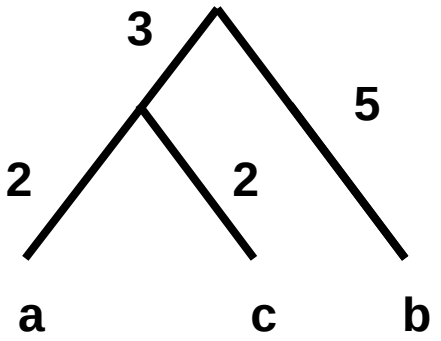
Exercises

1) Which methods assume a molecular clock: UPGMA,
   Which methods evaluate multiple trees: parsimony, maximum likelihood

2) Calculate branch lengths and tree (UPGMA) given the distance matrix:

	a	b	c
a	0		
b	13	0	
c	4	7	0



d((a,c), b) = (13 + 7)/2 = 10
1/2 d((a,c), b) = 5
5 - 2 = 3

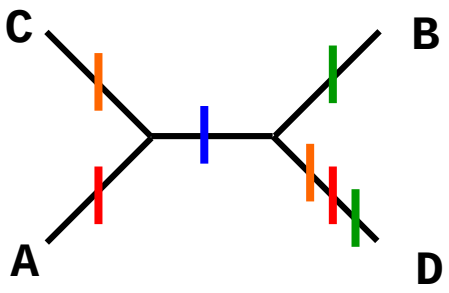
Exercises

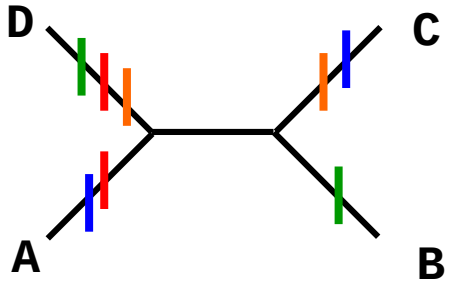
4) What causes long branch attraction in parsimony tree reconstruction? Homoplasy

5) What is the most parsimonious unrooted tree? What is the minimum number of mutations?

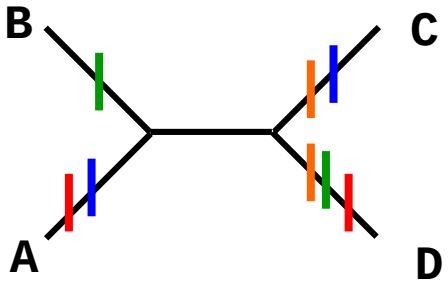
Sequence	Site 1	2	3	4
A	T	A	A	A
B	G	G	G	A
C	G	A	A	G
D	A	G	T	T

7





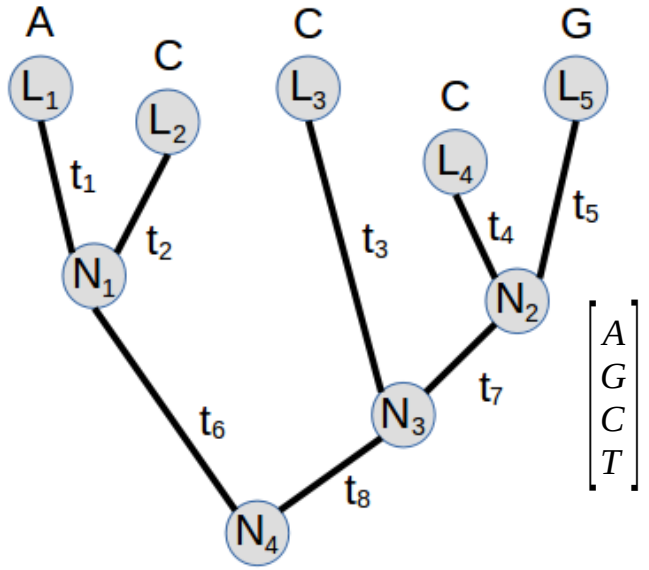
8



8

Exercises

4) What is the probability of $N_3 = A$, written in the form of $P_{ij}(t)$?



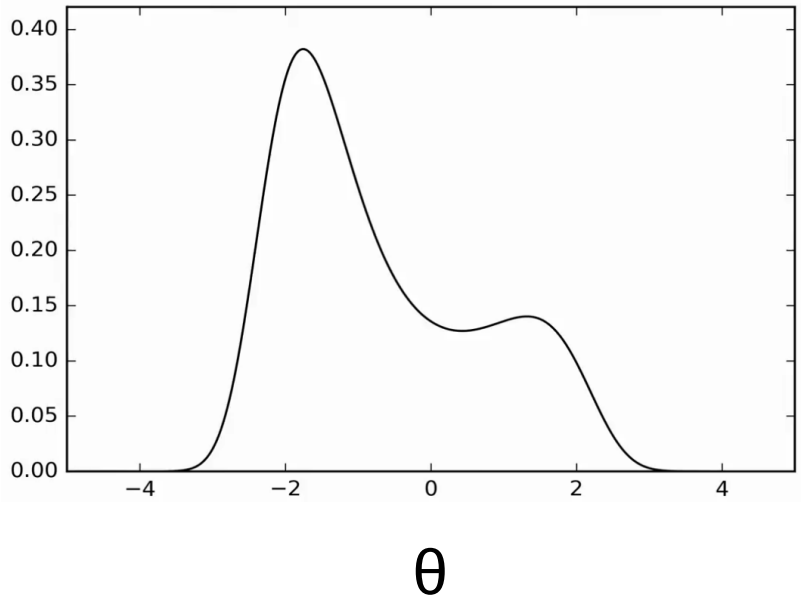
$$P(N_2 = \begin{bmatrix} A \\ G \\ C \\ T \end{bmatrix}) = \begin{bmatrix} P_{CA}(t_4)P_{GA}(t_5) \\ P_{CG}(t_4)P_{GG}(t_5) \\ P_{CC}(t_4)P_{GC}(t_5) \\ P_{CT}(t_4)P_{GT}(t_5) \end{bmatrix}$$

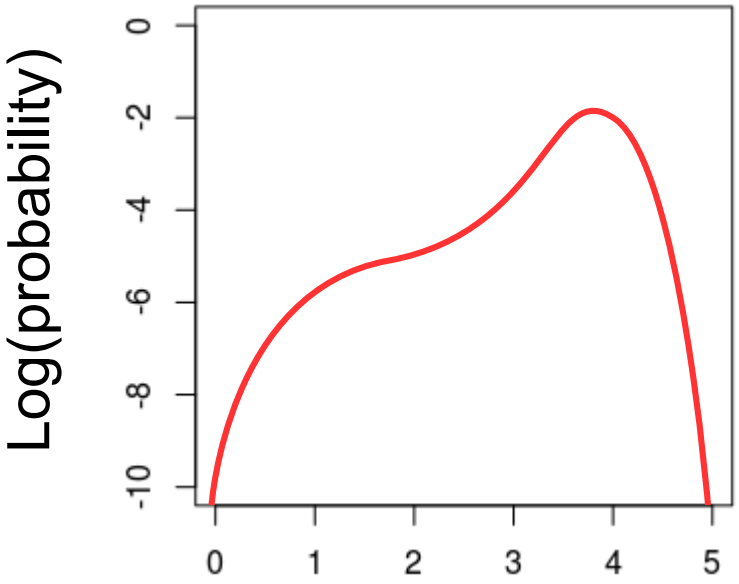
$$P(N_3 = A) = P_{CA}(t_3) \sum_{X=A,G,C,T} P_{XA}(t_7)P_{CX}(t_4)P_{GX}(t_5)$$

Exercises

5) Which algorithm would you use to find the maximum likelihood of the functions below: greedy or Monte Carlo Markov Chain

Left: MCMC to avoid local maxima
Right: greedy, no local maxima so faster





Schedule

Next week:

- 2/21 and 2/23 lectures will be on zoom (out of town)

Following week:

- 2/28 recitation, exercises, questions
- 3/2 in class exam 1:15 minutes
- Midterm exam
  - ~ 20 questions
  - Lectures 1-12 (end of next week)
  - Equations provided, calculator not needed
  - No phone, notes
- Lab06 is due Friday 3/4 midnight, not Tuesday

Today's objectives

- Selection ~ molecular evolution
- Selection ~ substitution rates (dN/dS)
- Selection ~ virus evolution

Where are we?

Alignment

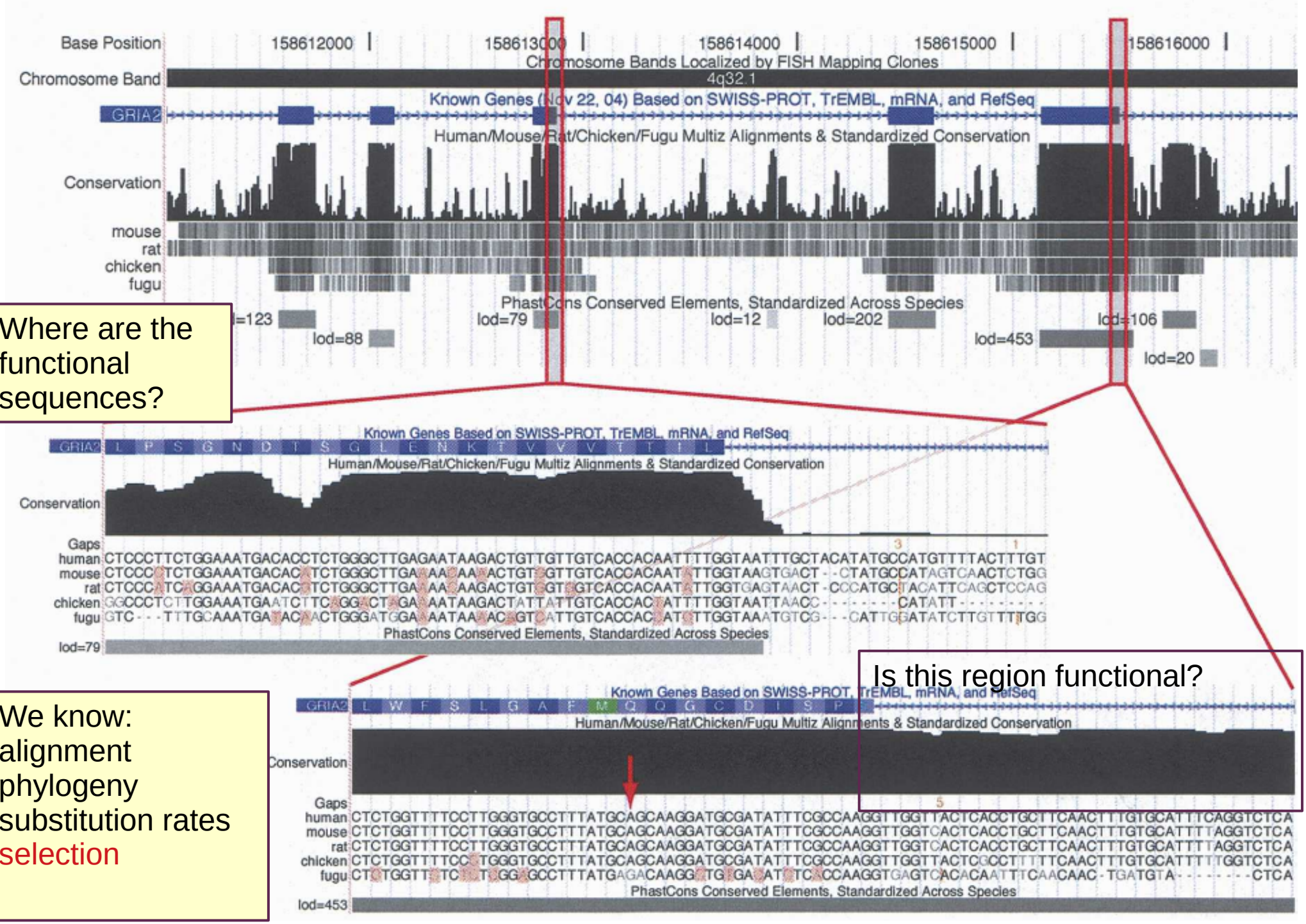
- Given a sequence, search for alignment (local, global)
- Speed vs memory
- What is the human homolog of X, where did this read come from

Substitution rate and phylogenetic trees

- Given an aligned sequence what is the substitution rate
- p(data | alignment), tree, P_{ij} transitions
- What is divergence time between X and Y, which is more closely related in a tree

Selection

- Given substitution rate has there been selection
- Is this sequence conserved (negative selection), or rapidly evolving (positive selection)
- Is this site functional (affects fitness) or not (neutral)


Base Position
158612000
158613000
158614000
158615000
158616000
Chromosome Band
Chromosome Bands Localized by FISH Mapping Clones
4q32.1
GRIA2
Known Genes (Nov 22, 04) Based on SWISS-PROT, TrEMBL, mRNA, and RefSeq
Human/Mouse/Rat/Chicken/Fugu Multiz Alignments & Standardized Conservation
Conservation
mouse
rat
chicken
fugu
PhastCons Conserved Elements, Standardized Across Species
lod=123
lod=88
lod=79
lod=12
lod=202
lod=453
lod=106
lod=20
Where are the functional sequences?
Known Genes Based on SWISS-PROT, TrEMBL, mRNA, and RefSeq
GRIA2 L P S G N D T S G L E N K T V V V T T T L
Human/Mouse/Rat/Chicken/Fugu Multiz Alignments & Standardized Conservation
Conservation
Gaps
human CTCCCTTCTGGAAATGACACCTCTGGGCTTGAGAATAAGACTGTTGTTGTCACCACAATTTTGGTAATTTGCTACATATGCCATGTTTTACTTTGT
mouse CTCCCTCTGGAAATGACACTCTGGGCTTGAAAAAAACTGTTGTTGTCACCACAATTTTGGTAAGTGACT--CTATGCCATAGTCAACTCTGG
rat CTCCCATCAGGAAATGACACTCTGGGCTTGAAAAAGACTGTTGTCGTCACCACAATTTTGGTGAGTAACT-CCCATGCTACATTCAGCTCCAG
chicken GGCCCTCTTGGAAATGAATCTTCAGGACTAGAAATAAGACTATTATTGTCACCACTATTTTGGTAATTAACC----------CATATT----------
fugu GTC---TTTGCAAATGATACAACTGGGATGGAAATAAACCGTCATTGTCACCACATCTTGGTAAATGTCG---CATTGGATATCTTGTTTTGG
PhastCons Conserved Elements, Standardized Across Species
lod=79
Is this region functional?
Known Genes Based on SWISS-PROT, TrEMBL, mRNA, and RefSeq
GRIA2 L W F S L G A F M Q Q G C D I S P
Human/Mouse/Rat/Chicken/Fugu Multiz Alignments & Standardized Conservation
Conservation
Gaps
human CTCTGGTTTTCCTTGGGTGCCTTTATGCAGCAAGGATGCGATATTTCGCCAAGGTTGGTTACTCACCTGCTTCAACTTTGTGCATTTTCAGGTCTCA
mouse CTCTGGTTTTCCTTGGGTGCCTTTATGCAGCAAGGATGCGATATTTCGCCAAGGTTGGTCACTCACCTGCTTCAACTTTGTGCATTTTAGGTCTCA
rat CTCTGGTTTTCCTTGGGTGCCTTTATGCAGCAAGGATGCGATATTTCGCCAAGGTTGGTCACTCACCTGCTTCAACTTTGTGCATTTTAGGTCTCA
chicken CTCTGGTTTTCCTTGGGTGCCTTTATGCAGCAAGGATGCGATATTTCGCCAAGGTTGGTTACTCGCCTTTTTCAACTTTGTGCATTTTTGGTCTCA
fugu CTCTGGTTTTCTTTGGGCCTTTATGAGACAAGGTGAGATTTCACCAAGGTGAGTCACACAATTTCAACAAC-TGATGTA-------CTCA
PhastCons Conserved Elements, Standardized Across Species
lod=453
We know:
alignment
phylogeny
substitution rates
selection

Likelihood Models of Molecular Evolution

Key Assumptions:

- Alignments are correct
- Sites are independent (not CpG sites)
- Stationarity (Q is constant over time)
- Time reversibility $\pi_x \mu_{xy} = \pi_y \mu_{yx}$
- Tree is correct
- Rate ~ Mutation, time & selection

Problem: How

Problem: how do we distinguish mutation from selection?

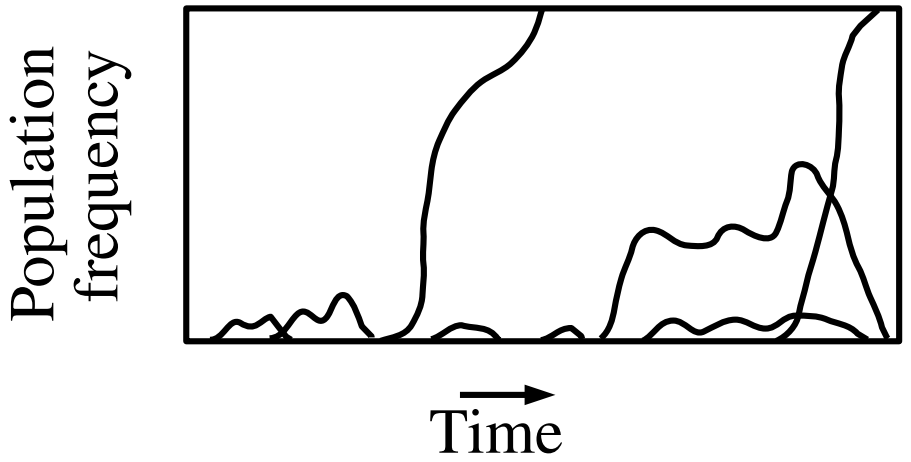
Mutation or selection? (time is the same)

```
ATGCATGCATGATATGCGCGTGCTTACCAGCTCGCGCGGTTATCGTCGCGC
....................G............T.G.T..A.T...A..G.T
```

Low mutation rate
or
Negative selection
(functionally conserved)

High mutation rate
or
Positive selection
(functionally diverged)

Difference between mutation rate and substitution rate.



Mutation rate the chance of a mutation occurring in each generation or cell division (does NOT depend on selection).
Mutation = variation created
Substitution rate the frequency at which mutations become fixed within a population (depends on mutation & selection).
Fixation = variation lost
Substitution rate = mutation rate * fixation probability * time
Fixation probability depends on selection only

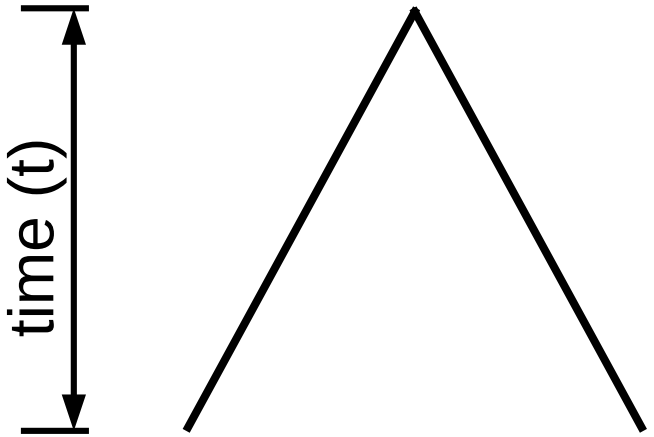
Substitution Rates with Selection

Substitution rate = mutation rate * fixation probability * time

Substitution rate for neutral mutations: $2N\mu * 1/2N * t = \mu t$

Substitution rate for adaptive mutations: $2N\mu_a * 2s * t = 4Ns\mu_a t$ for $4Ns > 1$

No selection: The substitution rate between two species is K = 2μt.



A B K(AB) = 2ut

Substitution Rates with Selection

Substitution rate = mutation rate * fixation probability * time

Substitution rate for neutral mutations: $2N\mu * 1/2N * t = \mu t$

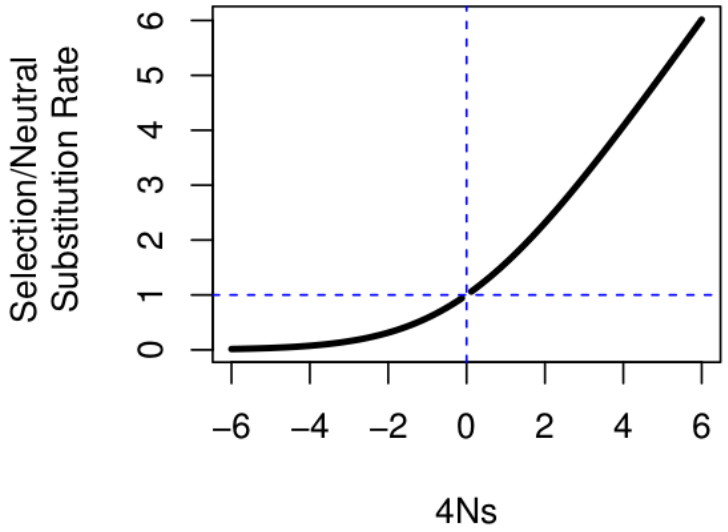
Substitution rate for adaptive mutations: $2N\mu_a * 2s * t = 4Ns\mu_a t$ for $4Ns > 1$

No selection: The substitution rate between two species is K = 2μt.

Fixation probability: $f = \frac{1-e^{-2s}}{1-e^{-4Ns}} \approx \frac{2s}{1-e^{-4Ns}}$

	AA	Aa	aa
Fitness	W_{AA}	W_{Aa}	W_{aa}
selection	1	1-hs	1-s

s = selection coefficient [0,1]
h = dominance [0,1]
W = fitness, reproductive output



|4Ns| > 5 is strongly influenced by selection
If N = 10^5 then |s| > 1.25 x 10^{-5} is strong selection

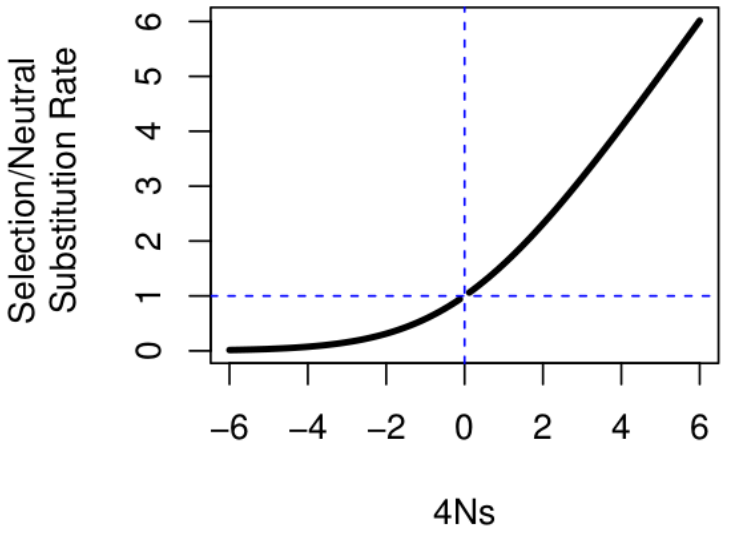
What if one region is neutral i.e. no selection?

Mutation or selection? (time is the same)

```
ATGCATGCATGATATGCGCGTGCTTACCAGCTCGCGCGGTTATCGTCGCGC
....................G............T.G.T..A.T...A..G.T
```

Low mutation rate
or
Negative selection

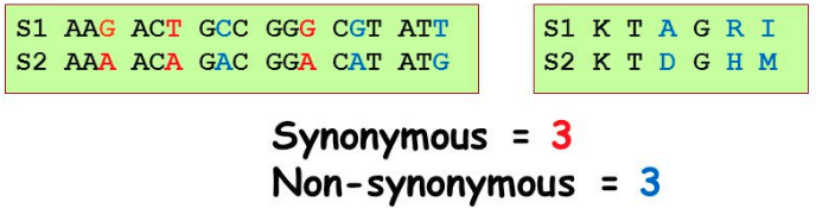
High mutation rate
or
Positive selection



Solution: Compare Nonsynonymous and synonymous sites

Table 1. The genetic code.

Codon	AA	Codon	AA	Codon	AA	Codon	AA
TTT	Phe	TCT	Ser	TAT	Tyr	TGT	Cys
TTC	Phe	TCC	Ser	TAC	Tyr	TGC	Cys
TTA	Leu	TCA	Ser	TAA	Stop	TGA	Stop
TTG	Leu	TCG	Ser	TAG	Stop	TGG	Trp
CTT	Leu	CCT	Pro	CAT	His	CGT	Arg
CTC	Leu	CCC	Pro	CAC	His	CGC	Arg
CTA	Leu	CCA	Pro	CAA	Gln	CGA	Arg
CTG	Leu	CCG	Pro	CAG	Gln	CGG	Arg
ATT	Ile	ACT	Thr	AAT	Asn	AGT	Ser
ATC	Ile	ACC	Thr	AAC	Asn	AGC	Ser
ATA	Ile	ACA	Thr	AAA	Lys	AGA	Arg
ATG	Met	ACG	Thr	AAG	Lys	AGG	Arg
GTT	Val	GCT	Ala	GAT	Asp	GGT	Gly
GTC	Val	GCC	Ala	GAC	Asp	GGC	Gly
GTA	Val	GCA	Ala	GAA	Glu	GGA	Gly
GTG	Val	GCG	Ala	GAG	Glu	GGG	Gly



Because synonymous and nonsynonymous sites are interspersed between one another, they should have the same mutation rate

dN or K_a = the nonsynonymous substitution rate = # nonsynonymous changes / # nonsynonymous sites.

dS or K_s = the synonymous substitution rate = # synonymous changes / # synonymous sites.

synonymous change – does not alter amino acid

nonsynonymous change – alters the amino acid

nondegenerate sites = codon positions where each nucleotide changes alters the amino acid

2-fold degenerate sites = codon positions where 4 nucleotides encode two amino acids

4-fold degenerate sites = codon positions where all for nucleotides encode for the same amino acid

Selection (dN/dS)

Detecting selection using the nucleotide substitution rate
dN = # nonsynonymous changes / # nonsynonymous sites.
dS = # synonymous changes / # synonymous sites.

Interpretation of dN/dS ratios (assuming synonymous sites are neutral):
dN/dS = 1
No constraint on protein sequence, i.e. nonsynonymous changes are neutral.
dN/dS < 1
Selection constrains changes in the protein sequence, i.e. nonsynonymous mutations are deleterious.
dN/dS > 1
Selection acts to change the function of the protein sequence, i.e. nonsynonymous mutations are adaptive.

Selection (dN/dS)

Interpreting dN/dS > 1; selection on dN or dS

dN is increased by positive selection

dS is decreased by negative selection

Interpreting dN/dS < 1; average rates in time/space

Syn sites neutral

10 syn sites (dS = 0.1) and 10 nsn sites

1 nsn site under positive selection (dN = 0.2)

9 nsn sites under negative selection (dN = 0)

average dN = 0.02, dN/dS = 0.2

2 nsn site neutral (dN = 0.1)

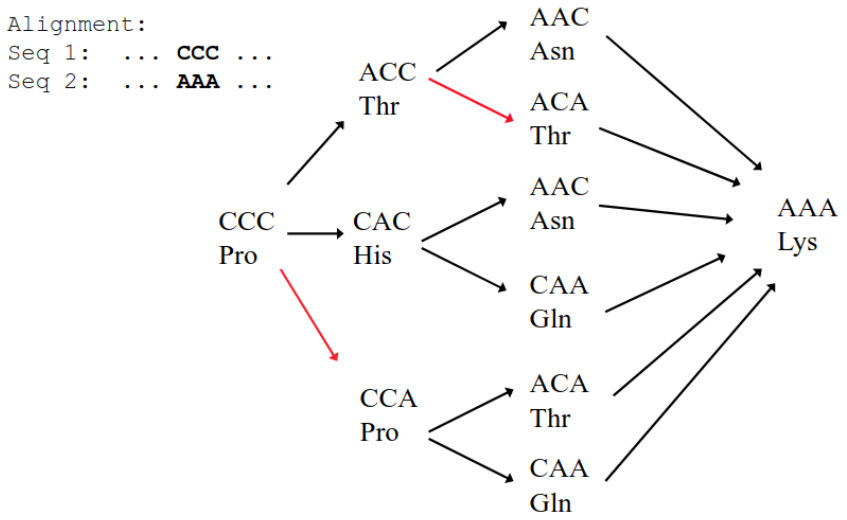
8 nsn sites under negative selection (dN = 0)

average dN = 0.02, dN/dS = 0.2

dN/dS is an upper bound on fraction of neutral nsn sites

Codon models
Goldman & Yang (1994)

$$Q_{ij} = \begin{cases} 0, \textit{more than one change} \\ \pi_j, \textit{synonymous transversion} \\ \kappa \pi_j, \textit{synonymous transition} \\ \omega \pi_j \textit{nonsynonymous tranversion} \\ \omega \kappa \pi_j \textit{nonsynonymous transition} \end{cases}$$



ω = dN/dS
κ = transition/transversion
π = equilibrium nucleotide freq.

Why more than one change per codon is not included:

- Some paths use 2 others use 3 nonsynonymous changes.

Nucleotide substitution models with selection

HKY85

$$Q=\begin{pmatrix} * & \pi_G \kappa & \pi_C & \pi_T \\ \pi_A \kappa & * & \pi_C & \pi_T \\ \pi_A & \pi_G & * & \pi_T \kappa \\ \pi_A & \pi_G & \pi_C \kappa & * \end{pmatrix}$$

	codon i	codon j
fitness	1	1+s

$$f_{ij}=\frac{2 s_{ij}}{1-e^{-2 N s_{ij}}}$$

Model with selection:

$$K = N \mu f t$$

$$Q_{ij} = N \mu_{ij} f_{ij}$$

Haploid
f = P(fixation)
t = time
N = pop. size
u = mutation

Codon substitution models with selection

HKY85

$$Q=\begin{pmatrix} * & \pi_G \kappa & \pi_C & \pi_T \\ \pi_A \kappa & * & \pi_C & \pi_T \\ \pi_A & \pi_G & * & \pi_T \kappa \\ \pi_A & \pi_G & \pi_C \kappa & * \end{pmatrix}$$

Model with selection: Haploid

$$K = N \mu f t$$

$$Q_{ij} = N \mu_{ij} f_{ij}$$

f = P(fixation)
t = time
N = pop. size
u = mutation

$$\mu_{ij} = \mu \kappa \pi_j \quad \text{for HKY85}$$

	codon i	codon j
fitness	1	1+s

$$f_{ij} = \frac{2 s_{ij}}{1 - e^{-2 N s_{ij}}}$$

$$Q_{ij} = N u_{ij} \frac{2 s_{ij}}{1 - e^{-2 N s_{ij}}}$$

$$Q_{ij} = u \kappa \pi_j \frac{S_{ij}}{1 - e^{-S_{ij}}}$$

$$S_{ij} = 2 N s_{ij}$$

Codon substitution models with selection

HKY85

$$Q=\begin{pmatrix} * & \pi_G \kappa & \pi_C & \pi_T \\ \pi_A \kappa & * & \pi_C & \pi_T \\ \pi_A & \pi_G & * & \pi_T \kappa \\ \pi_A & \pi_G & \pi_C \kappa & * \end{pmatrix}$$

Q can be:
- branch specific
- site specific

$$Q_{ij} = u \kappa \pi_j W_{ij}$$

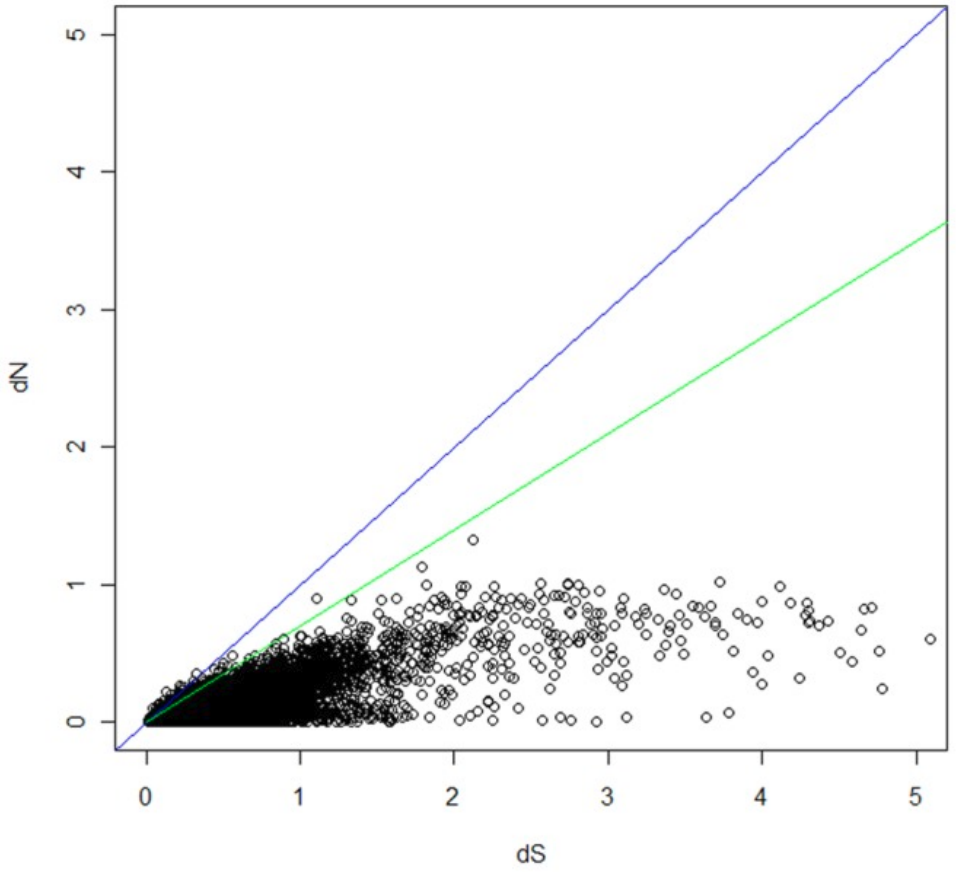
Mutation

Selection

$$W_{ij} = \frac{S_{ij}}{1 - e^{-S_{ij}}}$$

$$Q=\begin{pmatrix} * & \pi_G \kappa W_{AG} & \pi_C W_{AC} & \pi_T W_{AT} \\ \pi_A \kappa W_{GA} & * & \pi_C W_{GC} & \pi_T W_{GT} \\ \pi_A W_{CA} & \pi_G W_{CG} & * & \pi_T \kappa W_{CT} \\ \pi_A W_{TA} & \pi_G W_{TG} & \pi_C \kappa W_{TC} & * \end{pmatrix}$$

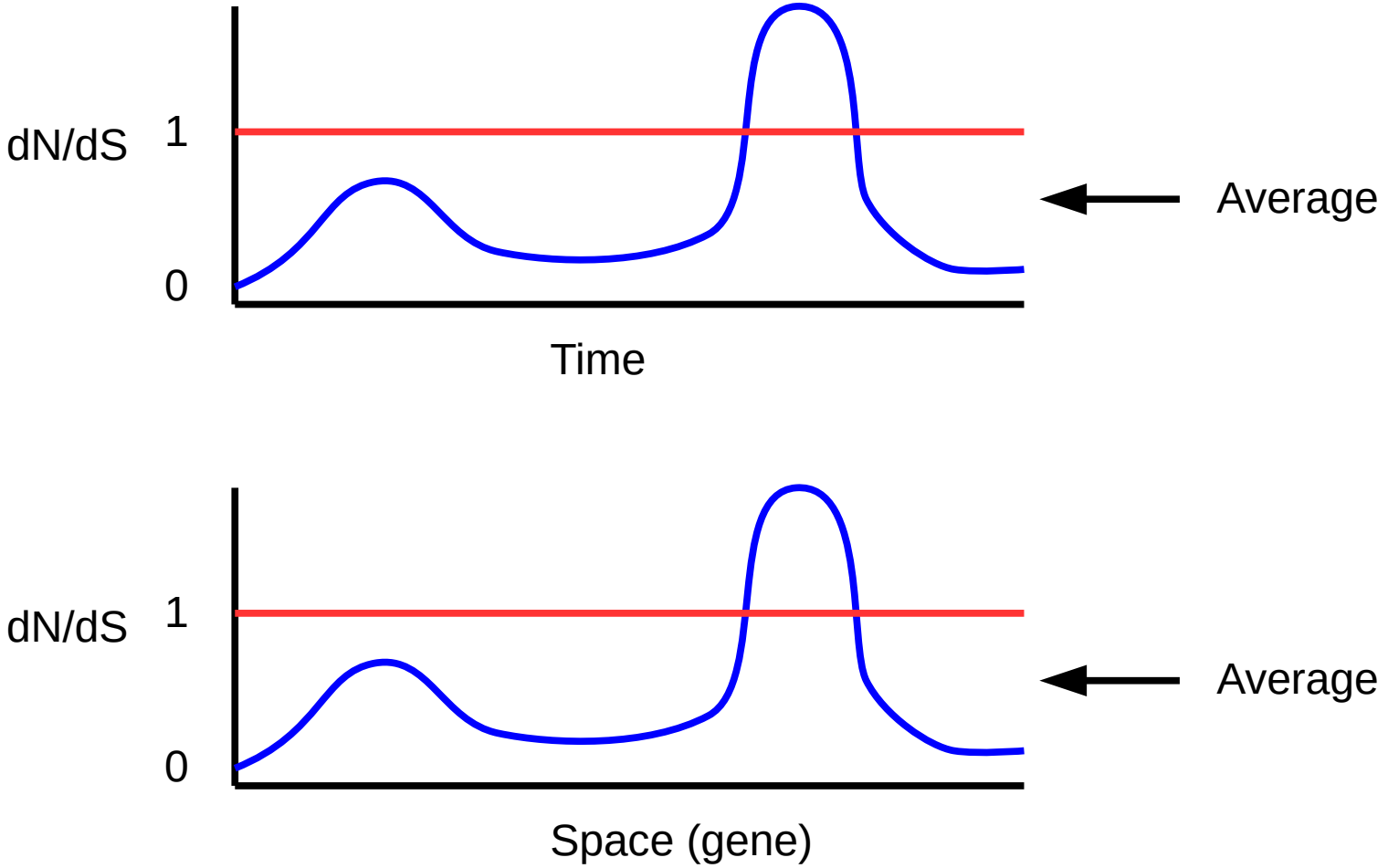
dN/dS across the genome



Results from genome-wide dN/dS:

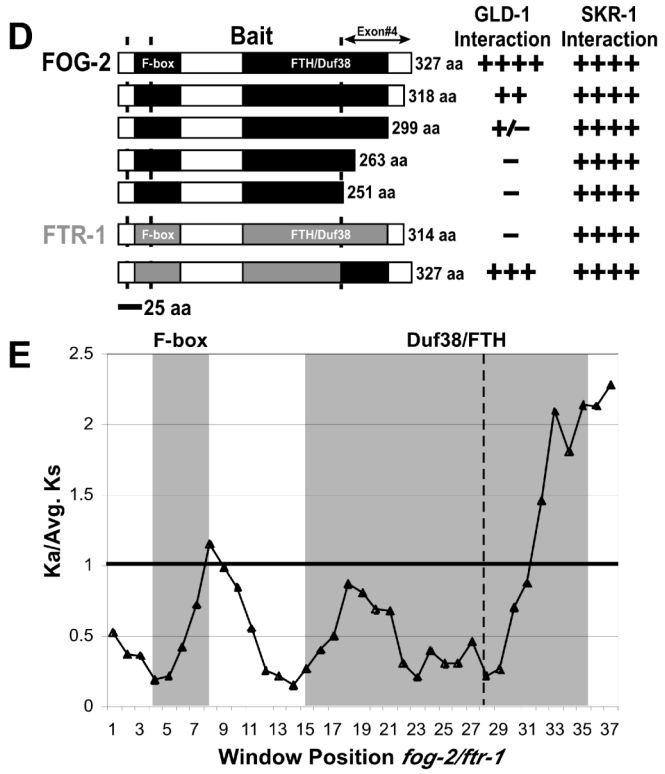
- most genes dN/dS << 1
  - positive selection is rare in space or time
- genes with dN/dS > 1
  - positive selection is common in space and time
  - sexual selection
  - host-pathogen interactions

Positive selection in space and time



An example: Rapidly Evolving Region = Loss of protein interaction

dN increased by positive selection
dN decreased by negative selection
Problem: dN of a gene may be influenced by both positive and negative selection and so can be less than dS



Solution:
- sliding window of dN/dS
- regions with dN/dS > 1 due to positive selection

Nayak et al. 2005

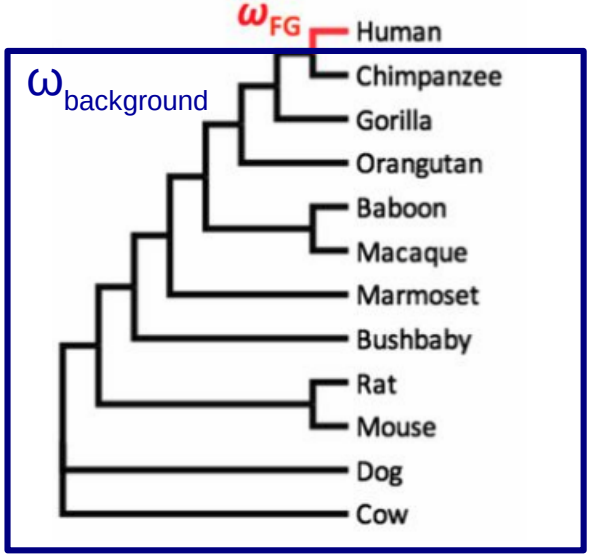
Detecting positive selection

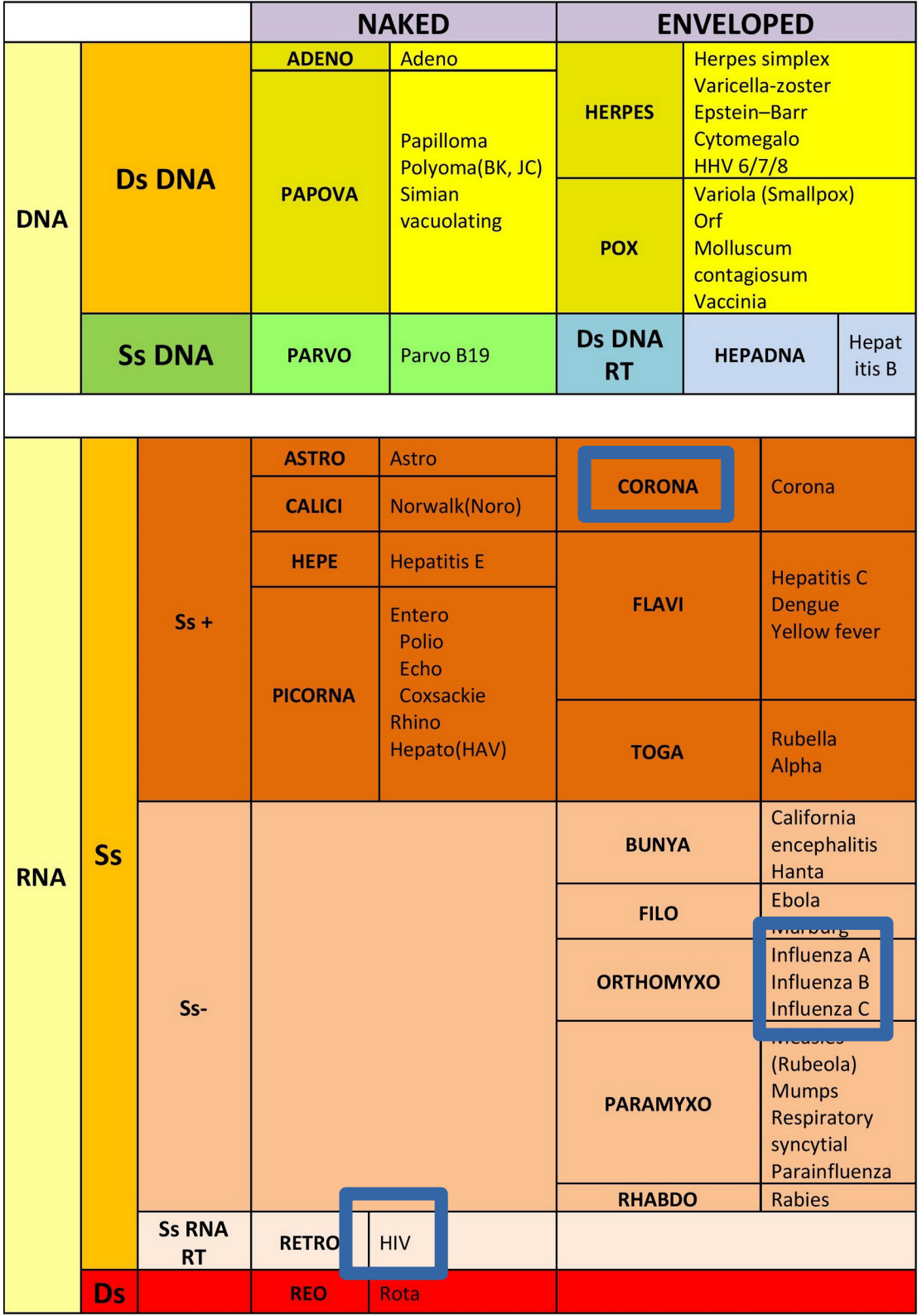
- Sliding window
- Branch model (e.g. dN/dS > 1 for human branch)
- Site model (e.g. dN/dS > 1 for a single codon)
- Branch*Site model

TPQLEKKIKRQS
TPQLEKKCKRQS
TPQLEKKLKRQS
TPQLEKKAKRQS
TPQLEKKIKRQS
TPQLEKKVKRQS
TPQLEKKVKRQS
TPQLEKKVKRQS
TPQLEKKIKRQS

ω is site specific
ω is branch specific

Interpretation?
dS = 1
dS = 12



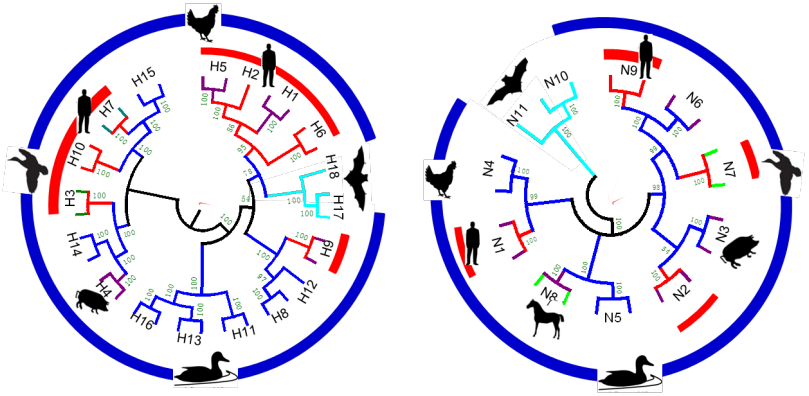


Virus classes

- Viruses must continuously evade immune system by positive selection
- Virus origins can be inferred using phylogenetics
- Virus epidemics (selection and spread) reflected in their genome

Influenza

hemagglutinin (HA) and neuraminidase (NA)

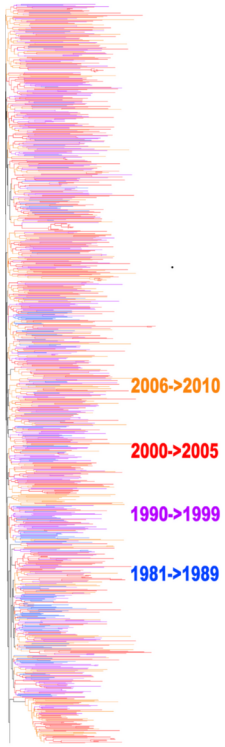


- Influenza (flu) is caused by single stranded RNA *Influenza virus* (13.6 kb)
- Influenza hemagglutinin (HA) is a glycoprotein found on the surface of influenza viruses and is responsible for binding the virus to cells with sialic acid on the membranes.
- Influenza neuraminidase (NA) is a surface protein that enables virus to be released
- HA and NA can recombine to make different serotypes (immune recognized types)
- H1N1 serotype – 1918 Spanish flu and 2009 Swine flu

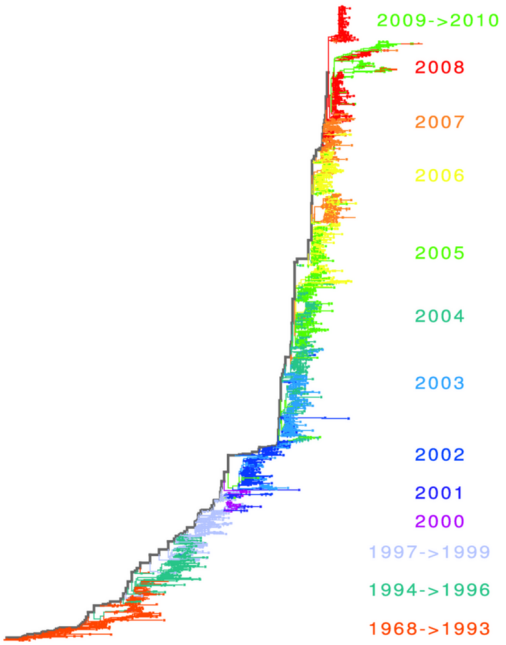
Influenza vs HIV: selection and tree

- Tree shows HA spread by year
- Trunk of tree shows rapid evolution
- Selection driven by immunological escape
- Prior year is predictive of subsequent years (vaccine)
- HIV shows rapid evolution on tips/leaves of tree

HIV-1 env gene



Influenza HA gene



Vaccine Predictive!

Each season Influenza must evolve to infect new hosts since last years hosts have immune protection: trunk = rapid

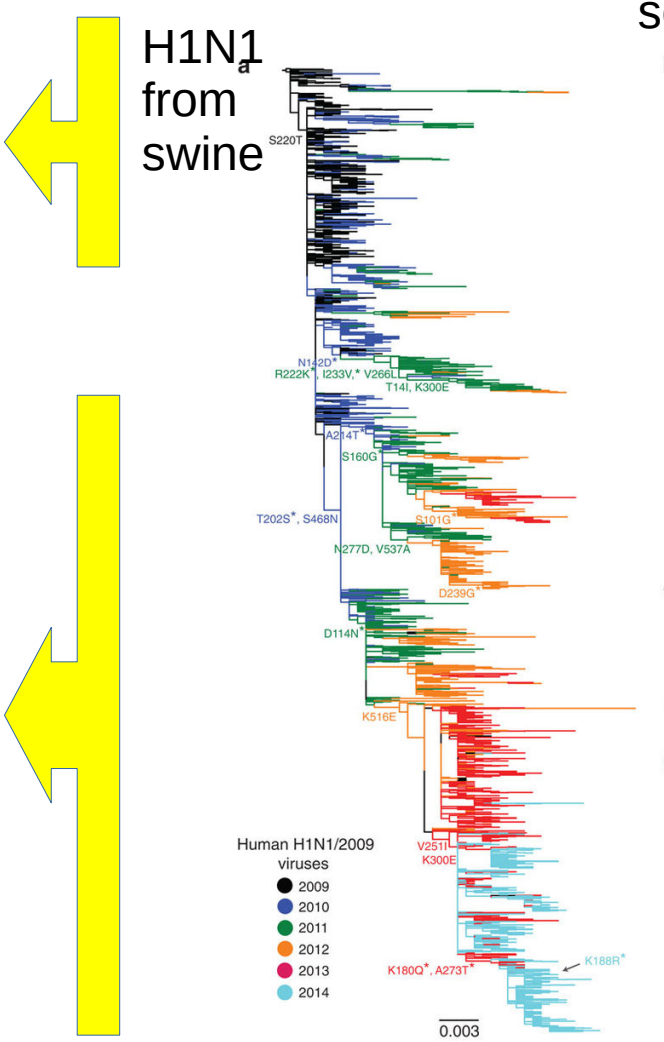
HIV adaptation differs; it evolves rapidly within hosts and suppresses immune response: leaves = rapid
Population is naive so no trunk evolution

Su et al. (2015); Pompei et al. (2012)

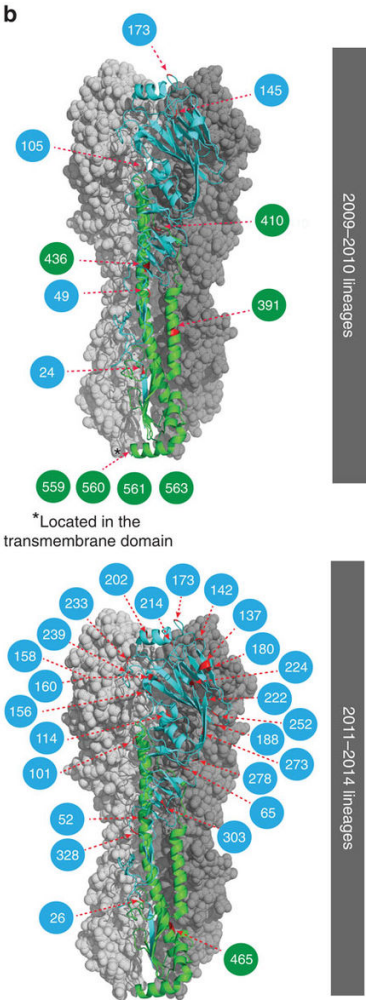
Influenza: H1N1 2009 outbreak

H1N1 2009 originated from Swine

- 2009 rapid adaptation to human host followed by diversification
- 2010 and later evolved ladder like immunological escape



Sites under positive selection

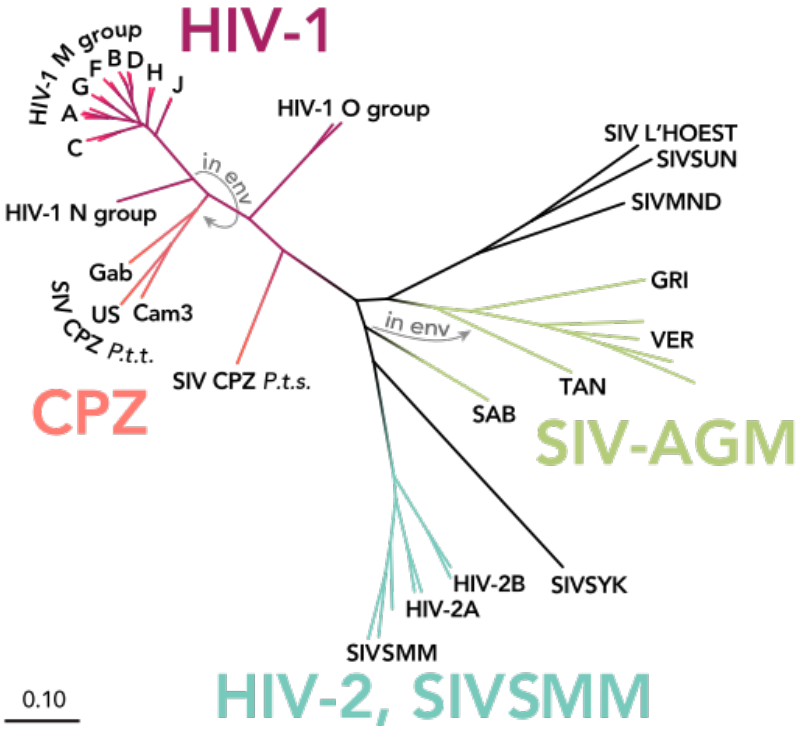


Yvonne et al (2015)

HIV-1 Phylogeny and Dates

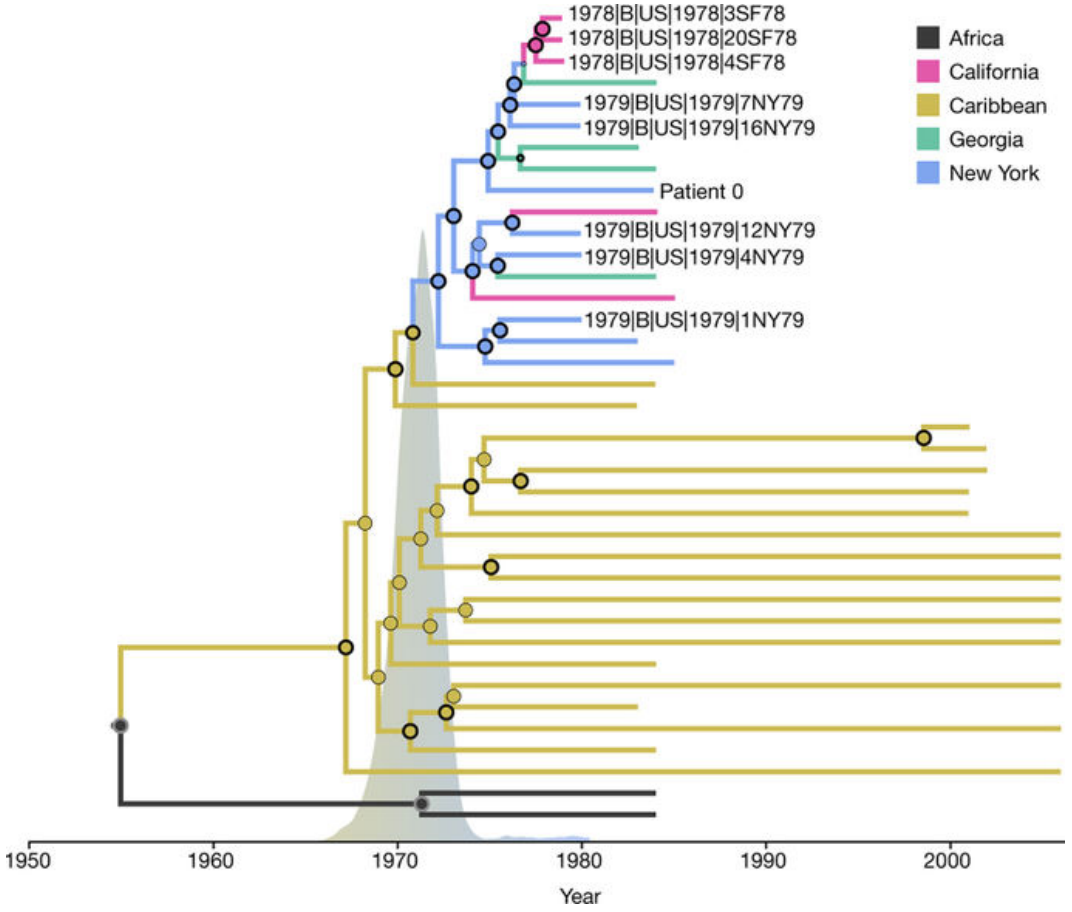
HIV is a lentivirus that causes AIDS

- Derived from SIV (simian immunodeficiency virus)
- High genetic variation and groups into subtypes
- Subtype M originated in 1910 and first known case was 1959
- Subtype N and O were also later transferred from chimps

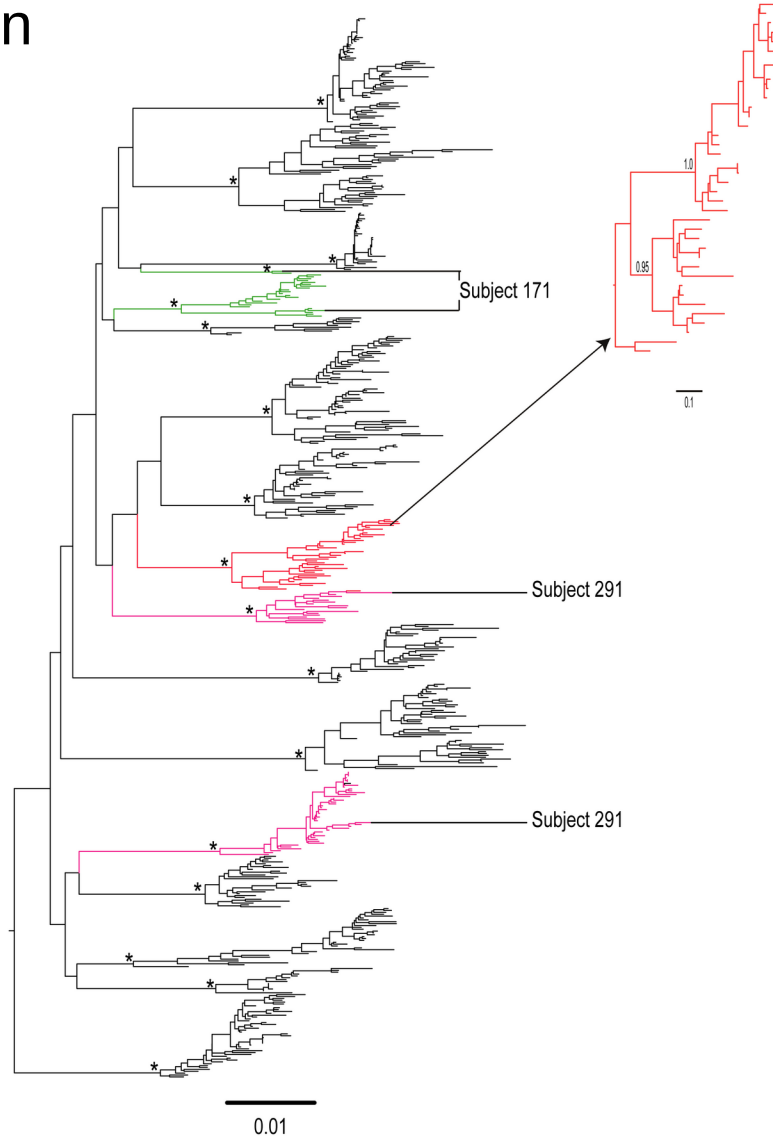


HIV-1 Rapid evolution

- Between individuals: migration, transmission
- Within individuals: selection evade host immune system (dN/dS > 1)



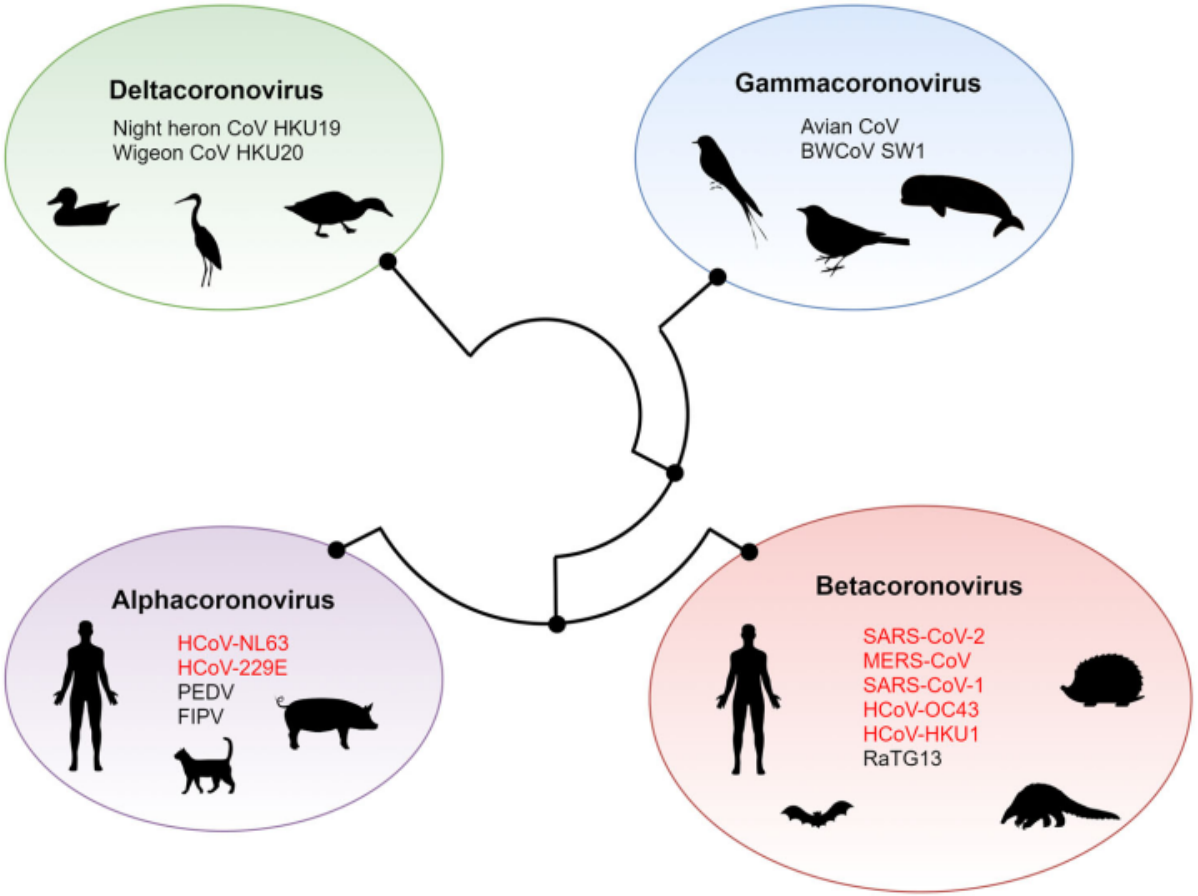
Between individuals
Geographic spread

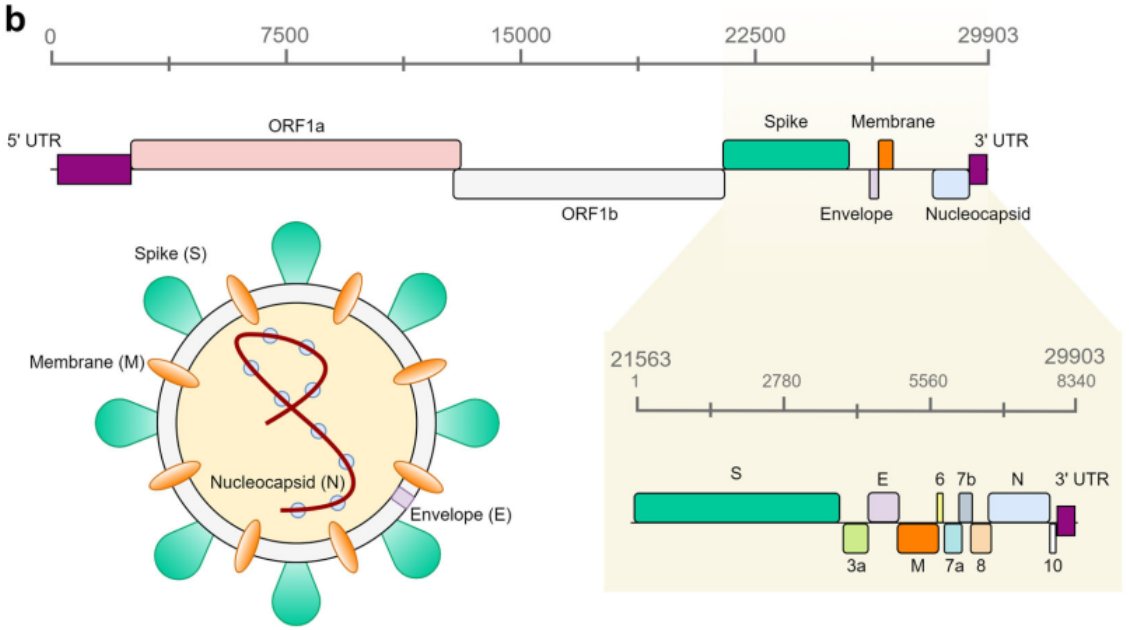


Within individuals

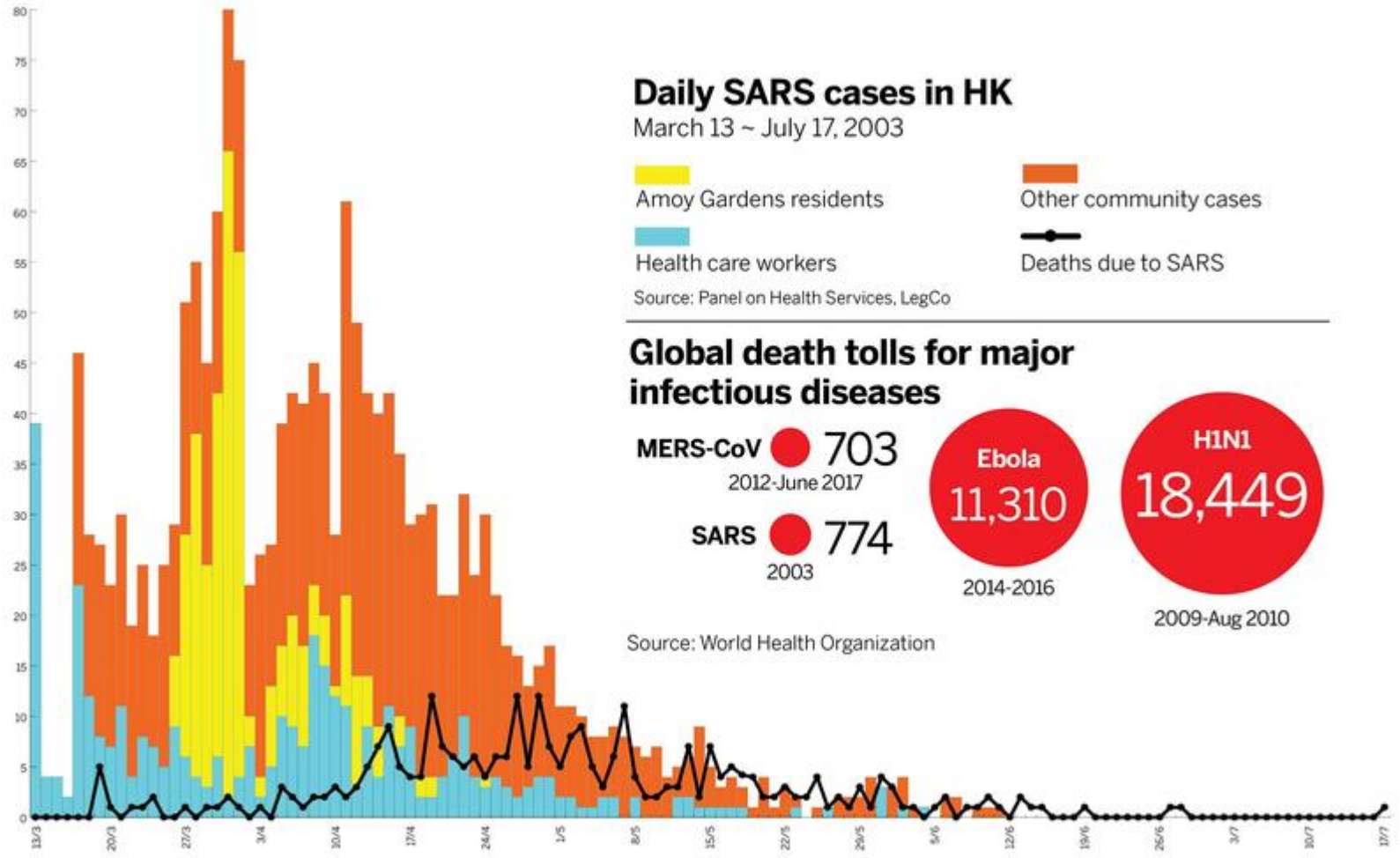
Worobey et al. (2016) Nature

Corona viruses



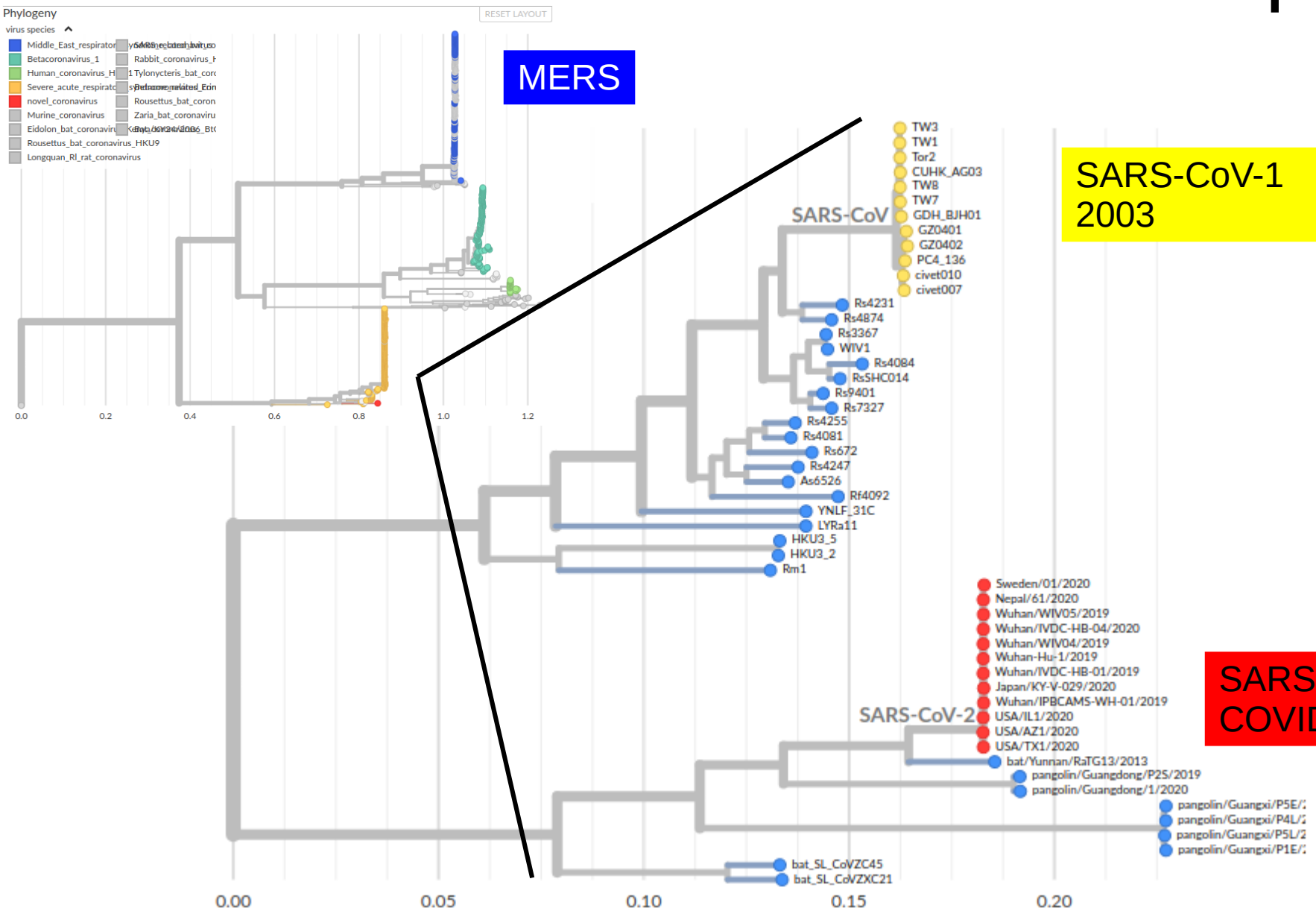


SARS and MERS and COVID-19



SARS: 2002/3 Severe Acute Respiratory Syndrome (coronavirus), 9.6% fatality, origin = horseshow bats to civets to humans
MERS: 2012 Middle East Respiratory Syndrome-related coronavirus, 30% fatality rate, origin = bats
COVID-19: 2019/20 Coronavirus disease 2019, 1-3% fatality, origin likely bats

Beta-Corona virus relationships

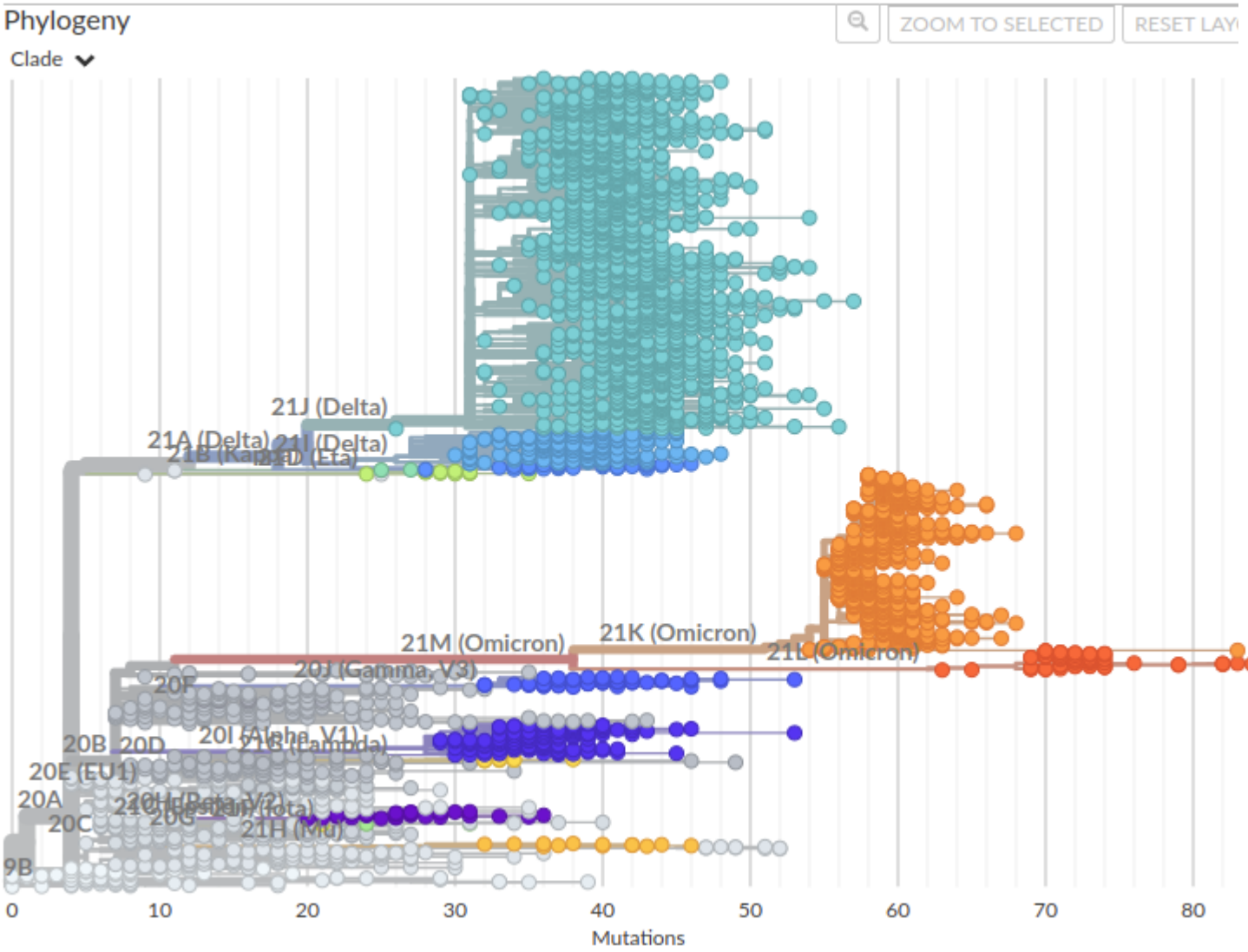


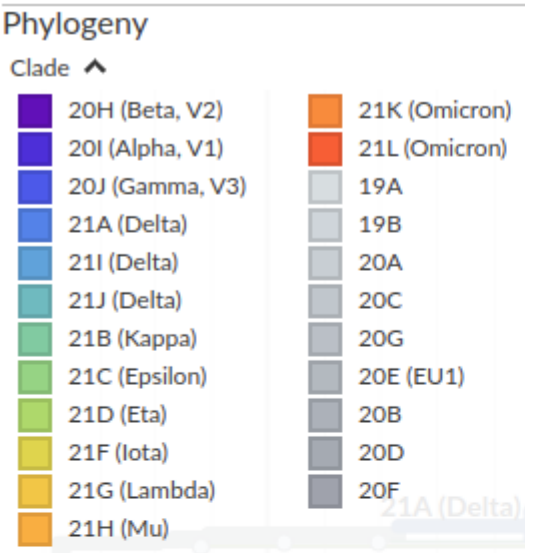
SARS-CoV-1
2003

SARS-CoV-2
COVID-19

https://nextstrain.org/ncov

SARS-CoV-2

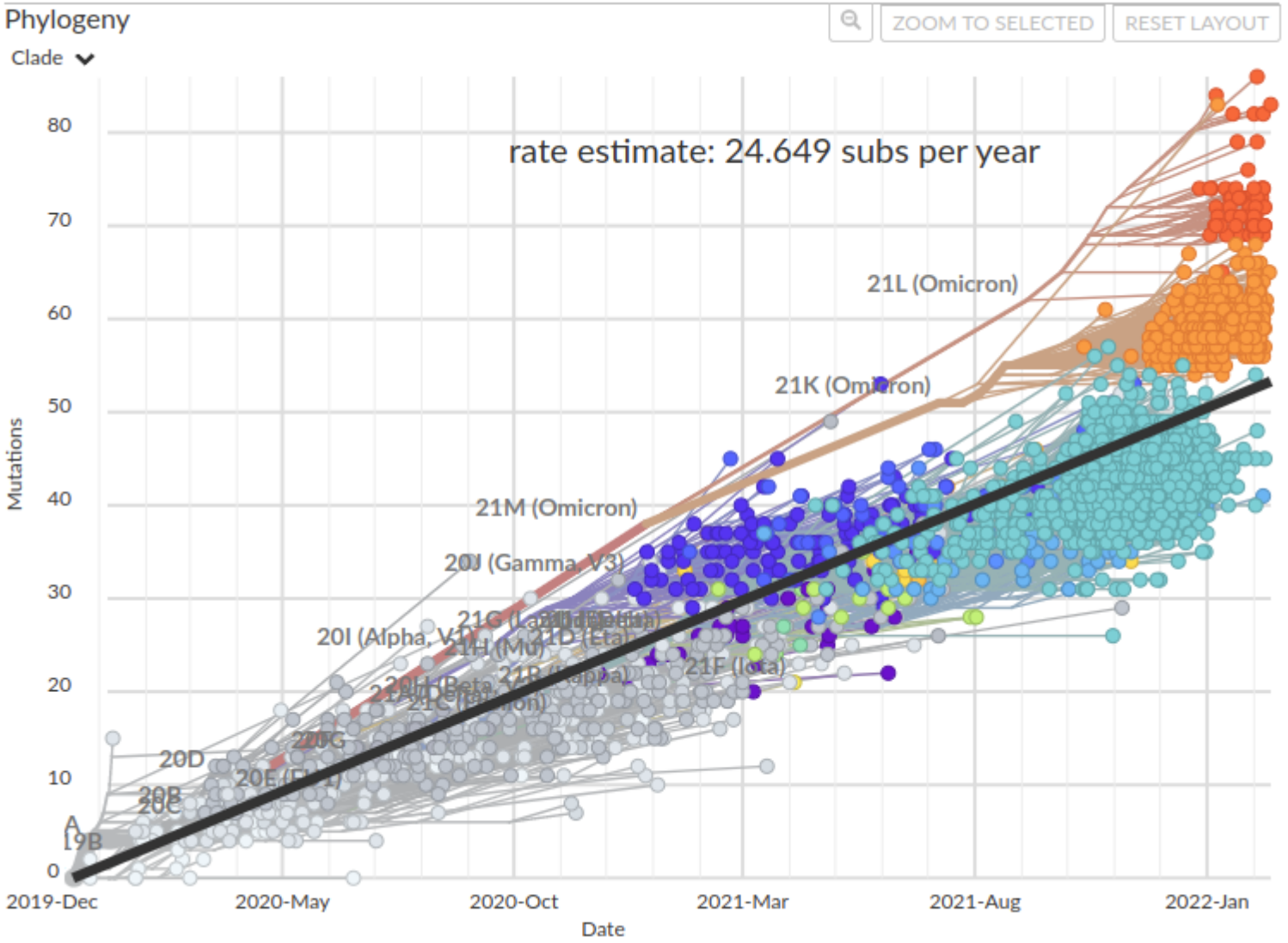




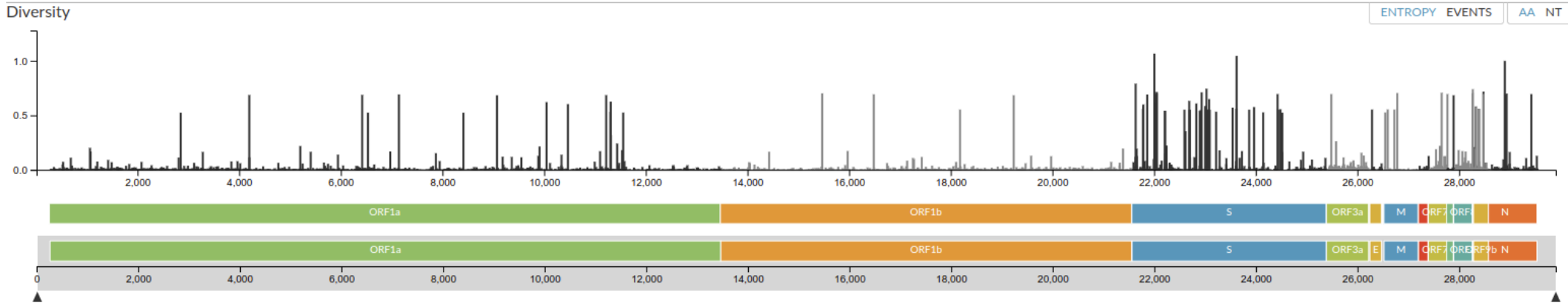
- Geographic patterns are present
- Tree is not ladderlike
- Major clades show burst of evolution (mutations)

nextstrain.org

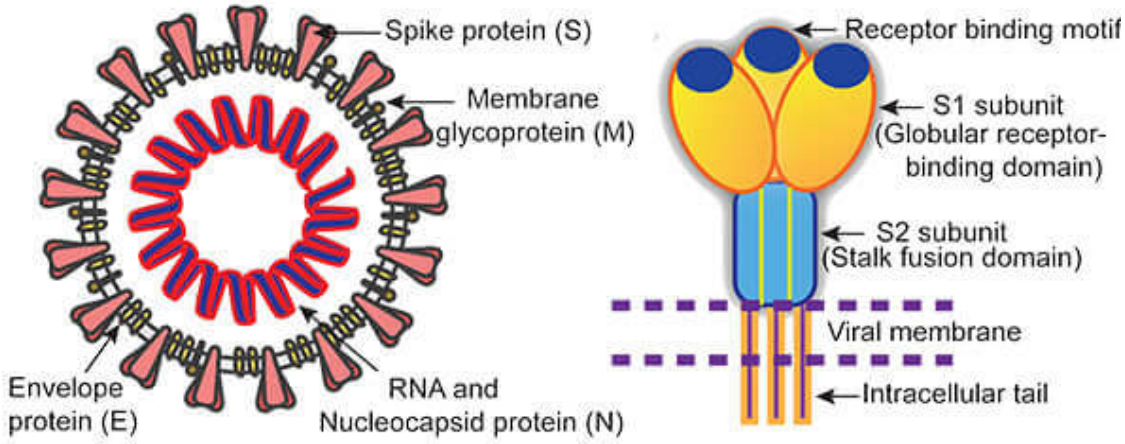
SARS-CoV-2

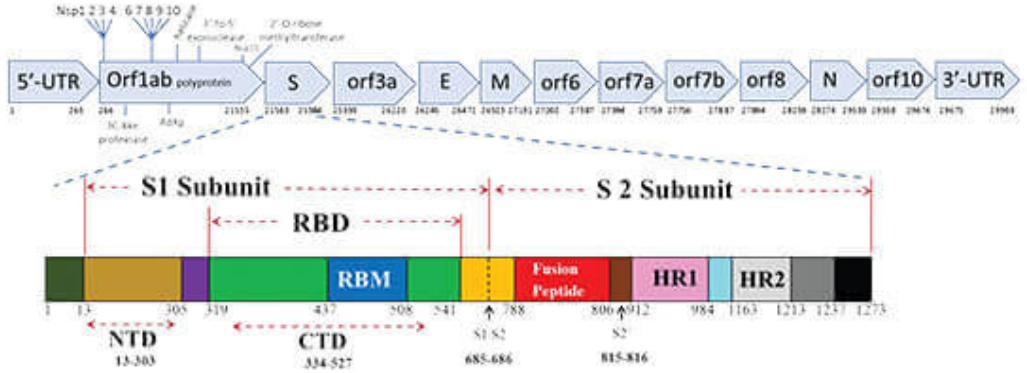

Phylogeny
Clade
ZOOM TO SELECTED
RESET LAYOUT
rate estimate: 24.649 subs per year
21L (Omicron)
21K (Omicron)
21M (Omicron)
20J (Gamma, V3)
20I (Alpha, V1)
21H (Mu)
21D (Eta)
21F (Iota)
20D
20B
20C
A
19B
Mutations
0
10
20
30
40
50
60
70
80
2019-Dec
2020-May
2020-Oct
2021-Mar
2021-Aug
2022-Jan
Date

Evolution across the genome

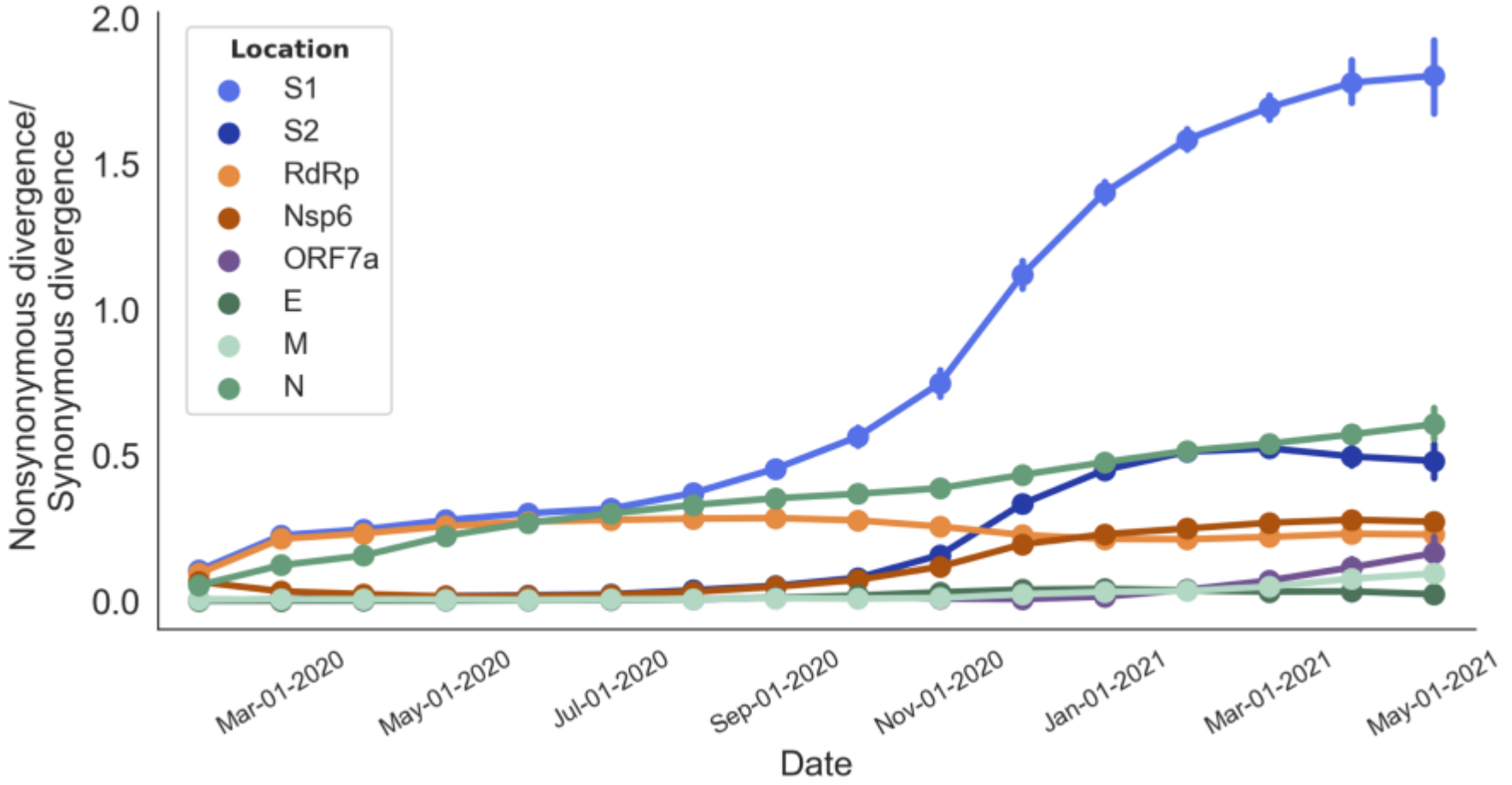


High rate of evolution in the S1 subunit of the spike protein





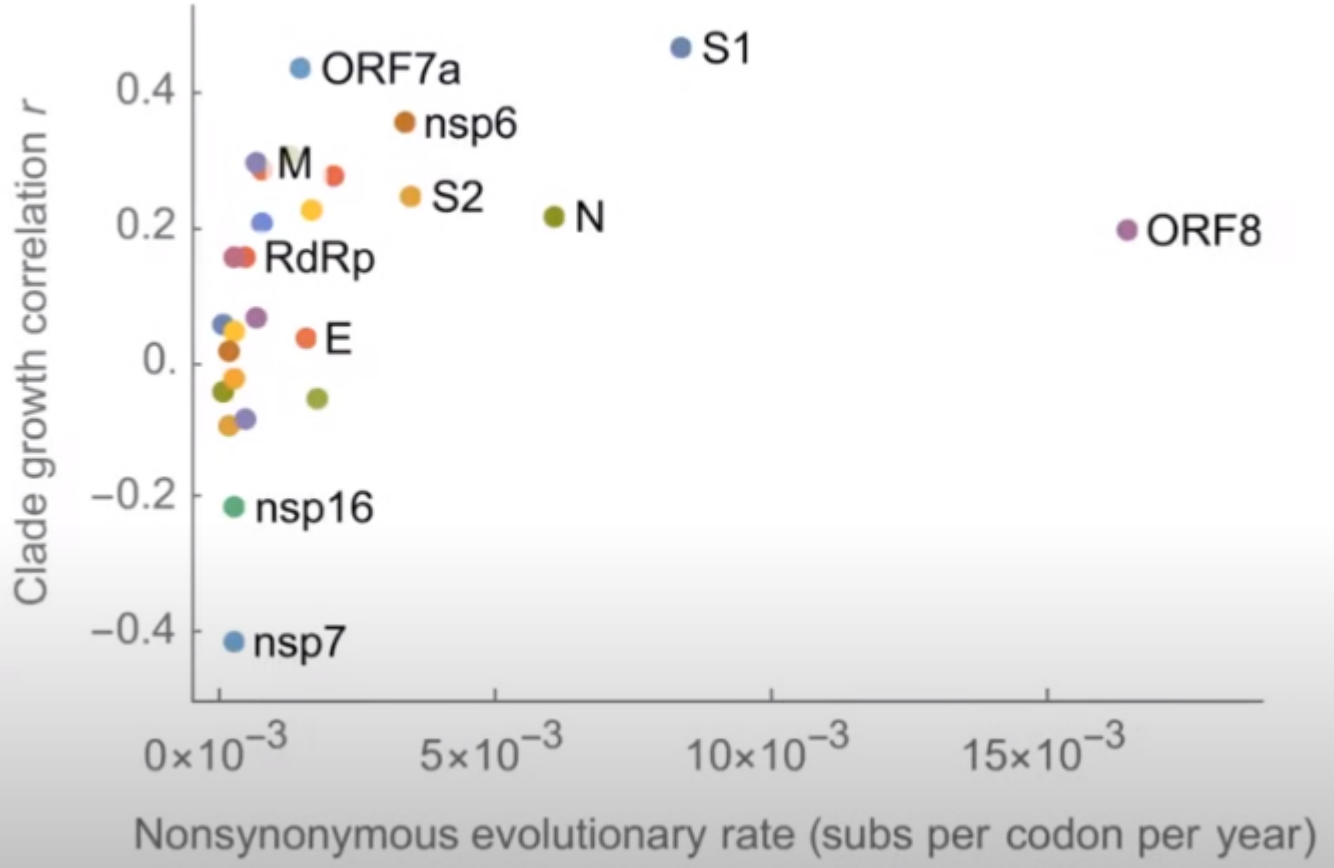
Positive selection



Spike protein has dN/dS > 1

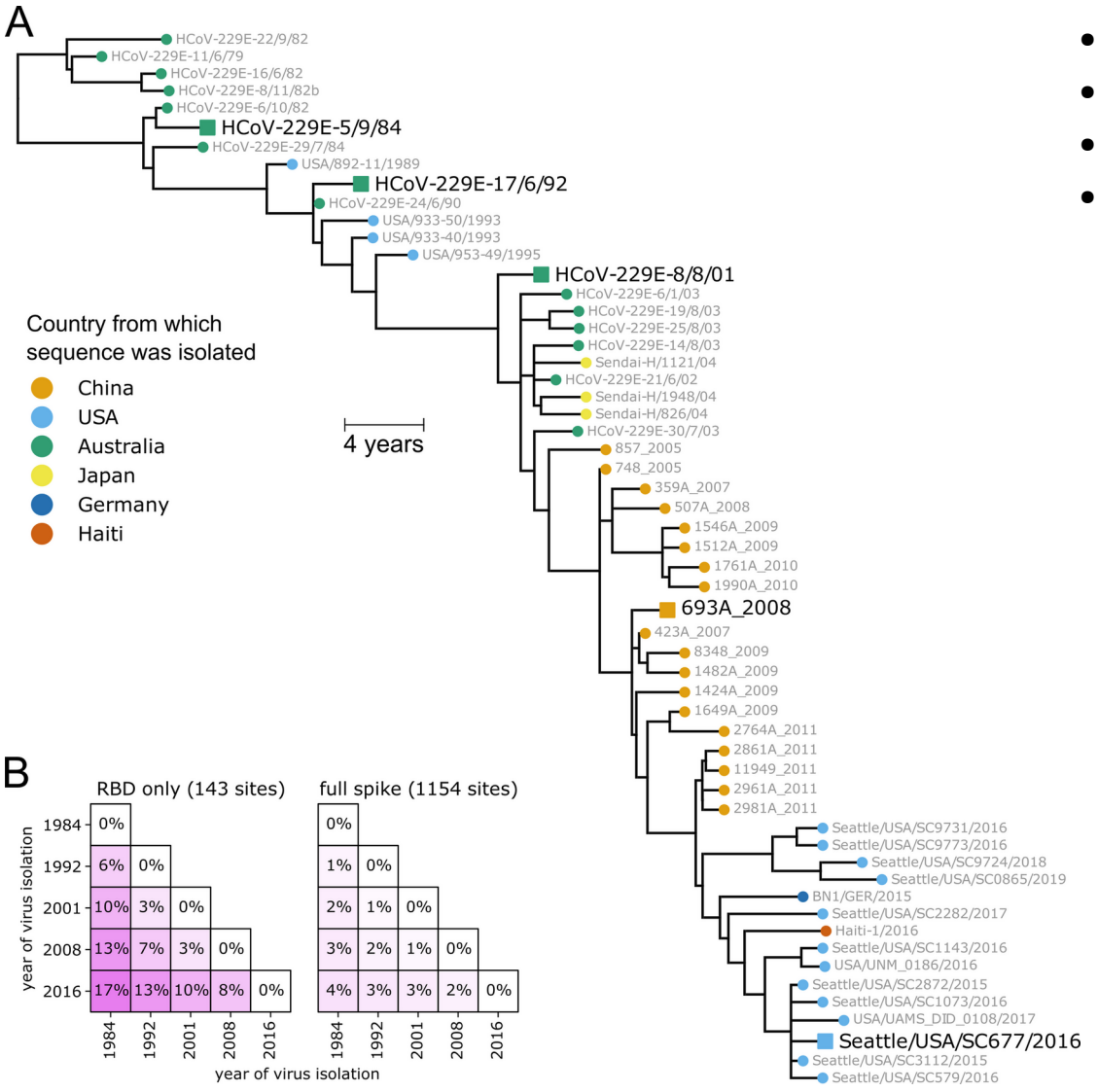
Selection ~ growth rate

S1 is quickly evolving and highly correlated with clade growth

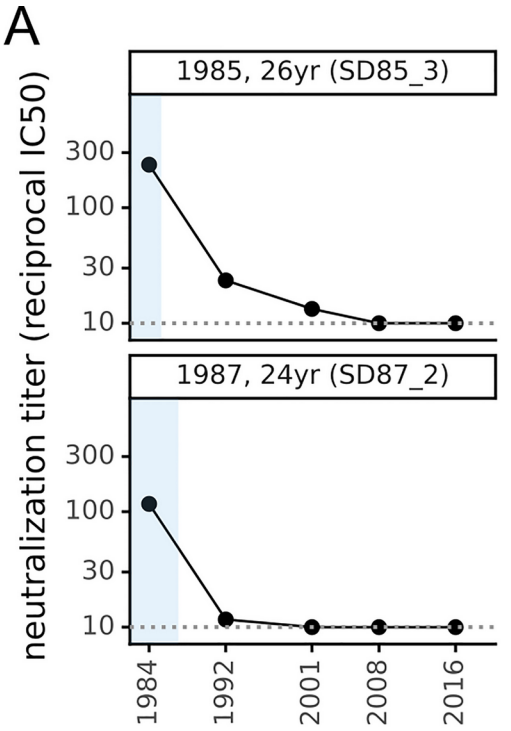


Kistler et al. 2021. bioRx

Corona 229E

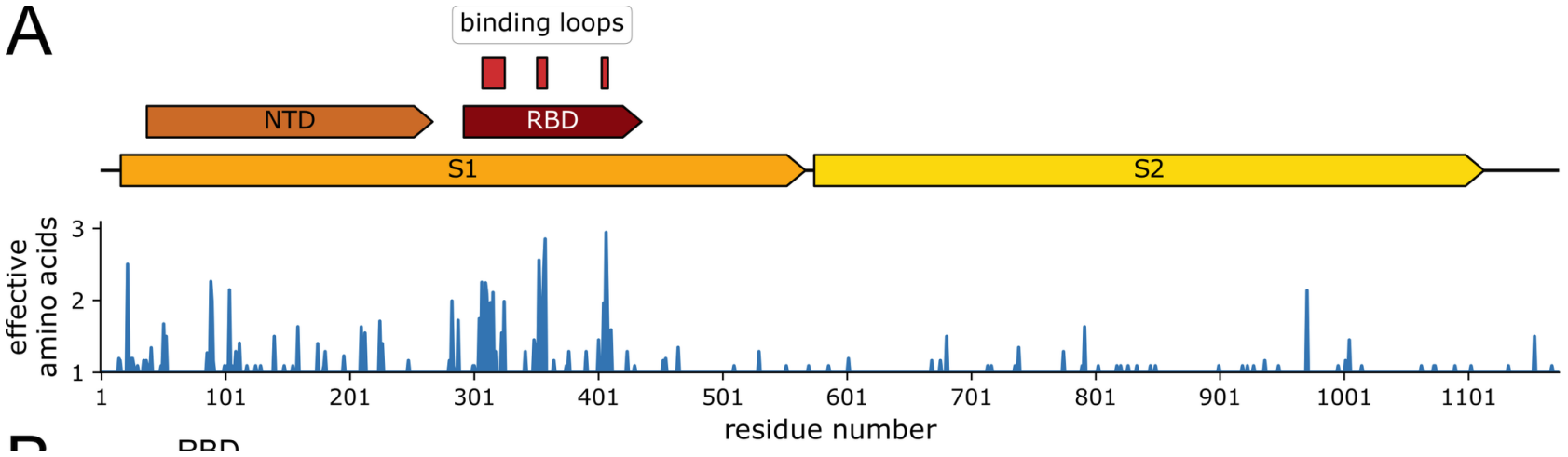


- Alphacoronavirus
- Infects humans and bats
- Causes common cold
- evolves via immunological escape



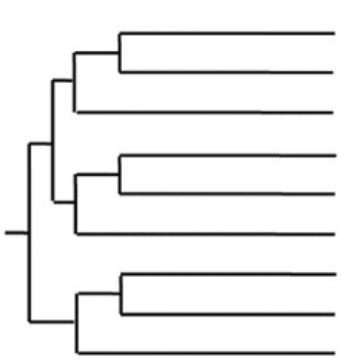
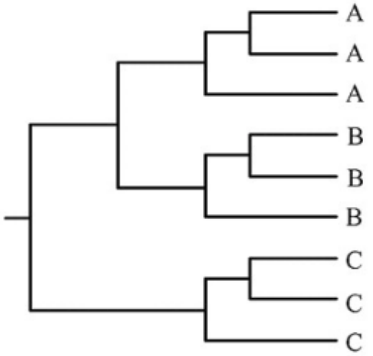
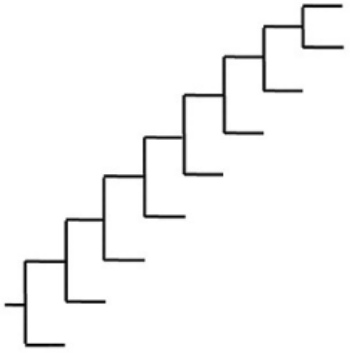
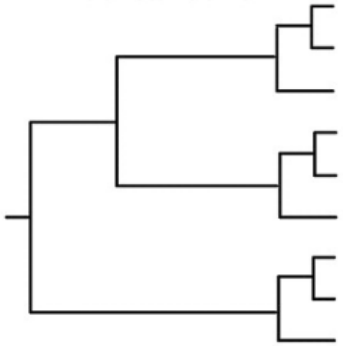
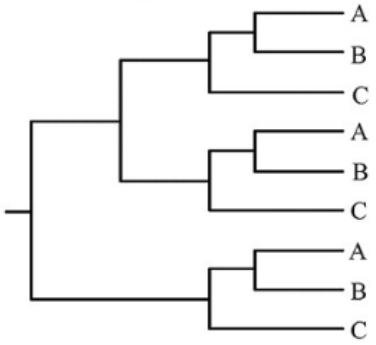
A human coronavirus evolves antigenically to escape antibody immunity
Rachel T. Eguia,Katharine H. D. Crawford,Terry Stevens-Ayers,Laurel Kelnhofer-Millevolte,Alexander L. Greninger,Janet A. Englund,Michael J. Boeckh,Jesse D. Bloom

Corona 229E



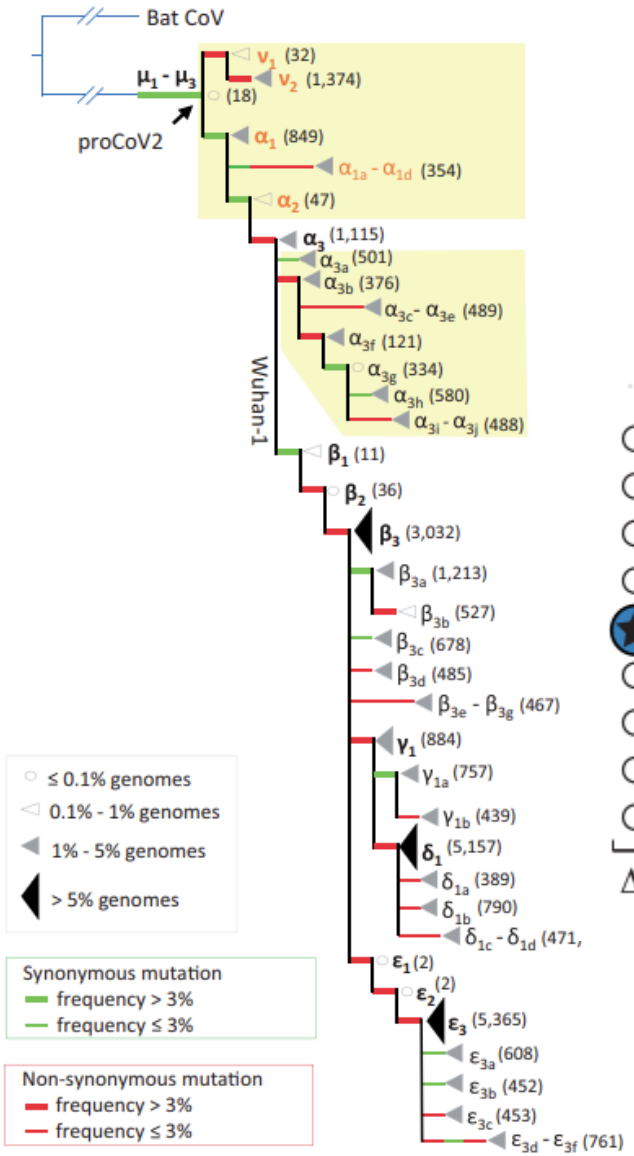
A human coronavirus evolves antigenically to escape antibody immunity
Rachel T. Eguia,Katharine H. D. Crawford,Terry Stevens-Ayers,Laurel Kelnhofer-Millevolte,Alexander L. Greninger,Janet A. Englund,Michael J. Boeckh,Jesse D. Bloom

Phylodynamics of Pathogens

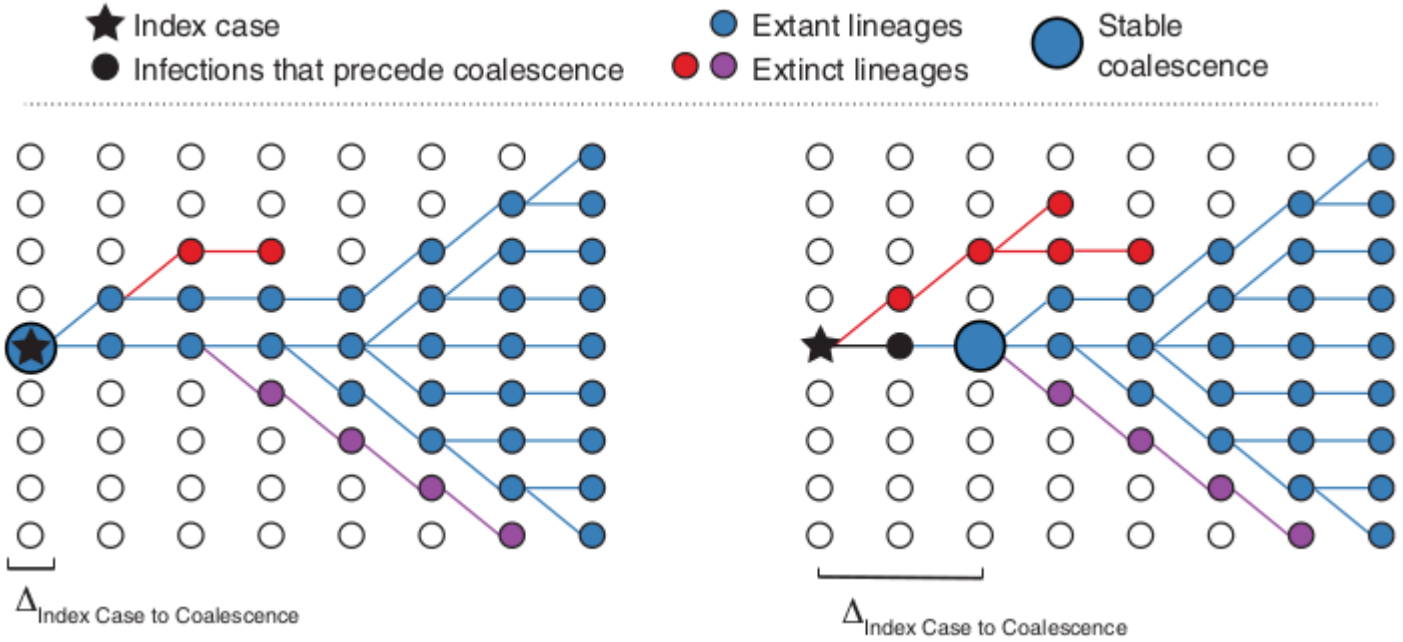
	Continual Immune Selection	Weak or Absent Immune Selection	
		Tree shape controlled by non-selective population dynamic processes	
Idealized Phylogeny Shapes		Population size dynamics	Spatial dynamics
		Exponential growth 	**Strong spatial structure** 
	 **Time →**	**Constant size** 	**Weak spatial structure** 
Examples	Human influenza A virus intra-host HIV	inter-host HIV inter-host HCV	Measles, rabies inter-host HIV
Tree Inferences	Detection of antigenic escape mutations	Estimation of population growth rates	Estimation of population migration rates

Grenfell et al. (2004)

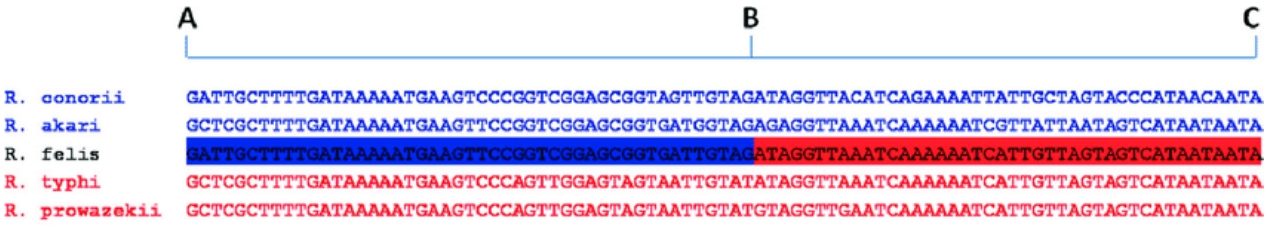
SARS-CoV-2 origin

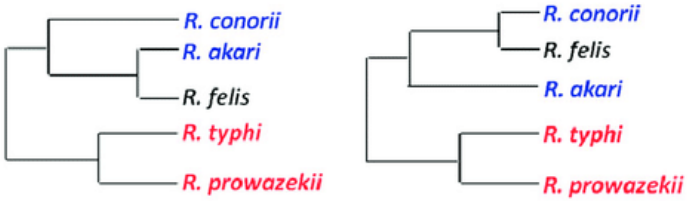


Infer ancestor (progenitor) genome
Earliest cases (Wuhan-1) are not index
Estimate of index case at late October 2019

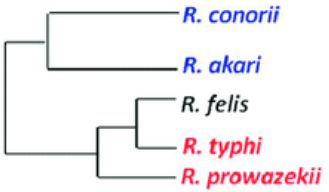


Recombination: Achille's heel

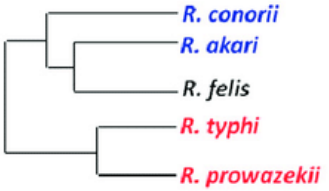




Alternative trees from fragment A-B



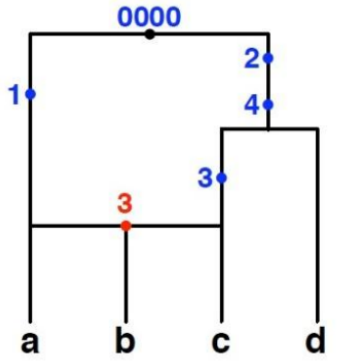
Tree from fragment B-C



Tree from fragment A-C

- Phylogenetics assumes no recombination among lineages, ie there is one tree
- Recombination causes homoplasies
- Networks can handle recombination

	c_1	c_2	c_3	c_4
a	1	0	0	0
b	1	0	1	1
c	0	1	1	1
d	0	1	0	1



Phylogenetic network: a directed acyclic graph (DAG)

Merhej, Vicky, et al. "The rhizome of life: the sympatric Rickettsia felis paradigm demonstrates the random transfer of DNA sequences." Molecular biology and evolution 28.11 (2011): 3213-3223.

Exercises

1) When does the substitution rate depend on population size?
2) dN/dS = 0.34 for a gene. Is this most consistent with positive or negative selection? If 66% of nonsynonymous sites are under strong purifying selection, what is dN/dS for the remaining sites?
3) What can cause dN/dS > 1?
4) Positive selection is acting on a gene, but dN/dS = 0.34 for the gene over the phylogenetic tree. What are three ways dN/dS could be used to detect selection?
5) Why is mutation not a good explanation for dN != dS?
6) Why are codons with multiple substitutions on a single branch not included when detecting selection?

Exercises

1) Why is dN/dS high on trunk (influenza) compared to leaves (HIV) of a tree? Why is influenza evolution predictable?

2) Why does recombination cause problems in phylogenetics?