Roadmap

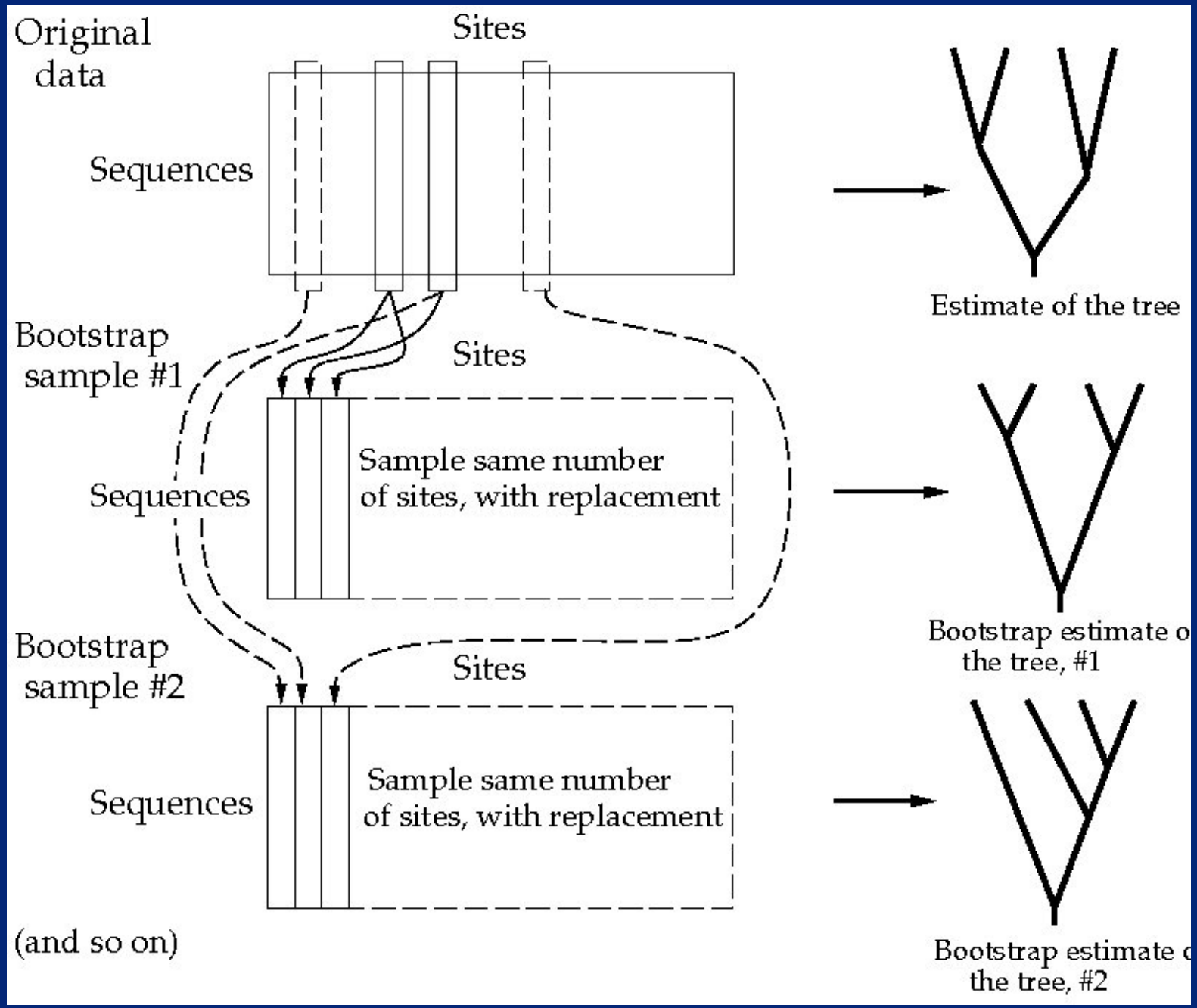
- Final exam schedule–reminder
- Teaching evaluations
- Bootstrap validation of phylogenies
- Coalescent inference within a population
  - General ideas
  - Case study–Benghazi virus case
  - (Another case study: red drum, presented in Lecture 2)

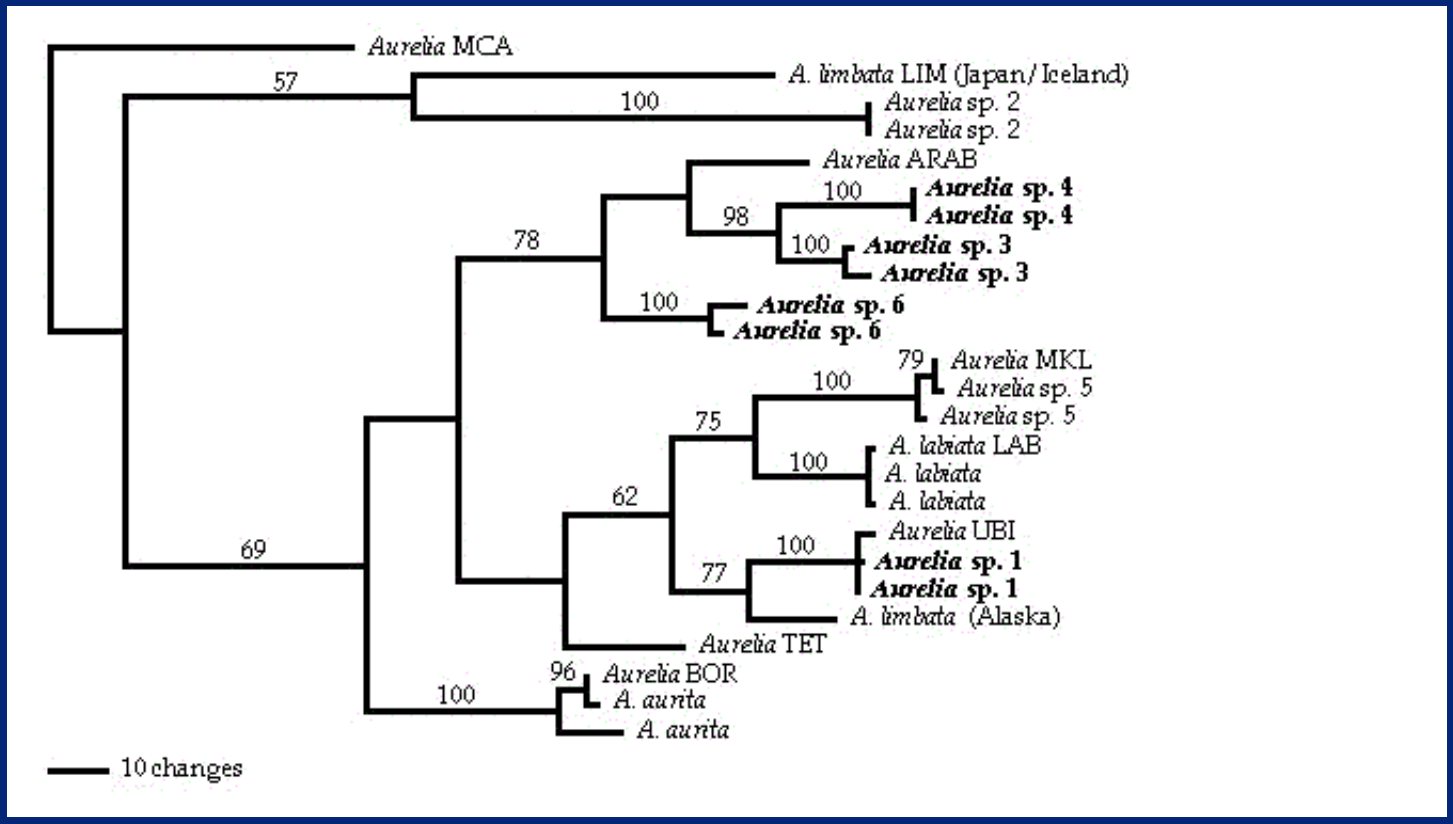
Final exam schedule

- You are welcome to take the final either:
  - Tuesday 3/19 at 10:30 am-12:30 pm
  - Wednesday 3/20 at 2:30 pm-4:30 pm (original schedule time)
- Both in S110, with any luck
- I ask the Tuesday group not to discuss the final until after Wednesday

Bootstrap

- The bootstrap is a general method for validating any type of phylogeny inference
- It answers the question: How sensitive are our conclusions to small variations in the data?
- Felsenstein's paper announcing bootstrap is #41 on "most cited papers of all time"


Original
data
Sites
Sequences
Estimate of the tree
Bootstrap
sample #1
Sites
Sequences
Sample same number
of sites, with replacement
Bootstrap estimate o
the tree, #1
Bootstrap
sample #2
Sites
Sequences
Sample same number
of sites, with replacement
(and so on)
Bootstrap estimate o
the tree, #2


Aurelia MCA
A. limbata LIM (Japan/ Iceland)
Aurelia sp. 2
Aurelia sp. 2
Aurelia ARAB
Aurelia sp. 4
Aurelia sp. 4
Aurelia sp. 3
Aurelia sp. 3
Aurelia sp. 6
Aurelia sp. 6
Aurelia MKL
Aurelia sp. 5
Aurelia sp. 5
A. labiata LAB
A. labiata
A. labiata
Aurelia UBI
Aurelia sp. 1
Aurelia sp. 1
A. limbata (Alaska)
Aurelia TET
Aurelia BOR
A. aurita
A. aurita
57
100
78
98
100
100
100
79
100
75
100
62
69
100
77
96
100
10 changes

Bootstrap

- Consider a problem data set:

Sites supporting human+chimp	51
Sites supporting gorilla+chimp	49

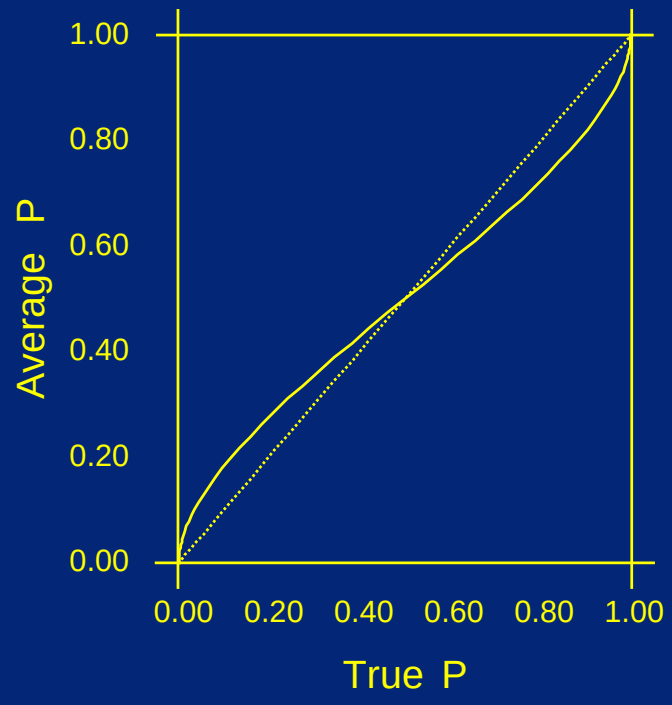
- Many of the resampled data sets will have 50-50 or 49-51 instead of 51-49.
- The human+chimp branch will not get strong bootstrap support
- This correctly reflects the poor signal of the data

Bootstrap

- Bootstrap assesses how sensitive your results are to random fluctuation in the data
- Does not detect violations of your assumptions
- Example: Method assumes a clock, but data are not clocklike
  - Original tree is systematically wrong
  - Bootstrap trees are systematically wrong too!

What do bootstrap values mean?

- Bootstrap values were originally interpreted as percent chance the branch was real
- This was disproven in the 1990's by computer simulation
- High values underestimate support; low values overestimate it



What do bootstrap values mean?

- There is no simple way to go from bootstrap value to percent support
- The relationship depends on number of tips and shape of tree
- Most people use a rough rule of thumb that 85% is a pretty good bootstrap and 65% is a definitely poor one
- It's best to publish the actual values and let readers draw their own conclusions

Other methods of validation

- Maximum likelihood algorithms come with built-in estimates of confidence
- These are only approximate for finite data
- Seldom used, I think because poorly understood

Bayes vs. bootstrap

- Bayesian "cloud of trees" can be treated like a bootstrap sample
- They answer different questions:
  - Bootstrap: would a slightly different data set prefer a different tree?
  - Bayesian support: would a slightly different tree fit this data set almost as well?
- It is easier to see that these are different than to understand how to use each one appropriately!
- If "cloud" is too small, results will be overly certain

Two hazards of phylogeny

- Garbage in, garbage out:
  - Long pieces of autosomal DNA
  - Misaligned sequences
  - Non-homologous traits
- Gene tree not necessarily the same as the species tree
  - Paralogs
  - Incomplete lineage sorting (ancestral polymorphism)
  - Horizontal gene transfer
  - Hybrid species

Coalescent inference

- Established in first half of course:
  - Rate of coalescence depends on N_e
  - Forces like growth, population subdivision can influence this
- Talked mainly about summary statistics like θ_π and F_{ST}
- Can we do better by inferring relationships among individuals?

Coalescent genealogy samplers

- $P(D|T)$ from a mutational model
- $P(T|\theta)$ from the coalescent
- Can add additional parameters:
  - growth rate
  - population structure
  - migration rate
  - recombination rate
- Sample genealogies based on $P(D|T)P(T|\theta)$

Sampling procedure

- Very similar to Bayesian phylogenetics, though developed separately
- Start with initial guesses for θ and T
- Iterate:
  - Change the tree slightly based on $P(T|\theta)$
  - Accept/reject based on $P(D|T)$
  - Record the current tree at intervals
- At the end, estimate θ (and other parameters if used)
- Note that the goal is parameter estimation, though you do get a cloud of trees as a side effect

Coalescent genealogy estimation

- Pros:
  - Much more informative than summary statistics like F_{ST}
  - Gives built-in estimates of certainty
- Cons:
  - Computationally cumbersome
  - Models are always oversimplified
  - Often works only for a narrow range of parameters (though other methods might not work outside that range either)

HIV epidemic in Libya

- In 1998-1999, over 400 children found infected with HIV
- All had been treated at El-Fatih Children's Hospital in Benghazi
- 43% were also infected with Hepatitis C

Benghazi, Libya. Image by Dennixo, from Wikipedia

Three theories

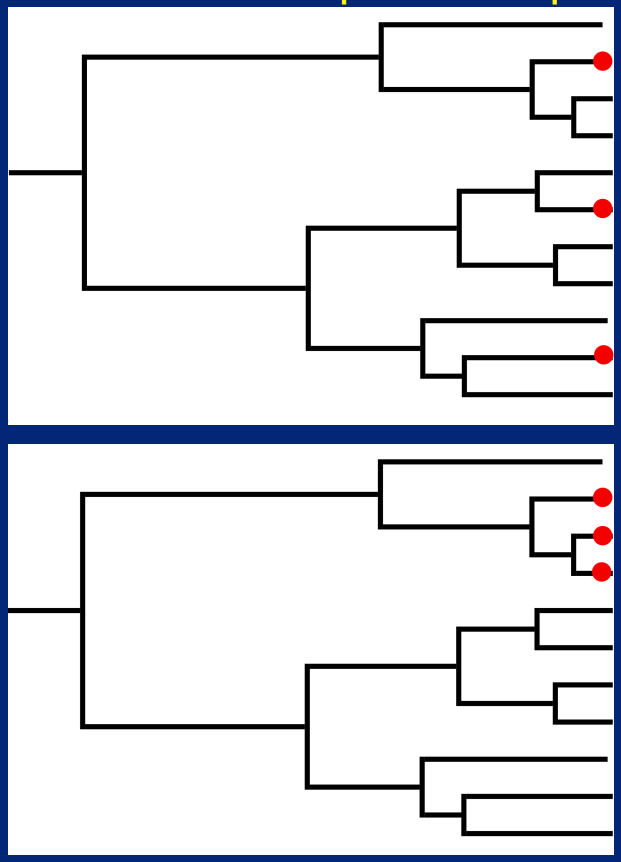
- The children became infected:

  1. ...in their communities ("community theory")
  2. ...at El-Fatih due to poor sanitary practices ("accident theory")
  3. ...at El-Fatih due to deliberate acts ("murder theory")

- Libya accepted the murder theory and sentenced 6 foreign medics to death

Expected relationships

Red dots = hospital samples, unmarked lines = community samples



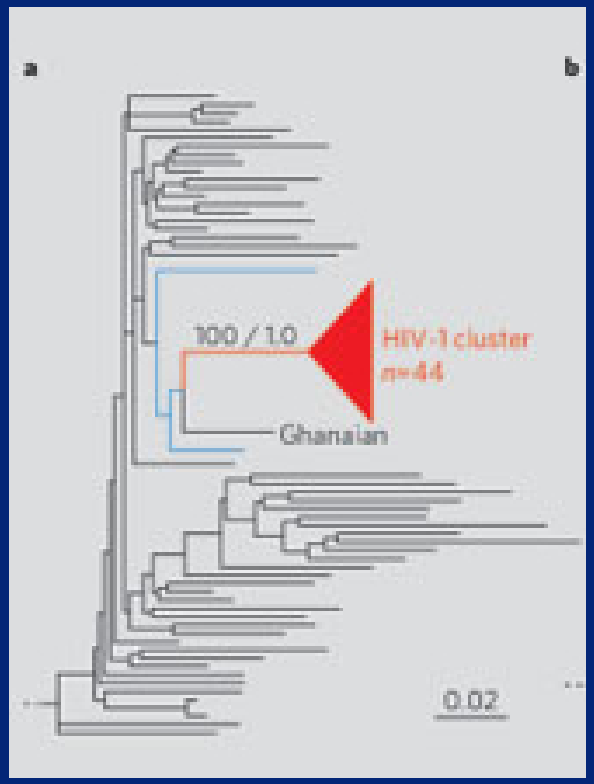
Community theory

Accident or murder theories

Subsample of 44 children analyzed (WGS of viruses) and compared to global database, with particular reference to Egypt, Cameroon and Ghana

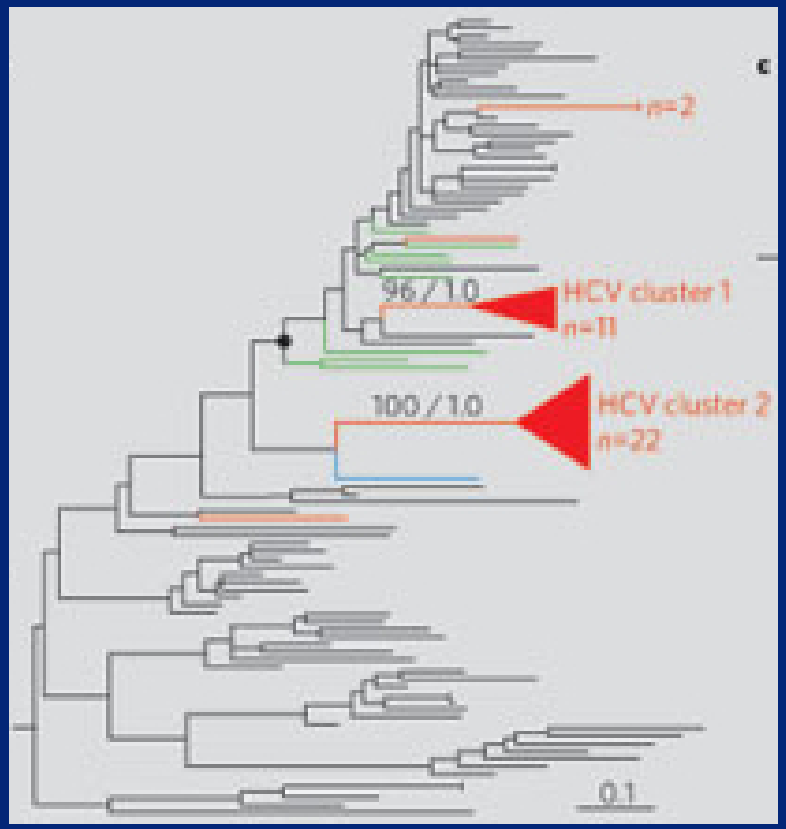
Northern Africa

Observed relationships: HIV-1



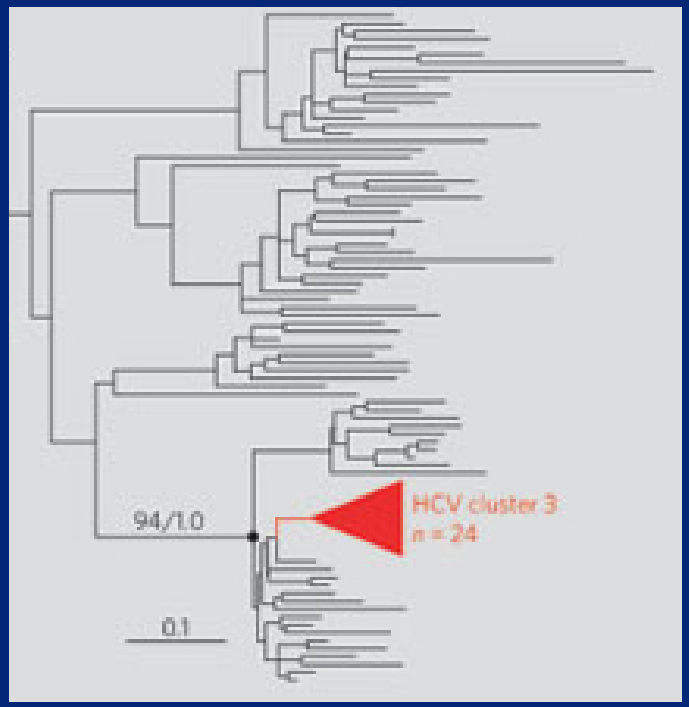
Red = hospital, blue = Cameroon
from de Oliveira et al, Nature 2006; used with permission

Observed relationships: HCV type 1



Red = hospital, green = Egypt, blue = Cameroon
from de Oliveira et al, Nature 2006; used with permission
This strain of HCV is epidemic worldwide

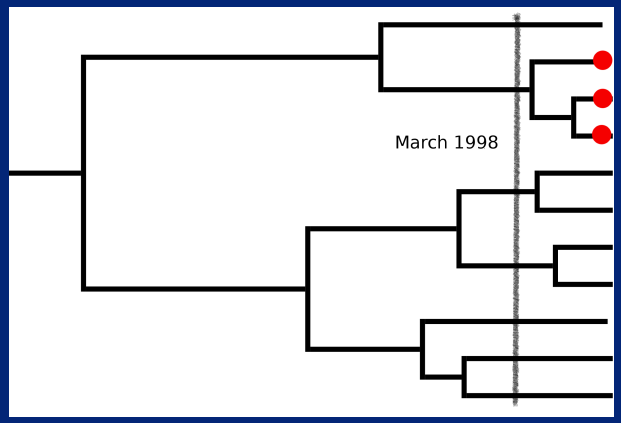
Observed relationships: HCV type 4



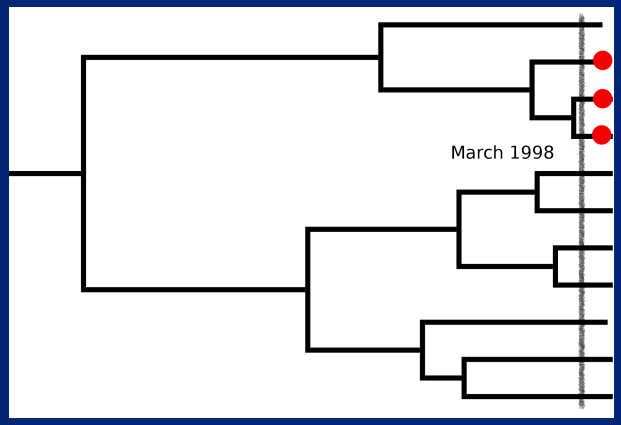
Red = hospital, green = Egypt, blue = Cameroon
from de Oliveira et al, Nature 2006; used with permission
This strain of HCV is epidemic in Egypt due to contamination of anti-worm medication in the 1970's

First conclusions from genetic analysis

- HCV type 1 results suggest community theory
- HCV type 4 and HIV results suggest a single origin of the virus in all 44 children
- Accident or murder?
- We know medics arrived in Libya in March 1998–



Common origin after medics' arrival: murder theory possible



Common origin before medics' arrival: murder theory impossible

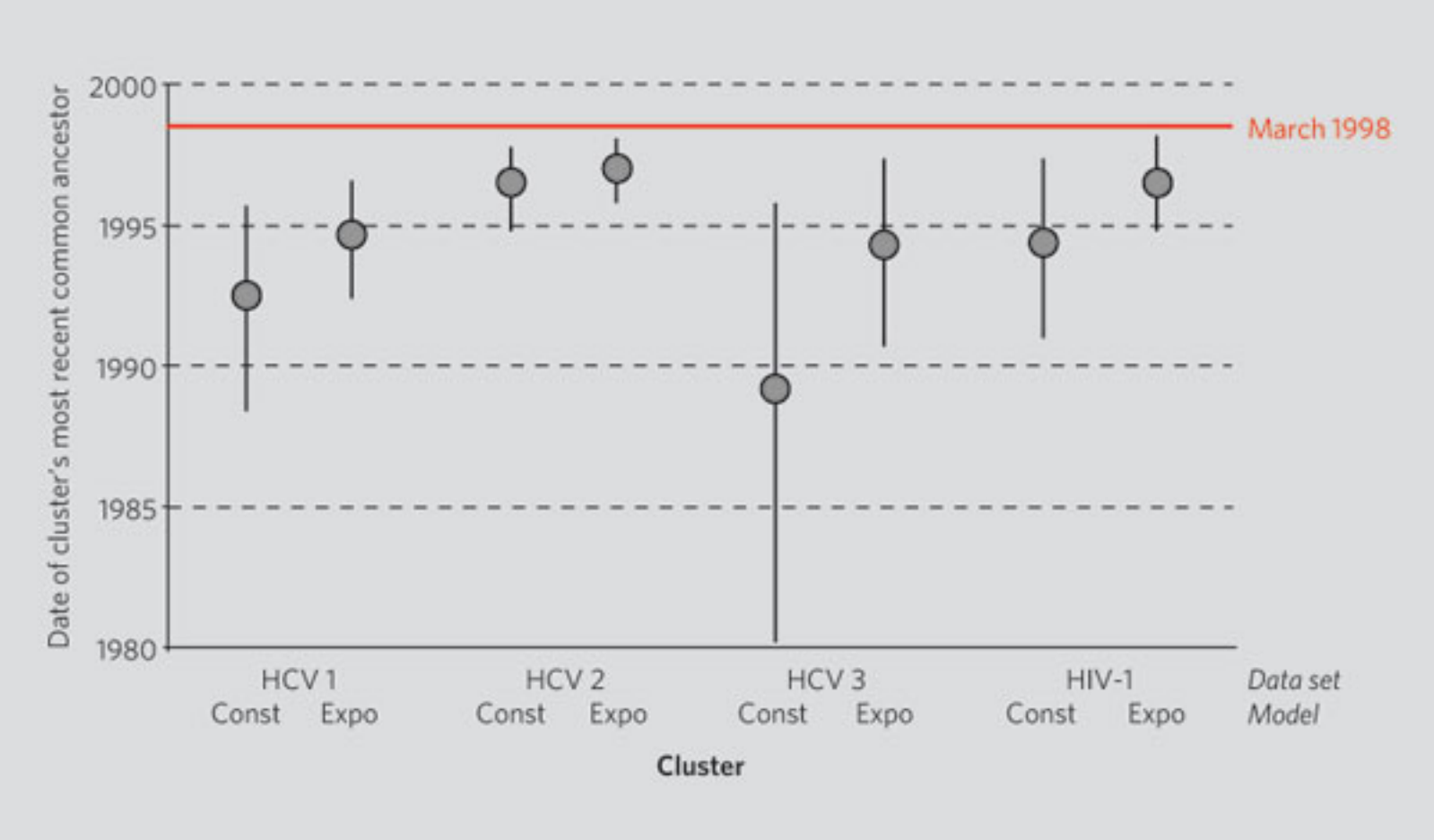
Mutation rate?

- Libya claimed the HIV virus was genetically engineered
- Genetic sequencing found:
  - It was a member of the CRF02_AG subtype
  - Number of mutations back to the common ancestor of CRF02_AG was similar to other strains
  - No evidence for engineering–mutation rate typical of HIV-1

When was the common ancestor of the childrens' viruses?

- Estimate mutations back to common ancestor with BEAST
  - Coalescent genealogy sampler
  - Specialized to allow relaxed molecular clock (important in virus data)
- Convert to years using estimated mutation rates

The viruses arose before March 1998



from de Oliveira et al, Nature 2006; used with permission

Vindication of the genetic analysis

- In August 2007 Saif al-Islam Gaddafi confirmed that some children had been infected prior to February 1998
- He also confirmed that the confessions were obtained via torture and threats to families

In July 2007 the medics were extradited to Bulgaria and freed by the Bulgarian government.

Monday

- Gene trees and species trees