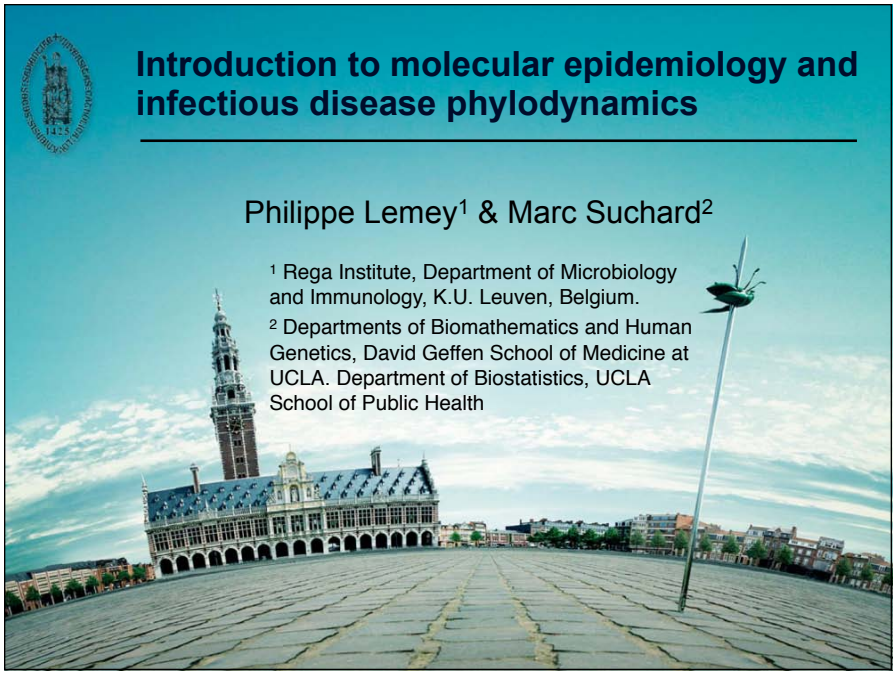


Introduction to molecular epidemiology and infectious disease phylodynamics

Philippe Lemey[1] & Marc Suchard[2]

[1] Rega Institute, Department of Microbiology and Immunology, K.U. Leuven, Belgium.

[2] Departments of Biomathematics and Human Genetics, David Geffen School of Medicine at UCLA. Department of Biostatistics, UCLA School of Public Health

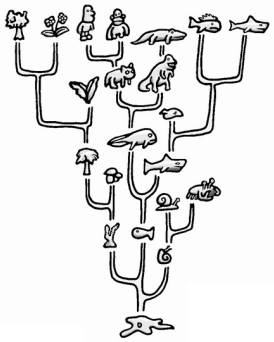
This course (SISMID module 2)

- Monday, July 9
  - ➡ Introduction
  - ➡ Alignment, substitution models and phylogenetic inference
- Tuesday, July 10
  - ➡ Phylogenetic inference practical
  - ➡ Bayesian phylogenetics
  - ➡ Molecular clocks: rates and dates
  - ➡ Introduction to BEAST
- Wednesday, July 11
  - ➡ Viral epidemiology and the coalescent
  - ➡ BEAST practical
  - ➡ Phylogeography
  - ➡ BEAST practical
- Bonus
  - ➡ Phylo-Alignment
  - ➡ Recombination
  - ➡ Robust Counting

http://perswww.kuleuven.be/philippe_Lemey/SISMID/
http://www.phylogeography.org

Molecular evolution and phylogenetics

- biological sequences (DNA, RNA, protein) contain information about the processes and events that formed them
- this information is often scrambled, fragmentary, hidden, or lost completely
- our aim is to use mathematical models to recover and decipher this information
- The central concept is a phylogeny: a diagram depicting the ancestral relationships among characters or genetic sequences



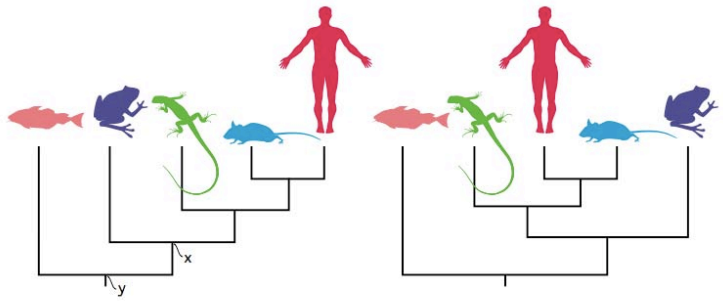
```
HIV-1 (UK)   ATC---TGCTAAAGCATATGACACAGAGGTACATAATGTTT
HIV-1 (USA)  ATCGGATGCTAGAGCTTATGATACAGAGGTACA---TGTTT
```

Phylogenetics

EVOLUTION

The Tree-Thinking Challenge

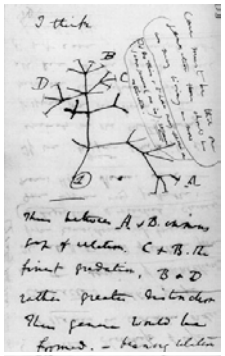
David A. Baum, Stacey DeWitt Smith, Samuel S. S. Donovan



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.

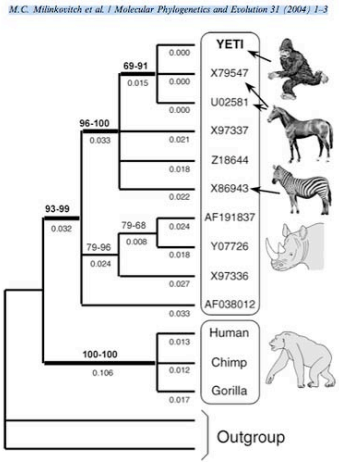
Phylogenetics

- Darwin, 1837



- Haeckel, 1866

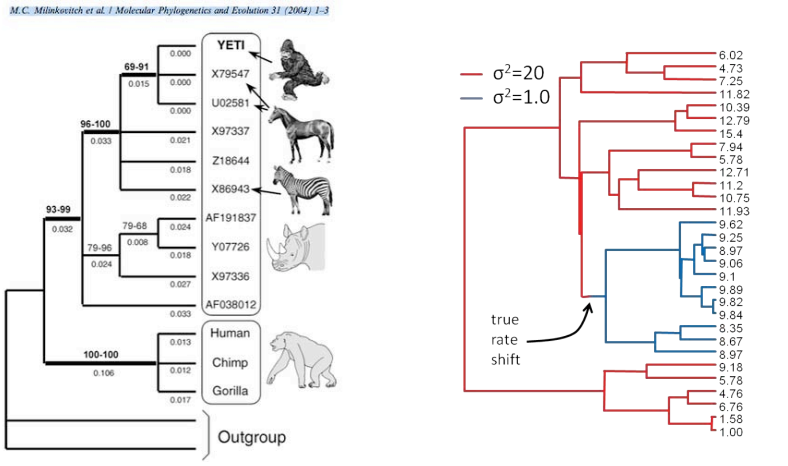
Phylogenetics



...Very little is known about the morphology of this enigmatic creature. The famous writer Peter Matthiessen notes that "The yeti is described most often as a hairy, reddish-brown creature with a rigid crown that gives it a pointed-head appearance; in size, despite the outsized foot 1/2. . . it has been likened to an adolescent boy, though much larger individuals have been reported" (Matthiessen, 1979, p. 119). This is perfectly consistent with the description given earlier by Haddock: "A sort of enormous monkey . . . with a huge head like a coconut" (Herge, 1960, p. 37)

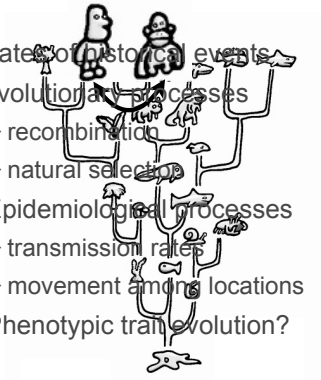
"All our analyses clearly indicate that the yeti is nested several nodes within a specific ungulate group (i.e., the perissodactyls, cf. Fig. 1) and, more specifically, forms a subclade with sequences U02581 and X79547 (cfr. figure legend). These results demonstrate that extensive morphological convergences have occurred between the yeti and primates."

Phylogenetics



Information in (viral) molecular sequences

- Genetic distances among strains
- Phylogeny
  - ➡ subtyping/classification
  - ➡ identification of transmission clusters
  - ➡ association with risk factors
  - ➡ forensics
- Dates of historical events
- Evolutionary processes
  - ➡ recombination
  - ➡ natural selection
- Epidemiological processes
  - ➡ transmission rates
  - ➡ movement among locations
- Phenotypic trait evolution?



```
HIV-1 (UK)  ATC---TGCTAAAGCATATGACACAGAGGTACATAATGTTT
HIV-1 (USA) ATCGGATGCTAGAGCTTATGATACAGAGGTACA---TGTTT
```

Our goal

MOLECULAR SEQUENCES

Alignment Methods — BIOINFORMATICS

ALIGNMENT

Sequence Evolution Models
Phylogenetic Methods — PHYLOGENETICS

EVOLUTIONARY TREE
(time scale = genetic distance)

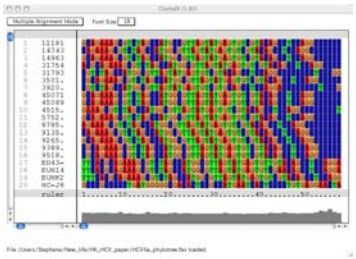
Molecular Clock Models — PHYLOGENETICS

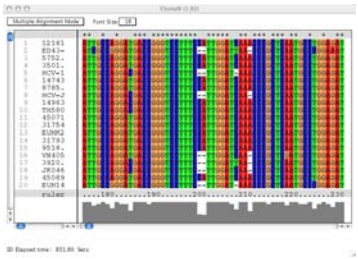
EVOLUTIONARY TREE
(time scale = years)

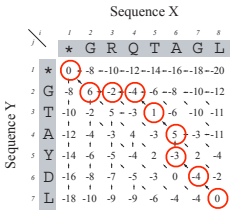
Phylodynamic Models — POPULATION GENETICS

EPIDEMIOLOGY

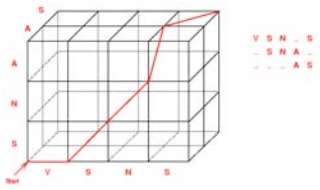
Sequence alignment



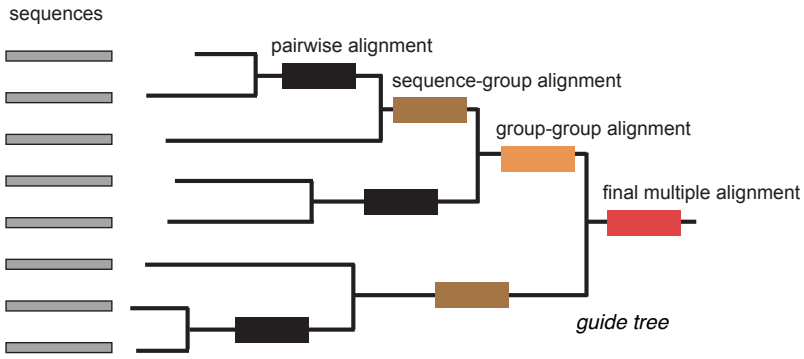




GRQTA-GL
G--TAYDL



Progressive alignment



http://www.kuleuven.be/aidslab/phylogenybook/Table3.1.html

Genetic distances

```
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 TACTAATTG 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
```

chimpanzee SIV vs HIV-1 envelope gene

Not all mutations are equally likely

- some point substitutions are more likely to occur than others:

  transitions are more likely than transversions

  - *transitions*:

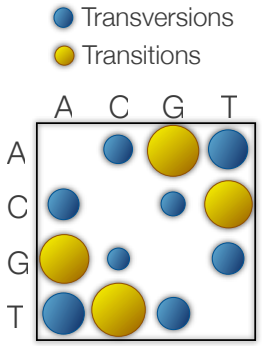
    purine↔purine or pyrimidine↔pyrimidine

    A↔G C↔T

  - *transversions*:

    purine↔pyrimidine

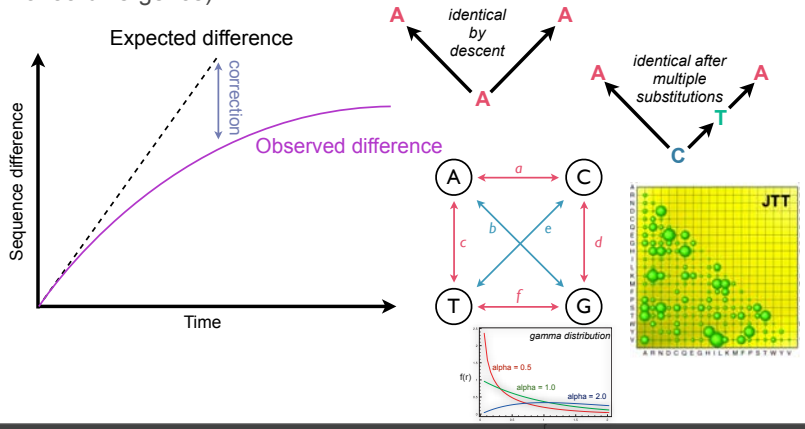
    A↔C A↔T
    G↔C G↔T



Unambiguous changes on most parsimonious tree of Ciliate SSUrDNA

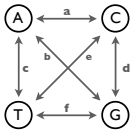
Substitution models

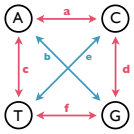
- During evolution, 'multiple hits' can occur at a single position: the evolutionary distance is almost always larger than the dissimilarity (% nt or aa divergence)

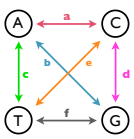


Nucleotide substitution models

Simplest (few parameters)







Most complex (many parameters)

1. Base frequencies are equal and all substitutions are equally likely (Jukes-Cantor) *(a=b=c=d=e=f)*

2. Base frequencies are equal but transitions and transversions occur at different rates (Kimura 2-parameter) *(a=c=d=f, b=e)*

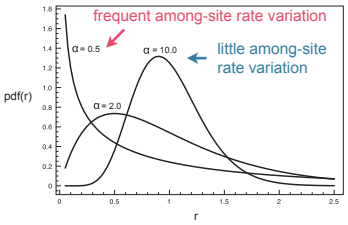
3. Unequal base frequencies and transitions and transversions occur at different rates (Hasegawa-Kishino-Yano) *(a=c=d=f, b=e)*

4. Unequal base frequencies and all substitution types occur at different rates (General Reversible Model) *(a, b, c, d, e, f)*

Does this matter?

Estimated genetic distances between SIVcpz and HIVlai, under different substitution models:

Observed % mismatches	= 0.406
JC (Jukes-Cantor)	= 0.586
HKY (Hasegawa-Kishino-Yano)	= 0.611
GTR (General Time Reversible)	= 0.620
GTR + gamma	= 1.017



Gene	α
Prolactin	1.37
Albumin	1.05
C-myc	0.47
Ctyochrome β (mtDNA)	0.44
Insulin	0.40
D-loop (mtDNA)	0.17
12S rRNA (mtDNA)	0.16

Phylogenetic reconstruction

MOLECULAR SEQUENCES

Alignment Methods — BIOINFORMATICS

ALIGNMENT

Sequence Evolution Models
Phylogenetic Methods — PHYLOGENETICS

EVOLUTIONARY TREE
(time scale = genetic distance)

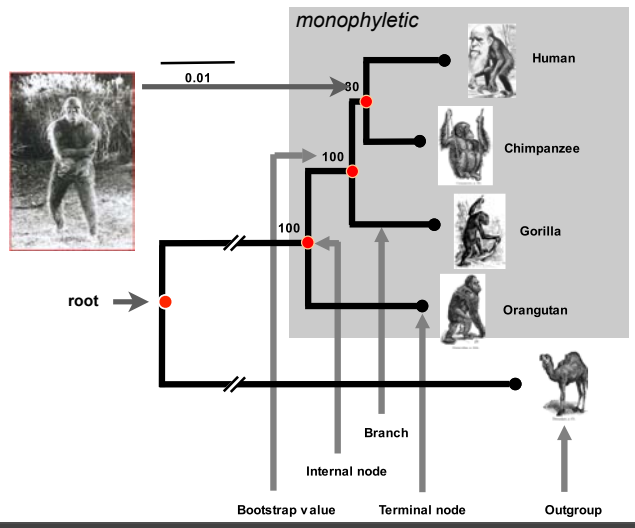
Molecular Clock Models — PHYLOGENETICS

EVOLUTIONARY TREE
(time scale = years)

Phylodynamic Models — POPULATION GENETICS

EPIDEMIOLOGY

What is a tree?



Terminology

Phylogeny

Genealogy

Phylogenetic reconstruction

- **CLUSTERING APPROACHES:** These begin with a genetic distance between each pair of sequences. A 'clustering algorithm' then transforms the genetic distances into a tree.
  - e.g. UPGMA, Neighbour-Joining
  - Simple, faster.
  - No measure of how good the estimated tree is (non-statistical)
- **OPTIMALITY METHODS**: These define a score for each possible tree. 'Search algorithms' are then used to find the tree with the highest score.
  - e.g. Parsimony, Maximum Likelihood, Bayesian Inference
  - More complex, slower. Search may not locate the 'best' tree.
  - Quality of each tree can be directly compared (statistical)

Phylogenetic reconstruction

- For n taxa, there are:
  $(2n-3)!/[(2^{n-2})*(n-2)!]$
  rooted, binary trees

# taxa	# trees	
4	15	enumerable by hand
5	105	enumerable by hand on a rainy day
6	945	enumerable by computer
7	10395	still searchable very quickly on computer
8	135135	a bit more than the number of hairs on your head
9	2027025	population of Glasgow
10	34459425	≈ upper limit for exhaustive searching; about the number of possible combinations of numbers in the National Lottery
20	8.20×10^{21}	≈ upper limit for branch-and-bound searching
48	3.21×10^{70}	≈ the number of particles in the universe
136	2.11×10^{267}	=number of trees to choose from in the "Out of Africa" data (Vigilant et al., 1991)

Phylogenetic inference

Books:

- Felsenstein J. (2003). *Inferring phylogenies.* Sinauer Associates
- Yang Z. (2003). *Computational Molecular Evolution*. Oxford University Press
- Nei M & Kumar S. (2000). *Molecular Evolution and Phylogenetics.* Oxford University Press.
- Graur D. (2000). *Fundamentals of Molecular Evolution.*
- 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.

Computer Software: http://evolution.genetics.washington.edu/phylip/software.html

- PAUP* (Phylogenetic Analysis Using Parsimony *and other methods)
  - http://paup.csit.fsu.edu/
- 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/

Information in viral molecular sequences

- Genetic distances a strains
- Phylogeny
  - ➡ subtyping/classification
  - ➡ identification of transmission clusters
  - ➡ association with risk factors
  - ➡ forensics

- Evolutionary processes
  - ➡ recombination
  - ➡ natural selection
- Dates of historical events
- Epidemiological processes
  - transmission rates
  - movement among locations
- Phenotypic trait evolution?

```
HIV-1 (UK)   ATC---TGCTAAAGCATATGACACAGAGGTACATAATGTTT
HIV-1 (USA)  ATCGGATGCTAGAGCTTATGATACAGAGGTACA---TGTTT
```

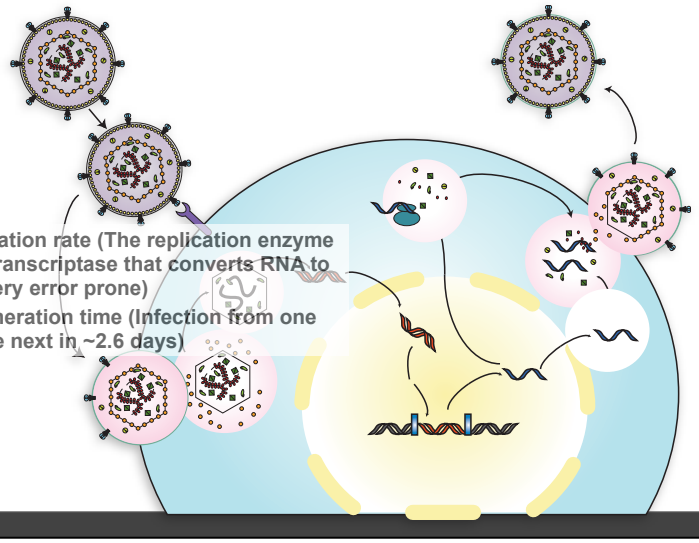
Comparative rates of evolution

"...the exceptionally high nucleotide mutation rate of a typical RNA virus — a million times greater than that of vertebrates — allows these viruses to generate mutations and adaptations de novo during environmental change.."

- Bacteria & large DNA viruses:
  ~0.003 mutations per genome, per replication
  ~10^{-7} to 10^{-9} substitutions/site/year
- Eukaryotes:
  ~0.01 mutations per genome, per replication
  ~10^{-9} to 10^{-11} substitutions/site/year
- RNA viruses (and ssDNA viruses):
  ~1 mutation per genome, per replication*
  ~10^{-3} to 10^{-4} substitutions/site/year

**Because RNA polymerase (RNA template) and reverse transcriptase (DNA template) have no proof-reading ability*

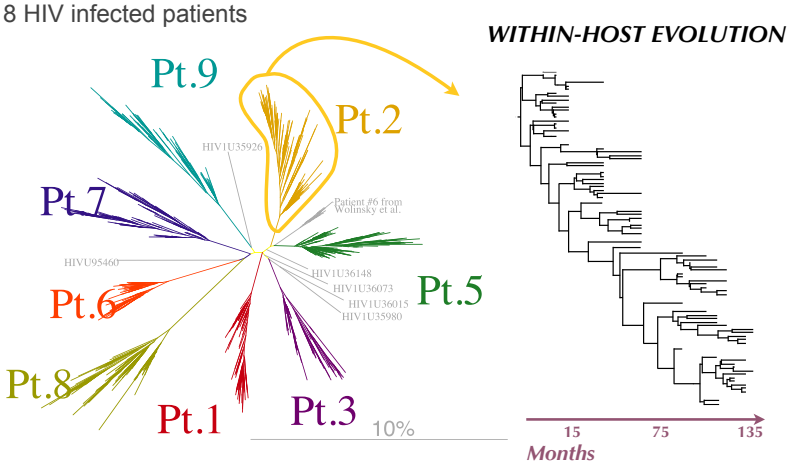
The Human Immunodeficiency virus



- High mutation rate (The replication enzyme reverse transcriptase that converts RNA to DNA is very error prone)
- Short generation time (Infection from one cell to the next in ~2.6 days)

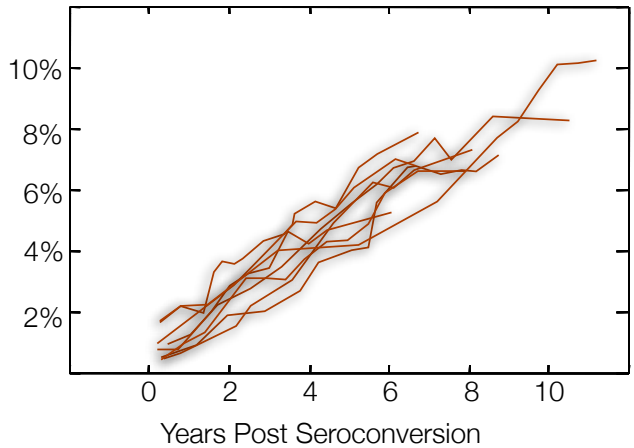
Measurably evolving populations

8 HIV infected patients

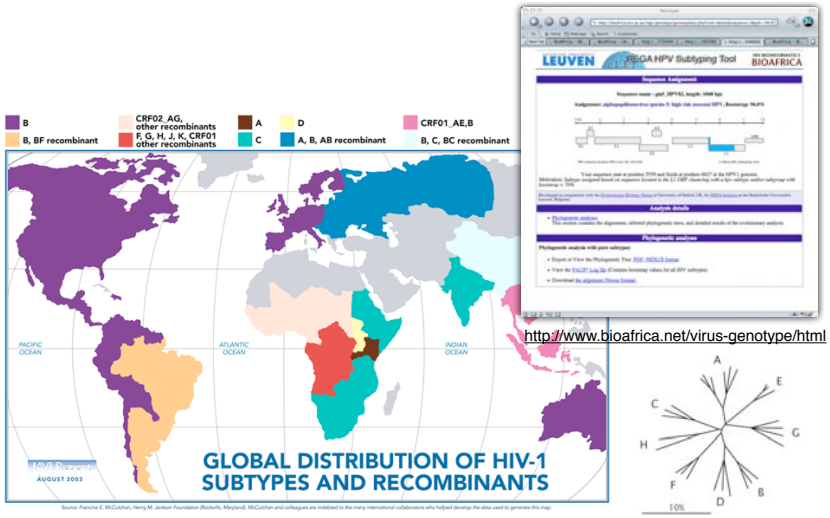


Measurably evolving populations

8 HIV infected patients

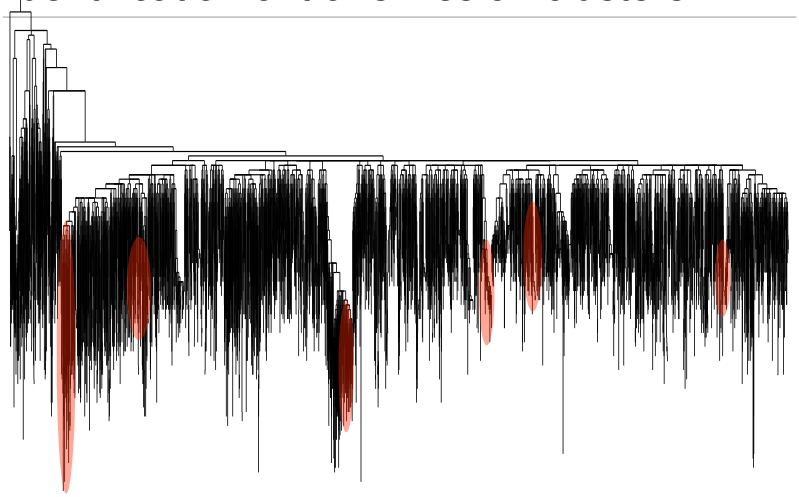


Subtyping/classification



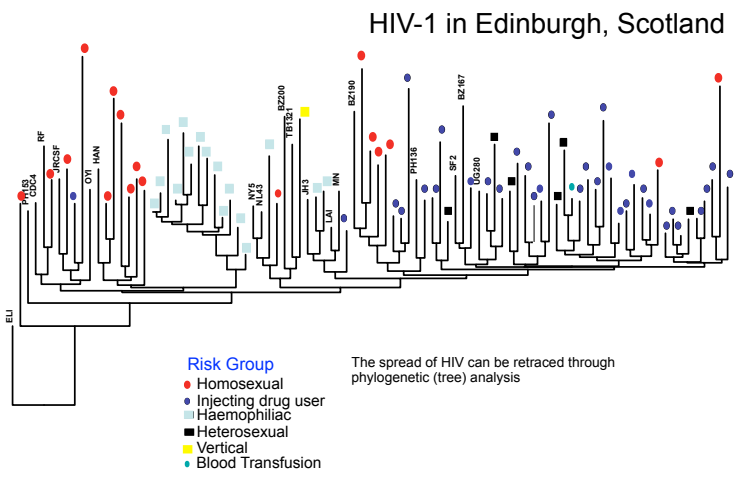
http://www.bioafrica.net/virus-genotype/html

Identification of transmission clusters

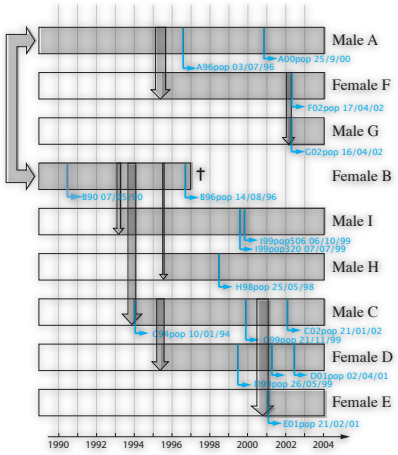


Hué et al. (2005) Proc. Natl. Acad. Sci. USA 102:4425-4429

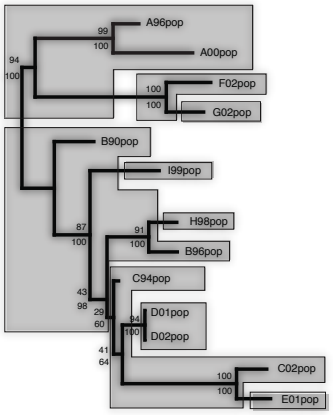
Association with risk factors



Reconstruction transmission chains



env gp41



Lemey et al., J. Virol (2005), 79(1): 11981-11989

Forensics

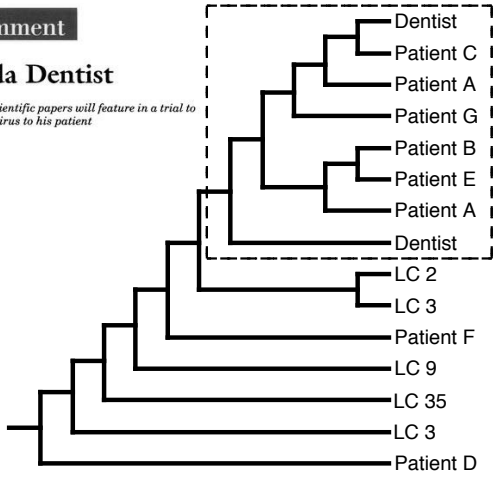
News & Comment

The Case of the Florida Dentist

Cutting-edge molecular biology and unpublished scientific papers will feature in a trial to determine whether a dentist tansmitted the AIDS virus to his patient

- Patients A, B, C, E, G - infected by the dentist
- Patients D, F - not infected by the dentist

LC - local controls (other HIV patients from Florida)



Dates of historical events

MOLECULAR SEQUENCES

Alignment Methods — BIOINFORMATICS

ALIGNMENT

Sequence Evolution Models
Phylogenetic Methods — PHYLOGENETICS

EVOLUTIONARY TREE
(time scale = genetic distance)

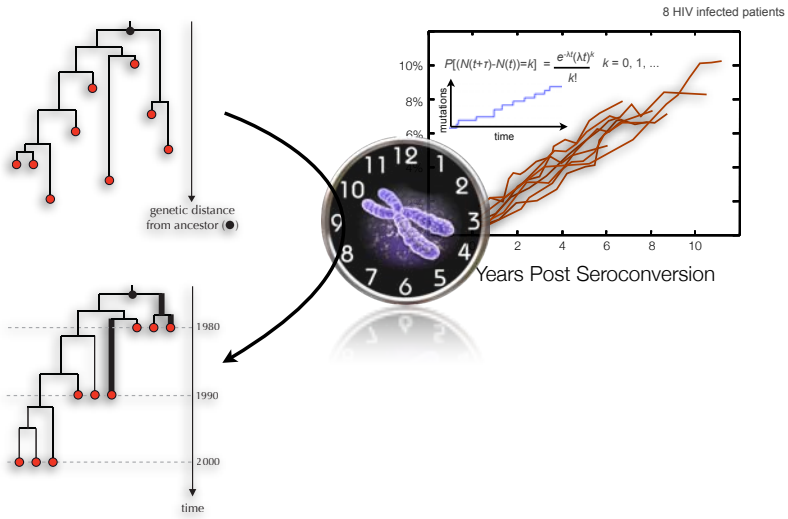
Molecular Clock Models — PHYLOGENETICS

EVOLUTIONARY TREE
(time scale = years)

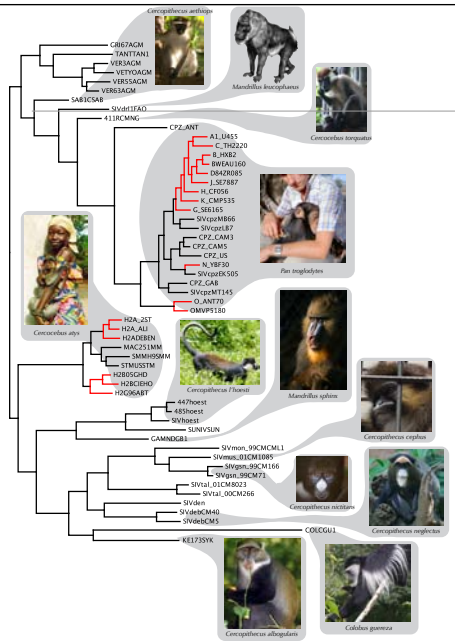
Phylodynamic Models — POPULATION GENETICS

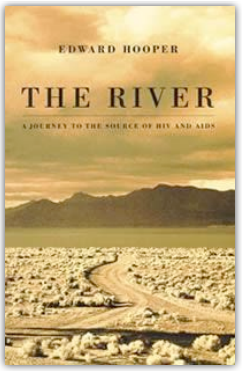
EPIDEMIOLOGY

Molecular clocks

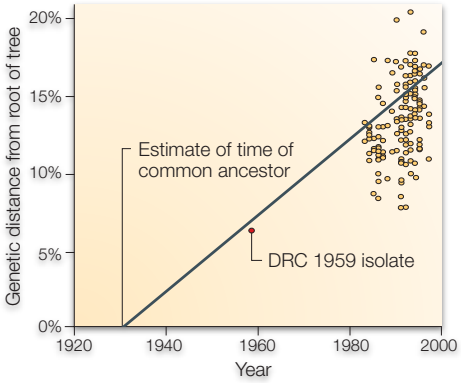


The origin of HIV-1





The origin of HIV-1



Korber et al. 2000. Science 288 1789-96

Molecular clock analysis Indicates that the most recent common ancestor of HIV-1 group M existed in the 1930s (± ~10 years).

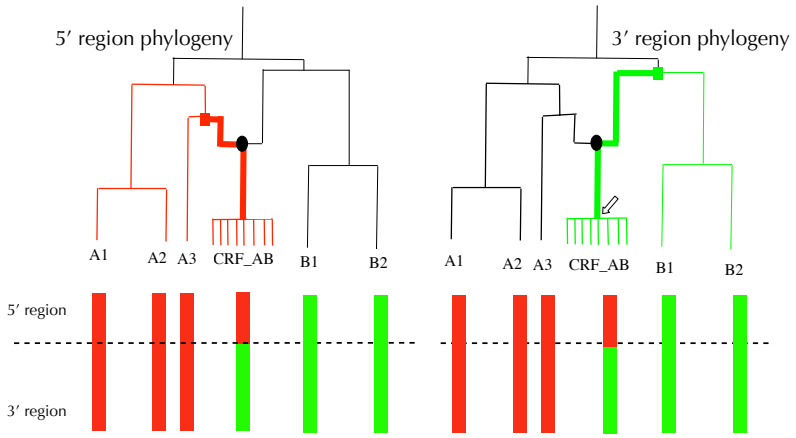
Although this estimate has been criticized on statistical grounds, it does correctly date the the oldest HIV-1 isolate known (sampled in 1959 from a man near Kinshasa, DRC).

Information in viral molecular sequences

- Genetic distances among strains
- Phylogeny
  - ➡ subtyping/classification
  - ➡ identification of transmission clusters
  - ➡ association with risk factors
  - ➡ forensics
- Dates of historical events
- Evolutionary processes
  - ➡ recombination
  - ➡ natural selection
- Epidemiological processes
  - ➡ transmission rates
  - ➡ movement among locations
- Phenotypic trait evolution?

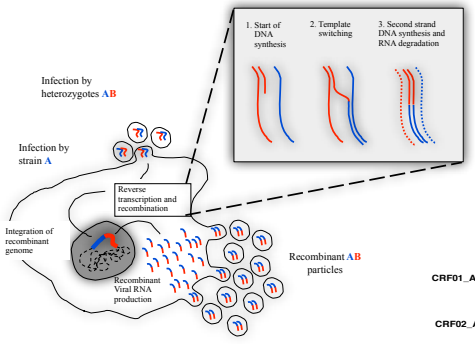
```
HIV-1 (UK)   ATC---TGCTAAAGCATATGACACAGAGGTACATAATGTTT
HIV-1 (USA)  ATCGGATGCTAGAGCTTATGATACAGAGGTACA---TGTTT
```

Evolutionary processes: recombination

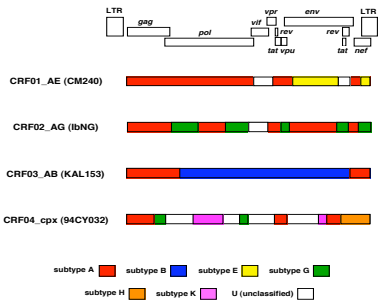


Recombination = genome regions with significantly different phylogenies

Evolutionary processes: recombination

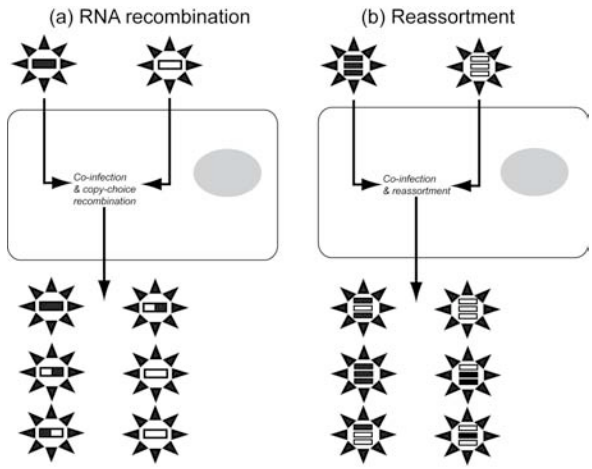


Importantly, recombinants that have the same recombinant form (designated Circulating Recombinant Forms - CRFs) have been found in many individuals.



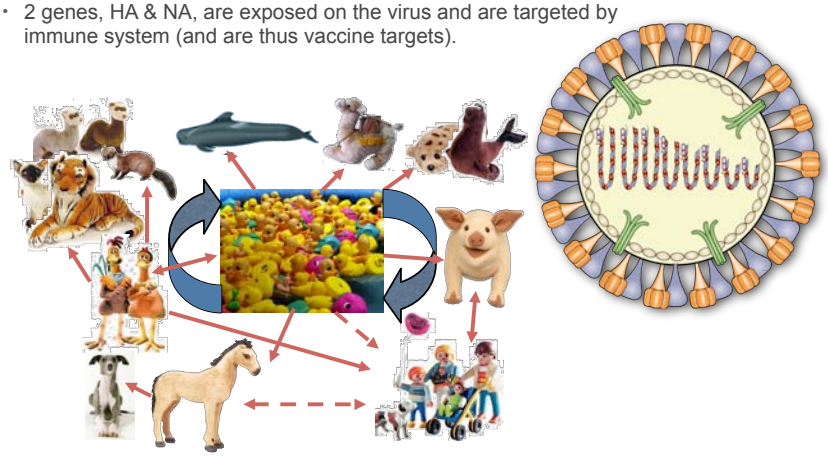
Recombination has been found to have occurred frequently between divergent viruses by the finding of discordant phylogenetic relationships in trees inferred from different genomic recombinants.

Sex in RNA viruses

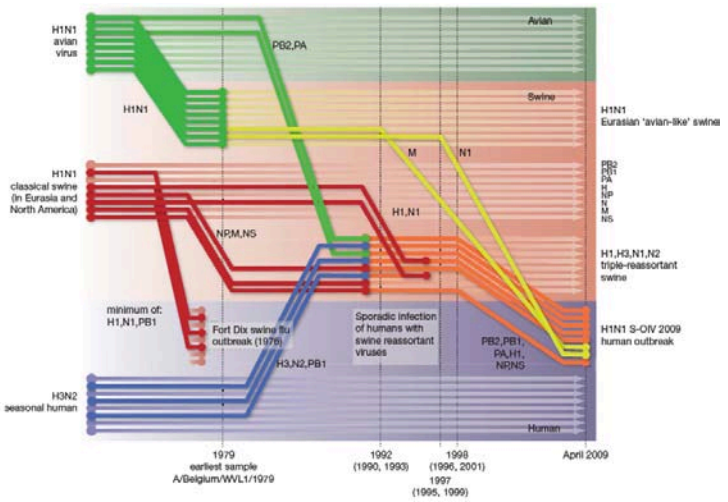


Influenza A Virus

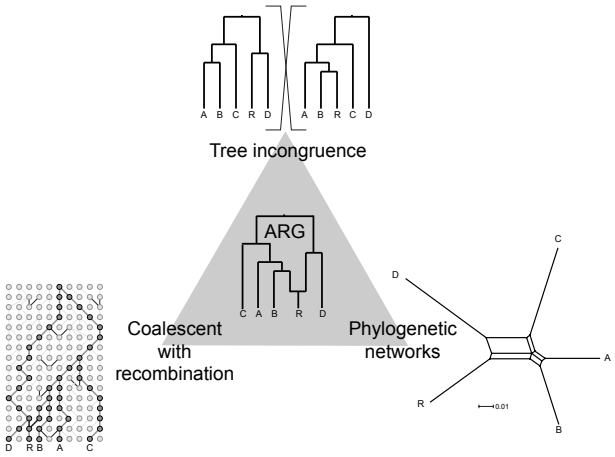
- segmented RNA virus, divided into 8 segments with 1 or 2 genes each
- 2 genes, HA & NA, are exposed on the virus and are targeted by immune system (and are thus vaccine targets).



The reassortment history of H1N1

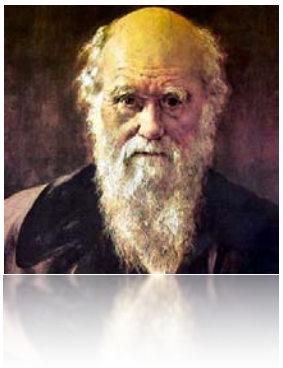


Analysis of recombination and reassortment



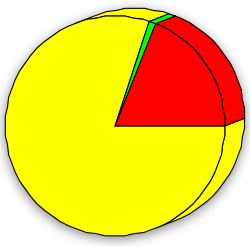
Evolutionary processes: natural selection

- "the preservation of favourable variations and the rejection of injurious variations, i call natural selection. variations neither useful nor injurious would not be affected by natural selection, and would be left a fluctuating element"
  - darwin, the origin of species



Evolutionary processes: natural selection

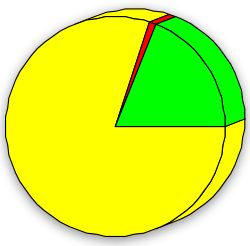
neutralist model
motoo kimura



most fixed mutations are neutral

$s>0$
$s\approx 0$
$s<0$

selectionist model
john gillespie

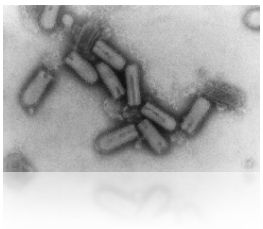


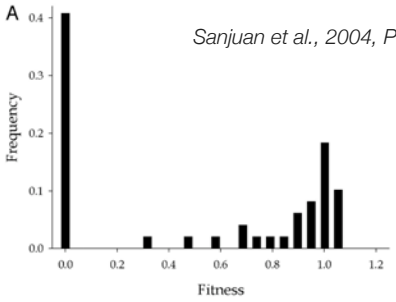
most fixed mutations are advantageous

Evolutionary processes: natural selection

	Random		Preobserved		Total	
	Proportion, %	Effect, %	Proportion, %	Effect, %	Proportion, %	Effect, %
Lethal	39.6 (19)	−100	11.6 (5)	−100	26.4 (24)	−100
Deleterious	29.2 (14)	−24.4	41.9 (18)	−16.4	35.2 (32)	−19.9
Neutral	27.1 (13)	−3.8	32.6 (14)	−0.9	29.7 (27)	−2.3
Beneficial	4.2 (2)	4.2	14.0 (6)	7.9	8.8 (8)	7.0
Total	100 (48)	−47.6	100 (43)	−17.7	100 (91)	−33.4

For each category, the mean fitness effect is shown.

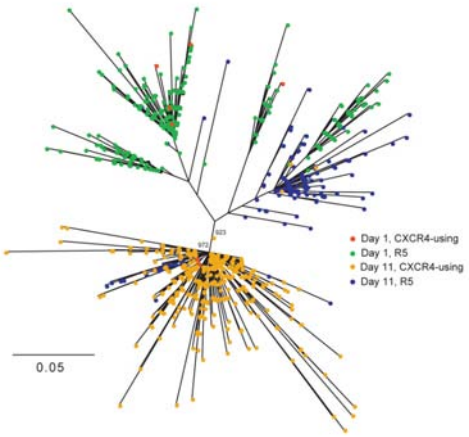




Sanjuan et al., 2004, PNAS

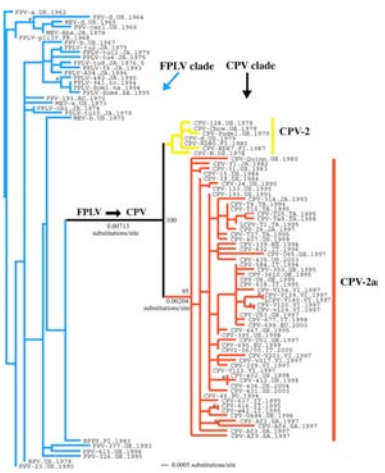
Evolutionary processes: natural selection

- Immune escape (antibodies*, T-cells*, innate immune responses)
- Antiviral drug resistance
- Vaccine escape mutations
- Cell & tissue tropism
- Inter-host viral transmission (i.e. for viral emergence)



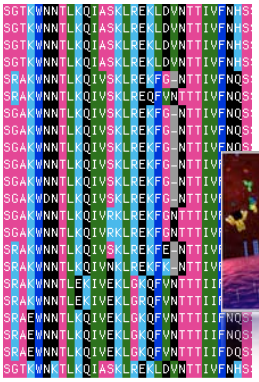
Evolutionary processes: natural selection

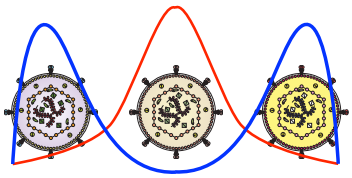
- Immune escape (antibodies*, T-cells*, innate immune responses)
- Antiviral drug resistance
- Vaccine escape mutations
- Cell & tissue tropism
- Inter-host viral transmission (i.e. for viral emergence)

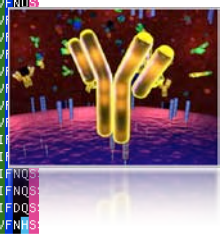


what exactly is d_N/d_S?

- diversifying selection

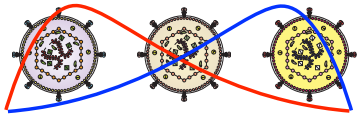




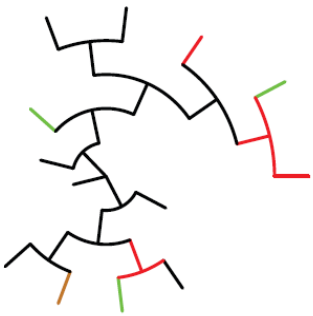


what exactly is d_N/d_S?

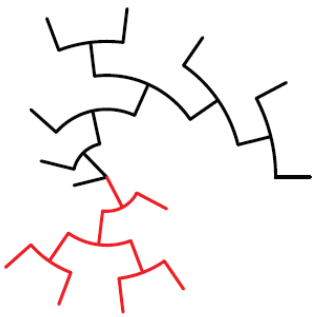
- directional selection



Diversifying and directional selection

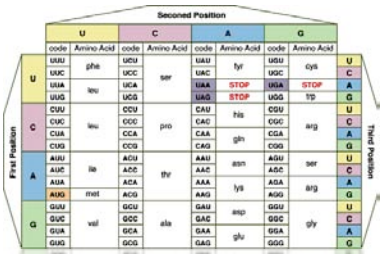


Diversifying/disruptive
Many non-synonymous substitutions, detected well by dN/dS analyses



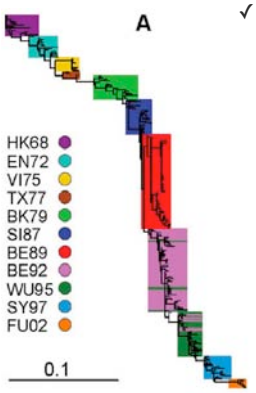
Directional/positive
Very few substitutions, but a significant change in allele frequencies. Confounds dN/dS methods, because there are very few substitutions and a frequency stationary model does not describe the biology well at all

Detecting molecular adaptation



- ✓ dN/dS: the relative rate of silent and replacement changes
  - codon substitution models
  - conservative (dN/dS < 1)

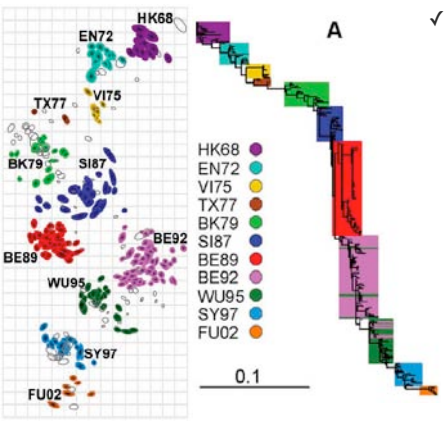
Detecting molecular adaptation



- ✓ dN/dS: the relative rate of silent and replacement changes
  - codon substitution models
  - conservative (dN/dS < 1)
  - site-specific, lineage specific

• Smith *et al.* (2004). *Science* **305**, 371-376.

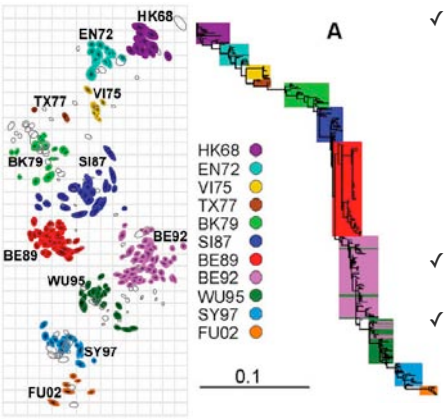
Detecting molecular adaptation



- ✓ dN/dS: the relative rate of silent and replacement changes
  - codon substitution models
  - conservative (dN/dS < 1)
  - site-specific, lineage specific
  - high dN/dS for recurrent rounds of positive selection (antagonist co-evolution)

• Smith *et al.* (2004). *Science* **305**, 371-376.

Detecting molecular adaptation



- ✓ dN/dS: the relative rate of silent and replacement changes
  - codon substitution models
  - conservative (dN/dS < 1)
  - site-specific, lineage specific
  - high dN/dS for recurrent rounds of positive selection (antagonist co-evolution)
- ✓ directional selection on protein sequences
- ✓ Genealogical statistics

• Smith *et al*. (2004). *Science* **305**, 371-376.

Epidemiological inference

MOLECULAR SEQUENCES

Alignment Methods — BIOINFORMATICS

ALIGNMENT

Sequence Evolution Models
Phylogenetic Methods — PHYLOGENETICS

EVOLUTIONARY TREE
(time scale = genetic distance)

Molecular Clock Models — PHYLOGENETICS

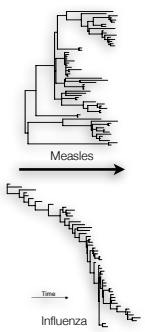
EVOLUTIONARY TREE
(time scale = years)

Phylodynamic Models — POPULATION GENETICS

EPIDEMIOLOGY

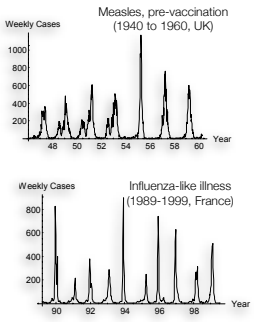
Phylodynamics™

GENETIC DIVERSITY
(phylogenetics &
molecular evolution)

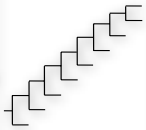
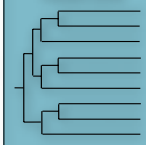
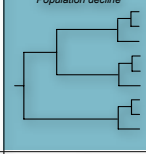
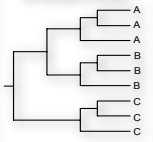
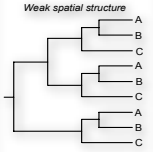


NATURAL SELECTION
(population genetics &
immunology)

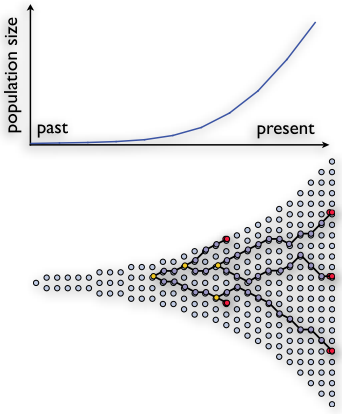
EPIDEMIC DYNAMICS
(mathematical epidemiology)



Phylodynamic Patterns

	Continual Immune Selection	Weak/No Immune Selection	
		Population dynamics	Spatial dynamics
Idealised Phylogeny Shapes		*Population growth*  *Population decline* 	*Strong spatial structure*  *Weak spatial structure* 
Examples	Human influenza A within-host HIV	among-host HIV among-host HCV	Measles Rabies, Dengue

Demography and coalescent theory



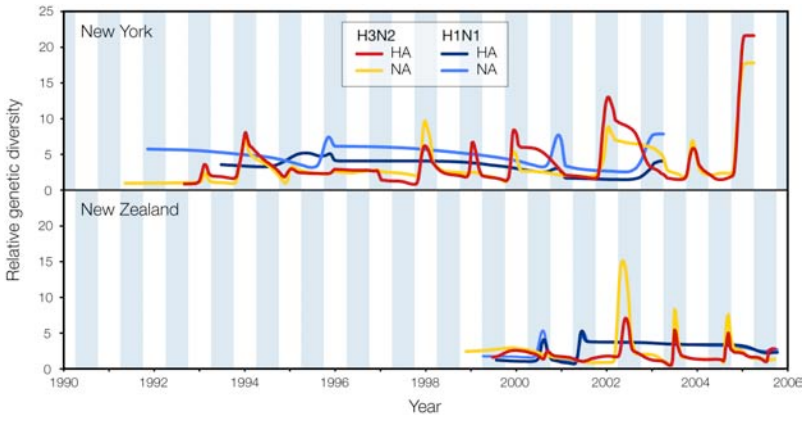
- The rate at which lineages 'coalesce' depends on population size and population structure.
- Population dynamics can be reconstructed using the 'skyline or skyride plot' method.

Pybus et al. (2000) Genetics 155:1429-37

Drummond, Rambaut, Shapiro & Pybus (2005) Mol Biol Evol 22:1185-92

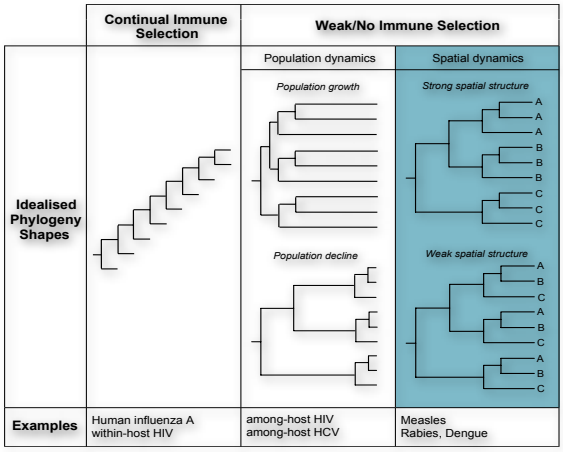
Minin, Bloomquist and Suchard (2008) Mol Biol Evol 25:1459-71

Influenza H3N2 epidemic dynamics

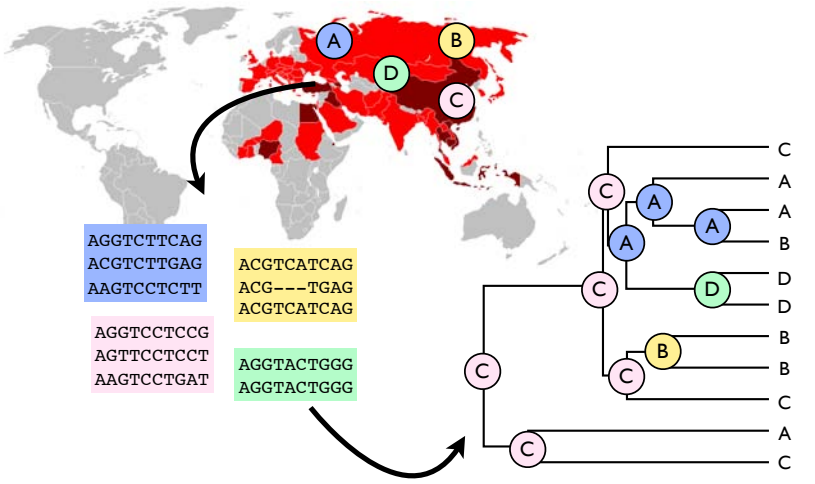


Rambaut et al. 2008. Nature

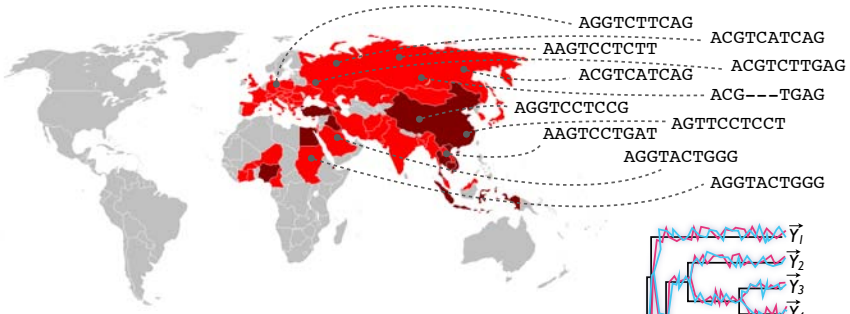
PhyloGEOdynamic Patterns



Phylogeography



Phylogeography



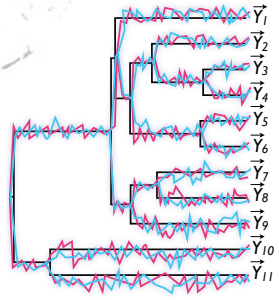
$\vec{Y} = (Y_{lat}, Y_{long})$

$\vec{Y}_{tip} = (\vec{Y}_1, \ldots, \vec{Y}_N)$

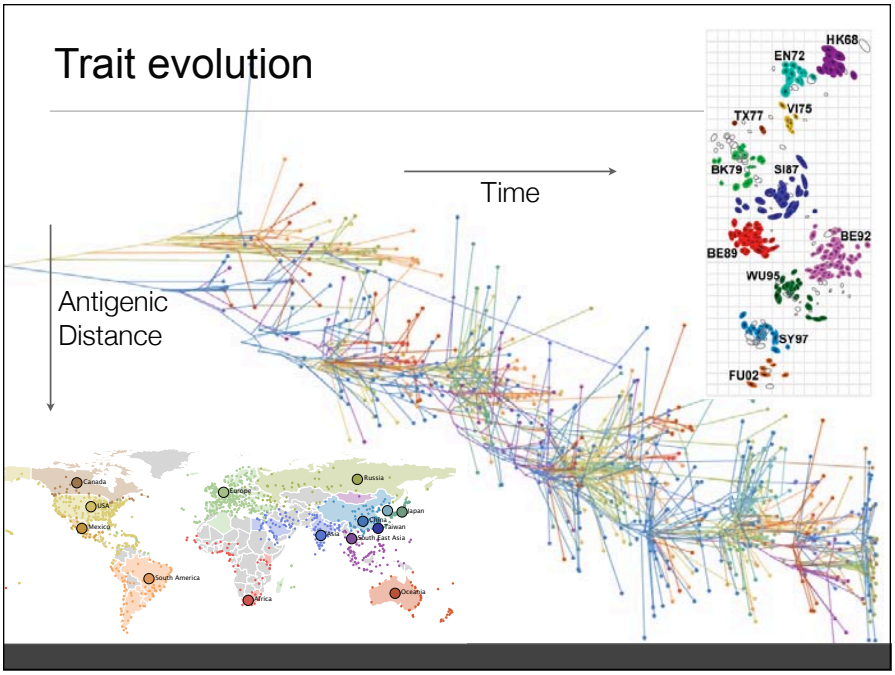
$\vec{Y}_{internal} = (\vec{Y}_{N+1}, \ldots, \vec{Y}_{2N-2})$

$\vec{Y}_{root}$

$\Delta\vec{Y} = BVN(0, \vec{St})$



Trait evolution



Bayesian Evolutionary Analysis Sampling Trees