



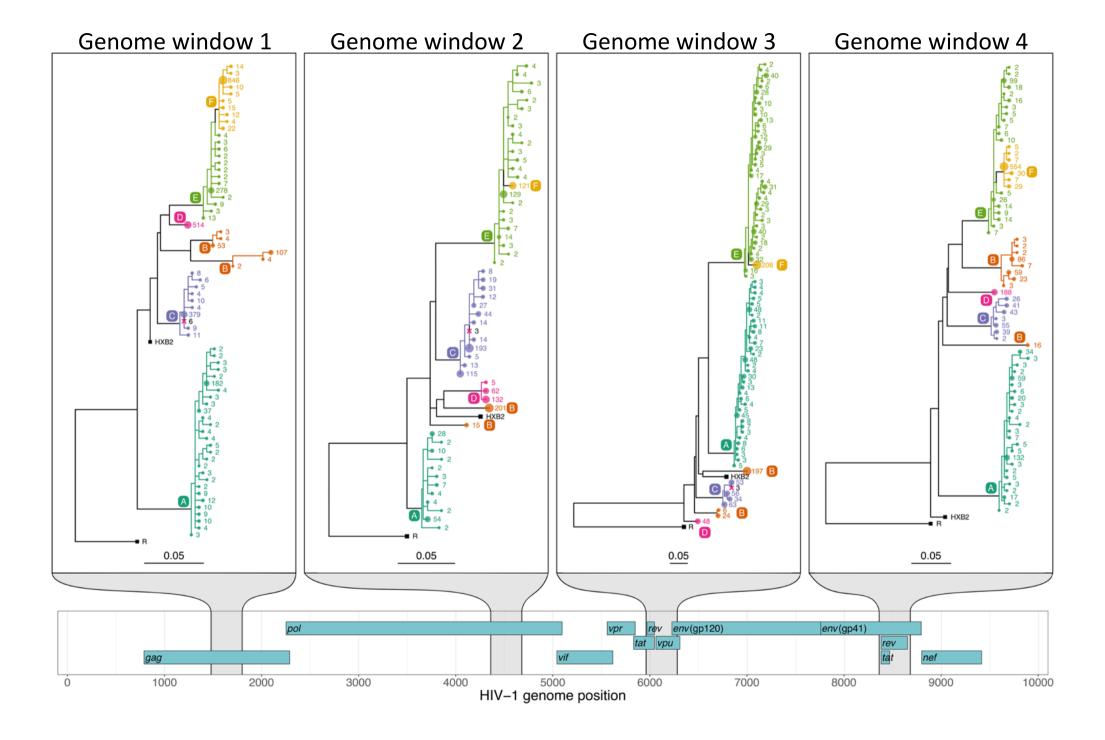




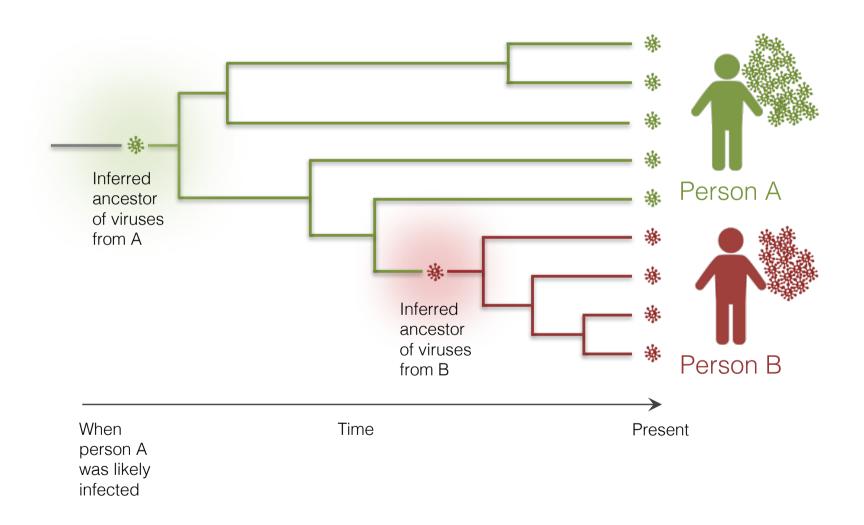
How to interpret a viral phylogeny

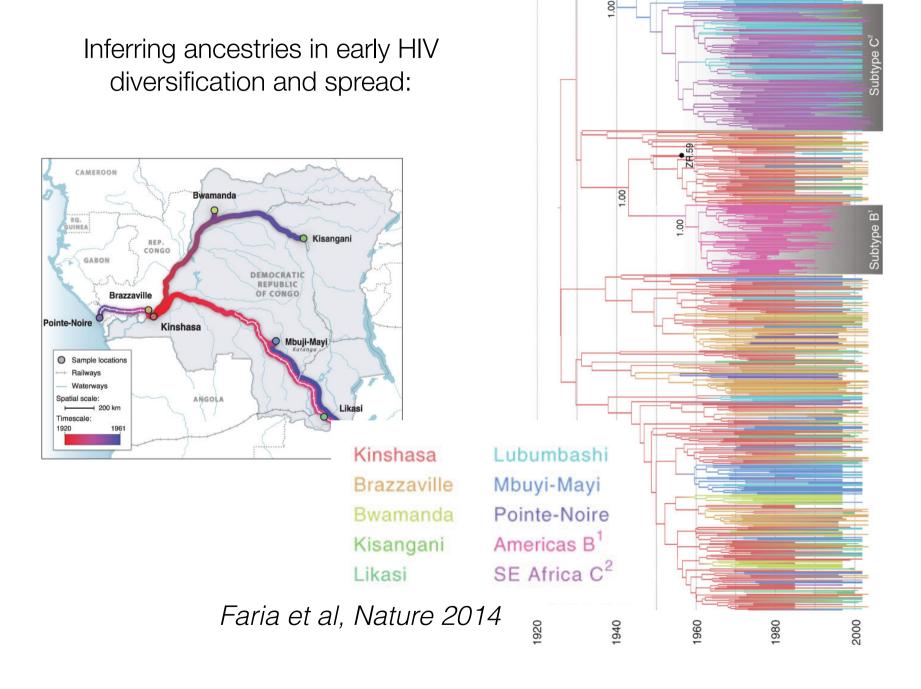
a PANGEA webinar

Christophe Fraser



Using within-host diversity to infer direction of transmission: Equivalent to inferring ancestral state of virus populations.

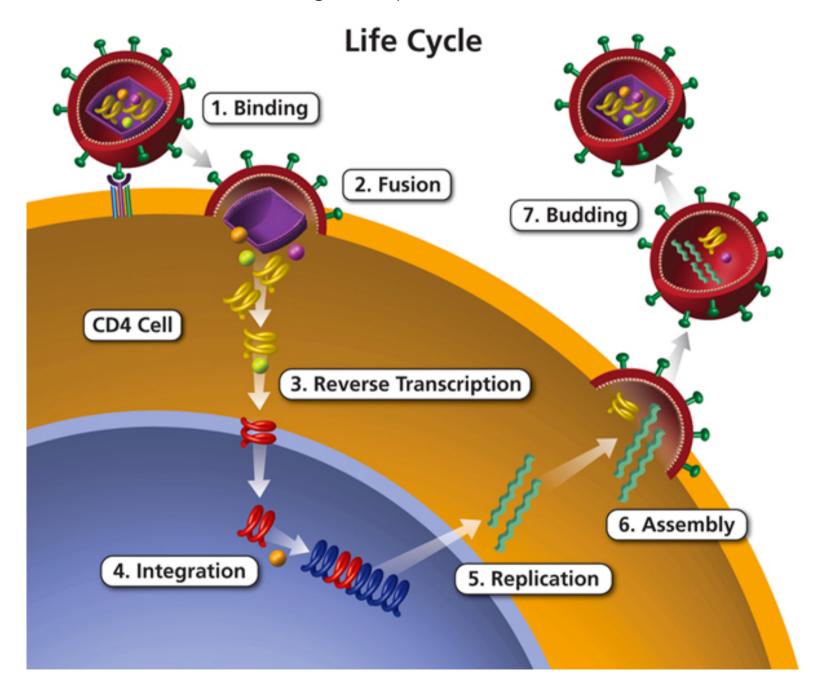




Why do we represent relatedness of viruses as trees?

Why not 'networks', 'clusters', 'maps'?

A: because HIV is a biological replicator.



A single virus







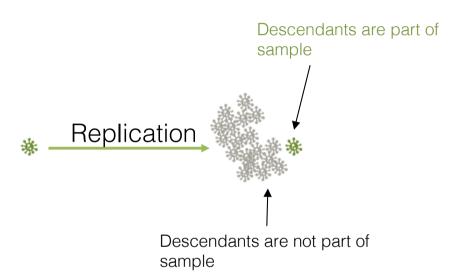
Day 2

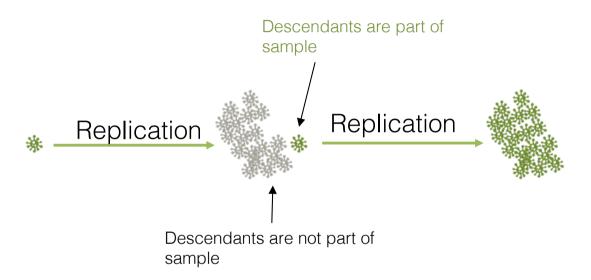
On average, each replication cycle takes ~2 days, and the virus acquires >1 mutation

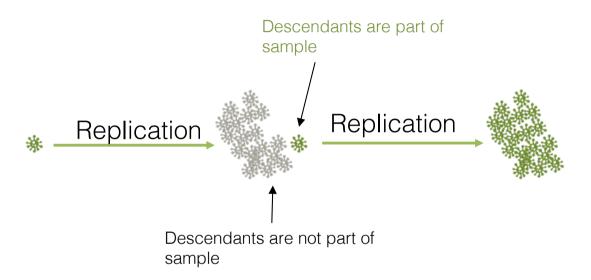


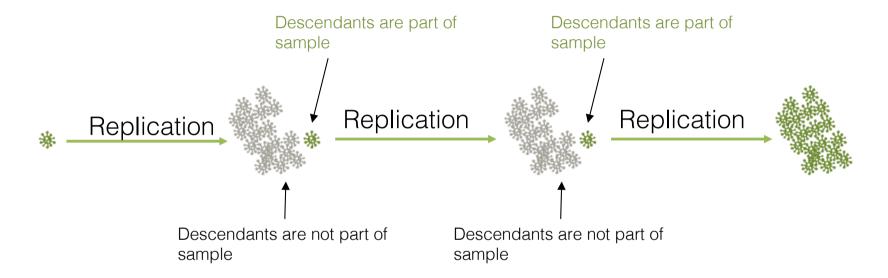
In tracing an ancestral lineage, we only keep track of viruses, whose descendants survive to become part of our sample of interest.



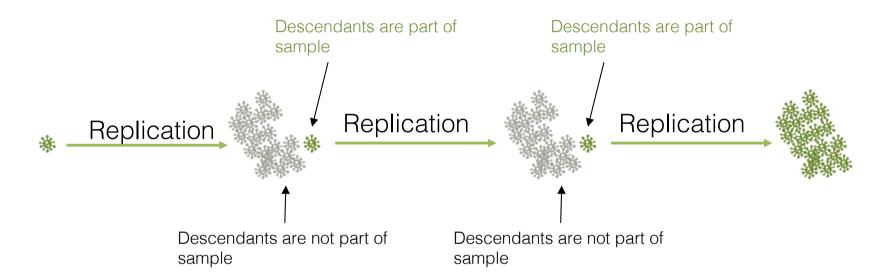




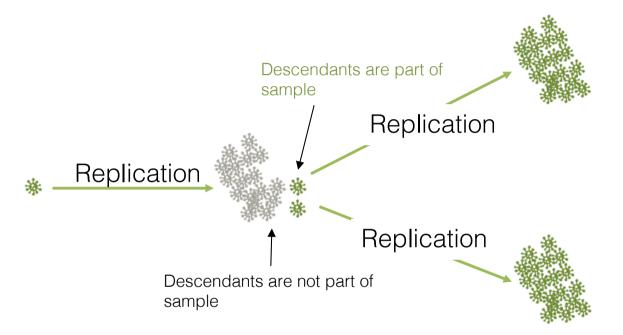




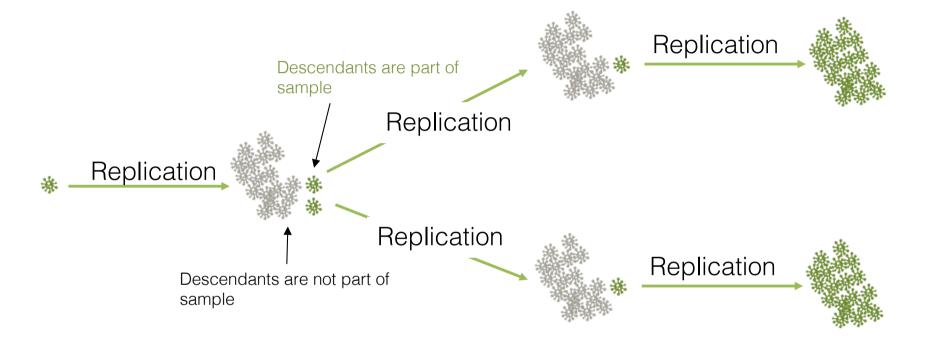
What happens if more than one virus has descendants in a sample?



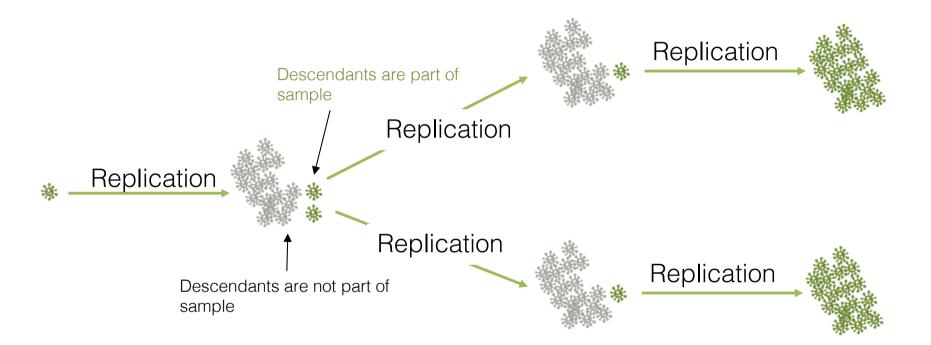
Day 4



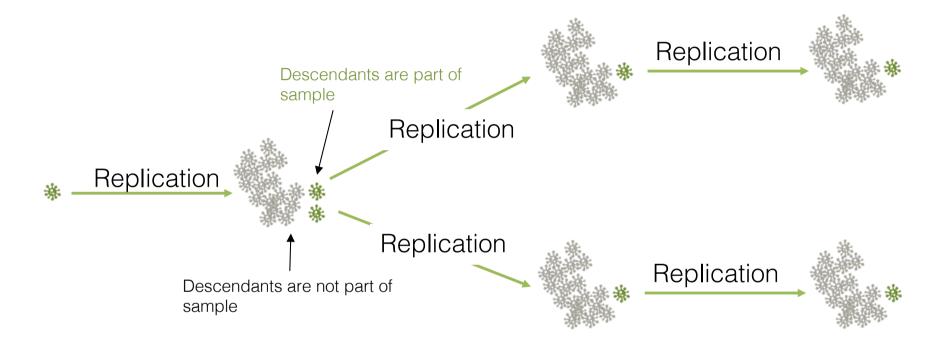
Day 6



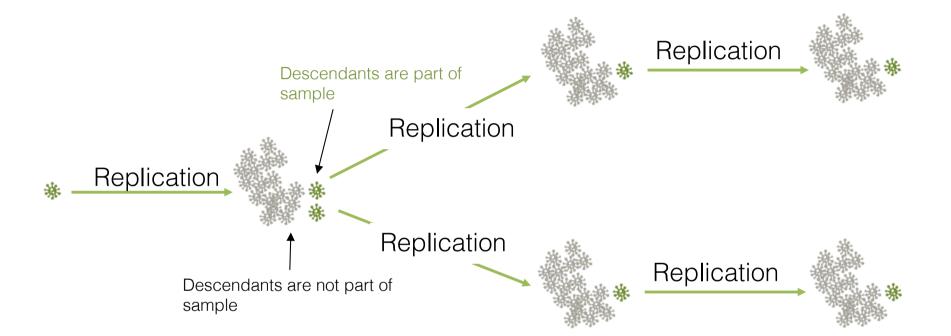
Tracing an ancestral lineage consists of keeping track of which viruses leave behind descendants in the sample



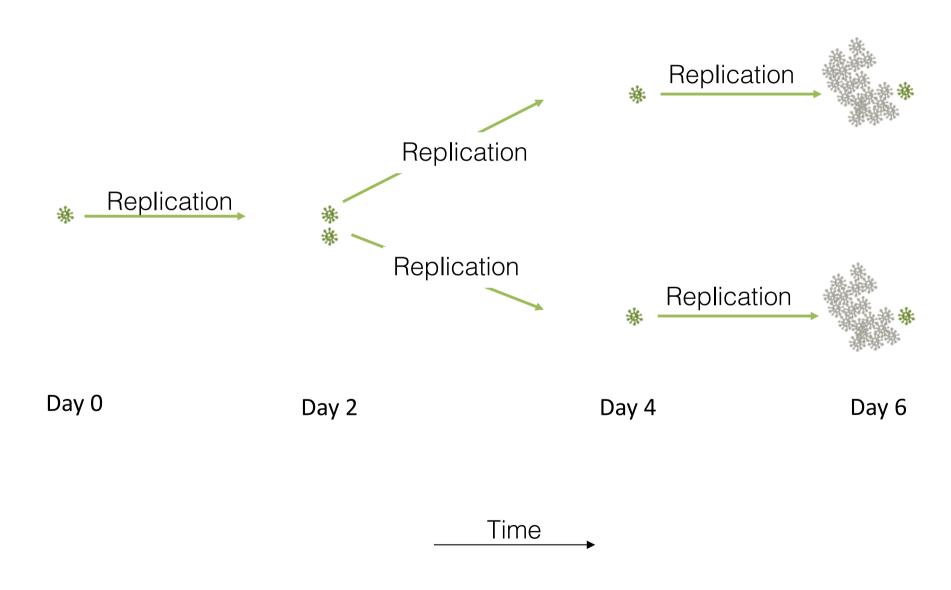
Tracing an ancestral lineage consists of keeping track of which viruses leave behind descendants in the sample.



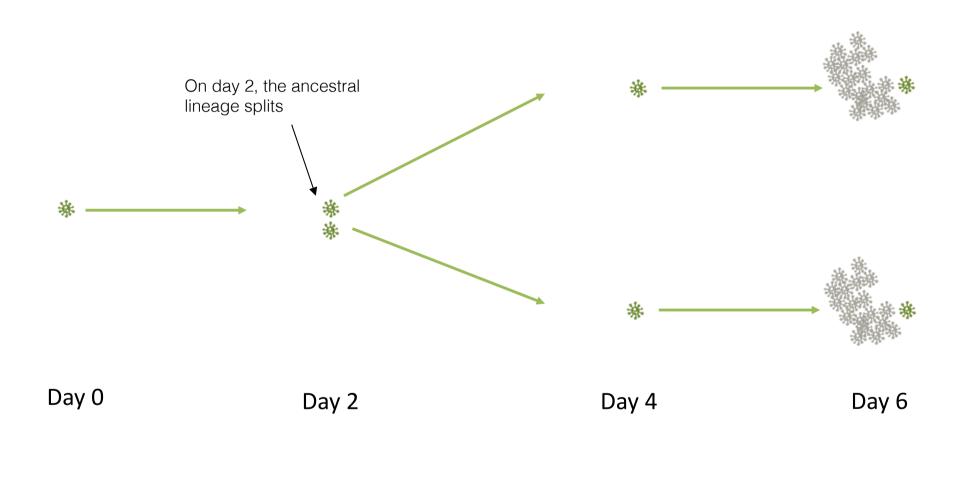
We ignore viruses that don't leave behind descendants:



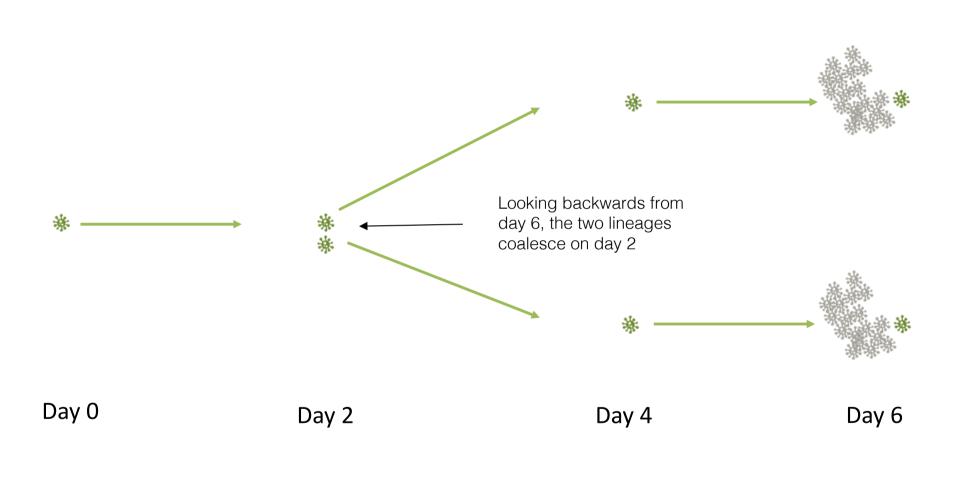
We ignore viruses that don't leave behind descendants:



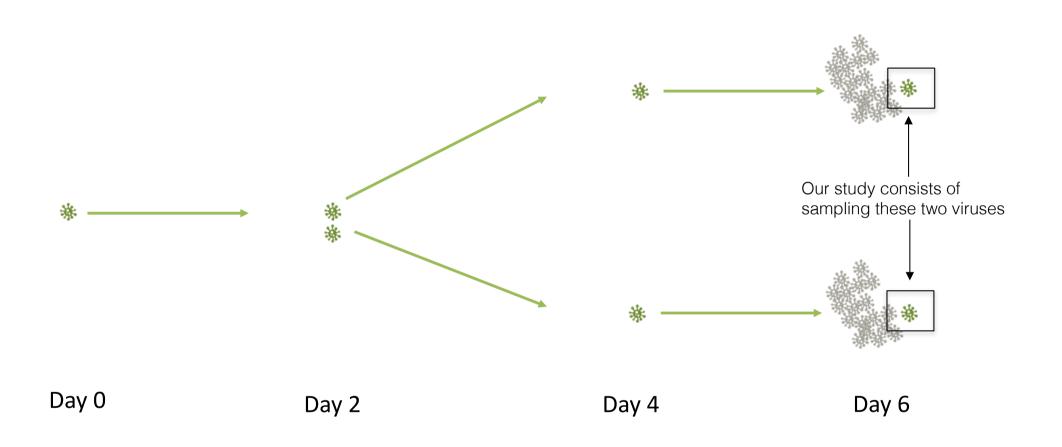
The history of the ancestral lineage has splits in time when more than one virus leaves behind descendants:



Because ancestral lineages are constructed backwards in time from the present, the splits are often called 'coalescences'

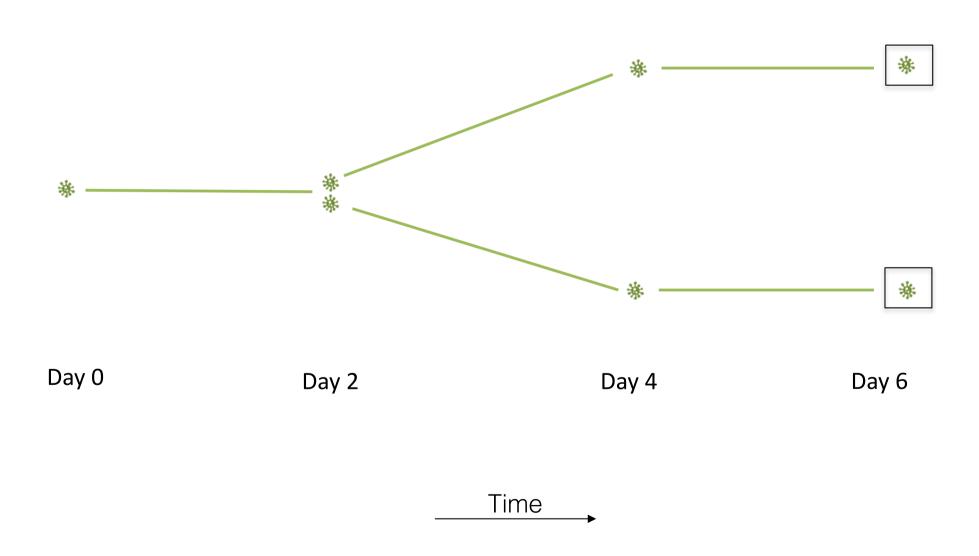


We usually infer this process indirectly from a sample:

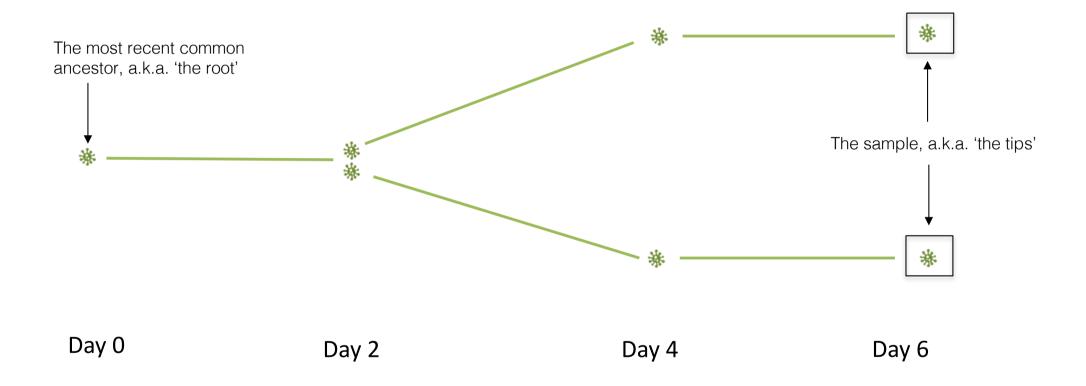


This object is the phylogeny of the two sampled viruses.

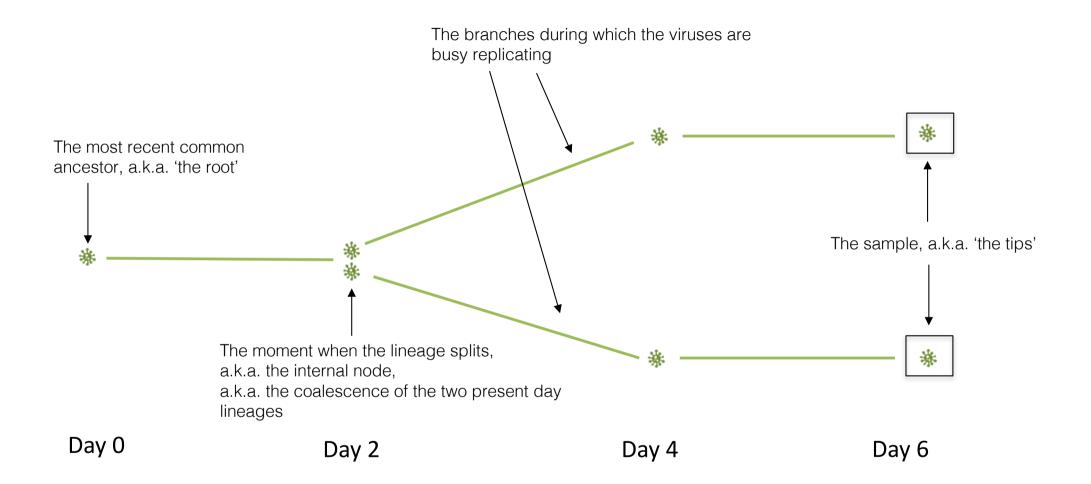
We have to infer this from data obtained from the two sampled viruses in the boxes.

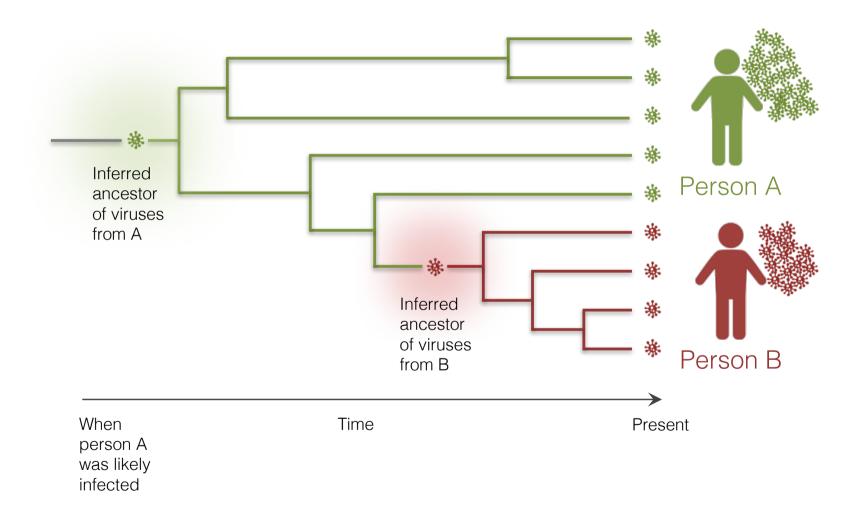


A simple phylogeny

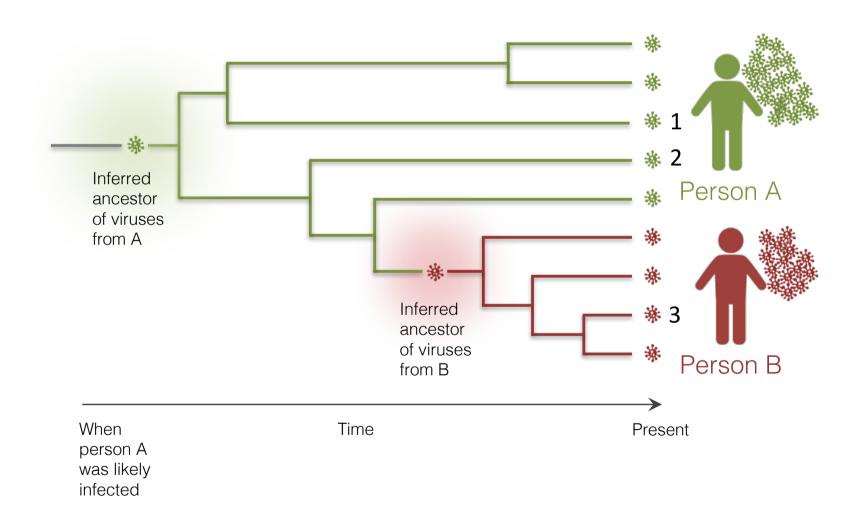


A simple phylogeny



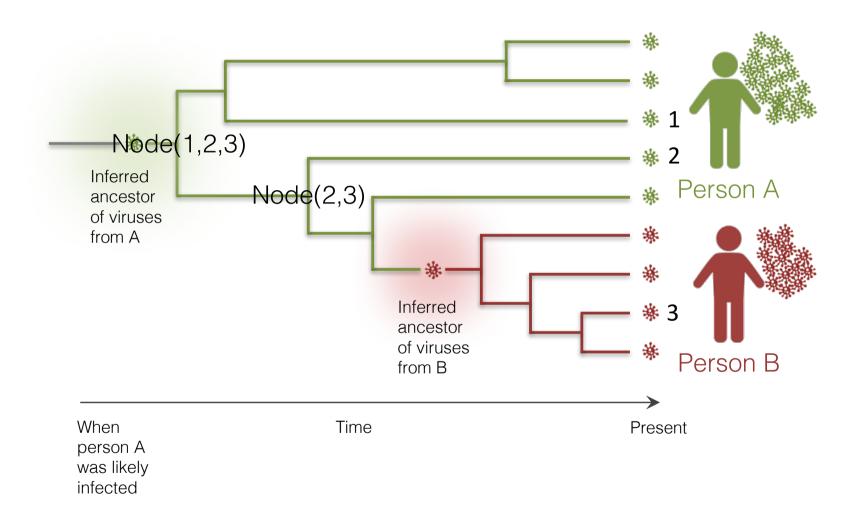


Correctly interpreting trees is not difficult, but requires a bit of practice as they are a bit visually misleading.



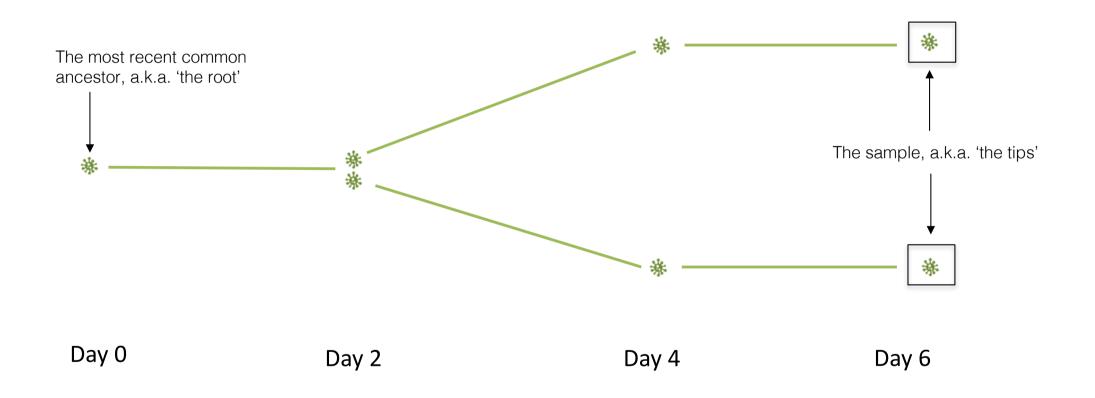
Q: which are more closely related of the viruses 1, 2 & 3?

Correctly interpreting trees is not difficult, but requires a bit of practice as they are a bit visually misleading.

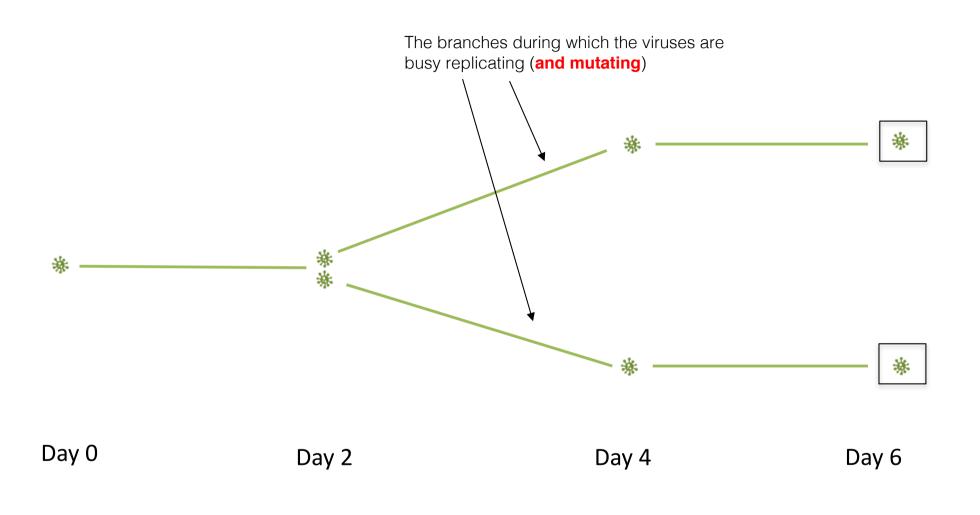


A: Viruses 2 and 3 share a more recent common ancestor than either shares with virus 1, and are expected to be more similar

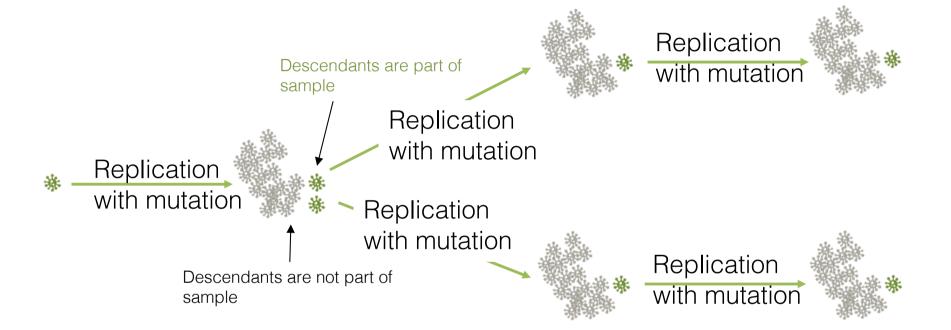
What makes our phylogeny uncertain is that we don't observe it. Instead, we infer it from data collected only at the tips



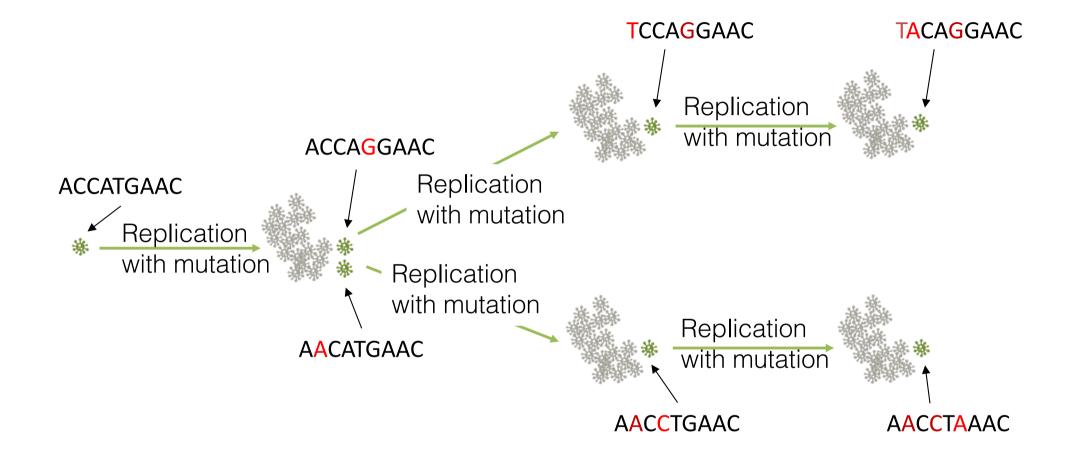
A simple phylogeny



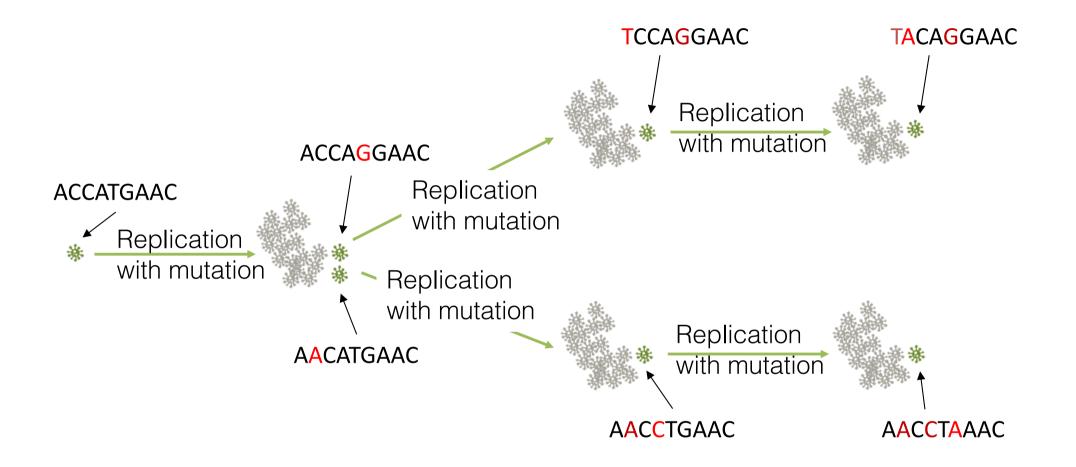
Replication with mutation



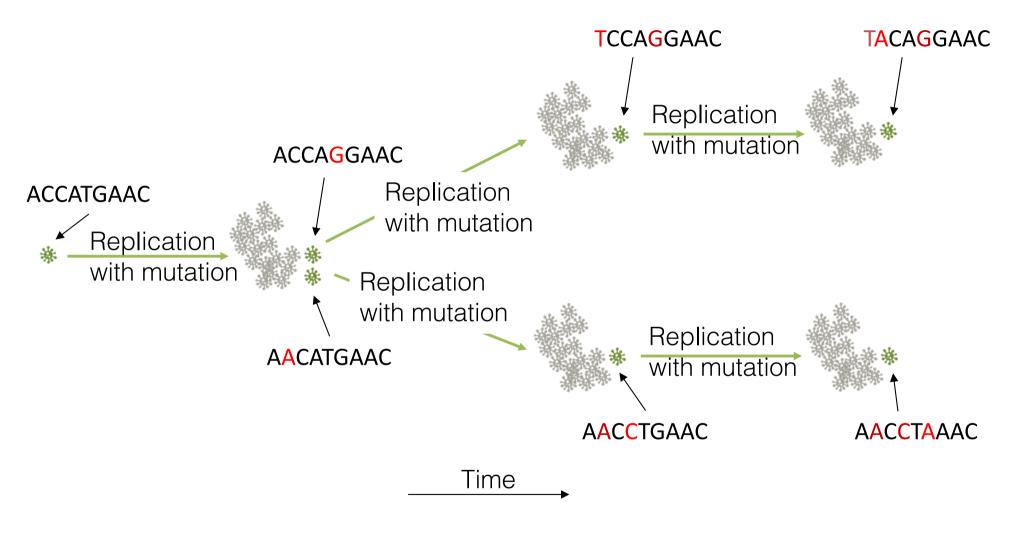
Look at the genome



A substitution is a mutation that survives in the ancestral lineage



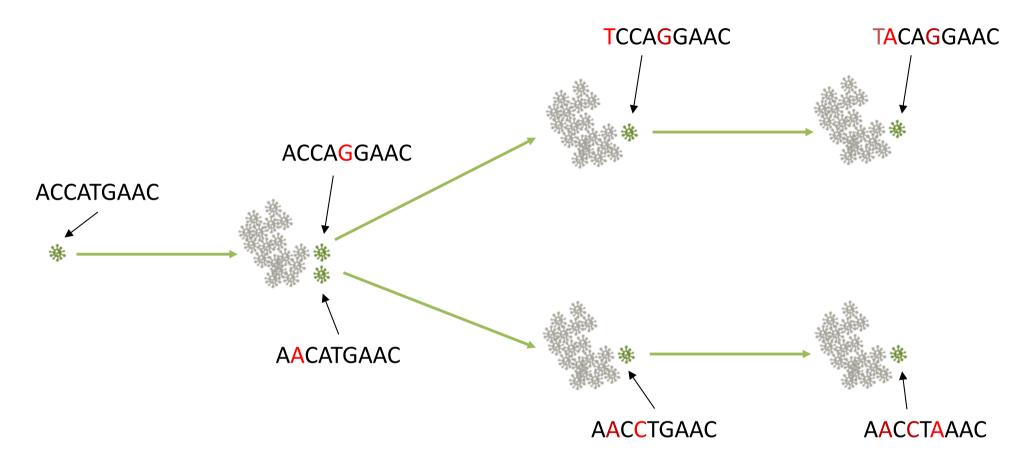
A substitution is a mutation that survives in the ancestral lineage



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

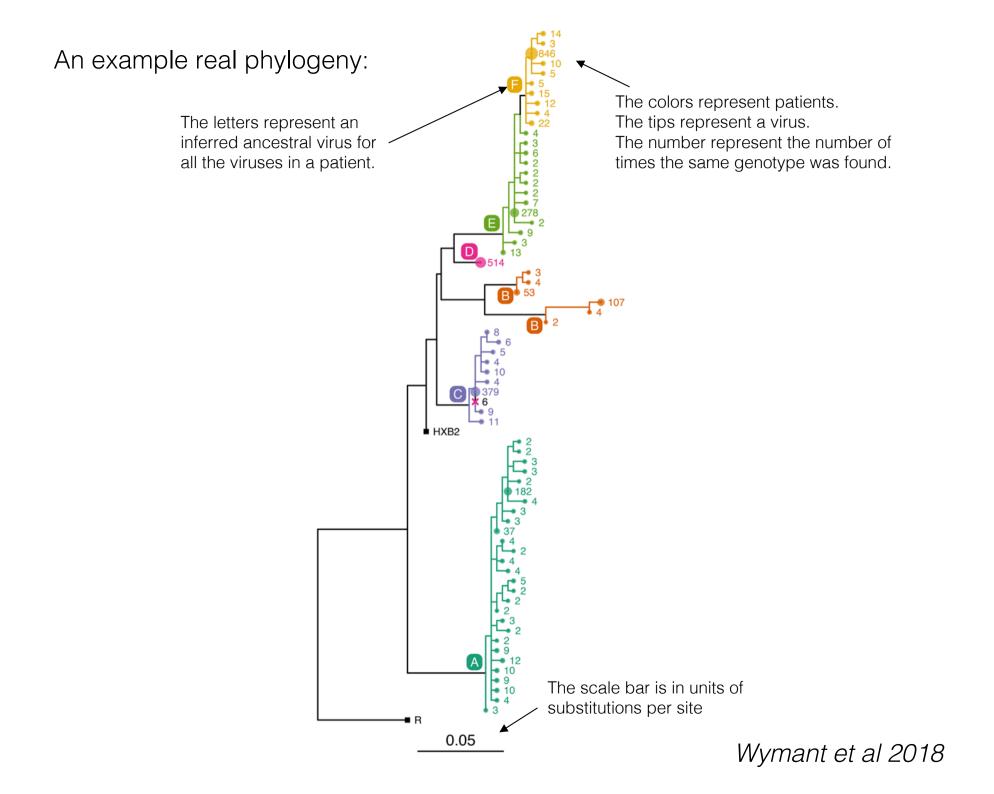
- Most mutations are harmful to the virus.
- The virus sometimes gets 'stuck' in a non-replicating latent state for years.

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.).

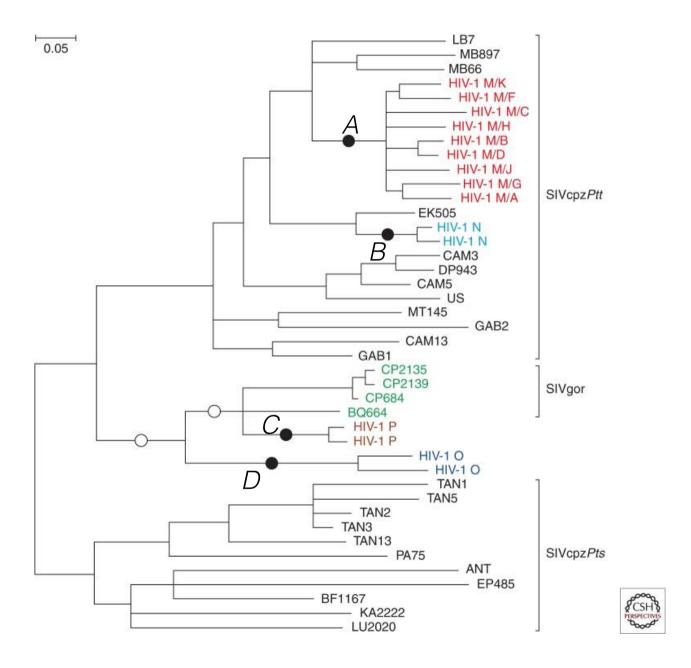


Example software include PhyML, RaxML, IQTree, Mr Bayes

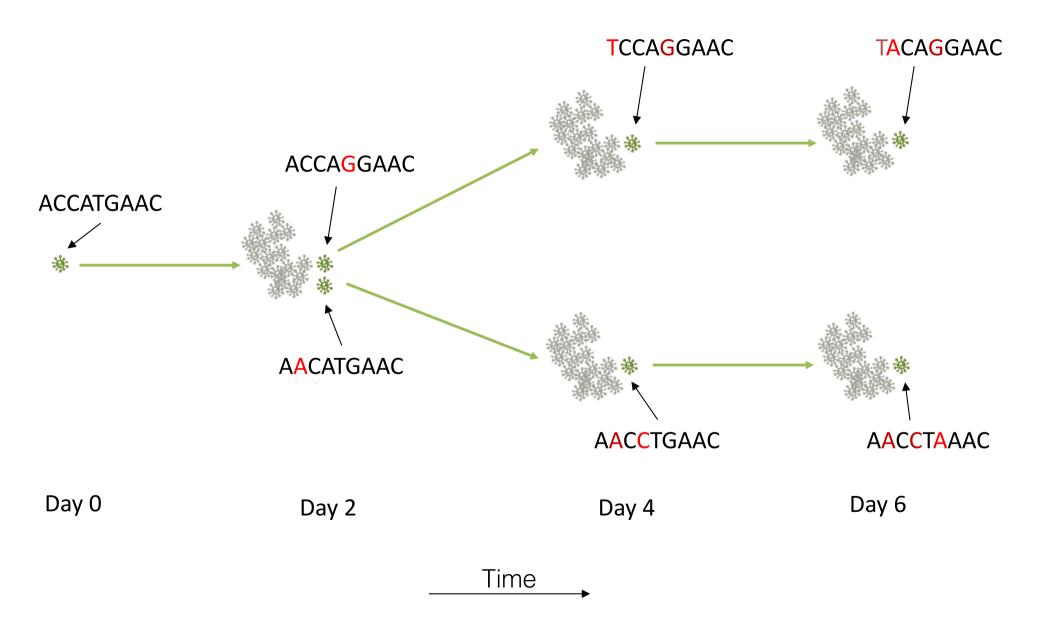




Phylogenies can provide information at very different scales

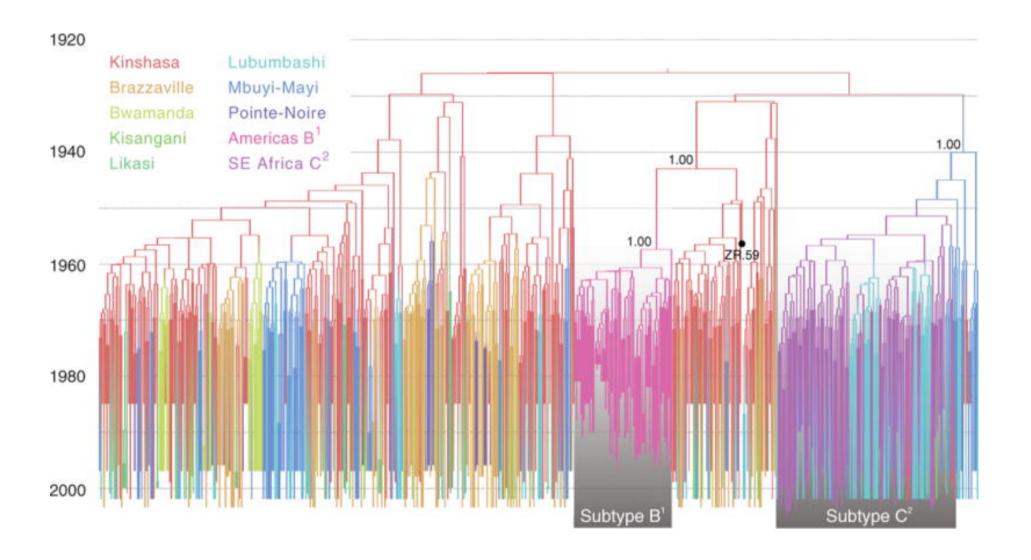


A tree can be dated if you have enough data to infer rates of substitution

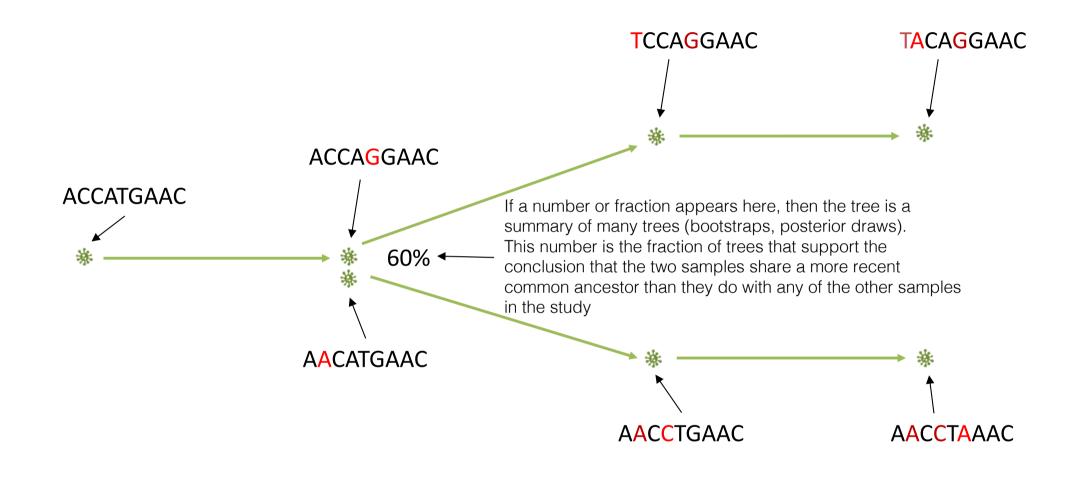


Example software include BEAST, TreeDater, TreeTime, Least Squares Dating (LSD)

This phylogeny has been rotated for clarity. The past is at the top.

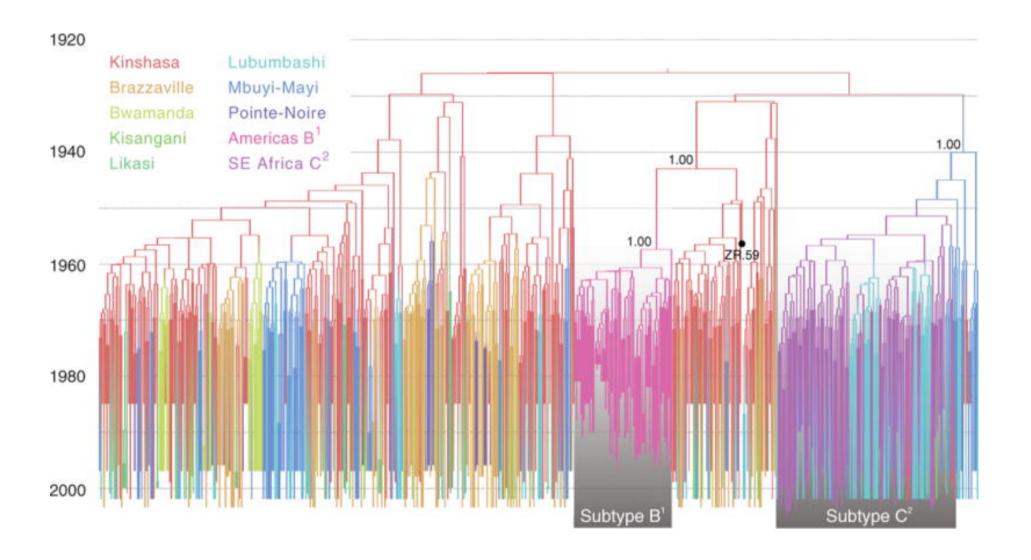


In practice, inferring phylogenies is full of uncertainty

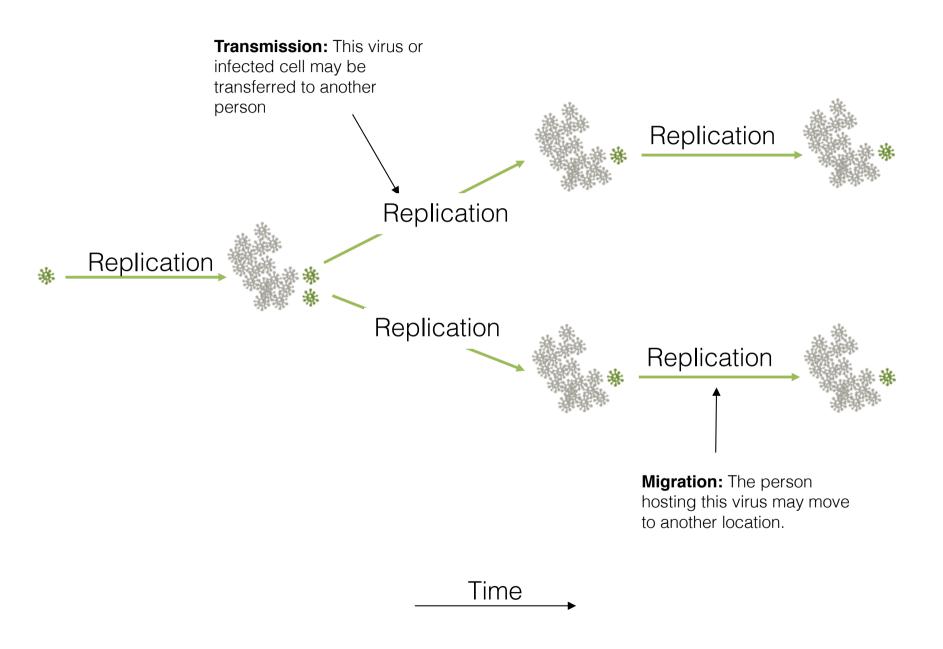


Time

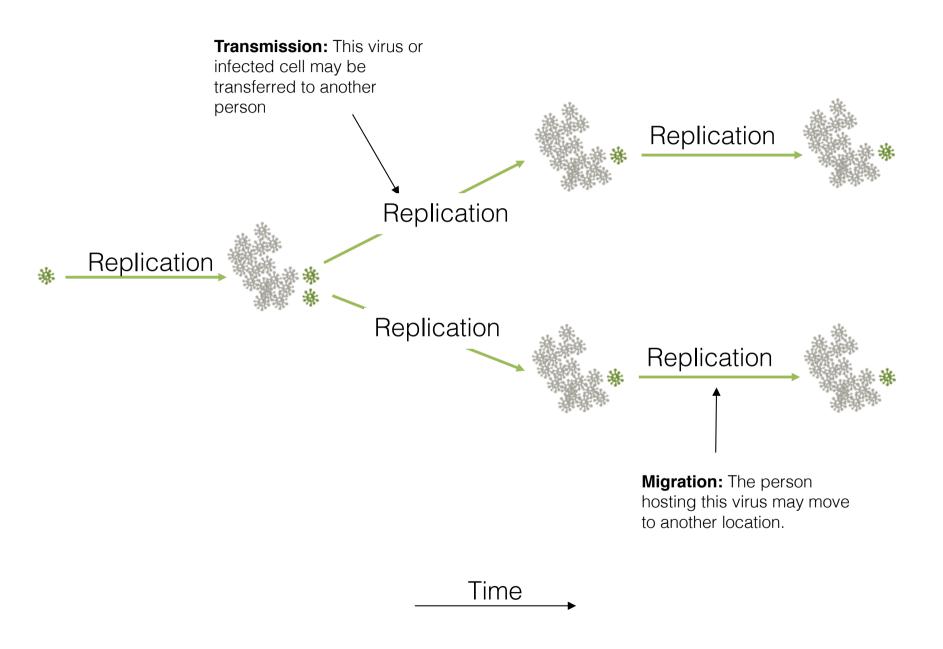
This phylogeny has been rotated for clarity. The past is at the top.

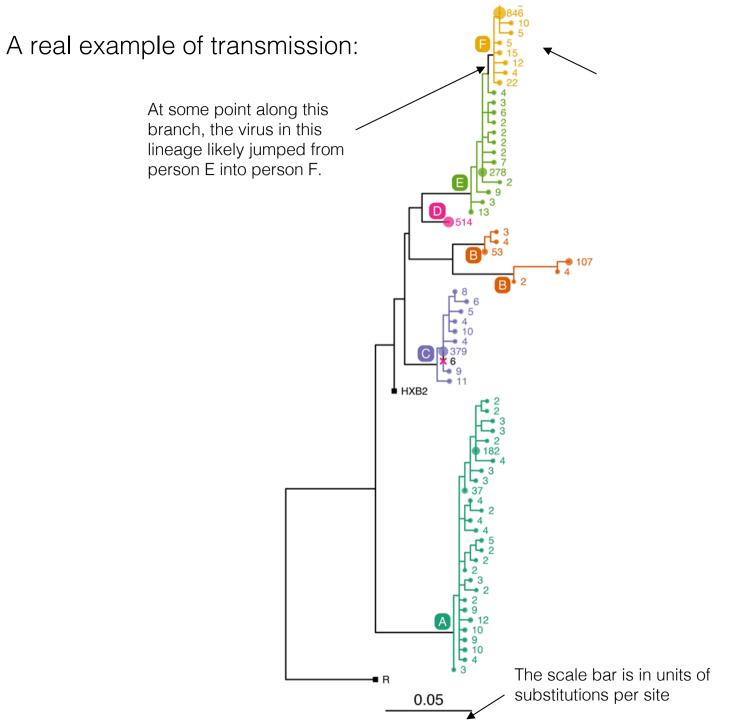


Interesting things may happen to viruses as they replicate:



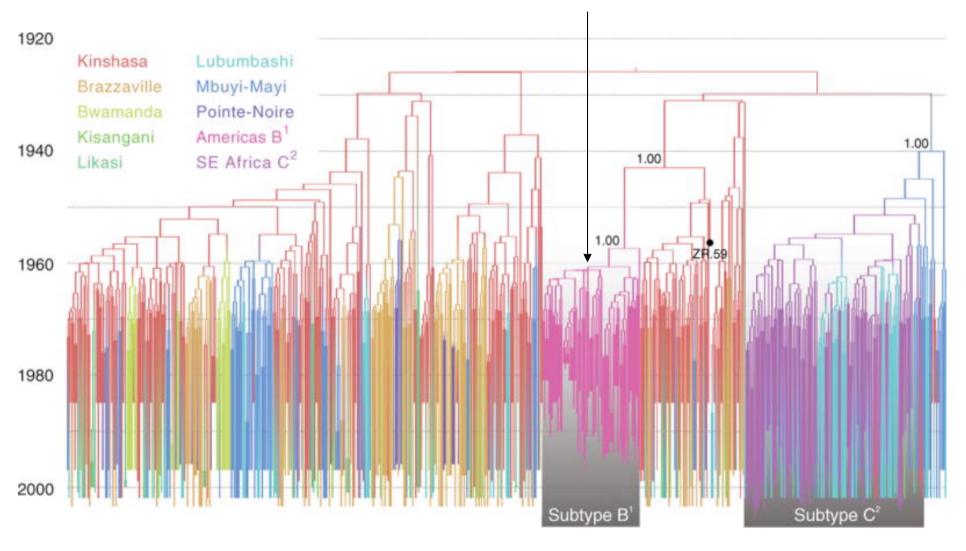
Interesting things may happen to viruses as they replicate:



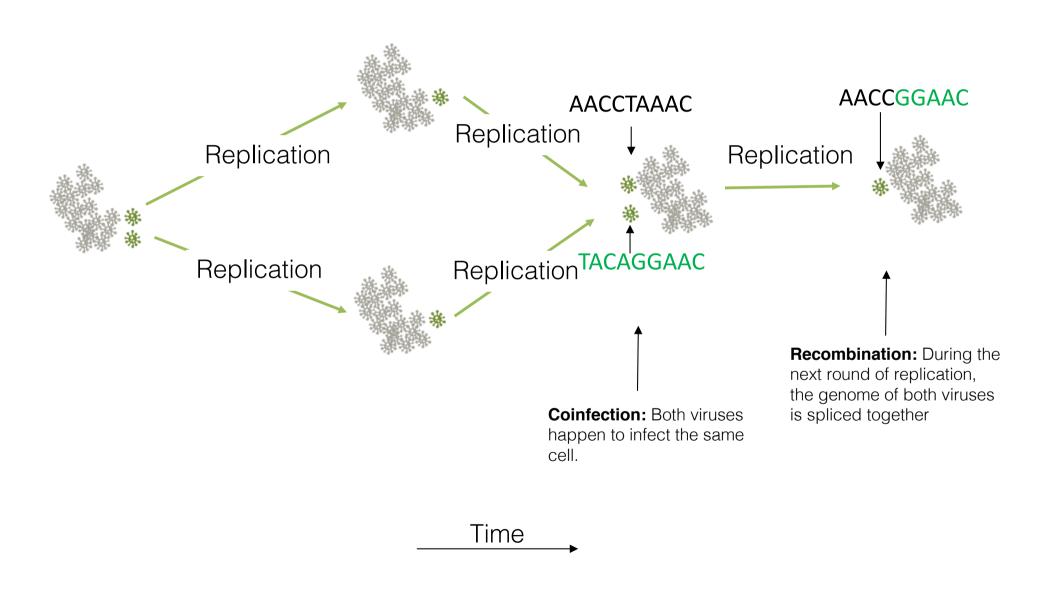




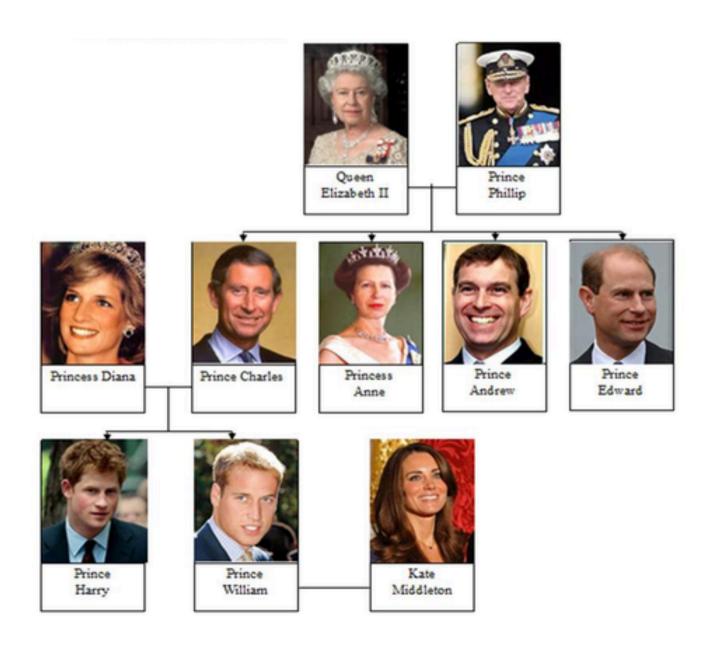
At some point along this branch, the person/people carrying this virus moved from Kinshasa to the Americas.



The final really complicated effect: recombination messes up tree-like structure, creating a 'ancestral graph instead'.



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.

GICTATE GGCATGGGTACCAGCACACAAAGGAATTGGAGGAAATGAACAAGTAGAT	
ATT ACCTECCACCI TAGTAGCAAAAGAATAGTAGCCAGCTGTGATAAATGTCAGAGAAACAGGGCAGGAAACAGCGGAAACAGCAGAAACAGGAAACAGGAAACAGCAG	
AGO LAC <mark>TO GATZ IL</mark> ATA G <mark>YAGCANA ACTTINI</mark> TO LAGCAGAAACAGGGCAGGAAACAGC	
GAATCAA GAGAAT TEGAAT COMA ON TOCCCAAAGTCAAGGAGTAGTAGAA	
AAGGGGGGATTGGGGGGTACAGTGCAGGGGAAAGAATAGTAGACATAATAGCAACAG	
AGGGGGGATTGGGGGGTACAGTGCAGGGGAAAGAATAGTAGACATAATAGCAACAG CTGGAAAGGTGAAGGGGCAGTAGTAATACAAGATAATAGTGACATAAAAGTAGTGCC	
AAGGGGGGATTGGGGGGTACAGTGCAGGGGAAAGAATAGTAGACATAATAGCAACAG CTGGAAAGGTGAAGGGGCAGTAGTAATACAAGATAATAGTGACATAAAAGTAGTGCC ACACCATATGTATGTTTCAGGGAAAGCTAGGGGATGGTTTTATAGACATCACTATGA	
AAGGGGGATTGGGGGTACAGTGCAGGGGAAAGAATAGTAGACATAATAGCAACAG TGGAAAGGTGAAGGGGCAGTAGTAATACAAGATAATAGTGACATAAAAGTAGCTGCC XCACCATATGTATGTTTCAGGGAAAGCTAGGGGATGGTTTTATAGACACACAC	
NAGGGGGATTGGGGGTACAGTGCAGGGAAAGAATAGTAGACATAATAGCAACAG TGGAAAGGTGAAGGGGCAGTAGTAATACAAGATAATAGTGACATAAAAGTAGTGCC ACACCATATGTTTCCAGGGAAAGCTAGGGATGGTTTTATAGACATCACTATGA TTGGGTCAGGAAGTCTCCCATAGAATGGAGGAAAAAAGAGATATATAGACACACAAGAGA	
AAGGGGGATTGGGGGTACAGTGCAGGGAAAGAATAGTAGACATAATAGCAACAG TGGAAAGGTGAAGGGGCAGTAGTAATACAAGATAATAGTGACATAAAAGTAGTGCC ACACCATATGTATGTTTCAGGGAAAGCTAGGGGATGGTTTTATAGACATCACTATGA TTGGGTCAGGGAGTCTCCATAGAATGGAGGAAAAAGAGATATAGCACACAAGTAGAC ATCAAGCAGGACATAACAAGGTAGGATCTCTACAATACTTGGCACTAGCAGCATTAA	
AGGGGGGATTGGGGGTACAGTGCAGGGGAAAGAATAGTAGACATAATAGCAACAG CTGGAAAGGTGAAGGGGCAGTAGTAATACAAGATAATAGTGACATAAAAGTAGTGCC ACACCATATGTATGTTTCAGGGAAAGCTAGGGGATGGTTTTATAGACATCACTATGA CTGGGTCAGGGAGTCTCCATAGAATGGAGGAAAAAGAGATATAGCACACAAGTAGAC ATCAAGCAGGACATAACAAGGTAGGATCTCTACAATACTTGGCACTAGCAGCATTAA	
AAGGGGGATTGGGGGTACAGTGCAGGGAAAGATAGTAGACATAATAGCAACAG TGGAAAGGTGAAGGGGCAGTAGTAATACAAGATAATAGTGACATAAAAGTAGTGCC ACACCATATGTATGTTTCAGGGAAAGCTAGGGGATGGTTTTATAGACATCACTATGA TTGGGTCAGGGAGTCTCCATAGAATGGGGAAAAAGAGATATAGCACACAAGTAGAC ATCAAGCAGGACATAACATA	
ANGGGGGATTGGGGGTACAGTGCAGGGGAAAGATAGTAGACATAATAGCAACAG TGGAAAGGTGAAGGGGCAGTAGTAATACAAGATAATAGTAACATAATAGCAACAG ACACCATATGTATGTTTCAGGGAAAGCTAGGGGATAGTTTTATAGACATCACTATGA TGGGTCAGGGAGTCTCCATAGAATGGAGGAAAAAGAGATATAGCACACAAGTAGAC ATCAAGCAGCAGTAGACAAGGTAGGATCTCTACAATACTTGGCACTAGCAGCATTTAA ATTAAGACATAGACTTTTAGAGGGGCTTAAAACTTCGCACTAGACATTACCAG ATTAAGACATAGACATTTTAGAGGGCTTAAAACAGCGTAAGACATTTCCACAGAAGAAGAAGAACACAA	
ARGEGGGATTEGGGGETACAGTGCAGGGAAAGATAGTAGACATAATAGCAACAG TEGAAAGGTGAAGGGCAGTAGTAATACAAGATAATAGTGACATAAAAGTAGTGCC ACACCATATGTATGTTTCAGGGAAAGCTAGGGGATGGTTTTATAGACATCACTATGA TEGGTCAGGGAGTCTCCATAGAATGGAGGAAAAAGAGATATAGCACACAAGTAGAC ATCAAGCAGGACATAACAAGGTAGGGATCTCTACAATACTTGGCACTTAGACATTTTCCT TAAATGAATAGAATGAGAGCTTAAAATGAGAAGAAGAAGAGAAGAAGAAGAAGAAGAAGA	
AGGGGGATTGGGGGTACAGTGCAGGGAAAGATAGTAGACATAATAGCAACAG CTGGAAAGGTGAAGGGGCAGTAGTAATACAAGATAATAGTAGCATAAAAGTAGTGCC ACACCATATGTATGTTTCAGGGAAAGCTAGGGGGATGGTTTTATAGACATCACTAGACAT ATTGGTCAGGGAGTCTCCATAGAATGGAGGAAAAGAGATATAGCACAACAAGTAGAC ATCAAGCAGGACATAACAAGGTAGGATCTCTACAATACTTGGCACTAGCAGCATTAA CAGCAGCACATAACAAGGAGCTTAAAAGCACTTTTCCT TTT AAATGAATAATAGCAATAGTAGATAGCAGCAGGAAGAAGCGGAGACAA ATTAGTAGCAATAATAATAGCAATAGTTGTGTGTCCATAGTAATCATAGAATATAG	
AGGGGGATTGGGGGTACAGTGCAGGGAAAGATAGTAGACATAATAGCAACAG TGGAAAGGTGAAGGGCAGTAGTAATACAAGATAATAGTAGCATAAAAGTAGTGCC ACACCATATGTATGTTTCAGGGAAAGCTAGGGGATAGTTTTATAGAACATCACTATGA TTGGGTCAGGGAGTCTCCATAGAATGGAGAAAAAGAGATATAGCACACAAGTAGAC ATCAAGCAGGACATAACAAGGTAGGATCTCTACAATACTTGGCACTAGCAGTTTTCCT TTT TAATGAGCTTTTAGAGGAGCTTAACTTGACAGAGGAGAGCCAA TTT TAATGAGATATACTAGAATAGTAGCACTAGGAGGAGAGCCAA AGTAGTAGCAATAATAATAGCAATAGTTGTTGGTCCATAGTAATCATAGAATATAG	
AAGGGGGATTGGGGGTACAGTGCAGGGAAAGATAGTAGACATAATAGCAACAG TGGAAAGGTGAAGGGCAGTAGTAATACAAGATAATAGTGACATAAAAGTAGTGCC ACACCATATGTATGTTCAGGGAAAGCTAGGGGATAGTATTATAGACATCACTATGA TTGGGTCAGGGAGTCTCCATAGAATGGAGGAAAAAGAGATATAGCACACAAGTAGAC ATCAAGCAGGAATAACAAGGTAGGAGCTTAAA TTTT AAA TTGATTAGAGAGGAGCTTAAAACAAGCTGTAGACATTTTCCT AAA TTGATAGATAACAAGATAGTAGAGCAGAGAGCAAAAAAGAGAAGAAGAAGAAGAAGAAGA	
AAGGGGGATTGGGGGTACAGTGCAGGGAAAGAATAGTAGACATAATAGCAACAG TGGAAAGGTGAAGGGCAGTAGTAATACAAGATAATAGTGACATAAAAGTAGTGCC ACACCATATGTATGTTTCAGGGAAAGCTAGGGGATAGTATTATAGCACACACA	
ANGEGEGATTEGEGGETACAGTGCAGGGGAAAGATAGTAGACATAATAGCAACAG CTEGAAAGGTGAAGGGGCAGTAGTAATACAAGATAATAGTAGCATAATAGTGCC CACACATATEGTTTCAGGGAAGCTAGGGGATGTTTTATAGACATCACTATGA TTEGGTCAGGGAGTCTCCATAGAATGGAGGAAAAAGAGATATAGCACACAGTAGAC ATCAAGCAGGACATAACAAGGTAGGAGTCTCTACAATACTTGGCACTAGCAGCATTAA CACACATAACACAGAGCTTTTAAGAGTAGTCCTAGAAGAGAAGAAGAAGAAGAAGAAGAAGAAGAAGAAGAA	
AGGGGGATTGGGGGTACAGTGCAGGGAAAGATAGTAGACATAATAGCAACAG TGGAAAGGTATATGTTTCAGGGAAAGCTAGATAATAGTAGCATAAAAGTAGTGCC ACACCATATGTATGTTCAGGGAAAGCTAGGGGATAGTTTTTATAGAACATCACTATGA TTGGGTCAGGGAGTCTCCATAGAATGGAGGAAAAAGAGATATAGCACACAAGTAGAC ATCAAGCAGGACATAACAAGAATAGTACATTAA TTTT TAAATGAGATTTAAAGTAGTGTTGTGTTAGACATTTCCT AGTAGTAGAGCTTTTAGAGAATAGTTGTTGTTGTAGACATATATAGAATATAG ATGGGGGTGGAGATGGGGCACCATGCTCTTGGGATGTTGATGATCTTAGAATATAG ATTTTTTTTTT	
ANGEGEGATTEGEGETACAGTECAGEGAAAGATAGTAGACATAATAGCAACAG TEGAAAGETGAAGGGCAGTAGTAATACAAGATAATAGTAGCATAAAAGTAGTGCC ACACCATATETTAGTTTCAGGGAAAGCTAGGGGATAGTATATAGCACACAAGATAGAC TEGAGCAGGAGTCTCCATAGAATGGAGGAAAAAGAGATATAGCACACAAGTAGAC TEGAGCAGGAACTAACAAGGTAGGAGCTTTAACATTTGCAGAGCATTTACA TETATAGACATTAACAAGGTAGGAGCTTTAACTTGCCAGAGAGAG	
AAGGGGGATTGGAGGGTACAGTGCAGGGGAAAGAATAGTAGACATAATAGCAACAG CTGGAAAGGTGAAGGGGCAGTAGTAATACAAGATAATAGTGACATAAAAGTAGCTGCC ACACCATATGTATGTTTCAGGGAAAGCTAGGGGATGGTTTTATAGACATCACTATGA TTGGGTCAGGGAGTCTCCATAGAATGGAGGAAAAAGAGATATAGCACAACAAGTAGAC ATCAAGCAGGACATAACAAGGTAGGATCTCTACAATACTTGGCACTAGCAGCATTAC TTTT TAA TTGATATTAGAGGAGCTTAAAGCGTTAGACAGTAGACAATTATCCT ATTTTATAATAATAATAGCAATAGTTGTGTGTCCATAGTAATCATAGAATATAGAATATAGAATATAAGCACACCCACC	
ANGGGGGATTGGGGGTACAGTGCAGGGAAAGATAGTAGACATAATAGCAACAG TGGAAAGGTGAAGGGGCAGTAGTAATACAAGATAATAGTGACATAAAAGTAGTGCC ACACCATATGTATGTTTCAGGGAAAGCTAGGGGATAGTATTTAAACACACAC	
ANGEGEGEATTEGEGEGETACAGTECAGEGEAAAGATAGTAGACATAATAGCAACAG TEGAAAGGTEAAGGGCAGTAGTAATACAAGATAATAGTGACATAAAAGTAGTGCC ACACCATATETATGTTTCAGGGAAAGCTAGGGGATAGTTTTATAGACATCACTATGA TTEGGTCAGGGAGTCTCCATAGAATGGAGGAAAAAGAGATATAGCACACAAGAGTAGAC ATCAAGCAGGACATAACAAGAGTAGACTTTACAATACTTGGCACTAGCAGCATTATACTTCAT TTATATAGAGCTTTTAGAGGAGCTTTAAAGTGCAGGAGAGAACAATAATAGCAGATAGAT	
LAGGGGGATTGGGGGTTACAGTGCAGGGAAAGAATAGTAGACATAATAGCAACAG CTGGAAAGTTGAAGGGGCAGTAGTAATACAAGATAATAGTGACATAAAAGTAGTGCC CACACATATGTATGTTTCAGGGAAAGCTAGGGGTTGGTTTTATAGACATCACTATGA CTGGGTCAGGGAGTCTCCATAGAATGGAGGAAAAAGAGATATAGCACACAAGTAGAC ATCAAGCAGGACATAACAAGGTAGGATCTCTACAATACTTGGCACTAGCAGCATTAA CAGCAGGACATAACAAGGTAGGATCTCTACAATACTTGGCACTAGCAGCATTAA CAGCAGCACATAACAAGGTAGGAGCTTAACAGCCCAGCAGAGAGAG	
AGGGGGATTGGAGGGTACAGTGCAGGGAAAGATAGTAGCATAATAGCAACAG CTGGAAAGGTATGTATGATAGTAGTAGTAGTAGCACACACA	
AGGGGGATTGGGGGTACAGTGCAGGGAAAGATAGTAGACATAATAGCAACAG TGGAAAGGTCAAGGGGCAGTAGTAATACAAGATAATAGTGACATAAAAGTAGTGCC ACACCATATGTATGTTTCAGGGAAAGCTAGGGGATAGTTTTATAGACATCACTATGA TTGGGTCAGGGACTCTCCATAGAATGGAGGAAAAAGAGATATTAGCACACAAGAGTAGAC ATCAAGCAGGACATAACAAGAGGTAGGAGTCTCTACAATACTTGGCACTAGACATTAA TTTT	
ANGGGGGATTGGGGGTACAGTGCAGGGAAAGATAGTAGCATAATAGCAACAG TGGAAAGGTTATGTTTCAGGGAAAGCTAGGGGTTGTTTTATAGACATCACTATGA TTGGGTCAGGGATCTCCATAGAATGGAGGAAAAAGAGATATAGCACACACA	