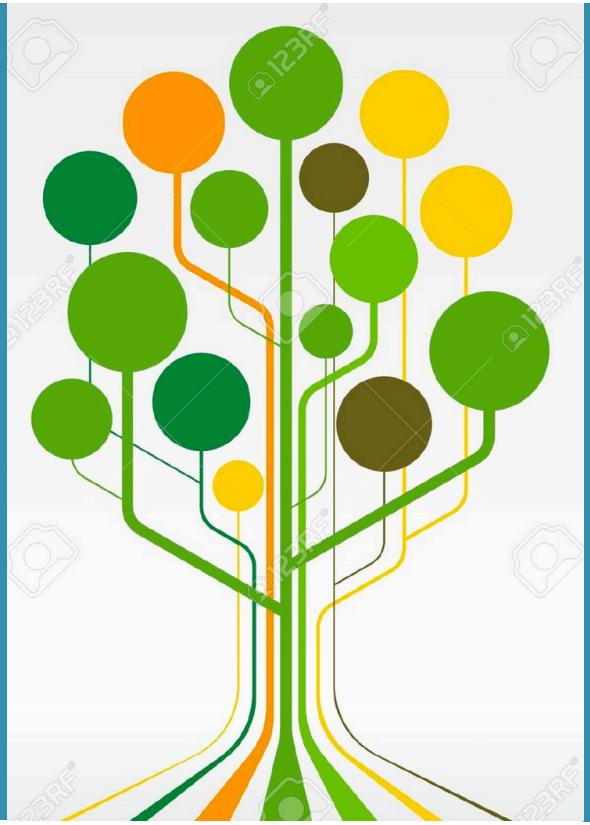


Trees and networks

Chiara Barbieri
Max Planck Institute for the Science of Human History, Jena

Phylogenetic networks

- “any” network in which taxa are represented by nodes and their evolutionary relationships are represented by edges. (For phylogenetic trees, edges are referred to as branches.) *Huson & Bryant 2006, Mol Biol Evol*

Phylogenetic networks

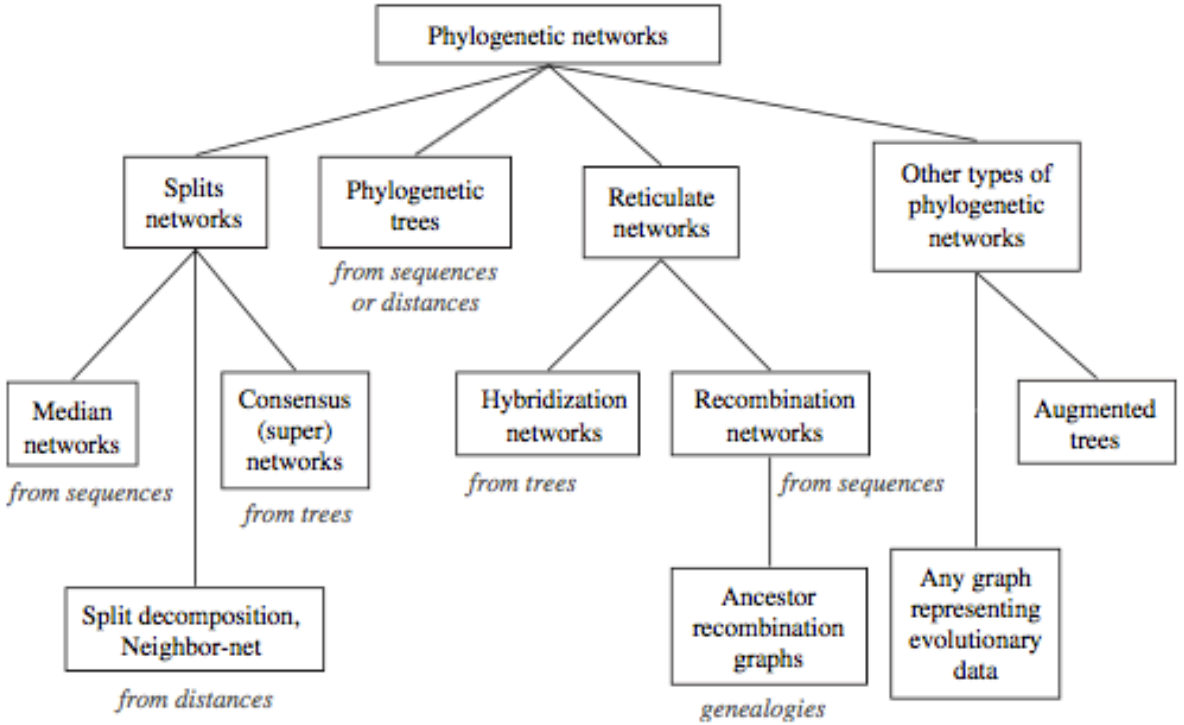
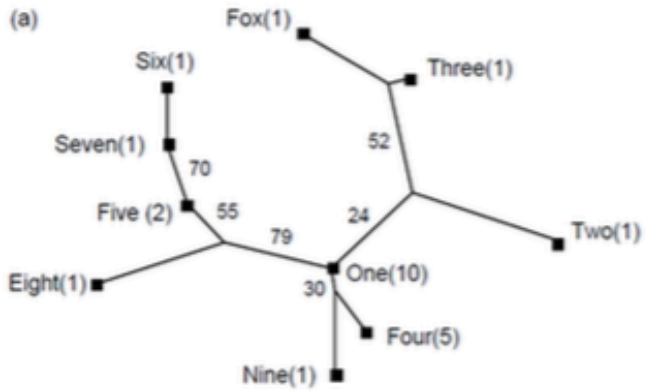


FIG. 1.—The term phylogenetic network encompasses a number of different concepts, including phylogenetic trees, split networks, reticulate networks, the latter covering both "hybridization" and "recombination" networks, and other types of networks such as "augmented trees." Recombination networks are closely related to ancestor recombination graphs used in population studies. Split networks can be obtained from character sequences, for example, as a median network, and from distances using the split decomposition or neighbor-net method or from trees as a consensus network or super-network. Augmented trees are obtained from phylogenetic trees by inserting additional edges to represent, for example, horizontal gene transfer. Other types of phylogenetic networks include host-parasite phylogenies or haplotype networks. Diagram adapted from Huson and Kloepper (2005).

Phylogenetic trees and networks

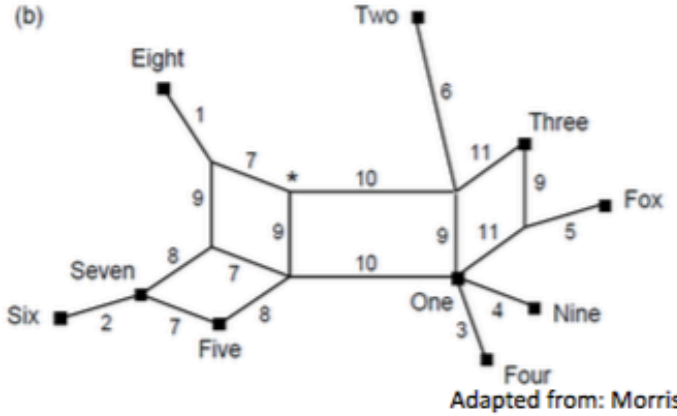
- Trees impose bifurcations
- Networks allow reticulations

Phylogenetic tree



is a tree for a set of taxa S with labels on all leaves, and possibly on some internal nodes. A phylogenetic tree may be rooted or unrooted, weighted or unweighted, binary or nonbinary.

Phylogenetic network

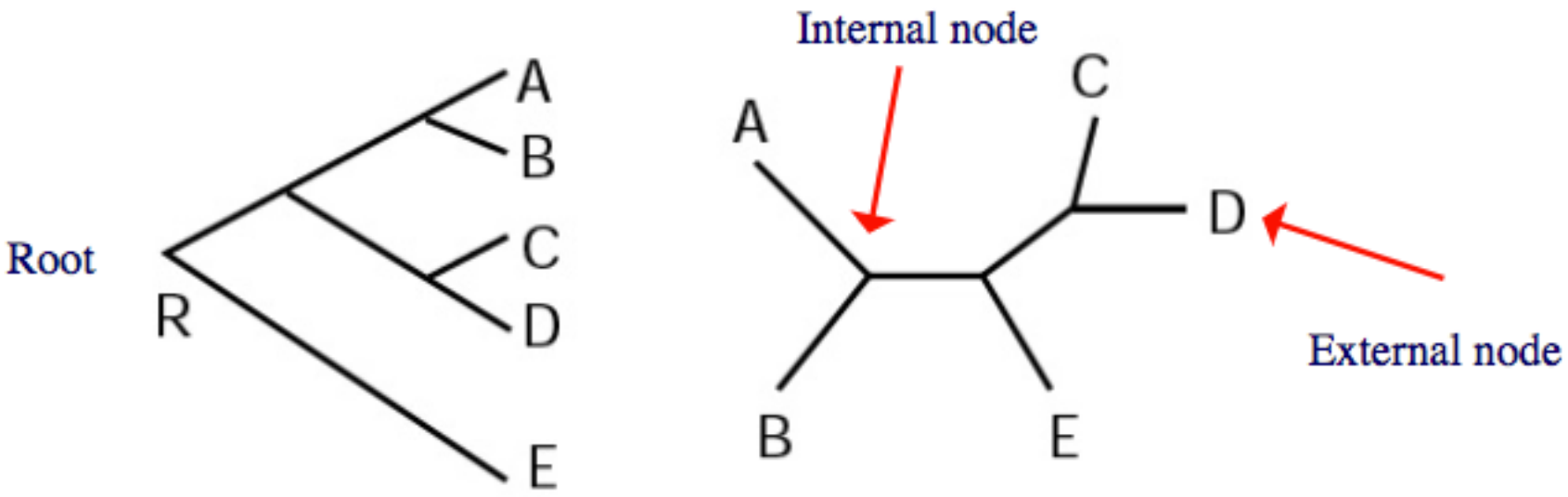


Adapted from: Morrison 2005

is a connected graph, again with some of the nodes labelled. In a network a set of (parallel) edges (branches) may be required to partition the graph into two connected subgraphs (so the graph appears 'box-like' as in figure.

Rooted vs unrooted trees

- If you have an outgroup you can root your tree



Rooted tree

Unrooted tree

Inferring phylogenies

- Inferring a tree is a combination of at least three components:

1. optimality criterion (parsimony, max likelihood, minimum evolution, least-squares fit, etc.).
2. search strategy (cluster methods, branch-and-bound, quartets, heuristic searches, etc.)
3. Assumptions about the mechanisms of evolution (JC, K2P, HKY, etc.)

Methods for building trees

- Distance-based
  - UPGMA
  - Neighbour Joining (NJ)
- Character-based
  - Maximum Parsimony (MP)
  - Maximum Likelihood (ML)
  - Bayesian methods (Markov Chain Monte Carlo MCMC)

Distance-based trees

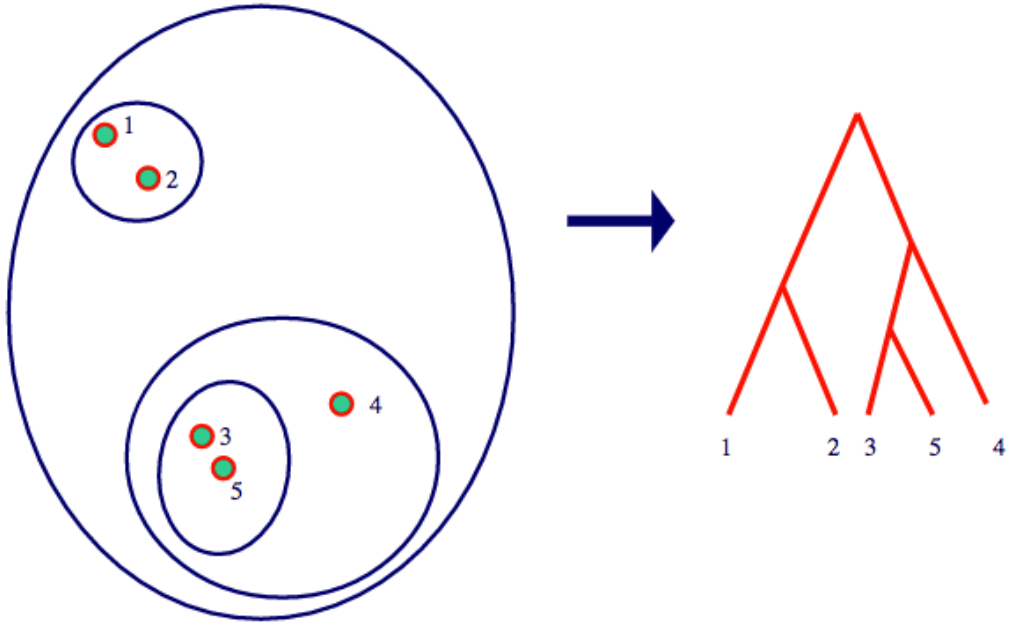
- First calculate distance matrix between pairs of sequences or populations
- Then build a tree

Distance-based trees

- UPGMA (Sokal and Sneath 1963): based on the molecular clock assumption generates ultrametric trees.
  - **Rooted** tree
  - all the end nodes are equidistant from the root
  - assuming a **molecular clock**.
- agglomerative (bottom-up) hierarchical clustering method. Picks the closest pair of neighbors, and adds the closest, and so on

Distance-based trees

- UPGMA (Sokal and Sneath 1963): based on the molecular clock assumption generates ultrametric trees.

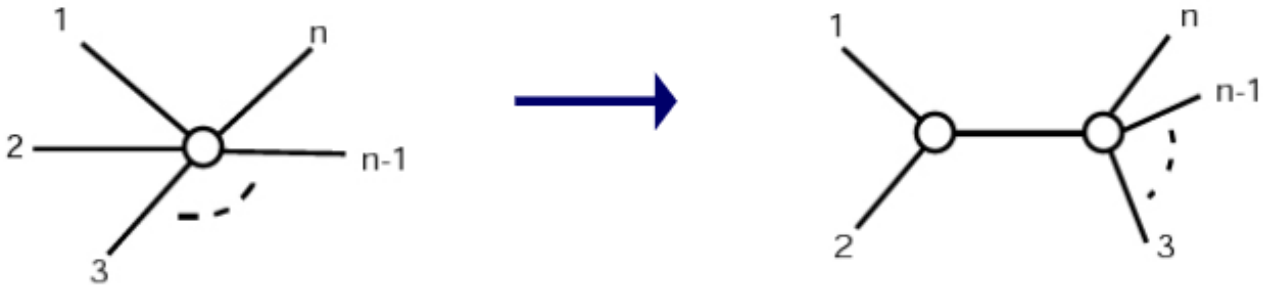


Distance-based trees

- **Neighbor-Joining NJ** (Saitou and Nei 1987)
  - **Unrooted** tree
  - Does not assume a **molecular clock**.
- Local search strategy using a Minimum Evolution (ME) optimality criteria
- Starts with an unresolved star-like tree, calculate the sum of branch length. Joins the pair with the closest branch length. And so on

Distance-based trees

- **Neighbor-Joining NJ** (Saitou and Nei 1987)



Start off with star tree; pull out pairs at a time

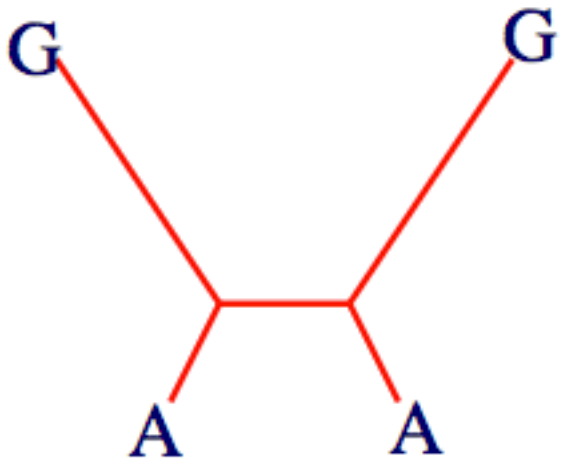
Character based trees

- Maximum parsimony (MP): choose tree that minimizes number of changes from a common ancestor
  - MP yields more than one tree with the same score
- Maximum likelihood (ML): find the tree which gives the highest likelihood of the observed data

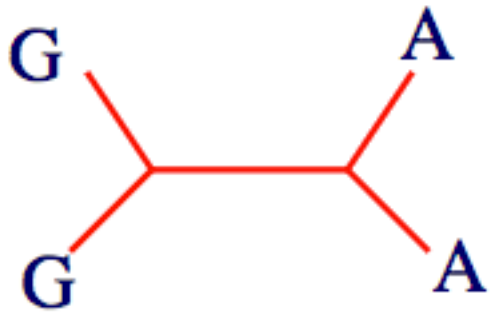
They both imply model of evolution

Parsimony weakness: long branch attraction

- Parsimony analysis implicitly assumes that rate of change along branches are similar



Real tree: two long branches where G has turned to A independently



Inferred tree

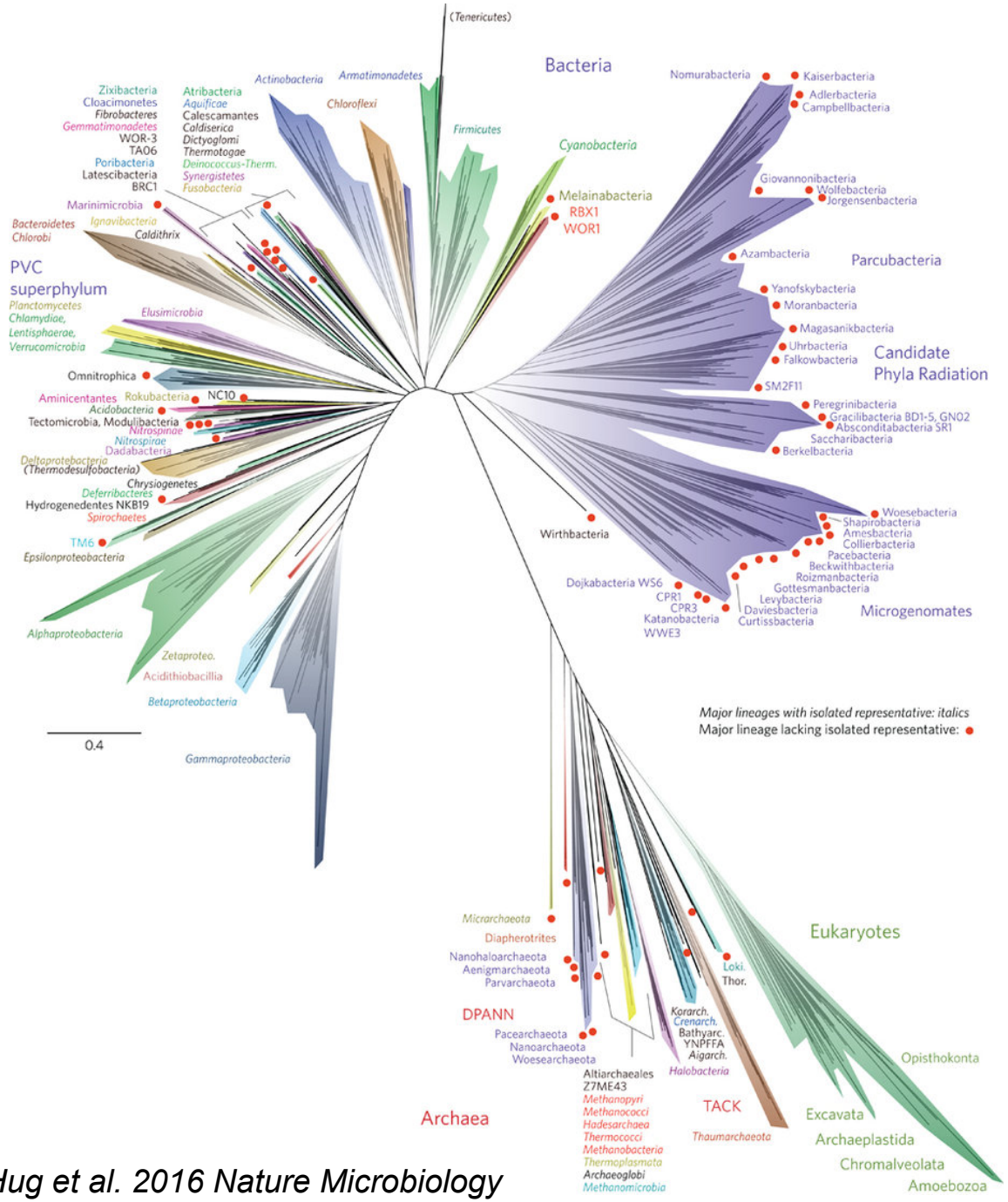
Summary: trees

- Distance methods are good for large data sets of highly similar sequences
- **Likelihood** and **Bayesian** methods often have more power and are more robust, especially for inferring deep phylogenies

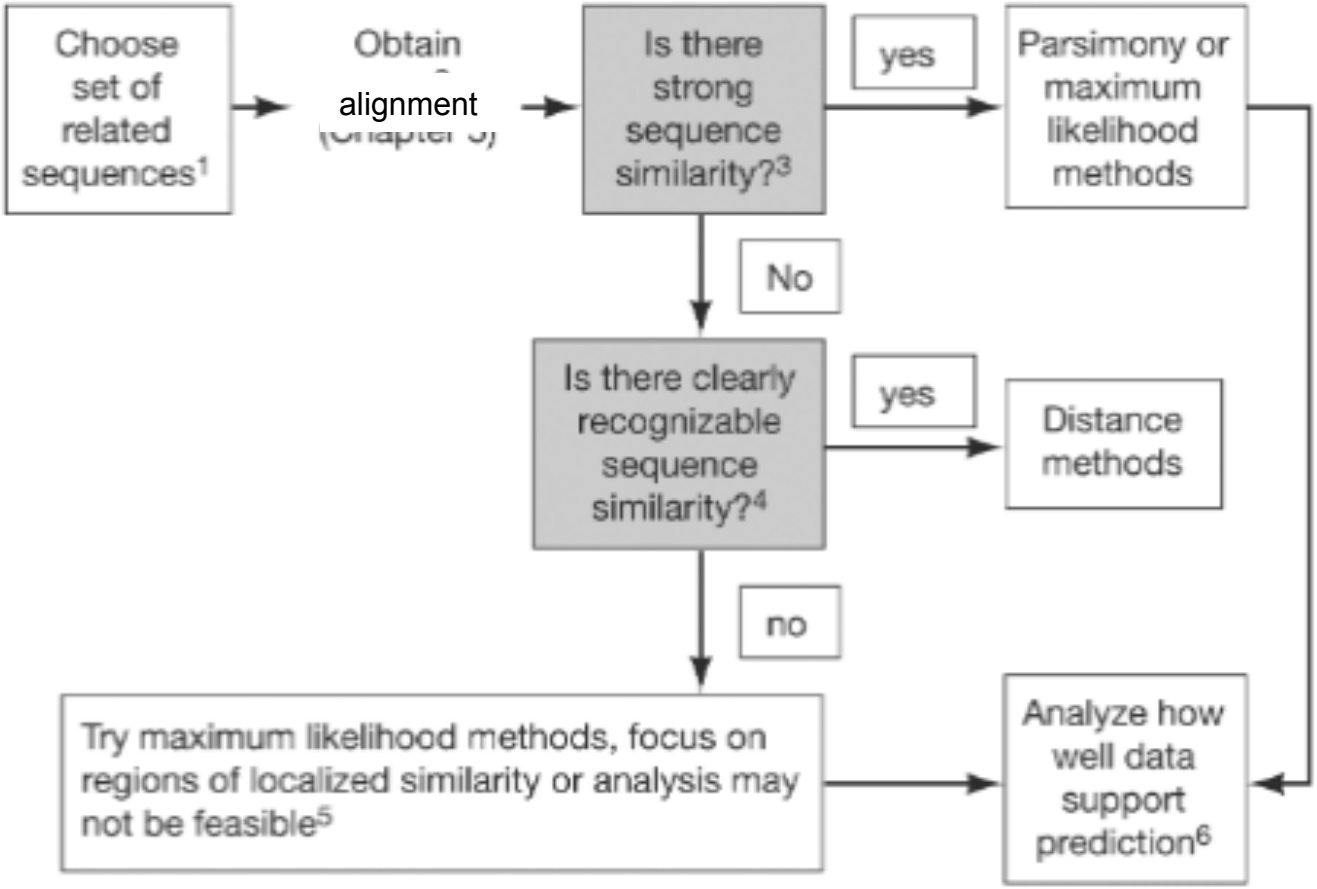
Battle between preferences:

- Many people like **Max Likelihood** based methods:
  - sensitive at large evolutionary distances
- Often a **BEAST** tree is the answer
  - But takes computational time
  - Advantage of including complex models with priori assumptions

ML tree of life

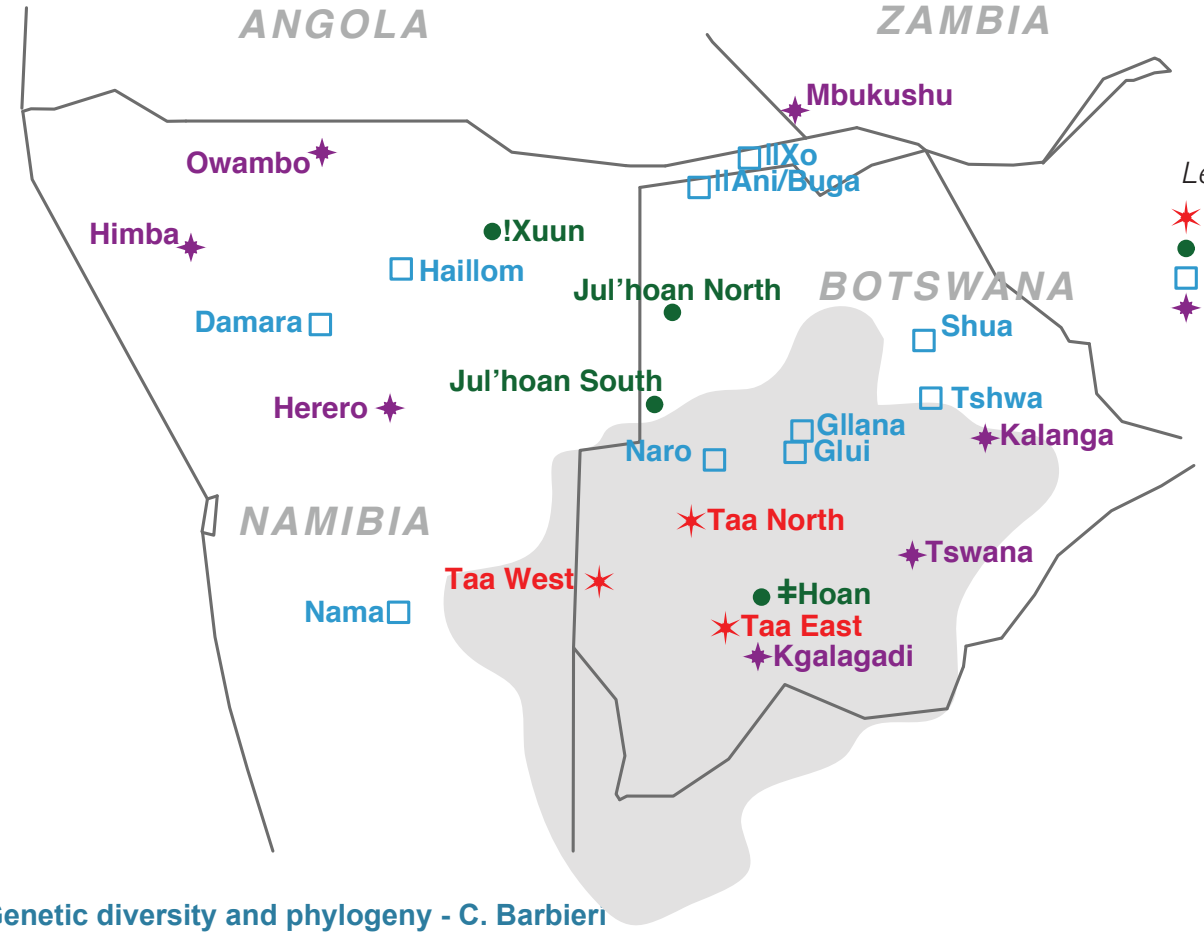


Hug et al. 2016 Nature Microbiology



Mount, 2008 Cold Spring Harbor Protocols

mtDNA genetic distances between populations from southern Africa Bantu and Khoisan speakers

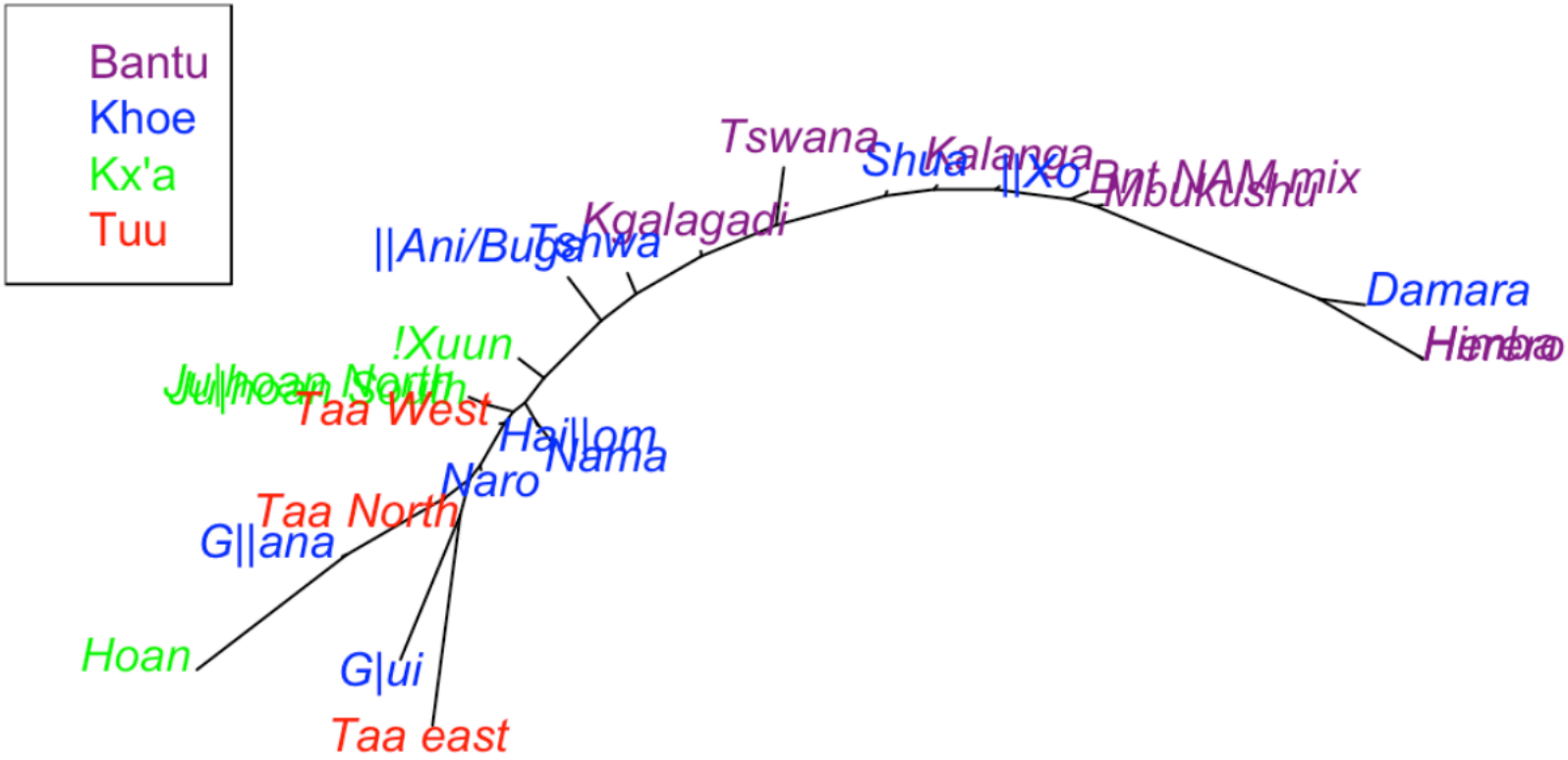


Legend: Language Family

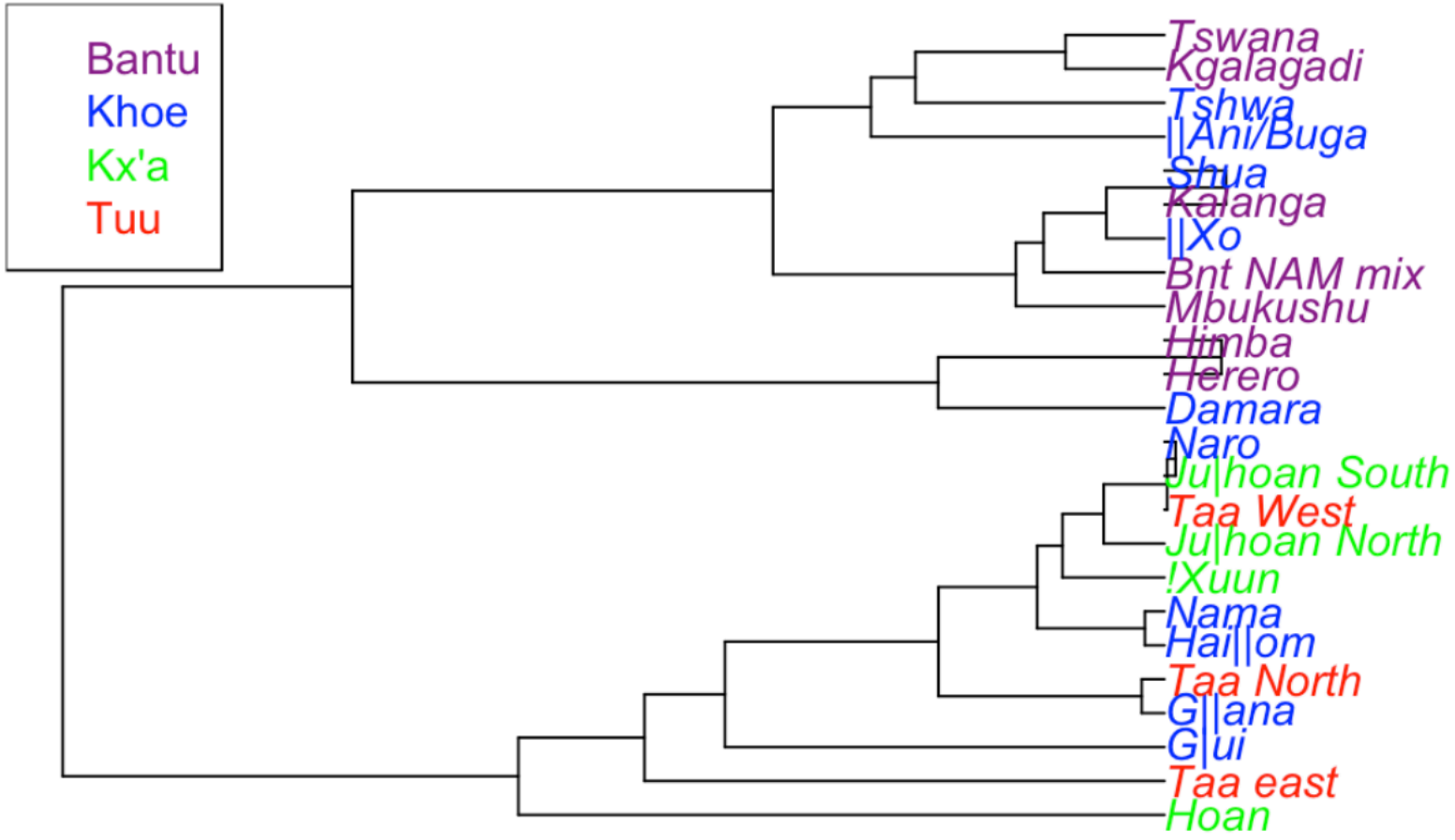
- Tuu
- Kx'a
- Khoe-Kwadi
- Bantu

Barbieri et al. 2014

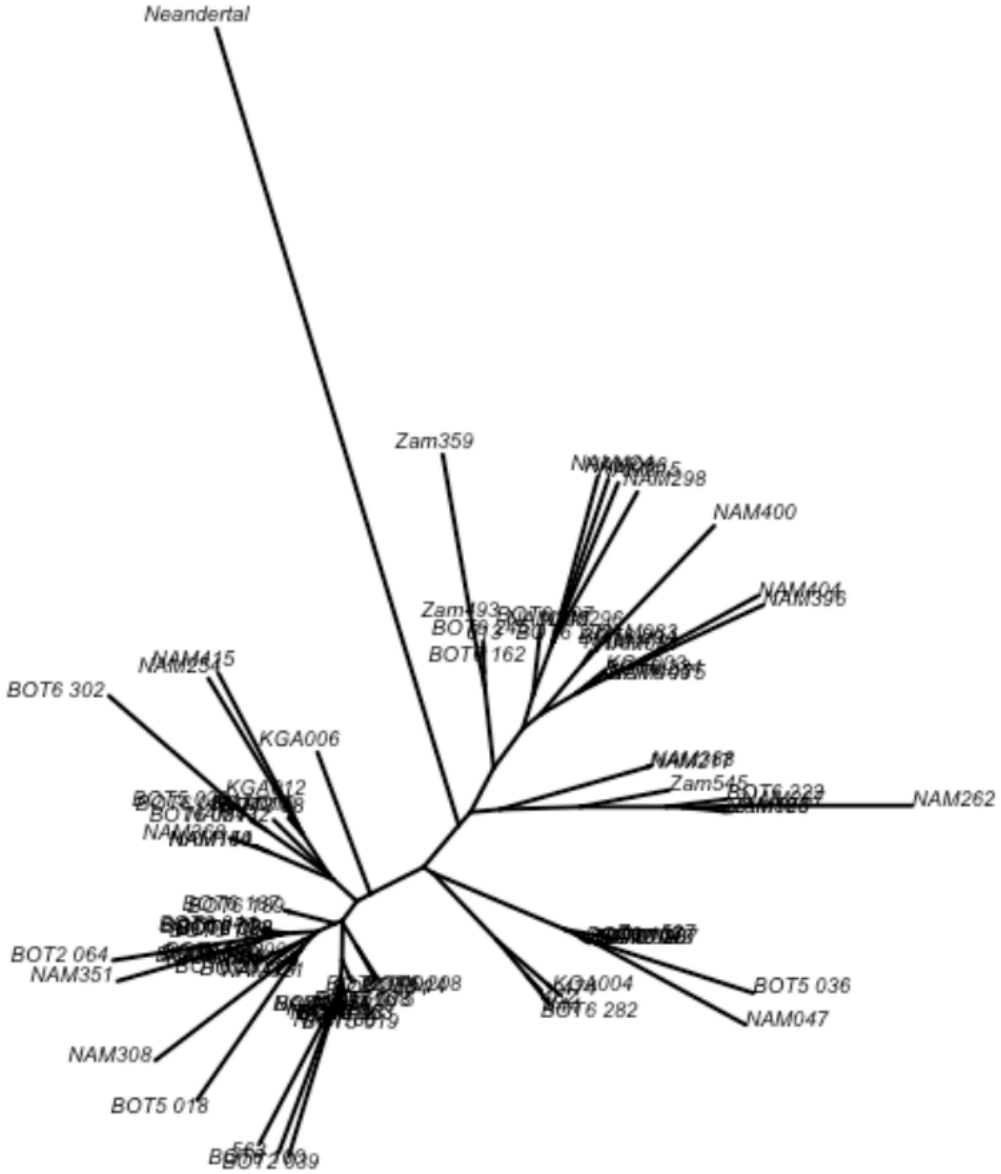
Neighbour Joining tree on population genetic distances (Fst)

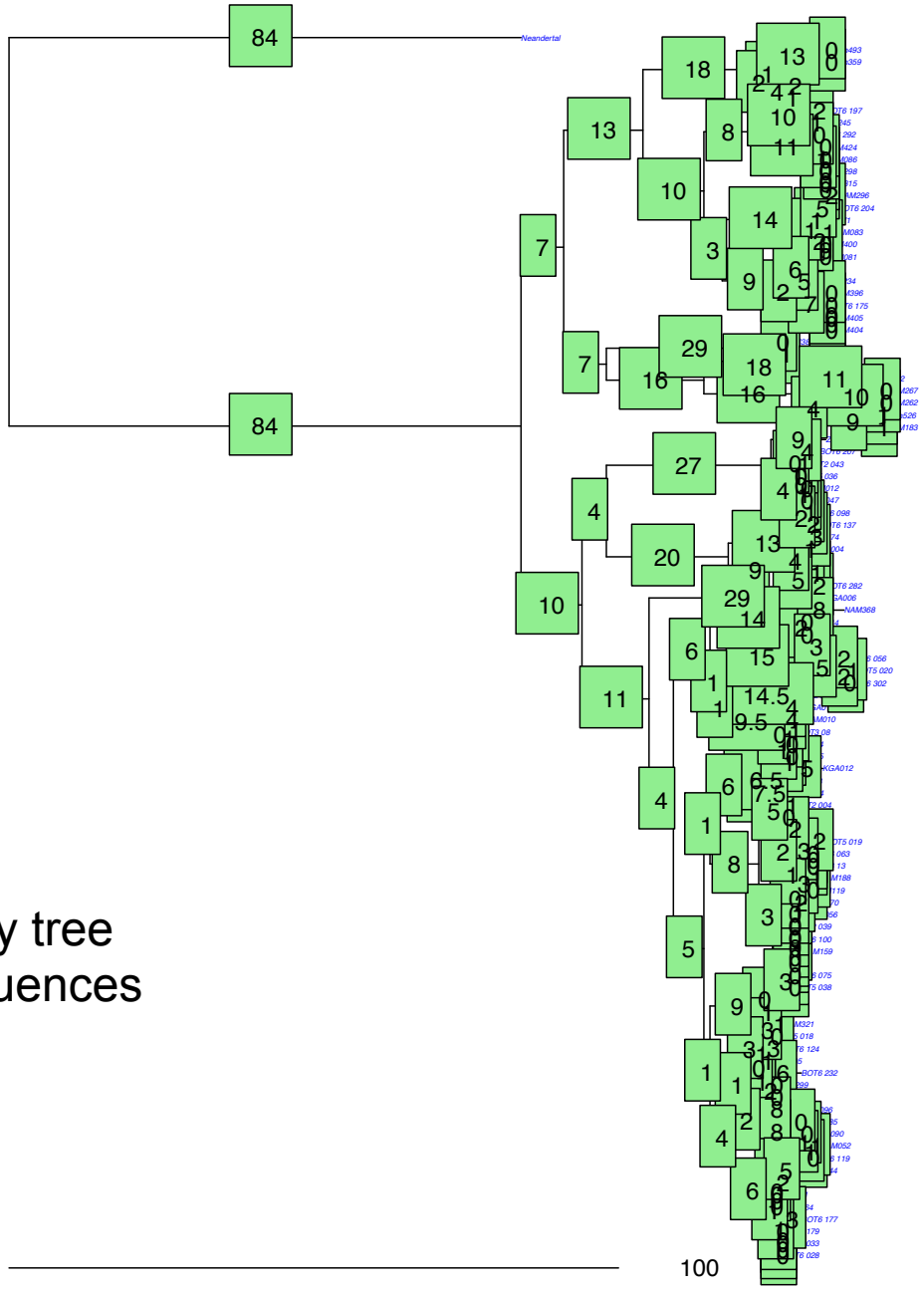


UPGMA tree on population genetic distances (Fst)



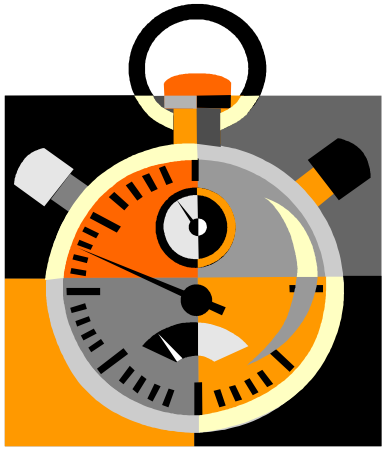
NJ tree on a set of 100 mtDNA sequences





Maximum parsimony tree
On 100 mtDNA sequences

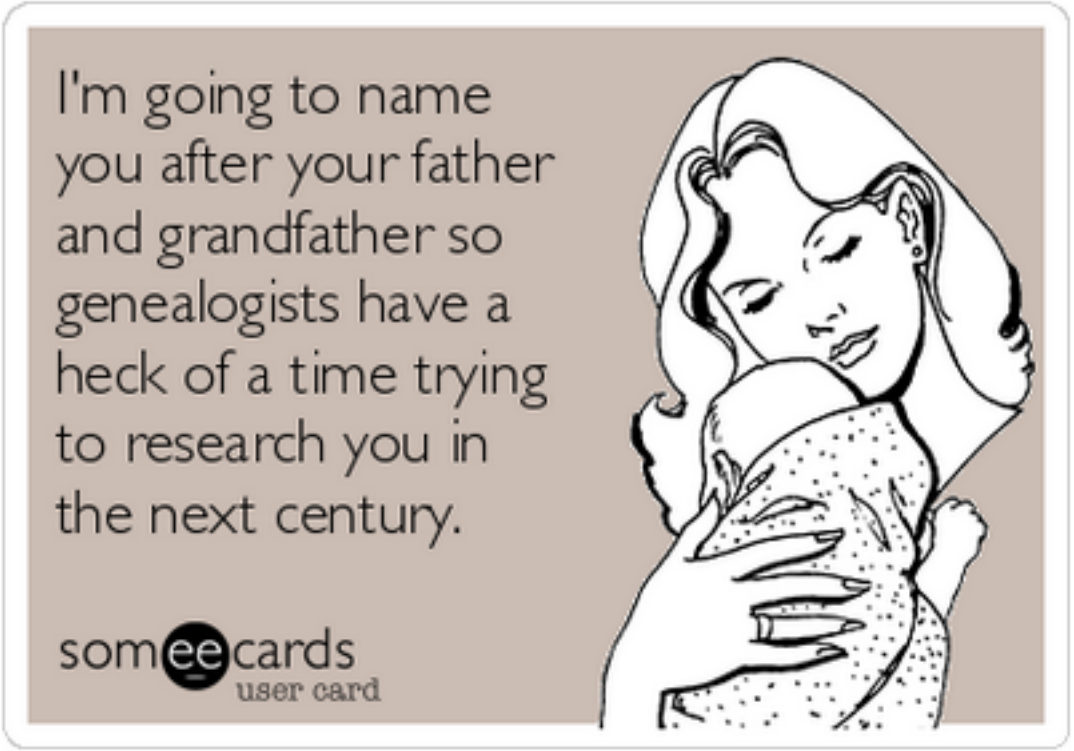
The Molecular Clock Hypothesis



- Amount of genetic difference between sequences is a function of time since separation.
- Rate of molecular change is constant (enough) to predict times of divergence

How to calibrate (in humans)

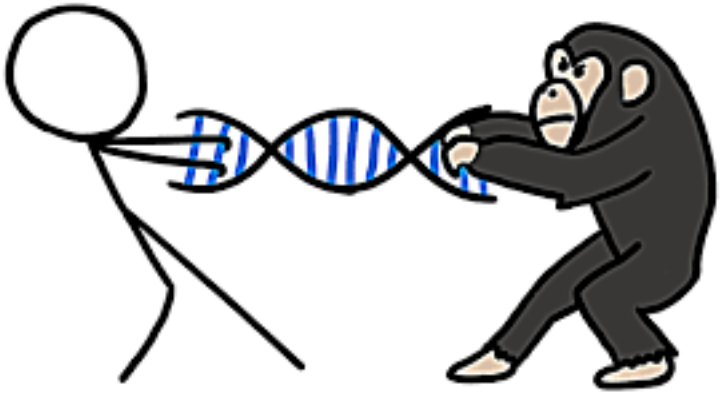
- Deep pedigree data
  - Count the mutations between n^{th} grade cousins



How to calibrate (in humans)

Calibrate nodes

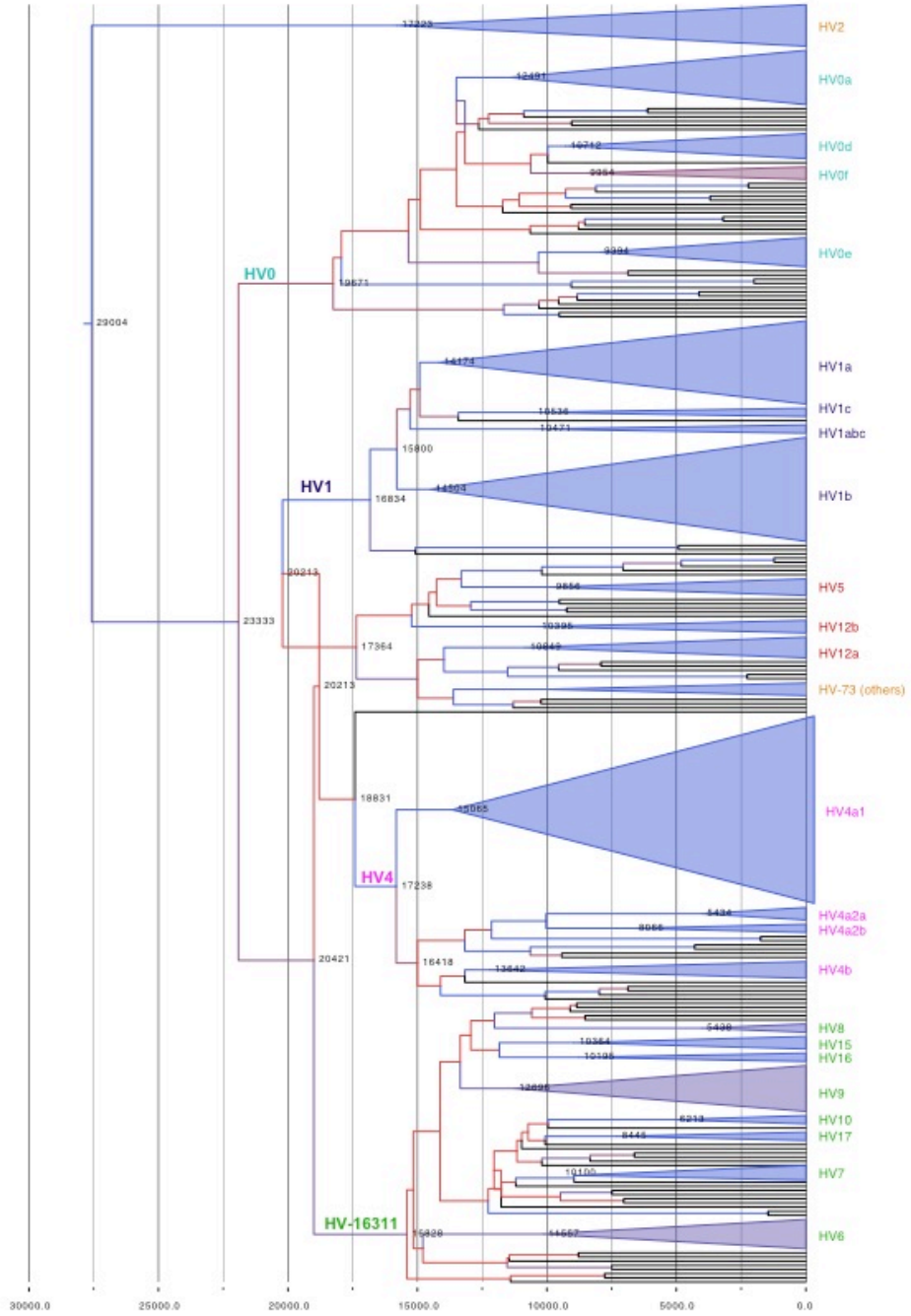
- Archaeological data
  - Species divergence (too old!)
    - *E.g Human-chimp split*
  - Historical events (too recent!)
    - *E.g Colonization of the pacific, specific lineage*



How to calibrate (in humans)

- Calibrate tips
- Direct calibration with inclusion of aDNA from properly dated fossils

S3 Figure. Phylogenetic tree with only the HV sequences, realized in BEAST.

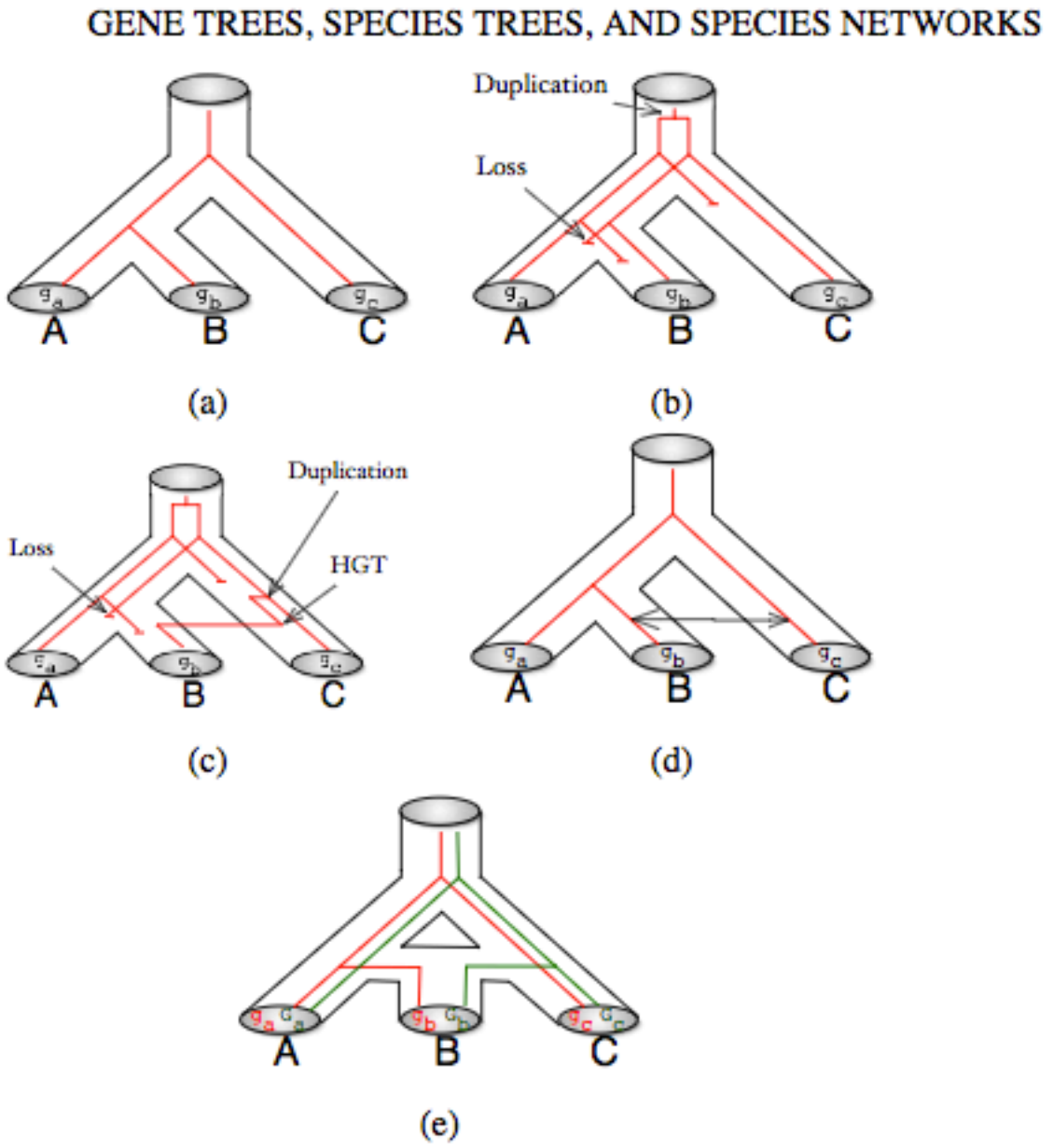


De Fanti, Barbieri 2016

Why trees might not work

- Noise: Data does fit a single tree, weak support is only a consequence of “noise”
- Trees in Trees: Data consists of multiple independent trees, genes and pops evolve treelike (e.g. incomplete lineage sorting, gene loss, gene duplication)
- Trees in Networks: Data consists of multiple independent trees, genes evolve treelike, pops don’t (e.g. hybridization, horizontal transfer)
- Reticulation: the data is not treelike

- A. gene tree agrees with species tree. B gene tree disagrees with species tree because of gene loss and duplication. C gene tree disagrees with species tree because of Horizontal Gene Transfer. D genetic material is exchanged between species B and C. E hybrid speciation results in two incongruent gene trees. (Nakhleh, Ruths, & Innan 2009).



Networks

Networks

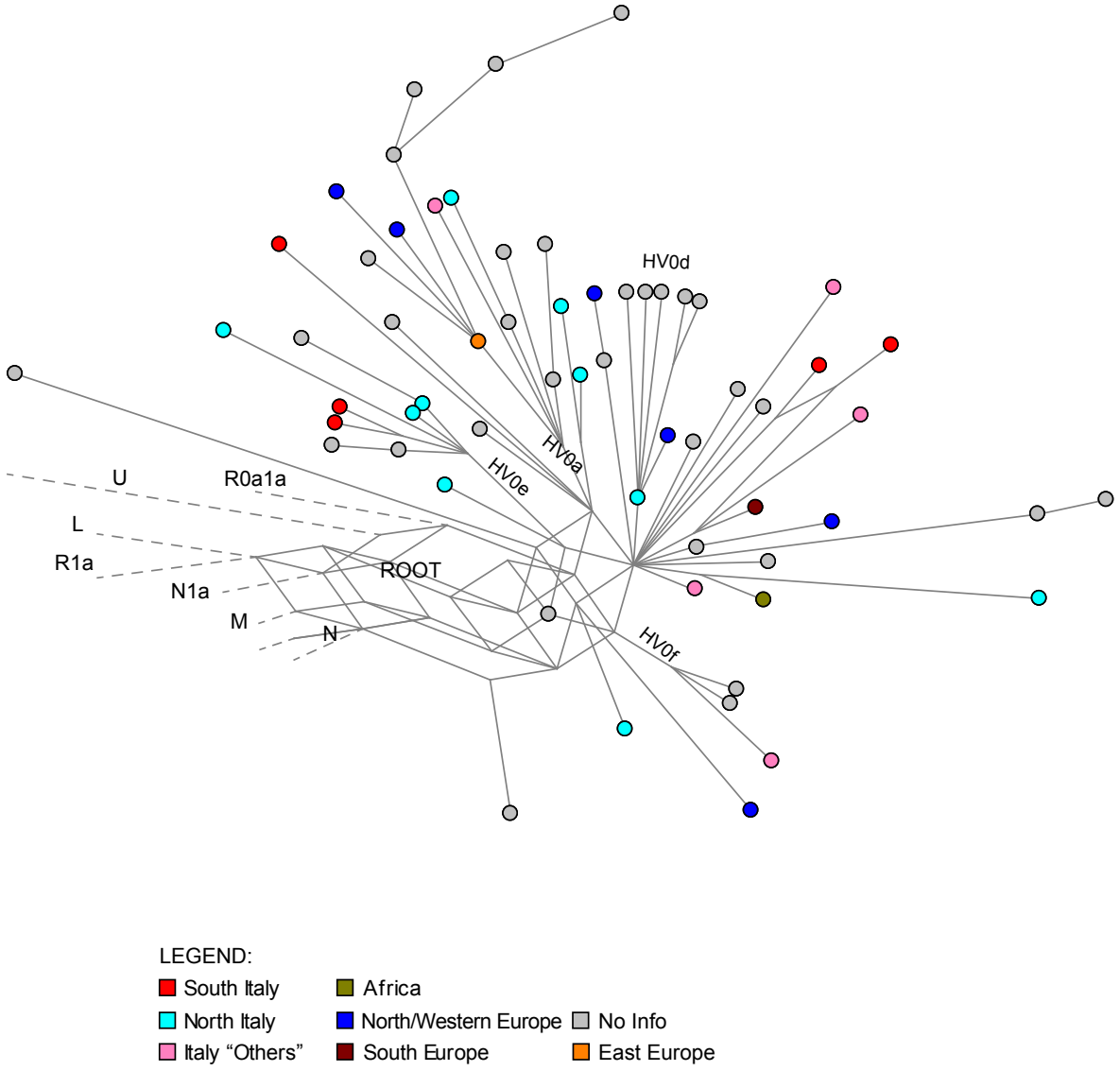
- Phylogenetic networks can be computed from multiple sequence alignments, distance matrices, sets of trees, clusters, splits, etc.
- Note: loss of evolutionary **direction**
- Note: the more tree-like the data are then the more tree-like will be the network

Median networks

- Character-based method (Bandelt, 1994 and Bandelt et al., 2000), usually applied to binary data.
- Simultaneously display all of the character-state differences among the taxa as separate branches in a network.
- Too much conflict can create undisplayable hypercubes

Use: Network http://www.fluxus-engineering.com/sharenet.htm

S6 Fig. Median-joining networks for major lineage blocks: Haplogroup HV0.
Mutations are given equal weights



De Fanti, Barbieri 2016

Splits network

Directly quantify the data incompatibilities and then try to display these incompatibilities, without ever explicitly inferring a tree.

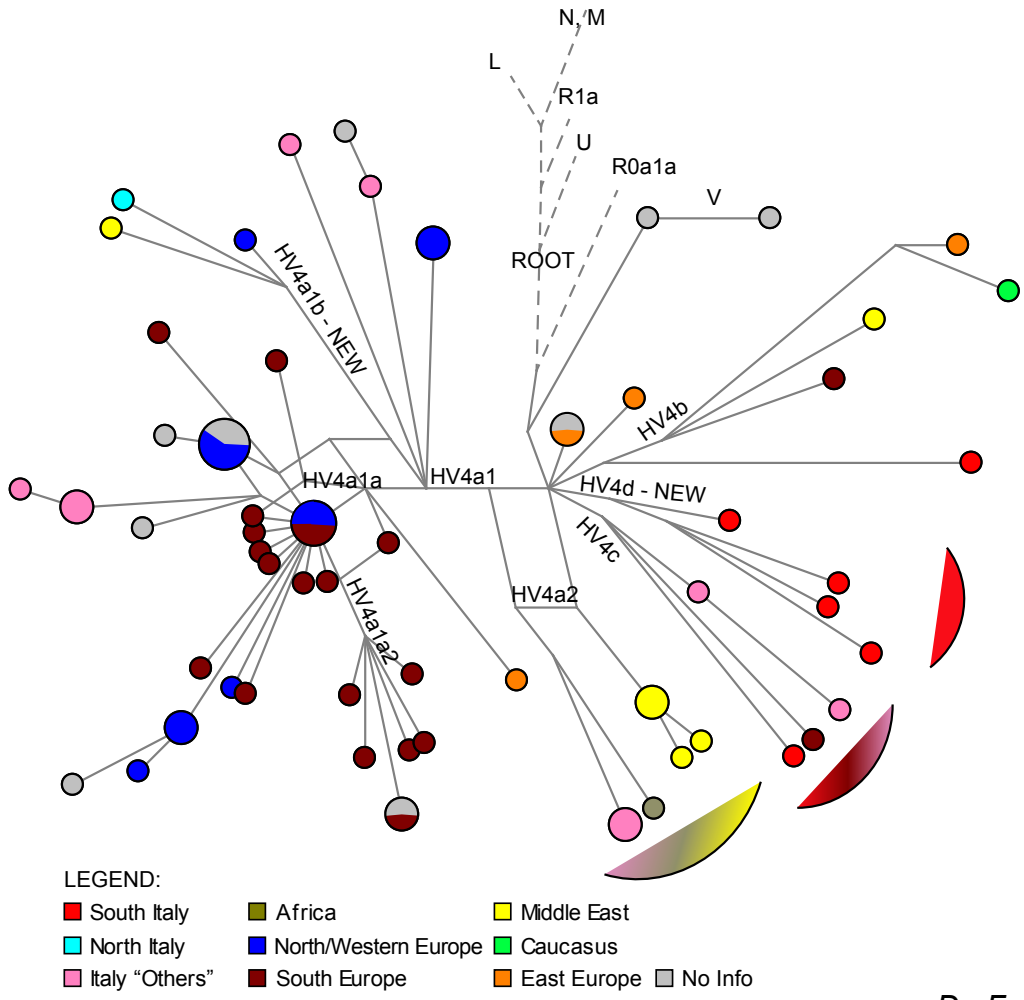
- Split decomposition: can be based on the raw data (called parsimony splits; Bandelt and Dress, 1993) or more usually on a distance measure (Bandelt and Dress, 1992).
- Neighbour-Net: distanced-based method (Bryant and Moulton, 2002 , 2004)
  - compromise between the preponderance of apparent false positives in median networks and the false negatives of split decomposition

Use: SplitsTree

Simplify a network

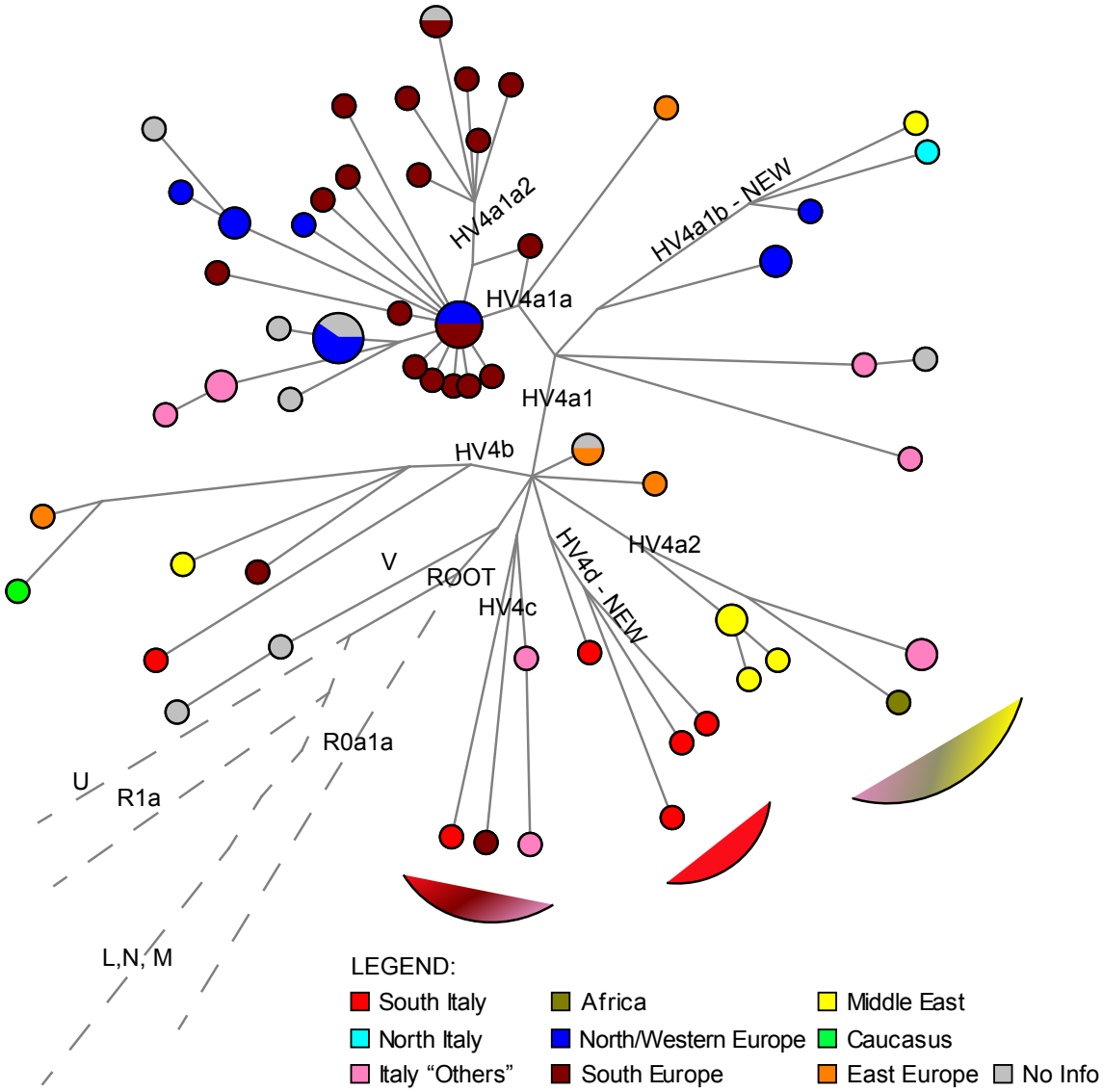
- For less reticulations, apply **weights** to the characters.
  - Positions with recurrent change of state are downweighted
- The software Network allows a few tricks: Reduced Median networks, post processing, star contraction

S5 Figure. Median-joining networks for major lineage blocks: Haplogroup HV4. Mutations are given equal weight.



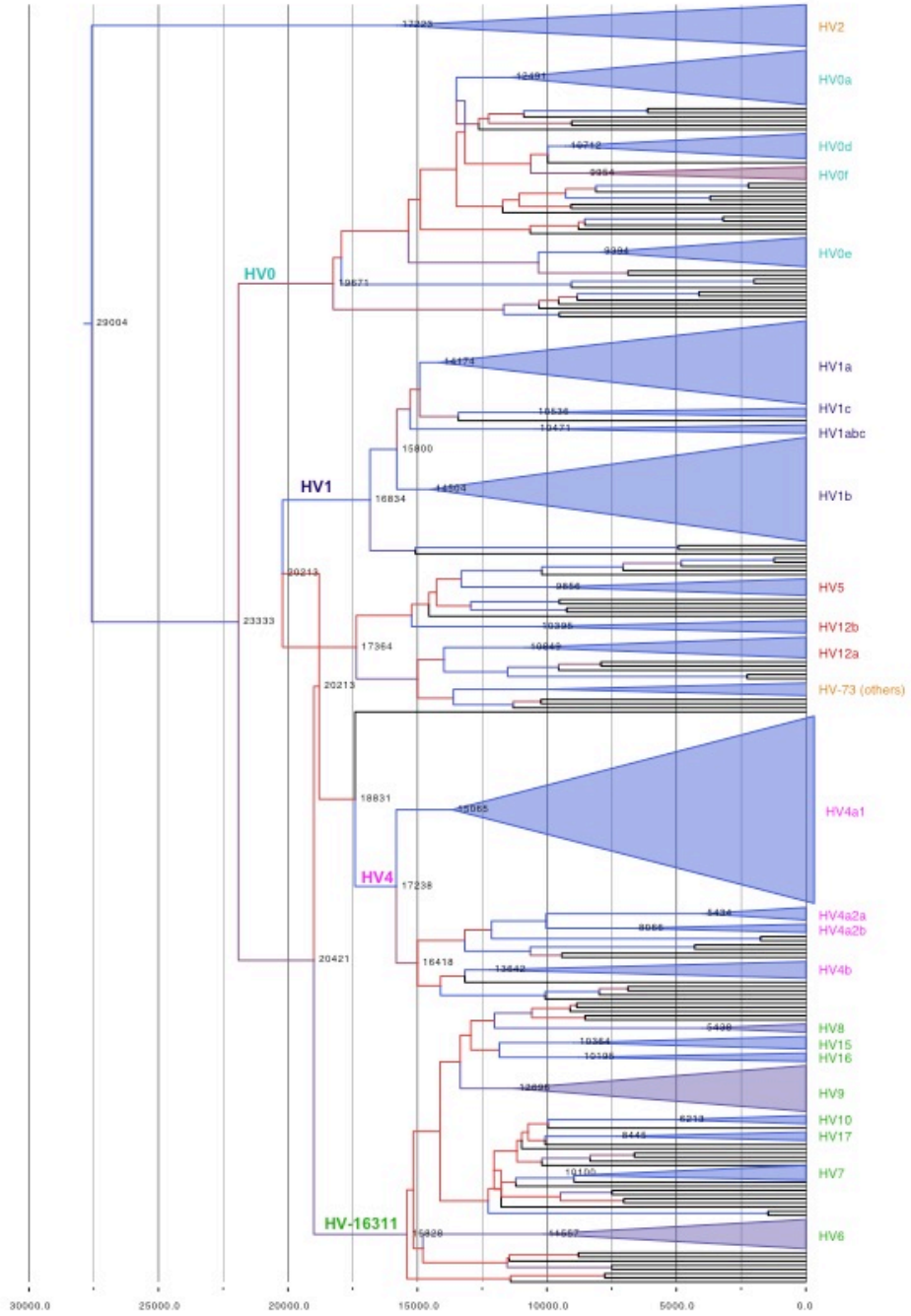
De Fanti, Barbieri 2016

Weighted network: resolve reticulations



De Fanti, Barbieri 2016

S3 Figure. Phylogenetic tree with only the HV sequences, realized in BEAST.



De Fanti, Barbieri 2016

Summarizing

Phylogenetic relationships: networks with an evolutionary meaning

- Trees
  - Based on distances (UPGMA, NJ) or character state (MP, ML, Bayesian MCMC)
  - Rooted with an outgroup
  - Find the best compromise between multiple possible reconstructions
- Networks: visualize all possible paths, allow reticulations and hypercubes
  - Loss of evolutionary direction

Practical session in R:

1. Import a mtDNA alignment

- Create a matrix of genetic distance between sequences
- Visualize it with a NJ tree
- Make a rooted MP tree

2. Import a matrix of genetic distance between populations

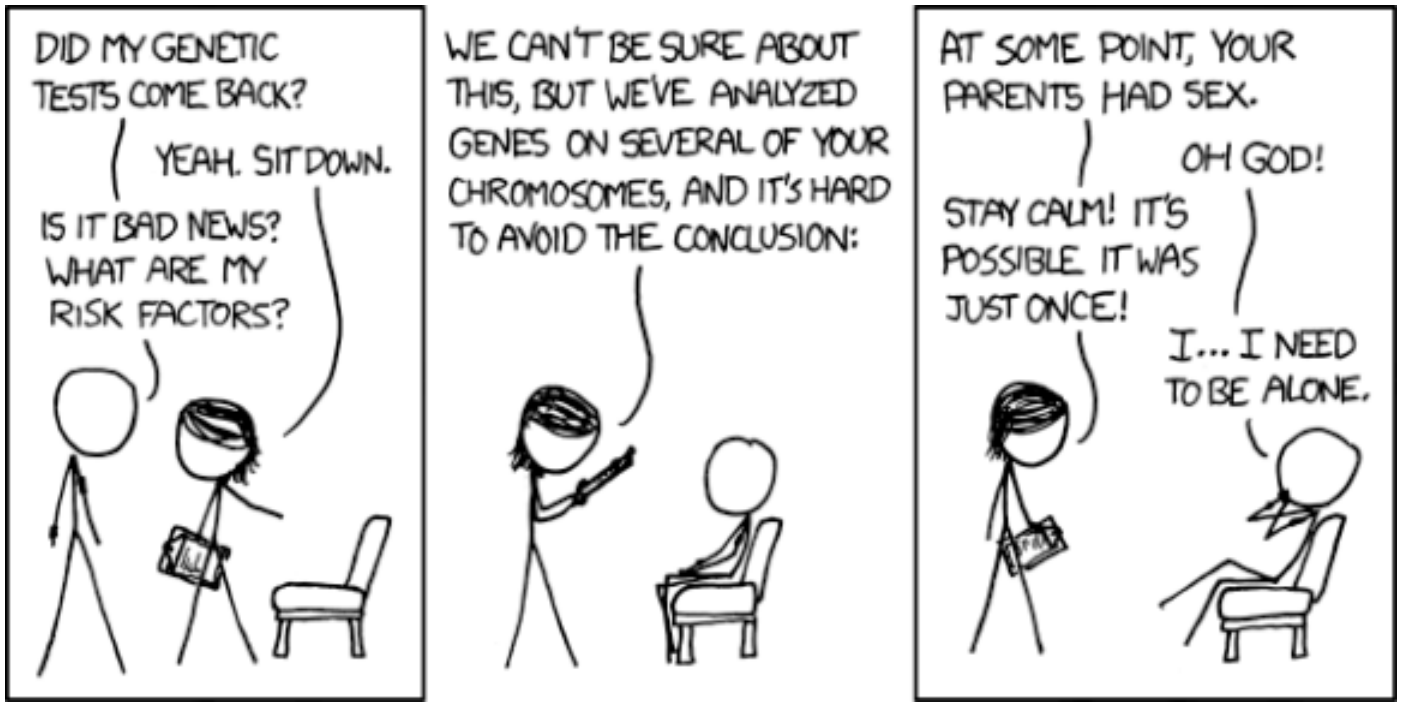
- Visualize relationships with a NJ tree and a UPGMA tree

Resources

- A good review of molecular trees: Yang, Z., & Rannala, B. (2012). Molecular phylogenetics: principles and practice. *Nature Reviews Genetics*, *13*(5), 303-314.
- Mount, D. W. (2008). Choosing a method for phylogenetic prediction. *Cold Spring Harbor Protocols*, *2008*(4), pdb-ip49.
- Morrison, D.A., 2005. Networks in phylogenetic analysis: new tools for population biology. *International journal for parasitology*, *35*(5), pp.567-582.
- http://ab.inf.uni-tuebingen.de/talks/pdfs/Phylogenetic%20Networks%20-%20GCB2006.pdf unfortunately in Comic Sans

Figures from:

- https://www.cs.princeton.edu/~mona/Lecture/phylogeny-slides.pdf
- www.cs.cmu.edu/~roseh/Slides/**durand**03-**molclock**.**ppt**


DID MY GENETIC TESTS COME BACK?
YEAH. SIT DOWN.
IS IT BAD NEWS? WHAT ARE MY RISK FACTORS?
WE CAN'T BE SURE ABOUT THIS, BUT WE'VE ANALYZED GENES ON SEVERAL OF YOUR CHROMOSOMES, AND IT'S HARD TO AVOID THE CONCLUSION:
AT SOME POINT, YOUR PARENTS HAD SEX.
OH GOD!
STAY CALM! IT'S POSSIBLE IT WAS JUST ONCE!
I... I NEED TO BE ALONE.