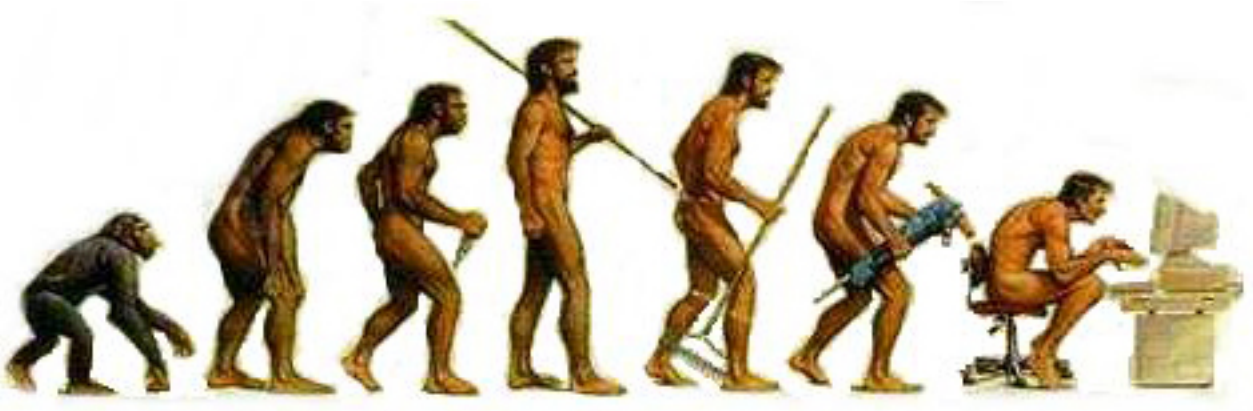
Genetic diversity: DNA, transmission, phylogeography

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

Why study population genetics

A multidisciplinary approach

- Anthropology as the study of our origin, history and current diversity
- Biological and cultural nature of humankind coevolve under the same demographic processes
- Hypotheses from genetics, archaeology, linguistics, cultural anthropology complement and validate each other



Genetic contributions to population prehistory

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

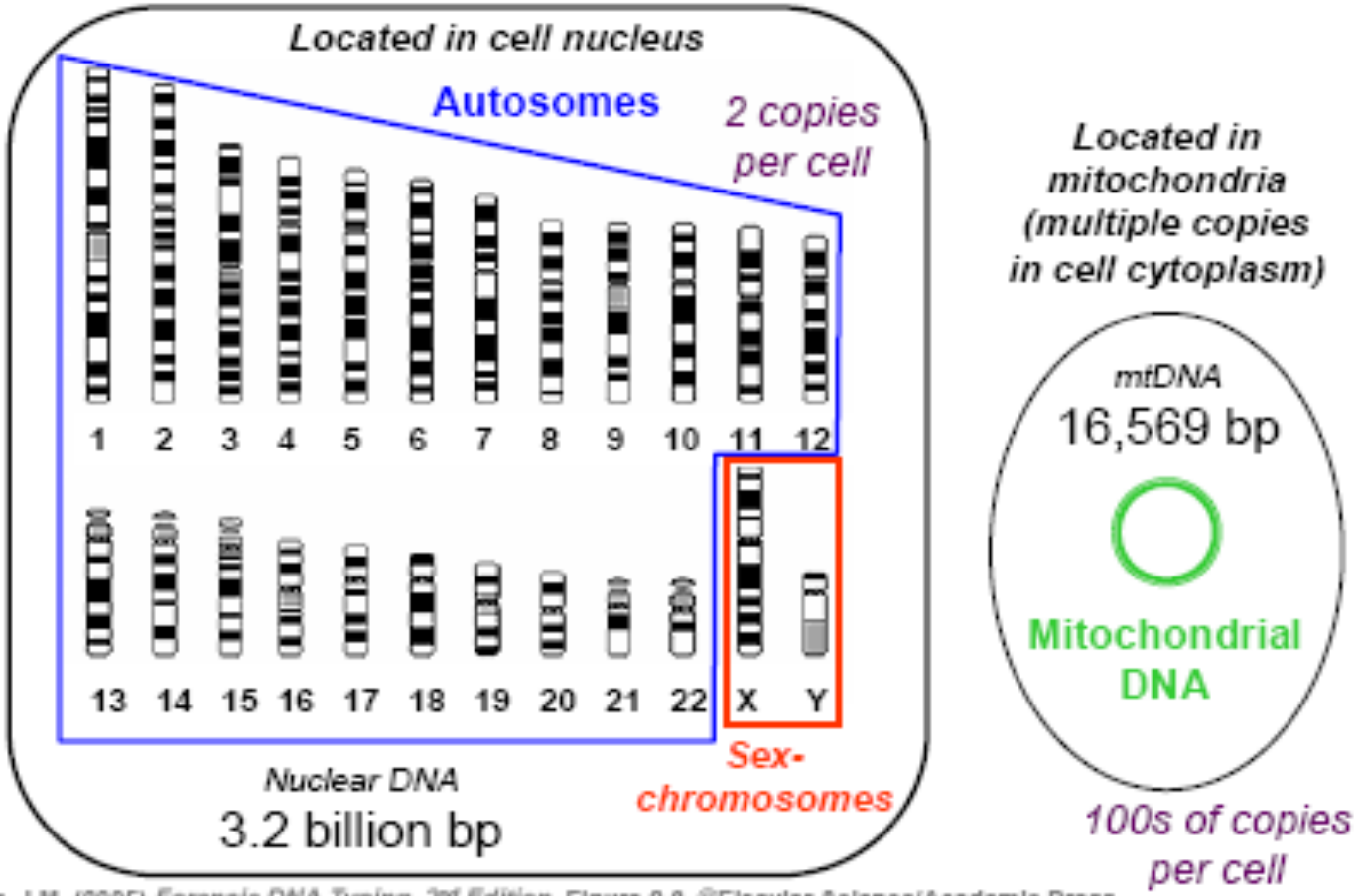
→DNA polymorphisms can retain traces of a group's prehistory

Genetic diversity a

Molecular markers

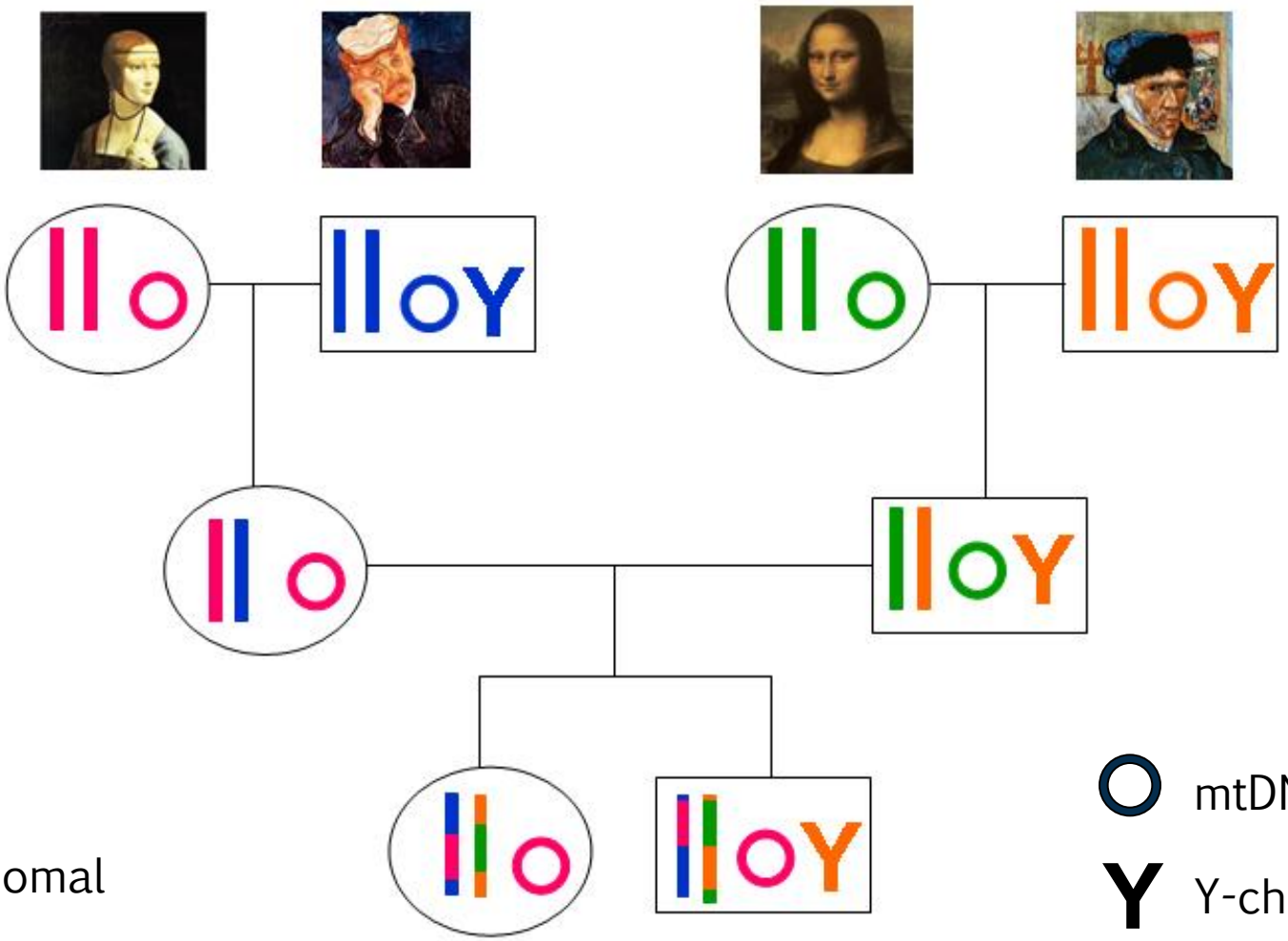
The human genome

- 23 pairs of chromosomes + mtDNA

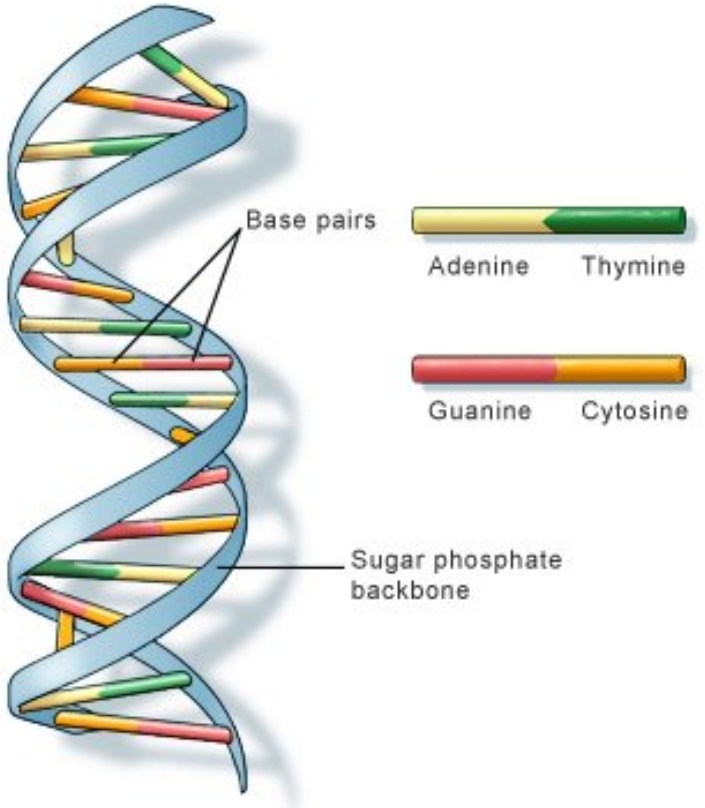


Butler, J.M. (2005) *Forensic DNA Typing, 2nd Edition*, Figure 2.3, ©Elsevier Science/Academic Press

Genetic markers: transmission



Measuring genetic diversity

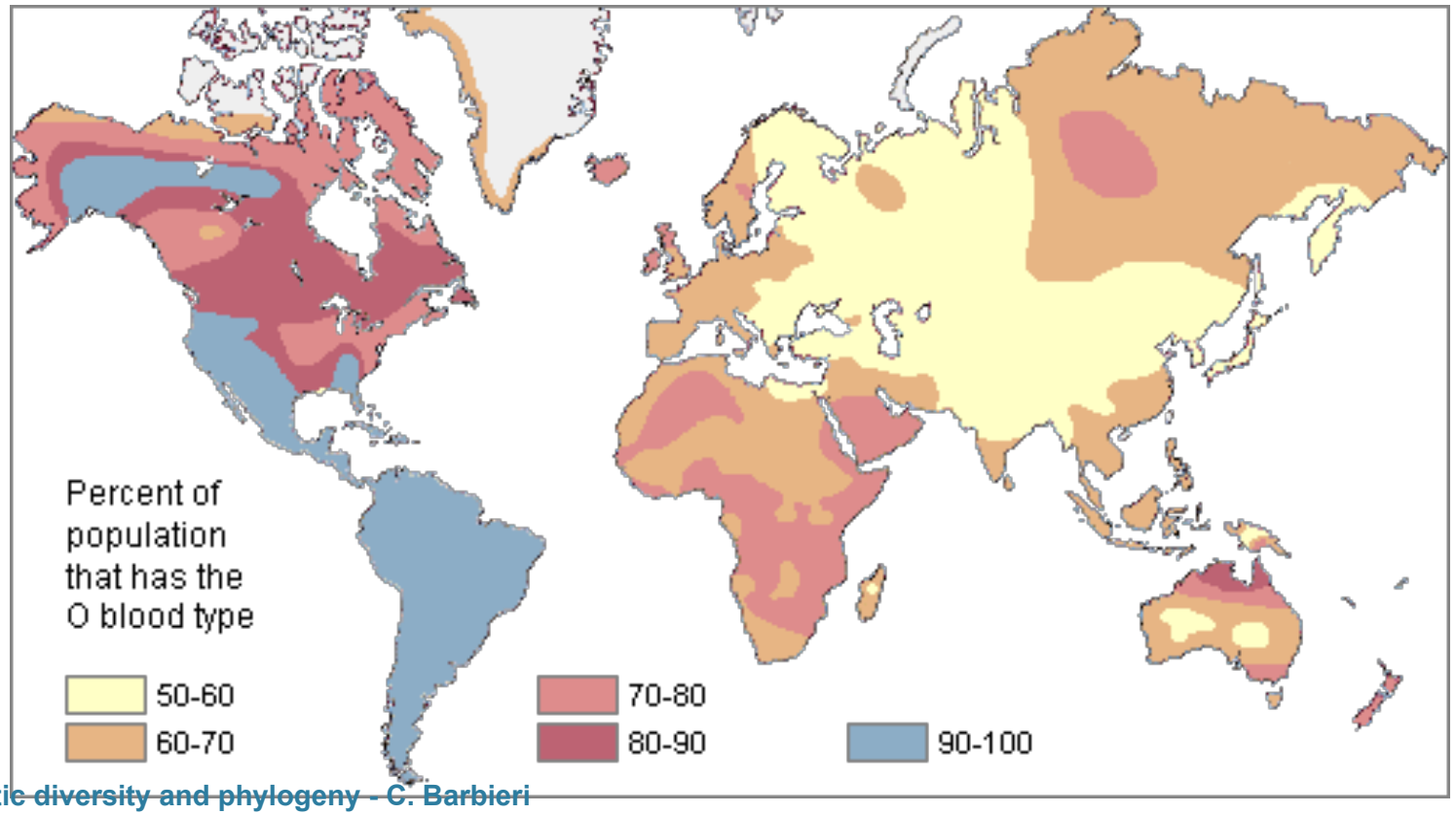


Different approaches:

- **Classical markers** give little information
- **Autosomal DNA** is subjected to recombination
- **Uniparental DNA** (**mtDNA**, **Y chromosome**) traces back to mother and father's lineages

Example of Classical marker: blood groups

- Percentage of Blood type 0



Genetic DNA markers

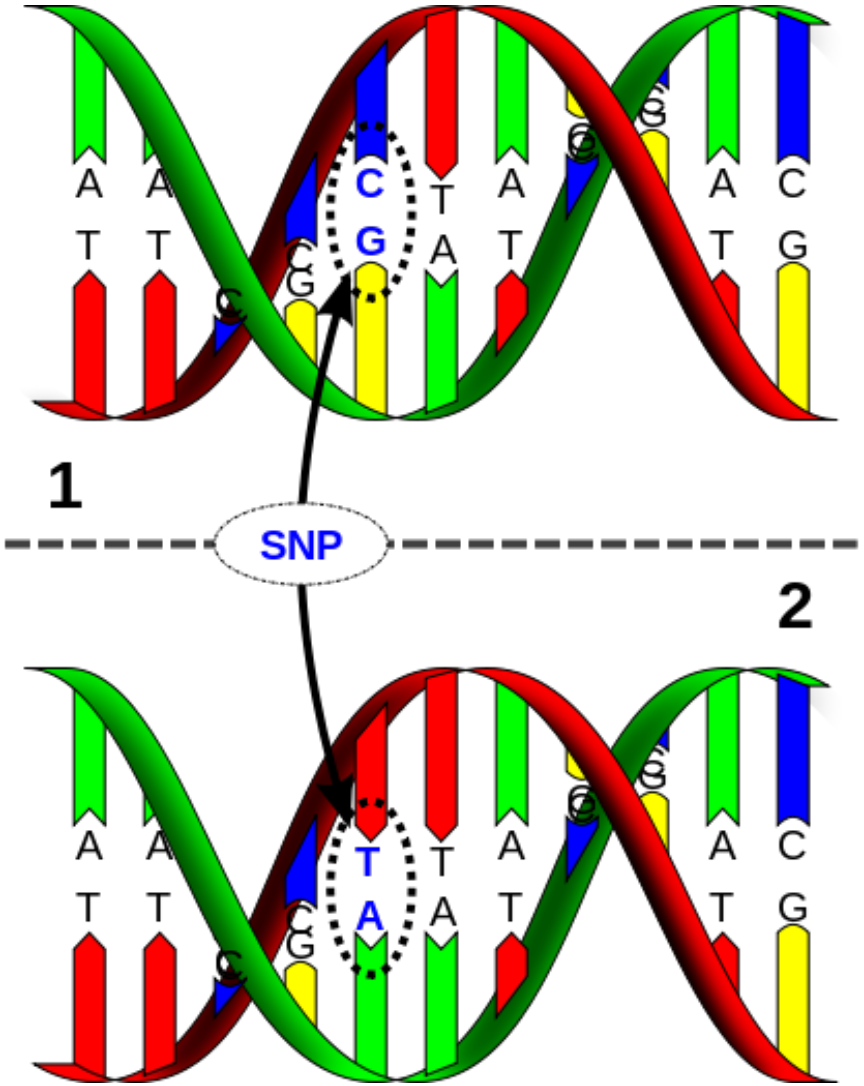
Uniparental markers (Y chromosome, mtDNA)

- We can detect sex differences in prehistoric events (sex-biased gene flow)
- Single lineages can be tracked back in time
- Geographic structure of variation (**phylogeography**)

Autosomal markers

- All the ancestors contribute DNA
- (still) more cost-intensive

DNA sequence data



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

Genetic markers

Uniparental markers:

- **mtDNA** is transmitted only by the mother
- **Y-chromosome** is transmitted only by the father

→ We can detect sex differences in prehistoric events (sex-biased gene flow)

Genetic markers

Uniparental markers:

- No recombination
- Mutations accumulate with time alone
- Shared mutations often indicate shared ancestry

LIMITS

- Small portion of human genome
- Only maternal/paternal view of history

Uniparental sequence data: some definitions

- **Haplotype =** sum of the variable (polymorphic) sites

```
Individual  1:  ACTTGGAAG
Individual  2:  AGTTGCTTG
Individual  3:  ACTTAGTTG
Individual  4:  AGTTACTTG
```

Uniparental sequence data: some definitions

- **Haplotype =** sum of the variable (polymorphic) sites

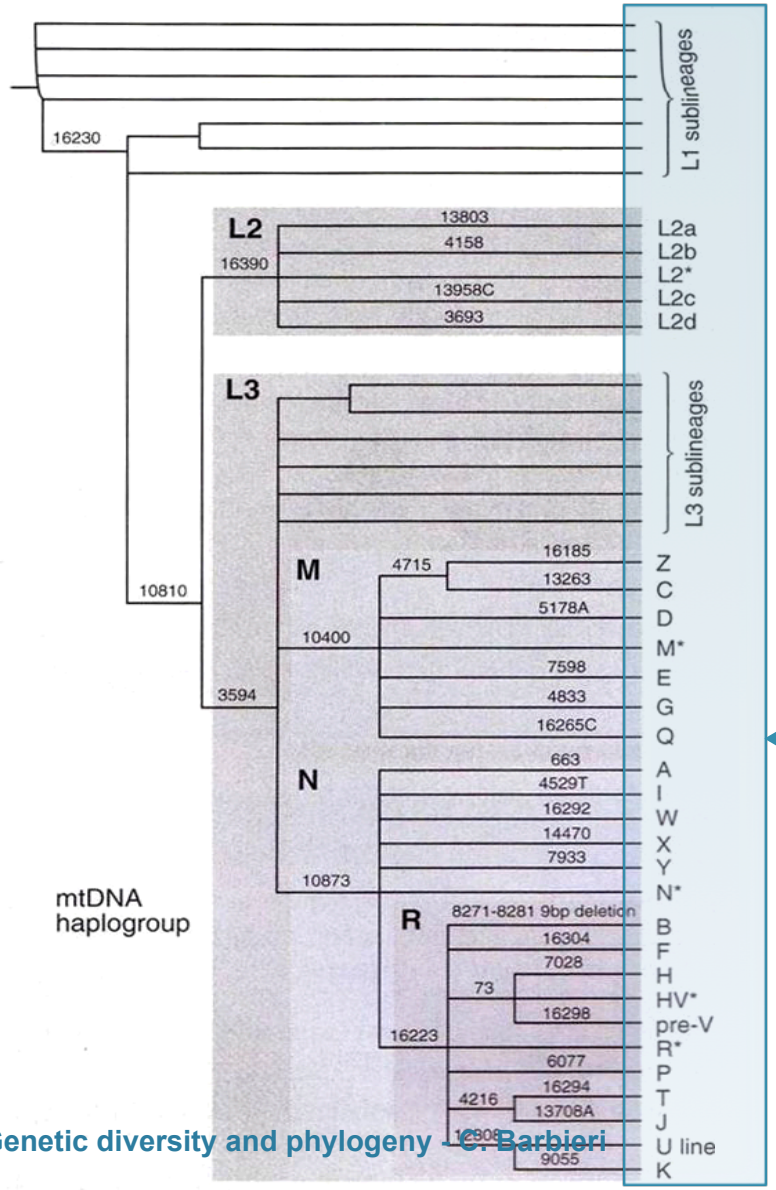
```
Individual 1: ACTTGGAAG
Individual 2: AGTTGCTTG
Individual 3: ACTTAGTTG
Individual 4: AGTTACTTG
```

4 individuals have 4 different haplotypes!

Uniparental sequence data: some definitions

- **Haplogroup =** a group of similar haplotypes that share a **common ancestor**.
  - Defined on the sharing of stable SNP mutations
  - Different haplotypes can belong to the same haplogroup

Haplogroup phylogeny

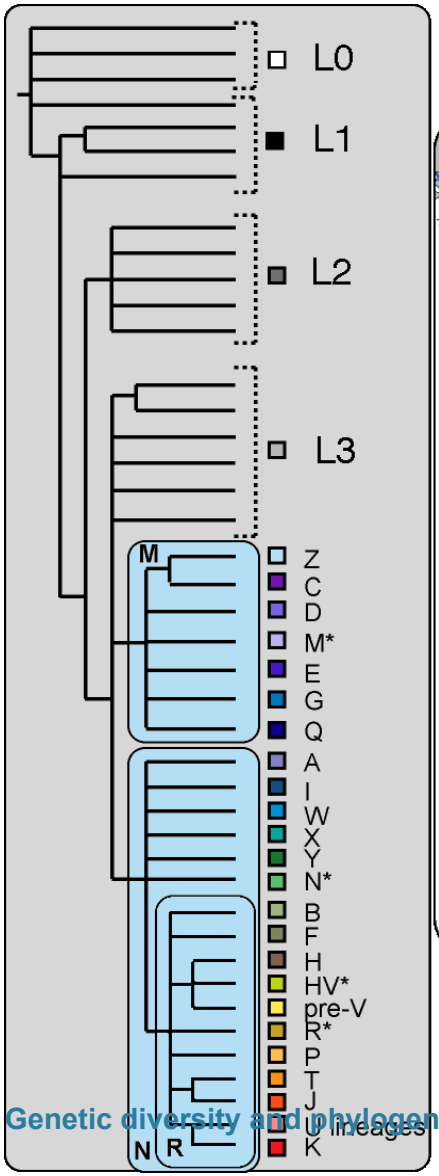


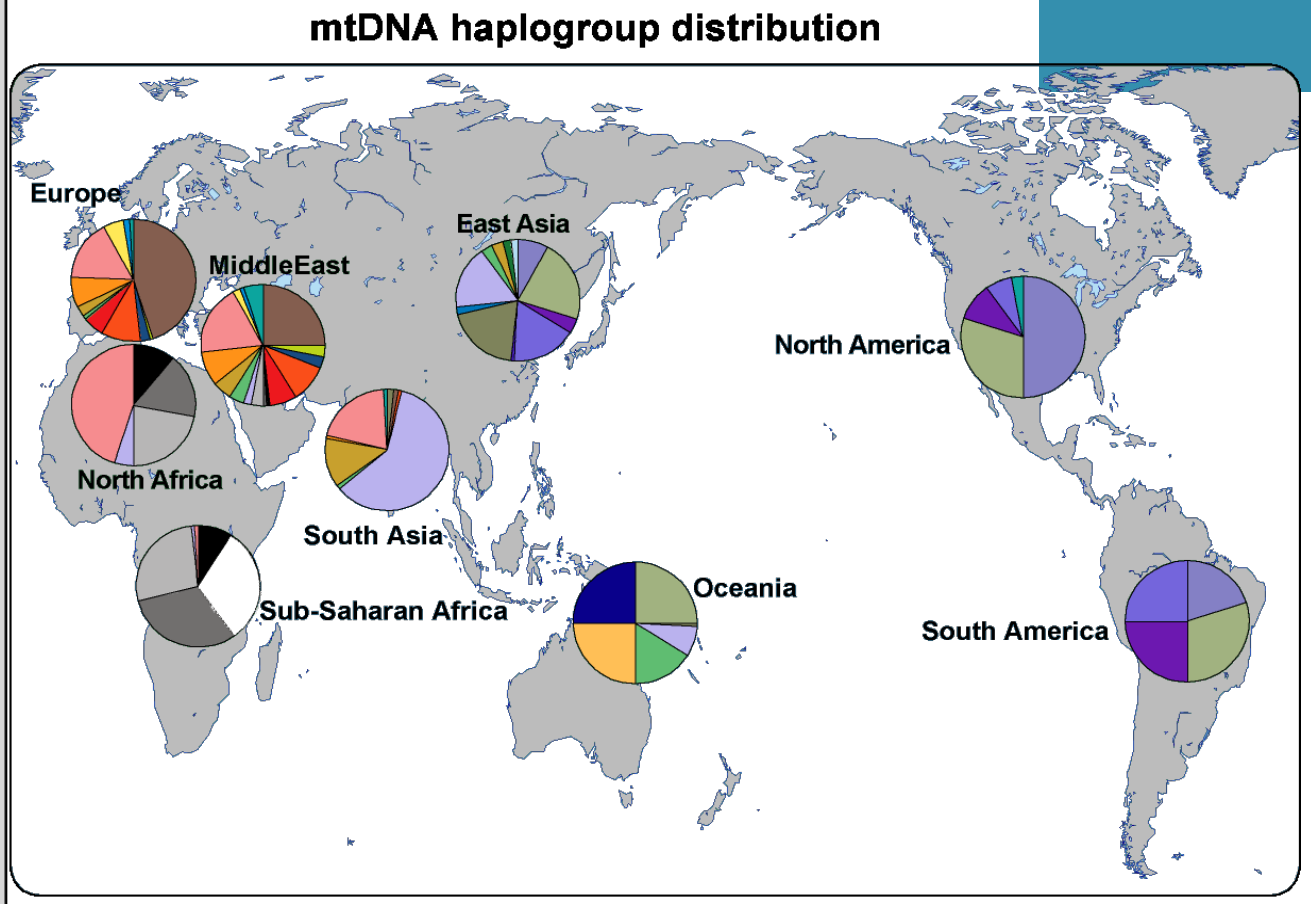
Phylogeny

Tree structure representing evolutionary relationships between clades

Haplogroups indicated with a capital letter

World mtDNA phylogeography

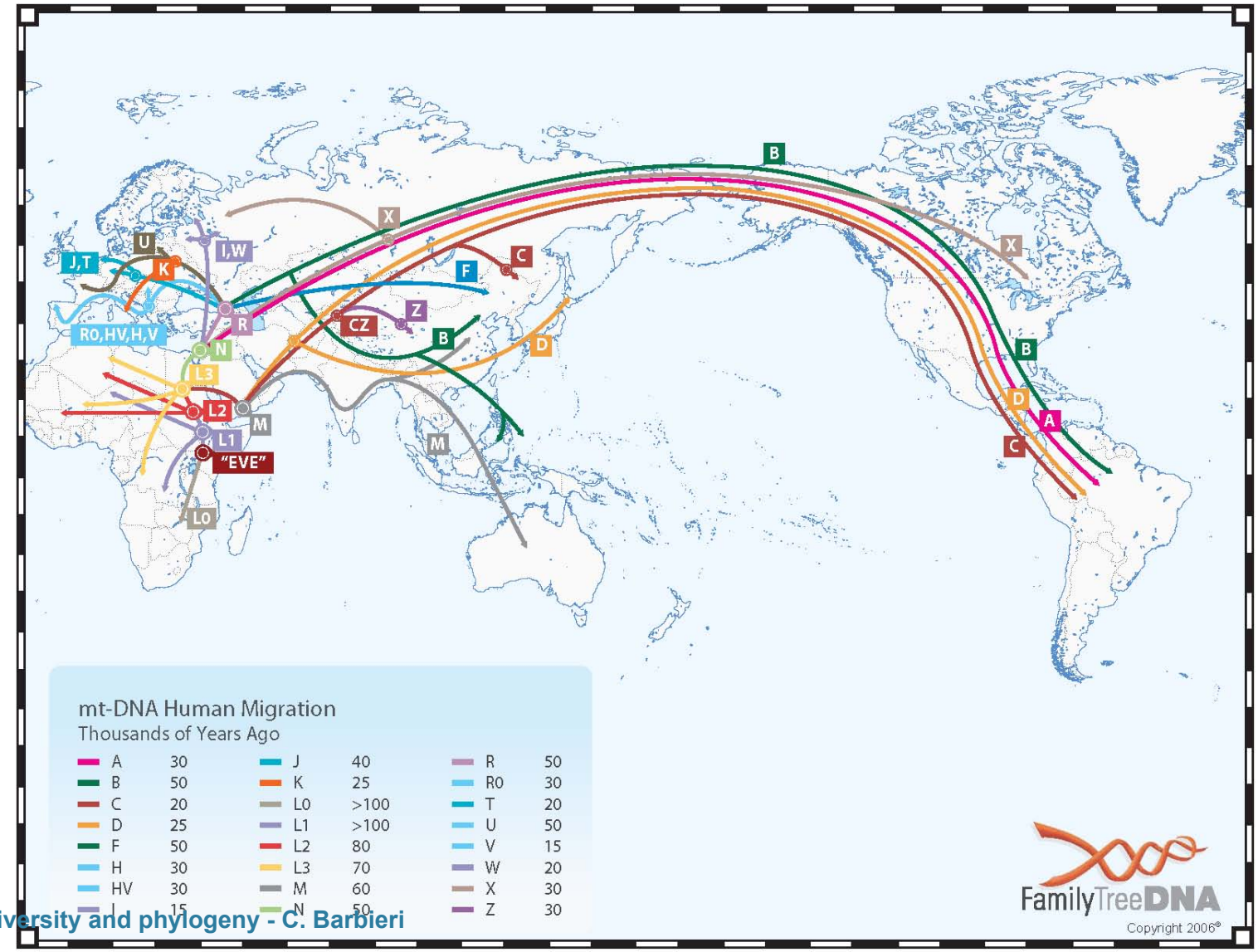




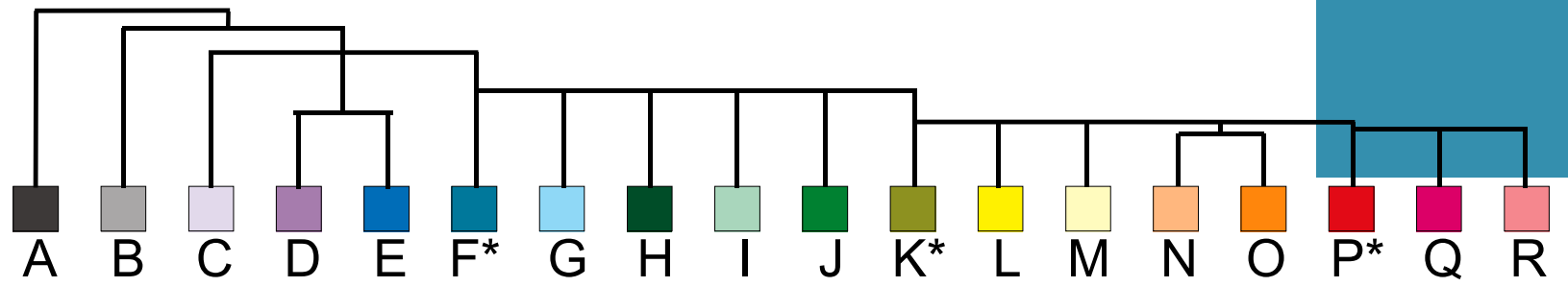
It combines temporal and evolutionary dimension of phylogeny with geographic distribution of haplogroups

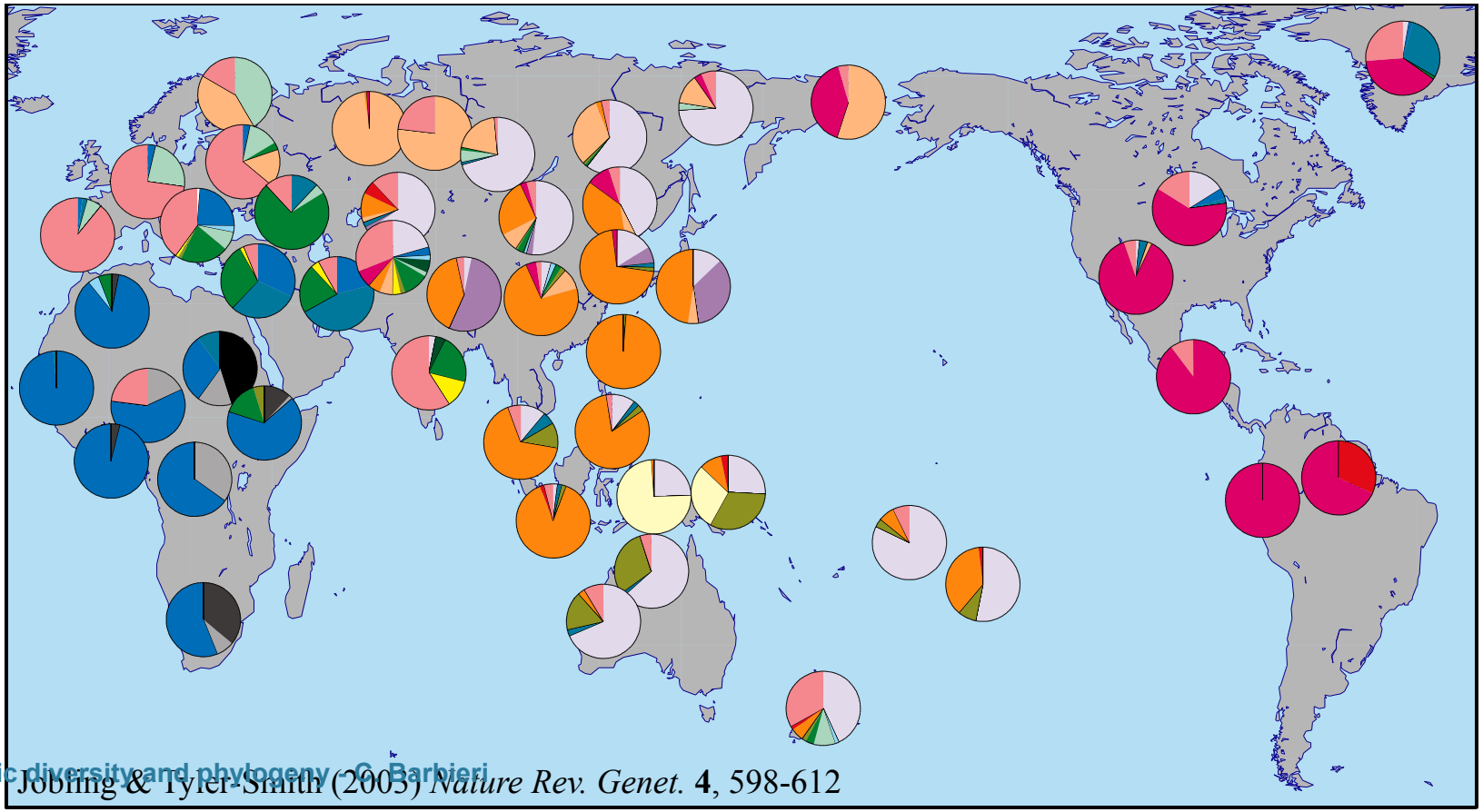
Modified from: www. http://statistics.arizona.edu

World mtDNA phylogeography



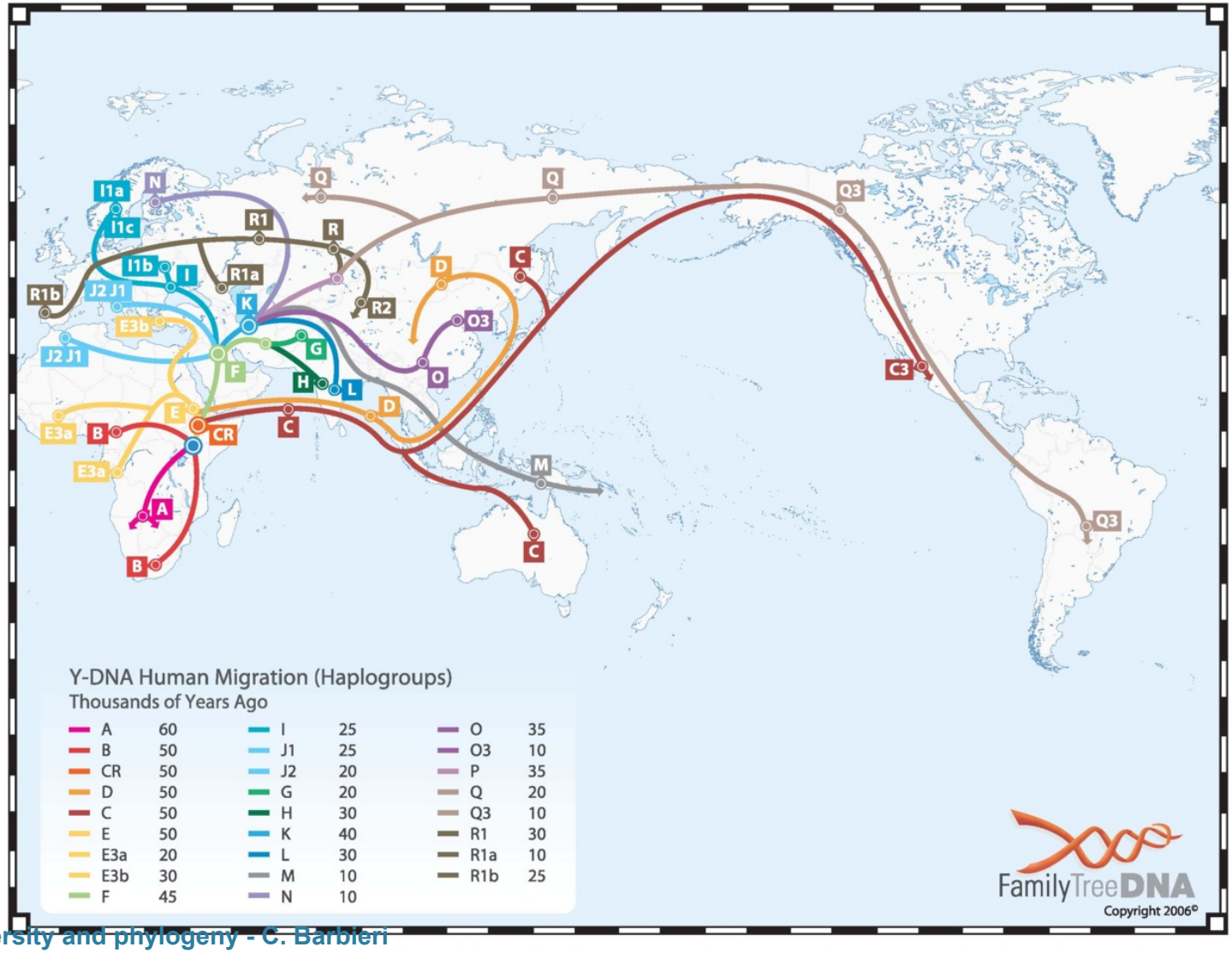
World Y chromosome phylogeography





Jobling & Tyler-Smith (2003) *Nature Rev. Genet.* **4**, 598-612

World Y chromosome phylogeography



Example of Y chromosome and mtDNA (and language)

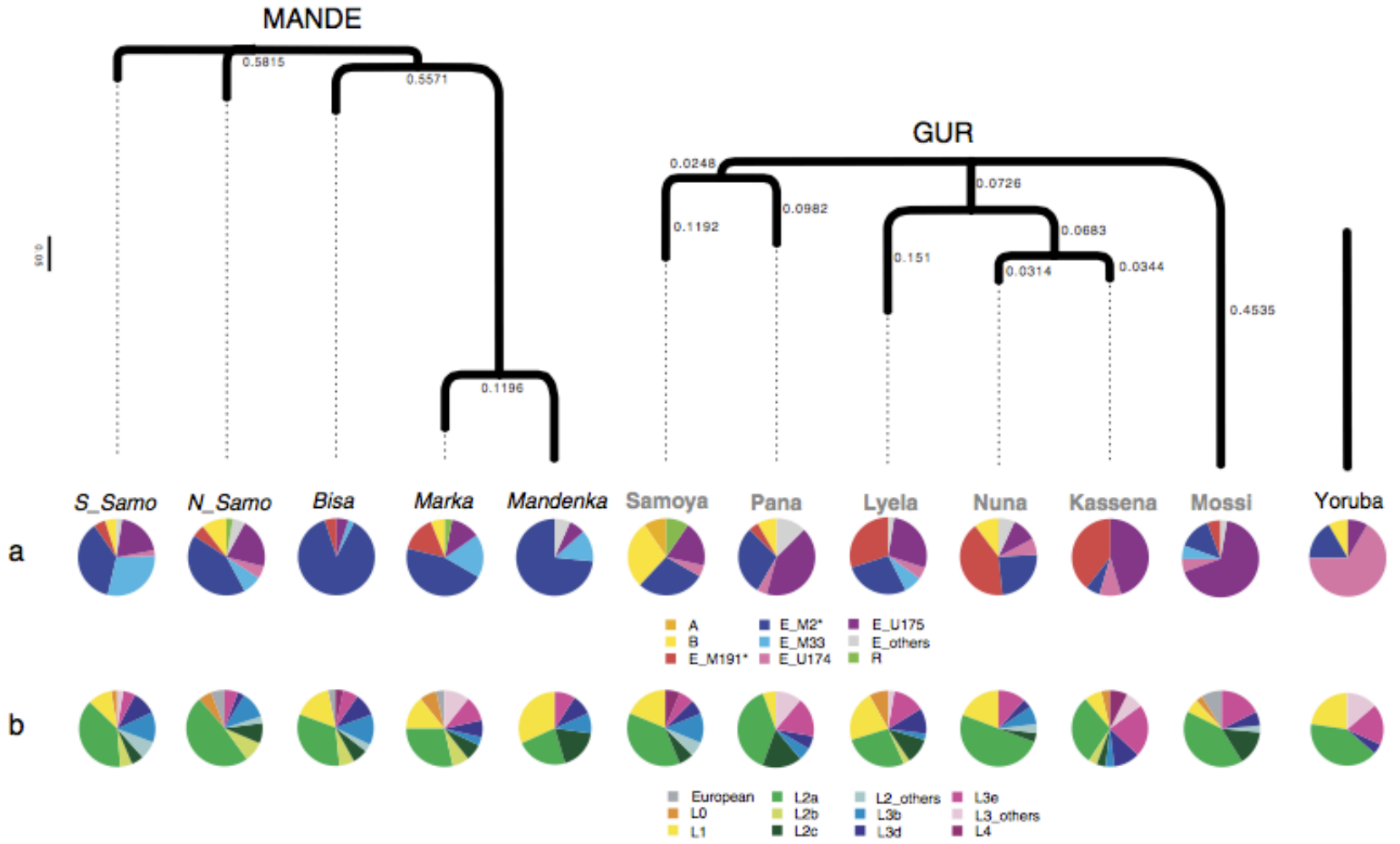
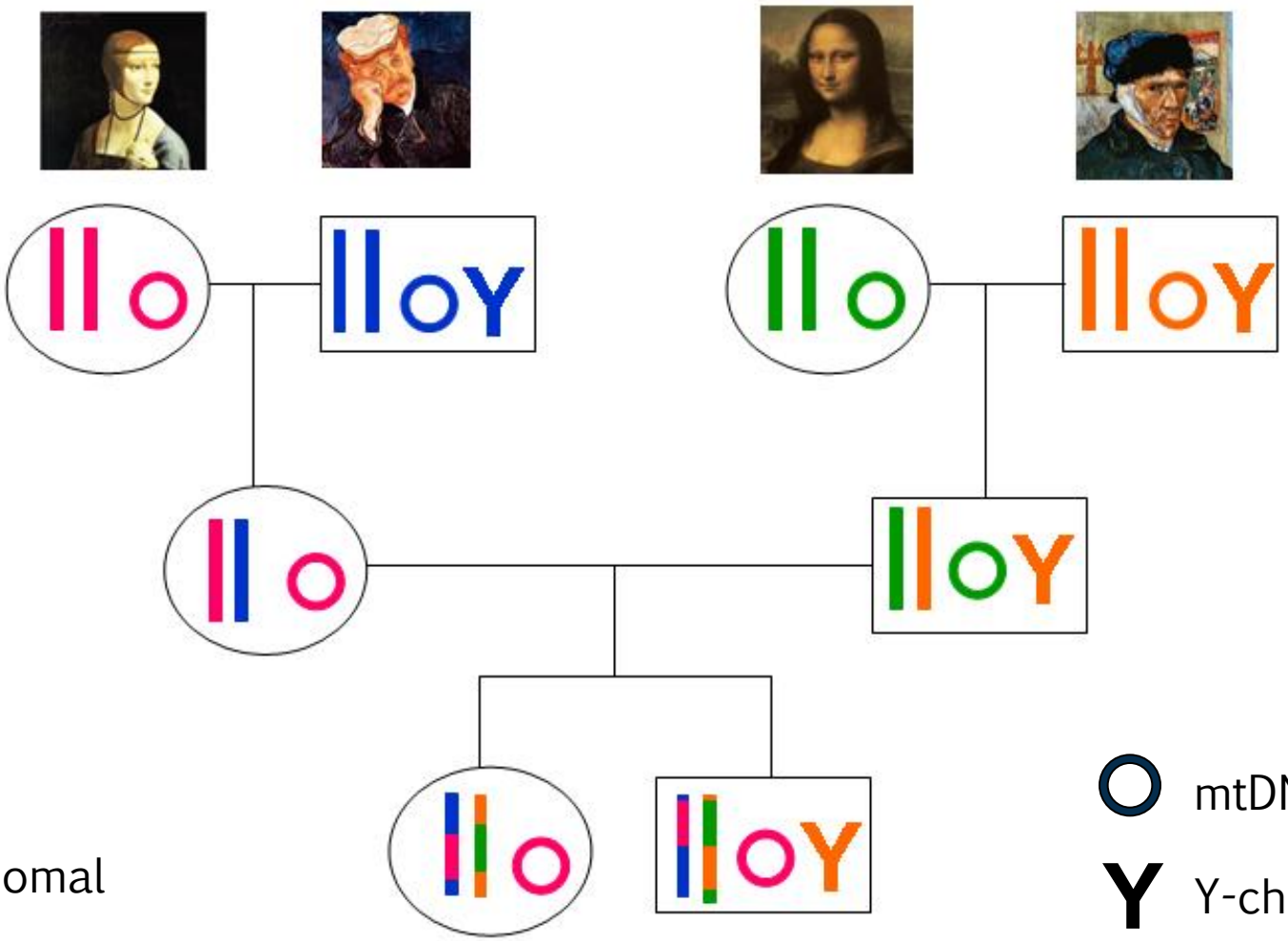


FIG. 3. Haplogroup composition of Mande and Gur populations as well as Yoruba, together with a Neighbour-Joining tree based on linguistic distances. (*a*) Y chromosome. (*b*)mtDNA. In the Y-chromosomal pie charts, haplogroups B-M150 and B-M181 were merged to B, and E-M35,

Barbieri et al. 2012, Mol Biol Evol

Genetic markers: transmission



Genetic markers: Autosomal

Autosomal markers:

- Large amount of genetic information
- Unbiased view of population prehistory (all the ancestors are taken into account)

LIMITS

- Recombination prevents the tracing back of individual mutations through space and time
- (still) more cost-intensive

Autosomal data

- Reconstruct the phylogeny of a certain gene or variant
- Capture a set of independent positions through the genome: SNP chip
  - Ascertained to be variable in human population
- Full genome sequencing

Phylogeny of the lactase allele in southern Africa

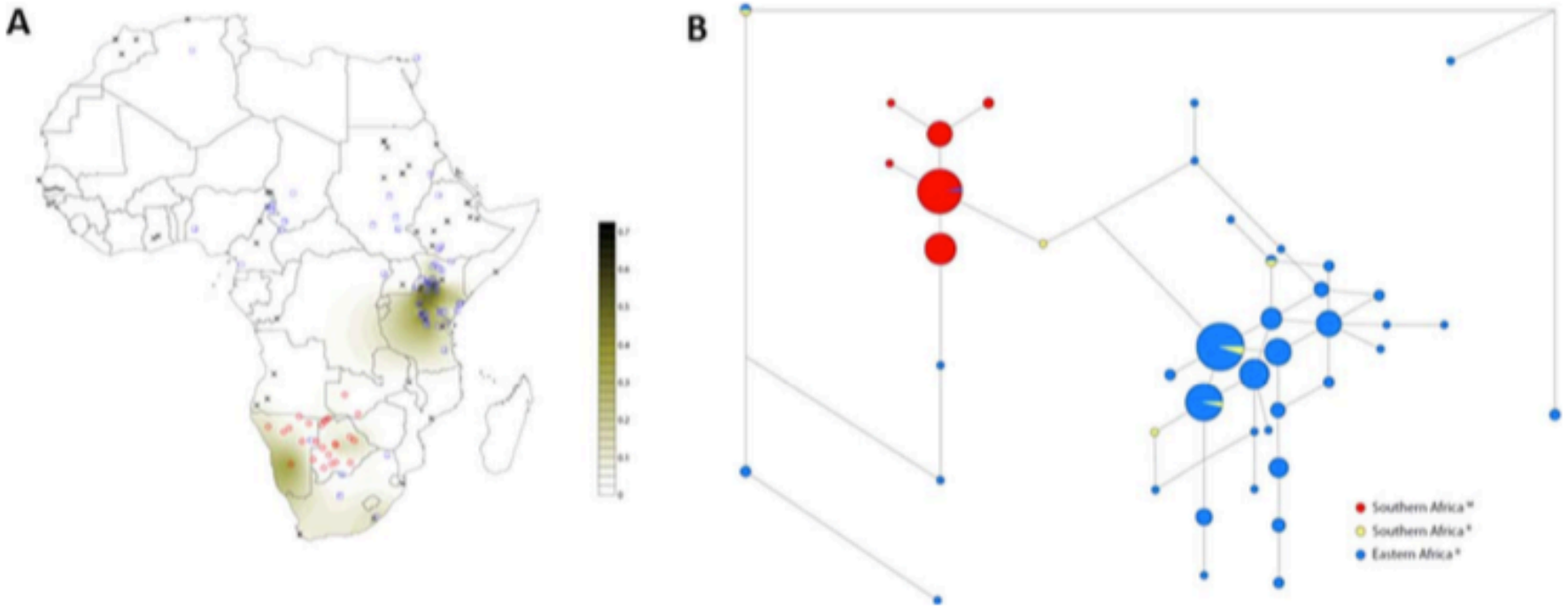
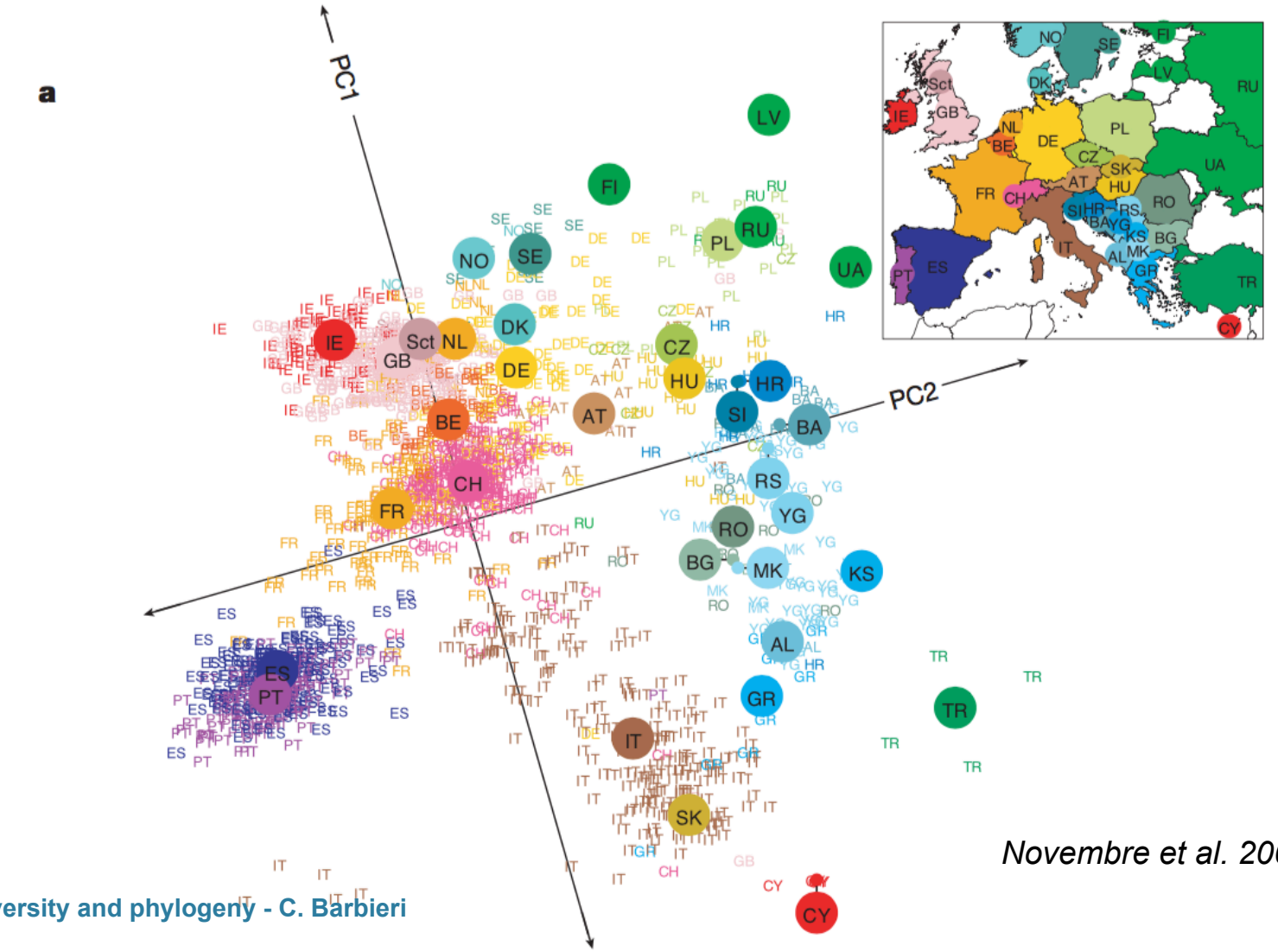


Fig. 1. Analyses of the C-14010 LP variant and associated STR haplotypes in Eastern and Southern African populations. **A**: Surfer map of the C-14010 allele frequency. Red circles denote sampling locations by Macholdt et al.; blue squares denote sampling locations by Ranciaro et al.; black crosses denote 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

Autosomal data: treemix in Europe

Lazaridis et al. 2014 Nature

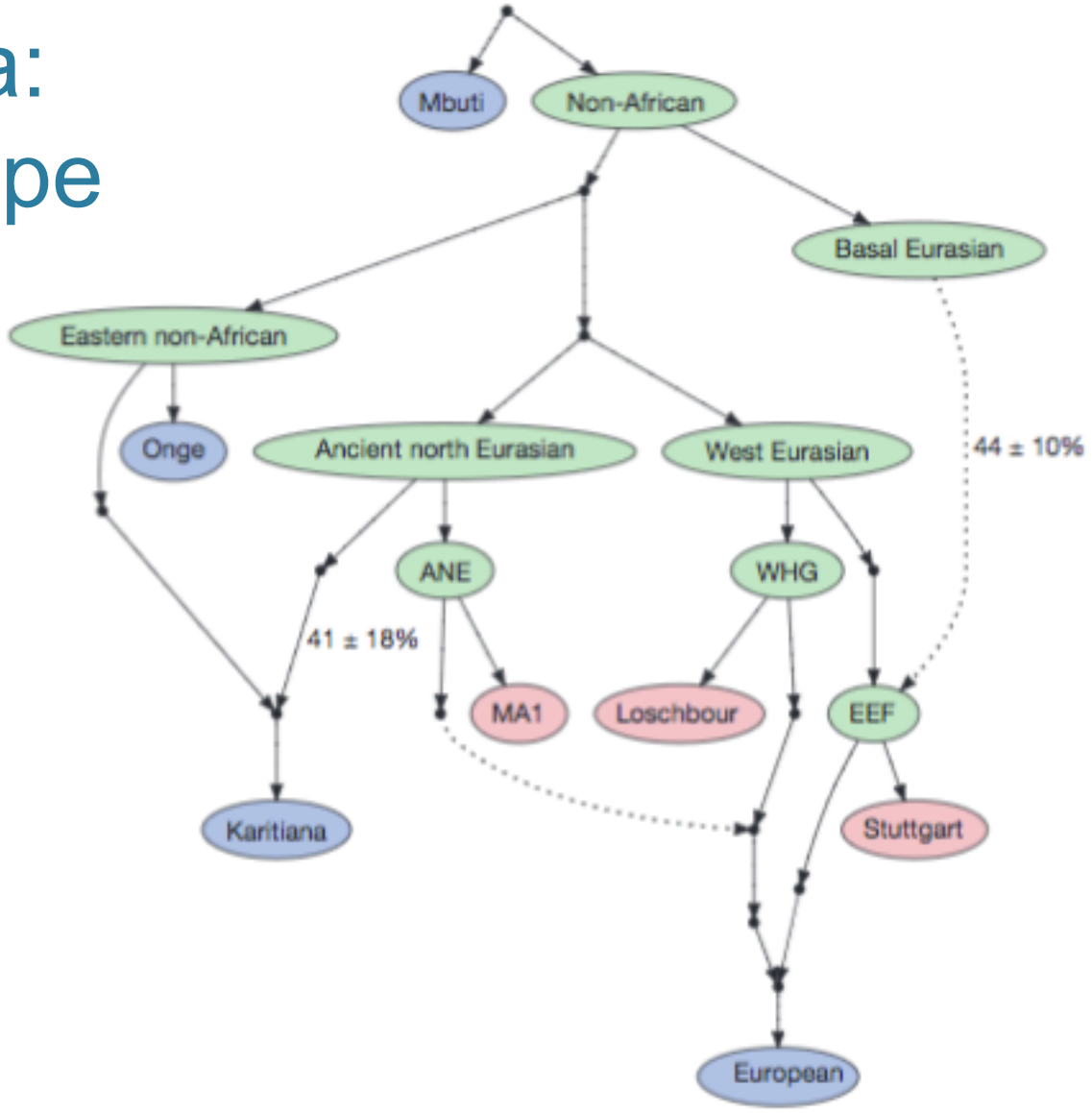
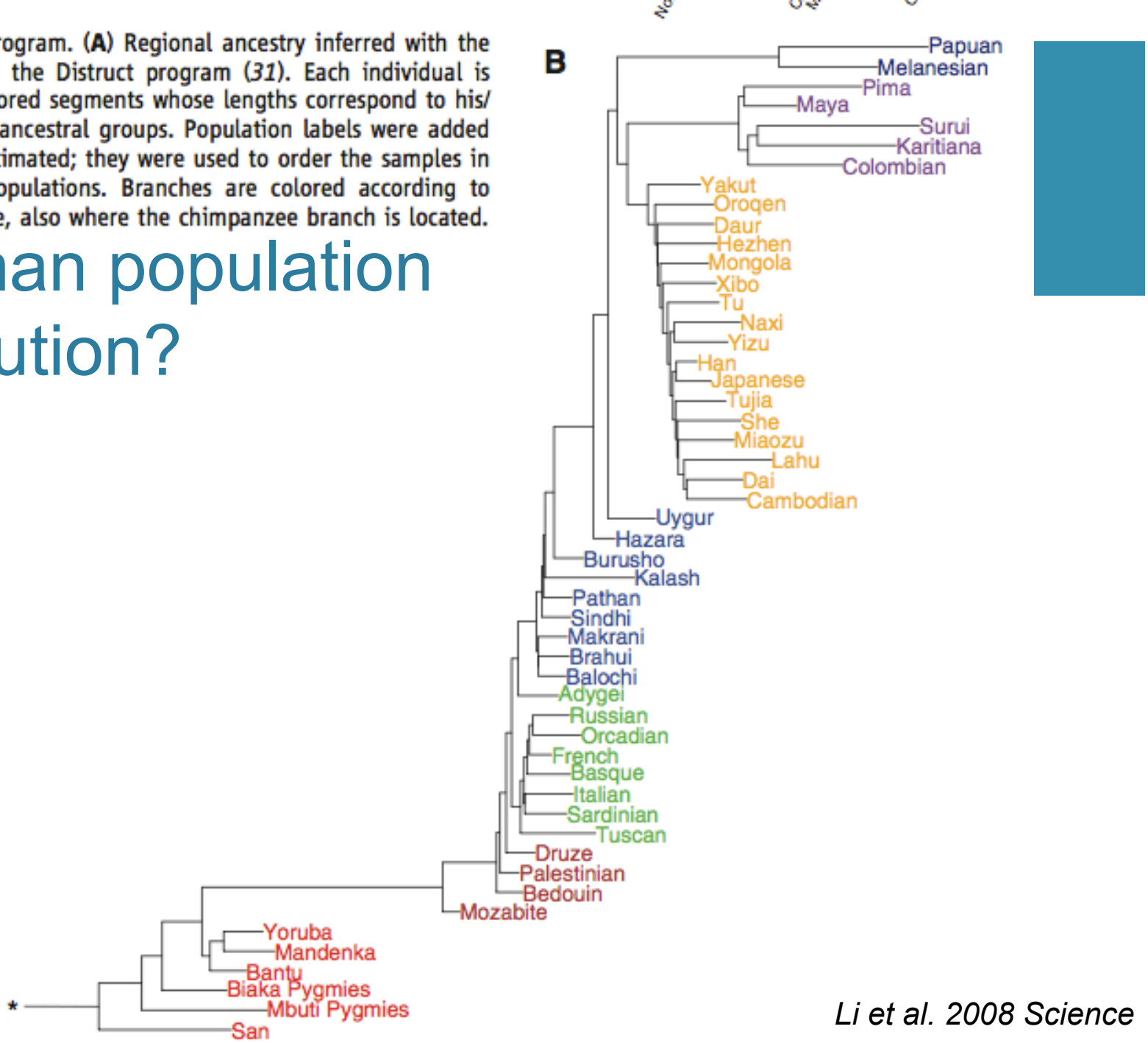


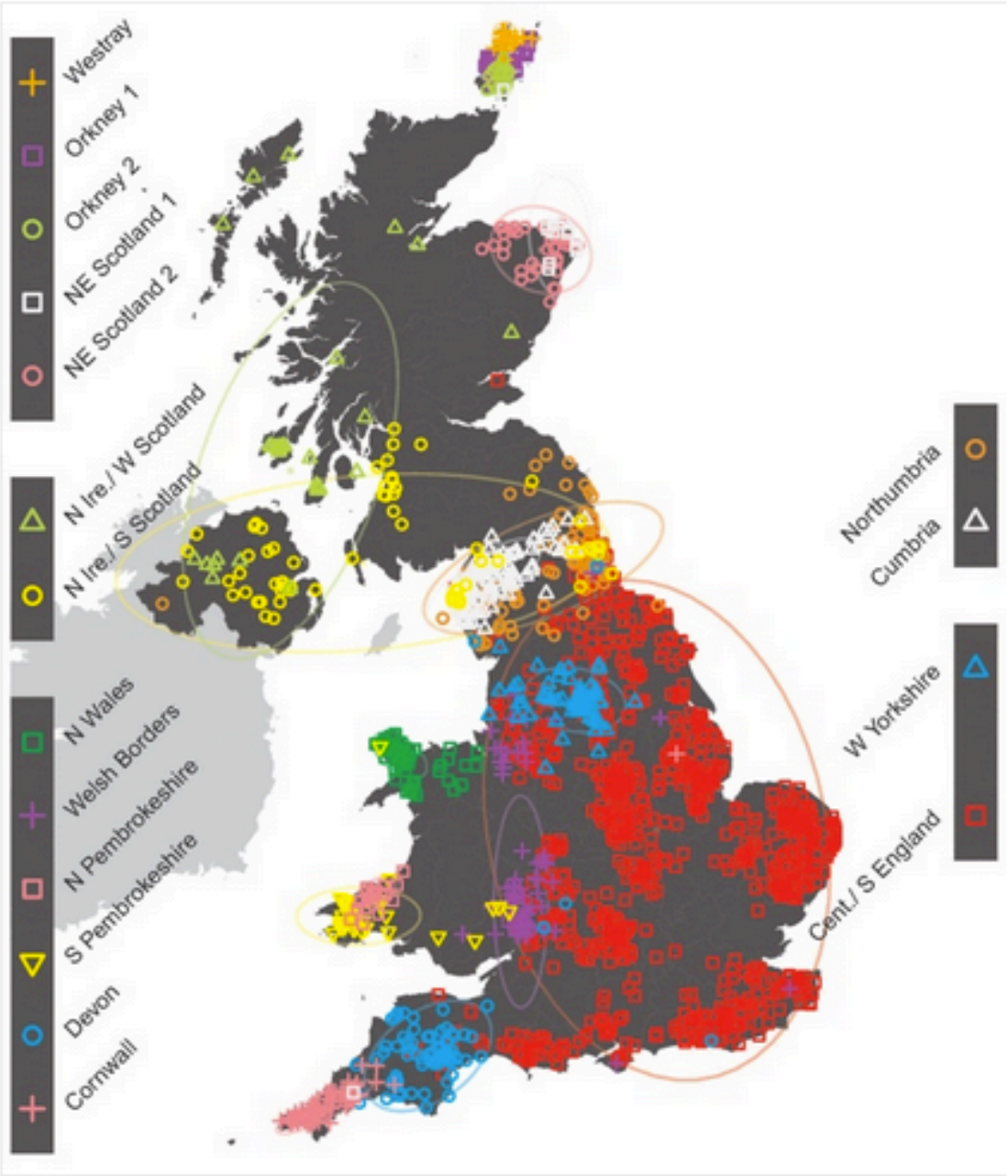
Figure 3 | Modelling the relationship of European to non-European populations. A three-way mixture model that is a fit to the data for many populations. Present-day samples are coloured in blue, ancient in red, and reconstructed ancestral populations in green. Solid lines represent descent without mixture, and dashed lines represent admixture. We print mixture

dendrogram. (**A**) Regional ancestry inferred with the
with the Distruct program (*31*). Each individual is
o colored segments whose lengths correspond to his/
rred ancestral groups. Population labels were added
n estimated; they were used to order the samples in
51 populations. Branches are colored according to
e tree, also where the chimpanzee branch is located.

Human population evolution?



Li et al. 2008 Science



Stephen Leslie

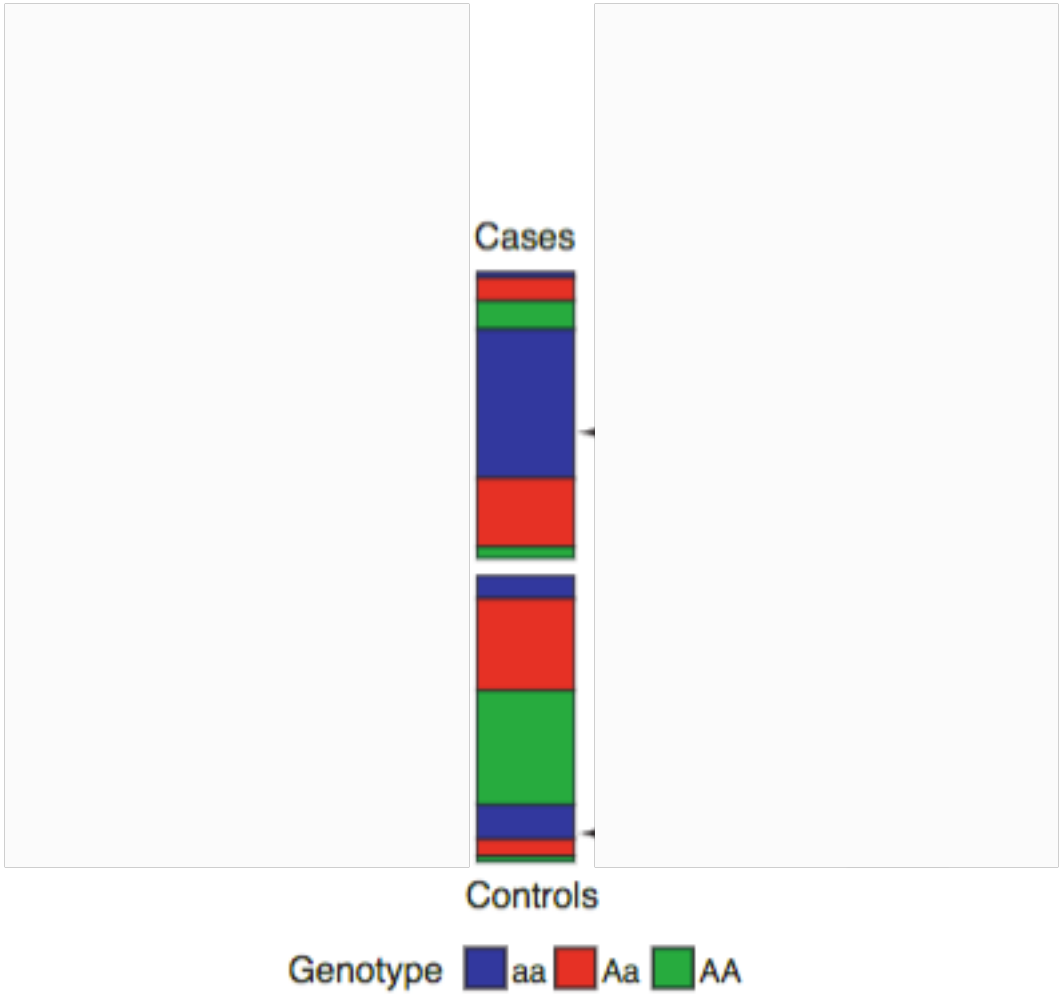
A map of the United Kingdom shows how individuals cluster based on their genetics, with a striking relationship to the geography of the country.

Gene

Leslie et al. 2015 Nature

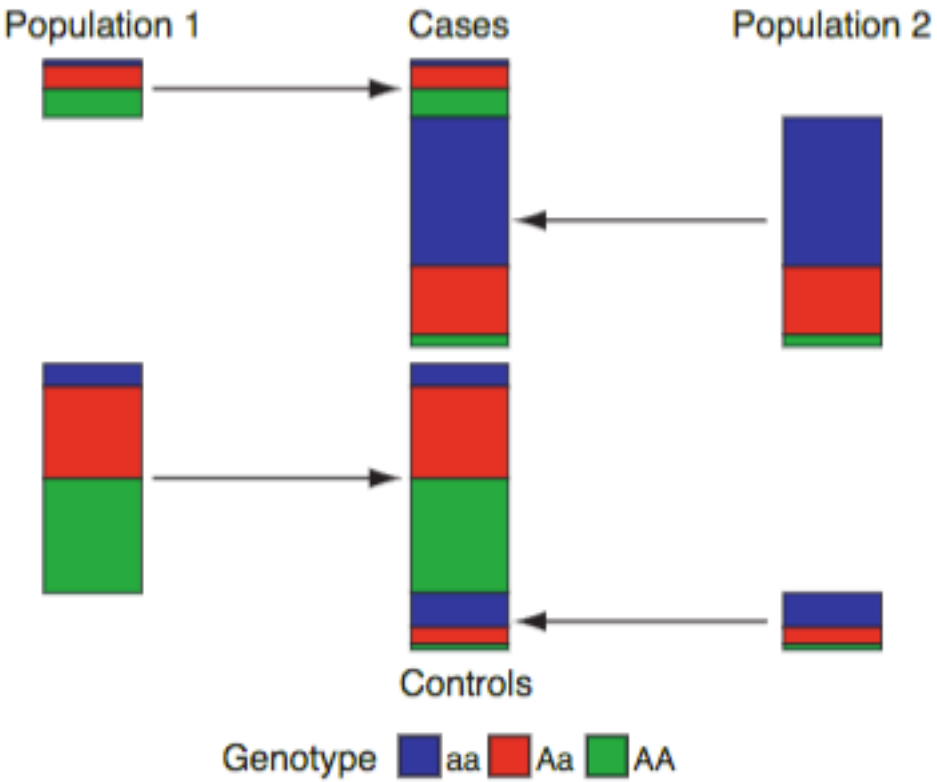
Population substructure

- Important for medical association studies
- Target gene(s) associated to an illness, or other phenotypic trait?
- Hidden structure (relatedness between separate groups) brings false positive results and failures to detect genuine associations
- Effects increase with sample size



Example of effects of pop structure a SNP locus. Two populations in which the cases have an excess of individuals from population 2 and population 2 has a lower frequency of allele A than population 1. In this example, the structure mimics the signal of association in that there is a significant difference in allele and genotype frequencies between cases and controls.

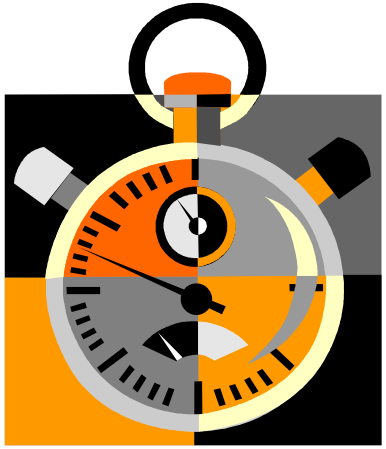
Marchini et al. 2004 Nature Genetics



Example of effects of pop structure a SNP locus. Two populations in which the cases have an excess of individuals from population 2 and population 2 has a lower frequency of allele A than population 1. In this example, the structure mimics the signal of association in that there is a significant difference in allele and genotype frequencies between cases and controls.

Marchini et al. 2004 Nature Genetics

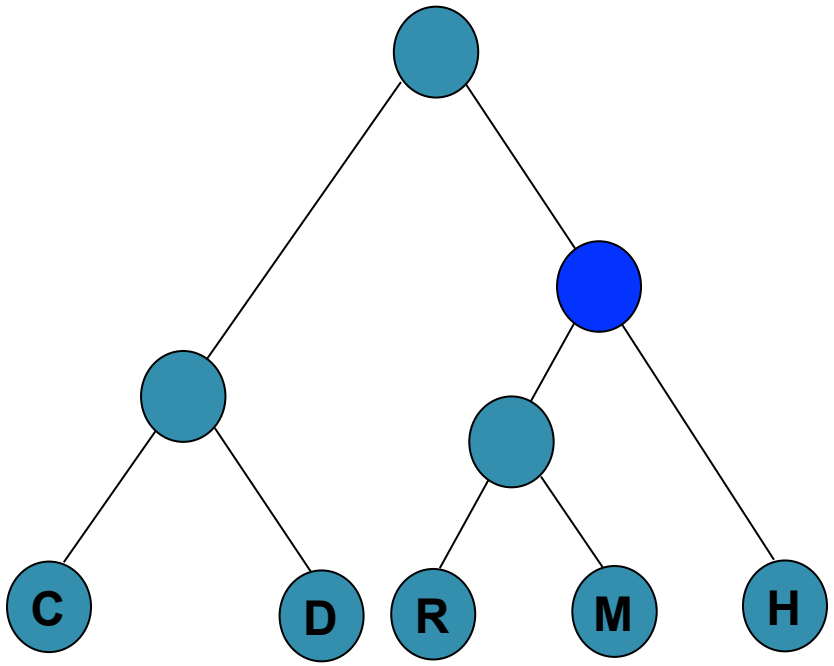
Considering relatedness is important



Molecular clock

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



110 MYA

Given

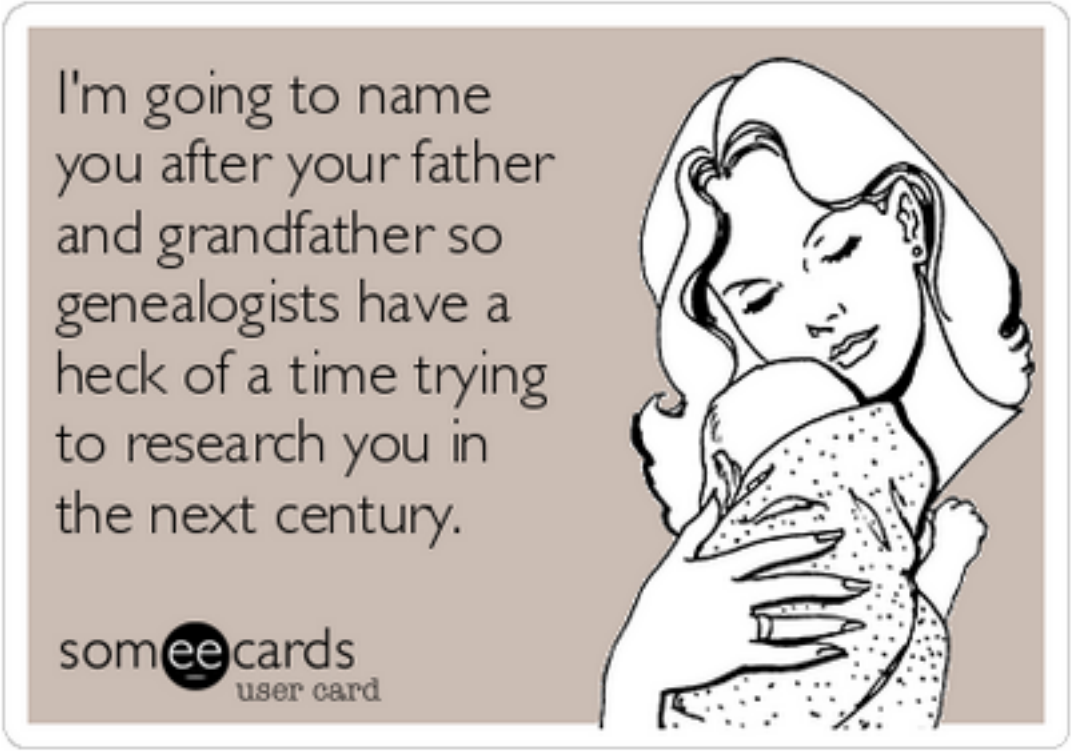
- a phylogenetic tree
- branch lengths (rt)
- a time estimate for one (or more) node (CALIBRATION!!!)

- Can we date other nodes in the tree?

- **Yes**... if the rate of molecular change is *constant* across all branches

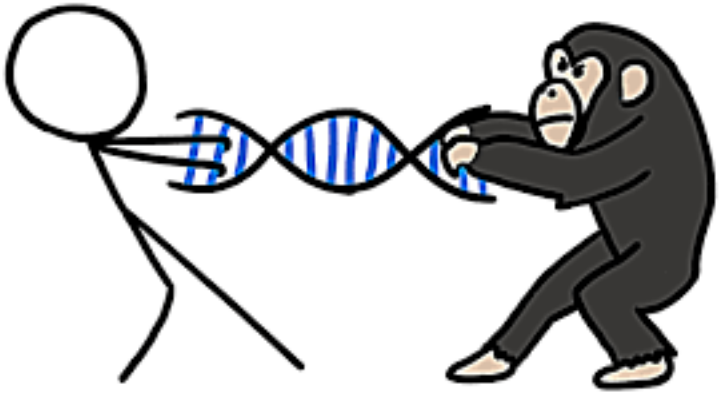
How to calibrate (in humans)

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



How to calibrate (in humans)

- 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)

- Direct calibration with inclusion of aDNA from properly dated fossils

Often, mutation rate is heterogeneous

- Along branches, or nodes
- Need the use of a relaxed clock model

Patrick Stewart and John Hurt — 74 years old

Rate Heterogeneity among Lineages

Cause	Reason
Repair equipment	e.g. RNA viruses have error-prone polymerases
Metabolic rate	More free radicals
Generation time	Copies DNA more frequently
Population size	Effects mutation fixation rate

Example of phylogeny with calibration: Y chromosome

Karmin et al. 2015 Genome Research

- Chr Y data of the 12.6-ky-old Anzick (Q1b) and 4-ky-old Saqqaq (Q2b) specimens (Rasmussen et al. 2010, 2014)
- Estimate mutation rate with a strict clock

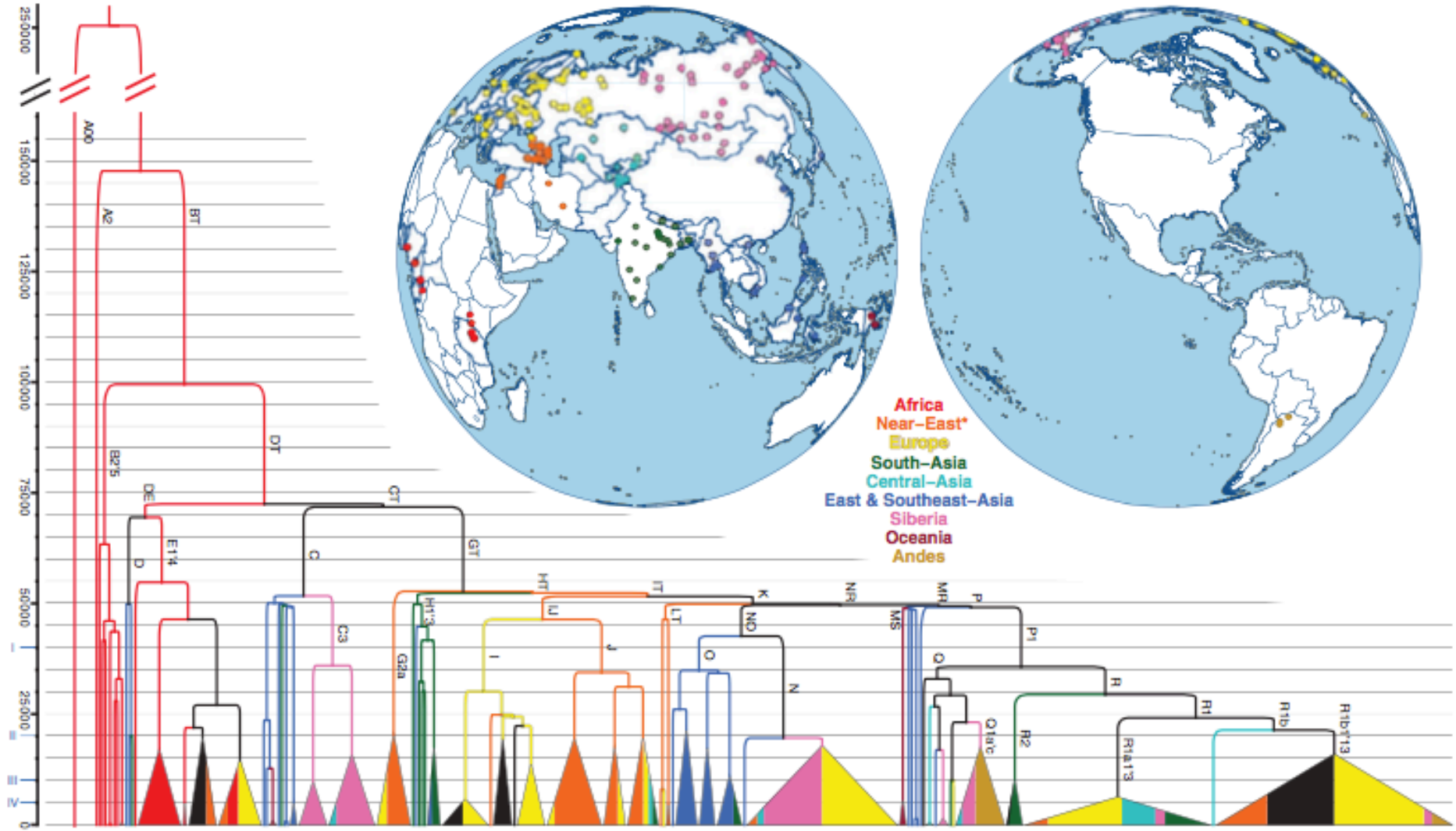


Figure 1. The phylogenetic tree of 456 whole Y chromosome sequences and a map of sampling locations. The phylogenetic tree is reconstructed using BEAST. Clades coalescing within 10% of the overall depth of the tree have been collapsed. Only main haplogroup labels are shown (details are provided in Supplemental Information 6). Colors indicate geographic origin of samples (Supplemental Table S1), and fill proportions of the collapsed clades represent the proportion of samples from a given region. Asterisk (*) marks the inclusion of samples from Caucasus area. Personal Genomes Project (http://www.personalgenomes.org) samples of unknown and mixed geographic/ethnic origin are shown in black. The proposed structure of Y chromosome haplogroup naming (Supplemental Table S5) is given in Roman numbers on the *y*-axis.

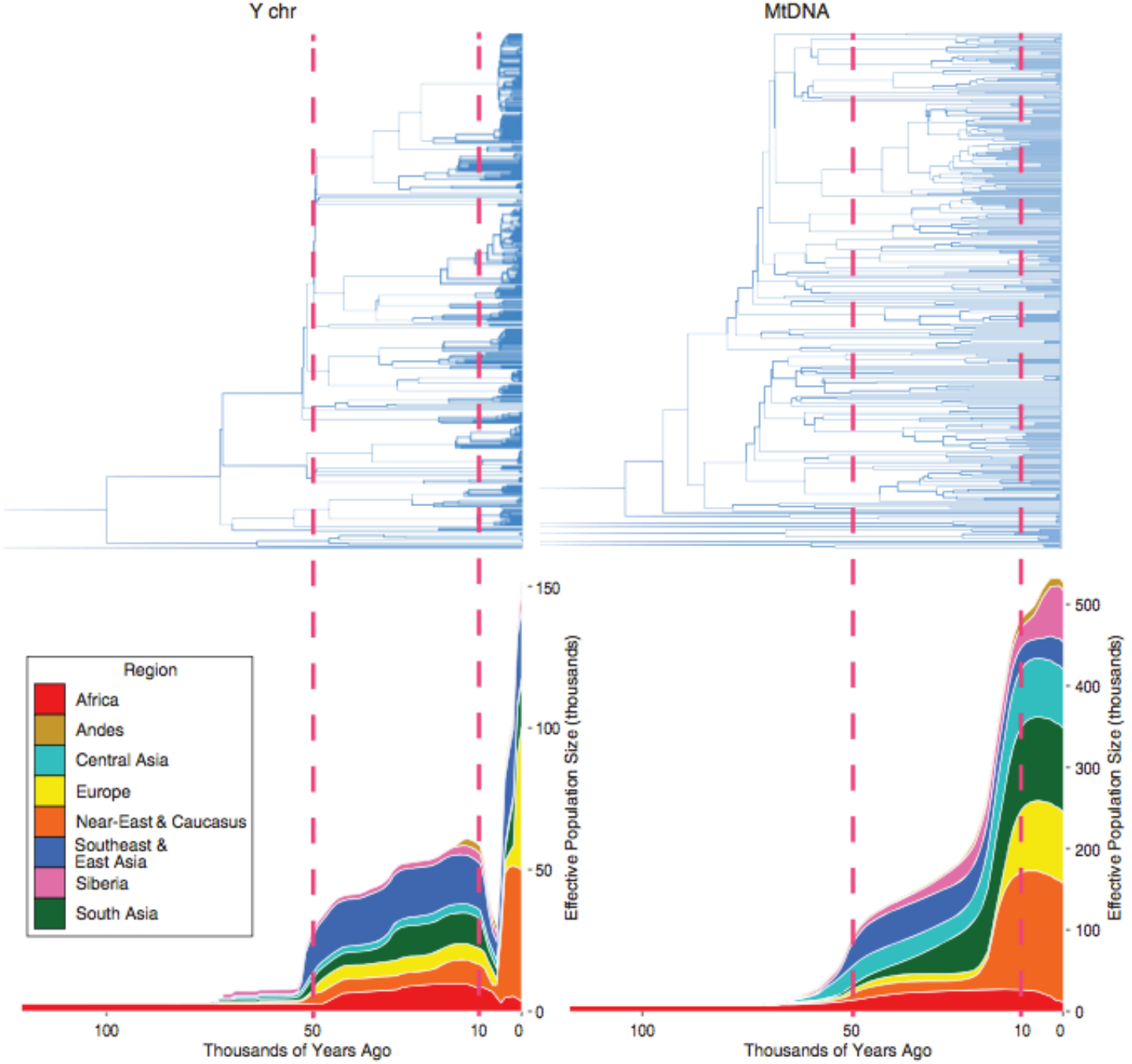


Figure 2. Cumulative Bayesian skyline plots of Y chromosome and mtDNA diversity by world regions. The red dashed lines highlight the horizons of 10 kya and 50 kya. Individual plots for each region are presented in Supplemental Figure S4A.

Example of phylogeny with calibration: primates

Prado-Martinez et al. 2013 Nature

- Full genomes of primates (including humans)
- Assumption of a human chimpanzee split of 6 million years, 25 years per generation in humans
- → Estimate divergence time between species and population sizes

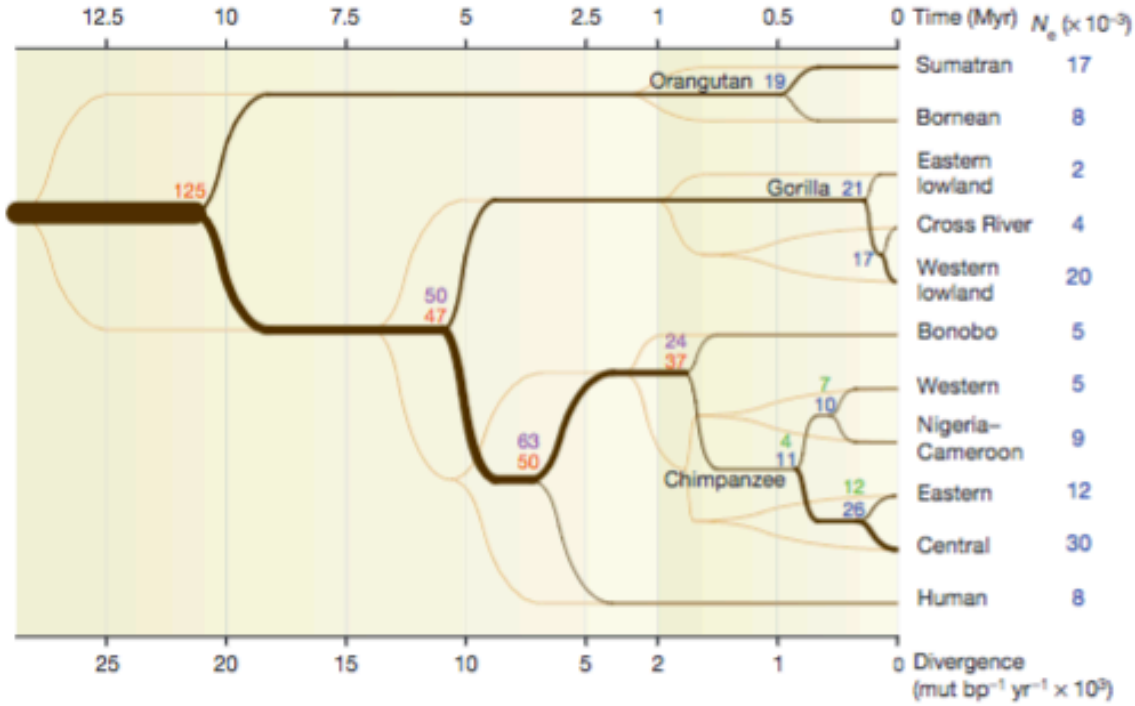
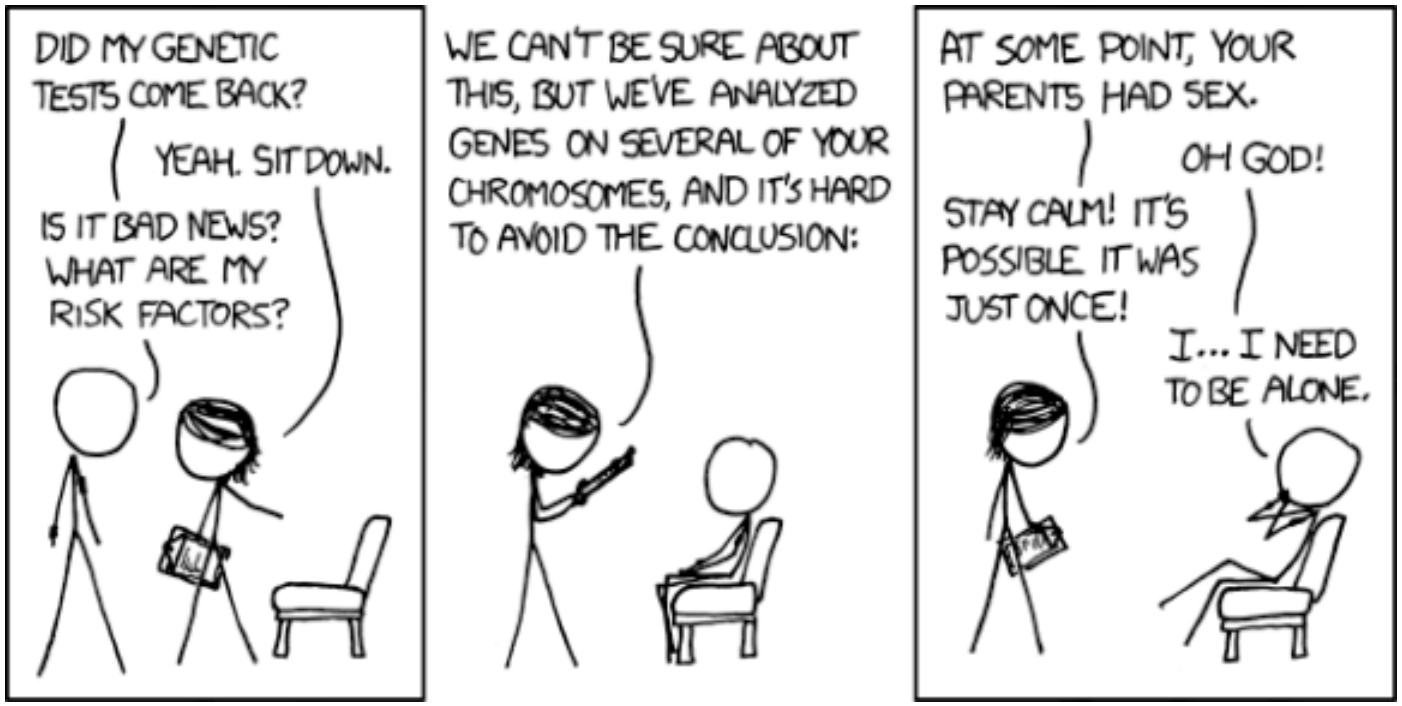


Figure 2 | Inferred population history. Population splits and effective population sizes (N_e) during great ape evolution. Split times (dark brown) and divergence times (light brown) are plotted as a function of divergence (d) on the bottom and time on top. Time is estimated using a single mutation rate (μ) of 1×10^{-9} mut bp^{-1} year^{-1}. The ancestral and current effective population sizes are also estimated using this mutation rate. The results from several methods used to estimate N_e (COALHMM, ILS COALHMM, PSMC and ABC) are coloured in orange, purple, blue and green, respectively. The chimpanzee split times are estimated using the ABC method. The *x* axis is rescaled for divergences larger than 2×10^{-3} to provide more resolution in recent splits. All the values used in this figure can be found in Supplementary Table 5. The terminal N_e correspond to the effective population size after the last split event.

Summarizing:

- Uniparental markers standardized lab techniques, easy to analyze
  - **mtDNA** often employed
  - Limited view of the whole demographic processes
  - Idea of **Haplogroup** and phylogeographic reconstruction
- Autosomal data: more information, complex phylogenies
  - The era of full genomes is next!
- More sequence data retrieved allows accurate hypothesis testing
- Molecular clock allows to date divergence, can be calibrated


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.

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

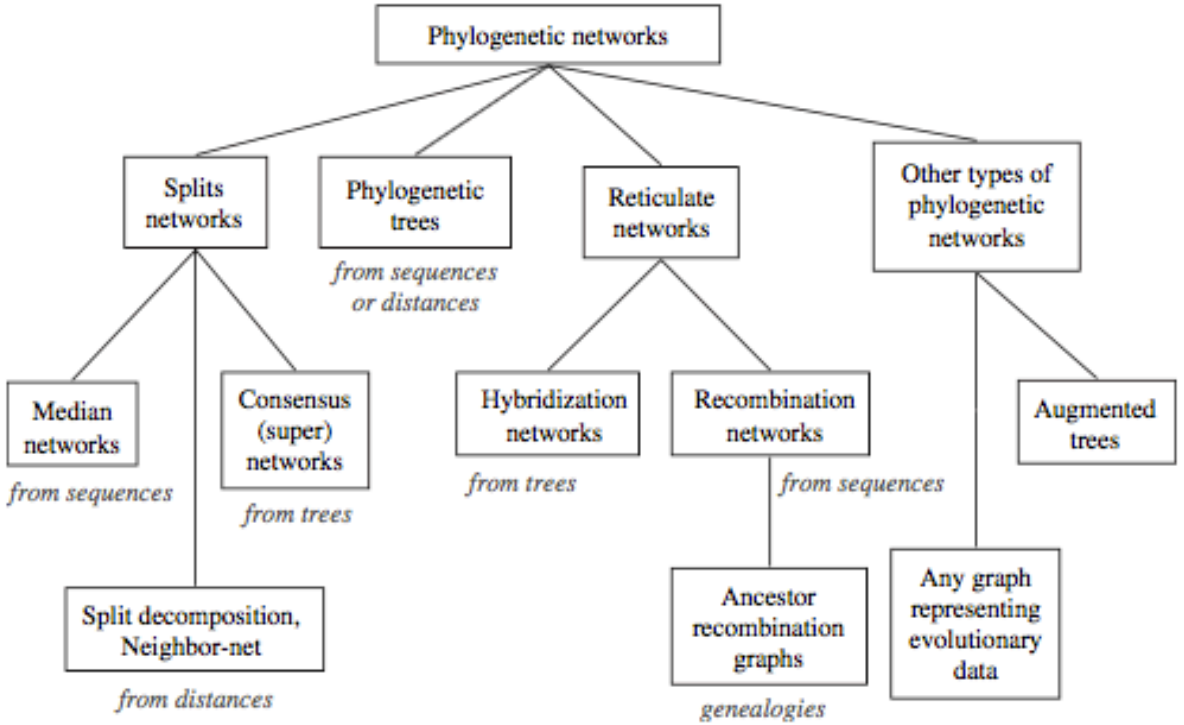
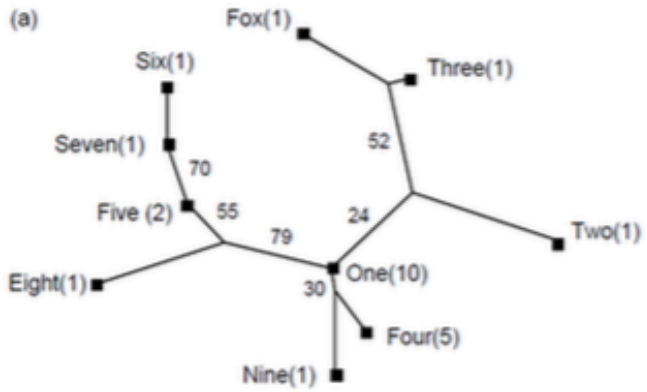


FIG. 1.—The term phylogenetic network encompasses a number of different concepts, including phylogenetic trees, split networks, reticulate networks, the latter covering both "hybridization" and "recombination" networks, and other types of networks such as "augmented trees." Recombination networks are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

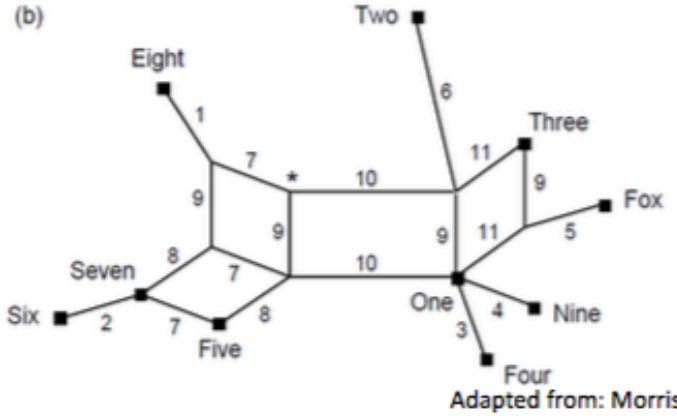
- Trees impose bifurcations
- Networks allow reticulations

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

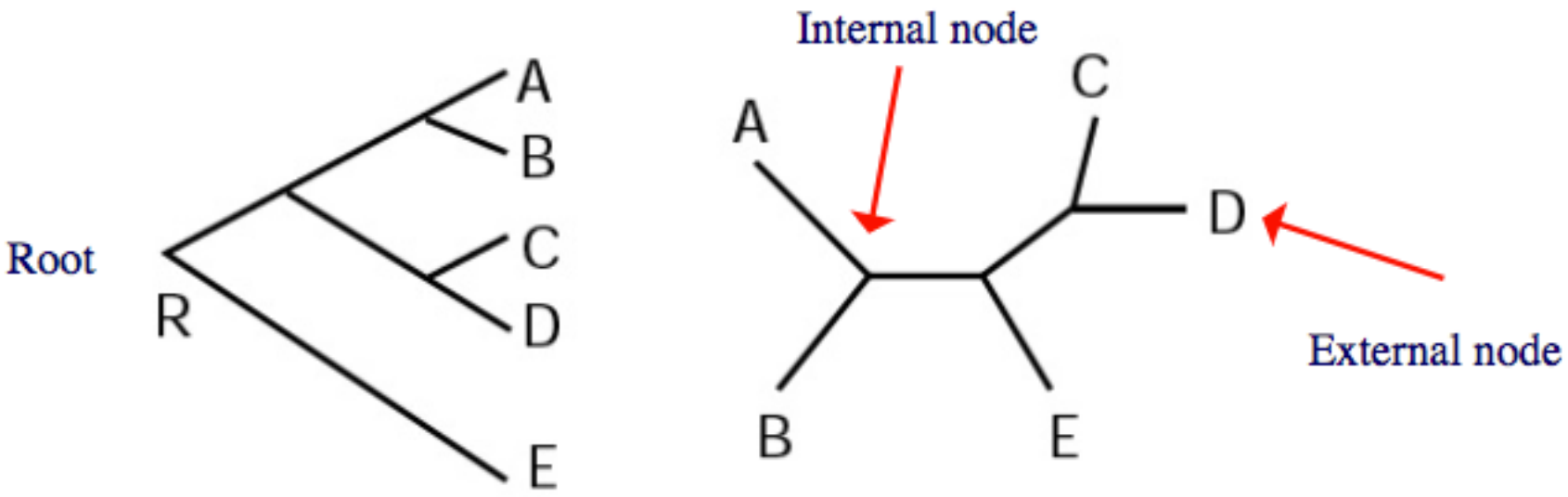
is a connected graph, again with some of the nodes labelled. In a network a set of (parallel) edges (branches) may be required to partition the graph into two connected subgraphs (so the graph appears 'box-like' as in figure.

Data for Phylogeny

- Numerical data
  - Distance between objects
  - e.g., distance(man, mouse)=500,
    distance(man, chimp)=100
- Discrete characters
  - Each character has finite number of states
  - e.g., number of legs = 1, 2, 4

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, minimum evolution, ML, 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

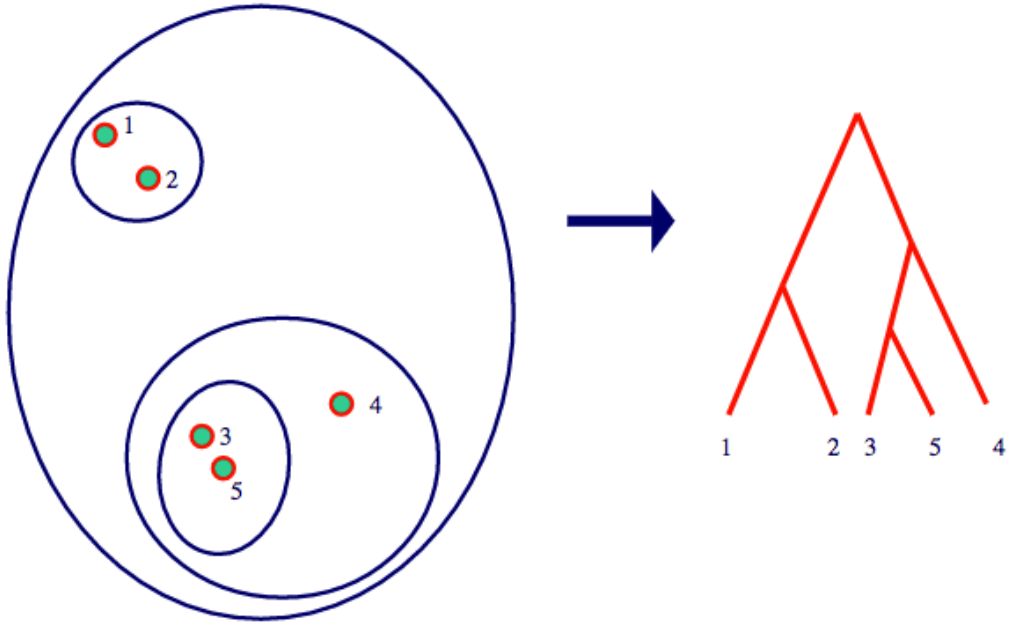
- First calculate distance matrix between pairs of sequences or populations
- Then build a tree

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.

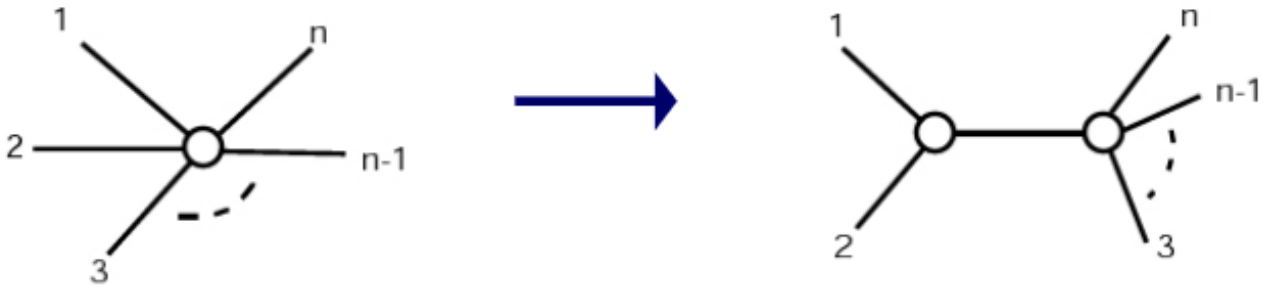


Distance-based trees

- **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)



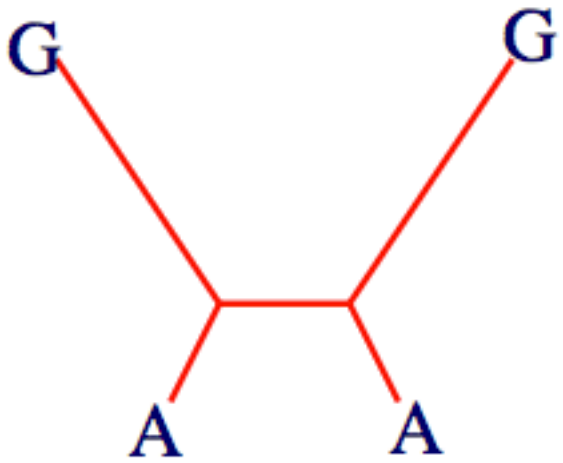
Start off with star tree; pull out pairs at a time

Character based trees

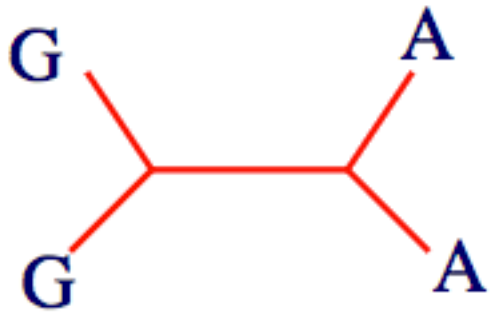
- Maximum parsimony (MP): choose tree that minimizes number of changes from a common ancestor
- MP yields more than one tree with the same score
- Maximum likelihood (ML): find the tree which gives the highest likelihood of the observed data
- They both imply model of evolution

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

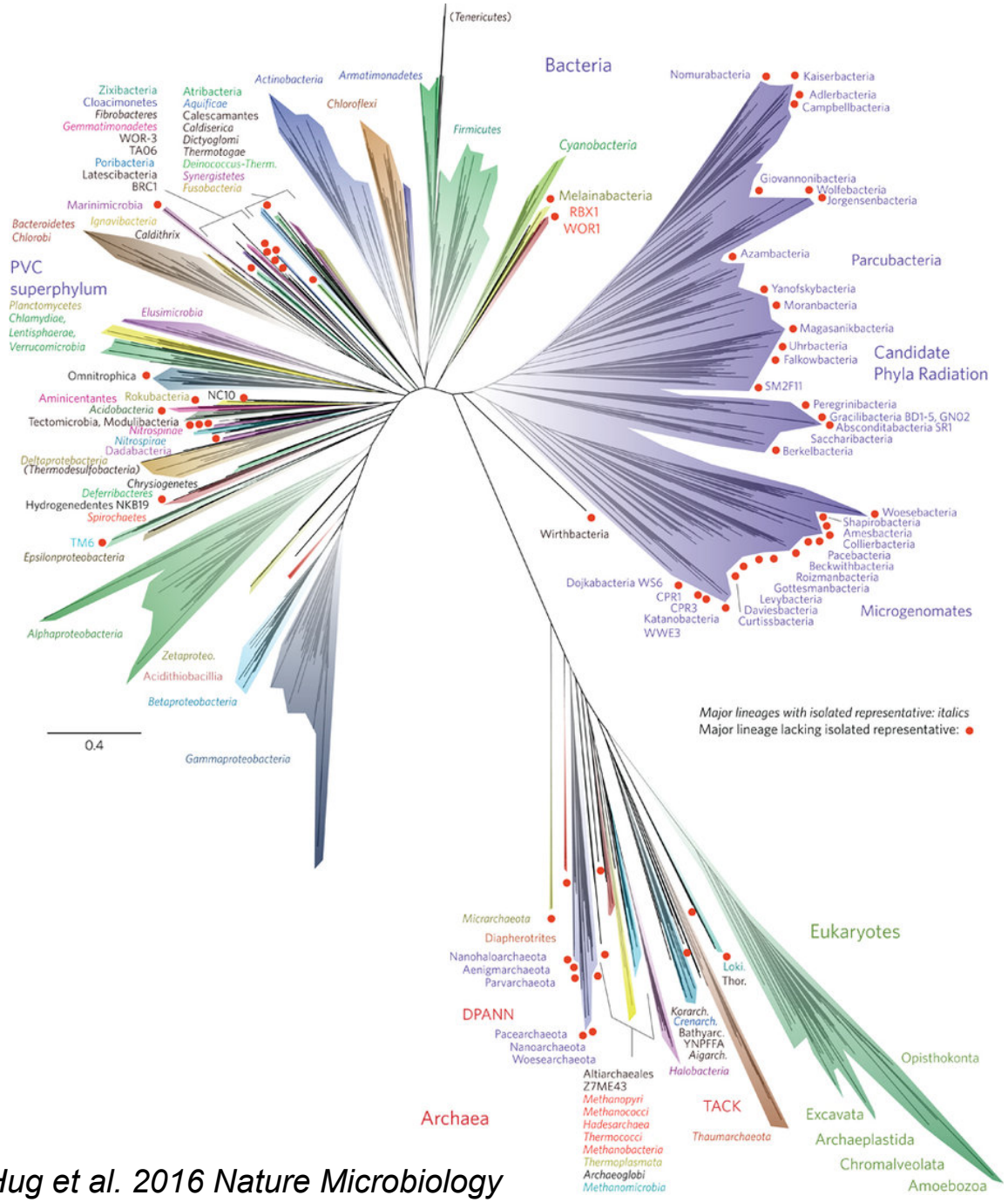


Inferred tree

Max Likelihood Makes 2 independence assumptions

- Different sites evolve independently
- Diverged sequences (or species) evolve independently after diverging

ML tree of life



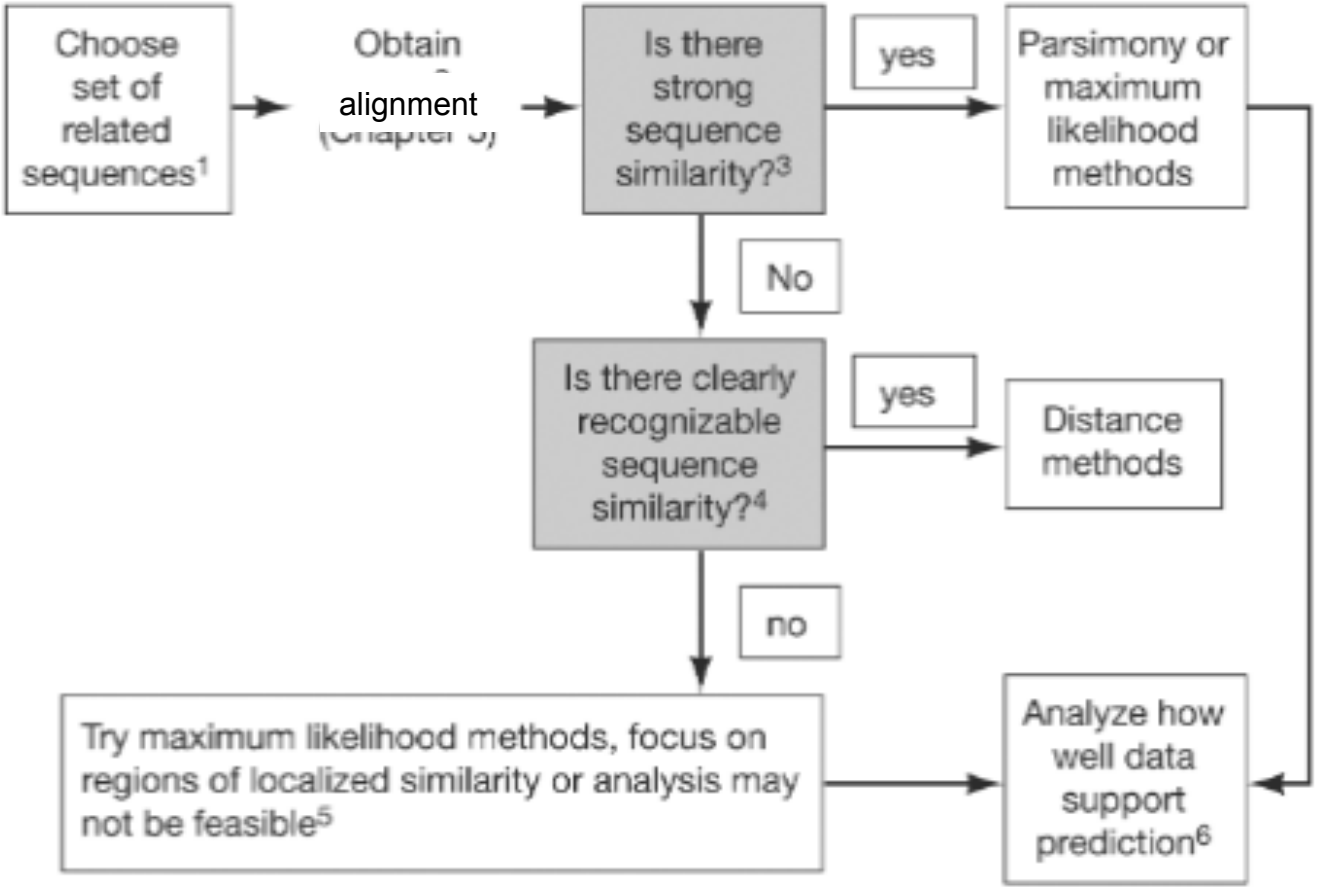
Hug et al. 2016 Nature Microbiology

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:

- Most people now believe **Max Likelihood** based methods are best:
- most sensitive at large evolutionary distances –
- but also most time-consuming & depend on specific model of evolution used

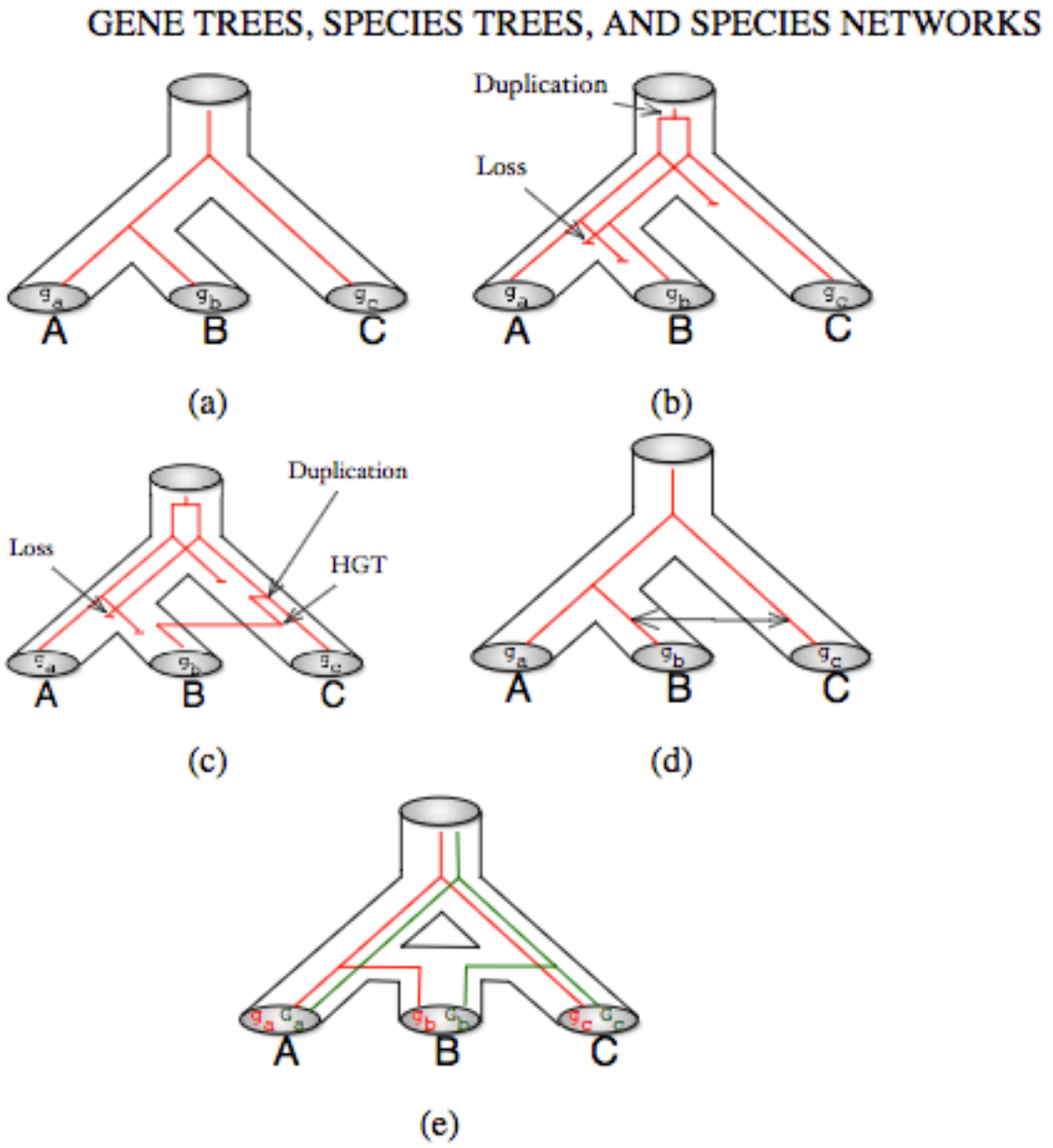


Mount, 2008 Cold Spring Harbor Protocols

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

- Illustrate evolutionary relationships when the evolutionary history of sequences or species may be poorly represented by a tree.
  - Reticulate events caused by recombination, hybridization, or lateral transfer events.
- 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. This approach is based on the idea that
- visually displaying all of the character differences between taxa will show all incompatibility
- Too much conflict can create undisplayable hypercubes

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

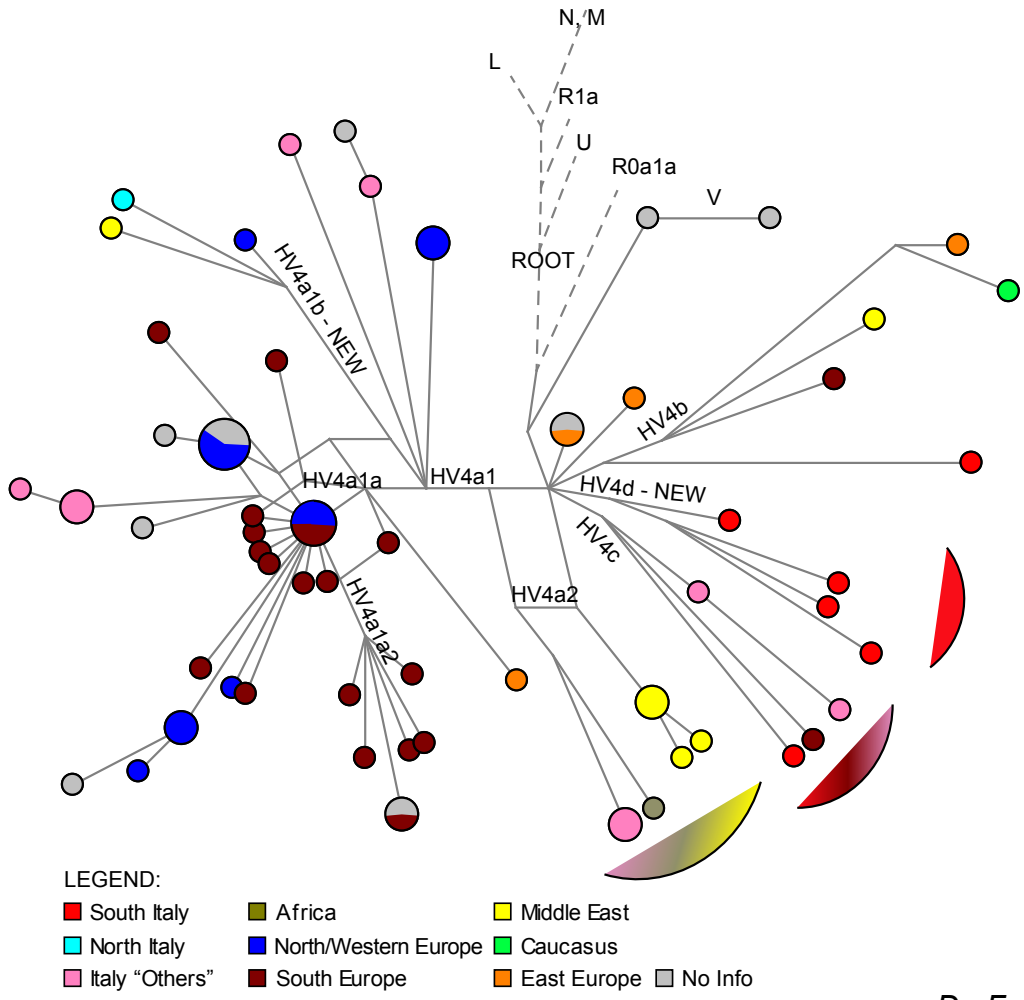
Median Joining

- Suited for very closely related sequences that have evolved without recombination, widely used in phylogeography and population studies

Simplify a network

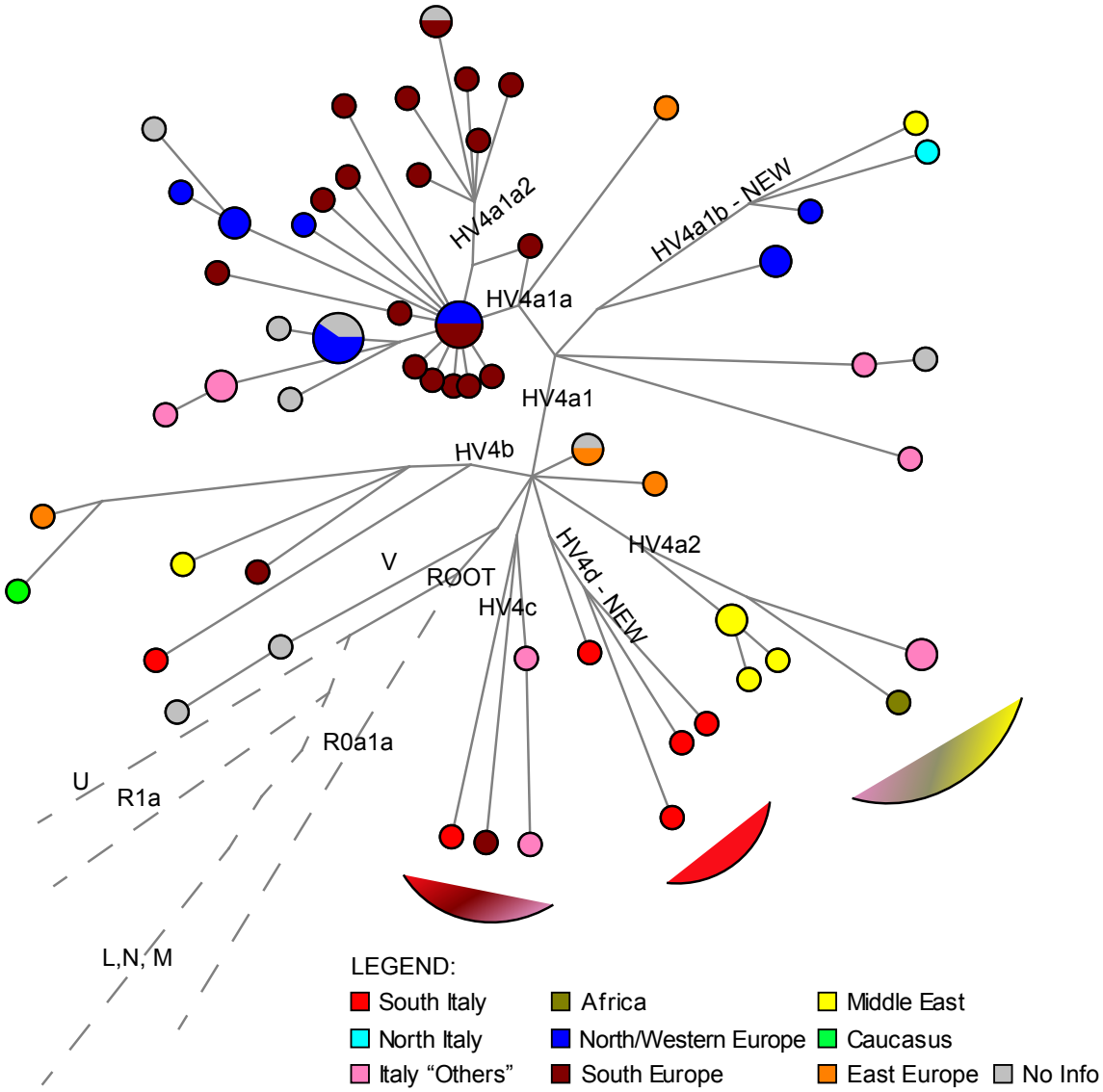
- For less reticulations, apply **weights** to the characters.
  - Positions with recurrent change of state are downweighted
- The software Network allows a few tricks: Reduced Median networks, post processing, star contraction

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).
  - Display of character conflict
  - increasing character-state complexity by producing uninformative multifurcations (i.e. false negatives)
- 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: SplitsTree4

Huson & Heyer 2006

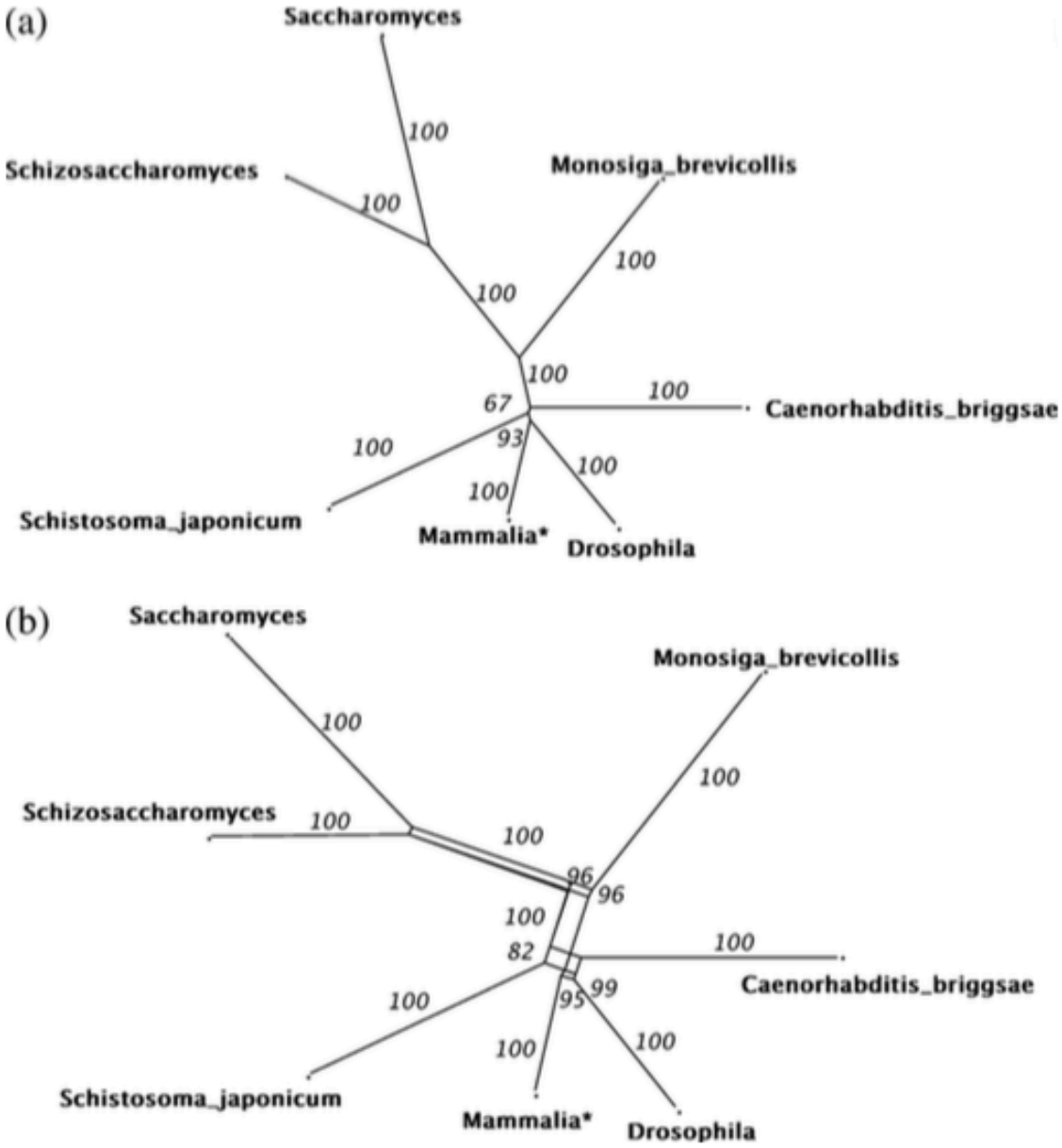


FIG. 9.—(*a*) The Bio-NJ tree (Gascuel 1997), with bootstrap values, for a smaller set of seven animals (following Philippe, Lartillot, and Brinkman 2005) using a concatenated alignment of 146 genes, and ML distances under a JTT + F + Γ model. The tree-based method gives reasonable, but not conclusive, support for the coelomate hypothesis, though the small bootstrap value, even with this large number of sites, already suggests that the clade is unreliable. (*b*) The neighbor-net network using the same distance estimates, with bootstrap values. Even without the cnidarian taxa, there is substantial (but not conclusive) support in the data for the ecdysozoa hypothesis.

Ge

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

Testing hypothesis: chicken and egg (reprise)

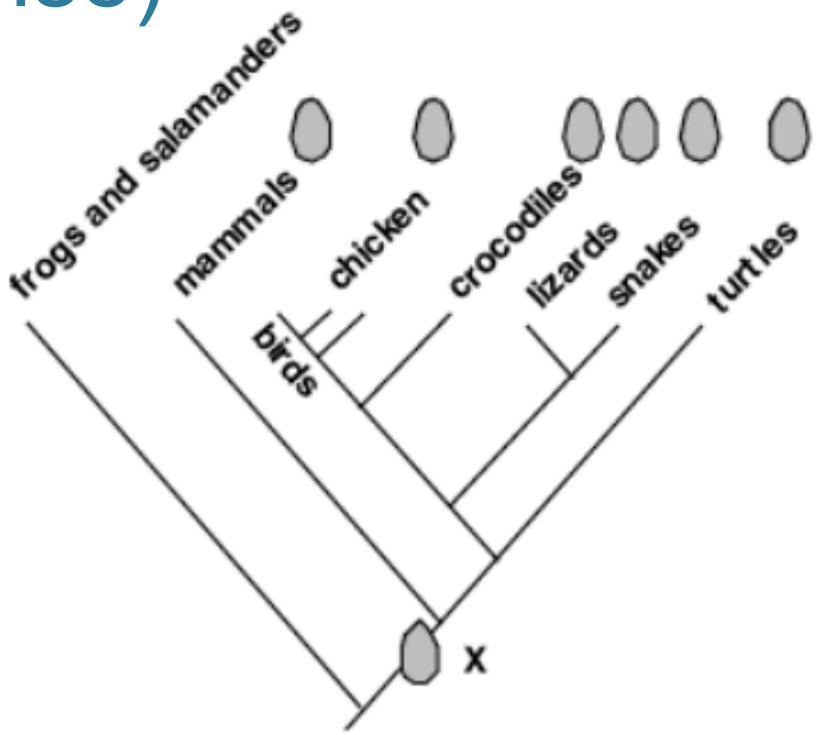


Fig. 7. Ordering the evolution of characters. *Setting the chicken in its phylogenetic context (here, a simplified tree of the tetrapods) quickly reveals that the amniote egg, of which the chicken egg is an example, is widely distributed (although it has been lost in higher mammals). The distribution implies that the amniote egg evolved at point X, well before the much later origin of the chickens. Fossil evidence suggests that the amniote egg appeared at least 310 million years ago and the first members of the Phasianidae (the family containing chickens) some 50 million years ago.*

Telford & Budd 2003

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**

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
- Check the haplogroups

2. Import a matrix of genetic distance between populations

- Visualize relationships with an MDS, a NJ tree, a UPGMA tree

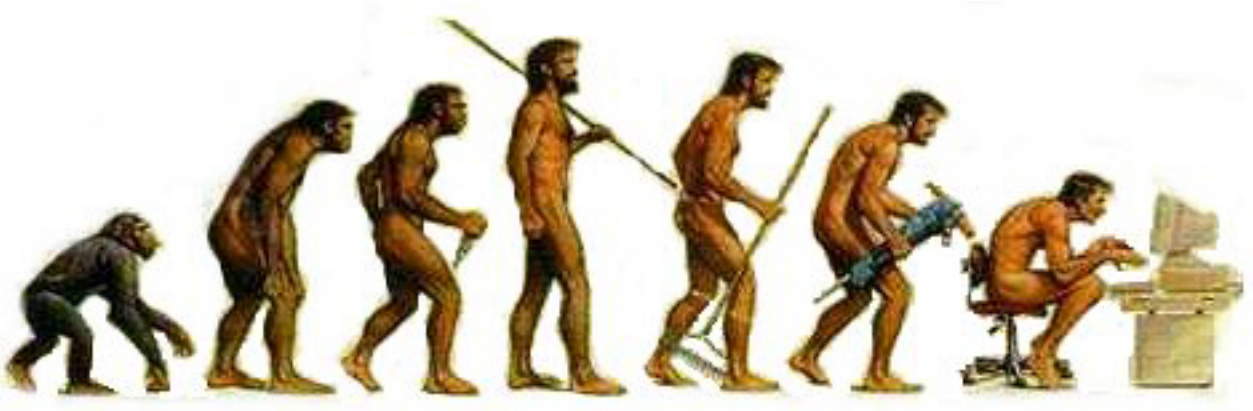
Bonus slides!!

The concept of “population” and the study design

Why study population genetics

A multidisciplinary approach

- Anthropology as the study of our origin, history and current diversity
- Biological and cultural nature of humankind coevolve under the same demographic processes
- Hypotheses from genetics, archaeology, linguistics, cultural anthropology complement and validate each other



Genetic contributions to population prehistory

EXAMPLES

- A mismatch between the genetic relationship and the linguistic affiliation of a group can indicate **language shift**
- With mtDNA and Y-chromosome ("uniparental markers"), we can detect sex differences in prehistoric events (**sex-biased gene flow**)
- Absence of **admixture** is indicative of some cultural or physical barrier between groups

Combining genetics and linguistics

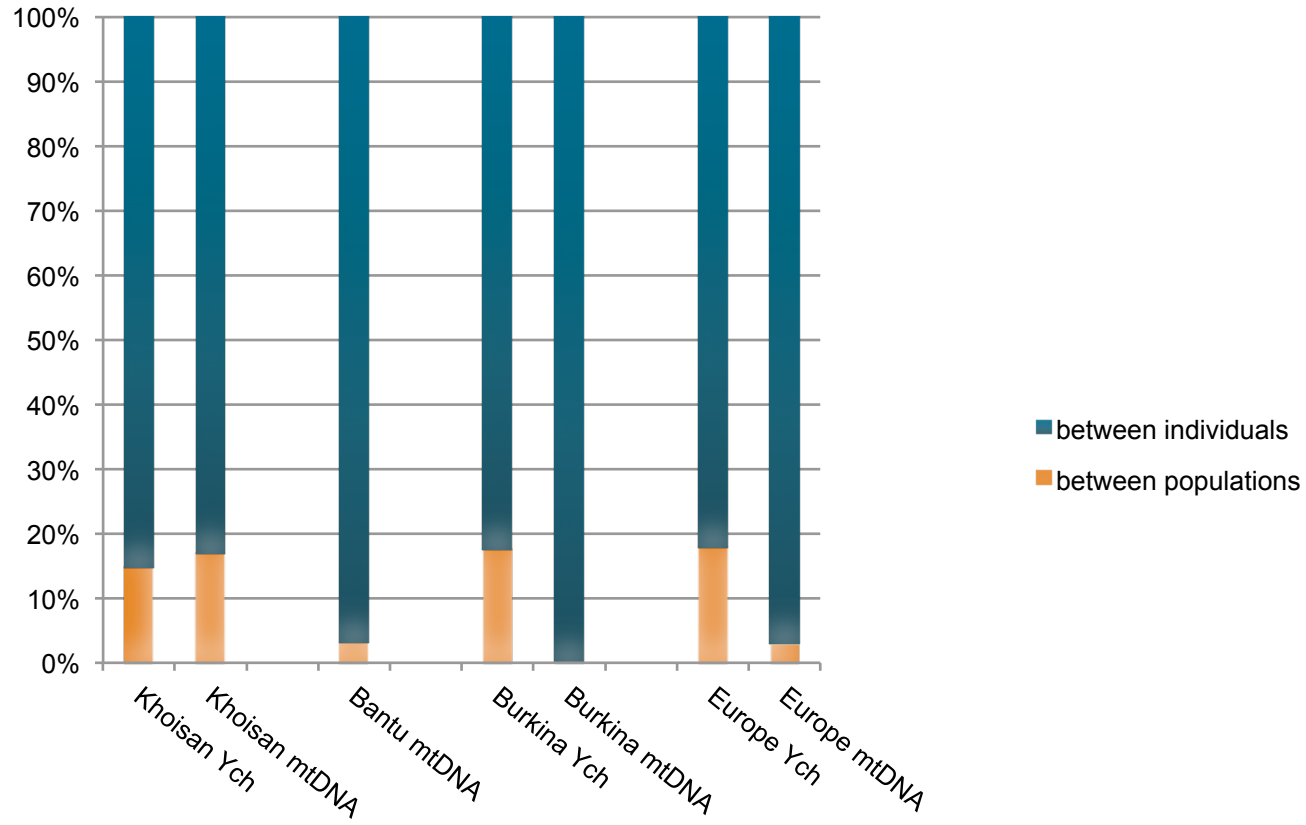
- **Molecular Anthropology** can detect cases of language shift, sex-biased gene flow, lack of admixture
- **Linguistics** can determine different types of language change (phonology, morphology, syntax, lexicon)

  → Understand population diversification and contact

Genes and culture

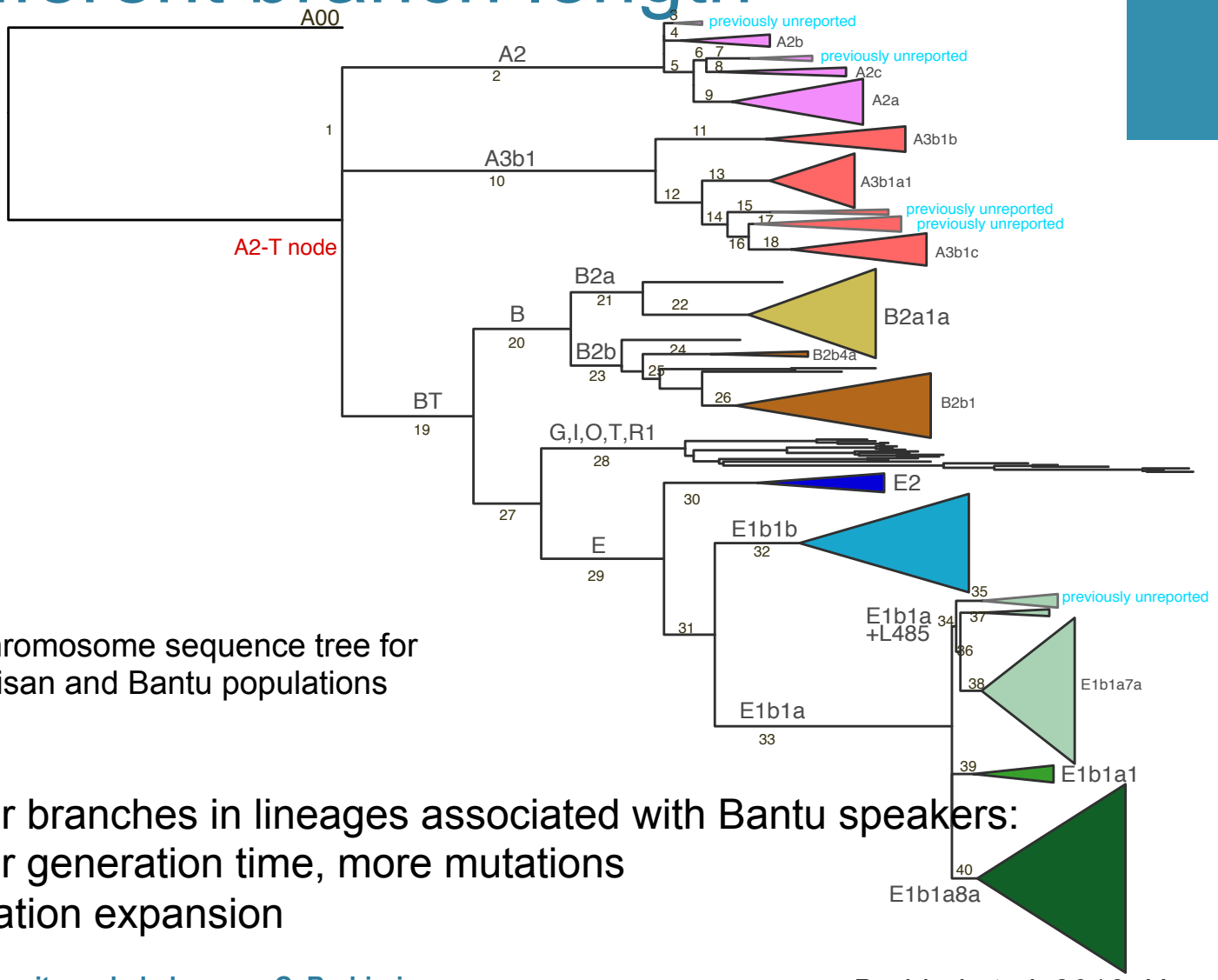
- Matrilocality vs. patrilocality
  - mtDNA vs Y chromosome structure
- Generation time
  - Affecting mutation rate

Sex biased pop variance



High mtDNA structure between Khoisan populations, which are multilocal or matrilocal, the other populations are patrilocal!

Different branch length



Y chromosome sequence tree for Khoisan and Bantu populations

Longer branches in lineages associated with Bantu speakers:
Longer generation time, more mutations
Population expansion

Barbieri et al. 2016, Human Genetics

Problems in comparison to other disciplines

- Transmission modalities
- Evolutionary time windows

Problems with genetics

- Y chromosome and mtDNA represent only a limited part of a population ancestry
  - Haplogroup data not always informative
- Autosomal SNP chip built on European genetic diversity suffer from Ascertainment Bias
  - Two European populations are more diverse than 2 African populations = impossible!
- …What is the population sample representative of? How was it sampled?

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

Continuous variability

On the Non-Existence of Human Races

by FRANK B. LIVINGSTONE☆

[*Ann Arbor, Mich., U.S.A.*, 12.10.61.]

In this paper I would like to point out that there are excellent arguments for abandoning the concept of race with reference to the living populations of *Homo sapiens*. Although this may seem

found among
wide-ranging
ally found
allopatric sp
Thus, alth
a group of r
units, namel
sible to div

Current Anthropology, 1962

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**

Study design

- Collect archeological, cultural, demographic, linguistic information over the populations of interest (diachronic and synchronic perspective)
- Focus on key research questions
  - Lack of written record?
  - Unique features?
  - Particular cases of contact and/or isolation?
  - Mysterious origin?

Study design: fieldwork

- Collect cultural data for quantitative comparison (if possible)
- Collect biologic material for DNA quantitative comparison
- Find methods that are suitable for both linguistic/cultural and genetic analysis and are directly comparable
- **Critical point: get appropriate ethical permits**!

Biological sampling

- Explain the aim of the project to the volunteer donor
- Saliva samples in Oragene kit (simple, non-invasive)

- Questionnaire to assess ethnolinguistic affiliation and place of birth up to two generations (grandparents rule)
- The biological material is kept anonymous with a numerical code

- The DNA is never considered at the individual level, but at the population level (genetic pool)

Genetic sampling in northern Botswana, 2009

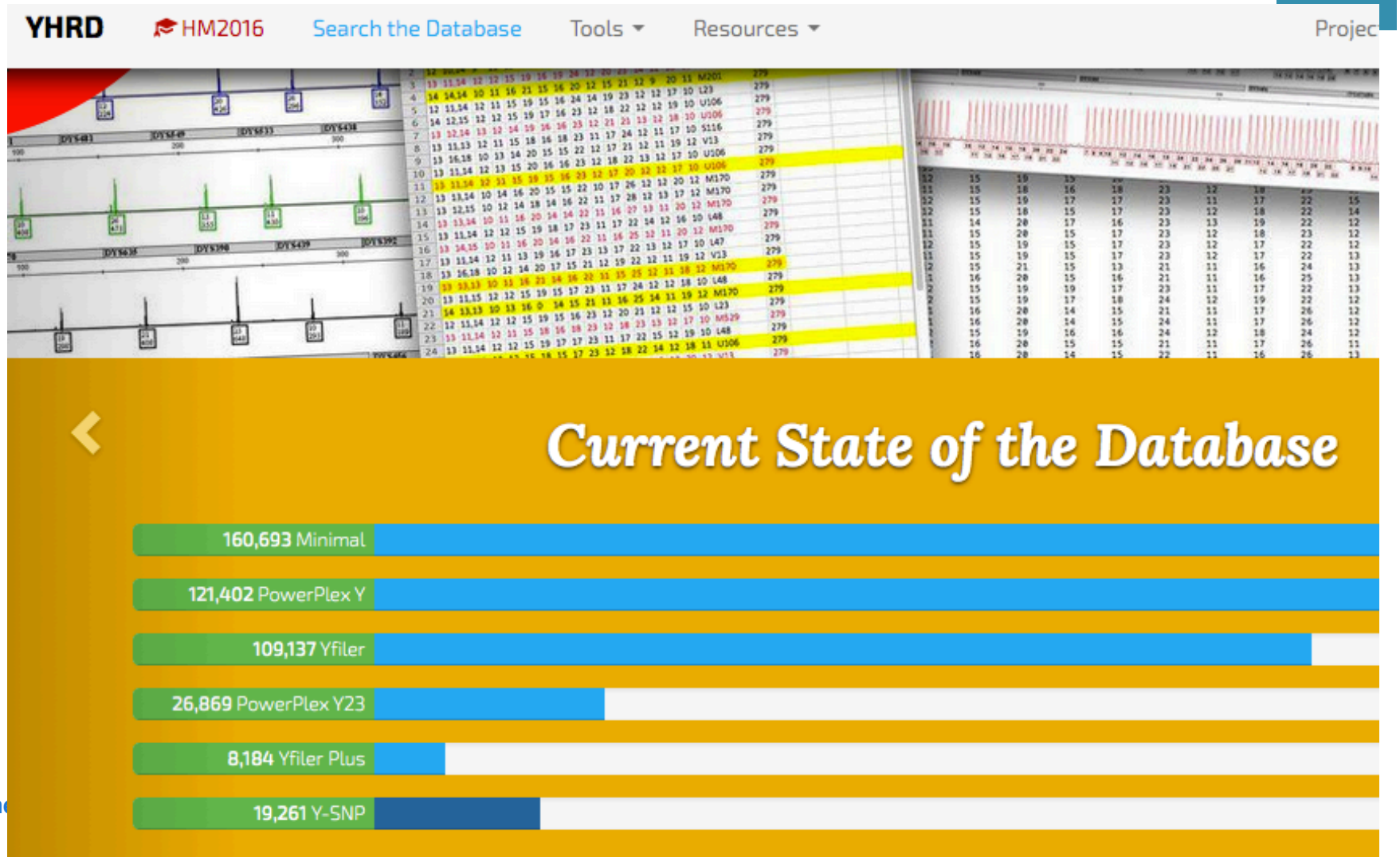
Available datasets for genetic diversity

- Forensic (autosomal, Y chromosome, mtDNA)
  - Standardized protocols and haplotypes
- World cell line banks
  - CEPH-HGDP (Human genome diversity panel)
  - 1000 genomes

Obviously all data available are anonymous.

forensic

Example Y chromosome: https://yhrd.org/



Genbank

- Catalogue of DNA sequences
- Submit queries, download data
- The problem is to merge: make alignments

CEPH-HGDP

Africans
1 Bantu
2 Mandenka
3 Yoruba
4 San
5 Mbuti pygmy
6 Biaka
7 Mozabite

Europeans
8 Orcadian
9 Adygei
10 Russian
11 Basque
12 French
13 North Italian
14 Sardinian
15 Tuscan

Western Asians
16 Bedouin
17 Druze
18 Palestinian

Central and Southern Asians
19 Balochi
20 Brahui
21 Makrani
22 Sindhi
23 Pathan
24 Burusho
25 Hazara
26 Uygur
27 Kalash

Eastern Asians
28 Han (S. China)
29 Han (N. China)
30 Dai
31 Daur
32 Hezhen
33 Lahu
34 Miao
35 Oroqen
36 She
37 Tujia
38 Tu
39 Xibo
40 Yi
41 Mongola
42 Naxi
43 Cambodian
44 Japanese
45 Yakut

Oceanians
46 Melanesian
47 Papuan

Native Americans
48 Karitiana
49 Surui
50 Colombian
51 Maya
52 Pima

Nature Reviews | Genetics

IGSR and the 1000 Genomes Project

Populations: - African; - American; - East Asian; - European; - South Asian;

The International Genome Sample Resource (IGSR) was established to ensure the ongoing usability of data generated by the 1000 Genomes Project and to extend the data set. More information is available about the IGSR.