Genetic diversity: Phylogeny and tree building

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

Genetic data

- Panels of population diversity

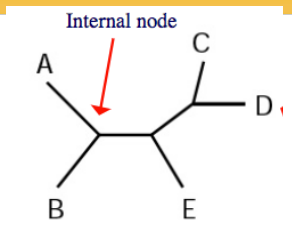
Exercise

Why population genetics

- Molecular anthropology and linguistics

Trees and networks

- How to build a tree



Building trees to understand evolution

Why study population genetics

Genetic contribution to diversity and prehistory

- DNA polymorphisms are transmitted vertically
- DNA polymorphisms are relatively stable through time and space

→DNA polymorphisms can retain traces of the demography at both individual and population level

Study population demographics

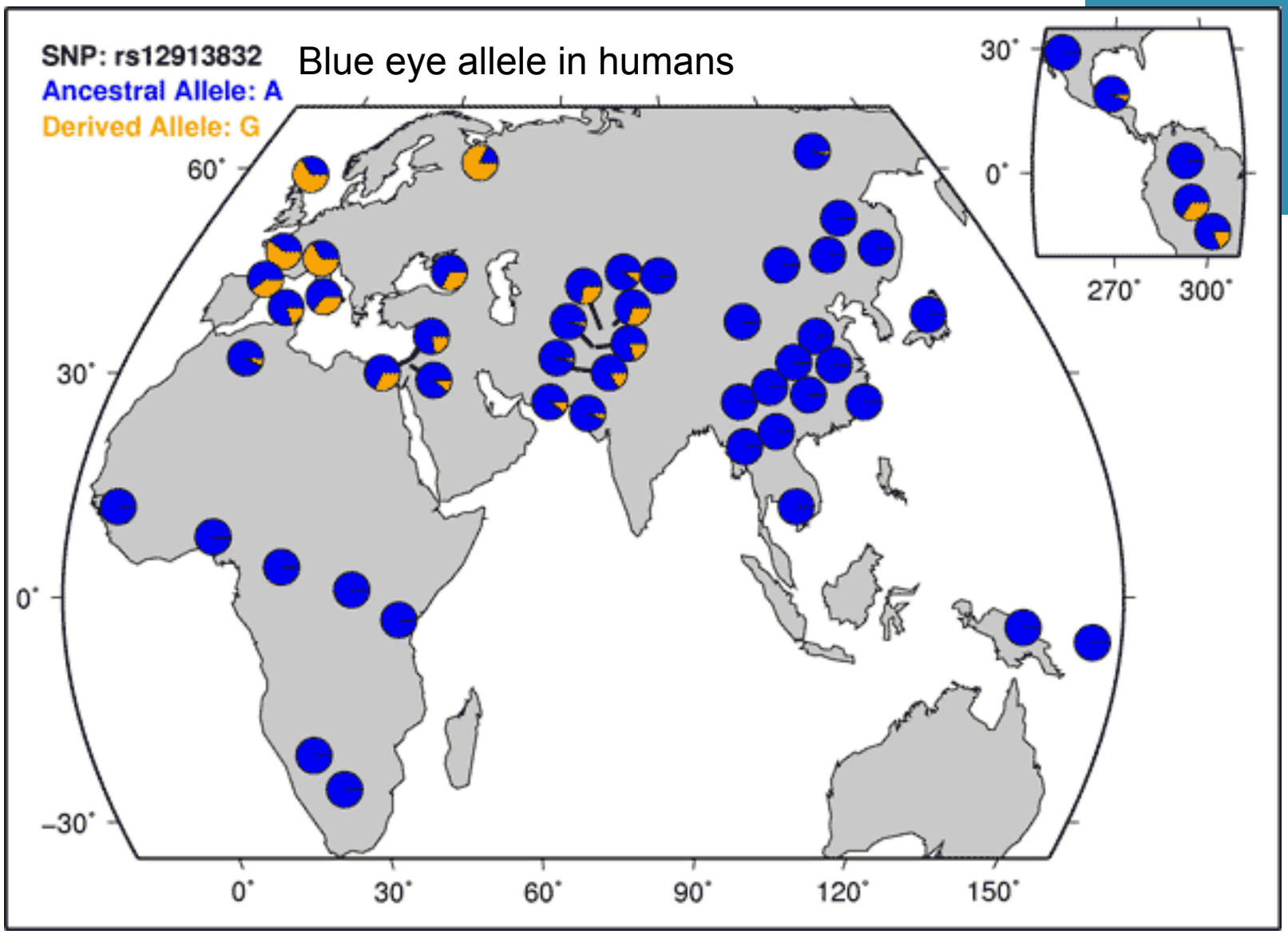
Population genetics:

- Study of genetic variation within a population and between populations

Phylogenetics:

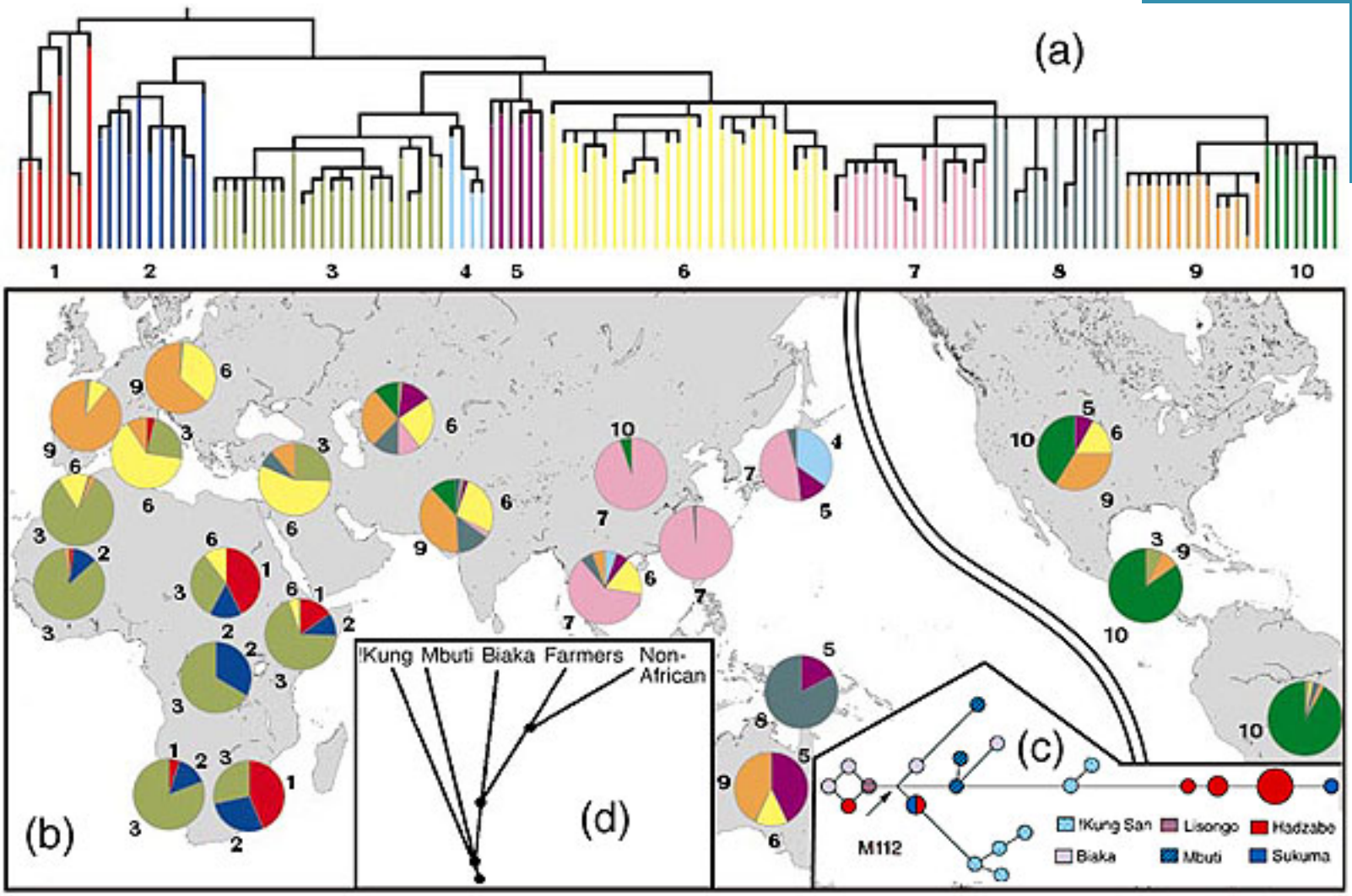
- Use genetic variation between taxa (species, populations) to infer evolutionary relationships

Present day variation



(HGDP Selection Browser)

Evolutionary history

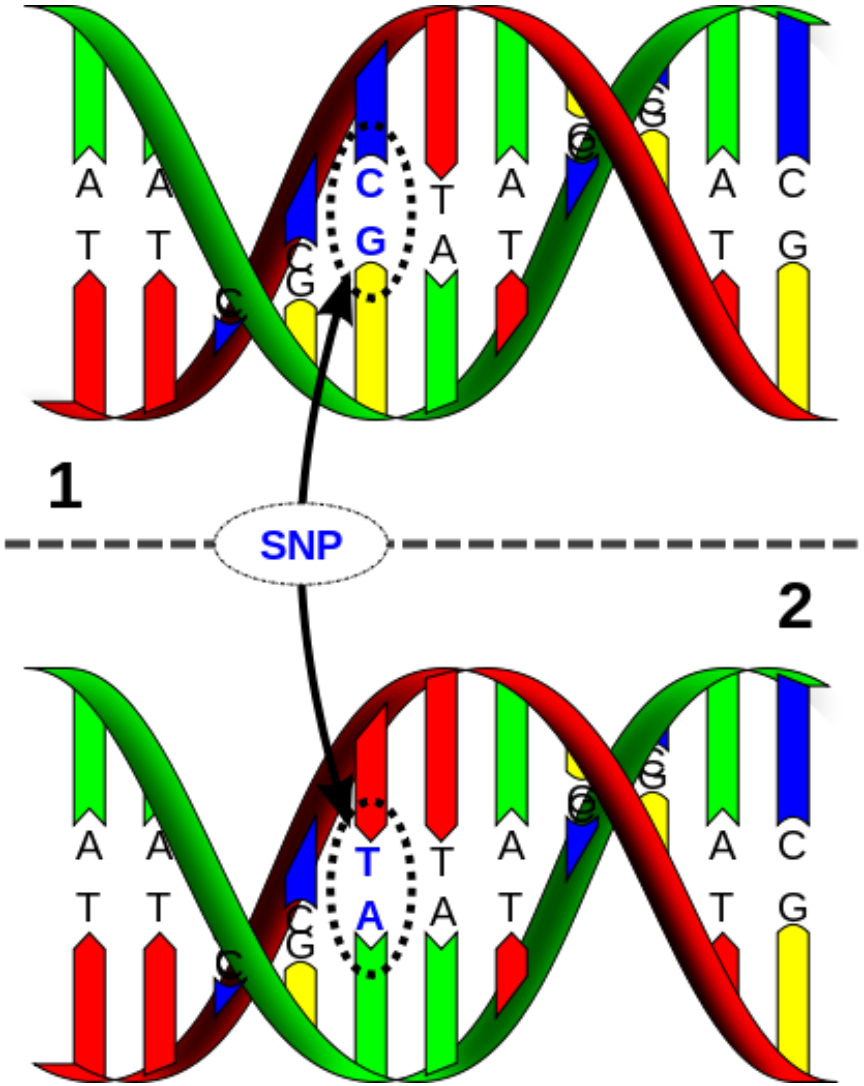


Y chromosome phylogeny

(Underhill et al. 2001)

Molecular markers

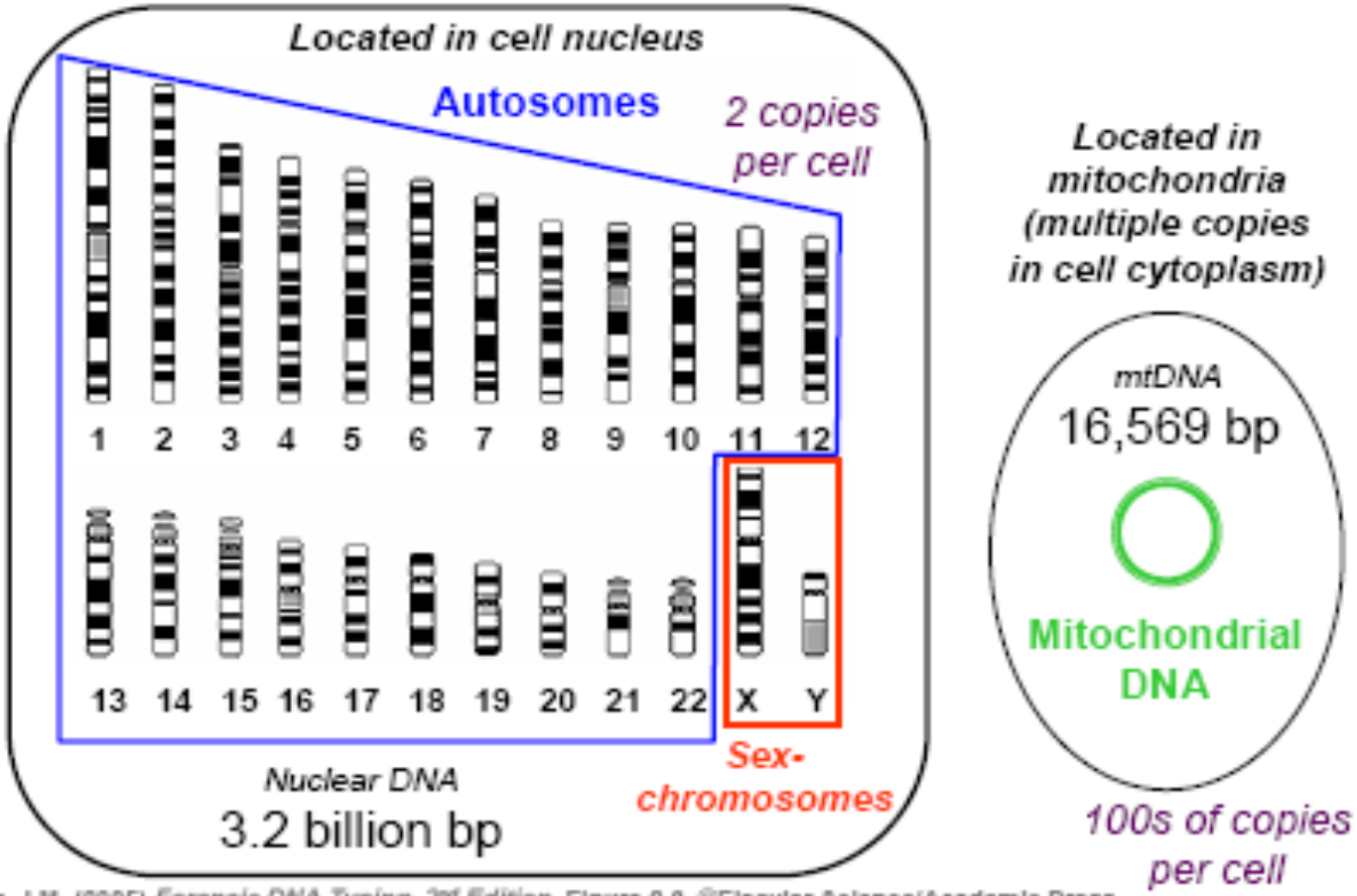
DNA sequence data



A mutation (SNP) is a change of one base in the sequence

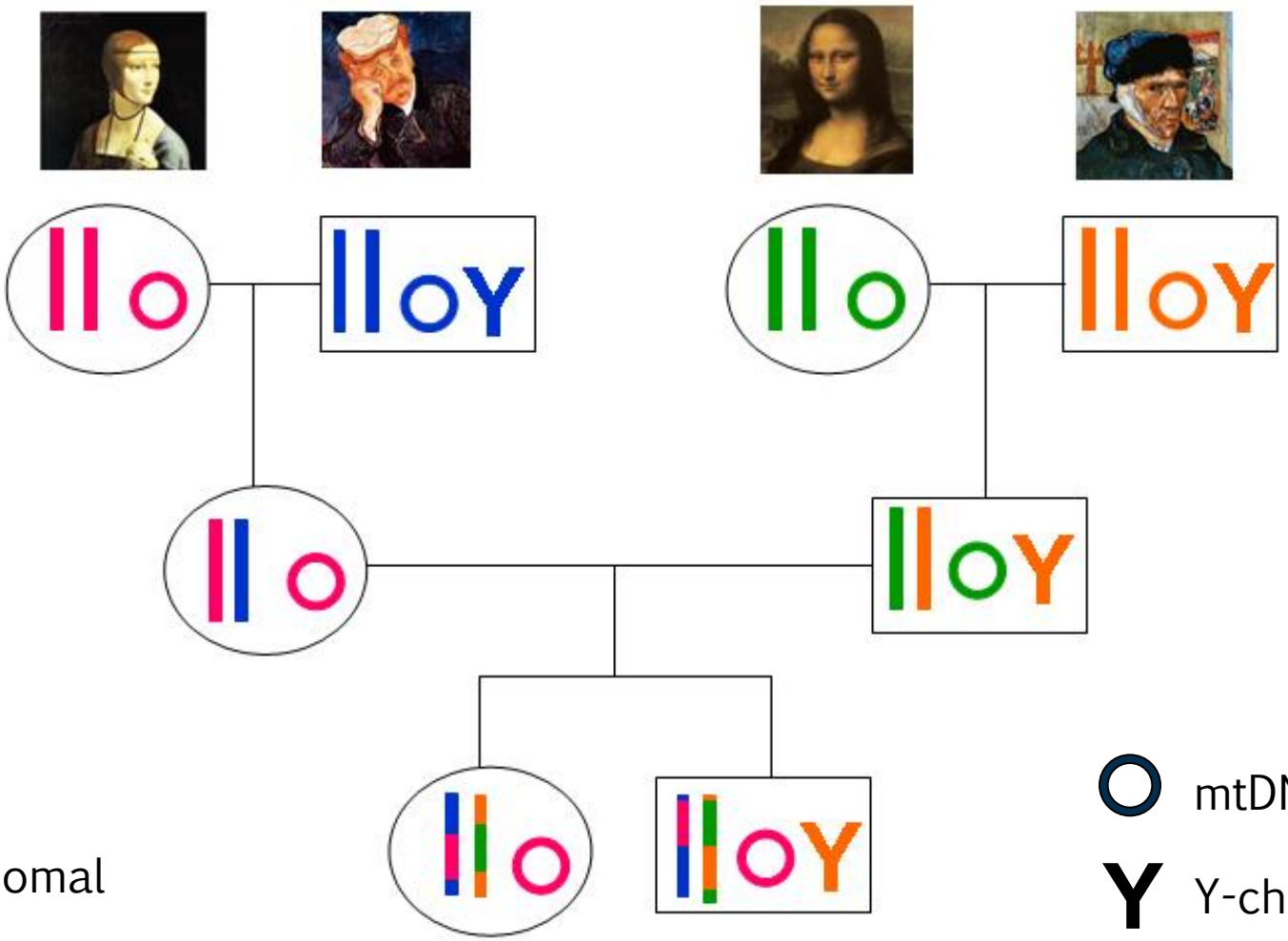
The human genome

- 23 pairs of chromosomes + mtDNA

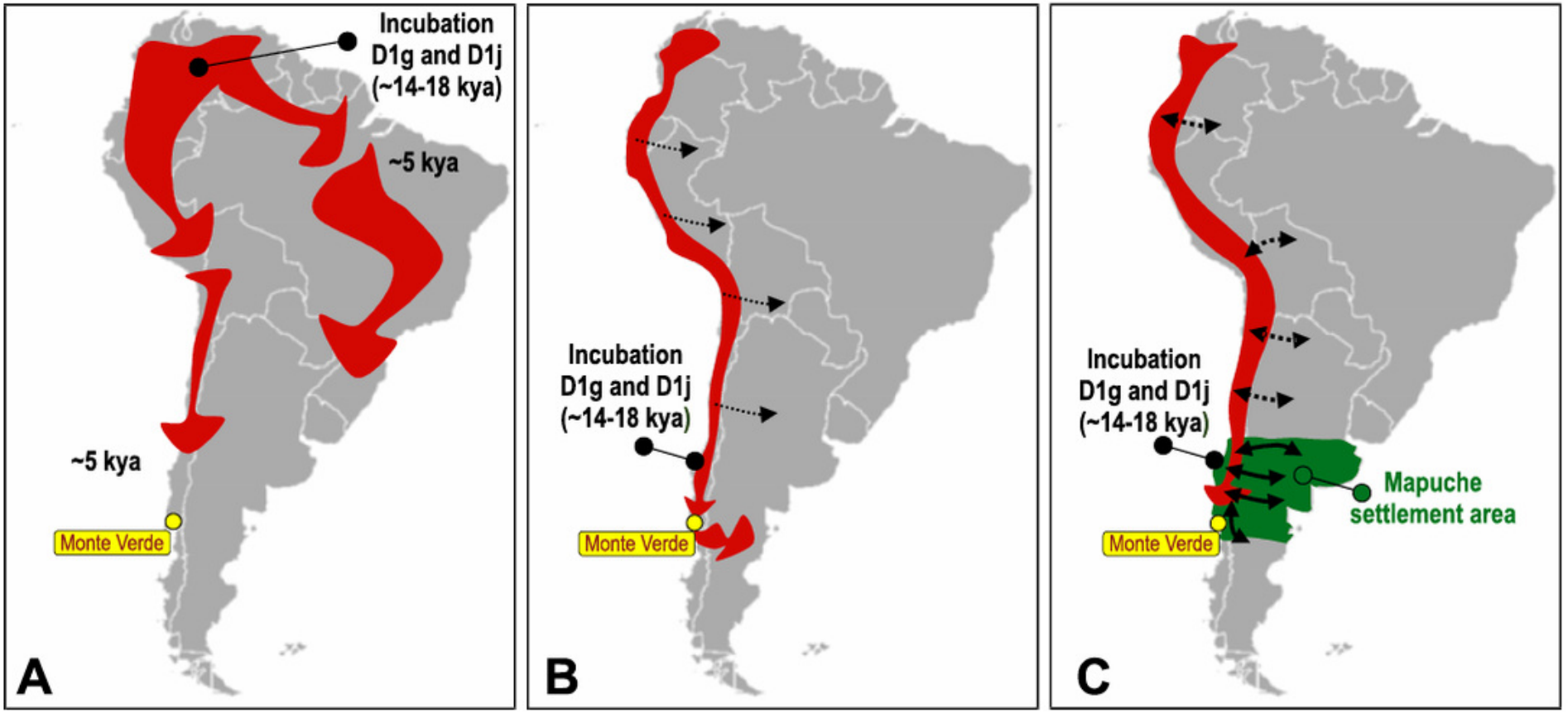


Butler, J.M. (2005) *Forensic DNA Typing, 2nd Edition*, Figure 2.3, ©Elsevier Science/Academic Press

Genetic markers: transmission



Tracing lineages through time and space (phylogeography)



Bodner et al. 2012 Genome Research

Phylogeny of the lactase allele in southern Africa

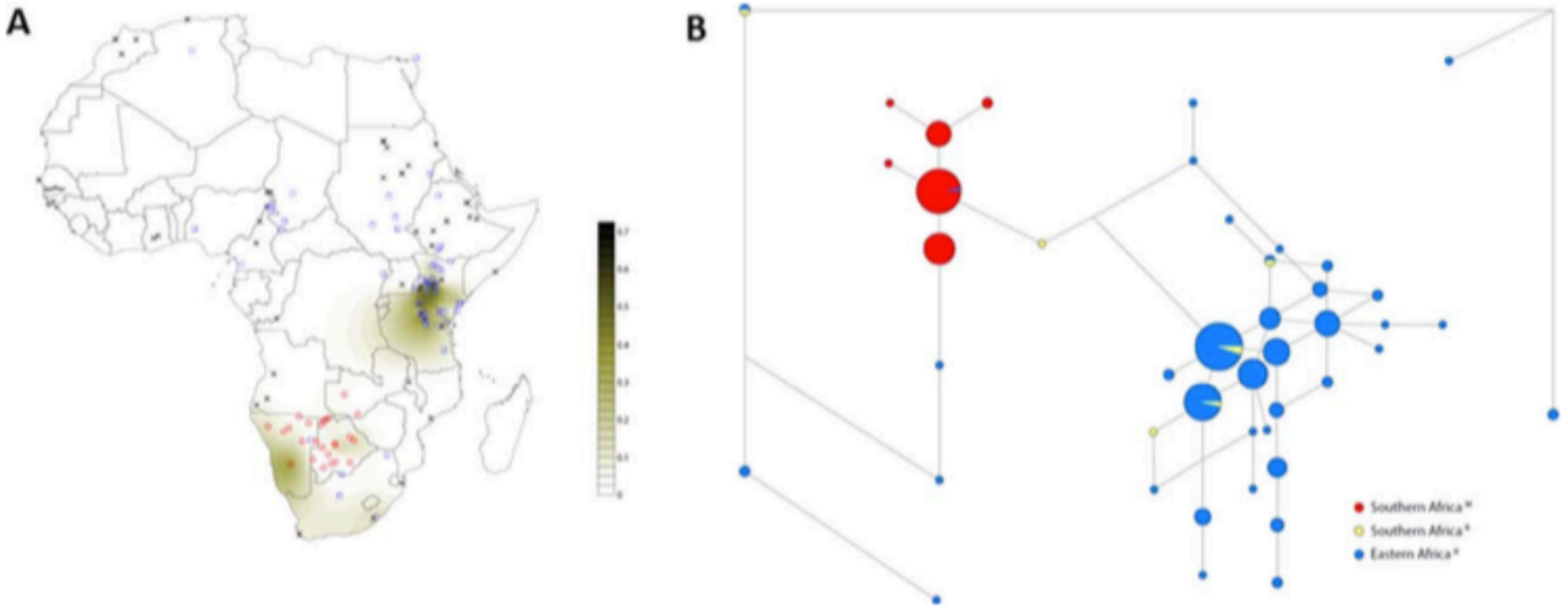
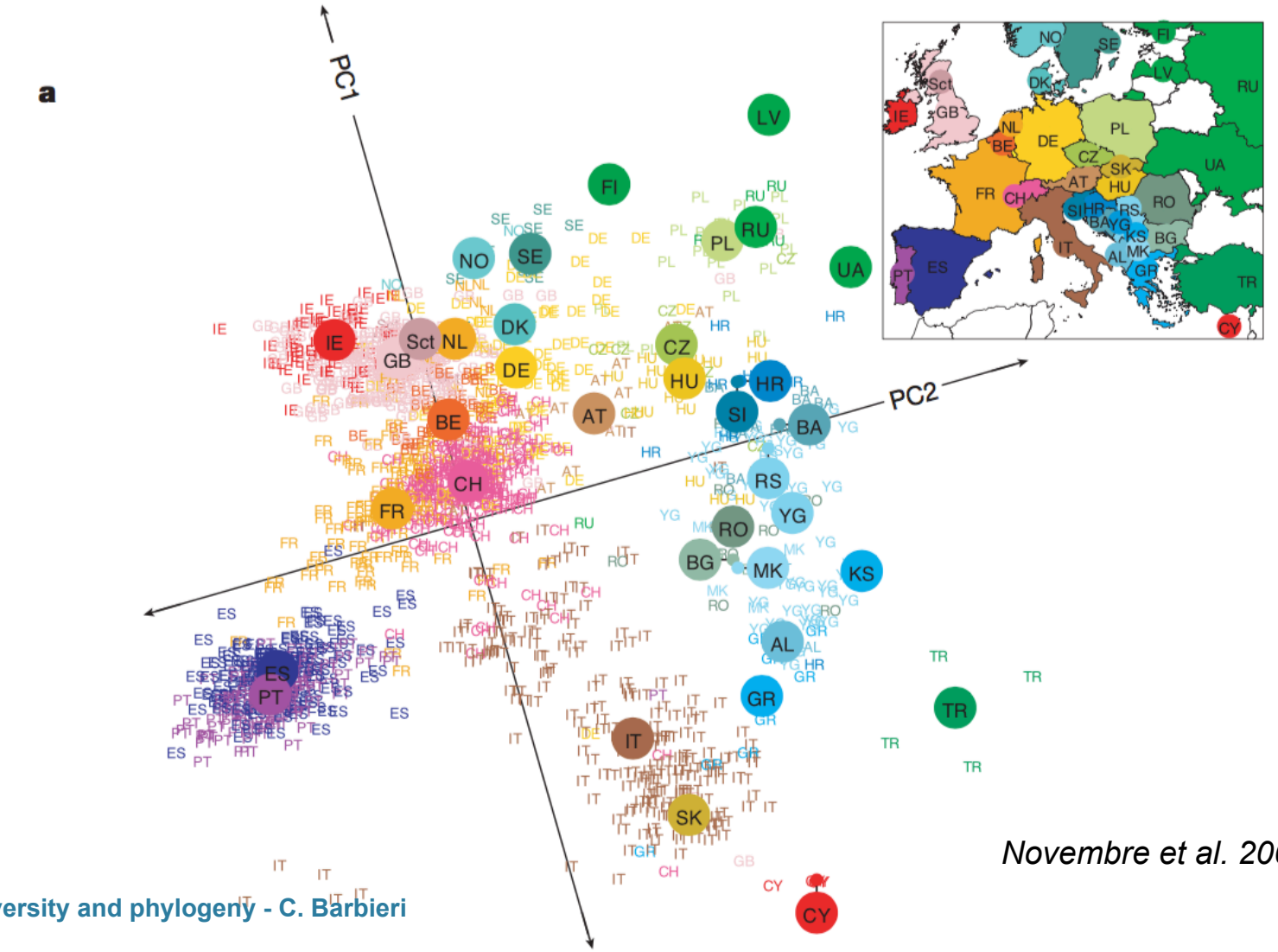


Fig. 1. Analyses of the C-14010 LP variant and associated STR haplotypes in Eastern and Southern African populations. **A**: Surfer map of the C-14010 allele frequency. Red circles denote sampling locations by Macholdt et al.; blue squares denote sampling locations by Ranciaro et al.; black crosses denote data from other published studies (taken from Macholdt et al. and Ranciaro et al.). **B**: Median-joining network of haplotypes associated with the C-14010 variant, based on four STR loci that flank the LP enhancer region. The M superscript denotes data by Macholdt et al., and the R superscript denotes data by Ranciaro et al. [Color

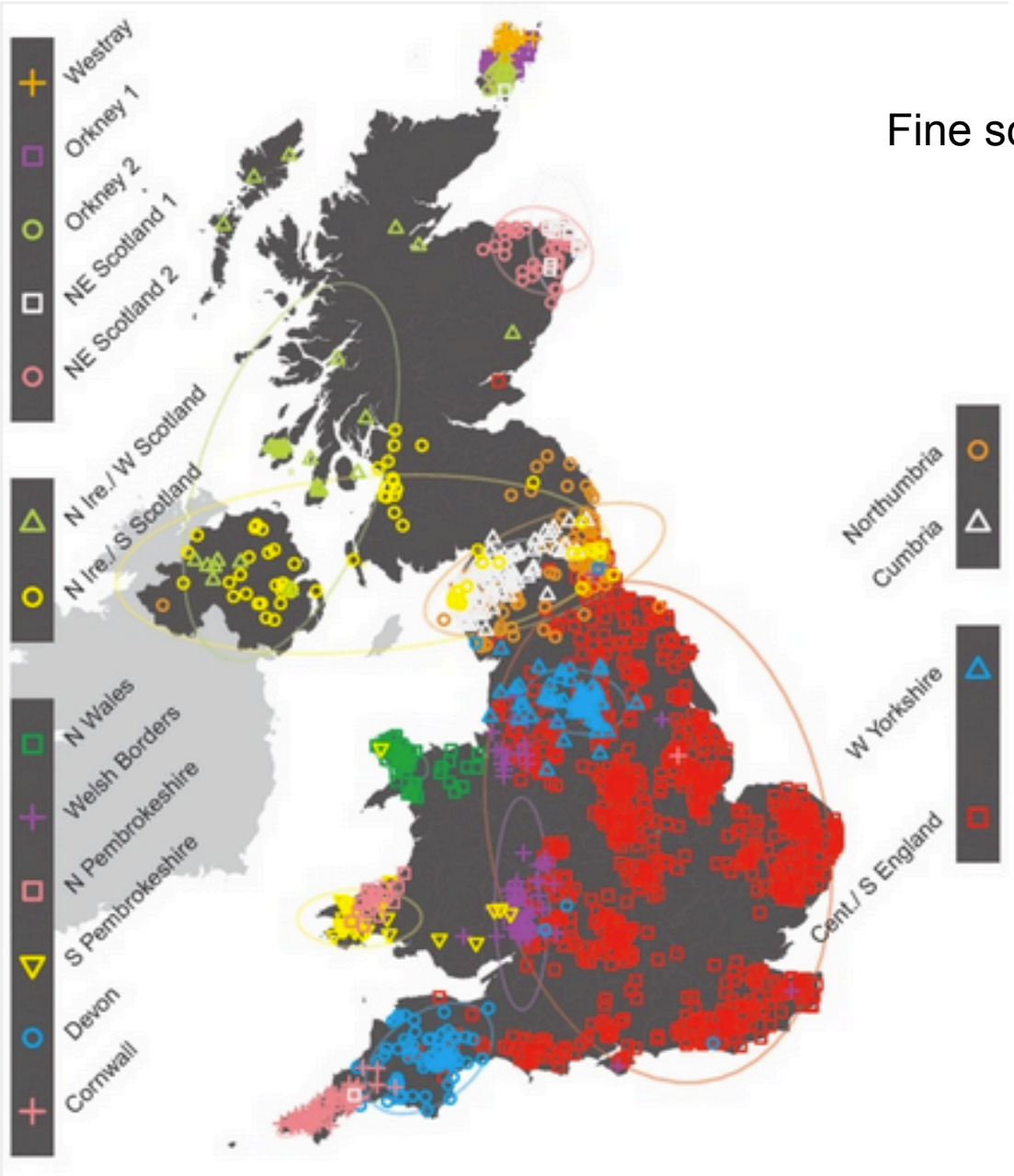
Macholdt et al. 2015 Am J Phys Anthropol

Autosomal data: Principal Component Analysis



Novembre et al. 2008 Nature

Fine scale PCA



Stephen Leslie

A map of the United Kingdom shows how individuals cluster based on their genetics, with a striking relationship to the geography of the country.

Leslie et al. 2015 Nature

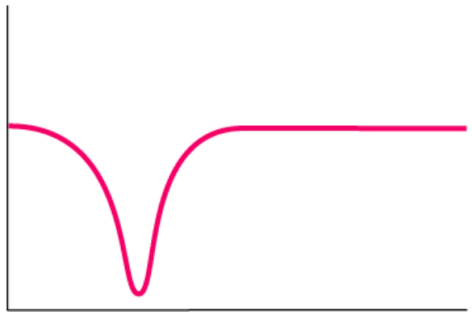
Reconstruct changes in population size

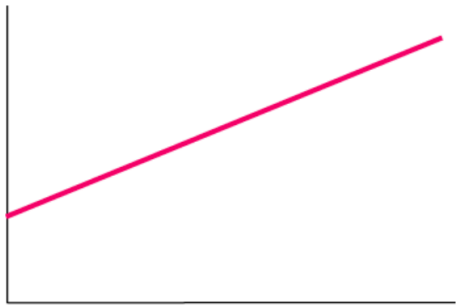
Changes to ancestral population sizes are of interest to evolutionary biologists.

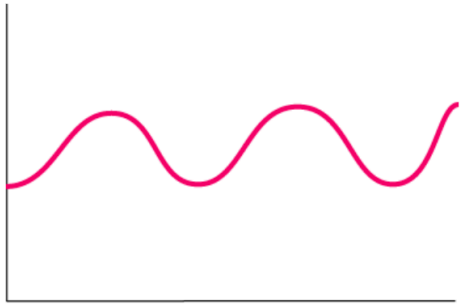
- Changes in population size induce changes in the coalescent rate
- Increasing population: more coalescent events branch differentiation in the tree.

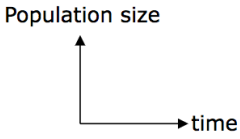
Inference of ancestral population history

We can use this method for any model of population size change that can be integrated with respect to t









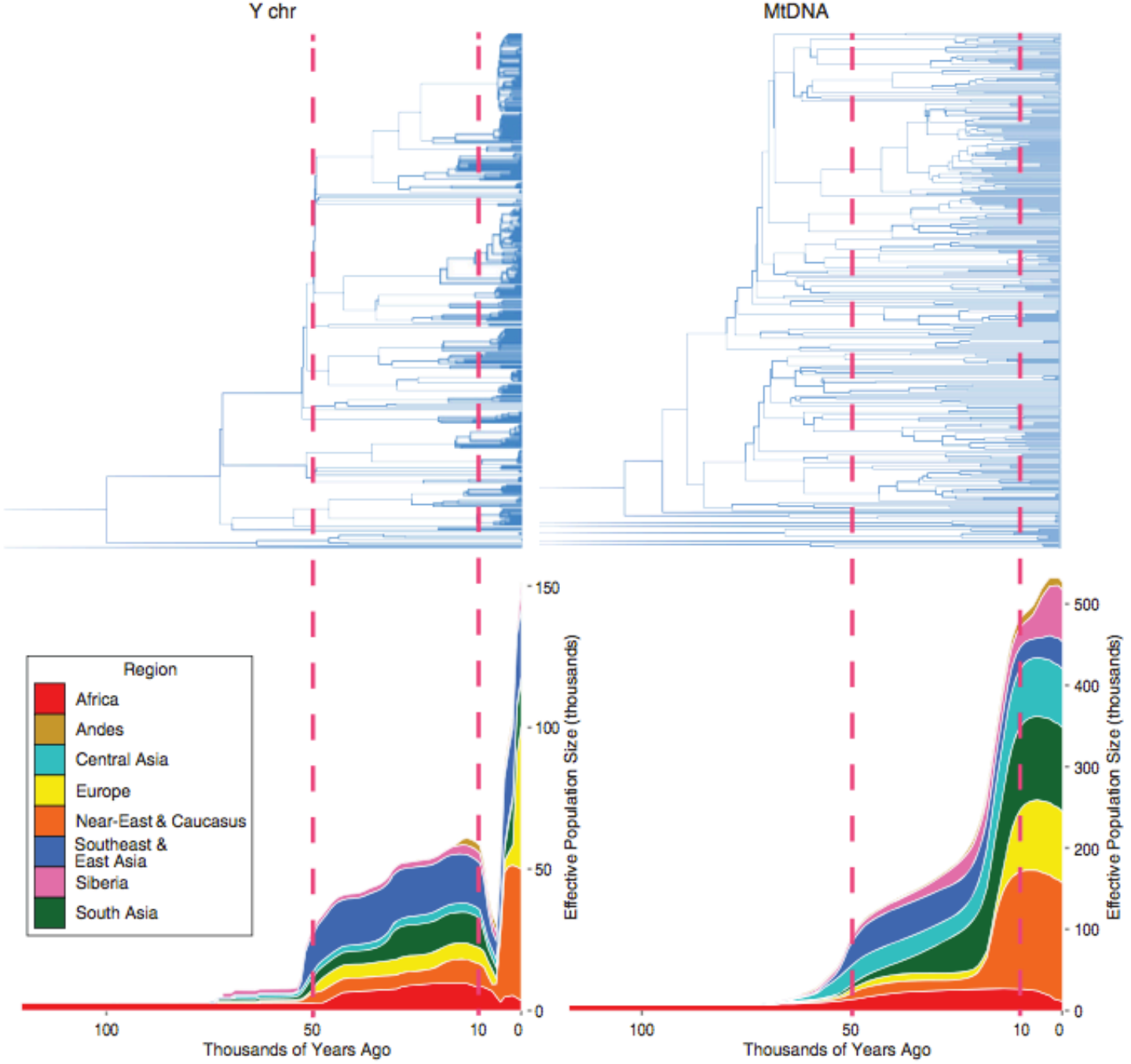
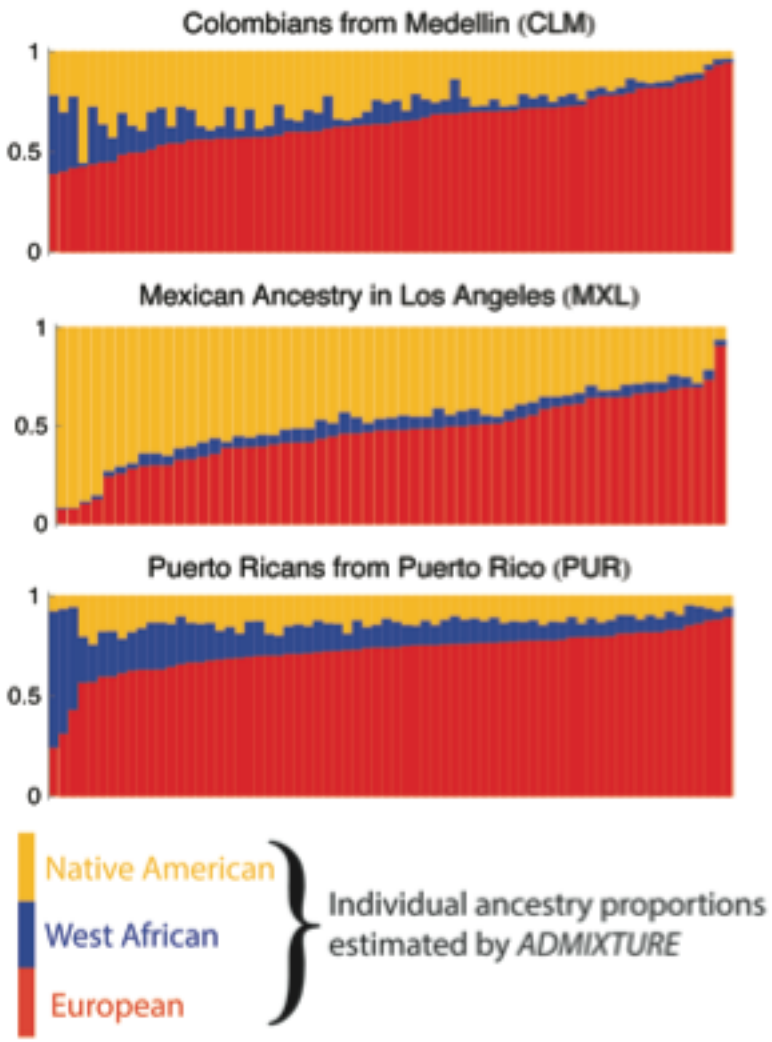


Figure 2. Cumulative Bayesian skyline plots of Y chromosome and mtDNA diversity by world regions. The red dashed lines highlight the horizons of 10 kya and 50 kya. Individual plots for each region are presented in Supplemental Figure S11.

(Karmin et al. 2015 Genome Research)

Admixture


Colombians from Medellin (CLM)
1
0.5
0
Mexican Ancestry in Los Angeles (MXL)
1
0.5
0
Puerto Ricans from Puerto Rico (PUR)
1
0.5
0
Native American
West African
European
Individual ancestry proportions
estimated by ADMIXTURE

Admixture graphs (trees + admixture events)

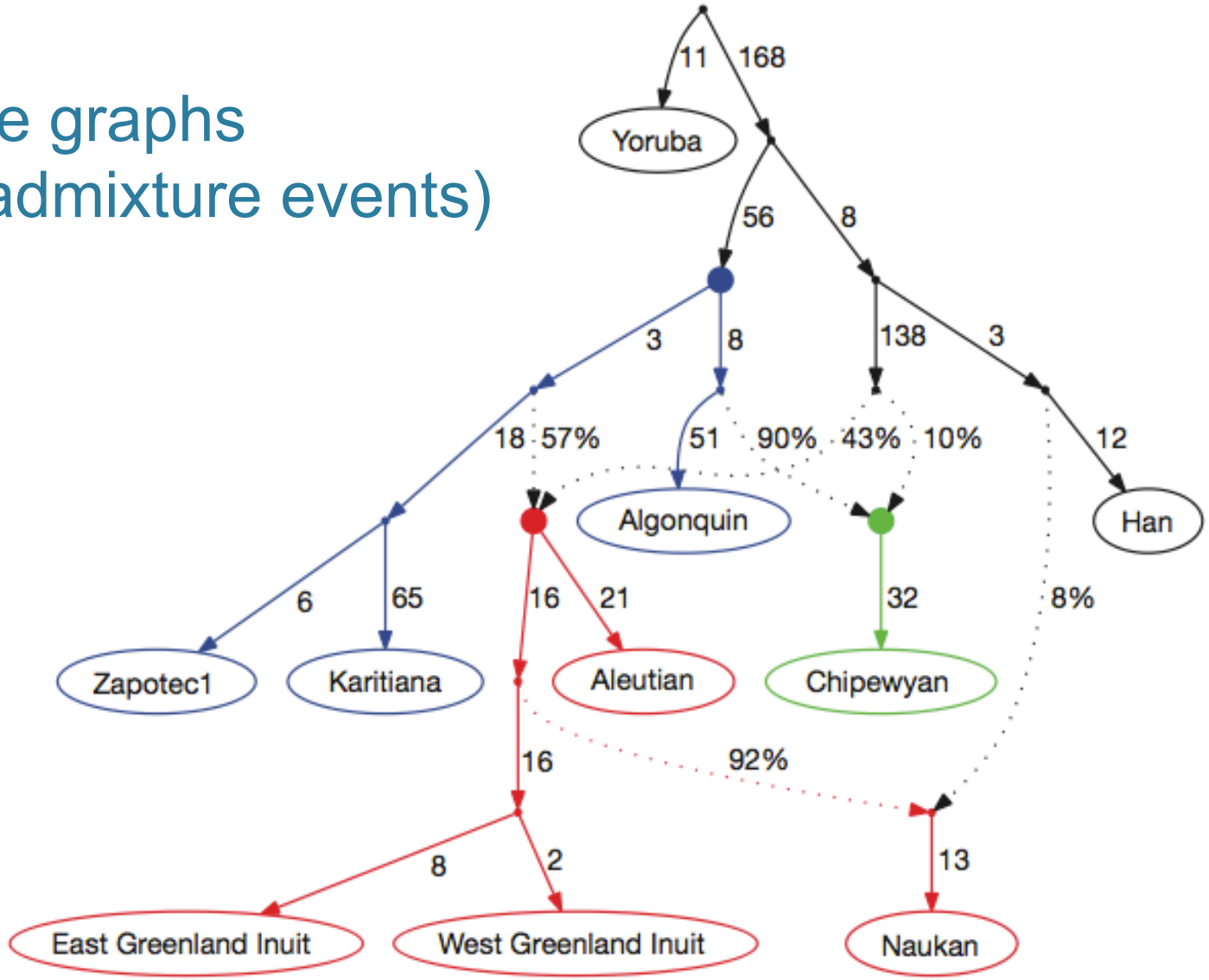


Figure 2 | Distinct streams of gene flow from Asia into America. We present

Reich et al. 2012, Nature

Considering relatedness is important

Genetics and Linguistics

Interpreting patterns of contact and population diversity within the linguistic perspective

- Identifying populations with ethno-linguistic affiliation
- Contact between populations, language shift, linguistic distances

What is a population?

- Population as a unit of research
  - Vertical transmission of traits
  - Unit stable in space and time
- Varies between disciplines
  - Social anthropology
  - Demography
  - Politics
  - Linguistics
  - Genetics
  - Molecular anthropology

What is a population?

- Population as a genetic pool
- Identifying human subgroups for understanding demographic trajectories
- Global human population is characterized by **gradients** of genetic and cultural diversity
- All human populations are the result of **admixture** occurring at a certain time depth

What is a population?

- All the individuals in a certain continent
- All the individuals in a certain country
- All the individuals in a certain region
- All the individuals who speak the same language
- All the individuals who speak the same language and are characterized by a common set of cultural features

  → **ethnolinguistic unit**

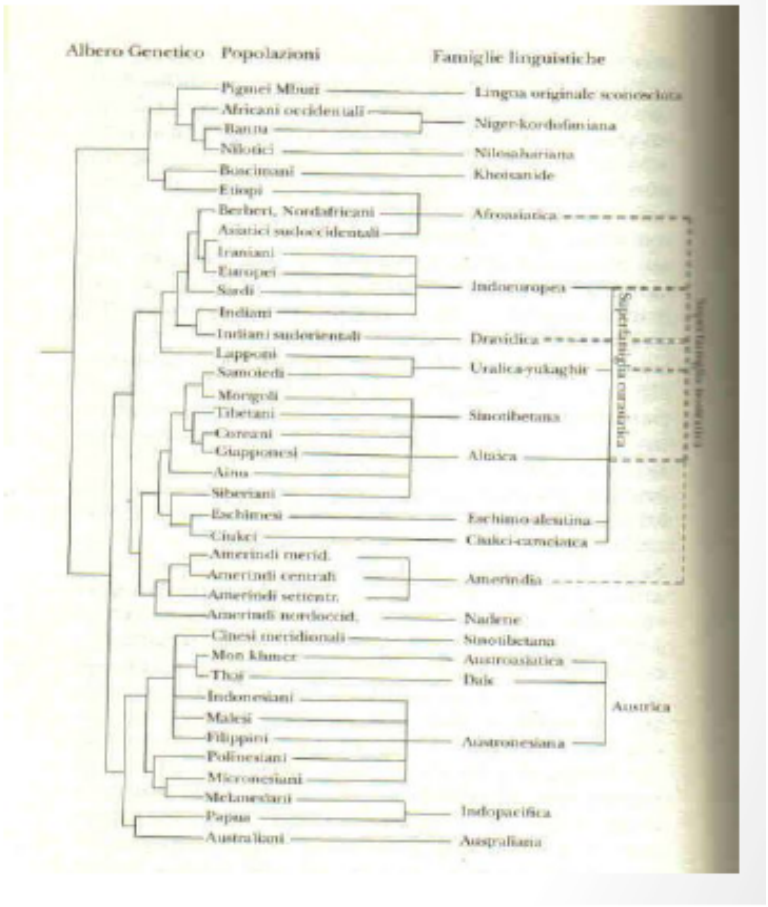
- The DNA is not considered at the individual level, but at the population level (genetic pool)

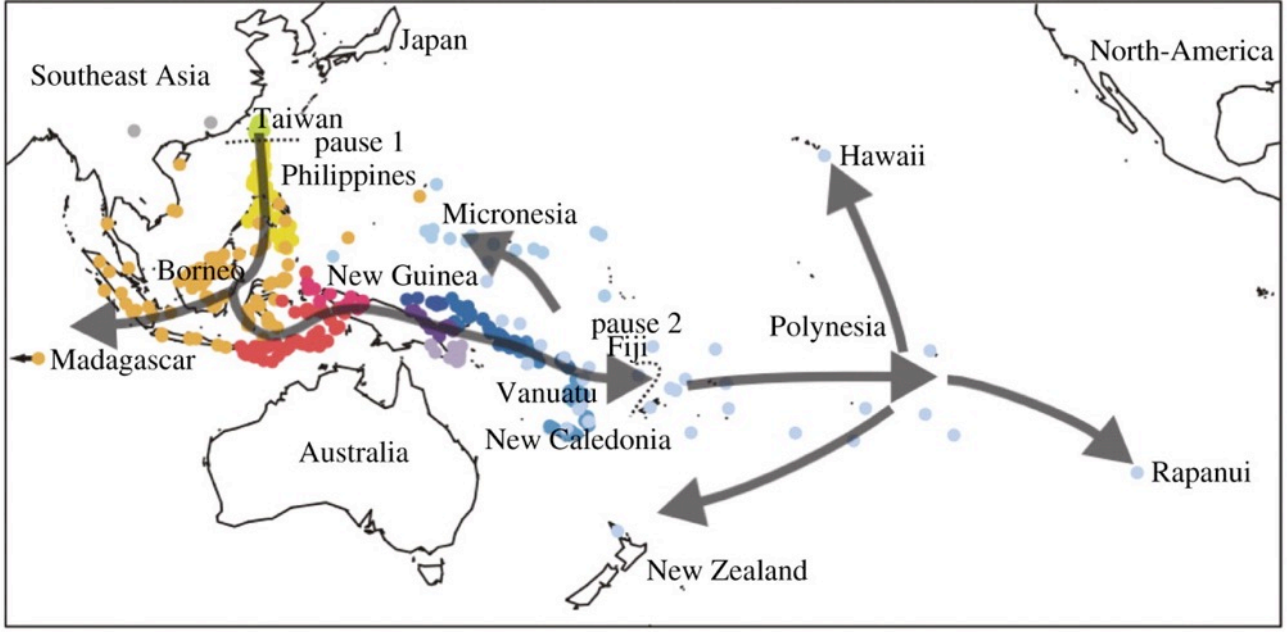
Population/language phylogenies

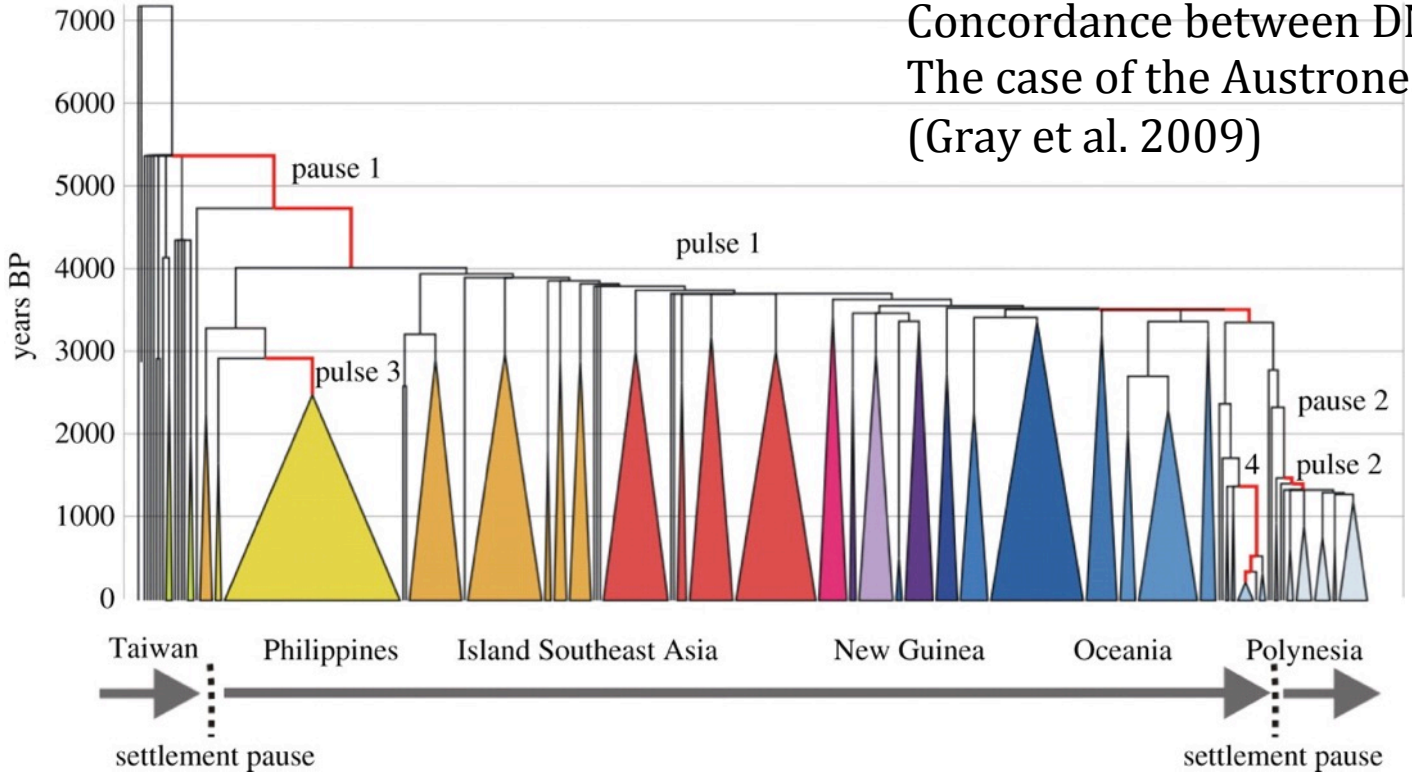
- Understanding human history
- Transmission of traits with modifications through generations

Genes, people, languages (L.L. Cavalli-Sforza, 1996)

Le popolazioni che appartengono alla stessa famiglia linguistica sono molto simili fra loro anche dal punto di vista genetico (216)







Concordance between DNA and Language:
The case of the Austronesian
(Gray et al. 2009)

Conclusions

- Multidisciplinary approach: genetics + linguistics + cultural anthropology + history + archaeology…
  - Reconstruct population history and understand present diversity
- Genetic data is shaped by demography
- Genetics works well with linguistics

Genetic data

- Panels of population diversity

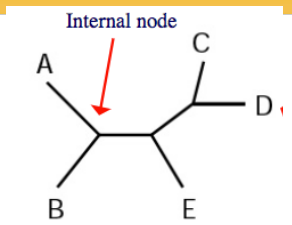
Exercise

Why population genetics

- Molecular anthropology and linguistics

Trees and networks

- How to build a tree



Building trees to understand evolution

A new genetic diversity panel: GeLaTo

GELATO: Genes and Languages Together

- **Language**: MPI SHH databases: glottobank (grambank, lexibank, parabank, phonobank, numeralbank), CoBL, soundcomparisons
  - Other language databases: WALS, Tsammalex, WOLD, AFBO...
- **Culture**: D-Place: Database of Places, Languages, Cultures and Environment, Pulotu: Database of Pacific Religions
- **DNA**: a new panel of genetic diversity

A new panel of genetic diversity

VALUE of GeLaTo:

- Interdisciplinary perspective on human history and diversity
- Link genetic samples with GlottoCodes and provide summary statistics
  - Non-geneticists can use standardized genetic values for their analysis
  - Molecular anthropologists/geneticists can use it as a reference and population geneticists will be pushed to give better ethnographic and linguistic characterization to their samples

Genetic datasets

- Autosomal STR (Pemberton)
  - Ready to use, GlottoCodes matched
- Mitochondrial full genomes (sequence)
- Y chromosome STR
  - Some data available to play with (Americas and Africa), ask me!
- Autosomal SNP (Human Origins)
  - Standardized panel. Work in progress with Irina Pugach (MPI EVA Leipzig) and Hiba Babiker (MPI SHH Jena)

GENETIC VALUES (data.csv)

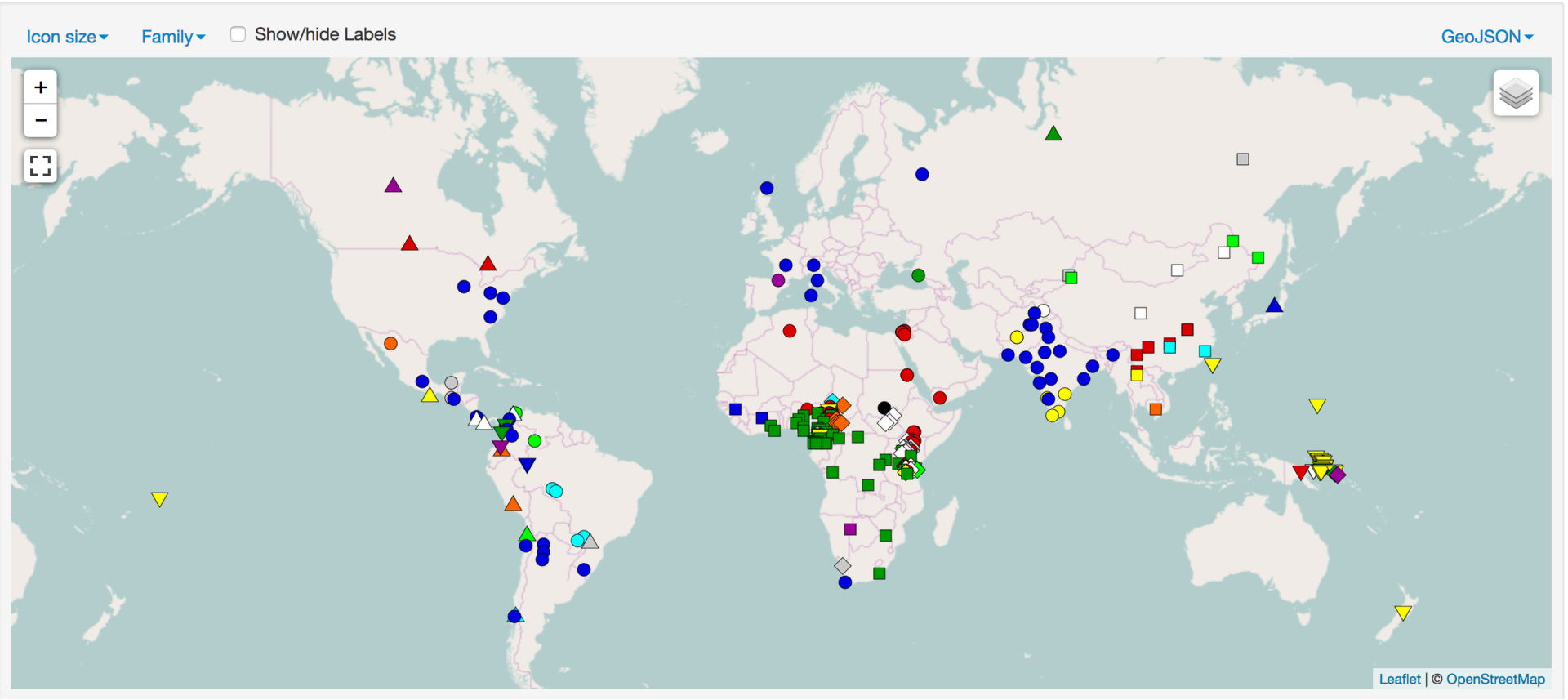
We want only robust values from standardized and widely used formulas

Within population

- Diversity
  - Heterozygosity – nucleotide diversity
  - Haplotype diversity
  - Average FST distance (also within geo regions)
- TMRCA
- Mismatch curve or BSP from BEAST - Indicative of population expansion or bottleneck

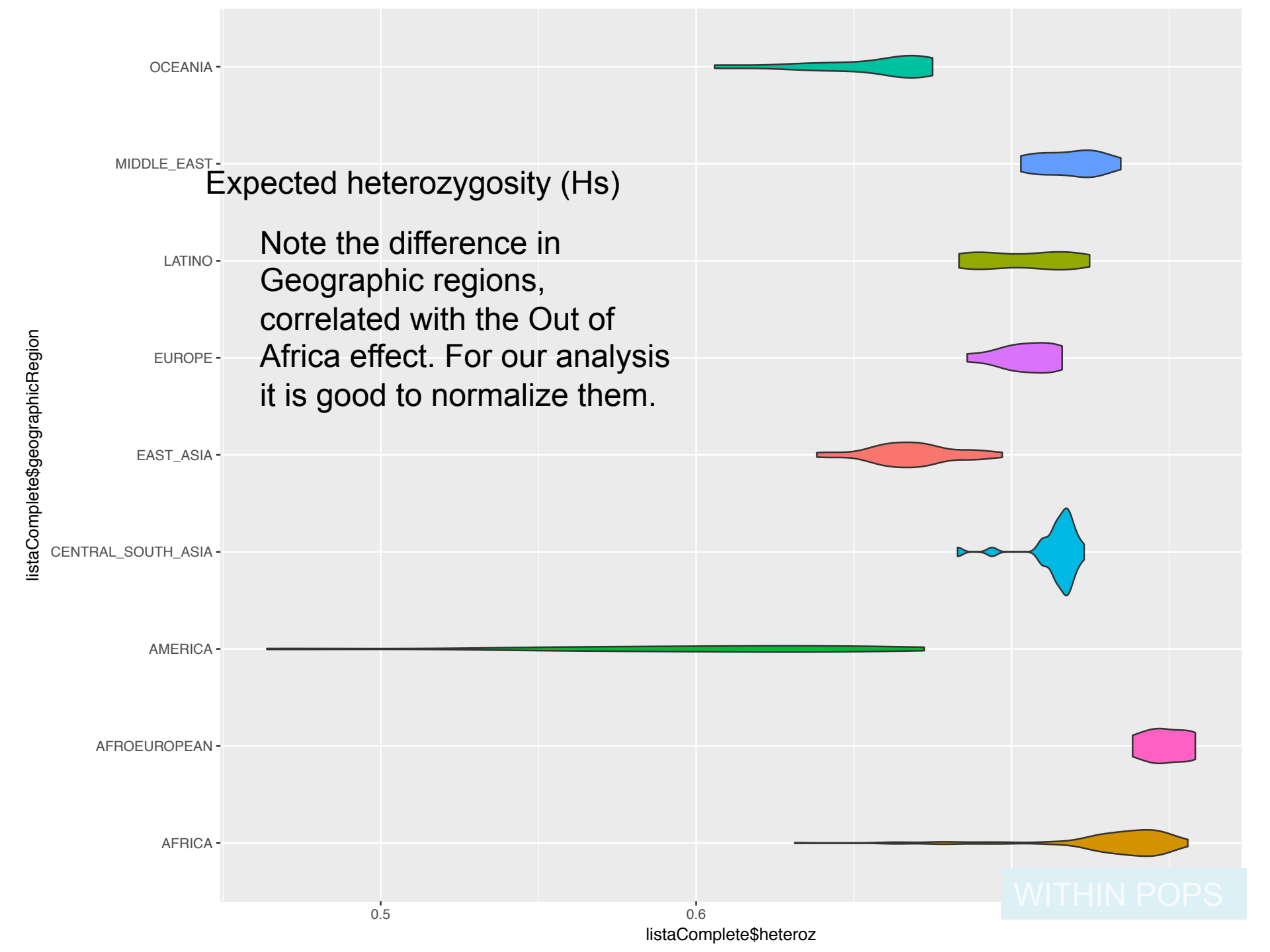
- Between populations
  - FST – genetic distance (medium-ancient patterns)
  - allele-sharing distance
  - Haplotype exchange (recent patterns)

Genetic Samples



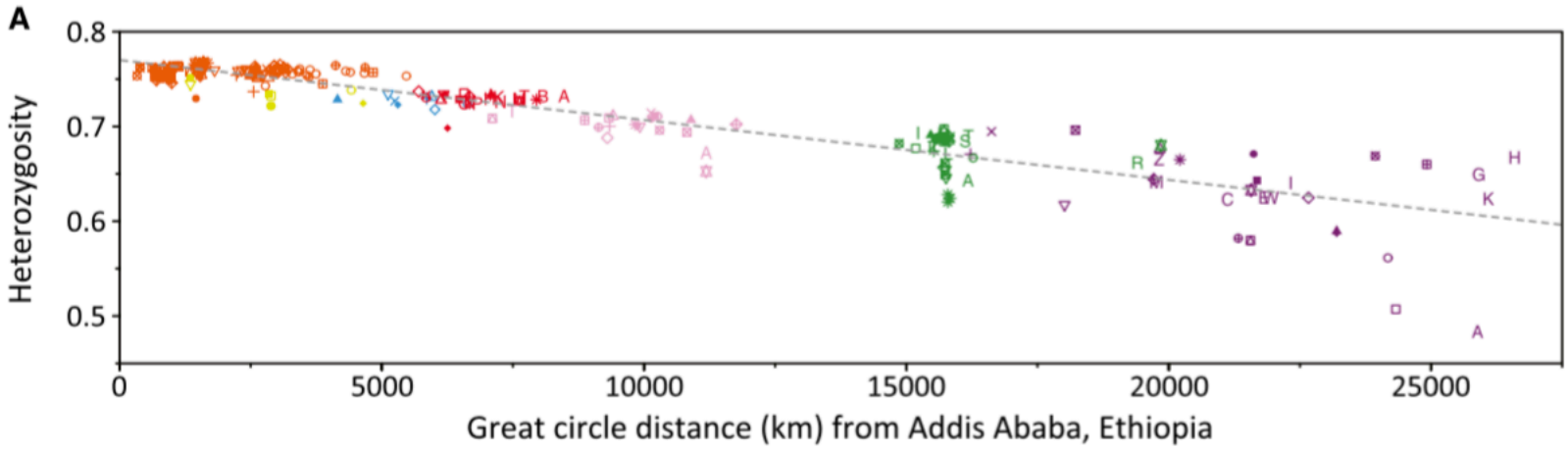
Geographic Regions: identified for their geographic but also genetic characteristics. Each region has different genetic characterization.

AFRICA	112
AFROEUROPEAN	5
AMERICA	29
CENTRAL_SOUTH_ASIA	24
EAST_ASIA	21
EUROPE	9
LATINO	13
MIDDLE_EAST	10
OCEANIA	38

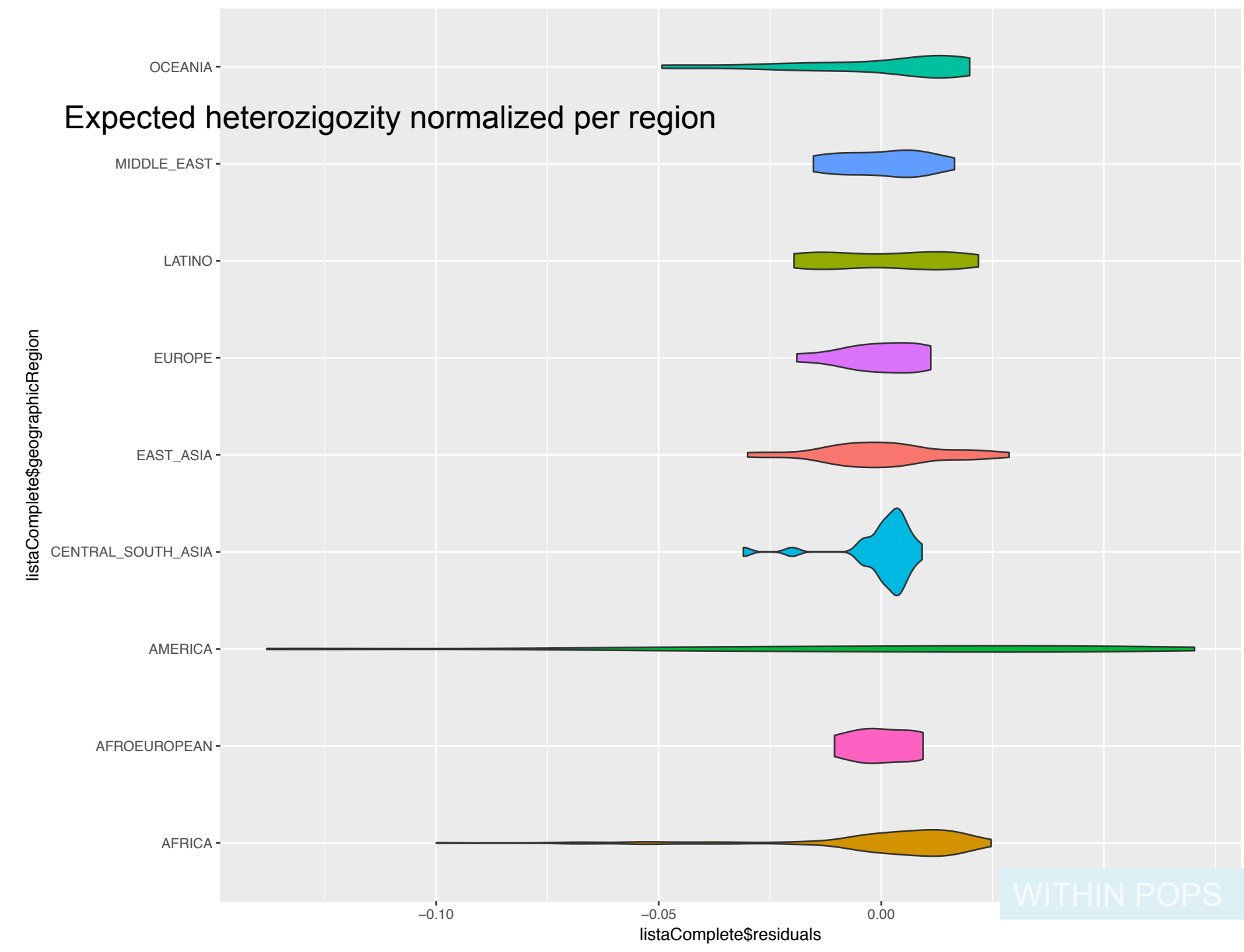

Expected heterozygosity (Hs)
Note the difference in Geographic regions, correlated with the Out of Africa effect. For our analysis it is good to normalize them.
OCEANIA
MIDDLE_EAST
LATINO
EUROPE
EAST_ASIA
CENTRAL_SOUTH_ASIA
AMERICA
AFROEUROPEAN
AFRICA
listaComplete$geographicRegion
0.5
0.6
listaComplete$heteroz
WITHIN POPS

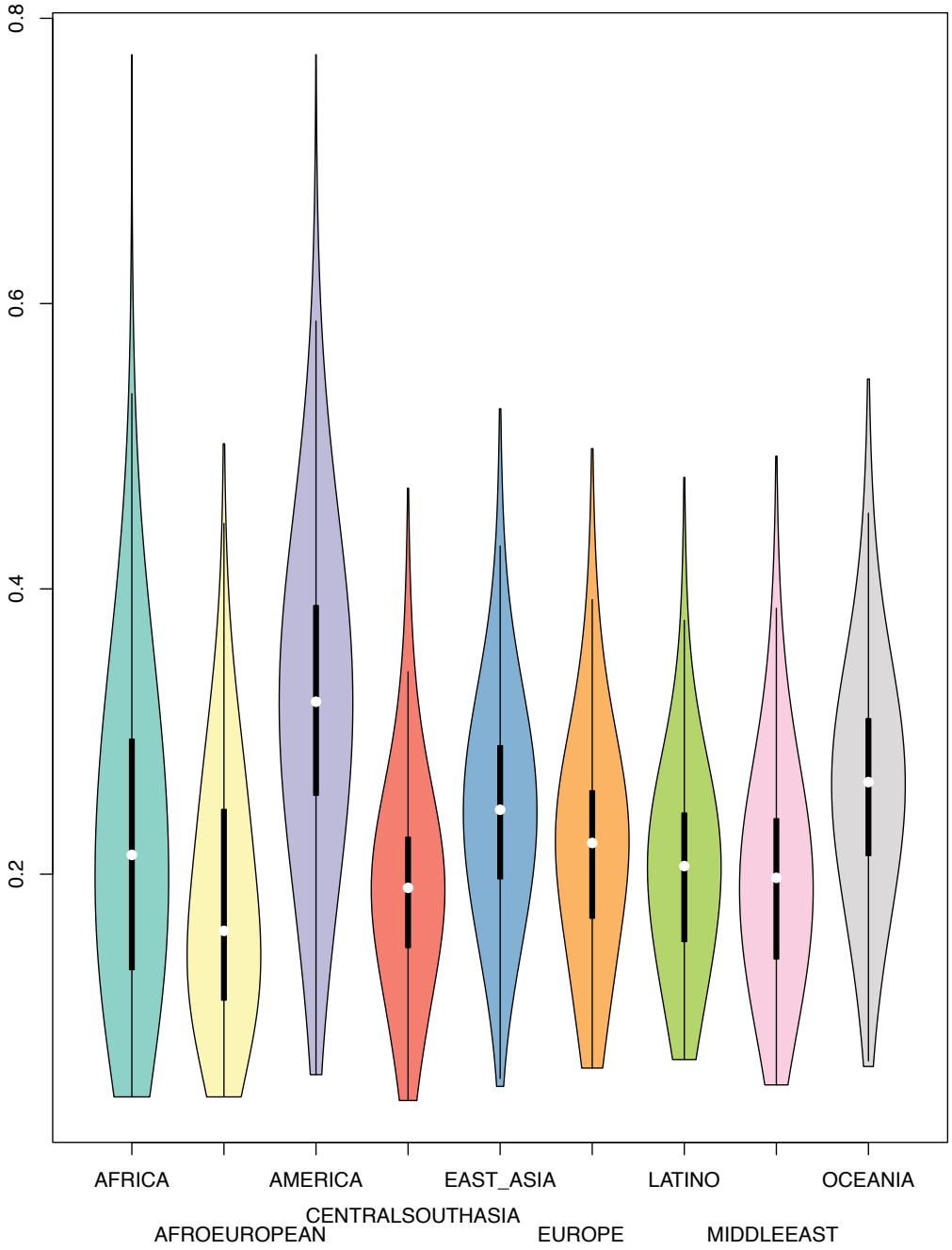
Ecological constraints

- The affinities between genetic and linguistic patterns of variation are primarily affected by geographic proximity, in a Isolation by Distance Wright model



Pemberton et al. 2013


Expected heterozigozity normalized per region
OCEANIA
MIDDLE_EAST
LATINO
EUROPE
EAST_ASIA
CENTRAL_SOUTH_ASIA
AMERICA
AFROEUROPEAN
AFRICA
listaComplete$geographicRegion
−0.10
−0.05
0.00
listaComplete$residuals
WITHIN POPS

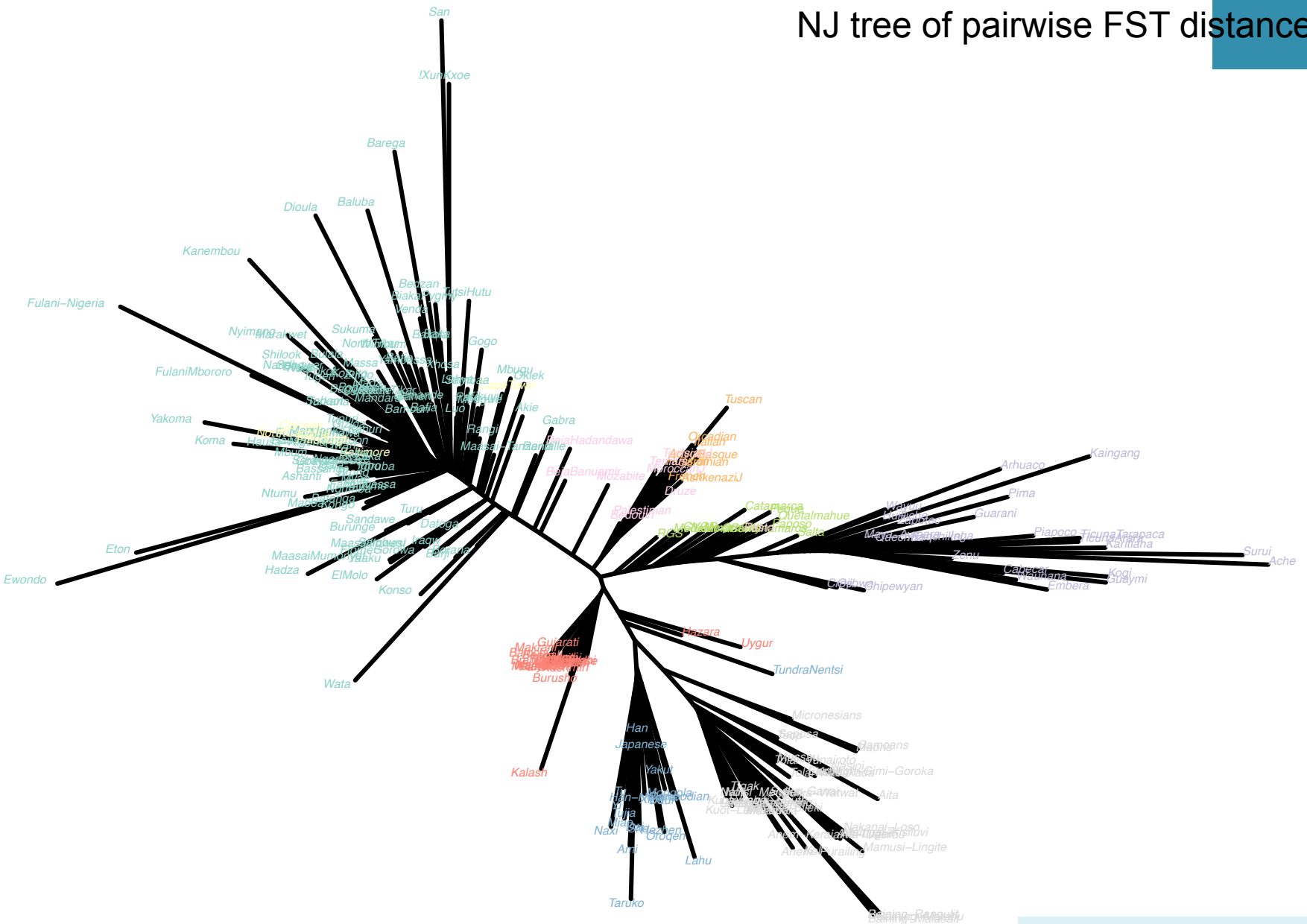


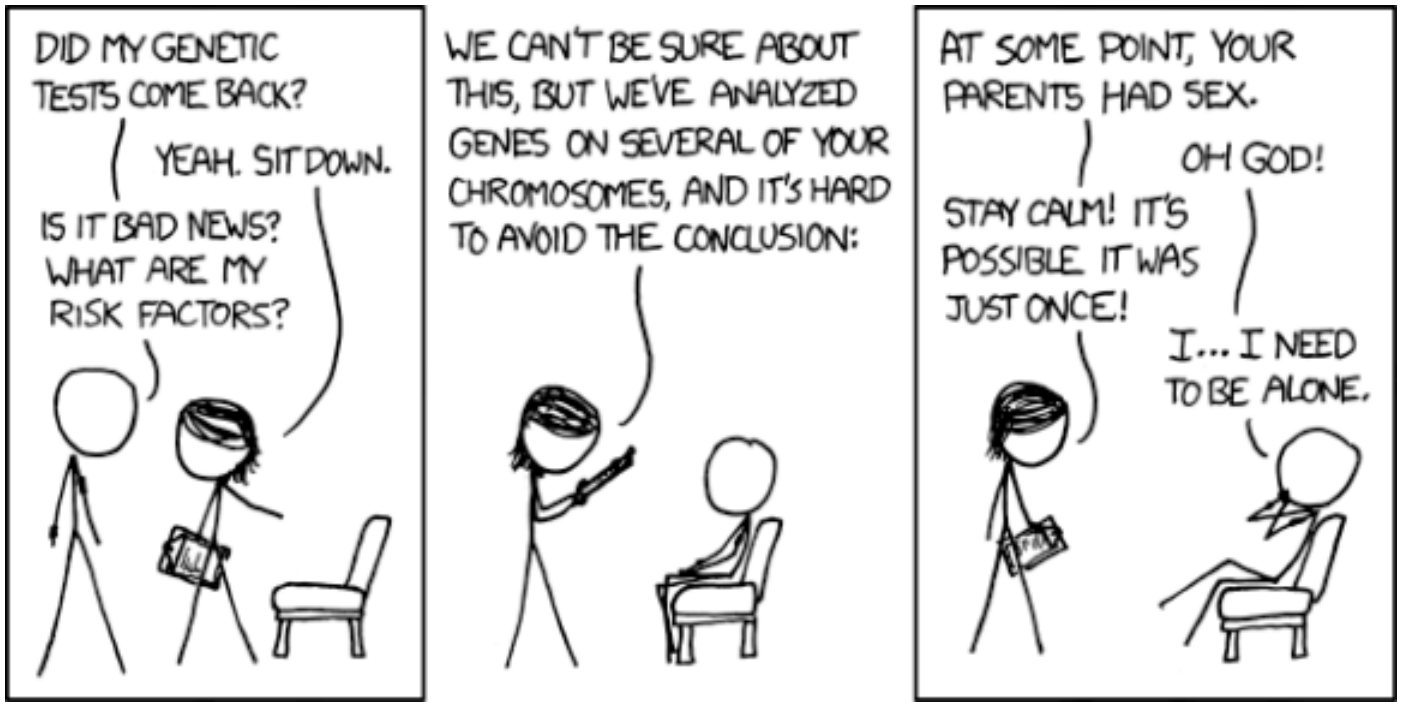
FST Genetic distance between pops

Population structure, how much population within a region are different from each other

BETWEEN POPS

NJ tree of pairwise FST distances




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.

Genetic data

- Panels of population diversity

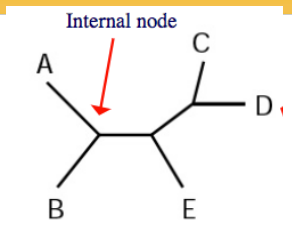
Exercise

Why population genetics

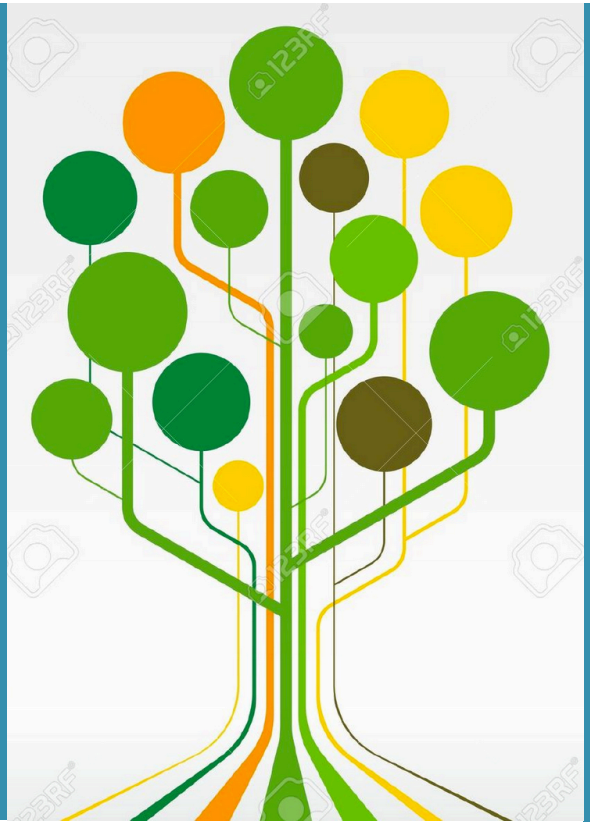
- Molecular anthropology and linguistics

Trees and networks

- How to build a tree



Building trees to understand evolution



Trees and networks

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

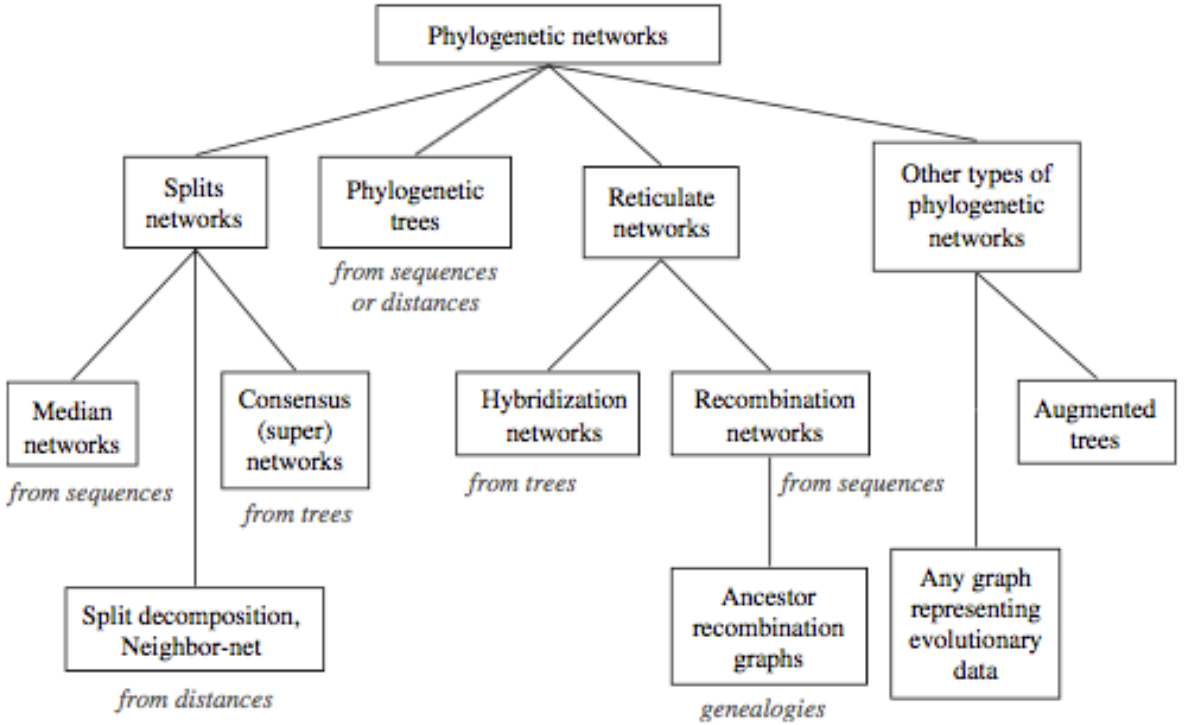
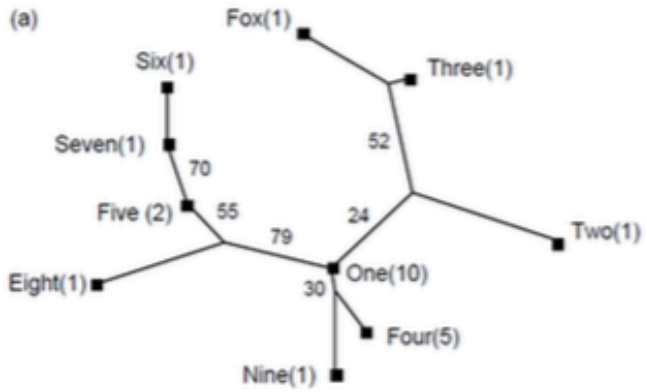


FIG. 1.—The term phylogenetic network encompasses a number of different concepts, including phylogenetic trees, split networks, reticulate networks, the latter covering both "hybridization" and "recombination" networks, and other types of networks such as "augmented trees." Recombination networks are 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

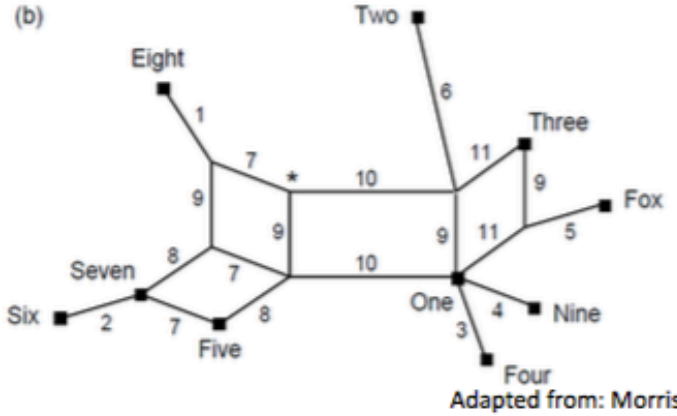
- Trees impose bifurcations
- Networks allow reticulations

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

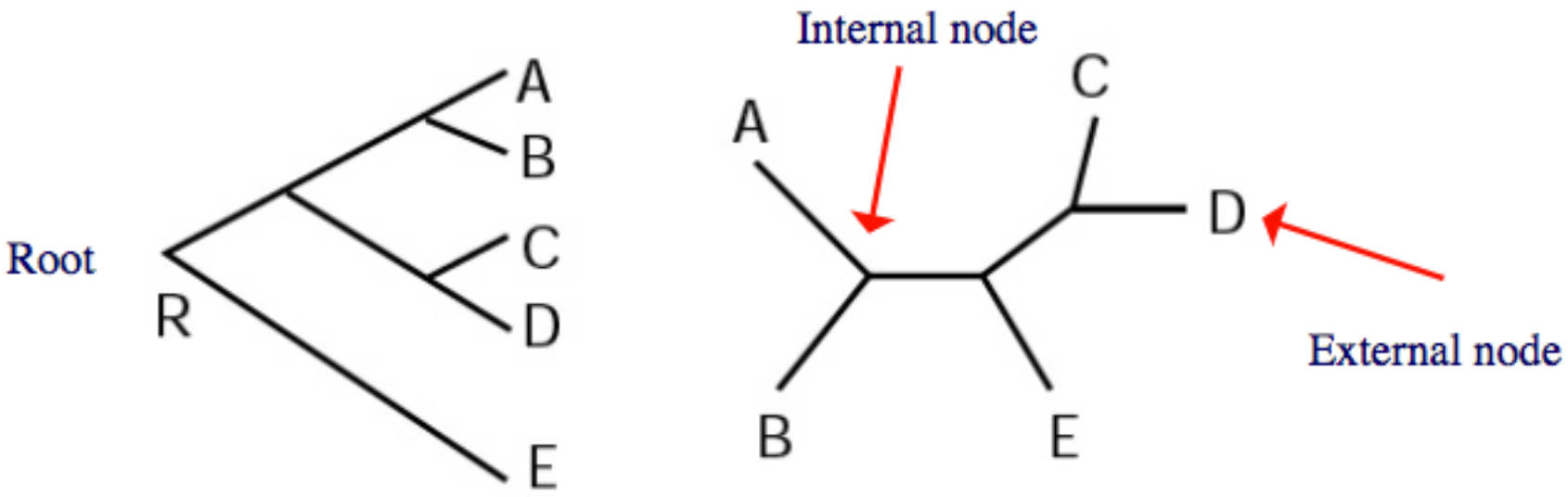


Adapted from: Morrison 2005

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

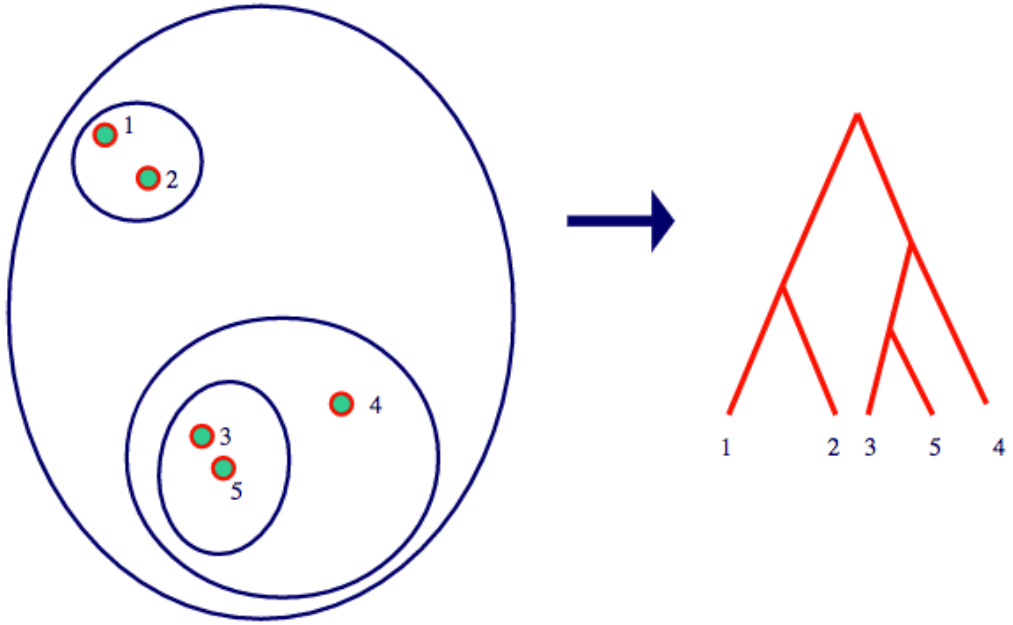
- First calculate distance matrix between pairs of sequences or populations
- Then build a tree

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.

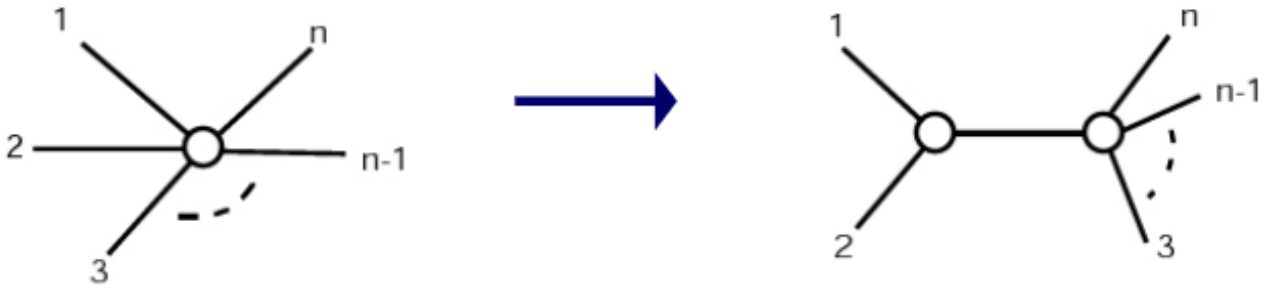


Distance-based trees

- **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

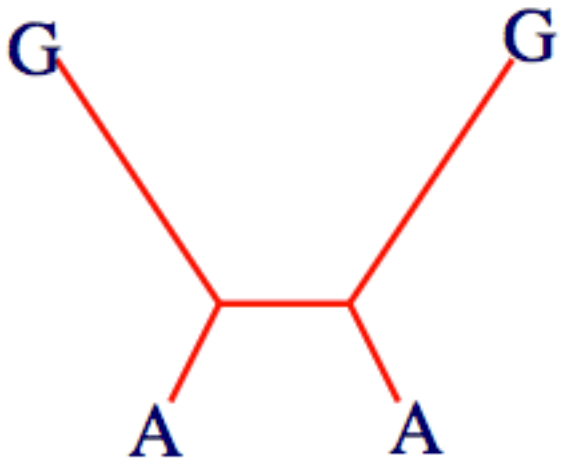
Character based trees

- 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

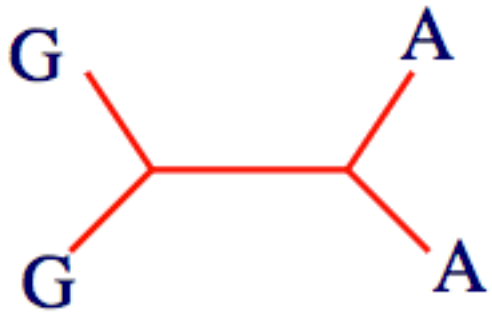
They both imply model of evolution

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

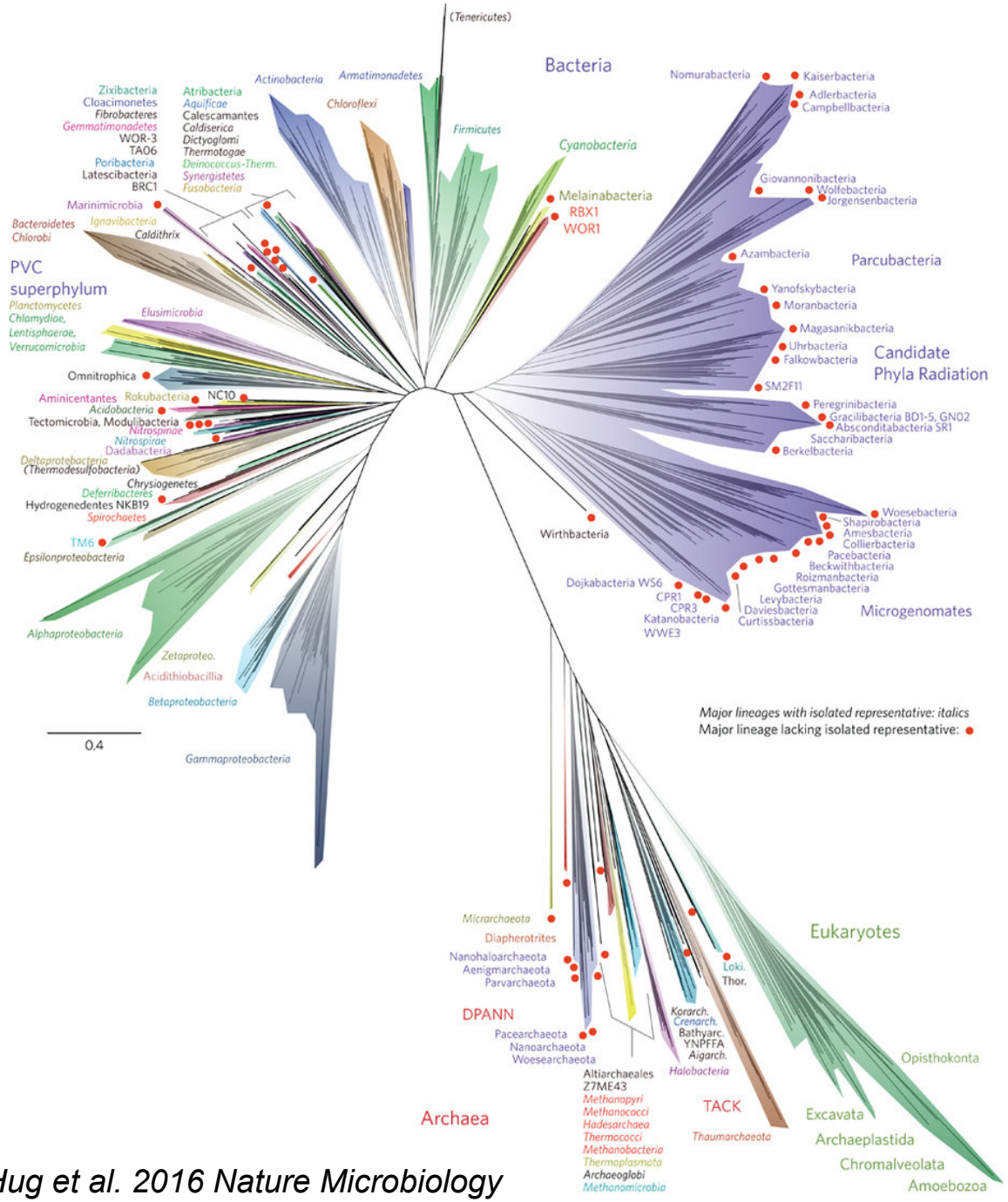
Summary: trees

- 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

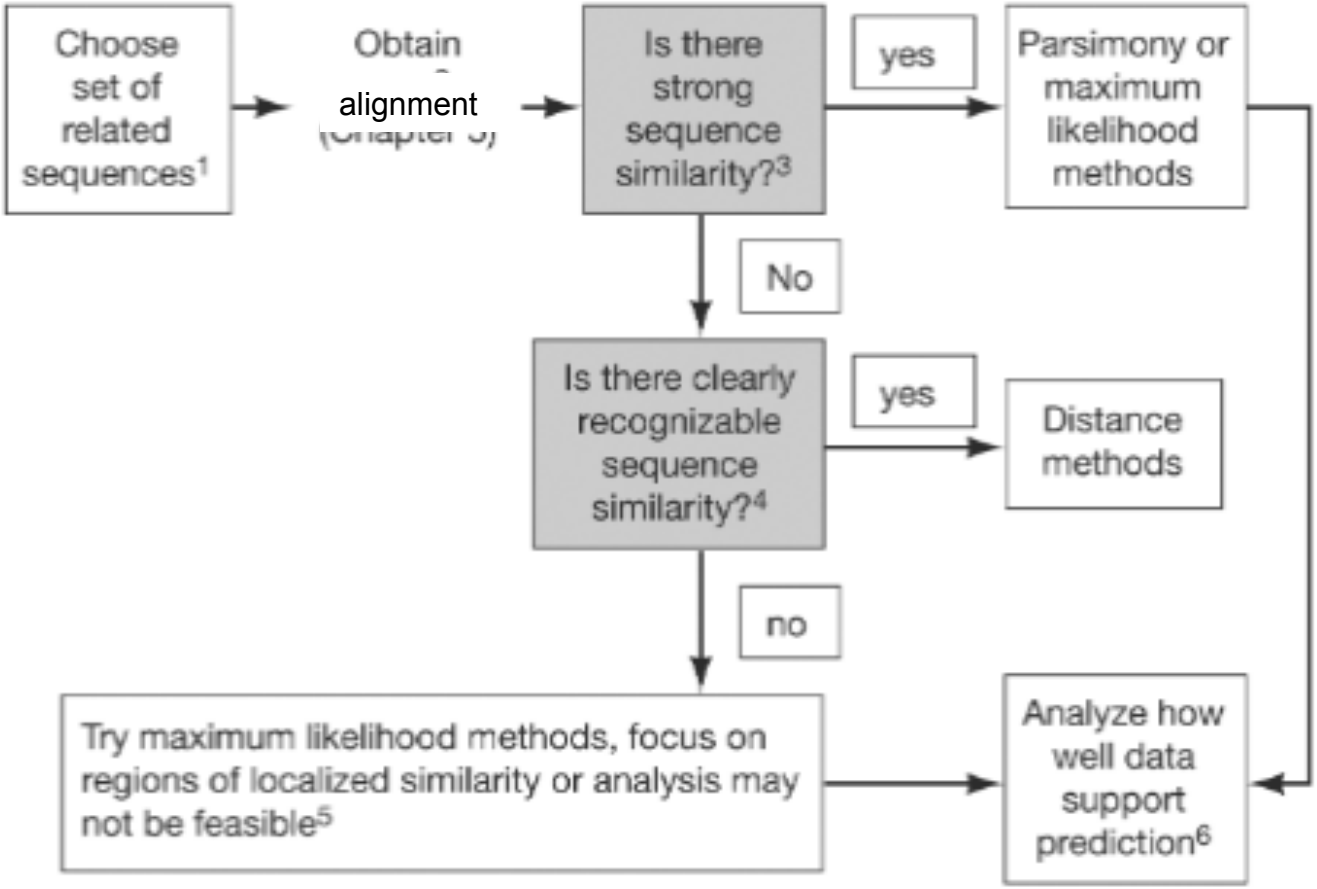
Battle between preferences:

- 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

Networks

Networks

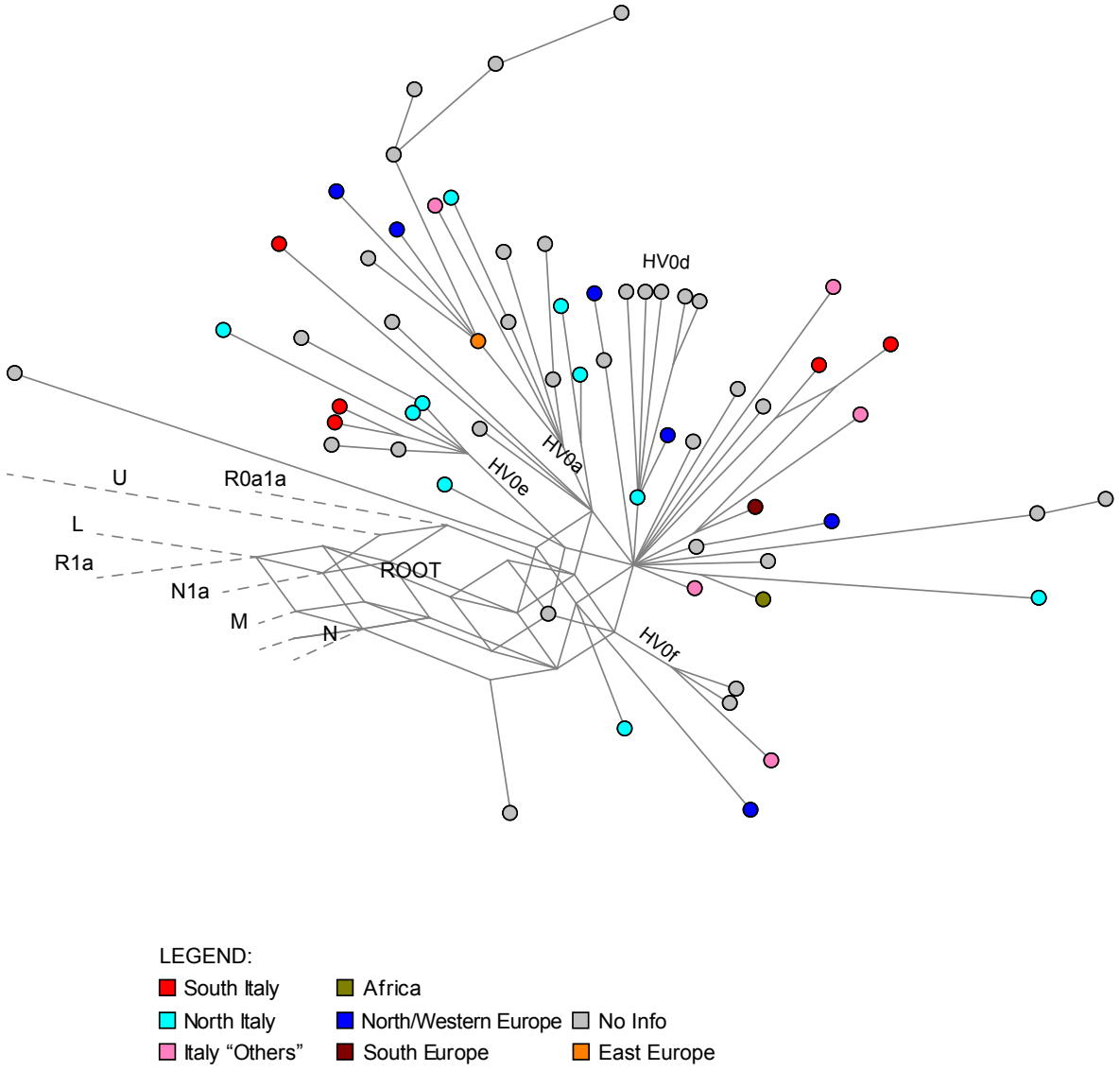
- Phylogenetic networks can be computed from multiple sequence alignments, distance matrices, sets of trees, clusters, splits, etc.
- Note: loss of evolutionary **direction**
- Note: the more tree-like the data are then the more tree-like will be the network

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

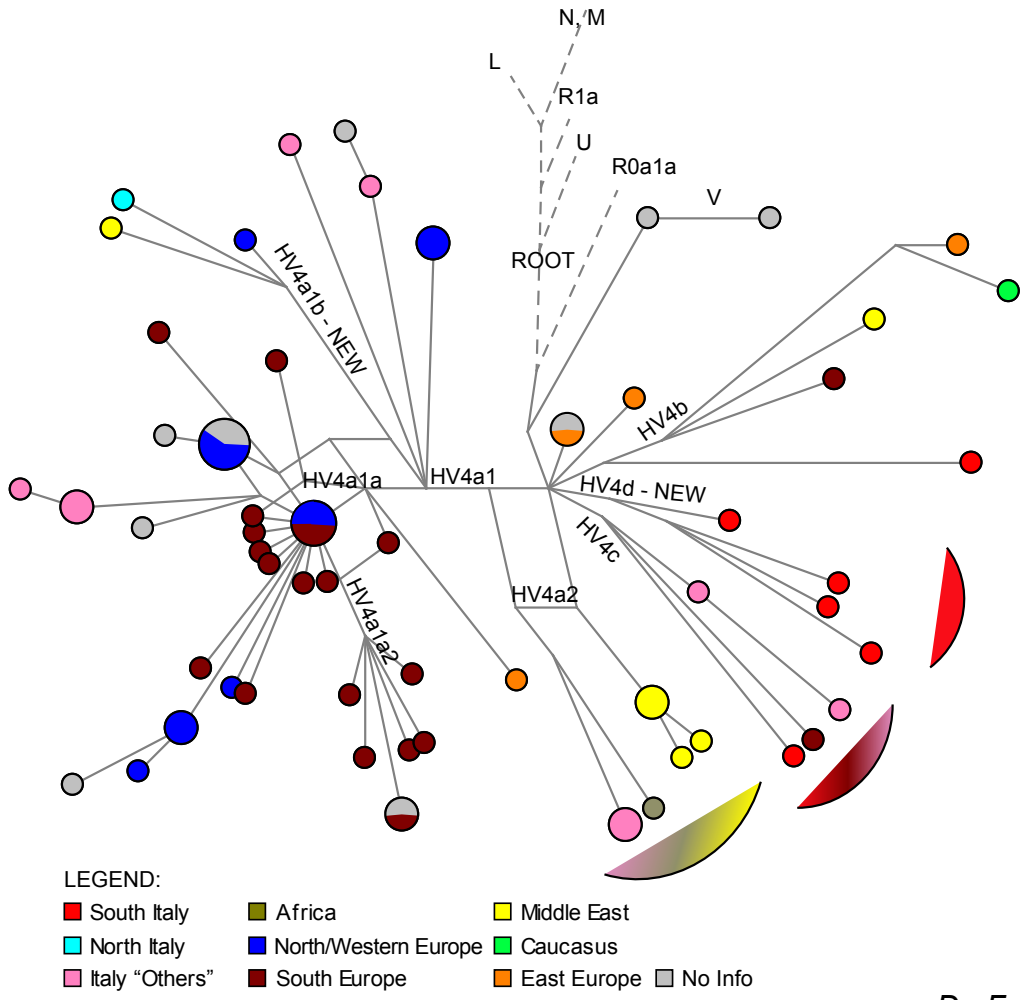


De Fanti, Barbieri 2016

Simplify a network

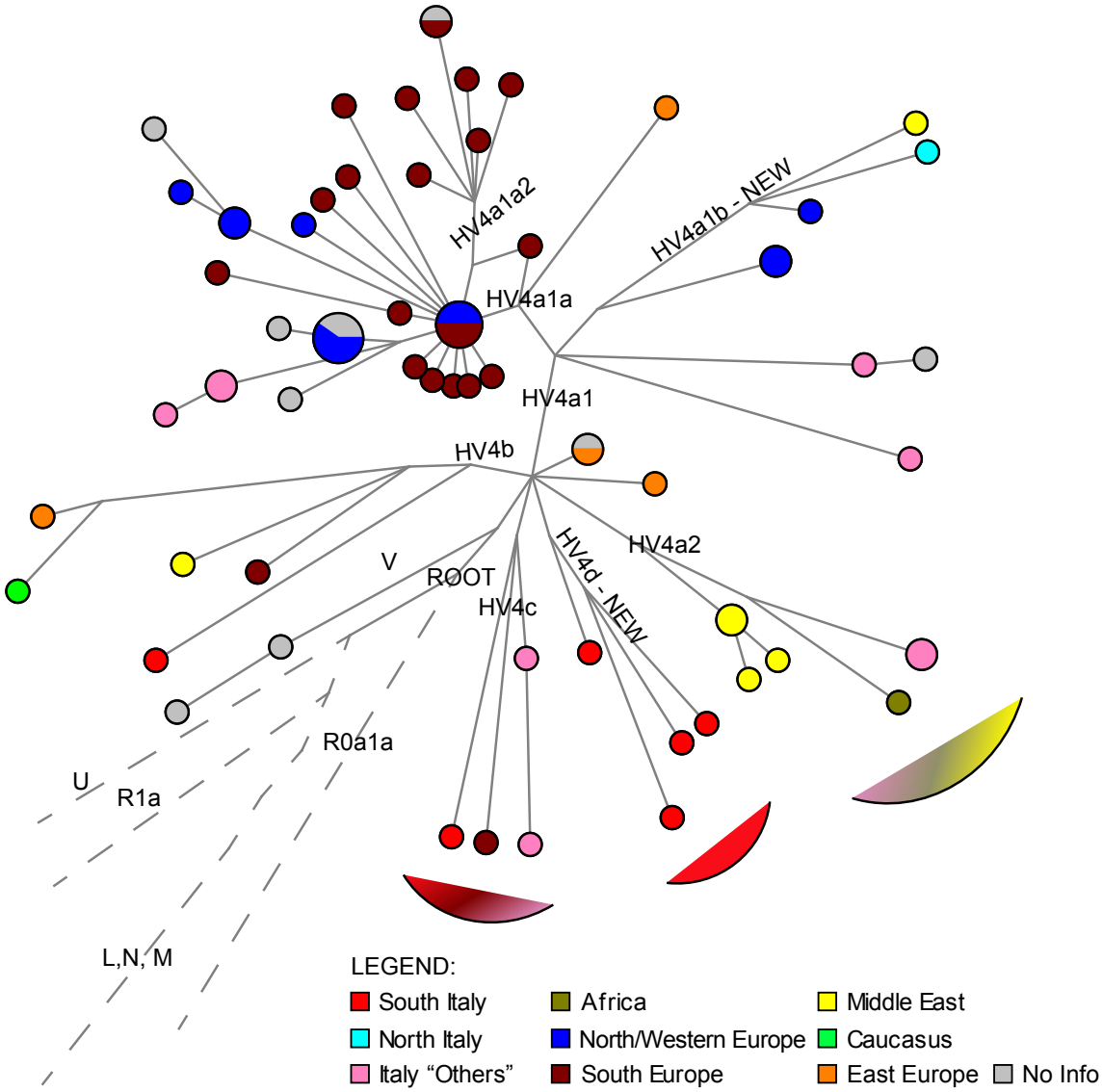
- For less reticulations, apply **weights** to the characters.
  - Positions with recurrent change of state are downweighted
- The software Network allows a few tricks: Reduced Median networks, post processing, star contraction

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

Splits network

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

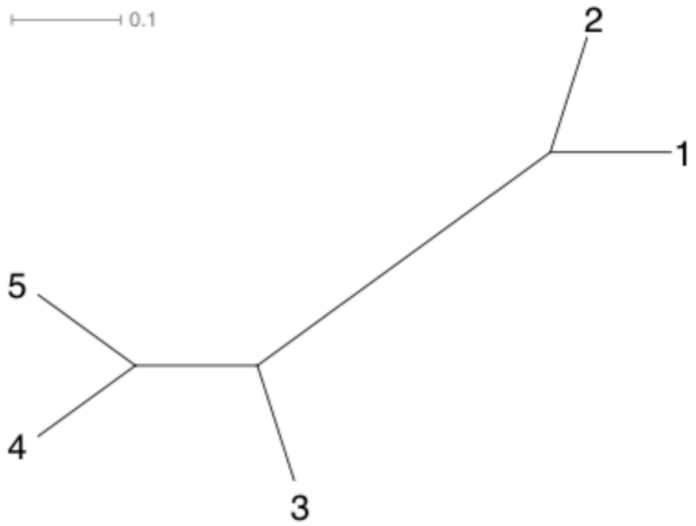
- Split decomposition: can be based on the raw data (called parsimony splits; Bandelt and Dress, 1993) or more usually on a distance measure (Bandelt and Dress, 1992).
- Neighbour-Net: distanced-based method (Bryant and Moulton, 2002 , 2004)
  - compromise between the preponderance of apparent false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

Use: **SplitsTree**

Understanding NeighborNet:

- simple data – with no conflicting signal, the tree fits perfectly
- The first 5 characters (A-E) are singletons – i.e. are a cognate set present in one language alone. Ignore them – they're just here to give the tree some branch-lengths.
- The characters F and G group language 1 and 2 together. Character H groups languages 4 and 5 together.
  Character I groups language 3, 4 and 5.

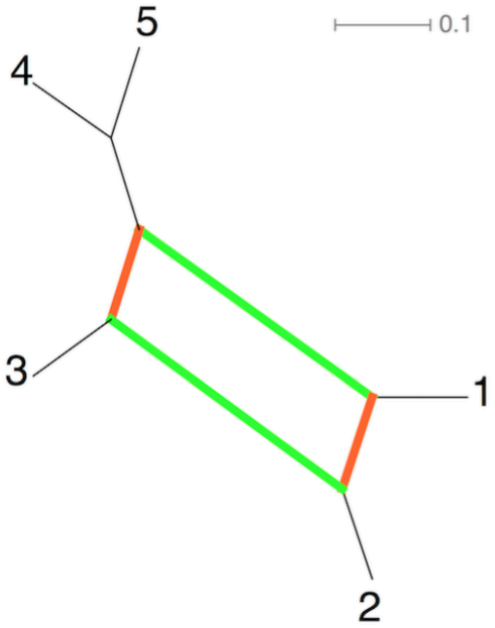
	A	B	C	D	E	F	G	H	I
language1	1	0	0	0	0	1	1	0	0
language2	0	1	0	0	0	1	1	0	0
language3	0	0	1	0	0	0	0	0	1
language4	0	0	0	1	0	0	0	1	1
language5	0	0	0	0	1	0	0	1	1



Courtesy S. Greenhill

- added one extra character to the data – J –
- this groups languages 2 and 3 together – in conflict with the groupings we saw above.
- We have 10 characters now, and 1/10 characters (i.e. J) says that 2 and 3 group together. This is the short edge of the box below (in orange), which measures about 1/10th in length. The other side of the box (green) is longer – about 2/10ths – which is the 2 character (F and G) that put language 1 and 2 together.

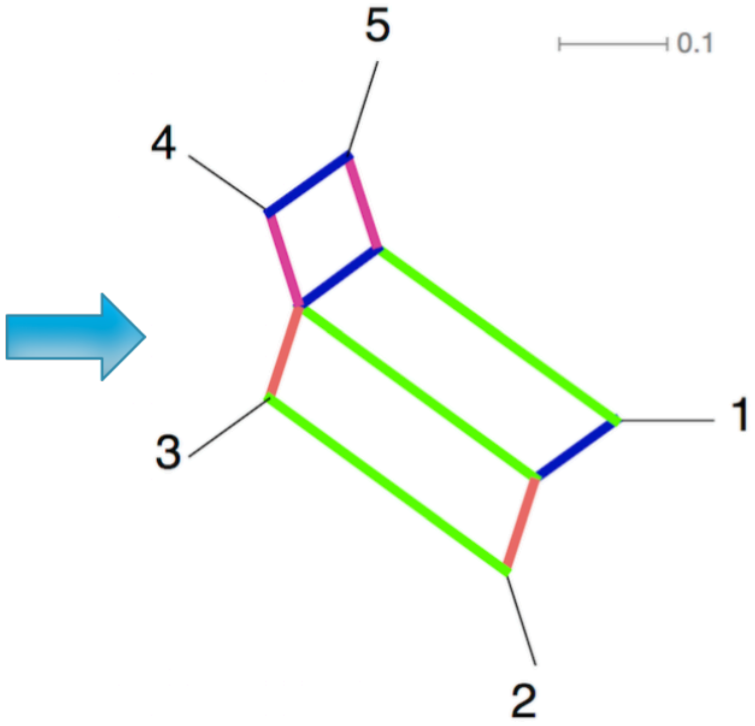
	A	B	C	D	E	F	G	H	I	J
language1	1	0	0	0	0	1	1	0	0	0
language2	0	1	0	0	0	1	1	0	0	1
language3	0	0	1	0	0	0	0	0	1	1
language4	0	0	0	1	0	0	0	1	1	0
language5	0	0	0	0	1	0	0	1	1	0



Courtesy S. Greenhill

- More complex: character K adds more conflict grouping languages 1 & 5 together.

	A	B	C	D	E	F	G	H	I	J	K
language1	1	0	0	0	0	1	1	0	0	0	1
language2	0	1	0	0	0	1	1	0	0	1	0
language3	0	0	1	0	0	0	0	0	1	1	0
language4	0	0	0	1	0	0	0	1	1	0	0
language5	0	0	0	0	1	0	0	1	1	0	1



Courtesy S. Greenhill

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

Summarizing:

- Genetic data
  - Study of “population” samples
  - Measures relatedness between individuals and populations
  - Can reconstruct population history, demography, origin, migration, contact
- Panels of genetic diversity not always informative
  - Use GeLaTo!
- Coalescence: infer phylogenies backwards, reconstruct history
  - can be calibrated in time with a clock assumption

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**