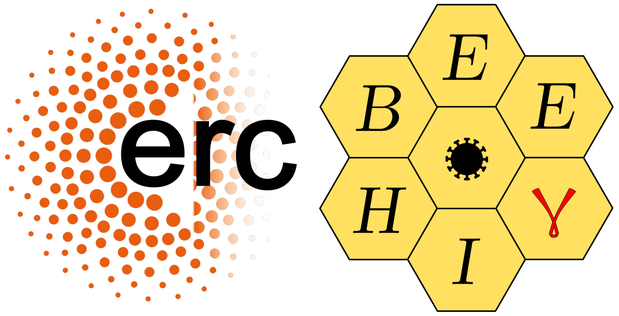
Phylogenetics as motivation for pathogen sequencing, and viral genome assembly e.g. with `shiver`

Chris Wymant[1,2]

[1] Big Data Institute, Nuffield Department of Medicine, University of Oxford
[2] Department of Infectious Disease Epidemiology, Imperial College London



Infectious diseases: motivation

HIV:

- 35,000,000 deaths so far
- In 2017:
  - 36,900,000 people living with HIV (no cure)
  - 1,800,000 new infections
  - 940,000 deaths

Tuberculosis in 2017: 10,000,000 cases, 1,600,000 died.

Malaria in 2016: 216,000,000 cases, 445,000 deaths, disproportionately children under 5.

Viral hepatitis in 2015: >328,000,000 chronic infections, 1,340,000 deaths

$O(10^7)$ deaths anually.
Make an $O(10^{-6})$ improvement: 10 lives saved each year.

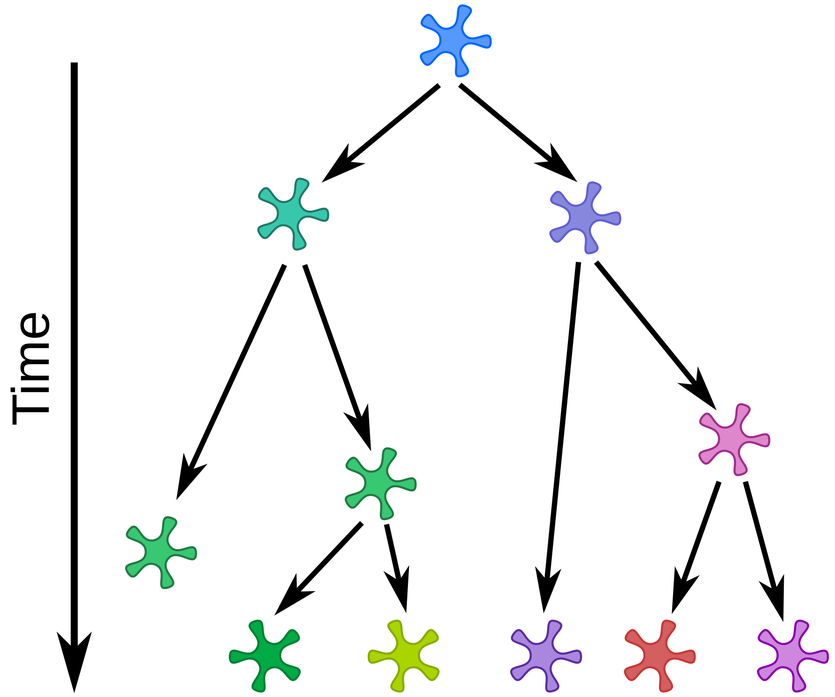
Source: WHO

Problem: how can we more effectively hinder the spread of an infectious disease?

A solution: learn more about the disease & its spread by analysing pathogen sequence data.

Genetic changes (substitution) accumulate over time

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

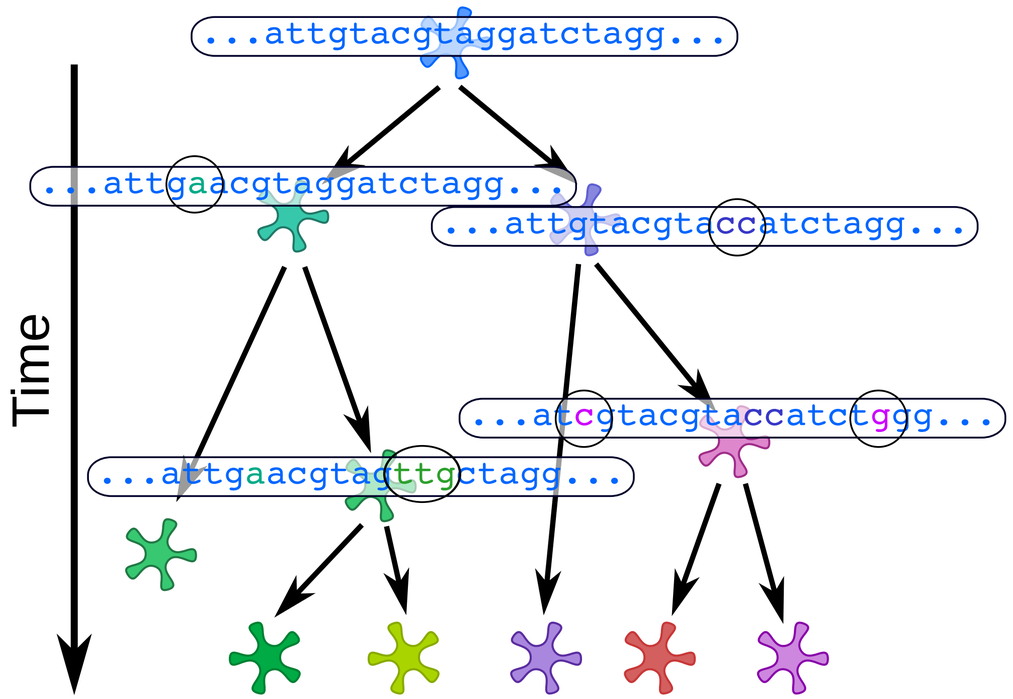


Original slide by
Thibaut Jombart

See also excellent webinar on understanding pathogen phylogenies by Christophe Fraser
https://www.dropbox.com/s/mia686o4oy1mhoj/181109%20Christophe%20Fraser.mp4?dl=0

Genetic changes (substitution) accumulate over time

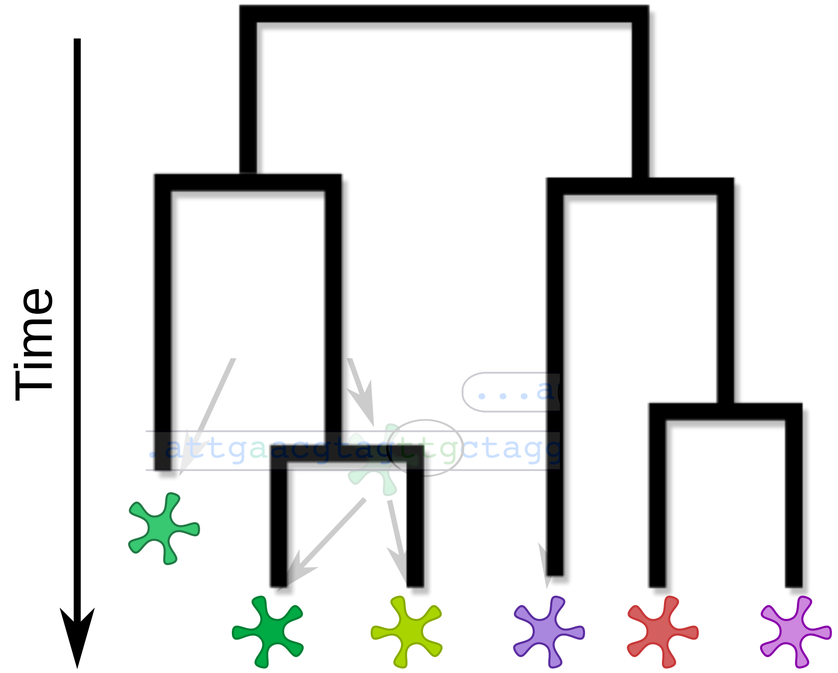
Substitution: replacement of a nucleotide (e.g. a $\rightarrow$ t)



Original slide by
Thibaut Jombart

Genetic changes (substitution) accumulate over time

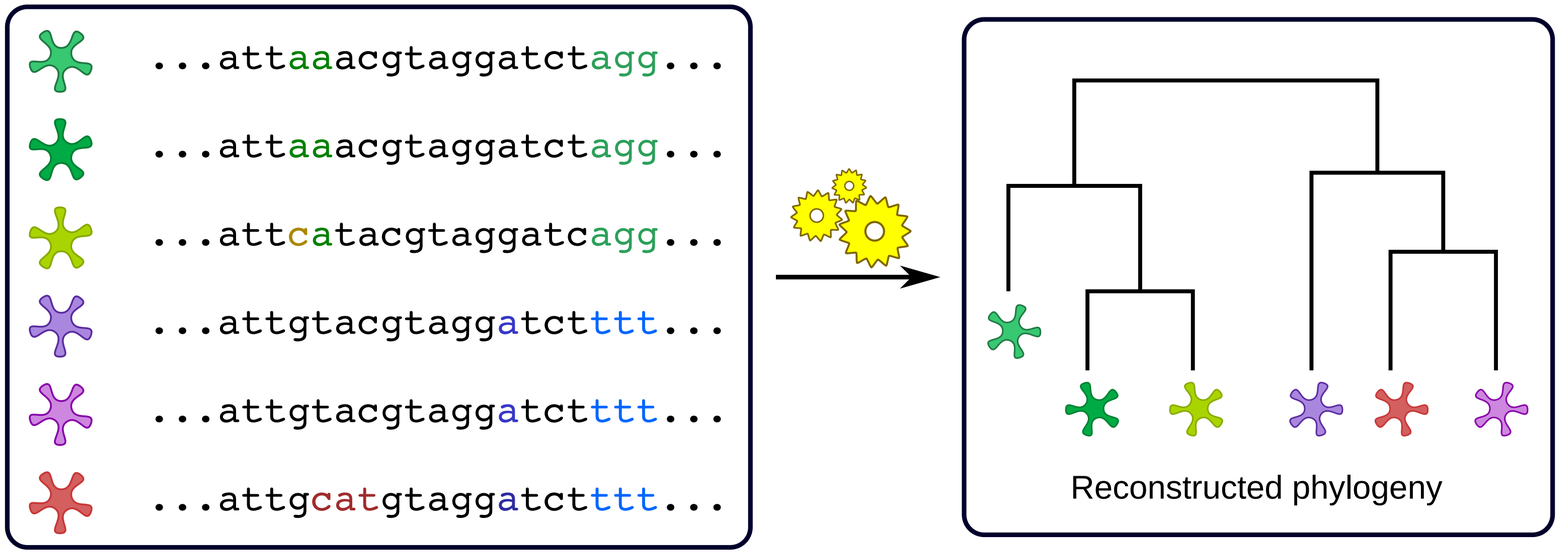
Substitution: replacement of a nucleotide (e.g. a $\rightarrow$ t)



Substitution patterns reflect the evolutionary history

Original slide by
Thibaut Jombart

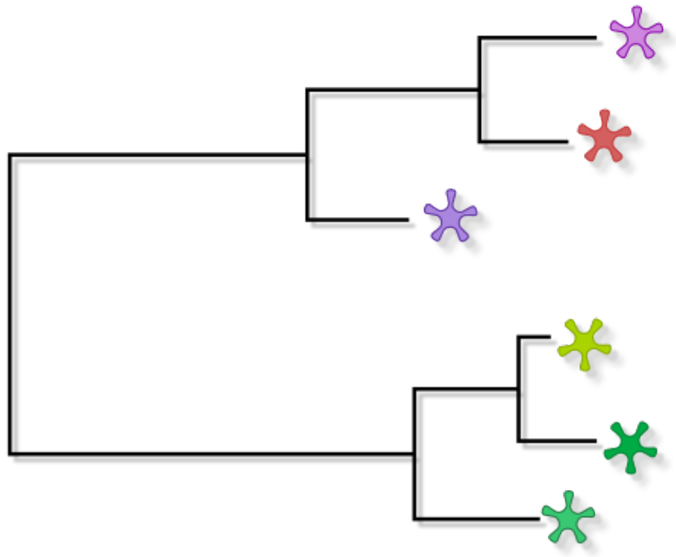
Using substitution patterns to reconstruct the evolutionary history



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

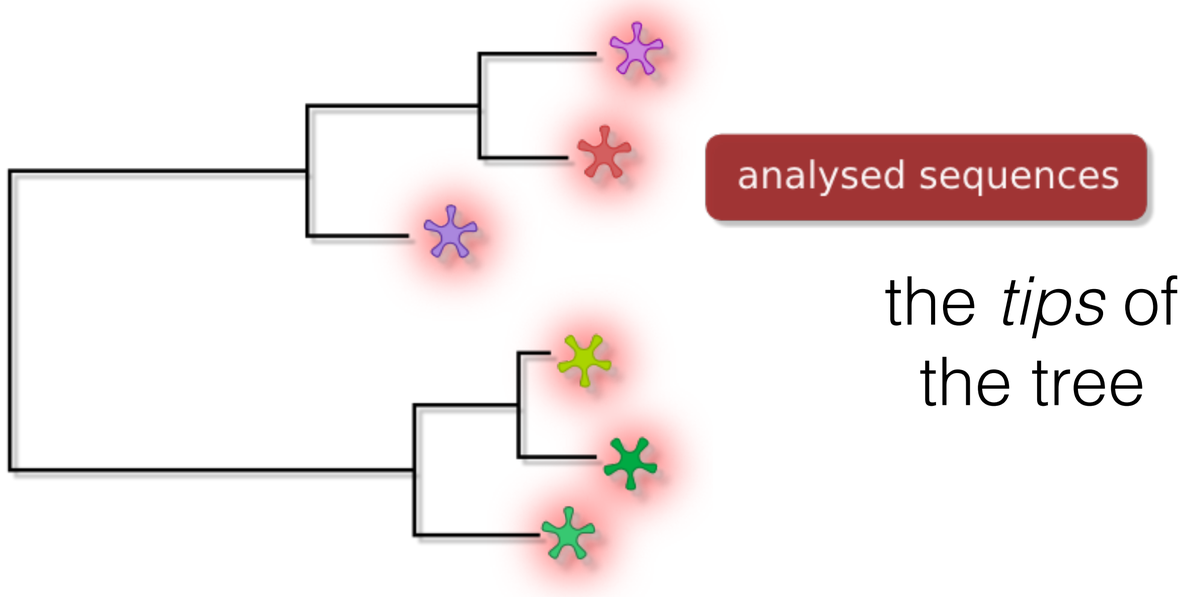
Original slide by Thibaut Jombart

Using trees to represent the evolutionary history



Original slide by
Thibaut Jombart

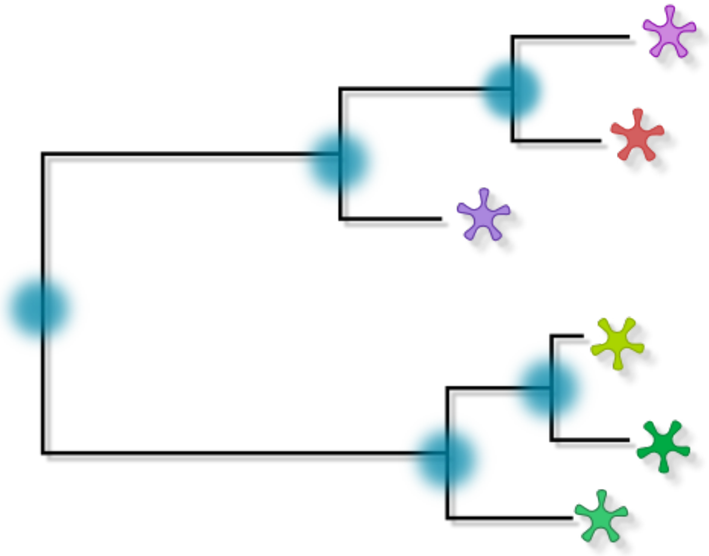
Using trees to represent the evolutionary history



Original slide by
Thibaut Jombart

Using trees to represent the evolutionary history

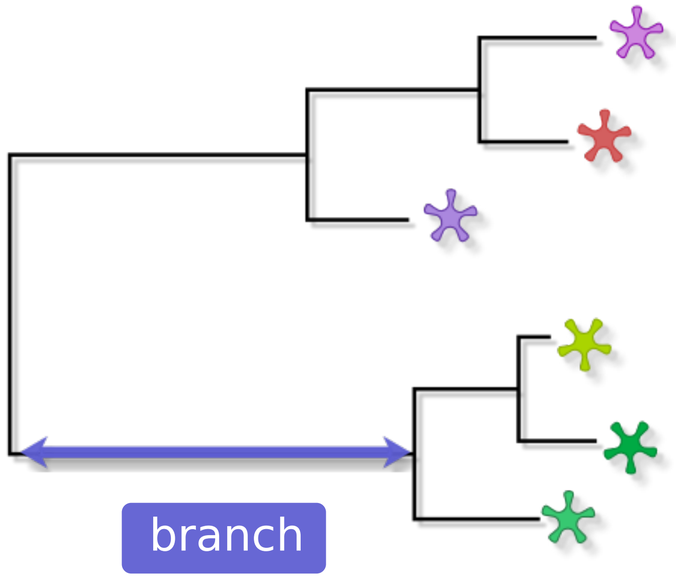
Most Recent Common Ancestors (MRCA)



the *nodes* of the tree

Original slide by Thibaut Jombart

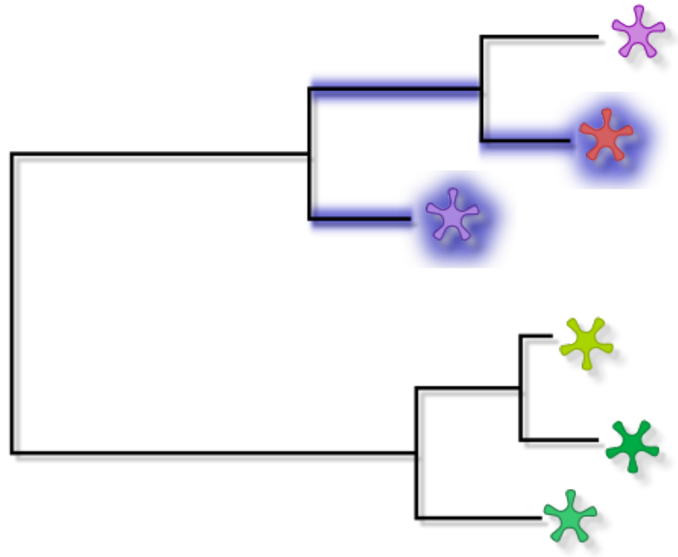
Using trees to represent the evolutionary history



length = amount of evolution

Original slide by
Thibaut Jombart

Using trees to represent the evolutionary history



distances between tips

"*patristic*" distance: sum of branch lengths

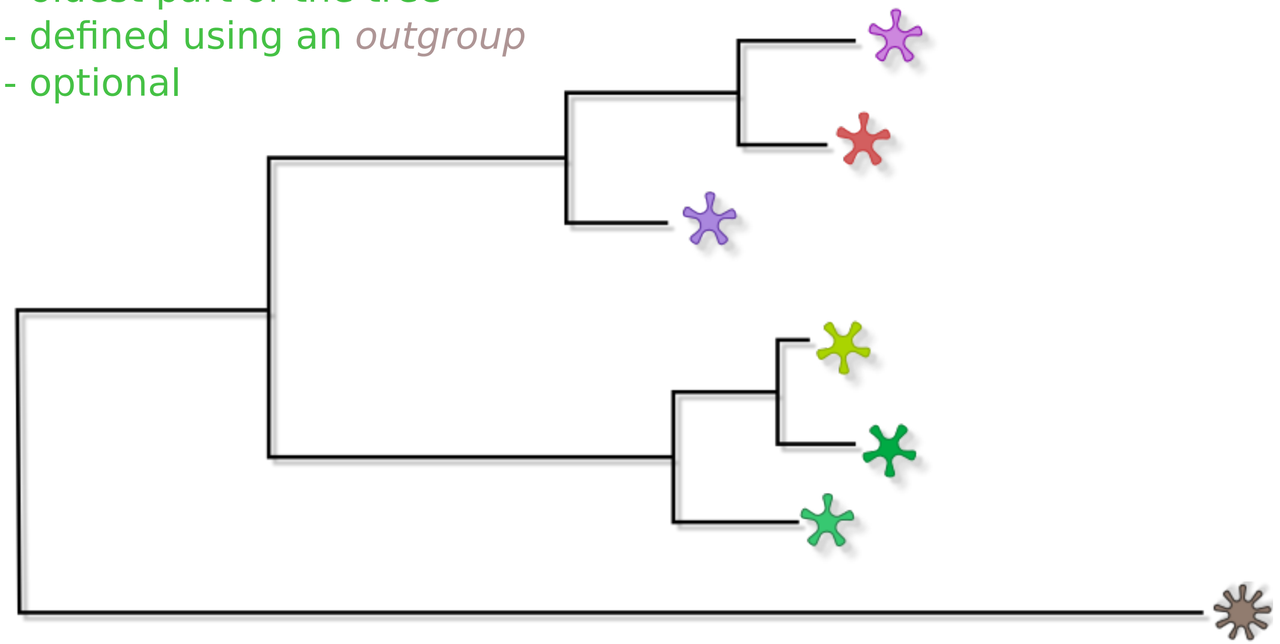
Original slide by
Thibaut Jombart

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

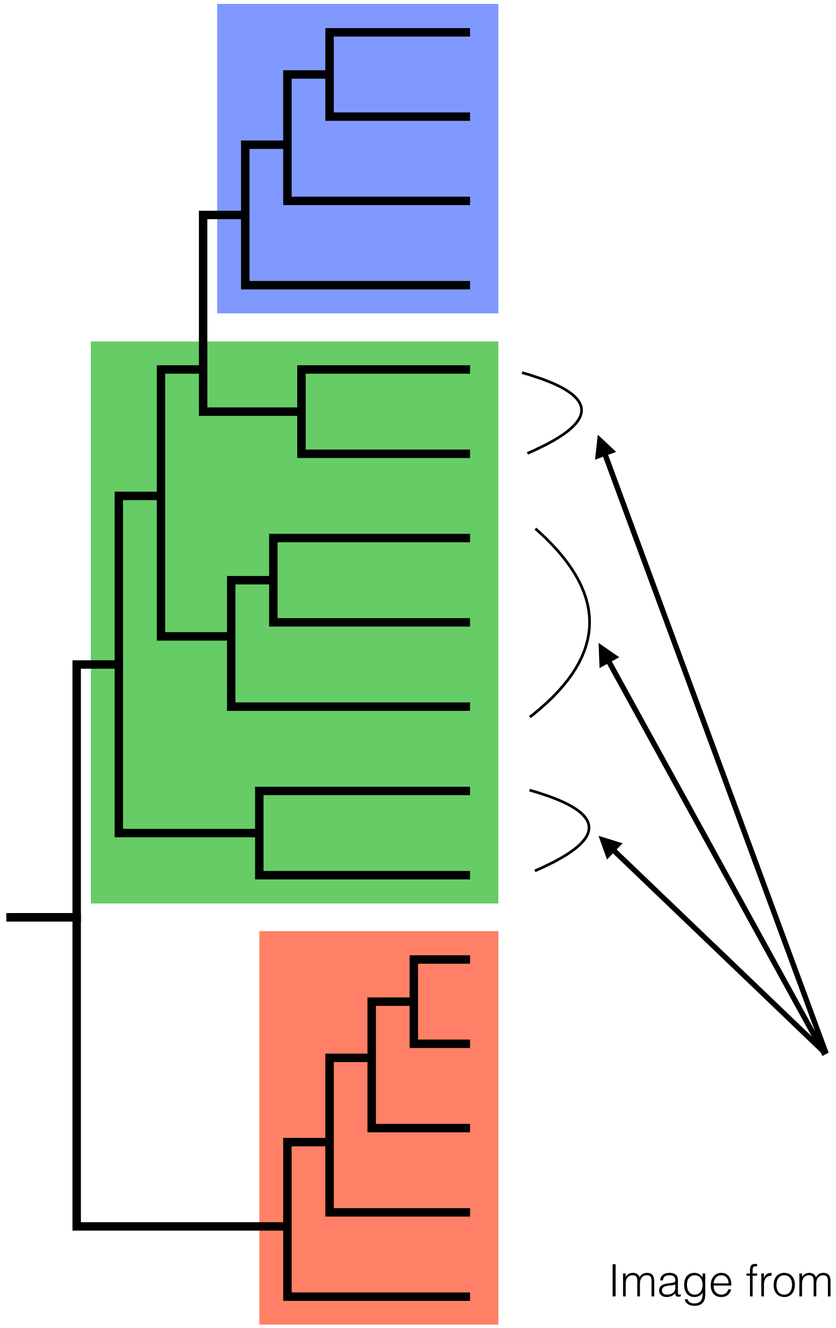
Using trees to represent the evolutionary history

Root

- oldest part of the tree
- defined using an *outgroup*
- optional



Original slide by
Thibaut Jombart



A *clade:* a group of organisms that consists of a common ancestor and all its descendants. a.k.a. a *monophyletic group*, a.k.a. a *monophylum*. The adjective is *monophyletic*. The blue and red boxes show two clades in this phylogeny (NB more clades are exist).

The group of tips in green is not monophyletic; it is called *paraphyletic*.
The set of green tips can be split into smaller groups, each of which is a clade; the minimum number needed is three, as illustrated.

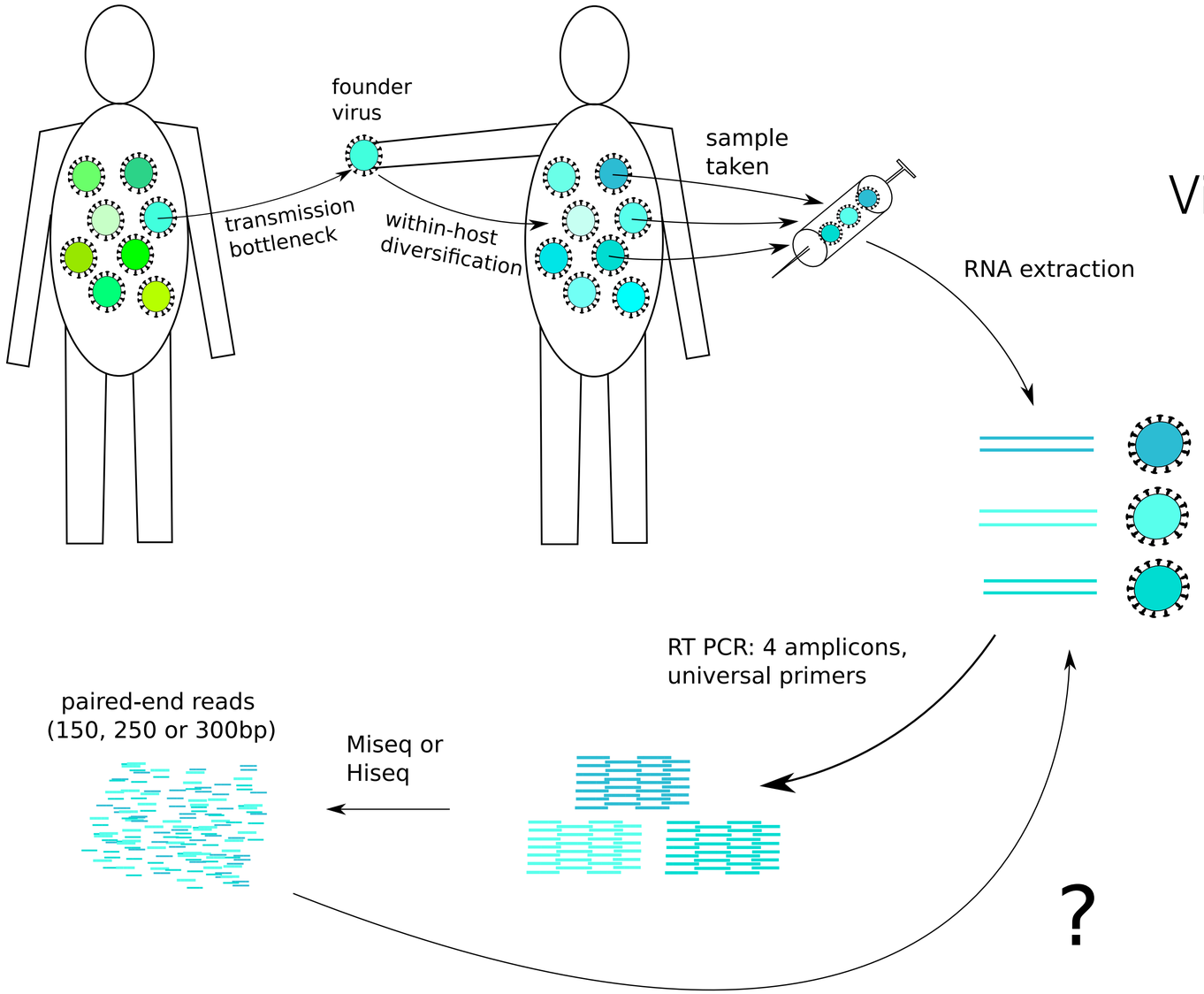
Image from https://en.wikipedia.org/wiki/Clade

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.

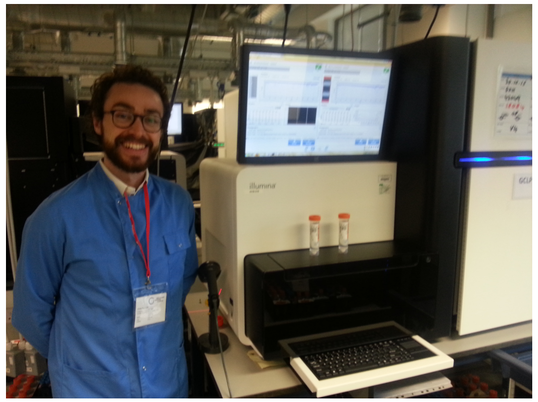
Our HIV Data*



Cornelissen *et al.*
Virus Research 2016

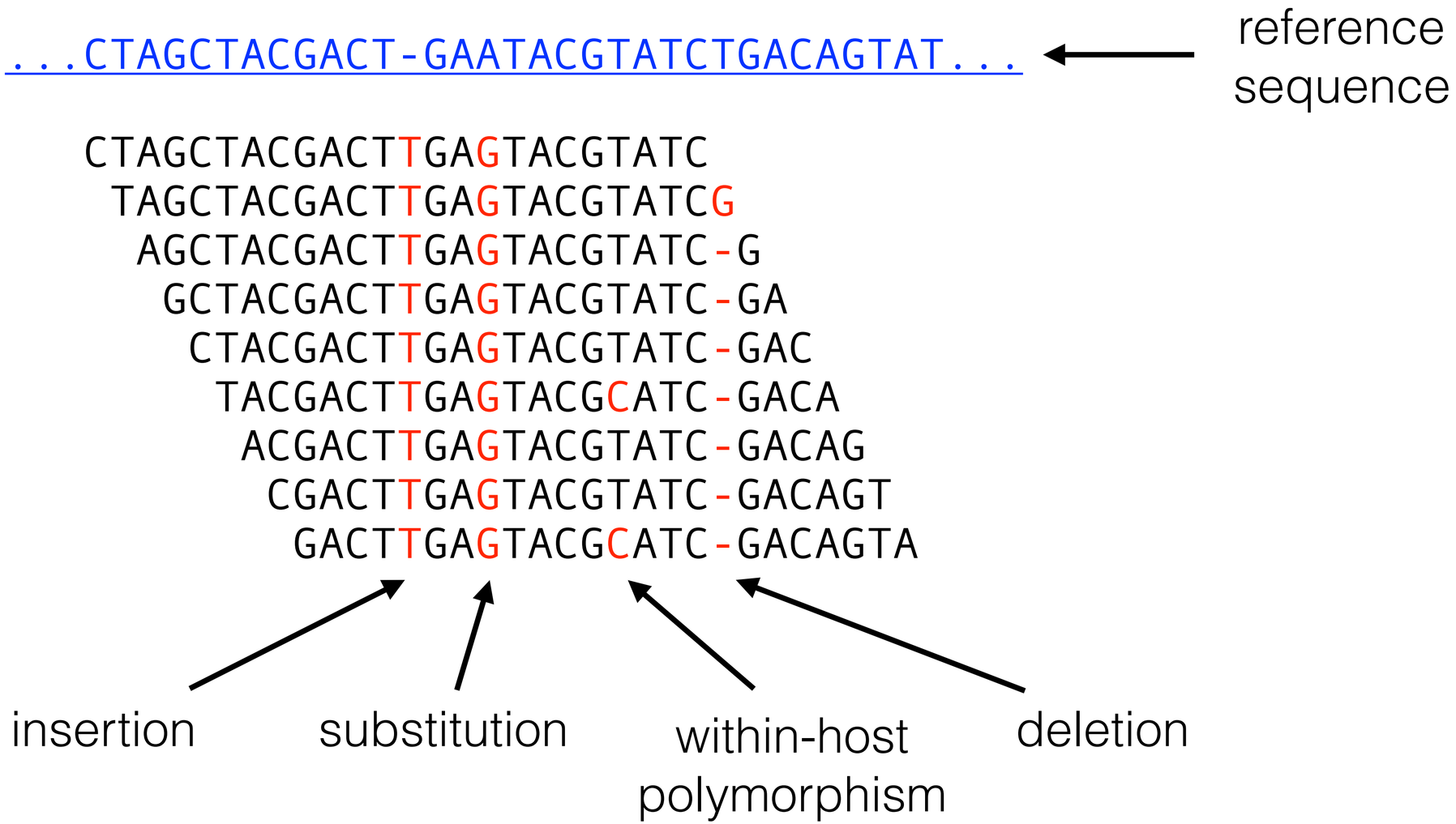
Gall *et al.*
J. Clin. Microbiol.
2012

* until now. See Bonsall, Golubchik et al. bioRxiv Aug 24th



↑ visiting, not helping

Mapping Reads to a Reference



Mapping Reads to a Reference

```
...CTAGCTACGACT-GAATACGTATCTGACAGTAT...
   ???GCTACGACTTGAGTACGTATC-GACA???
  CTAGCTACGACTTGAGTACGTATC
   TAGCTACGACTTGAGTACGTATCG
    AGCTACGACTTGAGTACGTATC-G
     GCTACGACTTGAGTACGTATC-GA
      CTACGACTTGAGTACGTATC-GAC
       TACGACTTGAGTACGCATC-GACA
        ACGACTTGAGTACGTATC-GACAG
         CGACTTGAGTACGTATC-GACAGT
          GACTTGAGTACGCATC-GACAGTA
```

consensus sequence, requiring 4X coverage (just as an example)

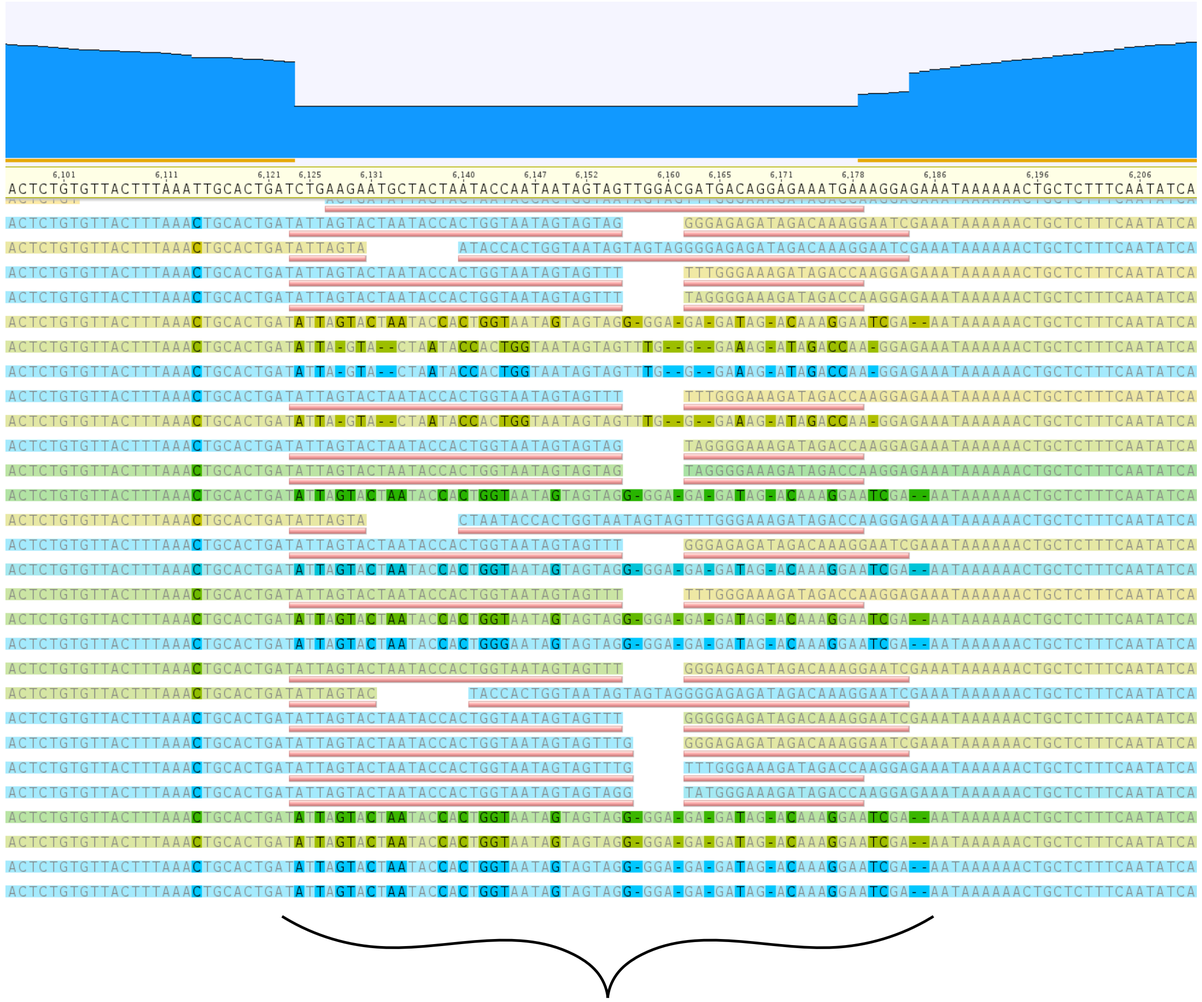
Note: Sanger sequencing, which was usually used when sequencing HIV until recently, only produces a consensus sequence. Having all these reads is very useful: stay tuned!

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**.

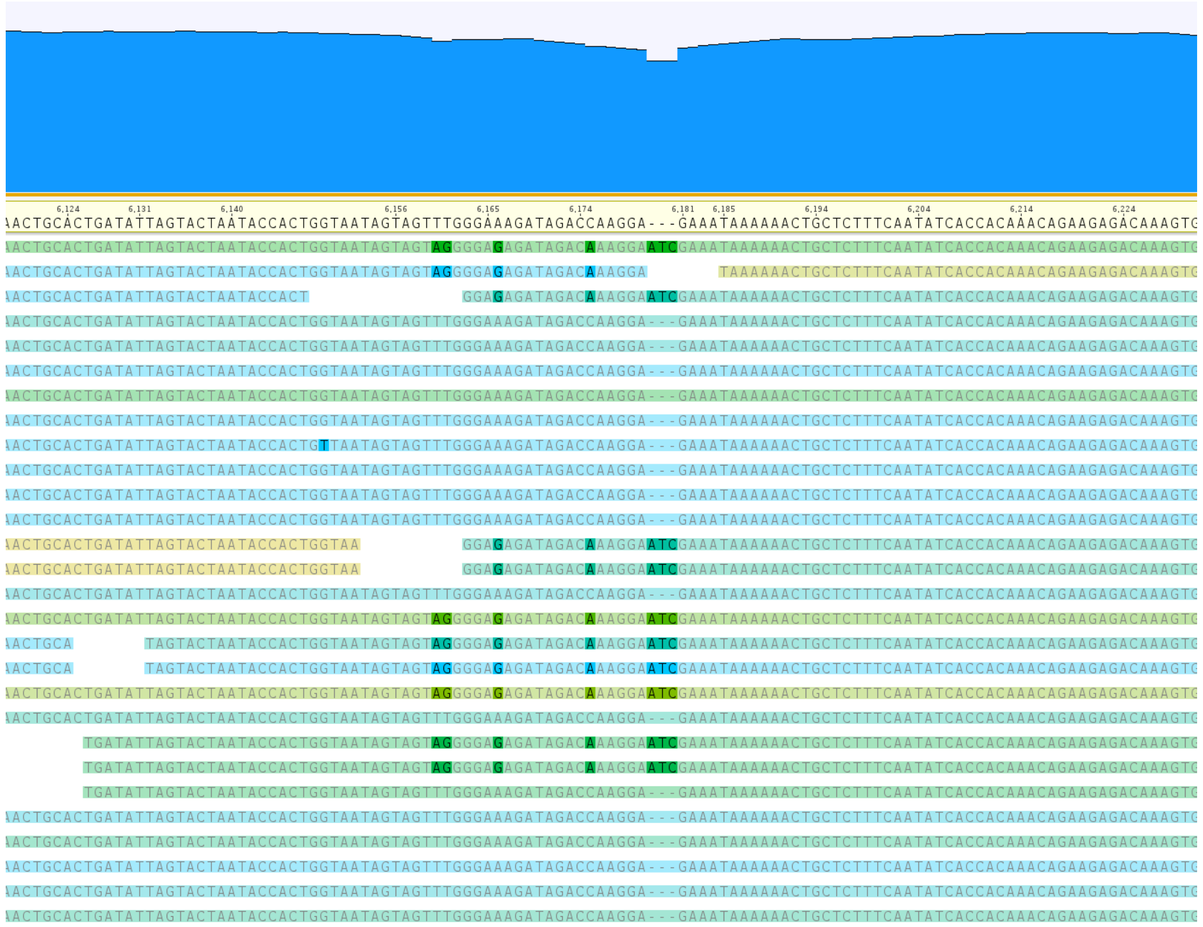
e.g. use the same reference for mapping for a group of individuals: bias their viruses to look similar to each other
→ false inference of transmission.

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

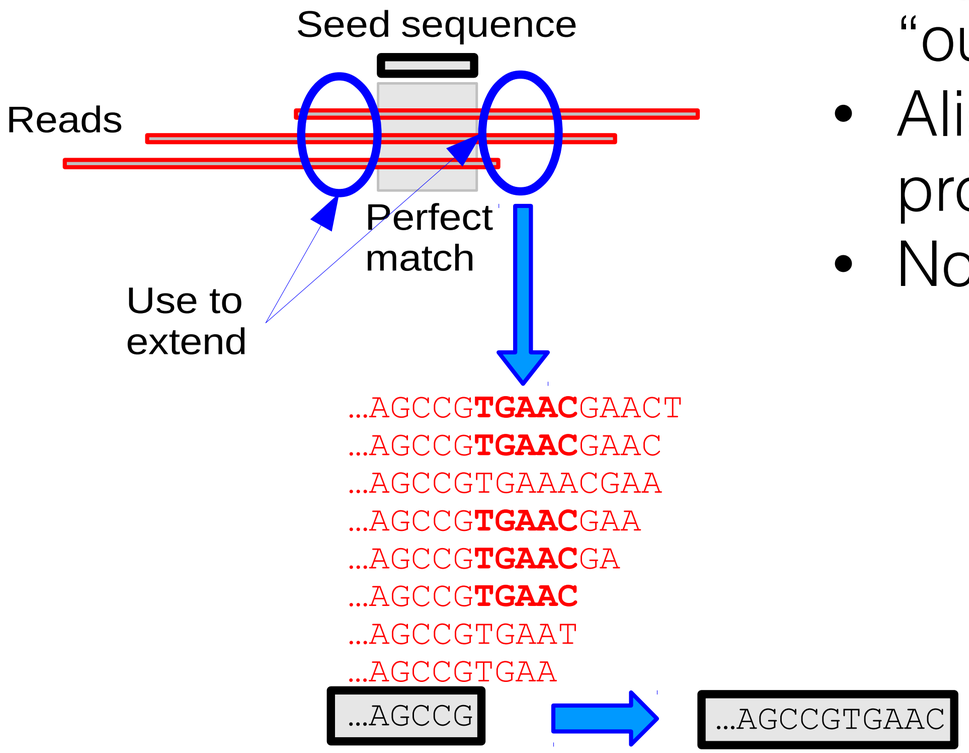
inappropriate reference ← same reads & mapping algorithm → appropriate reference



13 base miscalls, a false insertion and a false deletion



De Novo Assembly



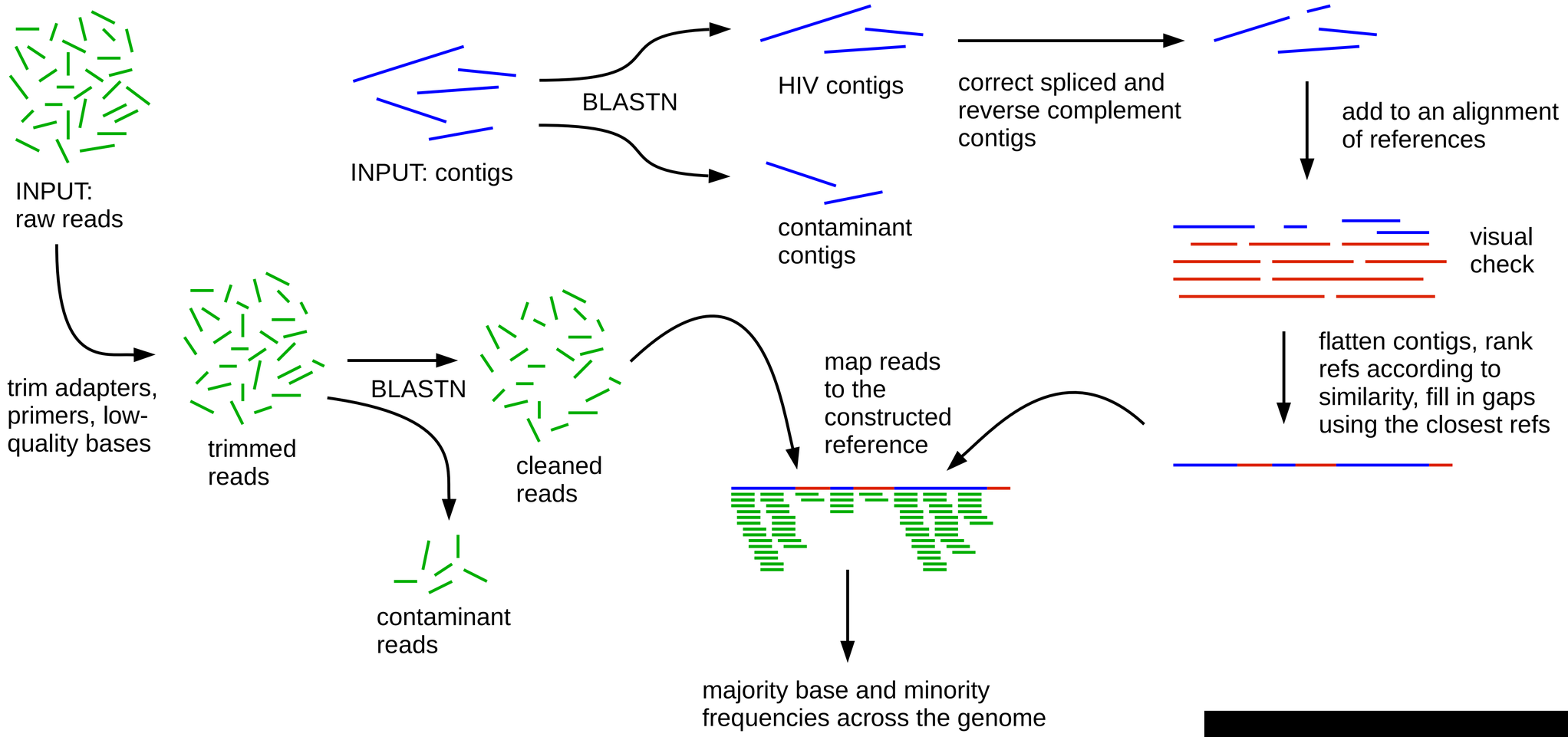
- e.g. IVA (*Hunt et al., Bioinformatics 2015*): on HIV, "outperforms all other virus de novo assemblers".
- Align the reads to themselves, iteratively extending, producing a number of summary sequences.
- No need for a reference: no bias.

but...

- Assembly failure (no contigs)
- Assembly errors (incorrect contigs)
- No minority variant information (contigs as summaries of reads)

Figure courtesy Martin Hunt

`shiver` - Sequences from HIV Easily Reconstructed



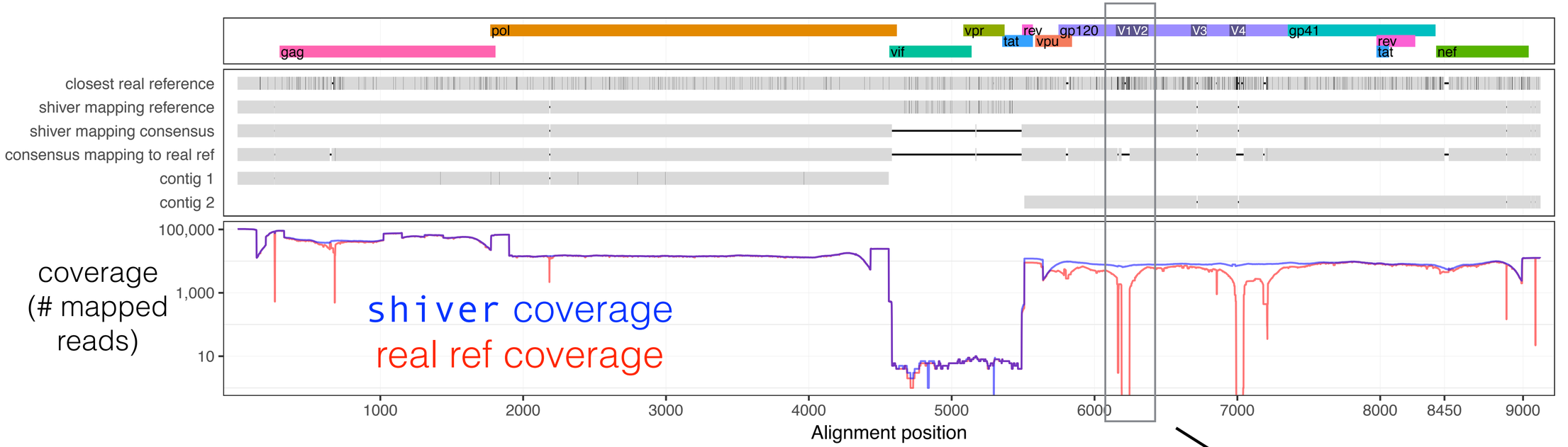
Wymant *et al.*, Virus Evolution 2018

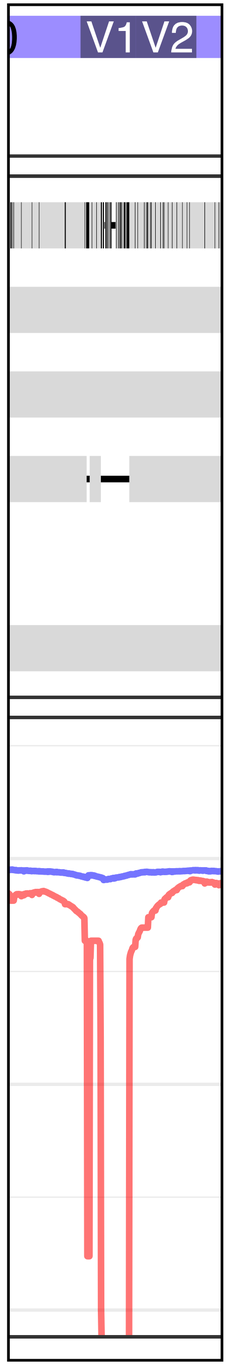
github.com/ChrisHIV/shiver

Works for other viruses too!

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

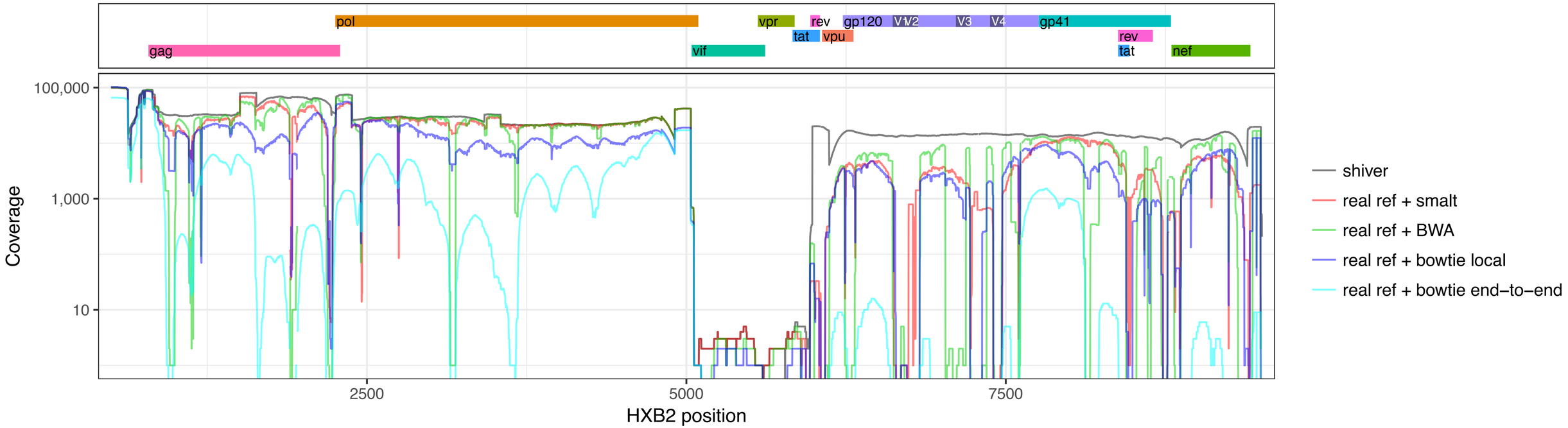
Map to `shiver`'s constructed reference,
also map to the closest of 3259 real references (from database),
compare.





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!).

Sometimes mapping to the closest real reference really sucks.



Application of `shiver` for 3 HIV projects led by Christophe:

Investigate the viral-genetic basis of virulence

BEEHIVE: N ~ 3,500
PopART phylo: N ~ 6,000 currently, 9,000 eventually
PANGEA(2): N ~ 13,000 currently, 23,000 eventually

Help interpret the effect of a population-level test-and-treat intervention in Zambia and South Africa

Generate new insights into transmission dynamics in generalised epidemics in Africa

Tanya Golubchik