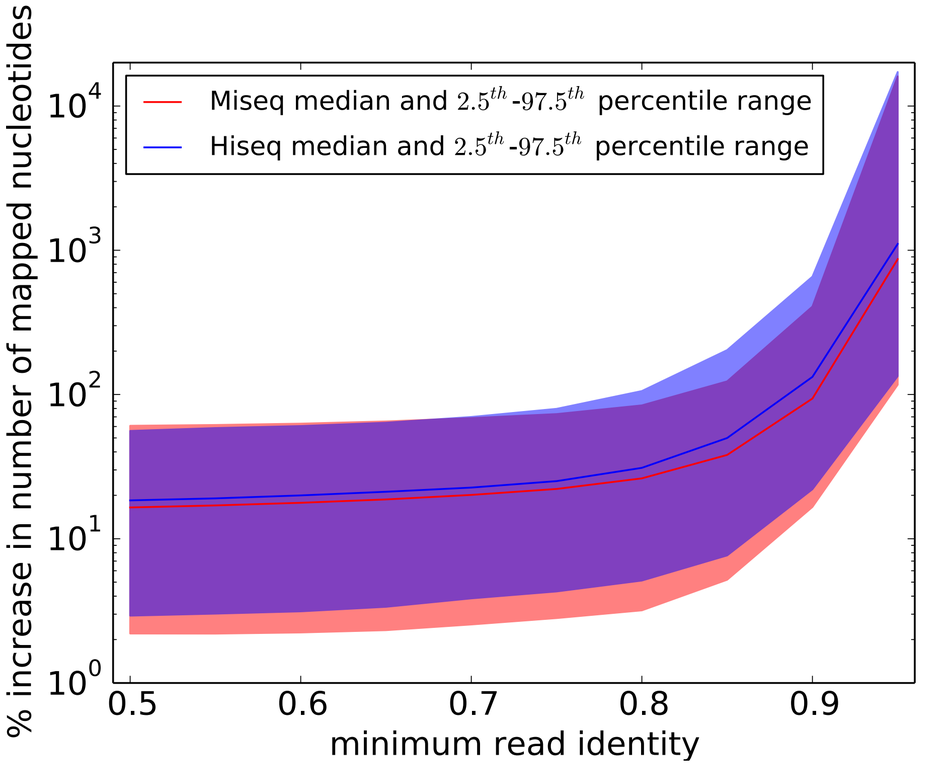
- **transmission**: one quasispecies as the descendant of another
- **dual infections**: two distinct quasispecies
- **recombination**: intermediates between two quasispecies
- **contamination**: repeated exact duplication of reads

Extra slides

Quantifying, not just arguing,
why you should use `shiver`

shiver maps more nucleotides





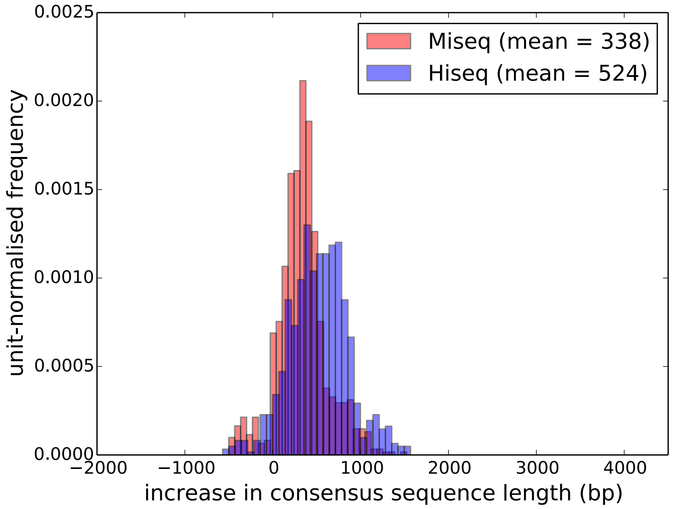
Map 904 Miseq and 864 Hiseq samples first to the closest of 160 existing references (chosen with `Kraken`, often HXB2) then to the custom reference constructed by `shiver`. Compare the number of nucleotides mapped. Repeat for different thresholds on read identity - the fraction of a read that agrees with the reference.

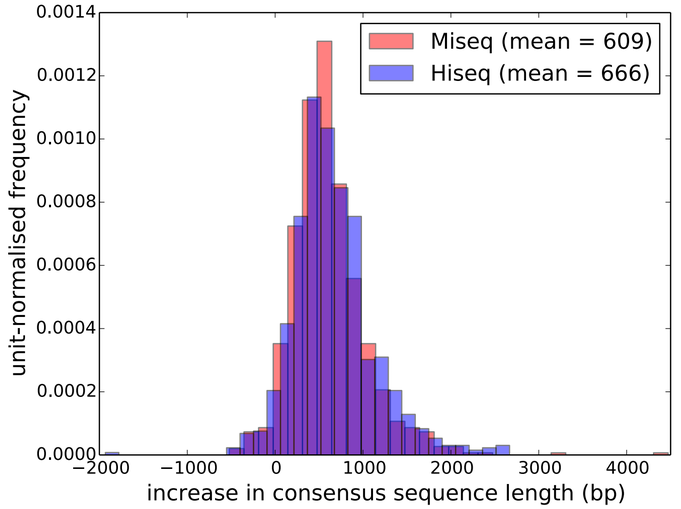
`shiver` reconstructs more of the genome

lenient on read identity (>50%)

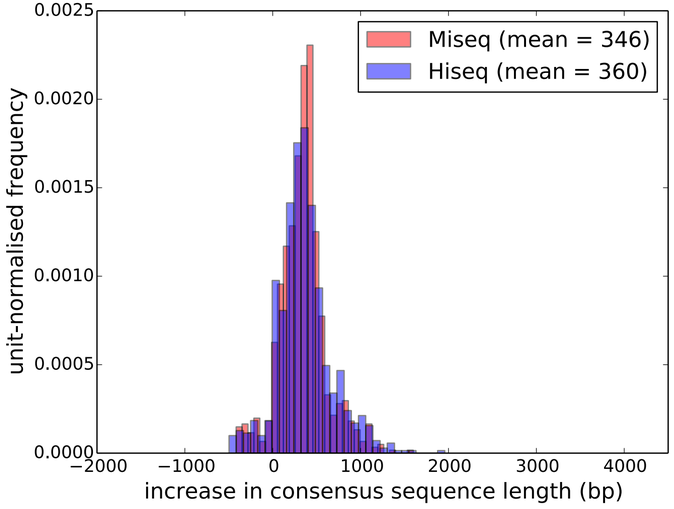
strict on read identity (>85%)

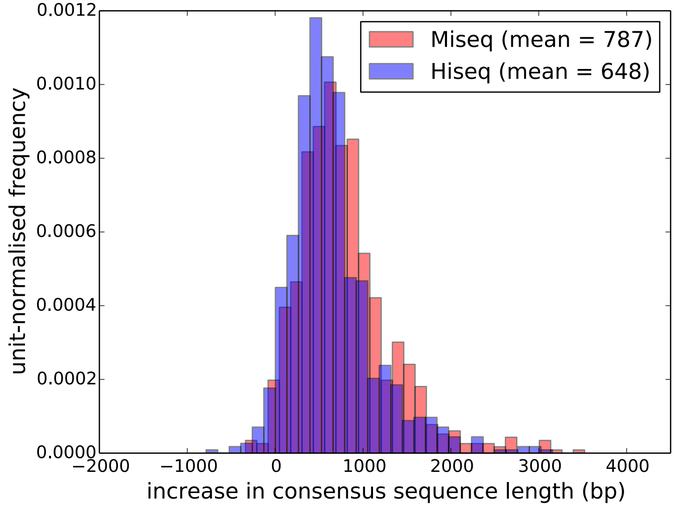
lenient on coverage (>10)



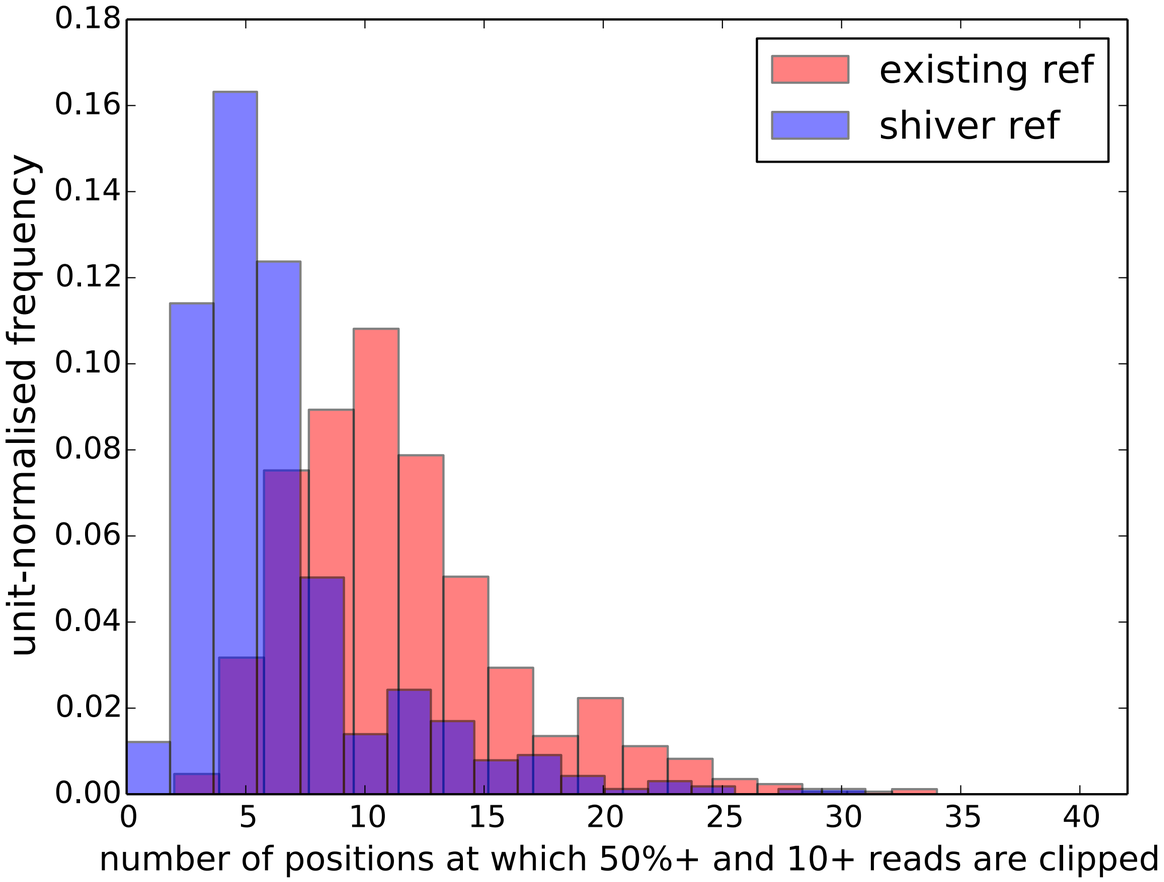


strict on coverage (>100)

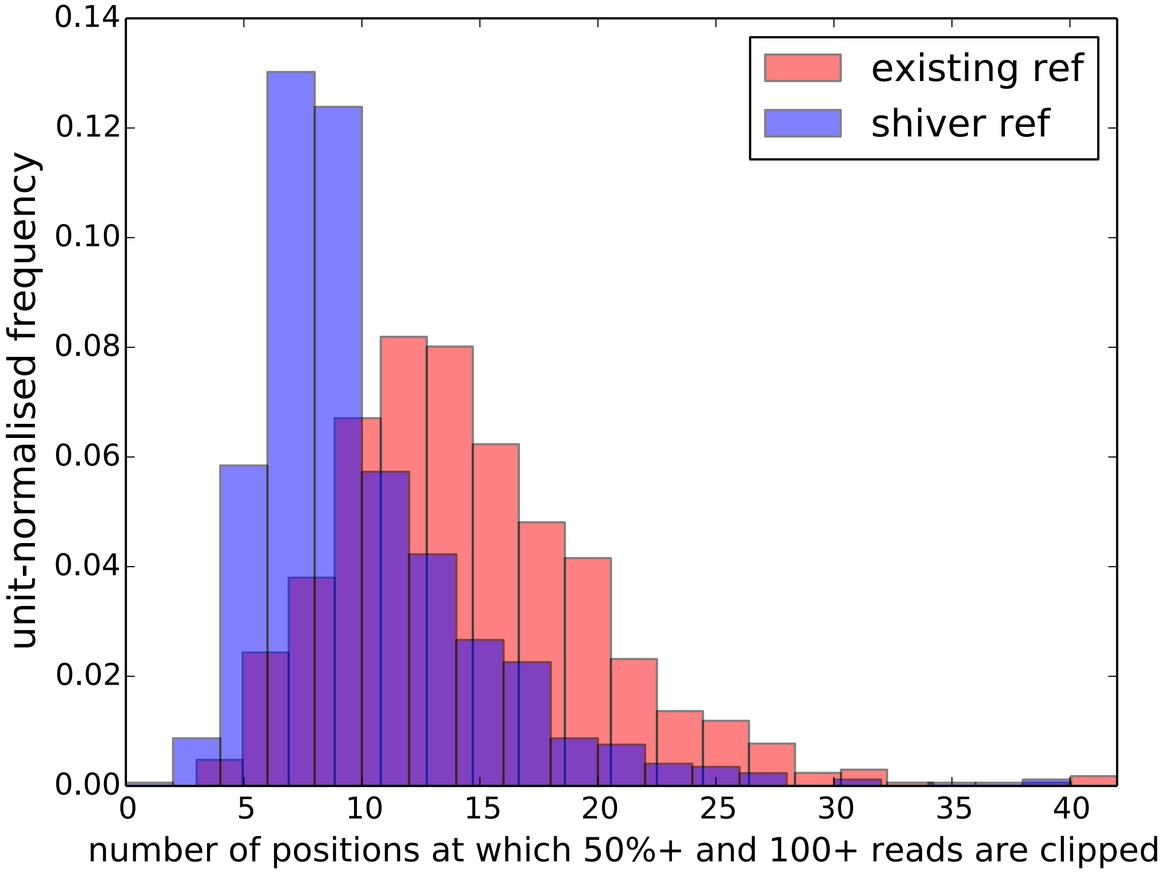




`shiver` decreases bias in the consensus



Miseq



Hiseq