How to interpret a viral phylogeny

a PANGEA webinar

Christophe Fraser
Using within-host diversity to infer direction of transmission: Equivalent to inferring ancestral state of virus populations.
Inferring ancestries in early HIV diversification and spread:

Faria et al, Nature 2014
Why do we represent relatedness of viruses as trees?

Why not ‘networks’, ‘clusters’, ‘maps’?
A: because HIV is a biological replicator.
Day 0

A single virus

Time
On average, each replication cycle takes ~2 days, and the virus acquires >1 mutation
In tracing an ancestral lineage, we only keep track of viruses, whose descendants survive to become part of our sample of interest.
Descendants are not part of sample

Descendants are part of sample

Replication

Time
Descendants are not part of sample

Descendants are part of sample

Replication

Day 4

Time
Replication

Day 4

Descendants are not part of sample

Descendants are part of sample

Time
Day 6

Replication

Descendants are not part of sample

Replication

Descendants are part of sample

Replication

Descendants are not part of sample

Time
What happens if more than one virus has descendants in a sample?
Replication

Descendants are not part of sample

Descendants are part of sample

Day 4

Time
Replication

Descendants are part of sample

Replication

Descendants are not part of sample

Replication

Replication

Replication

Day 6

Time
Tracing an ancestral lineage consists of keeping track of which viruses leave behind descendants in the sample.
Tracing an ancestral lineage consists of keeping track of which viruses leave behind descendants in the sample.
We ignore viruses that don’t leave behind descendants:
We ignore viruses that don’t leave behind descendants:
The history of the ancestral lineage has splits in time when more than one virus leaves behind descendants:

On day 2, the ancestral lineage splits

Day 0  Day 2  Day 4  Day 6

Time
Because ancestral lineages are constructed backwards in time from the present, the splits are often called ‘coalescences’.

Looking backwards from day 6, the two lineages coalesce on day 2.
We usually infer this process indirectly from a sample:

Our study consists of sampling these two viruses.
This object is the phylogeny of the two sampled viruses. We have to infer this from data obtained from the two sampled viruses in the boxes.
**A simple phylogeny**

The most recent common ancestor, a.k.a. 'the root'

The sample, a.k.a. 'the tips'

Day 0                Day 2                Day 4                Day 6

Time
A simple phylogeny

- The most recent common ancestor, a.k.a. ‘the root’
- The moment when the lineage splits, a.k.a. the internal node, a.k.a. the coalescence of the two present day lineages
- The branches during which the viruses are busy replicating
- The sample, a.k.a. ‘the tips’

Day 0  Day 2  Day 4  Day 6

Time
Inferred ancestor of viruses from A

Inferred ancestor of viruses from B

When person A was likely infected
Correctly interpreting trees is not difficult, but requires a bit of practice as they are a bit visually misleading.

Q: which are more closely related of the viruses 1, 2 & 3?
Correctly interpreting trees is not difficult, but requires a bit of practice as they are a bit visually misleading.

A: Viruses 2 and 3 share a more recent common ancestor than either shares with virus 1, and are expected to be more similar.
What makes our phylogeny uncertain is that we don’t observe it. Instead, we infer it from data collected only at the tips.
A simple phylogeny

The branches during which the viruses are busy replicating (and mutating)
Replication with mutation

Descendants are not part of sample

Descendants are part of sample

Replication with mutation

Time
Look at the genome

ACCATGAAC

Replication with mutation

ACCAGGAAC

Replication with mutation

AACATGAAC

Replication with mutation

AACCTGAAC

Replication with mutation

AACCTAAAC

TCCAGGAAC

Replication with mutation

TACAGGAAC

Time

ACCATGAAC
A substitution is a mutation that survives in the ancestral lineage.

Replication with mutation:

```
ACCATGAAC
```

```
ACCA
G
GAAC
```

```
A
A
CATGAAC
```

```
TCCAGGAAC
```

```
TCCA
G
GAAC
```

```
TACAGGAAC
```

```
AACATGAAC
```

```
AACATGAAC
```

```
AACATGAAC
```

```
AACCTGAAC
```

```
AACCTGAAC
```

```
AACCTAAAAC
```

Time

---
A substitution is a mutation that survives in the ancestral lineage.

Substitution rates are typically much lower than mutation rates (~10 substitutions/year versus ~1 mutation per replication cycle & 150 cycles per year).

- Most mutations are harmful to the virus.
- The virus sometimes gets ‘stuck’ in a non-replicating latent state for years.
Phylogenetic algorithms infer the phylogeny based on molecular models of how the viruses accumulate substitutions (e.g. the relative rate of A>C versus T>G, etc.).

Example software include PhyML, RaxML, IQTree, Mr Bayes.
An example real phylogeny:

The colors represent patients. The tips represent a virus. The number represent the number of times the same genotype was found.

The letters represent an inferred ancestral virus for all the viruses in a patient.

The scale bar is in units of substitutions per site.

Wymant et al 2018
Phylogenies can provide information at very different scales.
A tree can be dated if you have enough data to infer rates of substitution.

Example software include BEAST, TreeDater, TreeTime, Least Squares Dating (LSD)
This phylogeny has been rotated for clarity. The past is at the top.
In practice, inferring phylogenies is full of uncertainty.

If a number or fraction appears here, then the tree is a summary of many trees (bootstraps, posterior draws). This number is the fraction of trees that support the conclusion that the two samples share a more recent common ancestor than they do with any of the other samples in the study.
This phylogeny has been rotated for clarity. The past is at the top.
Interesting things may happen to viruses as they replicate:

**Transmission:** This virus or infected cell may be transferred to another person.

**Migration:** The person hosting this virus may move to another location.
Interesting things may happen to viruses as they replicate:

**Transmission:** This virus or infected cell may be transferred to another person.

**Migration:** The person hosting this virus may move to another location.

---

Time

---
A real example of transmission:

At some point along this branch, the virus in this lineage likely jumped from person E into person F.

The scale bar is in units of substitutions per site.

Wymant et al 2018
A real example of migration

At some point along this branch, the person/people carrying this virus moved from Kinshasa to the Americas.
The final really complicated effect: recombination messes up tree-like structure, creating a ‘ancestral graph instead’.

**Coinfection:** Both viruses happen to infect the same cell.

**Recombination:** During the next round of replication, the genome of both viruses is spliced together.
Luckily, HIV is not the only biological replicator that recombines. Some have been well studied.
Summary

• Phylogenetics provides powerful insights into dynamics of virus spread at different scales.

• Trees are a natural way to describe ancestry (recombination still challenging).

• Ancestral state reconstruction is the key link to epidemiology.

• We have talked a lot about phyloscanner, but there are many other tools and methods we should use with PANGEA to obtain insights.
Thank you.

Questions?