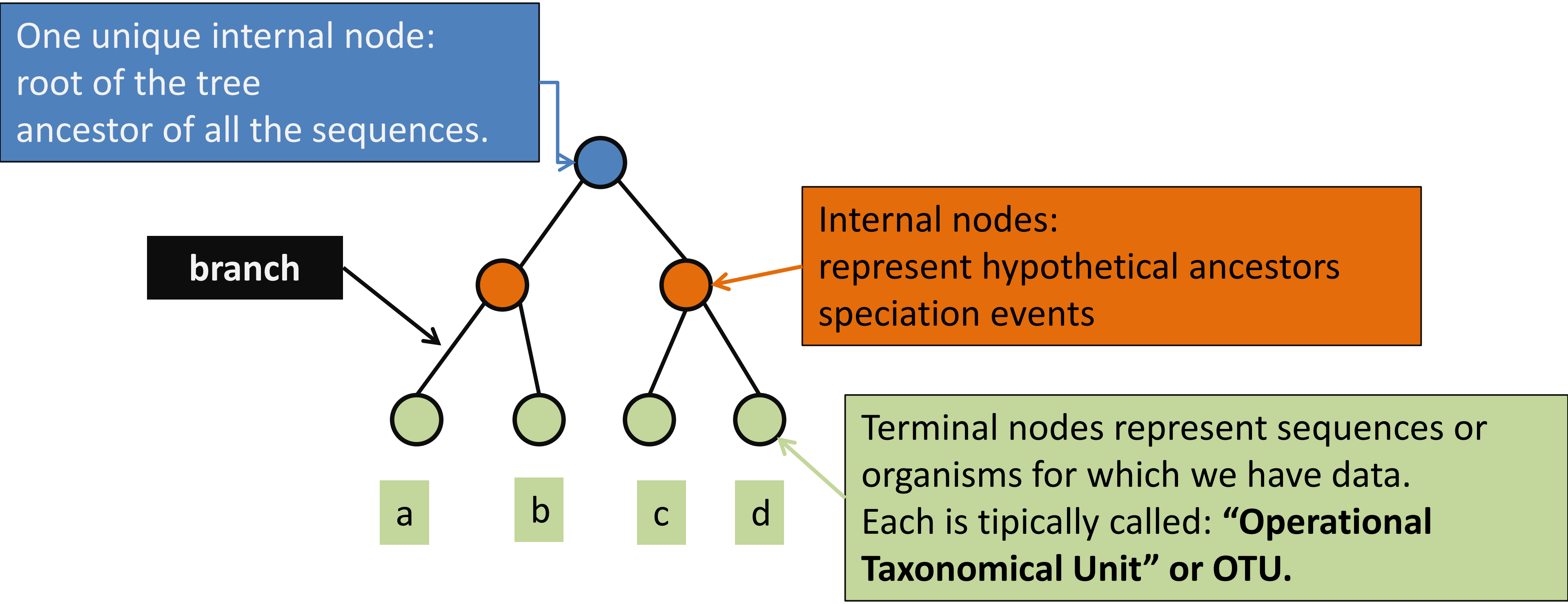
Phylogenetic analysis practical

What is a tree?

- A tree is a mathematical structure which represents a model of an actual evolutionary history of a group of sequences or organisms.
- In other words, it is an **evolutionary hypothesis.**



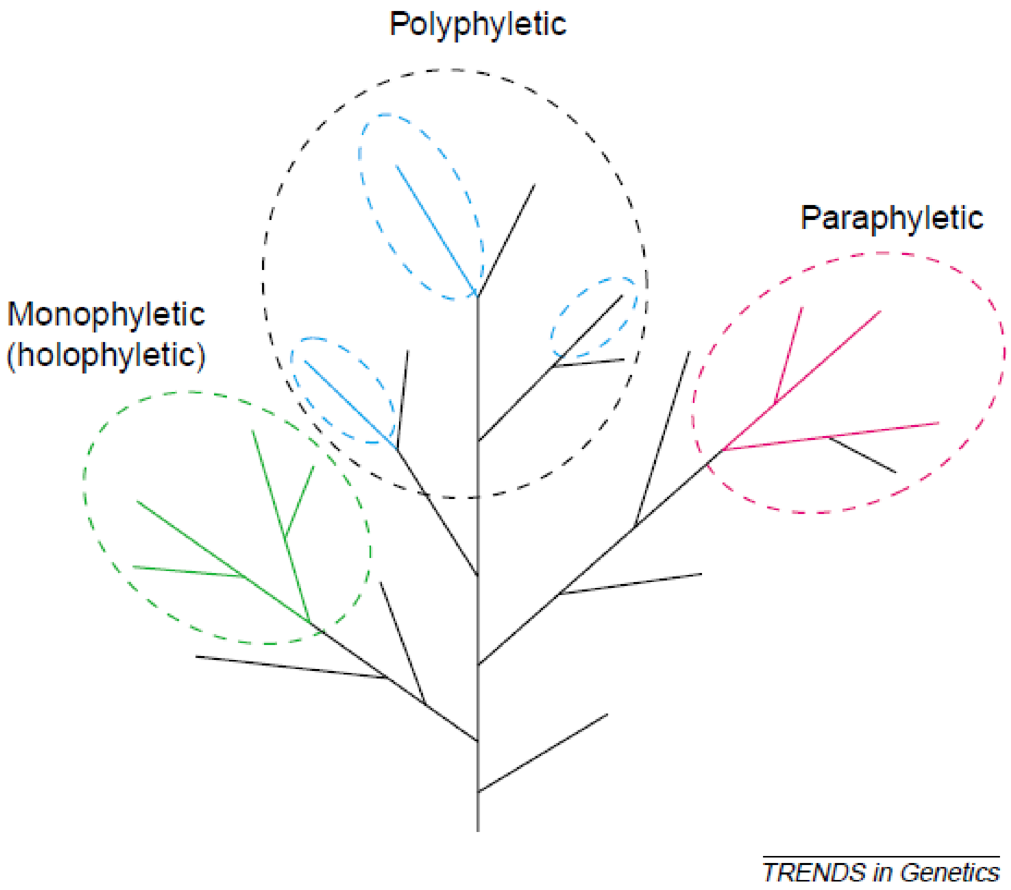
One unique internal node:
root of the tree
ancestor of all the sequences.

Internal nodes:
represent hypothetical ancestors
speciation events

Terminal nodes represent sequences or organisms for which we have data.
Each is tipically called: **“Operational Taxonomical Unit” or OTU.**

Groups

- node and everything arising from it is a ‘clade’ or a **monophyletic group** – all members are derived from a unique common ancestor
- **Paraphyletic group** - group excluding some of its descendents
- **Polyphyletic group** – group consisting from groups arosed from different ancestors – not a group at all



doi:10.1016/S0168-9525(03)00112-4

Rates and causes of molecular evolution

- Different parts of the genome are useful for different problems.
- Fast evolving sequences are useful for recent events, but become saturated and unrecognizable when comparing more distant relatives.
- Slow evolving sequences are useful around the base of the tree, but don’t have any variability at all among close relatives.

Different molecular regions, different rates

- DNA distant from genes evolves very quickly (at about one substitution per 10^8 years),
- Flanking regions upstream and downstream from a gene evolve less quickly than that,
- Introns evolve less quickly than those, though not much less,
- Third positions of codons evolve less quickly than introns,
- First and second positions of codons evolve less quickly than that
- Human Y-chromosome point mutation rate
  - **8.71×10^{-10} mutations per position per year (PPPY)**
  - **7.37×10^{-10} PPPY sequence from palindromes (PAL)**
  - **7.2×10^{-10} PPPY for paternally transmitted autosomes**

Helgason et al. *Nature Genetics* (2015) doi:10.1038/ng.317

Different molecular regions, different rates

Within a protein:

– active sites evolve very slowly,

– sites that bind heme, or interact with other proteins evolve a bit faster but also very slowly,

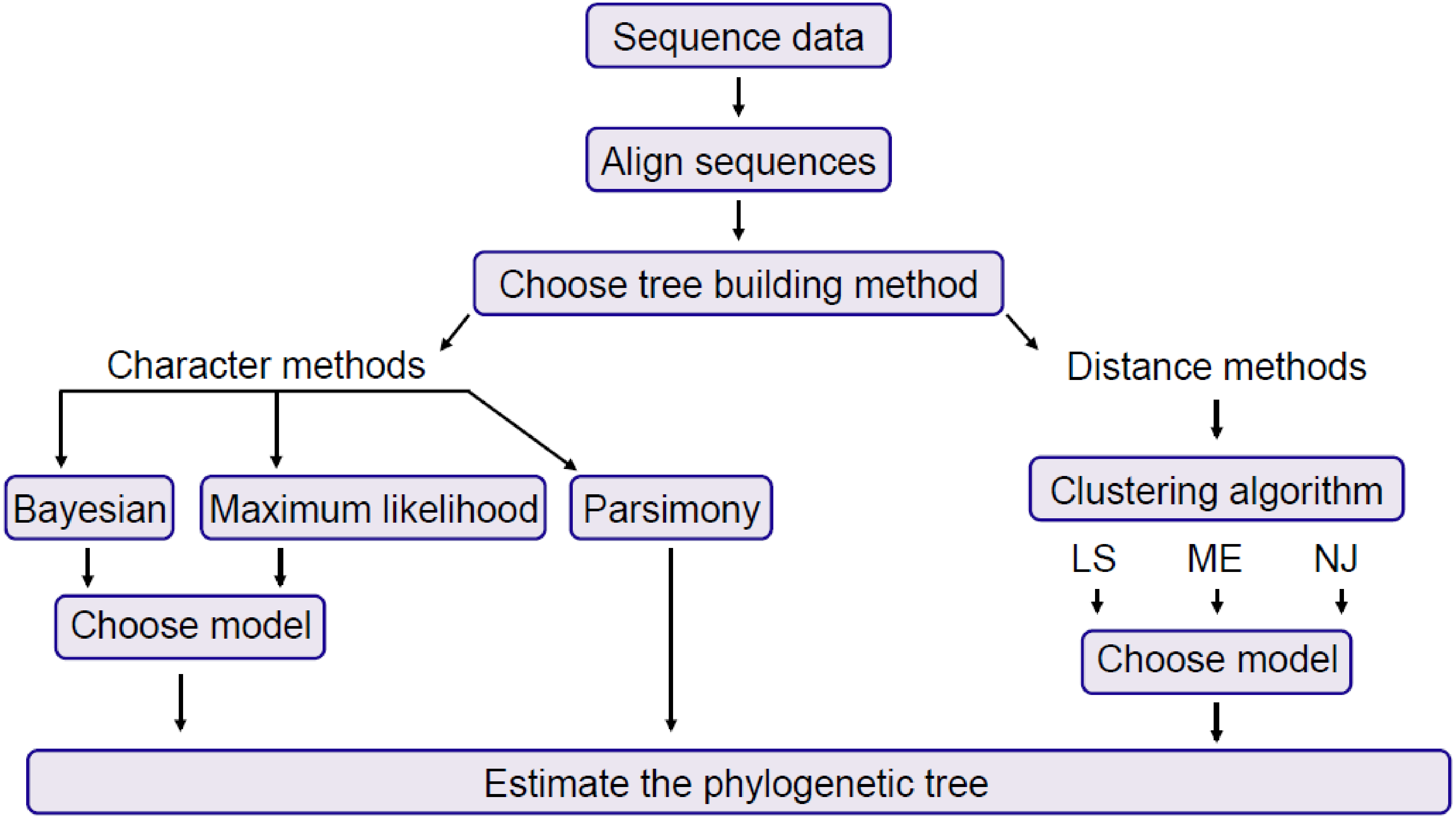
– interior sites evolve less quickly than exterior sites,

– substitutions that involve less radical changes of the amino acid (i.e. that change to a rather similar amino acid) happen more readily.

Of base changes, transitions (A -> G or C -> T) happen several times more readily than transversions (all other changes).

Between protein-coding loci, some (fibrinopeptide, for example) evolve rapidly, some less so (hemoglobins, cytochromes), and some (histones, for example) change very slowly.

Phylogenetic inference: From sequences to tree



Sequence data

Align sequences

Choose tree building method

Character methods

Distance methods

Bayesian

Maximum likelihood

Parsimony

Clustering algorithm

LS

ME

NJ

Choose model

Choose model

Estimate the phylogenetic tree

Credit: Jeff Silberman

Distance-based methods

The general idea

1. Calculate distances among all the sequences in a character matrix
2. Find a tree that best fits those observed distances – now dealing with a compressed distance matrix

Several distance methods/algorithms

- Neighbor Joining
- Fitch-Margoliash (least squares)
- Minimum Evolution

Summary

- Have the advantage of being extremely fast
- Not so good in terms of accuracy (ME's horrible)
- Need to specify the right model
- Need more data than character-based methods to achieve similar levels of accuracy
- Generally avoided

Character-based methods

The general idea

1. Evaluate each column in the alignment
2. Infer relationships based on the patterns in those columns

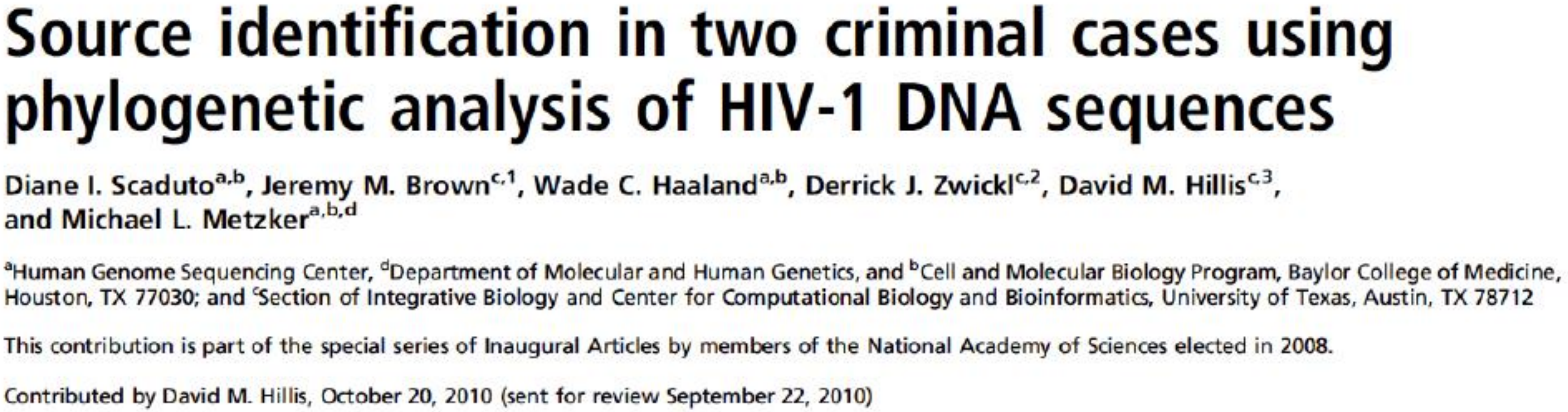
Two approaches/philosophies/schools of thought

- Maximum Parsimony – no *explicit* model of character evolution; minimize the overall number of character-state changes on the tree
- Model-based methods – specify an explicit model of character evolution (Maximum Likelihood and Bayesian Inference)

Summary

- Preserve more of the data than distance methods
- Outperform distance methods
- Model-based methods generally outperform parsimony methods
- All of the methods are sensitive to taxonomic sampling
- Model-based methods are guaranteed to work well when the model is properly specified (i.e., it properly accounts for the evolutionary process – this is hard)

Using Phylogeny to Disentangle the Criminal Spread of HIV



Source identification in two criminal cases using phylogenetic analysis of HIV-1 DNA sequences

Diane I. Scaduto[a,b], Jeremy M. Brown[c,1], Wade C. Haaland[a,b], Derrick J. Zwickl[c,2], David M. Hillis[c,3], and Michael L. Metzker[a,b,d]

[a]Human Genome Sequencing Center, [d]Department of Molecular and Human Genetics, and [b]Cell and Molecular Biology Program, Baylor College of Medicine, Houston, TX 77030; and [c]Section of Integrative Biology and Center for Computational Biology and Bioinformatics, University of Texas, Austin, TX 78712

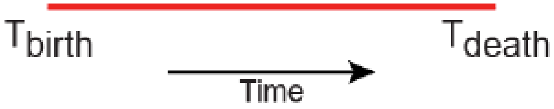
This contribution is part of the special series of Inaugural Articles by members of the National Academy of Sciences elected in 2008.

Contributed by David M. Hillis, October 20, 2010 (sent for review September 22, 2010)

Scaduto et al. 2010. PNAS [Link]

Using Phylogeny to Disentangle the Criminal Spread of HIV

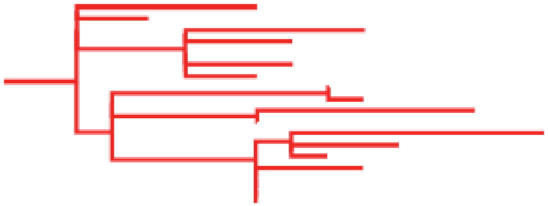
- common to use DNA profiling to identify individuals – made possible by the stability of human genomes over the course of a lifetime



T_{birth} T_{death}

Time

- HIV has high mutation and recombination rates, and extremely high replication rate (10^8 – 10^{10} virions per day)
- As a result, individuals with HIV contain a genetically diverse and rapidly evolving population of related genomes



HIV genetic variation within a single individual

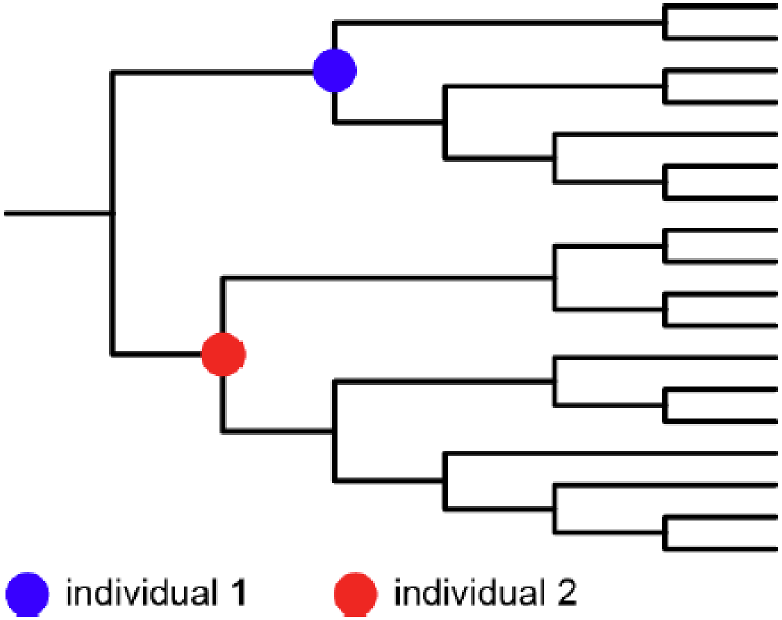
- It's therefore difficult (if not impossible) to use simple DNA profiling for HIV matching

Scaduto et al. 2010. PNAS [Link]

The Underlying Theory

HIV dynamics within and between individuals

- a single evolutionary lineage (monophyletic)
- within an individual, HIV strains share a most recent common ancestor
- HIV strains within an individual are more closely related to each other than they are to HIV strains in another individual



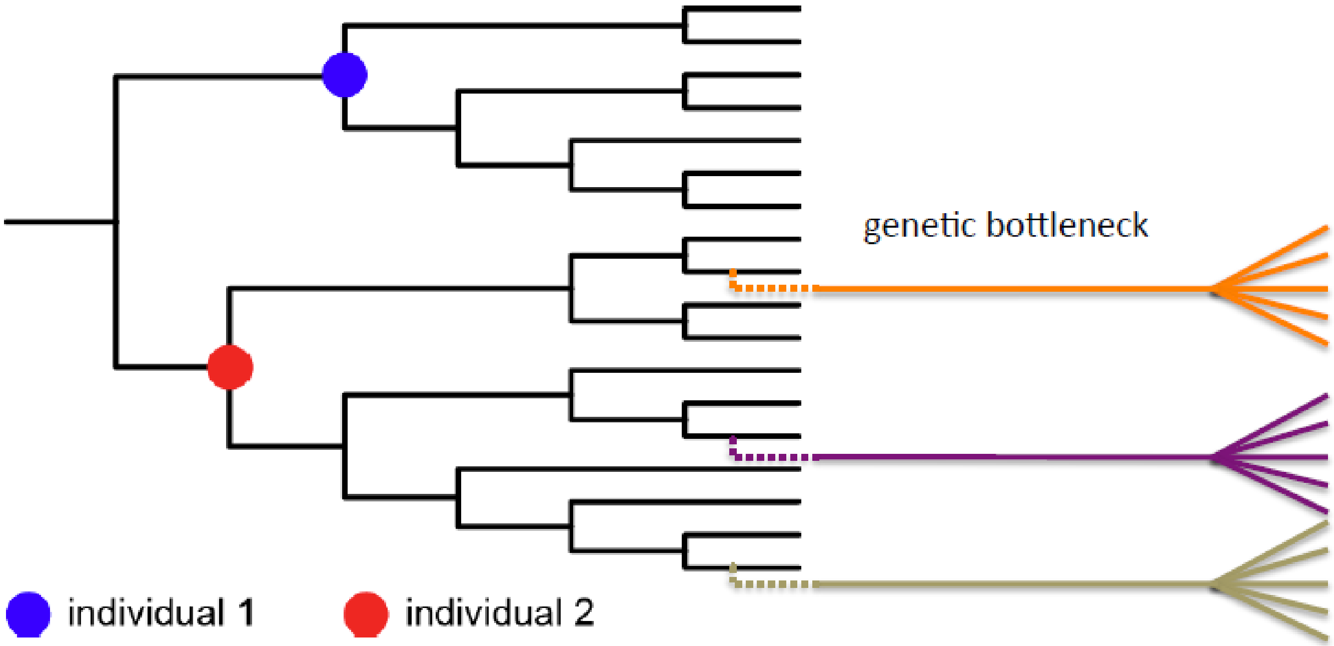
Question:
What would the tree look like if individual 2 began infecting others?

Scaduto et al. 2010. PNAS [Link]

The Underlying Theory

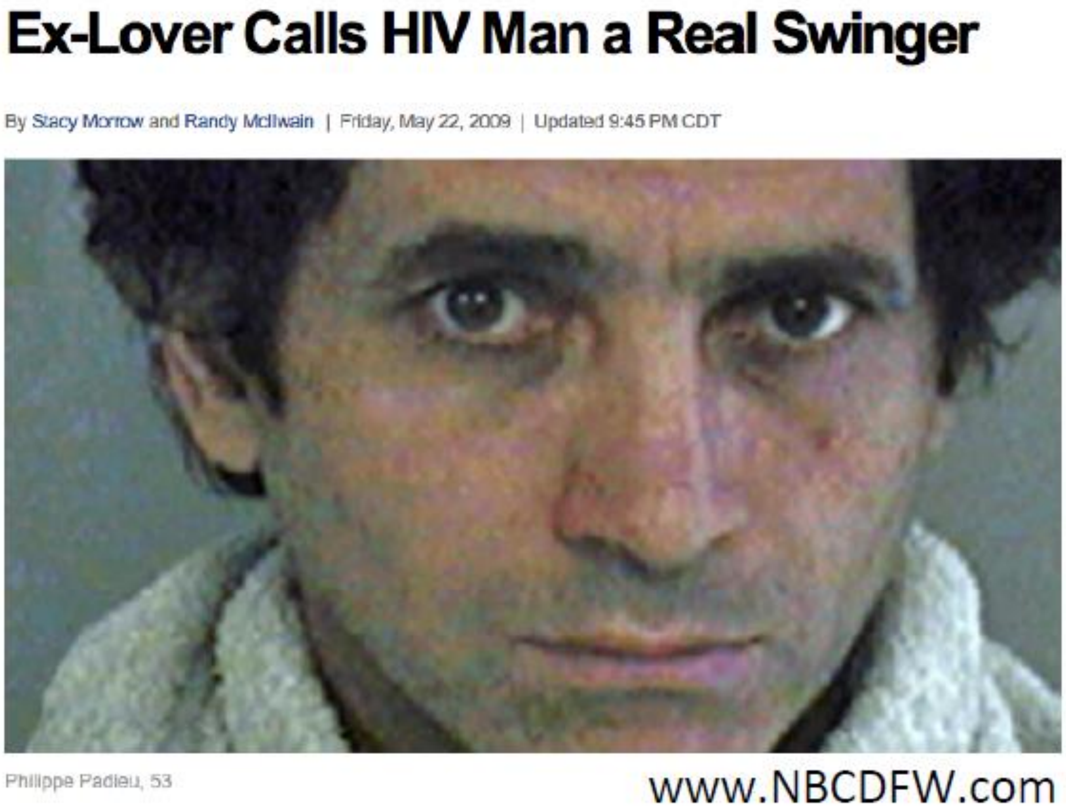
A subset of the donor's HIV population is passed to the recipient, resulting in a major genetic bottleneck at the time of transmission – a single virus is responsible for clinical infection in most cases

- the recipient's HIV population will diversify and evolve independently after the infection
- **Prediction:** monophyly of recipient HIV sequences, paraphyly of donor HIV sequences
- That is, a subset of source viral sequences is more closely related to all recipient sequences than to other source sequences – can identify the direction of transfer and therefore the source



Scaduto et al. 2010. PNAS [Link]

State of Texas *vs.* Phillipe Padieu



Ex-Lover Calls HIV Man a Real Swinger

By Stacy Morrow and Randy McIlwain | Friday, May 22, 2009 | Updated 9:45 PM CDT

Philippe Padieu, 53

www.NBCDFW.com

- six counts of aggravated assault with a deadly weapon
- knowingly infected six women he was dating with HIV
- July 2007
- Dallas/Fort Worth, TX

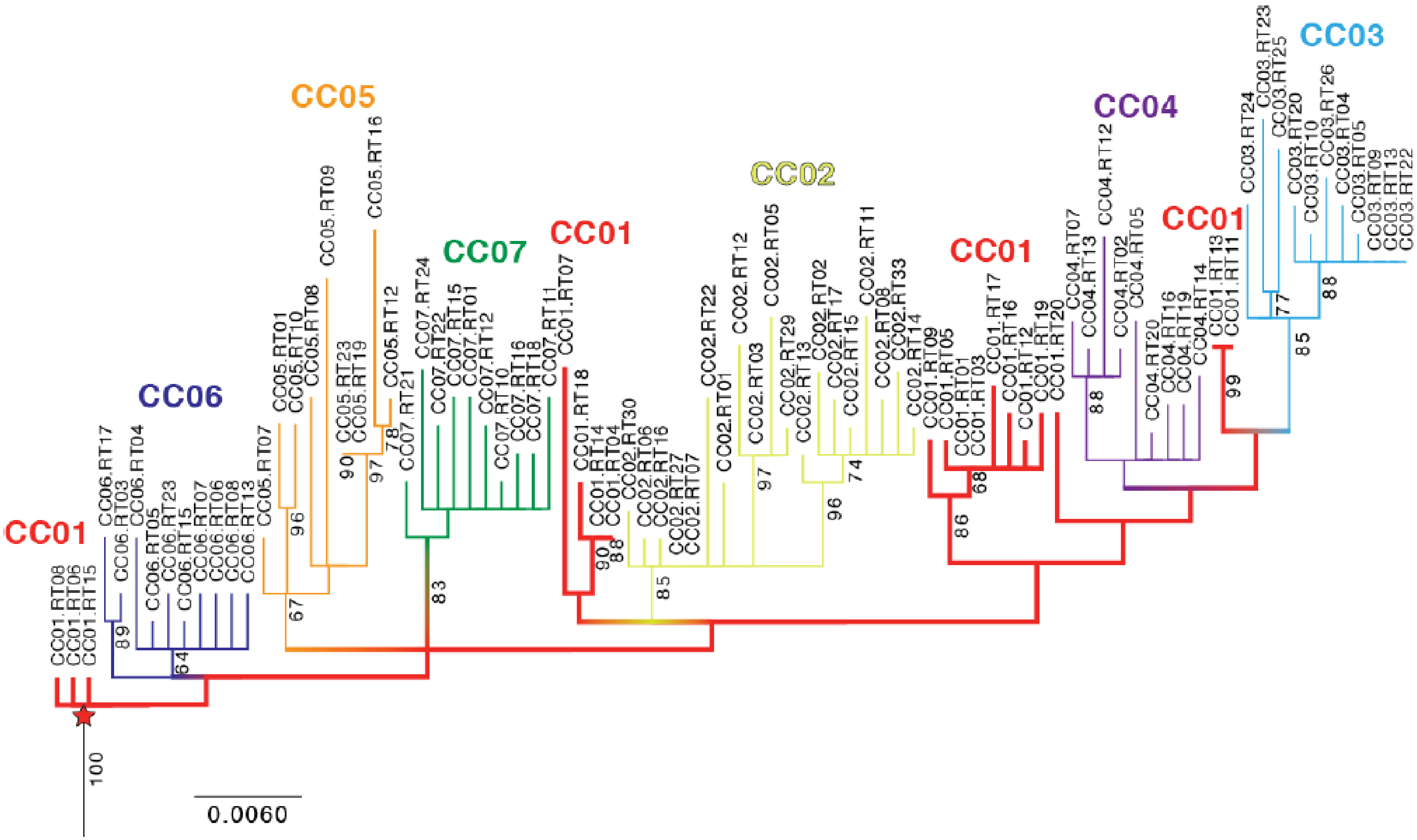
The approach

- 1. Conduct a "blind“ study by anonymously coding blood samples from the defendant and the alleged victims
    - Collin County, TX, samples: CC01 – CC07
- 2. PCR amplify the HIV *env* and *pol* genes
- 3. Clone the PCR products into bacterial vectors
- 4. Collect DNA sequences for a representative sample of the HIV population in each individual (ca. 20 sequences/individual)
- 5. Use BLAST to identify outgroup sequences from GenBank
- 6. Make an alignment, perform phylogenetic analysis...

Exercise

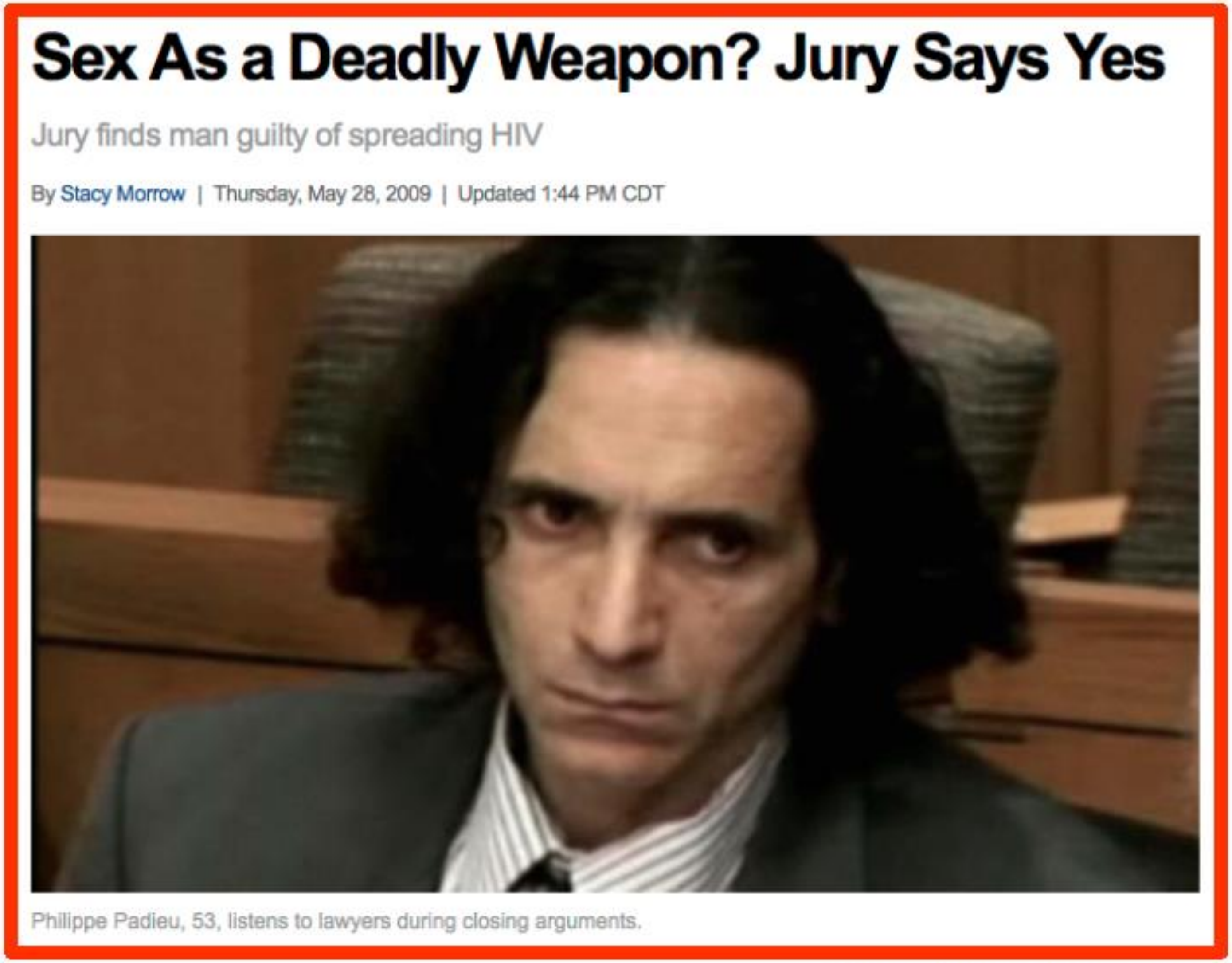
- Align the fasta file using ClustalX
- Look at the alignment
- Use phylip to create ML and distance tree
- View the tree in treeview
- Root the tree (outgroup)
- Who did it?

Maximum Likelihood Tree (Texas, *pol* gene)



Scaduto et al. 2010. PNAS [Link]

CC01: Phillipe Padieu



Sex As a Deadly Weapon? Jury Says Yes

Jury finds man guilty of spreading HIV

By Stacy Morrow | Thursday, May 28, 2009 | Updated 1:44 PM CDT

Philippe Padieu, 53, listens to lawyers during closing arguments.

www.NBCDFW.com