NIH Fogarty International Center
Advancing Science for Global Health

COV-IRT
COVID-19 International Research Team

Phylogenetics 101

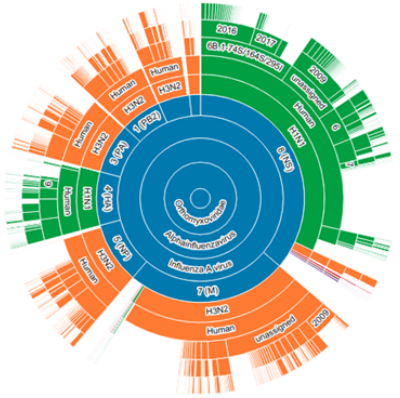
Nídia Trovão, PhD
Division of International Epidemiology and Population Studies
Fogarty International Center
National Institutes of Health

General data processing workflow

Obtain sample

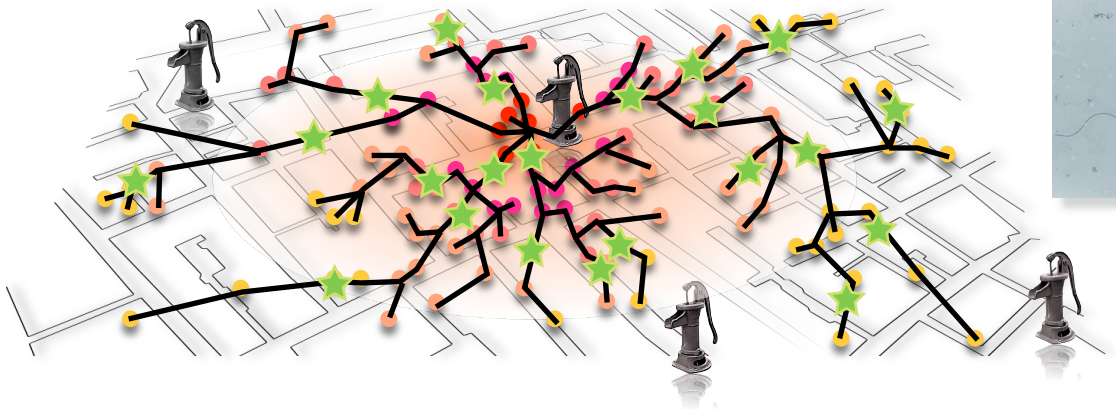
Prepare DNA/RNA for sequencing

Sequence DNA

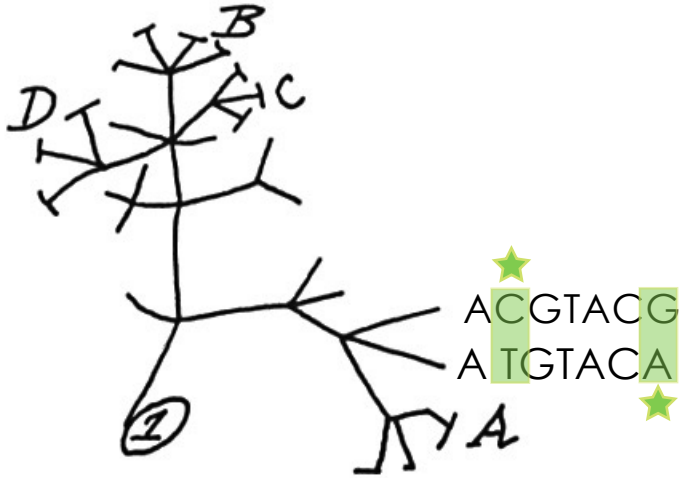


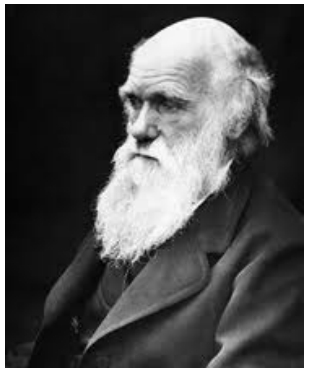
Initial sample processing

Molecular Epidemiology ~ Phylodynamics



John Snow
(1813-1858)



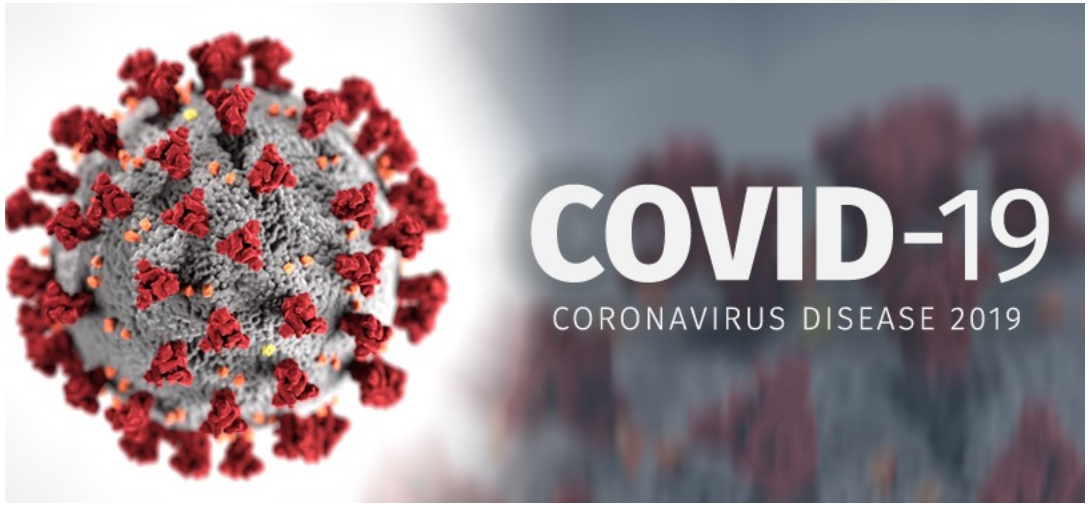


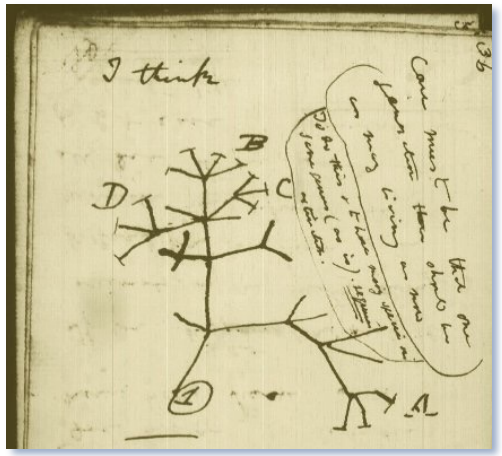
Charles Darwin
(1809-1882)

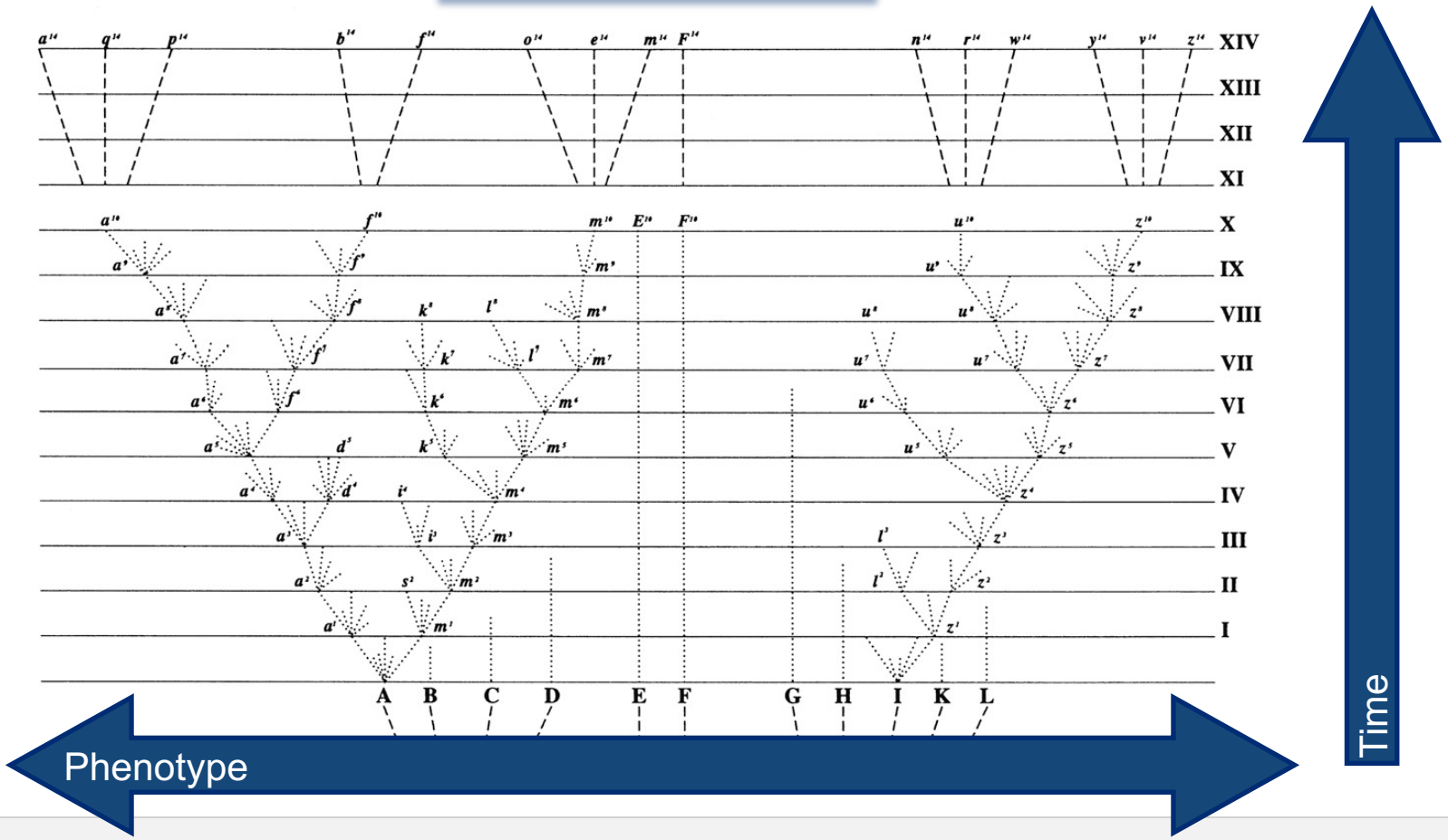

MERS-CoV


Zika virus

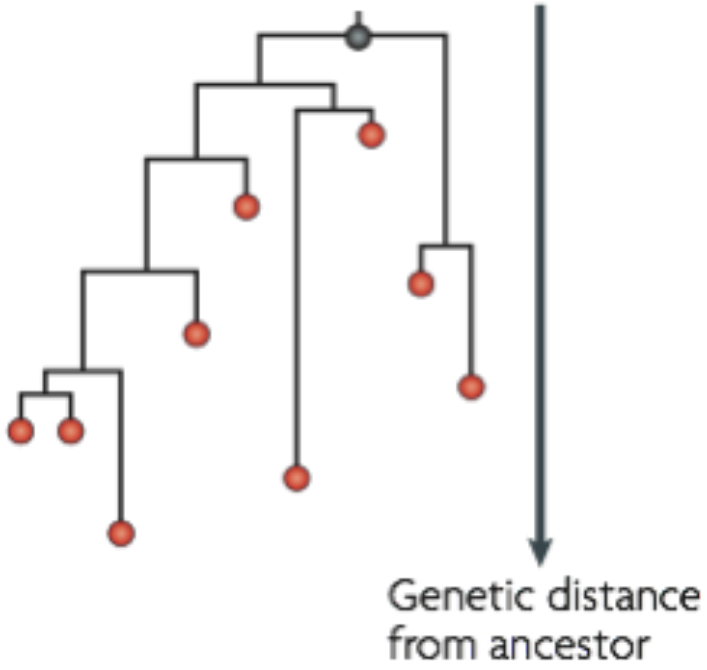

EBOLA
OUTBREAK


COVID-19
CORONAVIRUS DISEASE 2019


I think
B
D
C
A

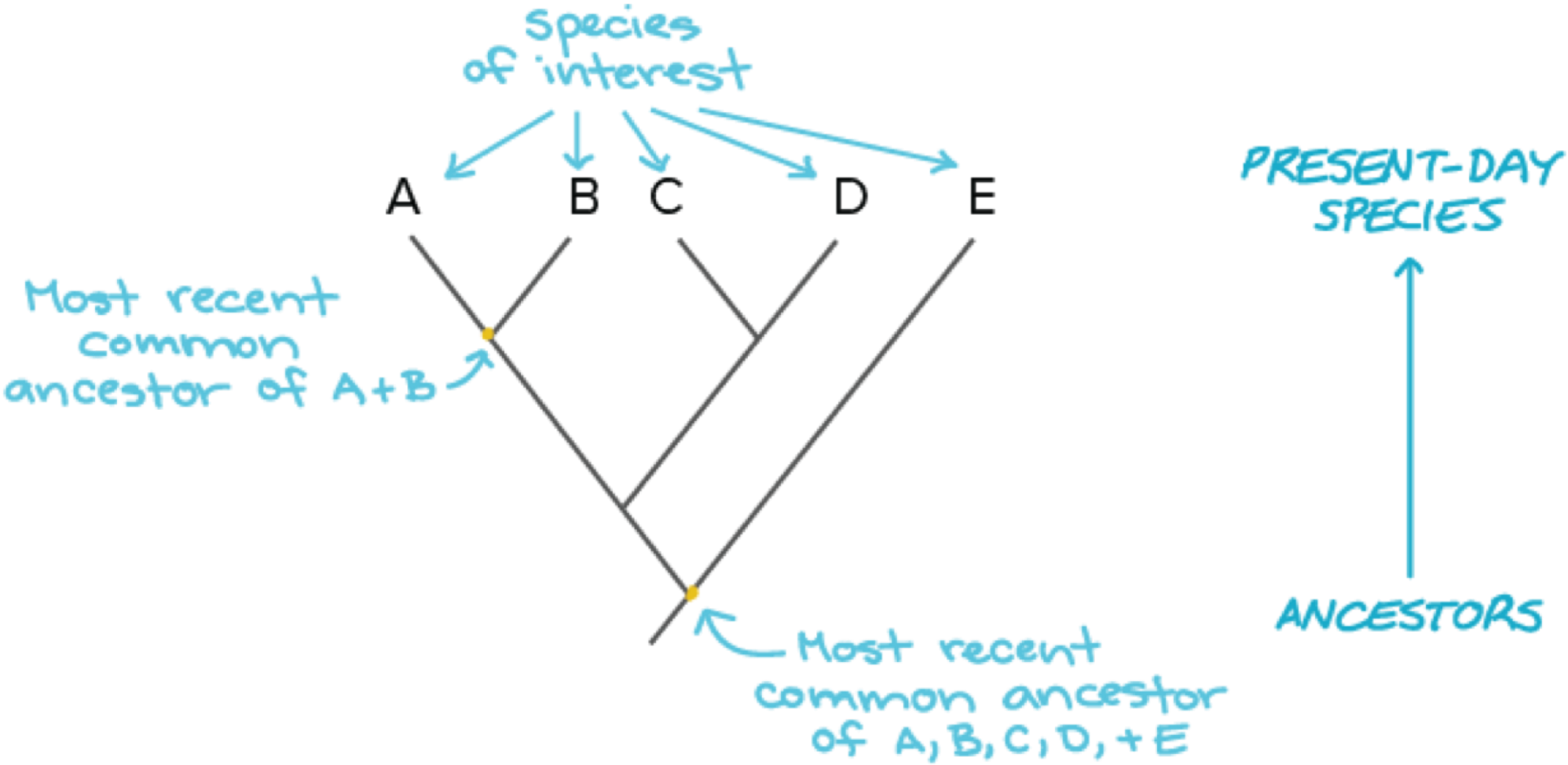

XIV
XIII
XII
XI
X
IX
VIII
VII
VI
V
IV
III
II
I
A
B
C
D
E
F
G
H
I
K
L
Phenotype
Time

Model for Molecular Epidemiology



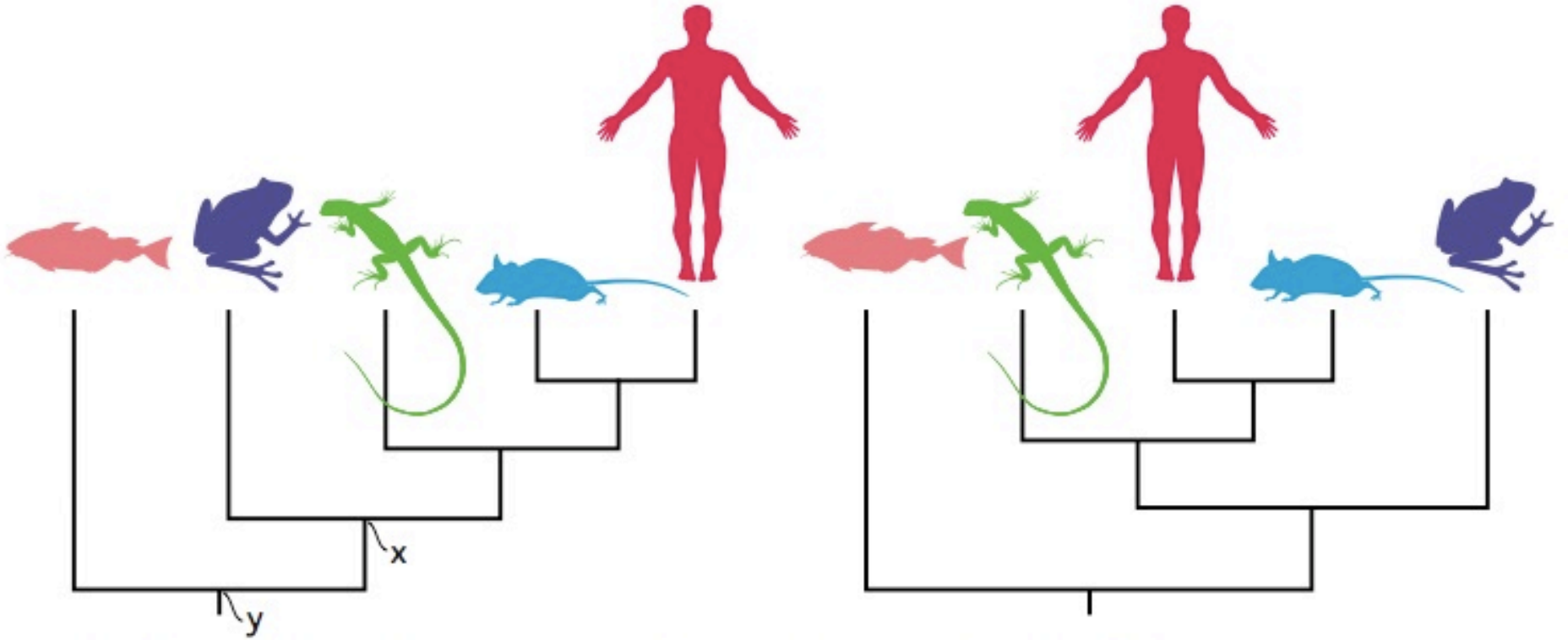
Pybus and Rambaut, NRG, 2009

Model for Molecular Epidemiology



The Tree-Thinking Challenge

David A. Baum, Stacey DeWitt Smith, Samuel S. S. Donovan



Which phylogenetic tree is accurate? On the basis of the tree on the left, is the frog more closely related to the fish or the human? Does the tree on the right change your mind? See the text for how the common ancestors (x and y) indicate relatedness.

Fundamental Phylodynamic Questions

- How genetically diverse is a pathogen population?
- How do pathogen genomes change through time?
- How does pathogen genetic diversity vary through space?
- What processes and/or events determine these changes?
- What are the effects of pathogen genetic diversity on virulence, transmissibility, resistance to treatment, etc.

Specific questions

- Where did it come from?
- How fast is it transmitting?
- In what direction is it spreading?
- Are hosts X, Y & Z epidemiologically linked?
- Are strains associated with particular transmission routes?
- What adaptations has it accrued?

Topics in Phylogenetics Reconstruction

- Estimating genetic distances between sequences
- Inferring phylogenetic trees
- Detecting recombination events

Mutation and Substitution

- Mutation rate (m)

  -The rate at which mutational errors are incorporated into the genome during replication.

  -Depends largely on replication mechanism.

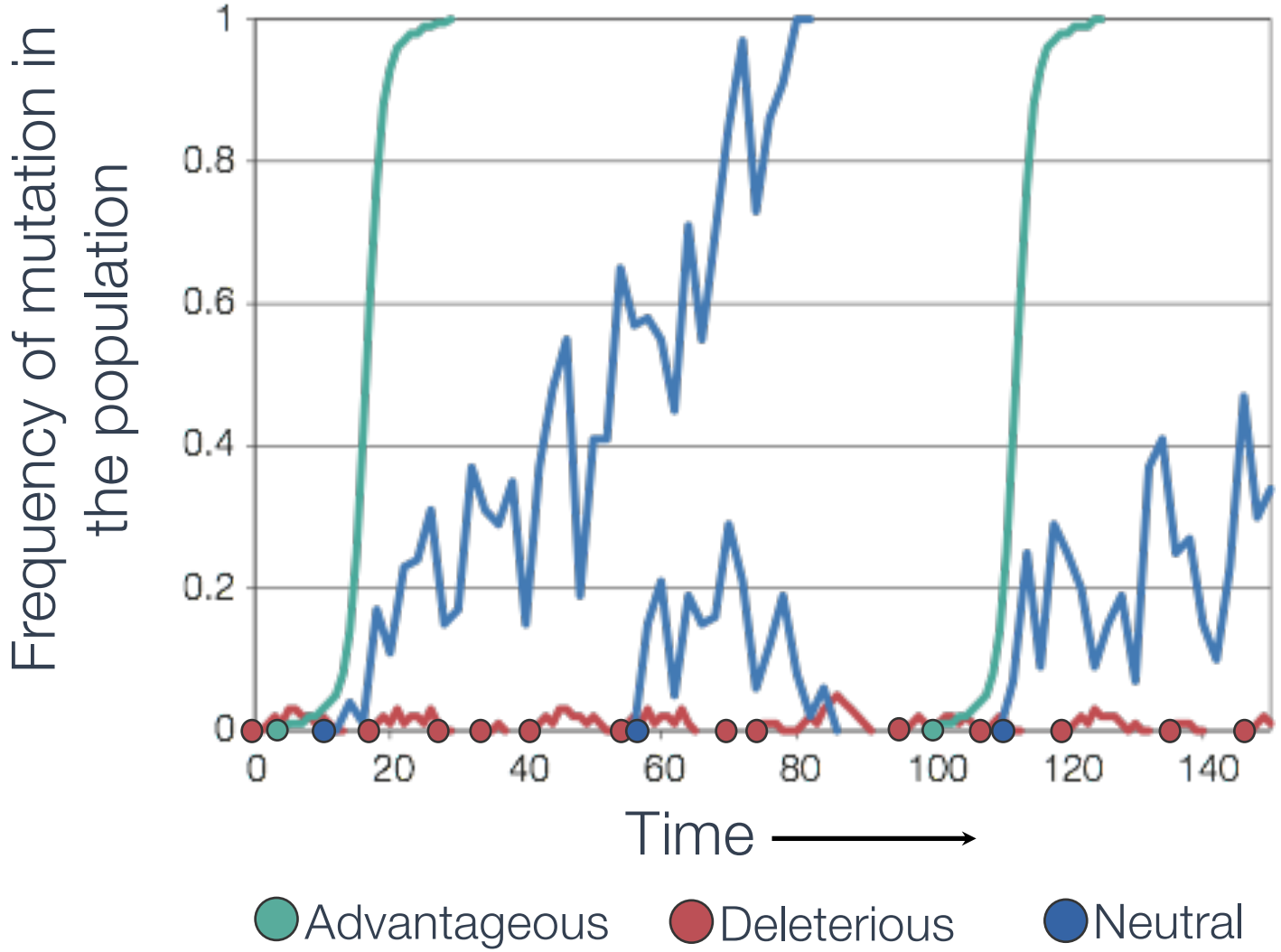
  -Can be expressed as mutations per nucleotide site per replication event.

  -Can be measured *in vitro* and *in vivo* using molecular biology techniques.

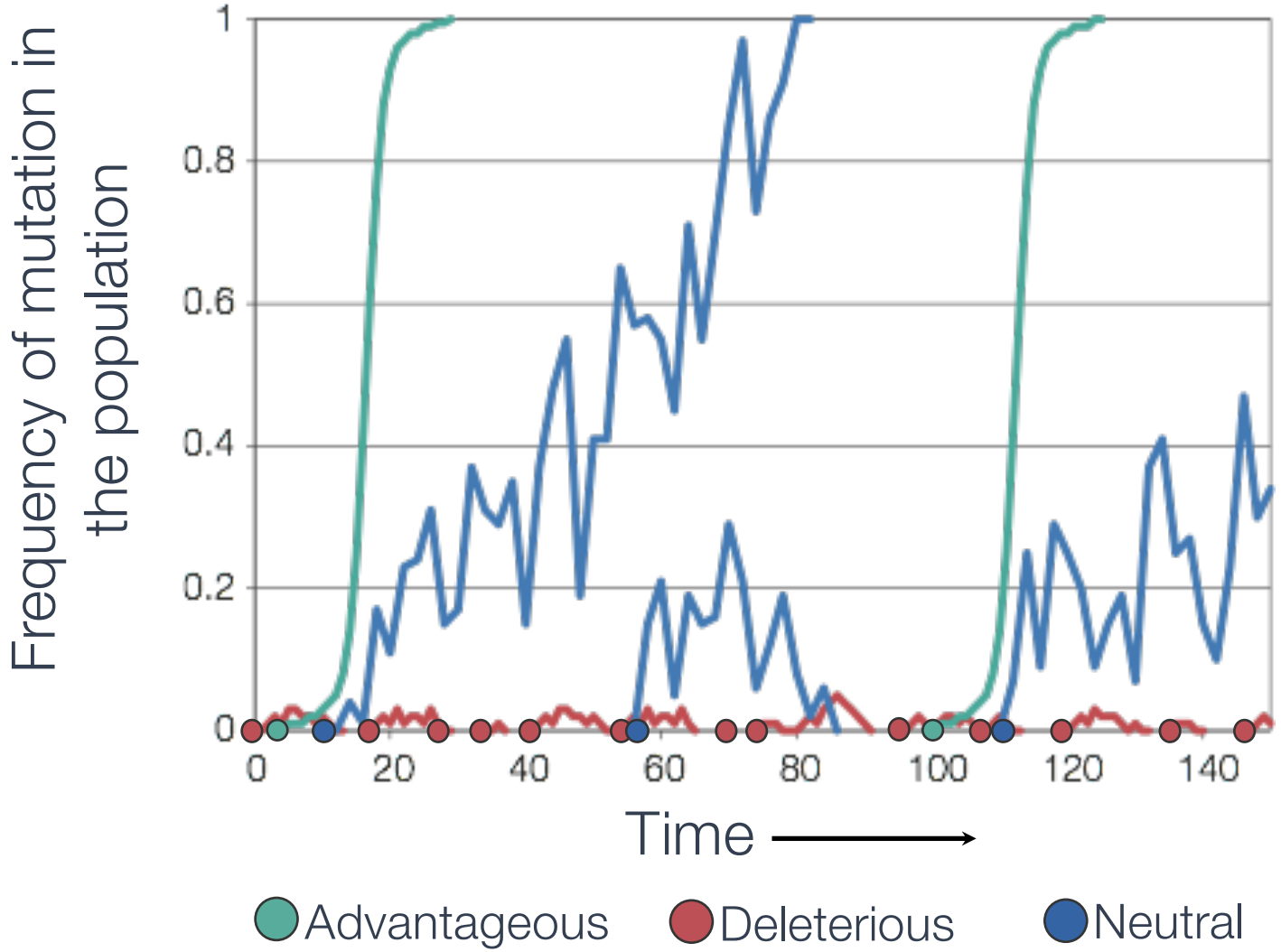
Mutation and Substitution

- Substitution rate (μ)
  - -The rate at which mutations spread and become fixed in the population.
  - -Depends on many factors such as population dynamics, natural selection, genetic drift, generation time.
  - -Measured in number of substitutions per nucleotide site per unit time (days, years, generations).
  - -Can be estimated from virus sequences sampled over time.

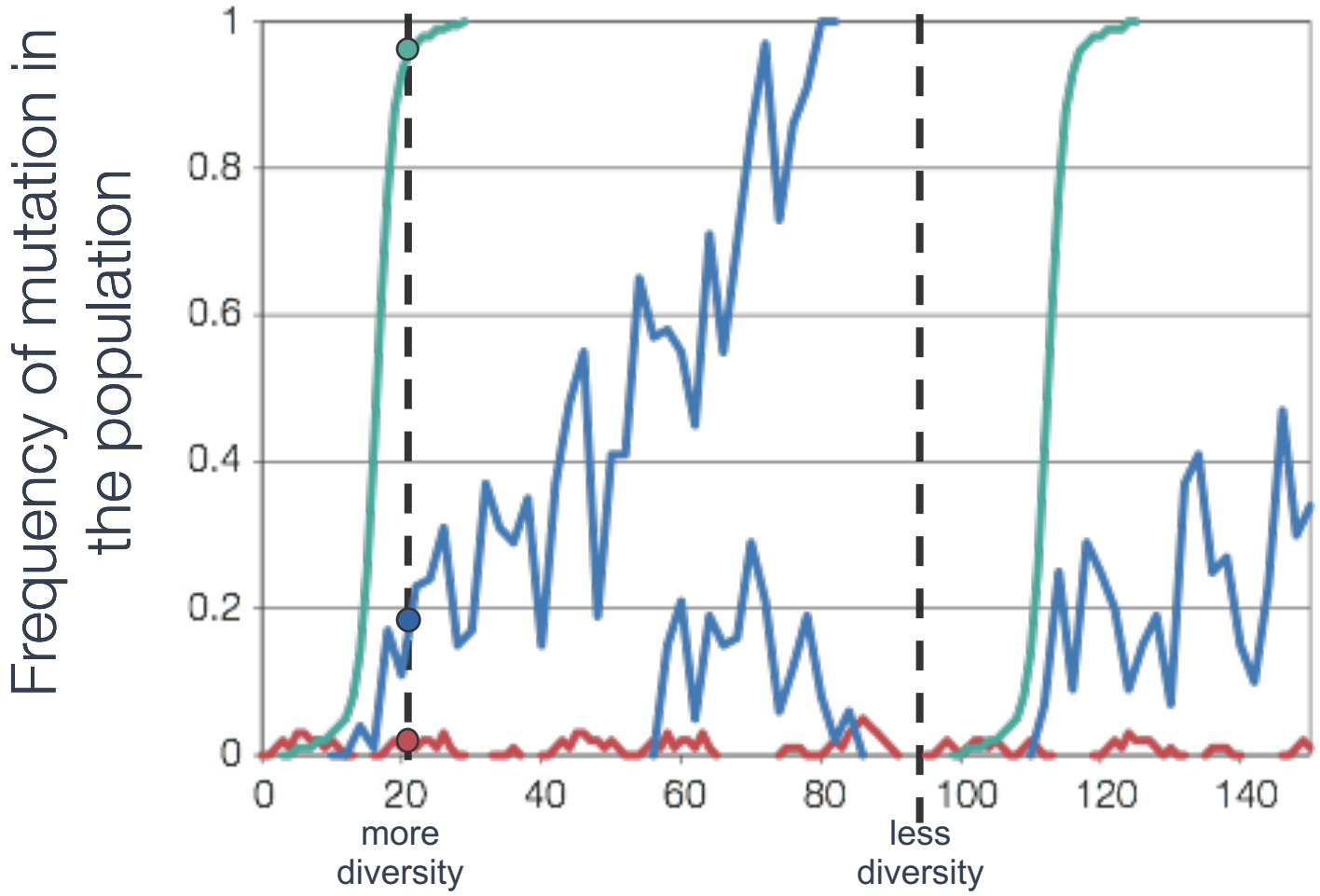
Mutation rate



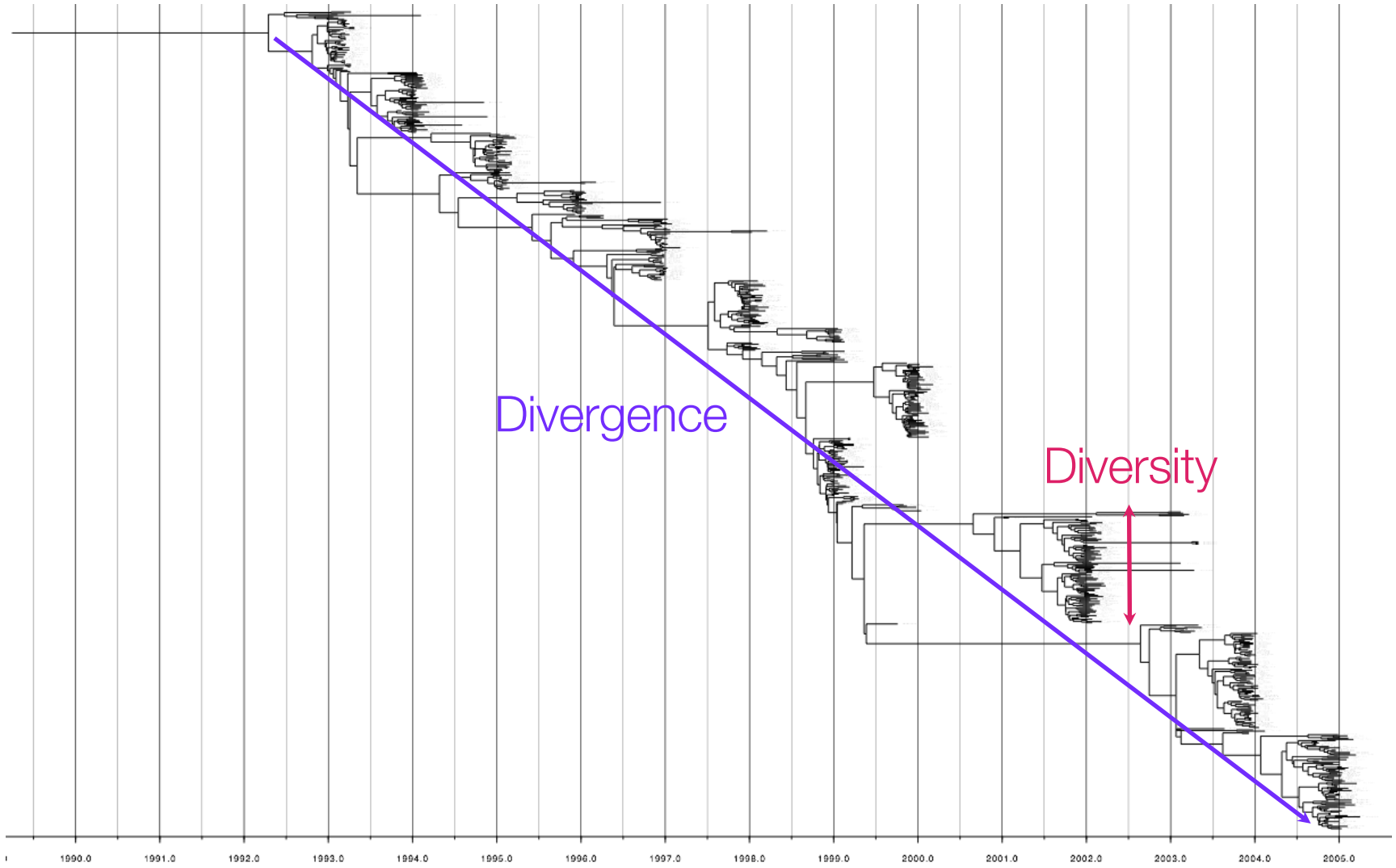
The accumulation of substitutions through time is typically called divergence



Diversity depends on a very complex interplay of selection (positive and negative), genetic drift and gene flow.



Example: H3N2 influenza (HA gene) sampled weekly over 10 years in New York, USA

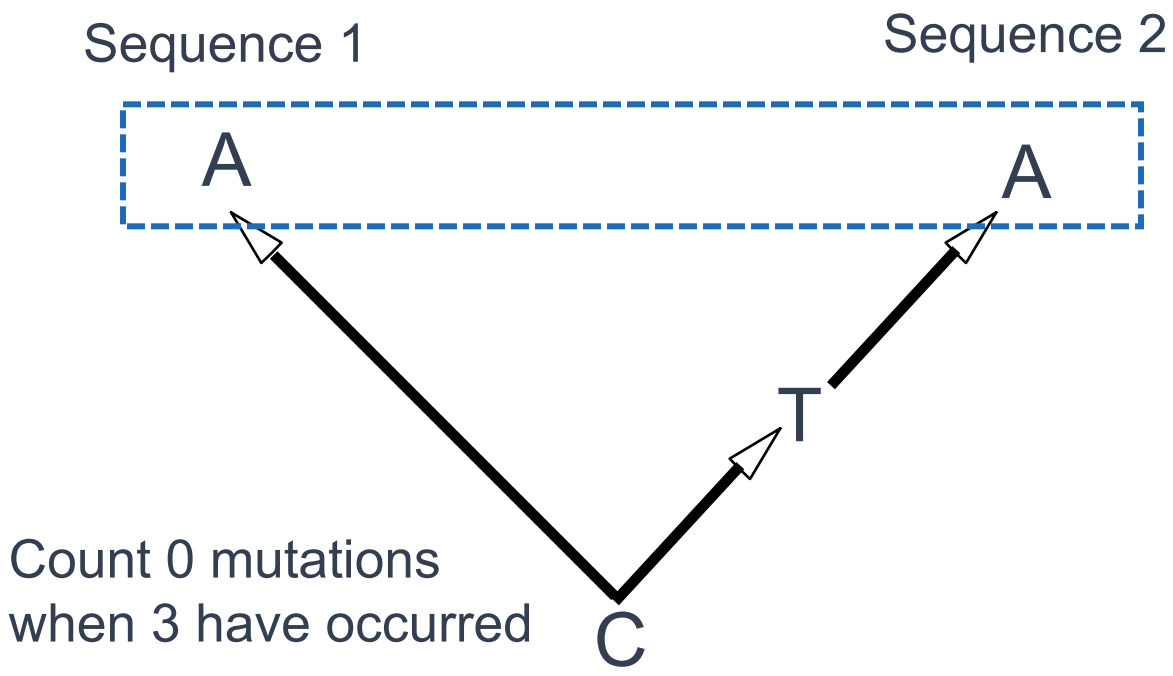


Estimating Genetic Distances Between Sequences

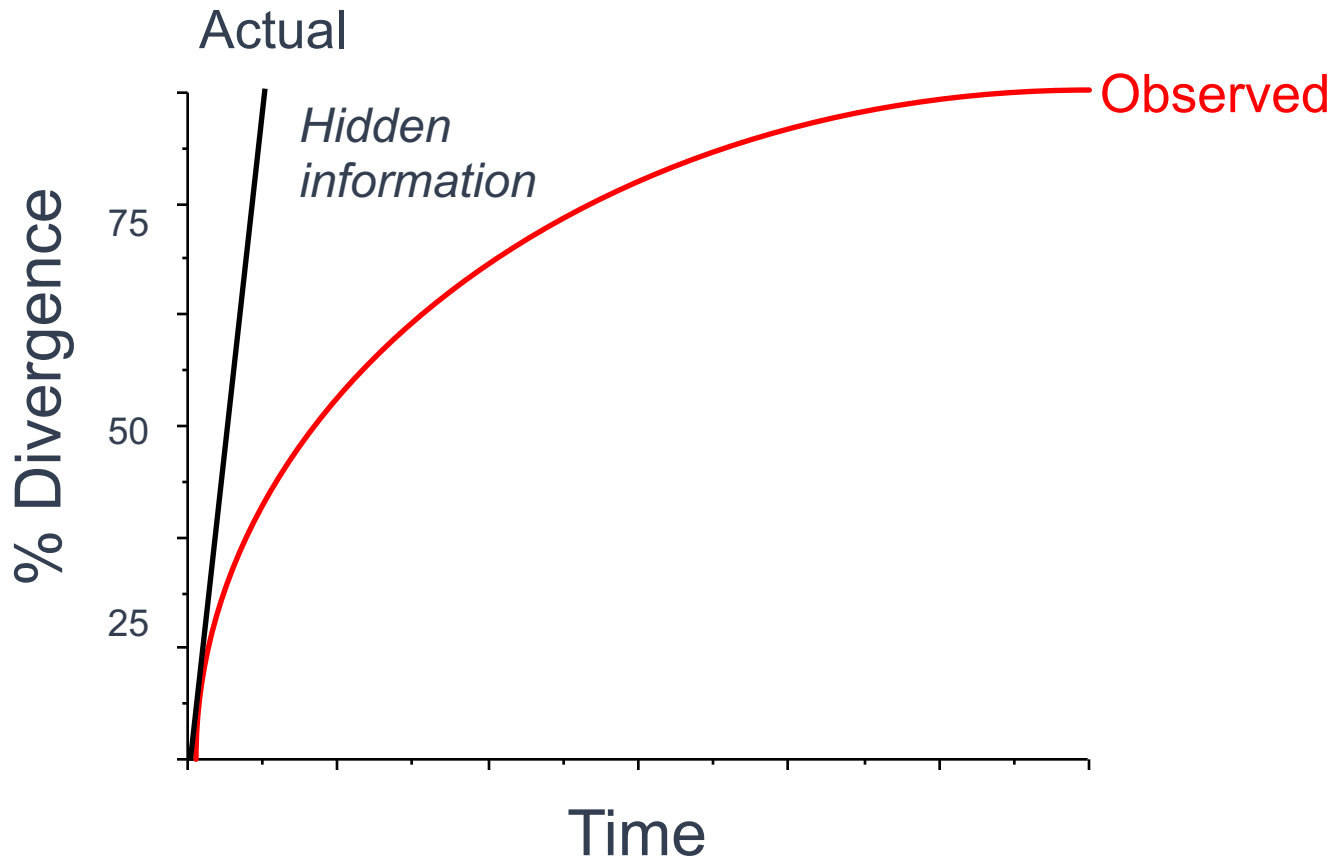
Estimating Genetic Distance

```
SIVcpz    ATGGGTGCGA GAGCGTCAGT TCTAACAGGG GGAAAATTAG ATCGCTGGGA
HIV-1     ATGGGTGCGA GAGCGTCAGT ATTAAGCGGG GGAGAATTAG ATCGATGGGA

SIVcpz    AAAAGTTCGG CTTAGGCCCG GGGGAAGAAA AAGATATATG ATGAAACATT
HIV-1     AAAAATTCGG TTAAGGCCAG GGGGAAAGAA AAAATATAAA TTAAAACATA

SIVcpz    TAGTATGGGC AAGCAGGGAG CTGGAAAGAT TCGCATGTGA CCCCGGGCTA
HIV-1     TAGTATGGGC AAGCAGGGAG CTAGAACGAT TCGCAGTTAA TCCTGGCCTG

SIVcpz    ATGGAAAGTA AGGAAGGATG TACTAAATTG TTACAACAAT TAGAGCCAGC
HIV-1     TTAGAAACAT CAGAAGGCTG TAGACAAATA CTGGGACAGC TACAACCATC

SIVcpz    TCTCAAAACA GGCTCAGAAG GACTGCGGTC CTTGTTTAAC ACTCTGGCAG
HIV-1     CCTTCAGACA GGATCAGAAG AACTTAGATC ATTATATAAT ACAGTAGCAA

SIVcpz    TACTGTGGTG CATACATAGT GACATCACTG TAGAAGACAC ACAGAAAGCT
HIV-1     CCCTCTATTG TGTGCATCAA AGGATAGAGA TAAAAGACAC CAAGGAAGCT

SIVcpz    CTAGAACAGC TAAAGCGGCA TCATGGAGAA CAACAGAGCA AAACTGAAAG
HIV-1     TTAGACAAGA TAGAG--GAA -----GAGCA AAACAAAAGT AA---GAAAA

SIVcpz    TAACTCAGGA AGCCGTGAAG GGGGAGCCAG TCAAGGCGCT AGTGCCTCTG
HIV-1     AAGCACAGCA AGC-----AG CAGCTGACA- -CAGGACAC- AG--CAGC--

SIVcpz    CTGGCATTAG TGGAAATTAC
HIV-1     CAGG--TCAG CCAAAATTAC
```

Multiple Substitutions at a Single Site - Hidden Information



The Problem of Multiple Substitution



- When % divergence is low, observed distance (*p*) is a good estimator of genetic distance (*d*)
- When % divergence is high, *p* underestimates *d* and a “correction statistic” is required i.e. a model of DNA substitution

Models of DNA Substitution

• Models of DNA sequence evolution are required to recover the missing information through correcting for multiple substitutions.

i. The probability of substitution between bases (e.g. A to C, C to T…)

ii. The probability of substitution along a sequence (different sites/regions evolve at different rates)

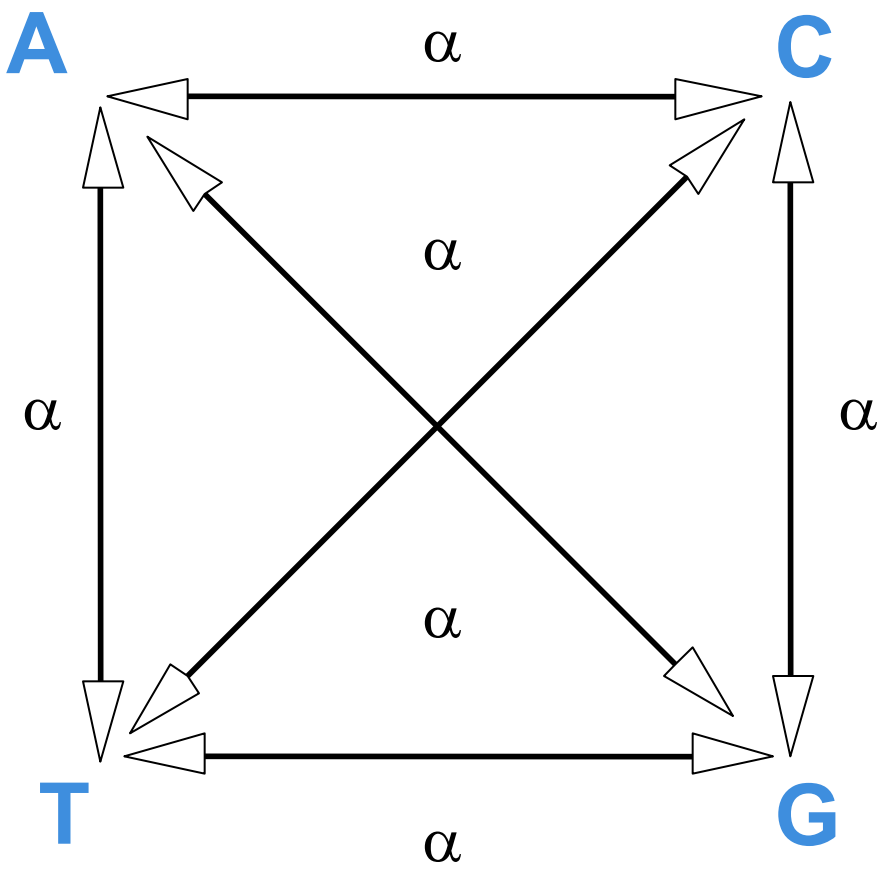
Models of DNA Substitution 1 - Jukes-Cantor, 1969

- Assumptions:
  - i. All bases evolve independently
  - ii. All bases are at equal frequency
  - iii. Each base can change with equal probability (α)
  - iv. Mutations arise according to a Poisson distribution (rare and independent events)

- From this the number of substitutions per site (d) can be estimated by;

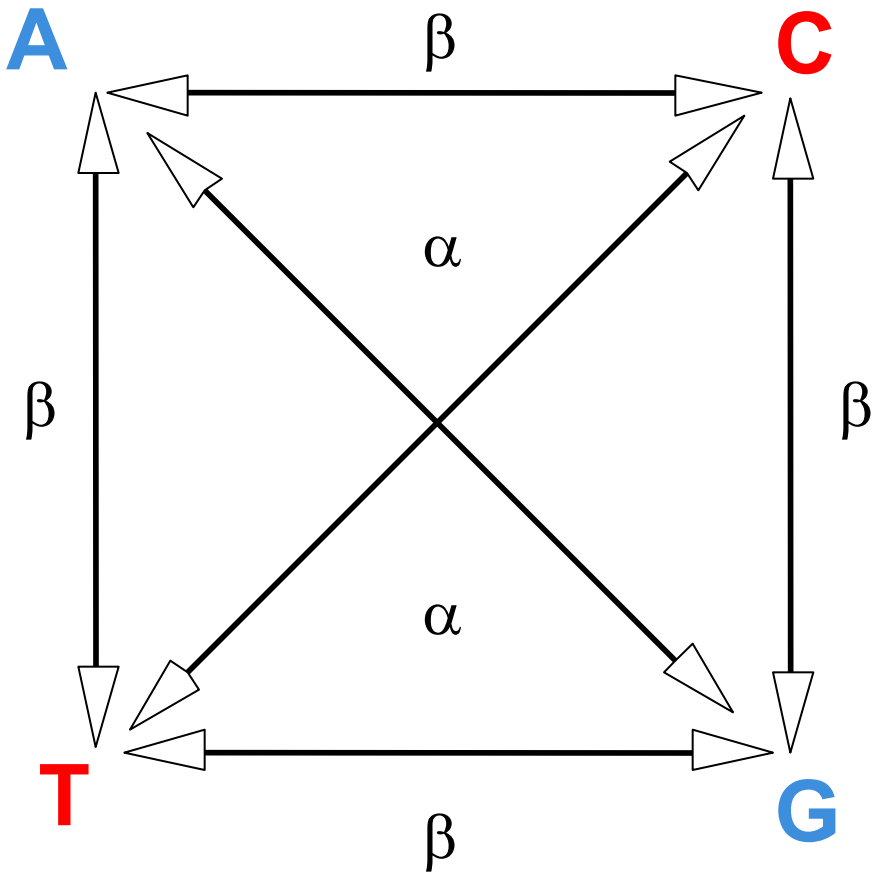
$$d = -3/4 \ln (1-4/3P)$$

where P is the proportion of observed nucleotide differences between 2 sequences.



All substitutions occur at the same rate (α)

Is this model too simple for real data?



Transitions (α) and transversions (β) occur at a different rate

Models of DNA Substitution 2 - Kimura 2-parameter, 1980

- Assumptions:
  - i. All bases evolve independently
  - ii. All bases are at equal frequency
  - iii. Transitions and transversions occur with different probabilities (α and β)
  - iv. The Jukes-Cantor model is applied to transitions and transversions independently

- From this the expected number of substitutions per site (d) can be estimated by;

$$d = -1/2 \ln (1-2P-Q)\sqrt{1-2Q}$$

where P is the proportion of observed transitions and Q the proportion of observed transversions between 2 sequences

Models of DNA Substitution

Simplest (few parameters)

1. Base frequencies are equal and all substitutions are equally likely (Jukes-Cantor)

2. Base frequencies are equal but transitions and transversions occur at different rates (Kimura 2-parameter)

3. Unequal base frequencies and transitions and transversions occur at different rates (Hasegawa-Kishino-Yano)

4. Unequal base frequencies and all substitution types occur at different rates (General Reversible Model)

Most complex (many parameters)

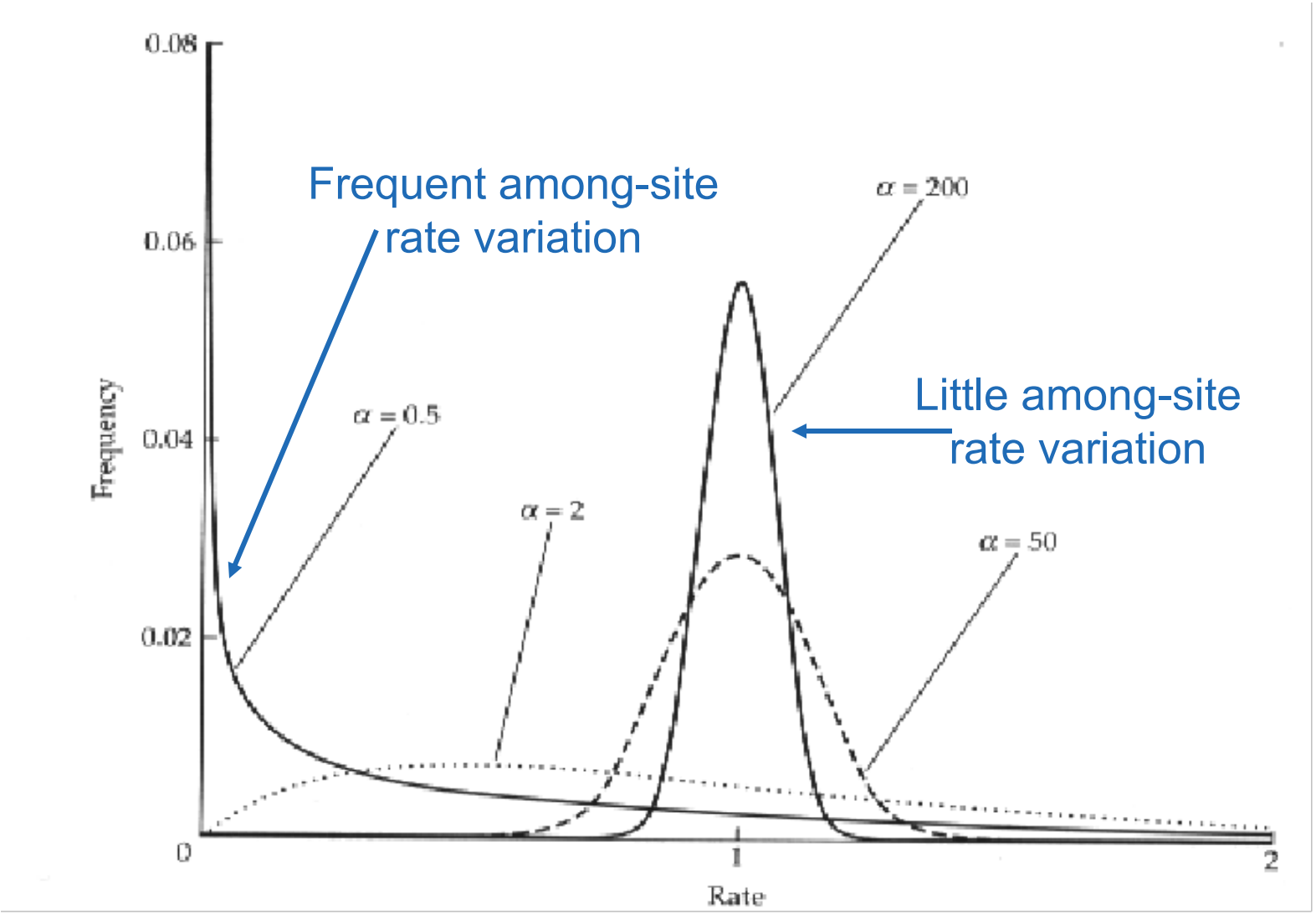
All these models can be tested using the program jMODELTEST (darwin.uvigo.es/software/jmodeltest.html)

Models of DNA Substitution

i. The probability of substitution between bases (e.g. A to C, C to T…)

ii. The probability of substitution along a sequence (different sites/regions evolve at different rates)

Gamma Distribution Helps Model Among-Site Rate Heterogeneity

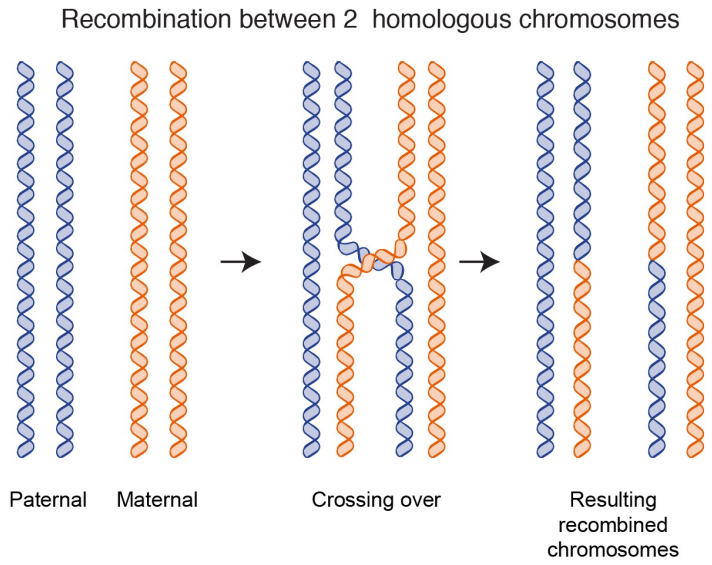


- Viruses are usually characterized by extensive among-site rate variation ($\alpha <<< 1$).

Estimating Genetic Distance: SIVcpz vs HIVlai

- Uncorrected (*p*-distance) = 0.406
- Jukes-Cantor = 0.586
- Kimura 2-parameter = 0.602
- Hasegawa-Kishino-Yano = 0.611
- General reversible = 0.620
- **General reversible + gamma = 1.017**

Detecting Recombination



Recombination & Reassortment

- Reassortment of Segmented Viruses: entire genome segments swapped during co-infection (influenza, rotavirus)

- Recombination:
  (a) breakage and rejoining of DNA/RNA molecules
  (b) template switching by RNA polymerases during replication (homologous and non-homologous)

- Major driving force in evolution of genetic diversity in viruses

Recombination & Reassortment

- The Problems:
  - Generates new genetic configurations
  - Complicates our attempts to infer phylogenetic history and other evolutionary processes (e.g. positive selection)

- The Solutions:
  - Find recombinants and remove them from the data set (usual plan)
  - Incorporate recombinants into an explicit evolutionary model (far harder)

Methods for Recombination Detection

- Measure level of linkage disequilibrium:
  - *LDhat, D'*

- Look for changes in patterns of sequence similarity (often pairwise):
  - *GENECOV, RDP, Max Chi-Square, SimPlot, SiScan, TOPAL*

- Look for incongruent phylogenetic trees:
  - *BOOTSCAN, 3SEQ, LARD, PLATO, LIKEWIND*

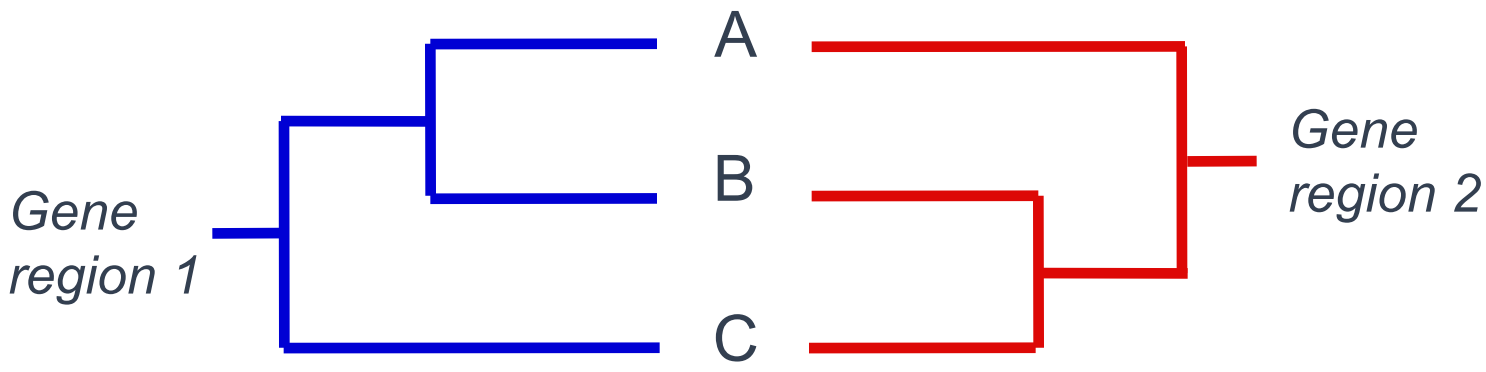
- Look for "networked" evolution
  - *SplitsTree, NeighborNet*

- Look for excessive convergent evolution:
  - *Homoplasy test, PIST*

- See http://www.bioinf.manchester.ac.uk/recombination/programs.shtml for a more complete list
- Many of these methods are available in the Recombination Detection Program (RDP3) – http://darwin.uvigo.es/rdp/rdp.html

Detecting Recombination: Looking for Incongruent Trees

- Different genes produce different trees



- "Topological incongruence", where different gene regions (or genes) produce different phylogenetic trees, is the strongest signal for recombination

Major influenza epidemics associated with reassortant viruses

- 1947 post-WWII total vaccine failure

Reassortant virus
'Old' segments: PB1, NA, M
'New' segments: PB2, PA, HA, NP, NS

- 1951 severe epidemic

Reassortant virus
'Old' segments: PB2, HA
'New' segments: PB1, PA, NP, NA, M, NS

Trouble-shooting

The 7 most common errors in phylogenetic analysis

1. Not using enough background data in the analysis

- Start BIG!
  - Better to begin by using too many background sequences from GenBank than too little: you can always remove sequences later in your analysis if they are not relevant
- Not including enough background data can make interpretation of your data difficult
  - Ex. What appears to be a single viral lineage is really multiple introductions

2. Not fully aligning the sequence data

- Programs like MUSCLE & CLUSTAL are very effective for difficult alignments, but you also need to manually go over the alignment afterwards and check it
- Sometimes a lot of sequencing errors occur at the ends of a sequence and these short sections should be removed

3. Not removing recombinants

- Recombination is frequent for many RNA viruses (eg, HIV, HBV), and not removing recombinants can make noise in your phylogenetic signal

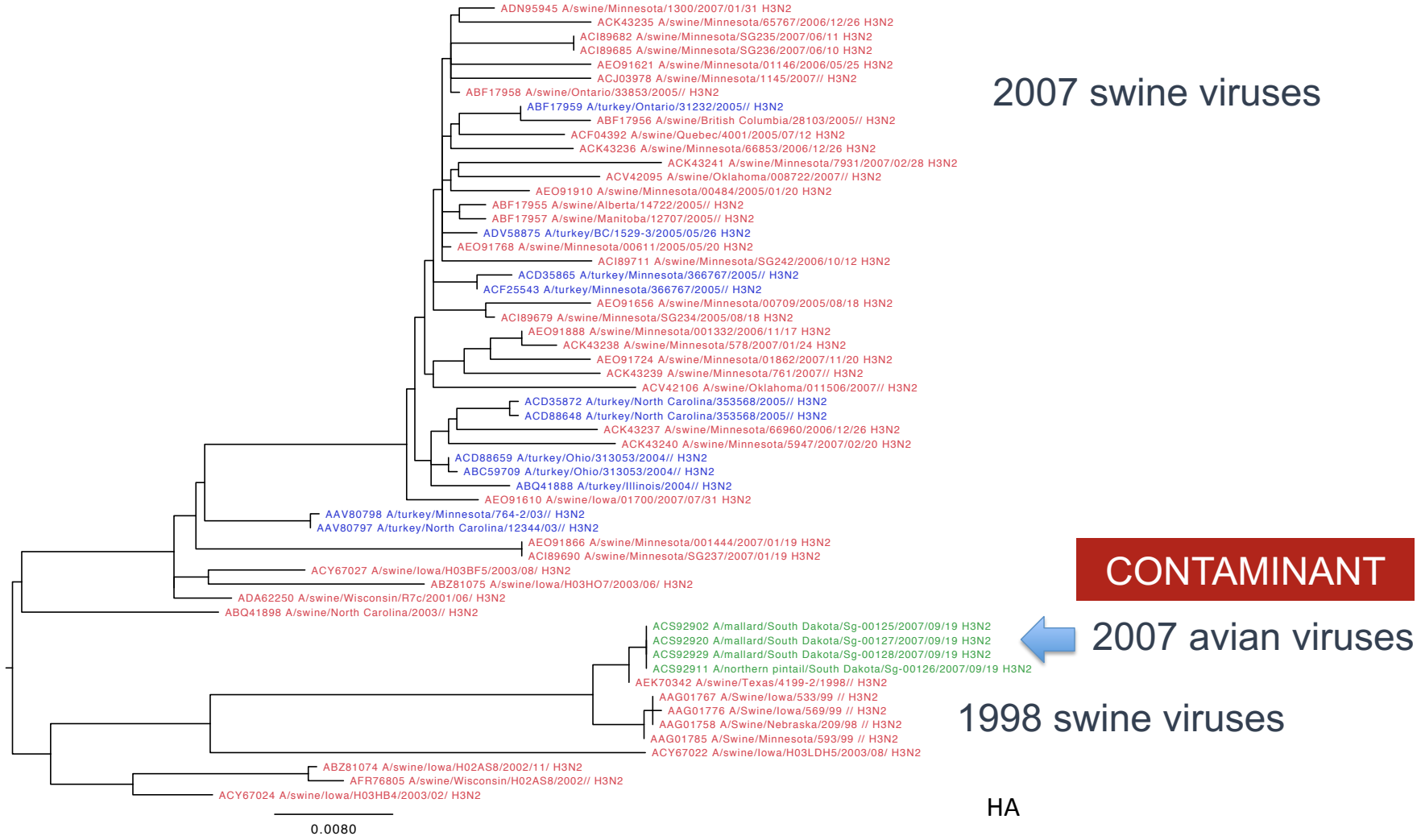
4. Not using multiple methods to build trees

- Important to confirm that multiple approaches to tree-building arrive at the same tree
- Different ML approaches (RAxML, iQTree, FastTree)
- Bayesian approaches (MrBayes, BEAST)

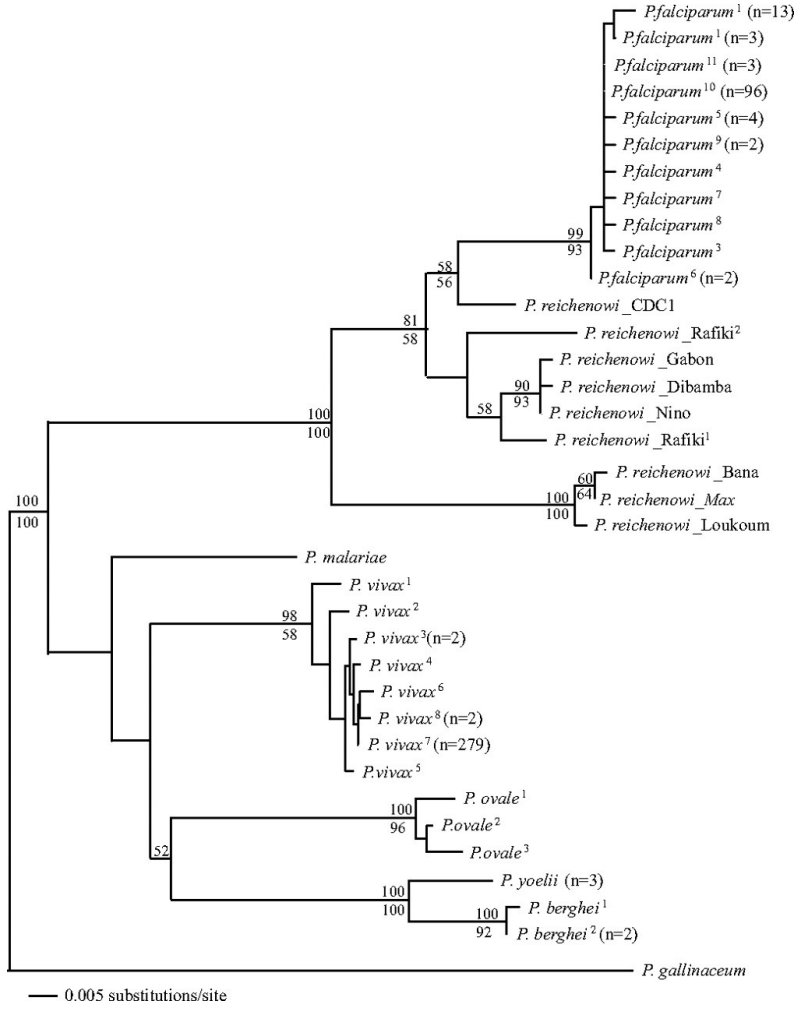
5. Trying to use advanced phylodynamic models before studying data using simpler methods

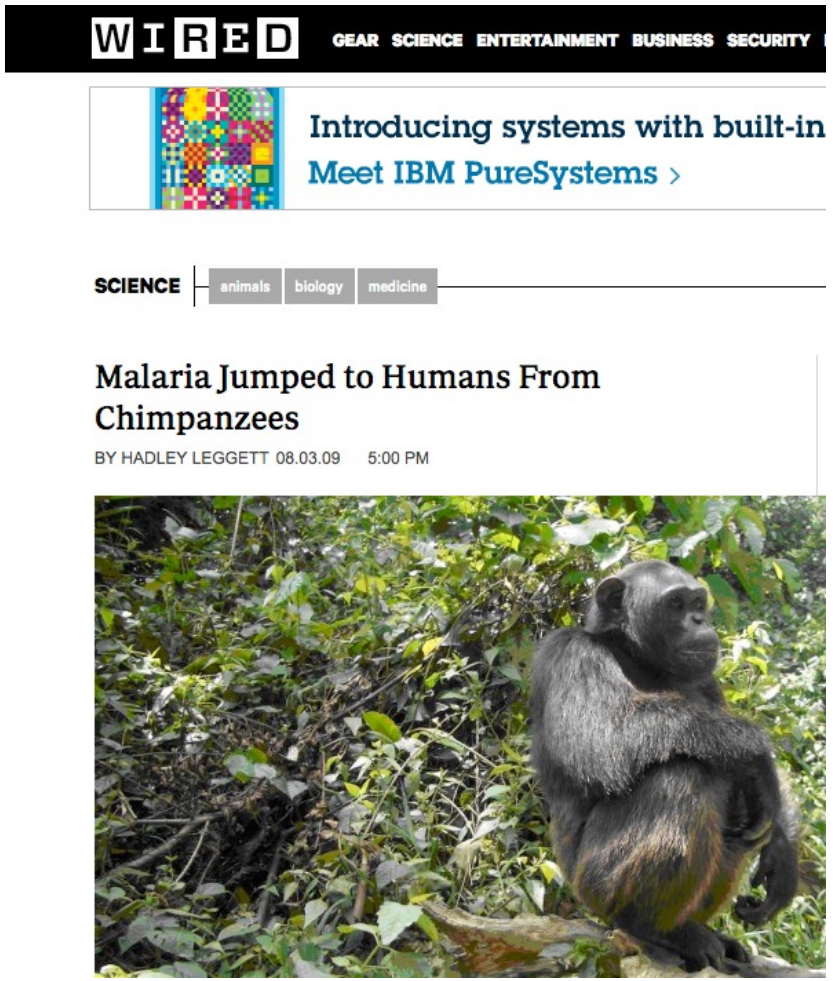
- You really want to thoroughly study your data using NJ + ML methods before doing a more complicated analysis in BEAST

6. Not identifying contaminants/sequencing errors



7. Not accounting for bias and gaps in data when interpreting the phylogeny





Rich S M et al. PNAS 2009;106:14902-14907

Summary

- Use adequate background data
- Make sure your alignment is correct
- Be mindful of gaps in the data (don’t over-interpret) and possibility of sequencing error

Useful Textbooks & Software

Books:

- Page RDM & Holmes EC. (1998). *Molecular Evolution: A Phylogenetic Approach*. Blackwell Science Ltd, Oxford.
- Lemey P, Salemi M & Vandamme A-M. (2009). *The Phylogenetic Handbook, 2nd Edition*. Cambridge University Press.
- Holmes EC. (2009). *The Evolution and Emergence of RNA Viruses*. Oxford University Press, Oxford.

Computer Software:

- BEAST (Bayesian Evolutionary Analysis Sampling Trees)
  - http://beast.bio.ed.ac.uk/
- MEGA (Molecular Evolutionary Genetics Analysis)
  - http://megasoftware.net/
- MrBayes (Bayesian inference of phylogeny)
  - http://mrbayes.csit.fsu.edu/
- PhyML (Maximum likelihood phylogenetics)
  - http://www.atgc-montpellier.fr/phyml/
- HyPhy/DATAMONKEY (Selection, recombination & hypothesis testing)
  - http://datamonkey.org/
- RDP3 (Recombination detection program)
  - darwin.uvigo.es/rdp/rdp.html
- PAUP* (Phylogenetic Analysis Using Parsimony *and other methods)
  - http://paup.csit.fsu.edu/

COVID-19 International Research Team

Fogarty International Center

Advancing Science for Global Health