CABANA
Capacity building for bioinformatics in Latin America

USMP
UNIVERSIDAD DE SAN MARTÍN DE PORRES
FACULTAD DE MEDICINA HUMANA

ANÁLISIS Y FILOGENIA DE POBLACIONES HUMANAS

Chiara Barbieri, University of Zurich, Suiza
Max Planck Institute for the Science of Human History, Jena, Alemania

MAX-PLANCK-GESELLSCHAFT

University of Zurich[UZH]

Wenner-Gren Foundation

barbieri.chiara@gmail.com

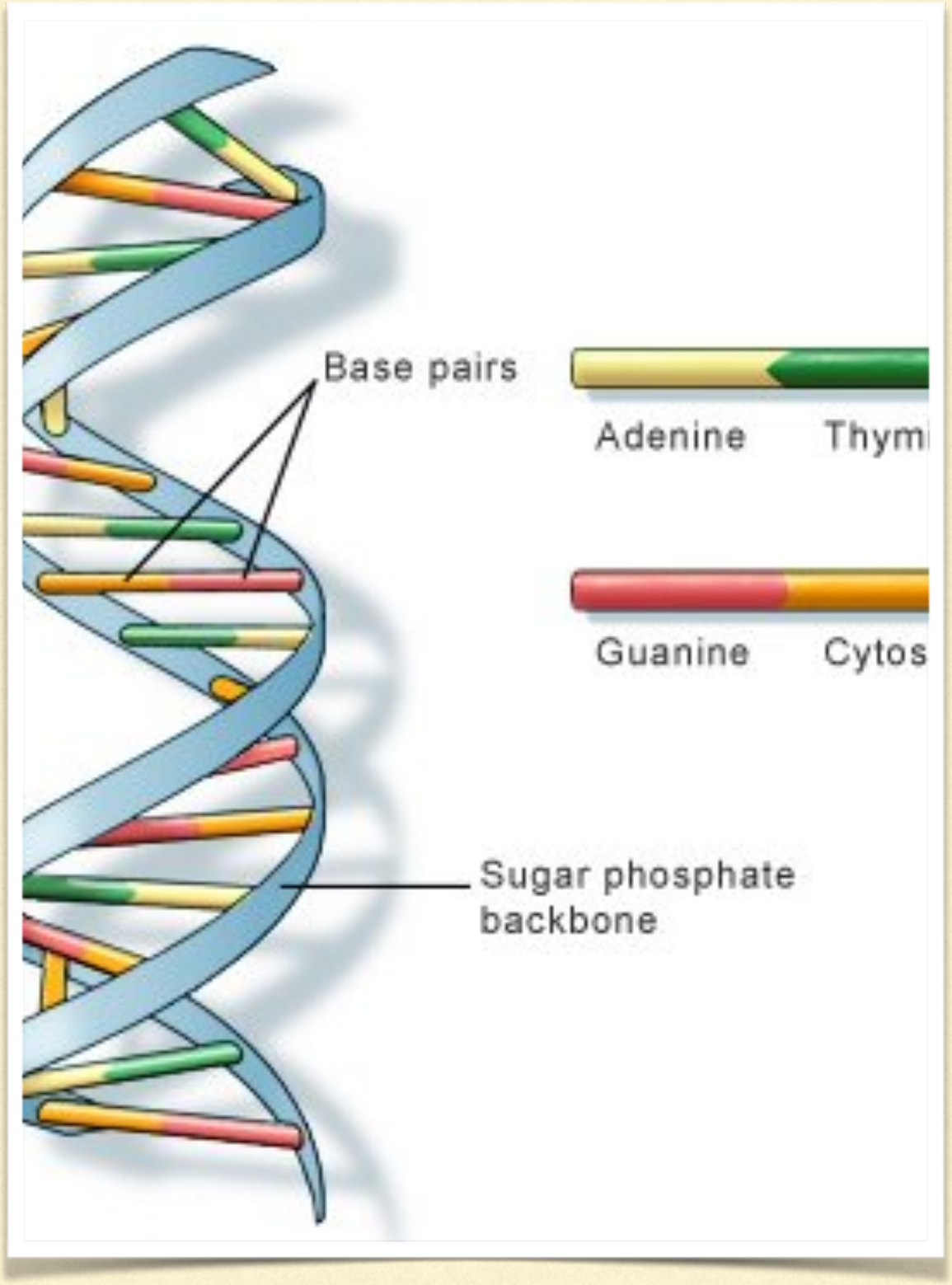
OUTLINE

- Marcadores genéticos en estudios de diversidad humana
- Population structure, diversidad y admixture
- Filogenia
- Case study: historia de poblaciones en Africa

EJERCICIO:

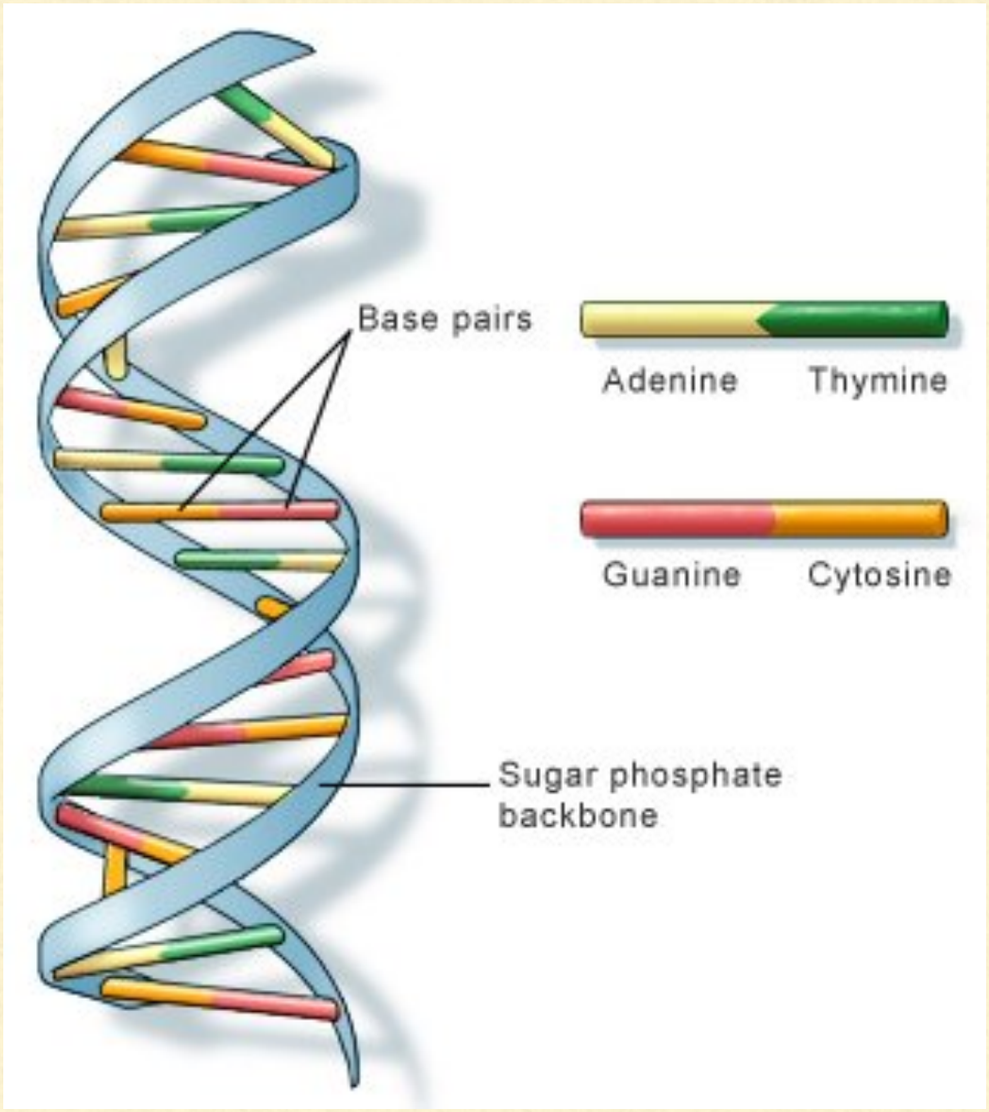
- https://github.com/chiarabarbieri/CABANA_Lima_2019

MARCADORES GENÉTICOS EN ESTUDIOS DE DIVERSIDAD HUMANA



Chiara Barbieri

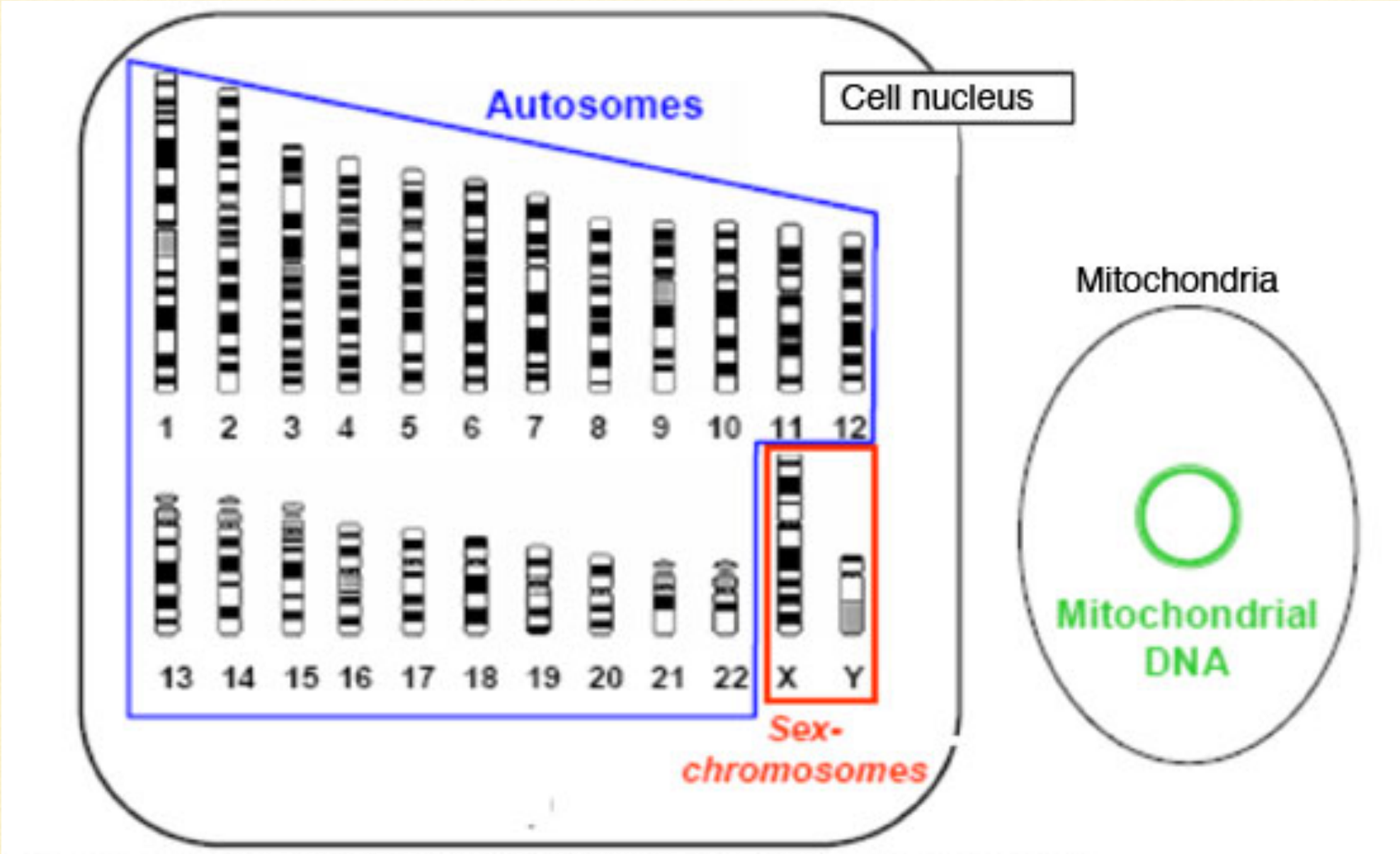
GENETICS AND POPULATION HISTORY



- **<u>Classical markers</u>**
  - little information

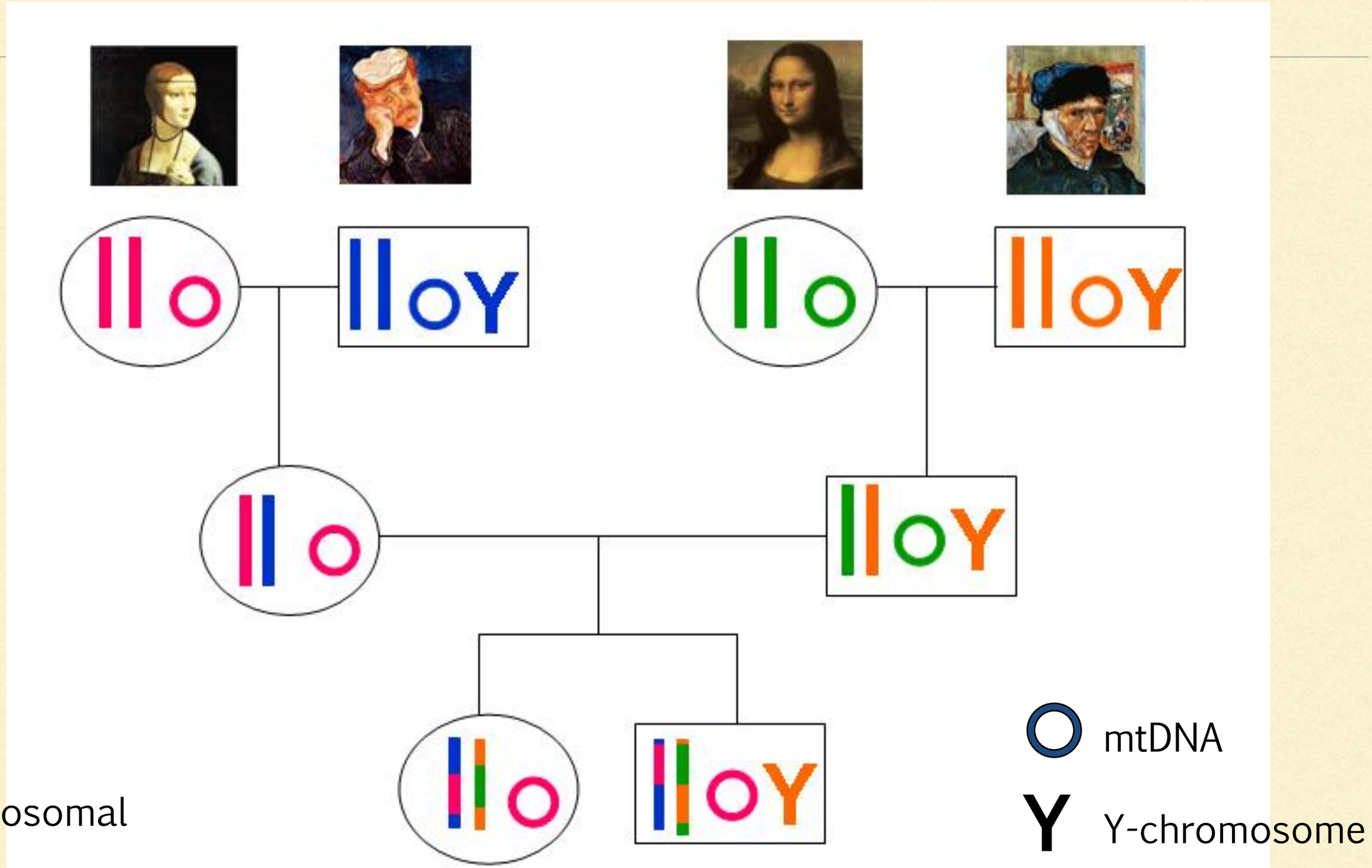
- **<u>Molecular markers</u>**
- **Autosomal DNA**
  - ancestors history
  - SNPs
  - Genome sequencing
- **Uniparental DNA (mtDNA, Y chromosome)**
  - mother and father's lineages
  - SNPs
  - Sequences

THE HUMAN GENOME
23 pairs of Chromosomes + mtDNA



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

GENETIC MARKERS: TRANSMISSION



MARCADORES GENÉTICOS AUTOSOMICOS

Autosomal markers:

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

LIMITS

- Recombination prevents the tracing back of individual mutations through space and time
- Can be more cost-intensive

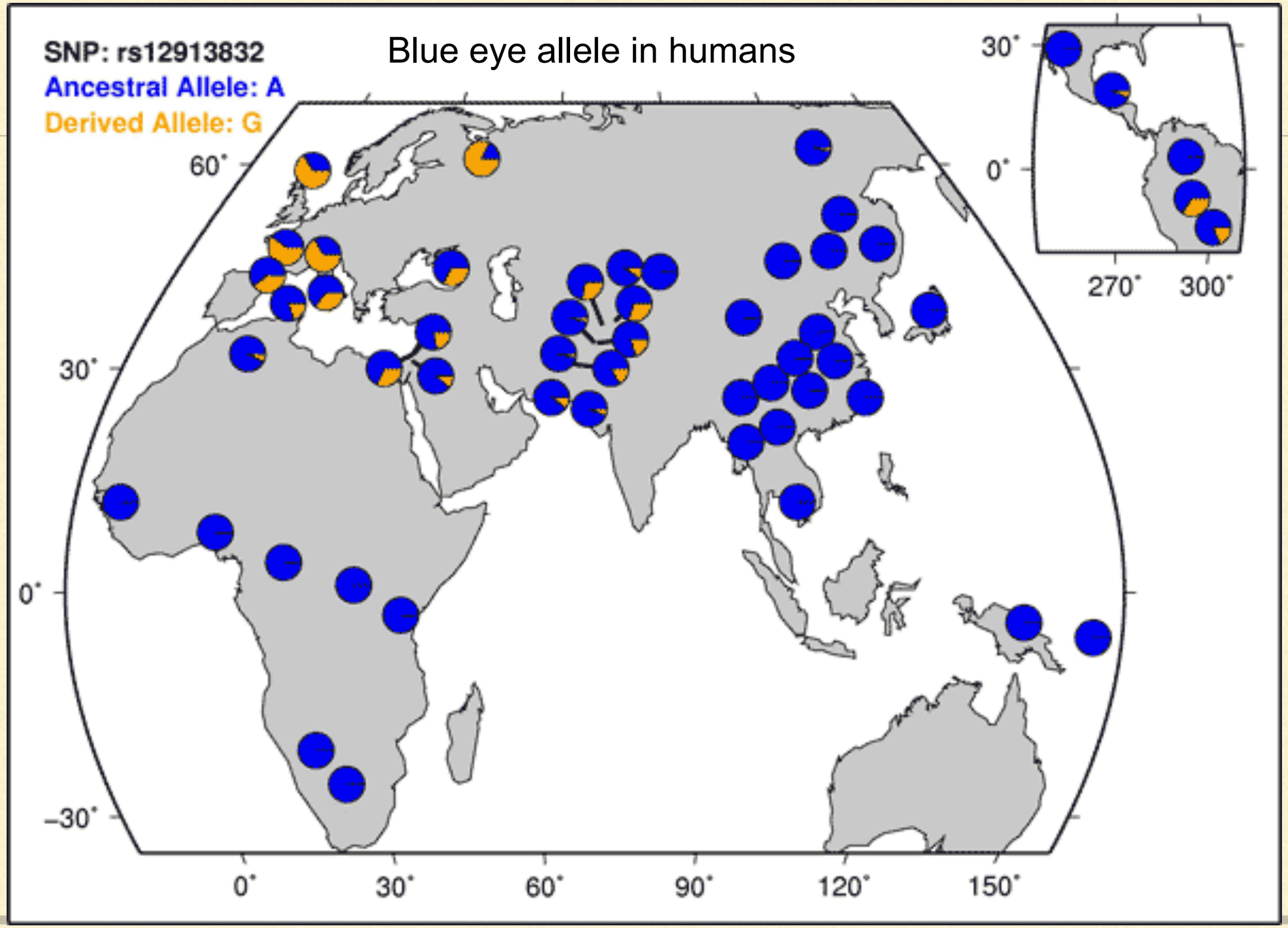
AUTOSOMAL DATA

- Focus on one gene or variant
  - Population variation
  - Reconstruct the phylogeny
- Capture a set of independent positions through the genome: SNP chip
  - Ascertained to be variable in human population
-

AUTOSOMAL DATA

- Focus on one gene or variant
  - Population variation
  - Reconstruct the phylogeny
- Capture a set of independent positions through the genome: SNP chip
  - Ascertained to be variable in human population
-

PRESENT DAY VARIATION



(HGDP Selection Browser)

PHYLOGENY OF THE LACTASE ALLELE IN SOUTHERN AFRICA

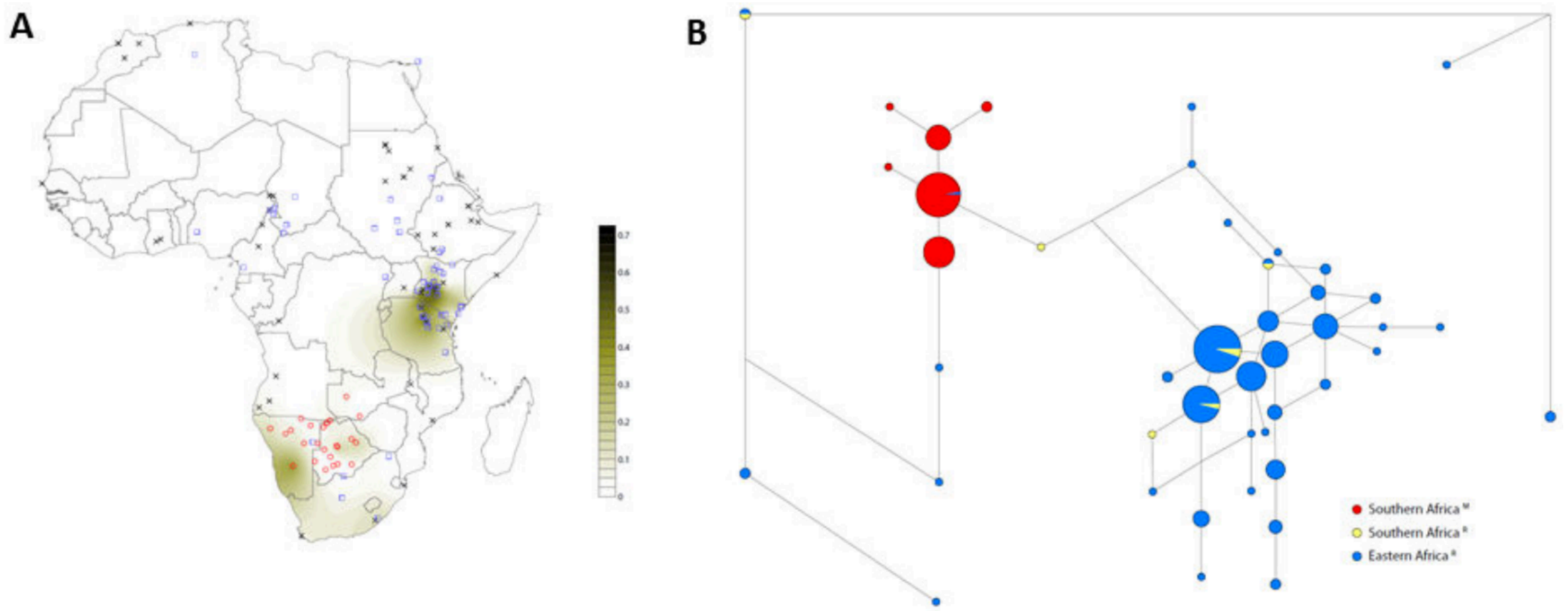


Fig. 1. Analyses of the C-14010 LP variant and associated STR haplotypes in Eastern and Southern African populations. **A**: Surfer map of the C-14010 allele frequency. Red circles denote sampling locations by Macholdt et al.; blue squares denote sampling locations by Ranciaro et al.; black crosses denote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

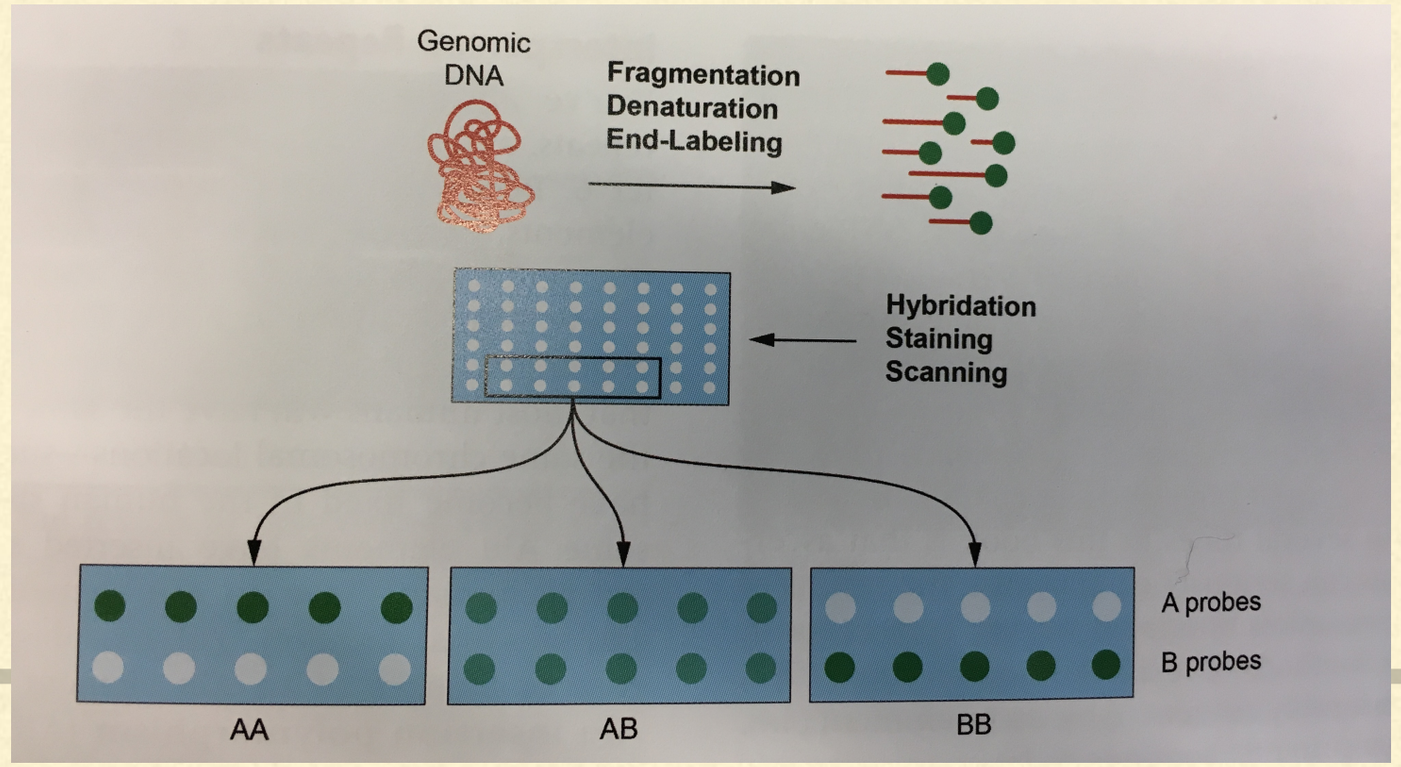
Macholdt et al. 2015 Am J Phys Anthropol

AUTOSOMAL DATA

- Focus on one gene or variant
  - Population variation
  - Reconstruct the phylogeny
- Capture a set of independent positions through the genome: SNP chip
  - Ascertained to be variable in human population
-

ANALIZAR SNPS

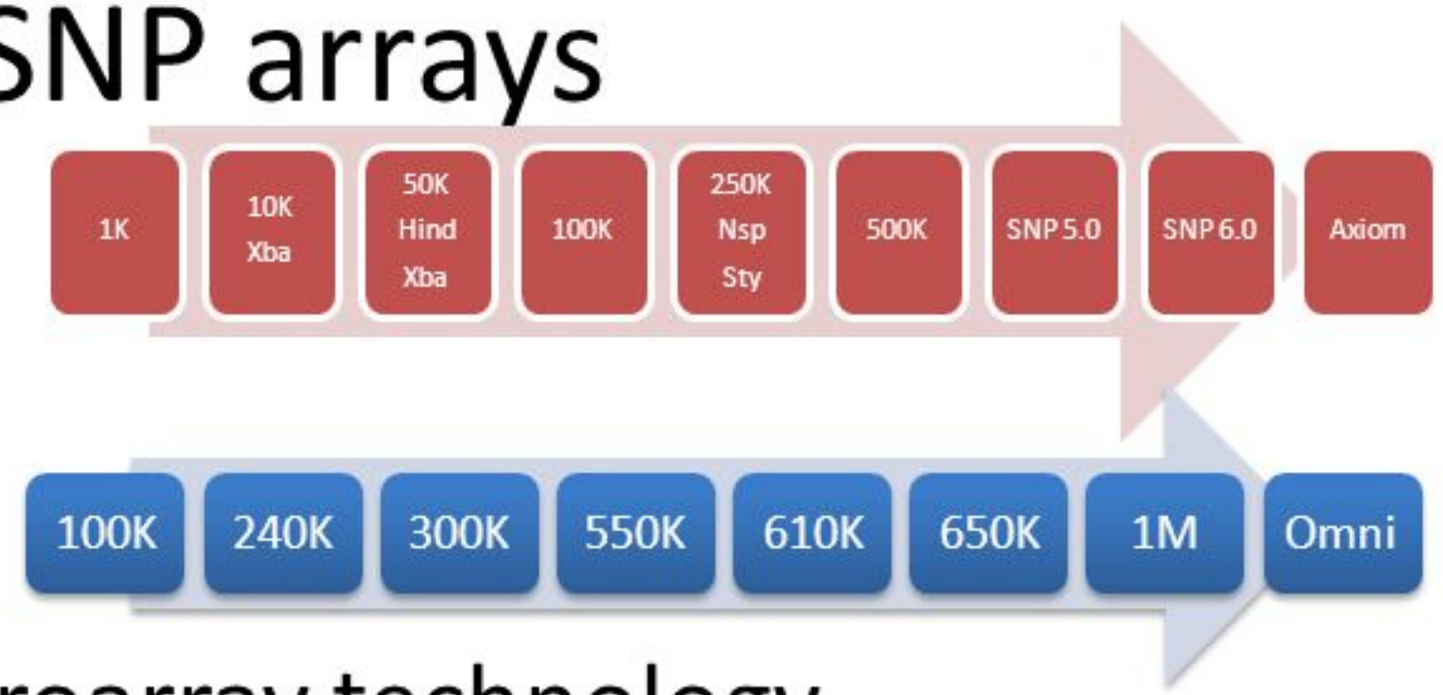
- SNP chip (early 2000) : microarray that contain synthetic DNA probes to the allele for a selection of SNPs
- Fragmented labelled DNA hybridizes to the DNA probes, the rest is washed away
- Only the probes bound to DNA are read and their position in the array is recognised



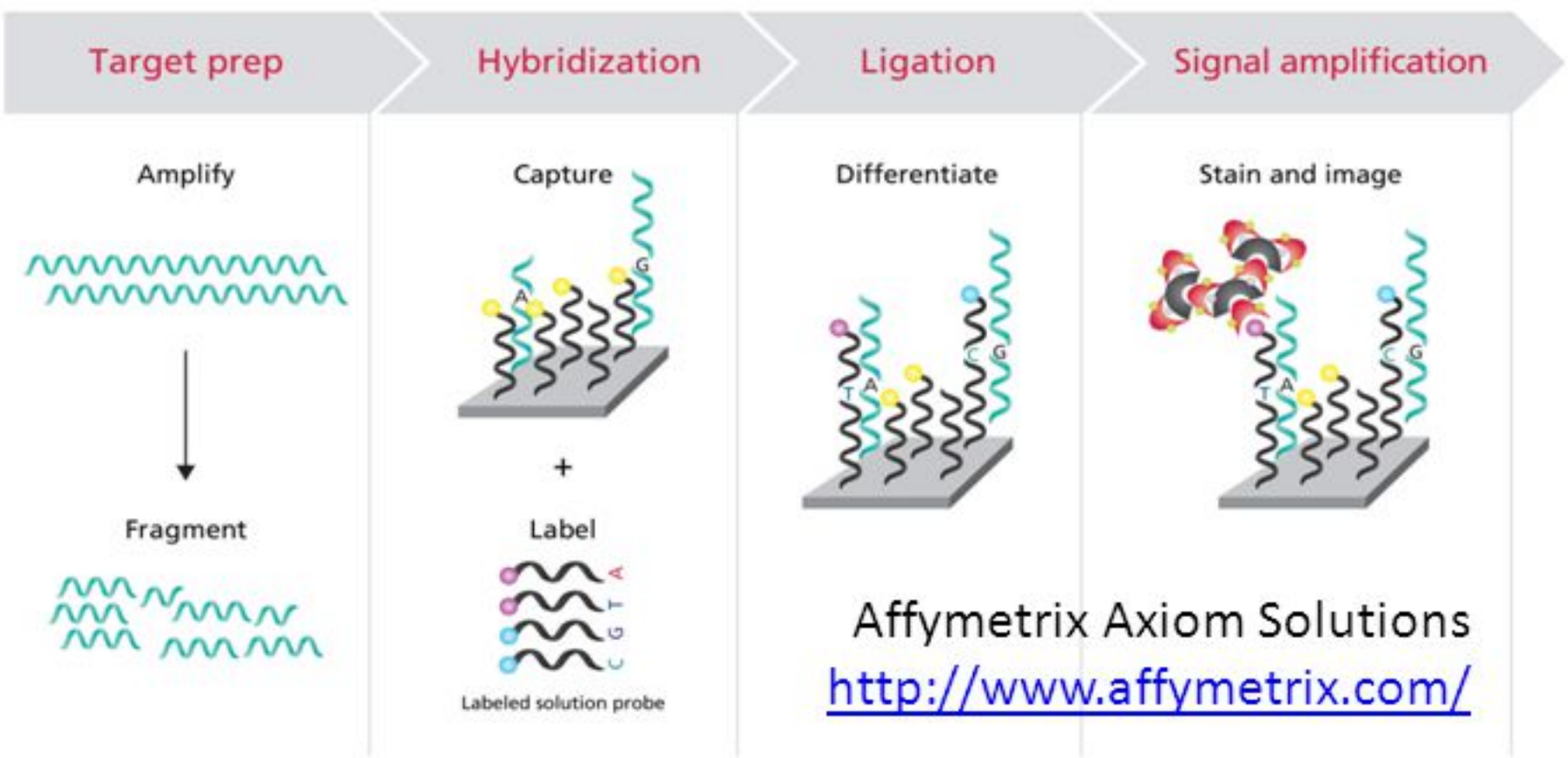
Stoneking 2016

SNP arrays

- Affymetrix
- Illumina



- Probes on microarray technology



Credits: J. Chen

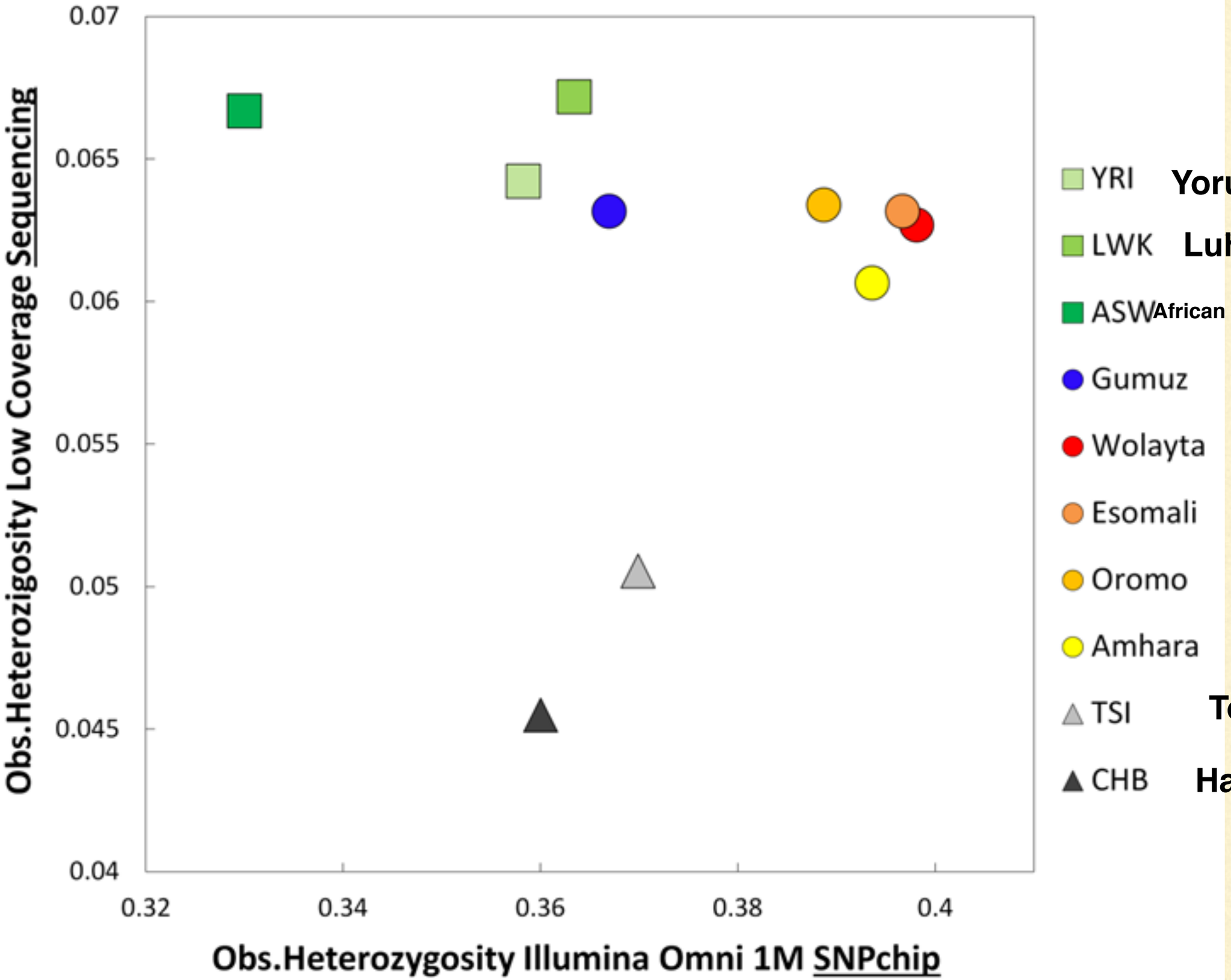
SNP CHIPS AND ASCERTAINMENT BIAS

- The diversity recovered does not correspond to a true measure of diversity in terms of population genetics

- can be caused by sampling a nonrandom set of individuals or by biased SNP discovery protocols.

  - SNPs have been identified by sequencing European individuals - only known SNPs can be genotyped!!

- pre-ascertained SNPs are biased toward older SNPs



Yoruba

Luhya

African ancestry in Southwest USA

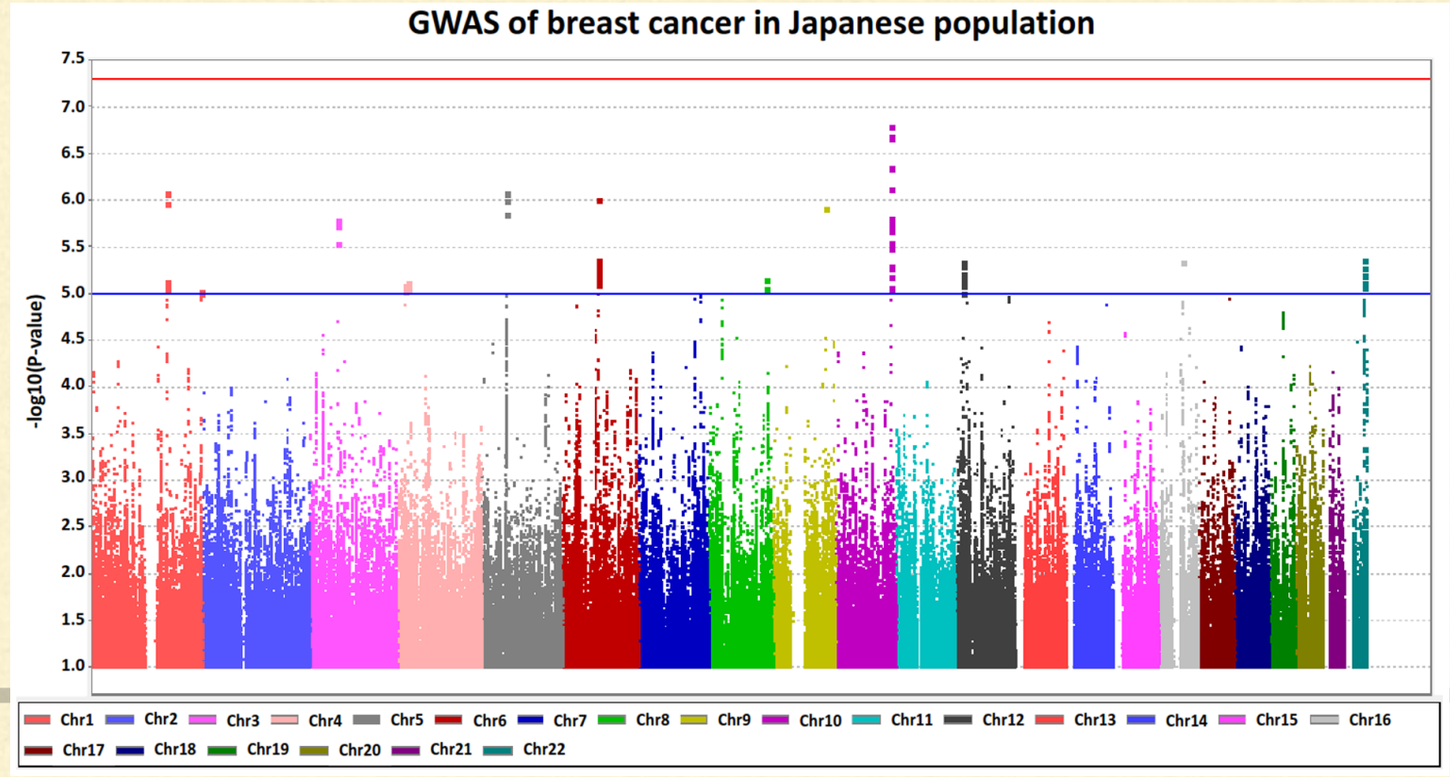
Toscani, Italy

Han Chinese

Figure credits: Luca Pagani

GENOME WIDE ASSOCIATION STUDIES - GWAS

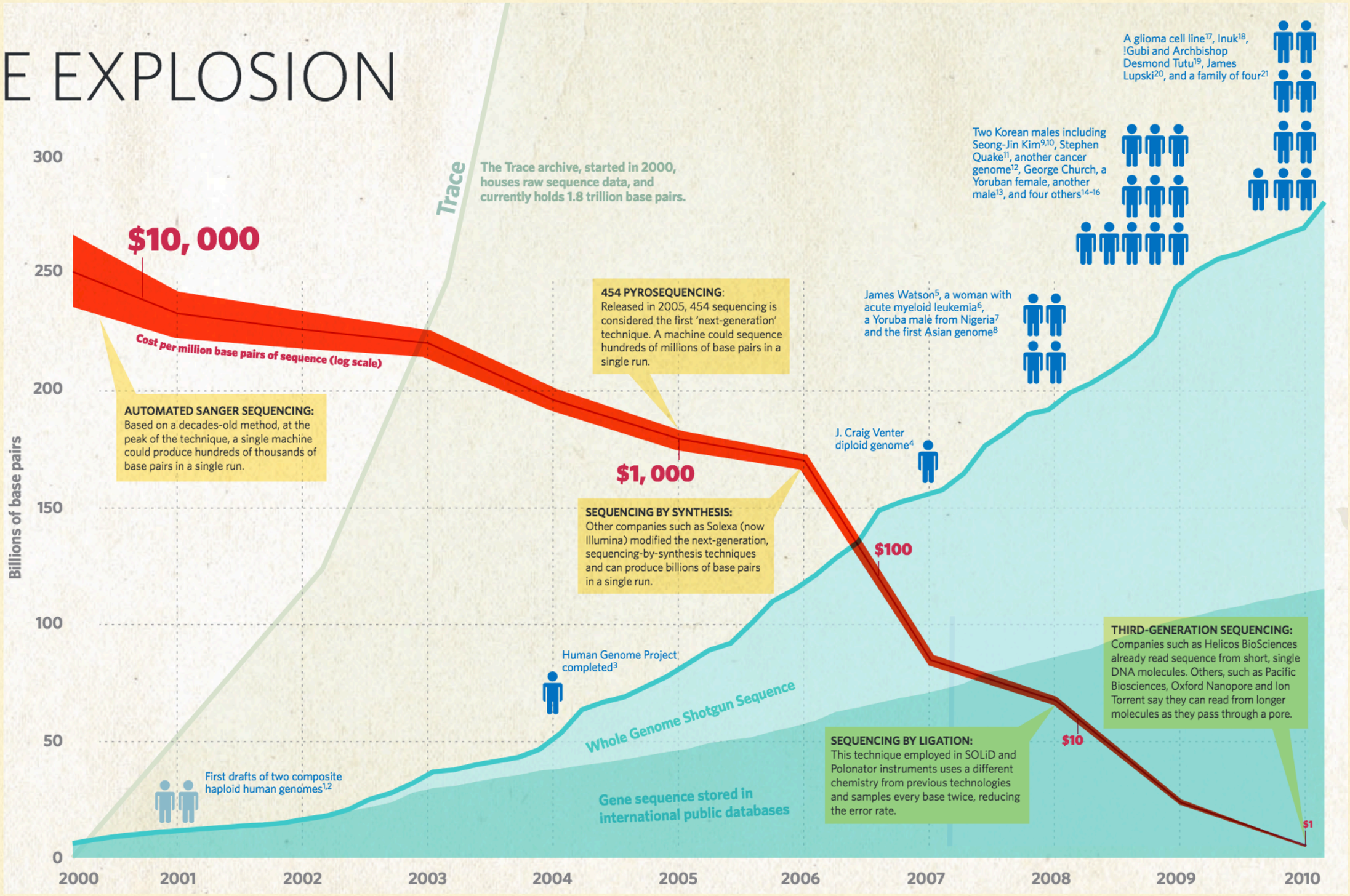
- Based on SNP chip (not really genome wide)
- Hypothesis free testing for association to SNP variants and diseases (or other traits)
- Problems: replicate failures, false positives, missing rare variants, Linkage Disequilibrium, population specific variants



Low et al. Plos ONE 2013

SECUENCIA DE ADN

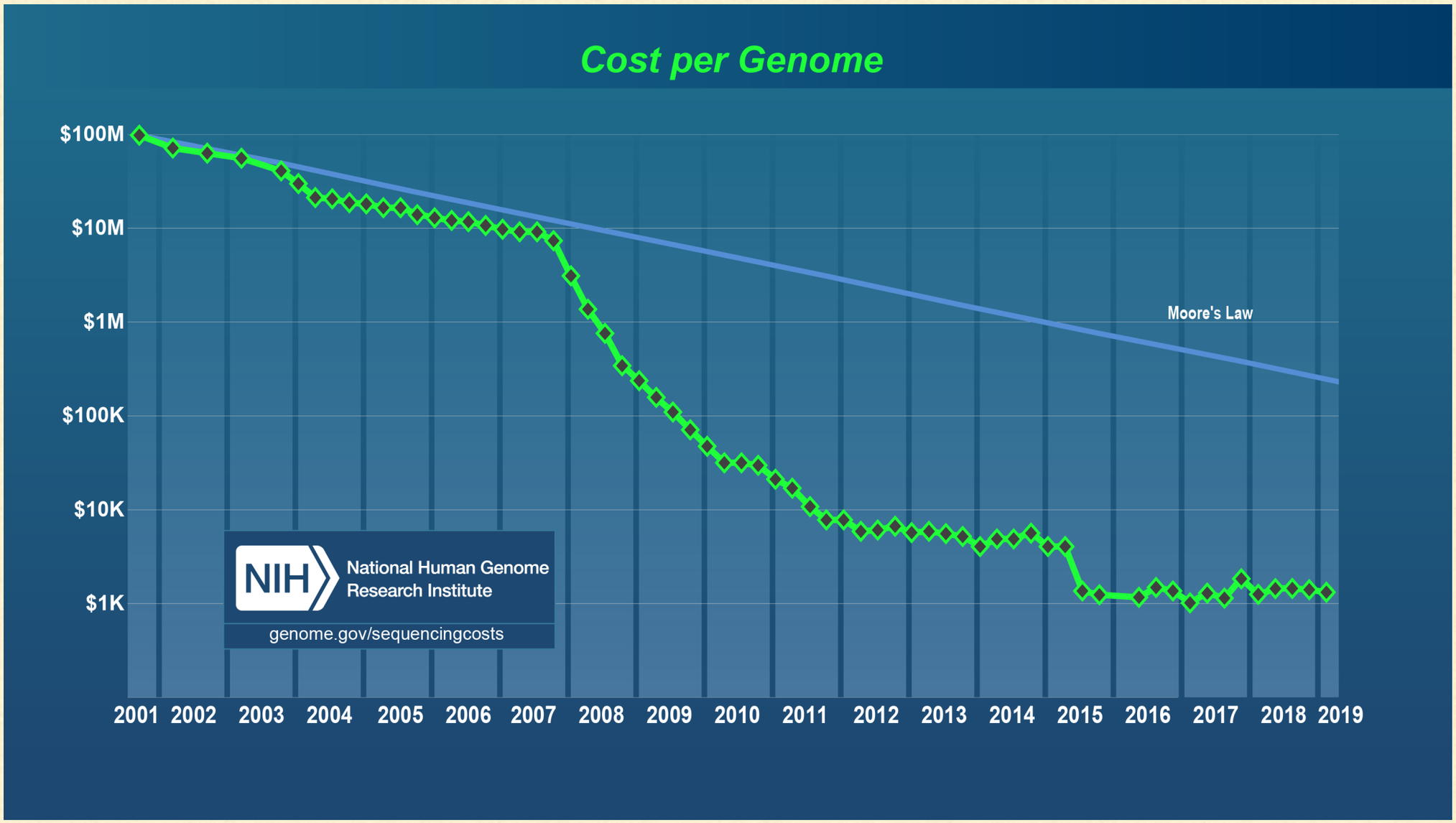
- Next Generation sequencing (first papers out in 2005)
- Amplify DNA on fragments (Illumina: attached to a flow cell)
  - Higher error rate compensated by coverage
  - Shorter read fragments make difficult to cover genomic regions such as high repetitive DNA
  - A complete genome sequence at 30X is only 85% complete



The sequence explosion - Nature, 2010

Chiara Barbieri

CUANTO CUESTA SECUENCIAR UN GENOMA?



"Our ability to sequence human genomes has vastly outpaced our ability to interpret genetic variation"

REVIEW

doi:10.1038/nature24286

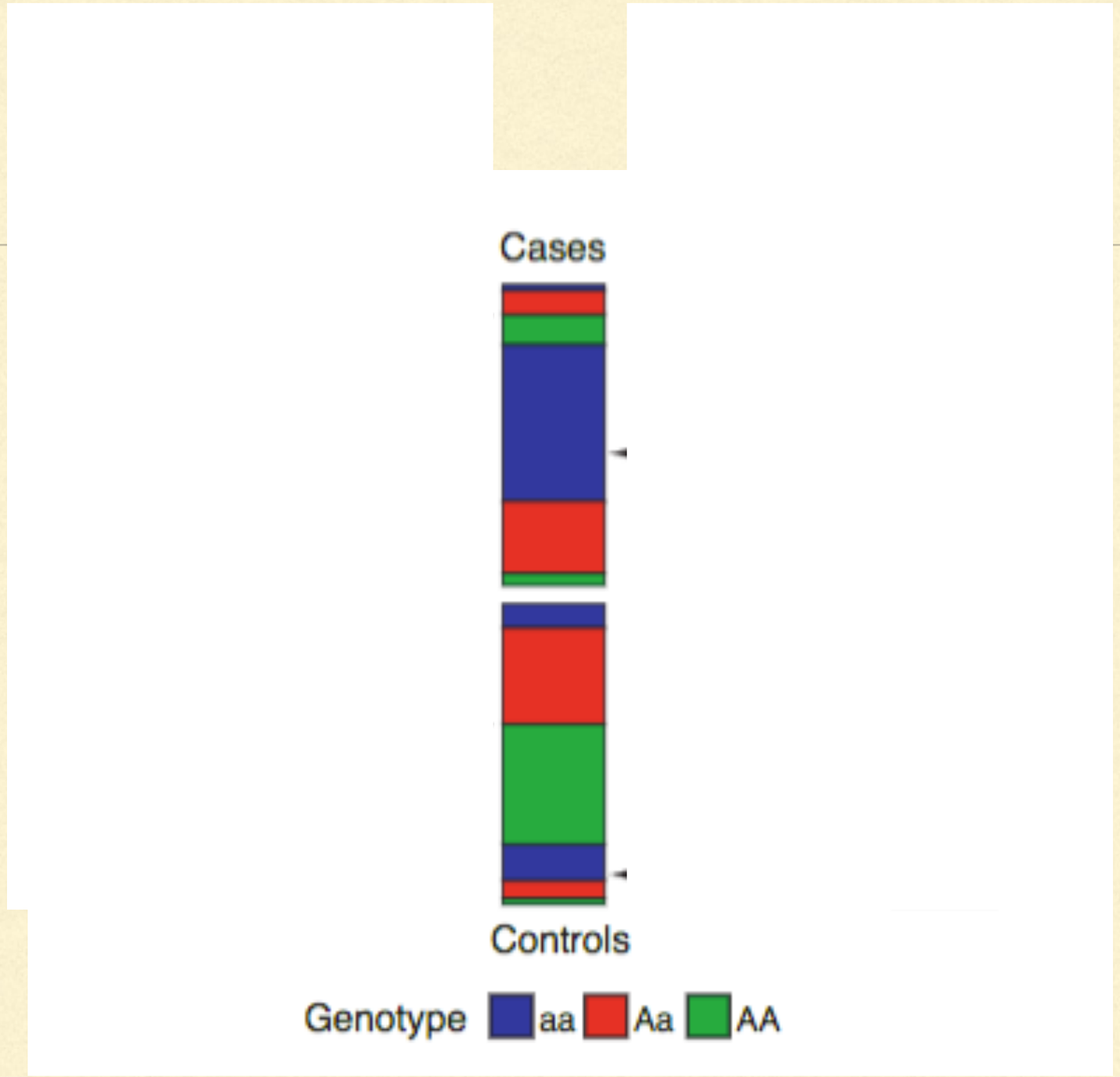
DNA sequencing at 40: past, present and future

Jay Shendure[1,2], Shankar Balasubramanian[3,4], George M. Church[5], Walter Gilbert[6], Jane Rogers[7], Jeffery A. Schloss[8] & Robert H. Waterston[1]

https://www.nationalgeographic.com/news/2010/3/human-genome-project-tenth-anniversary/

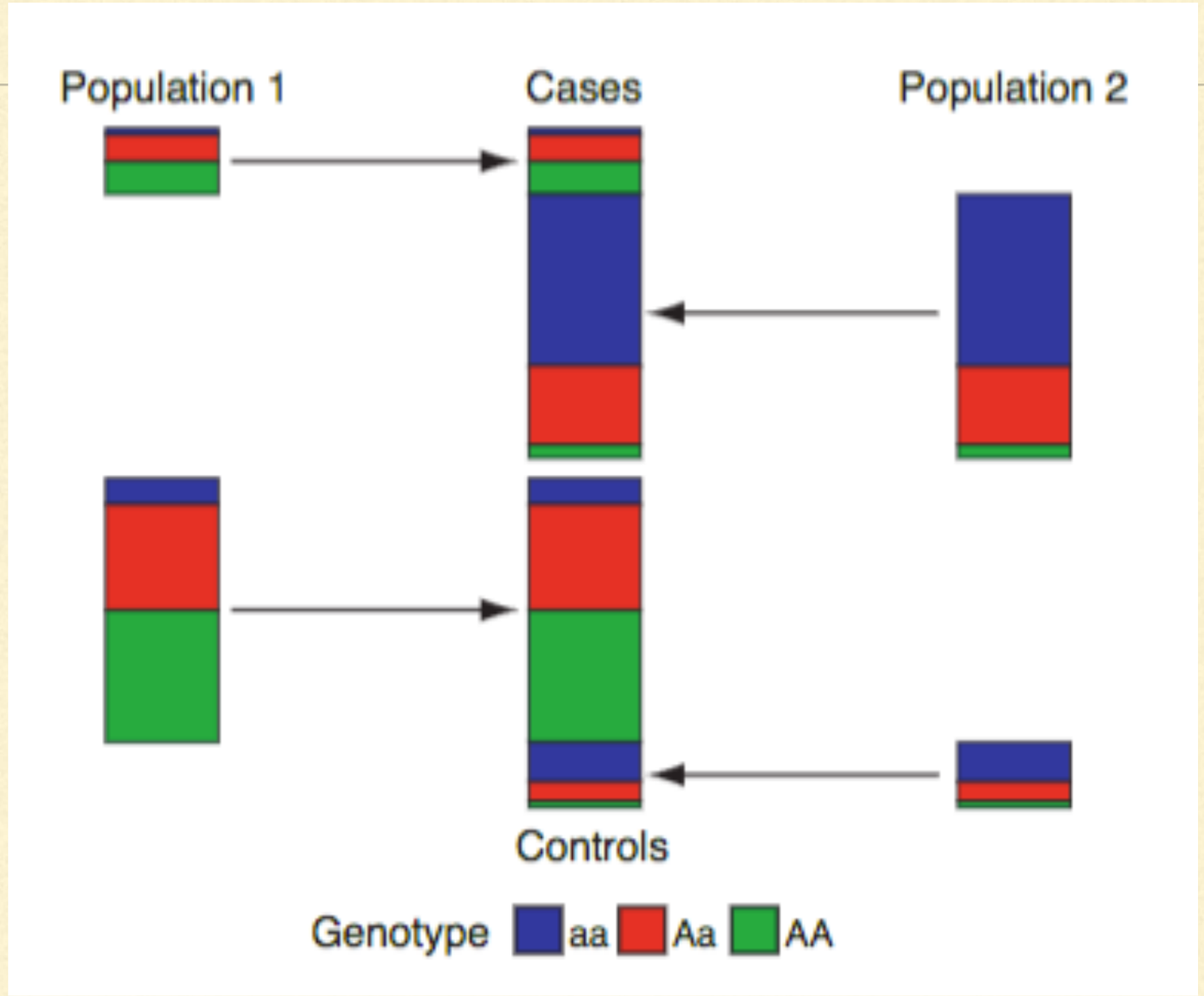
POPULATION STRUCTURE / STRATIFICATION

- 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 at 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.

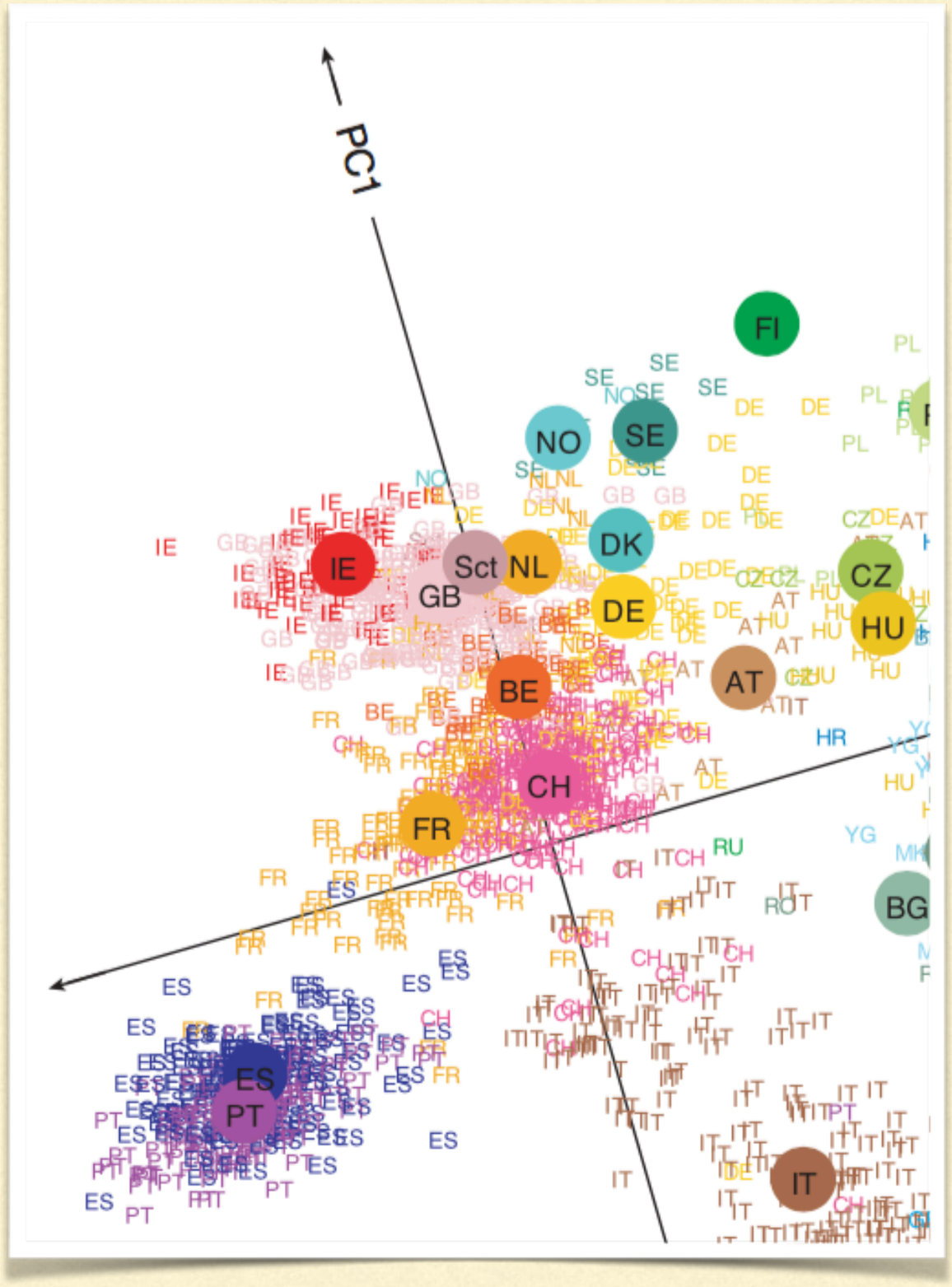
Marchini et al. 2004 Nature Genetics



Example of effects of pop structure at 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

POPULATION STRUCTURE: PCA AND ADMIXTURE



Chiara Barbieri

PCA - PRINCIPAL COMPONENT ANALYSIS

Principal component analysis (**PCA**) is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables (entities each of which takes on various numerical values) into a set of values of linearly uncorrelated variables called **principal components**. (Wikipedia)

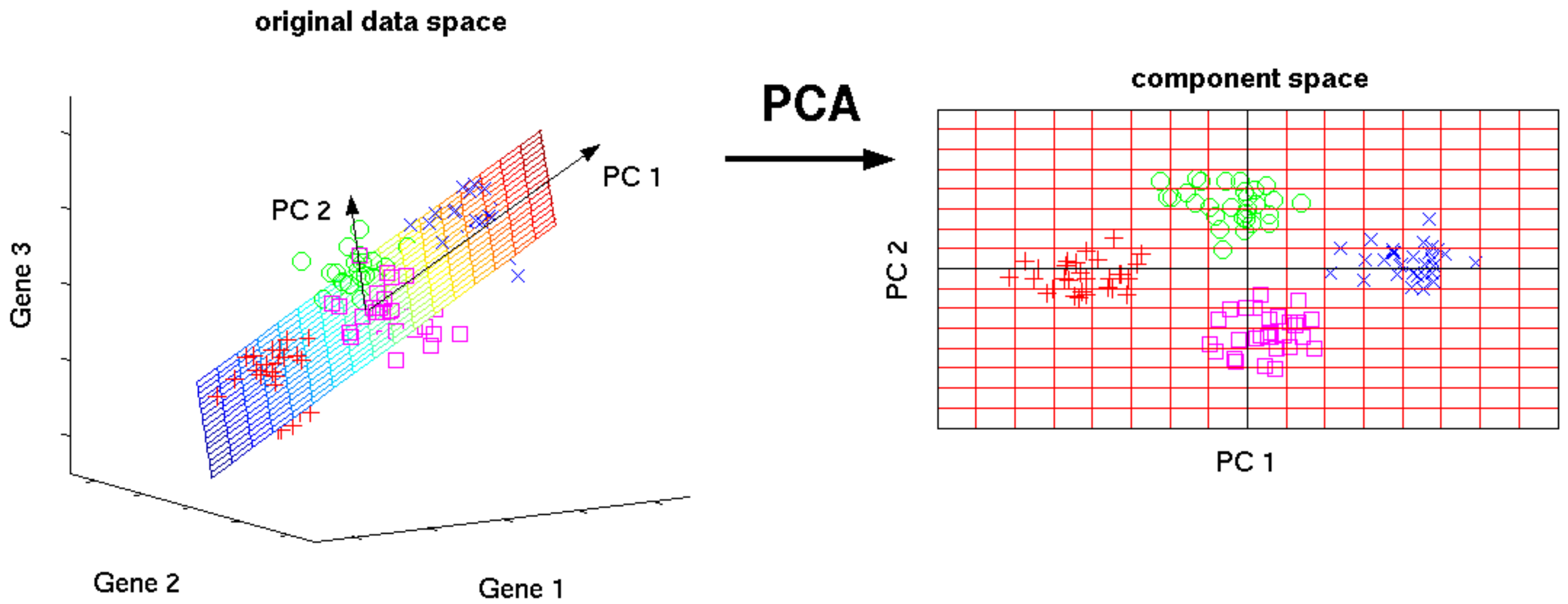
Each PC projection is designed to capture as much variance as possible for a linear transformation. The first dimension is the component which explains the highest variation.

Eigenvalues decomposition of a covariance matrix from uncorrelated source of variation (genetic markers).

The amount of eigenvectors/values that exist equals the number of dimensions the data set has.

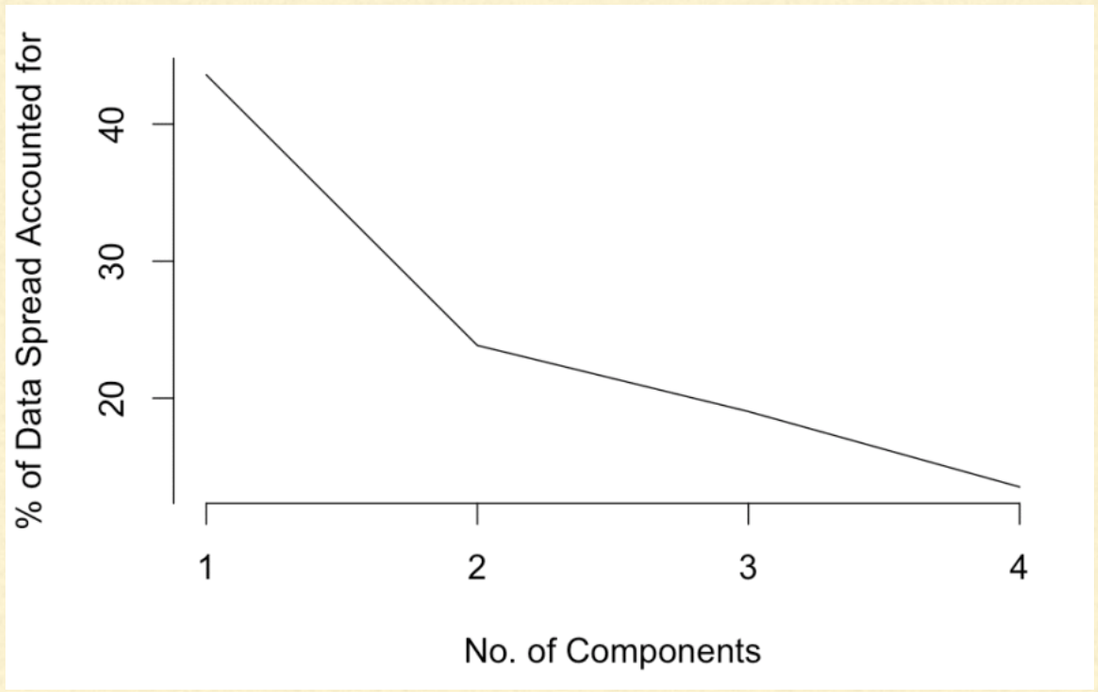
PCA

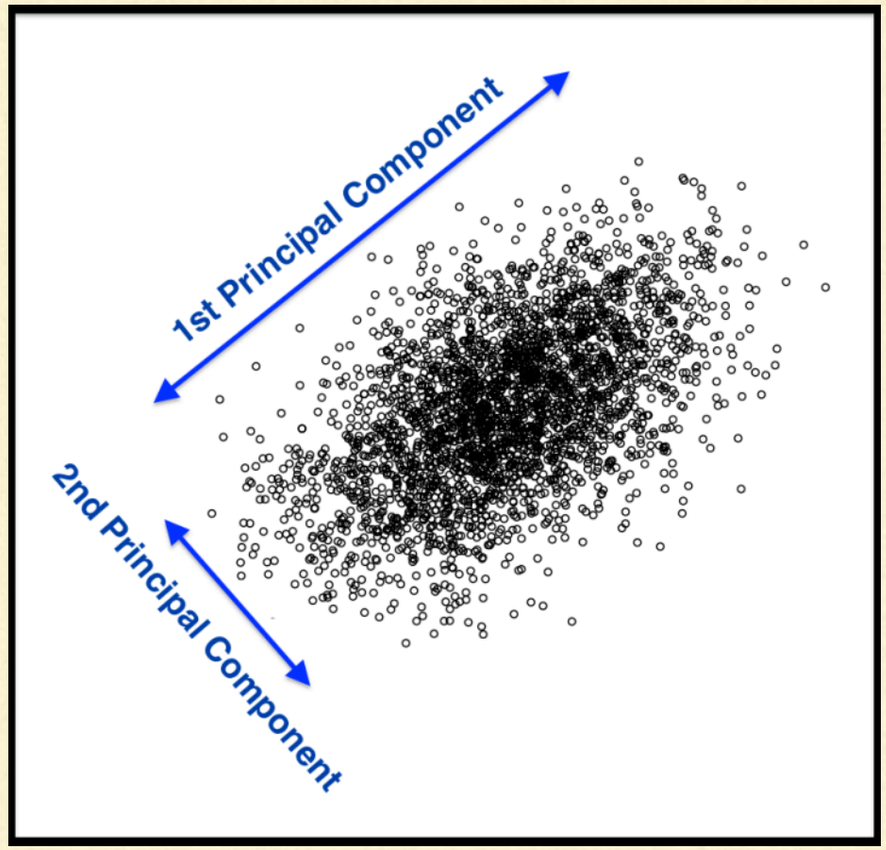
PCA is a standard technique for visualizing high dimensional data and for data pre-processing. PCA reduces the dimensionality (the number of variables) of a data set by maintaining as much variance as possible.



credits: www.nlpca.org

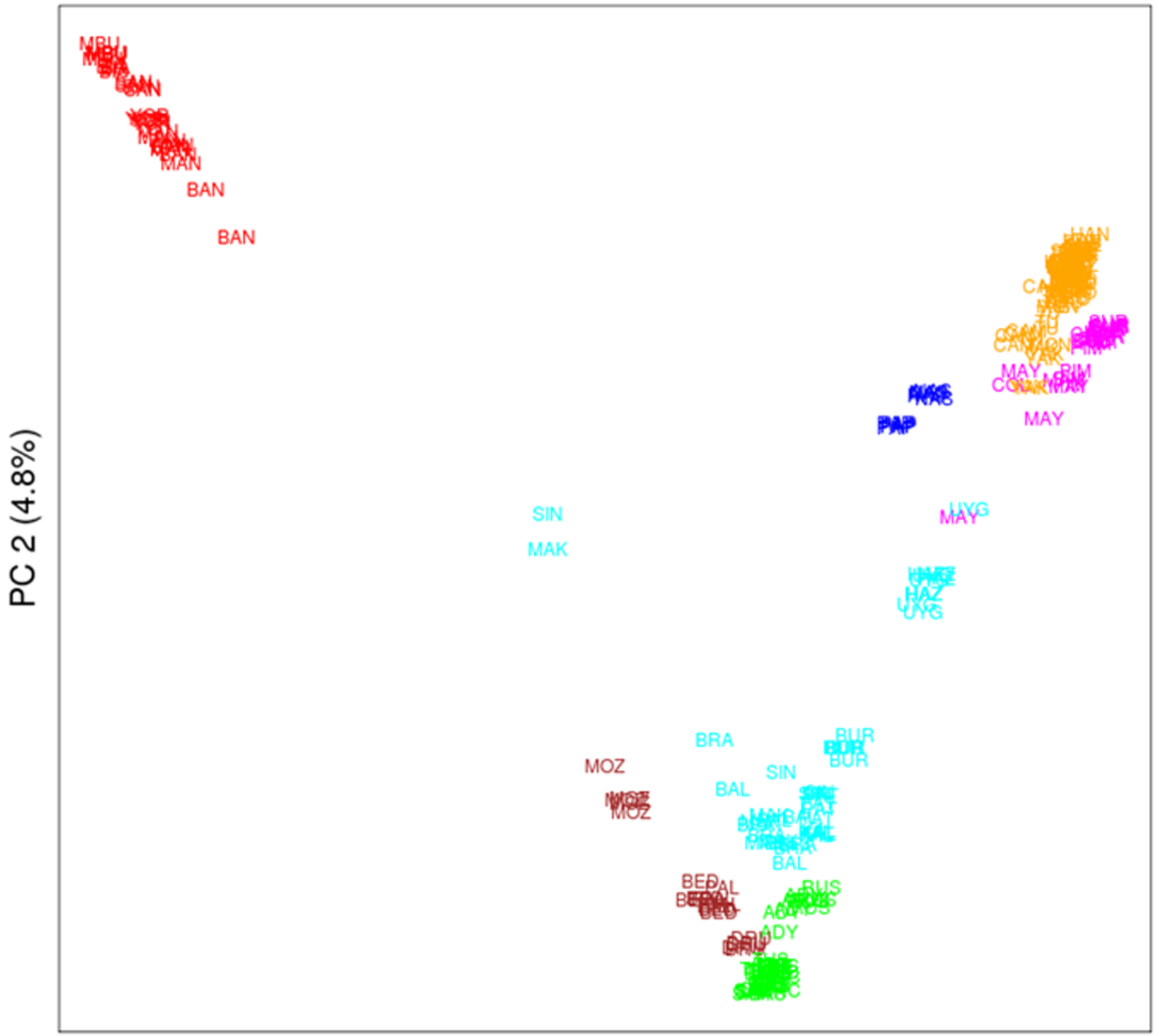
Principal components are also ordered by their effectiveness in differentiating data points, with the first principal component doing so to the largest degree. To keep results simple and generalizable, only the first few principal components are selected for visualization and further analysis. The number of principal components to consider is determined by something called a *scree plot*:





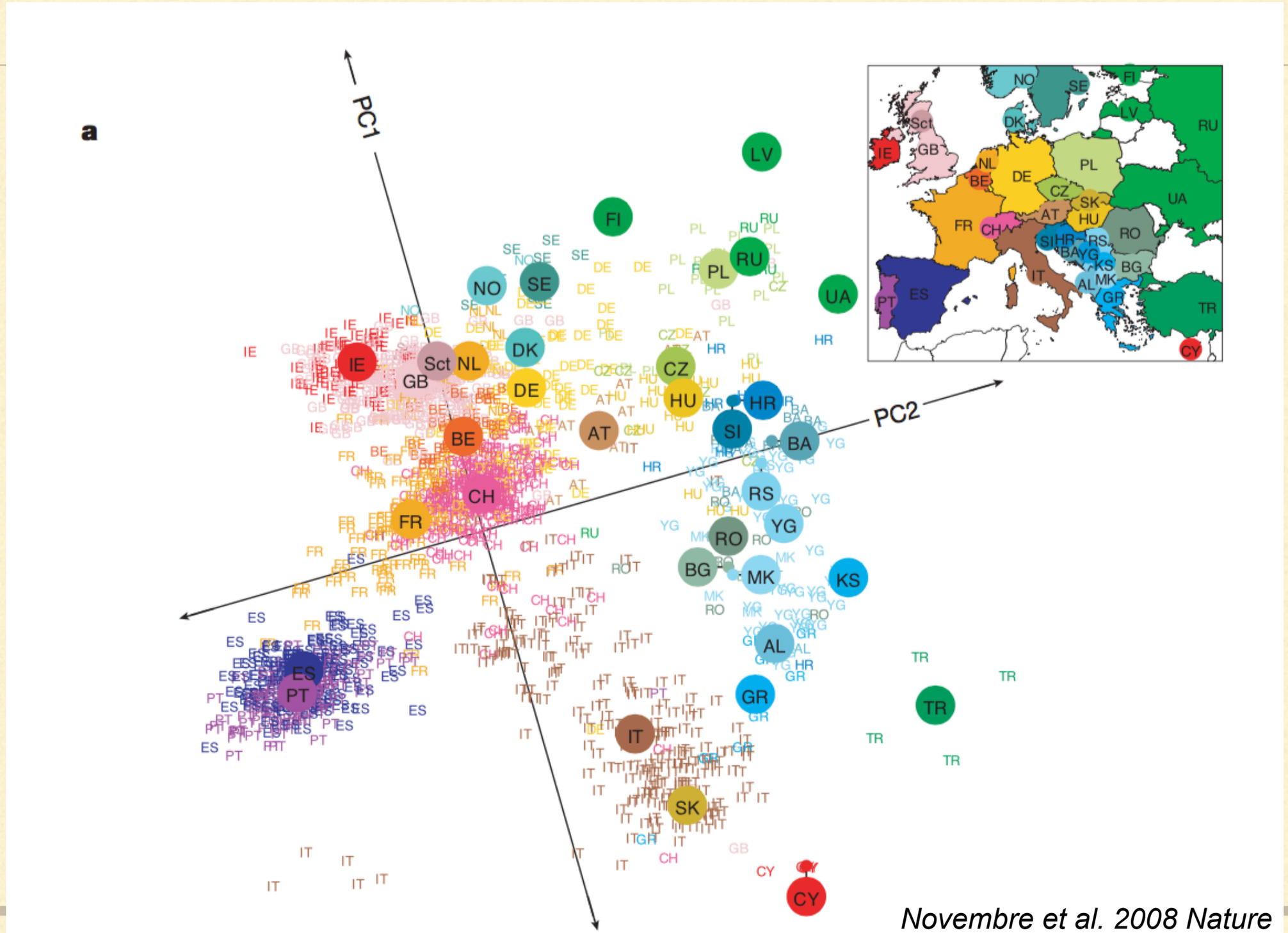
credits: algobeans.com

COL	AM - Colombian (Arawak)	TU	EA - Tu
KAR	AM - Karitiana	TUJ	EA - Tujia
MAY	AM - Maya	XIB	EA - Xibo
PIM	AM - Pima	YAK	EA - Yakut
SUR	AM - Surui	YIZ	EA - Yizu
BAL	CSA - Balochi	ADY	EUR - Adygei
BRA	CSA - Brahui	BAS	EUR - Basque
BUR	CSA - Burusho	BER	EUR - Bergamo
HAZ	CSA - Hazara	FRE	EUR - French
KAL	CSA - Kalash	ORC	EUR - Orcadian
MAK	CSA - Makrani	RUS	EUR - Russian
PAT	CSA - Pathan	SAR	EUR - Sardinian
SIN	CSA - Sindhi	TUS	EUR - Tuscan
UYG	CSA - Uygur	BED	ME - Bedouin
CAM	EA - Cambodia	DRU	ME - Druze
DAI	EA - Dai	MOZ	ME - Mozabite
DAU	EA - Daur	PAL	ME - Palestinian
HAN	EA - Han	NAS	OC - Nasioi
HEZ	EA - Hezhen	PAP	OC - Papuan
JAP	EA - Japanese	BAN	SSA - Bantu
LAH	EA - Lahu	BIA	SSA - BiakaPygmy
MIA	EA - Miaozu	MAN	SSA - Mandenka
MON	EA - Mongola	MBU	SSA - MbutiPygmy
NAX	EA - Naxi	SAN	SSA - San
ORO	EA - Oroqen	YOR	SSA - Yoruba
SHE	EA - She		



Lopez-Herraez et al. Plos ONE 2009

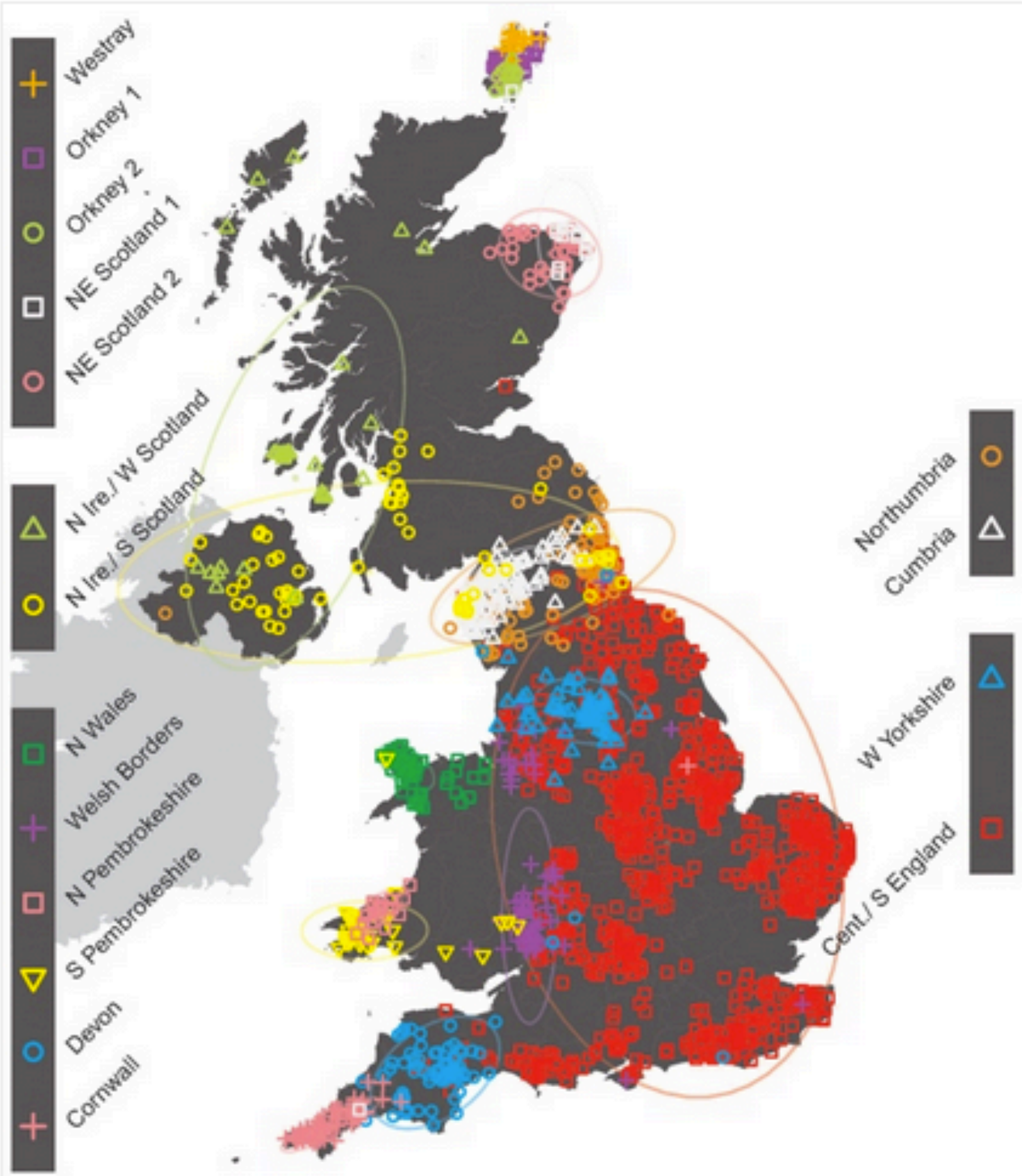
RELATIONSHIP BETWEEN INDIVIDUALS



Novembre et al. 2008 Nature

RELATIONSHIP BET

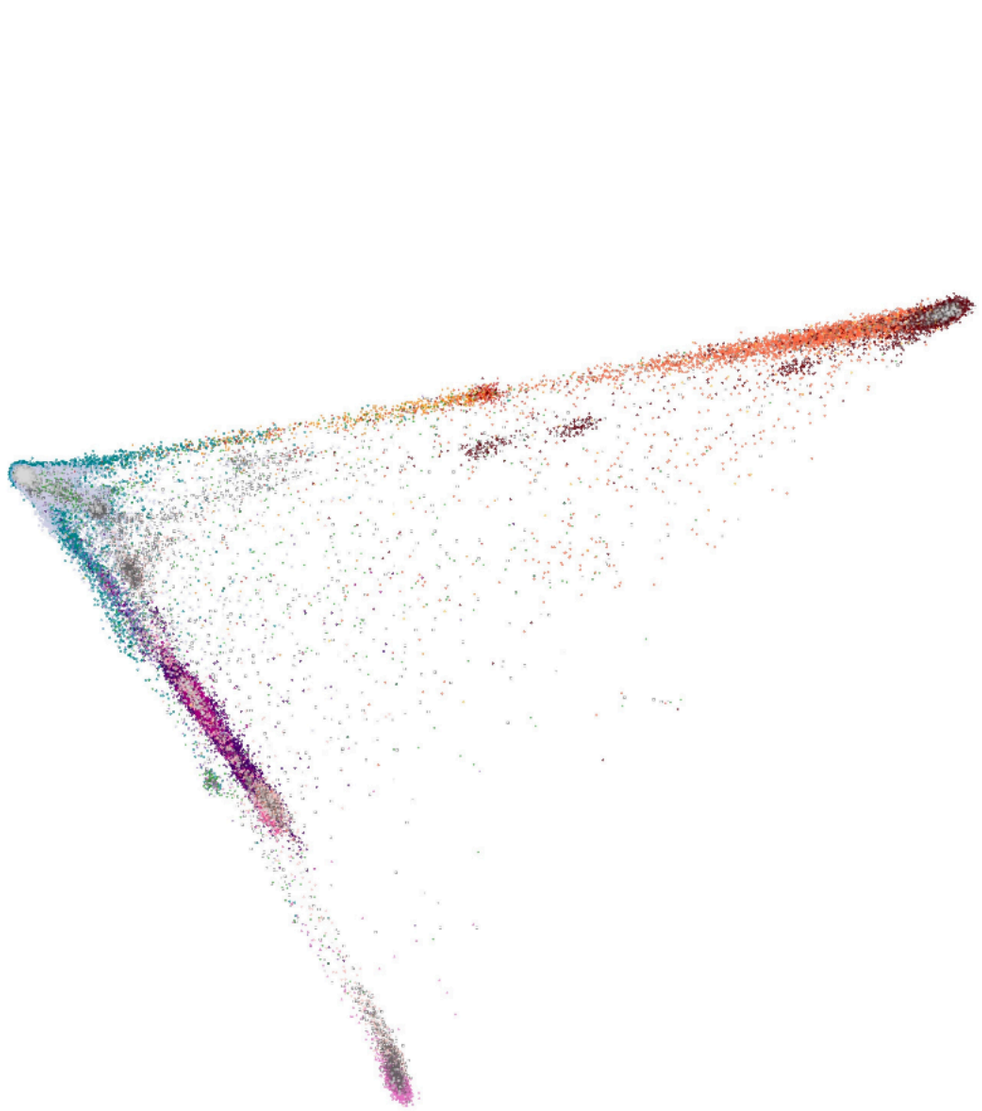
- …At finer scale



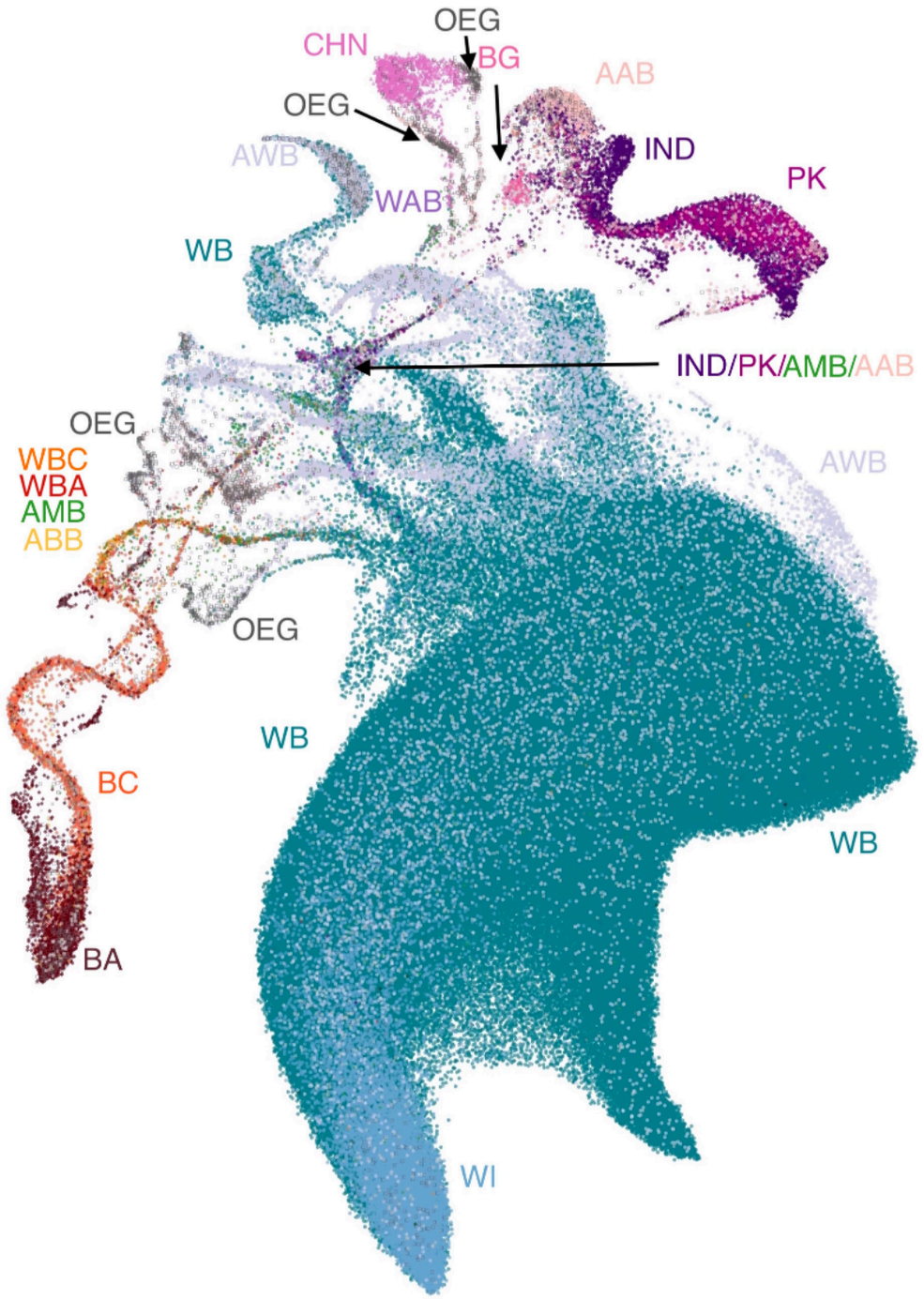
Stephen Leslie

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



(a) Principal components 1 and 2



(b) UMAP on first 10 principal components

UK BioBank dataset plotted with different PCA methods

Diaz-Papkovich et alo. 2018 BioRXiv

OTHER NON-LINEAR DIMENSIONAL REDUCTION METHODS

- MDS - multidimensional scaling
- Starts from a distance matrix to force individual relationships onto a fixed sets of dimensions (usually bi or tridimensional)
- The geometrical transformation aims at minimising the Stress (an estimate of how the data is fitting the space)

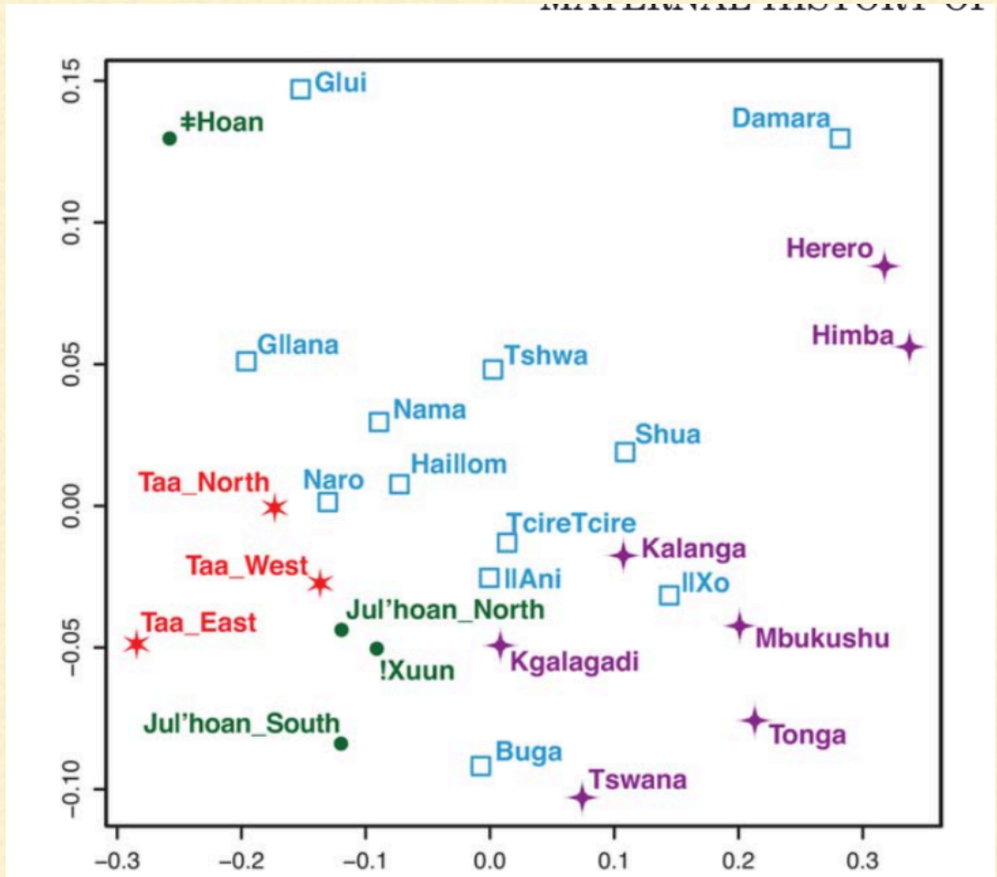
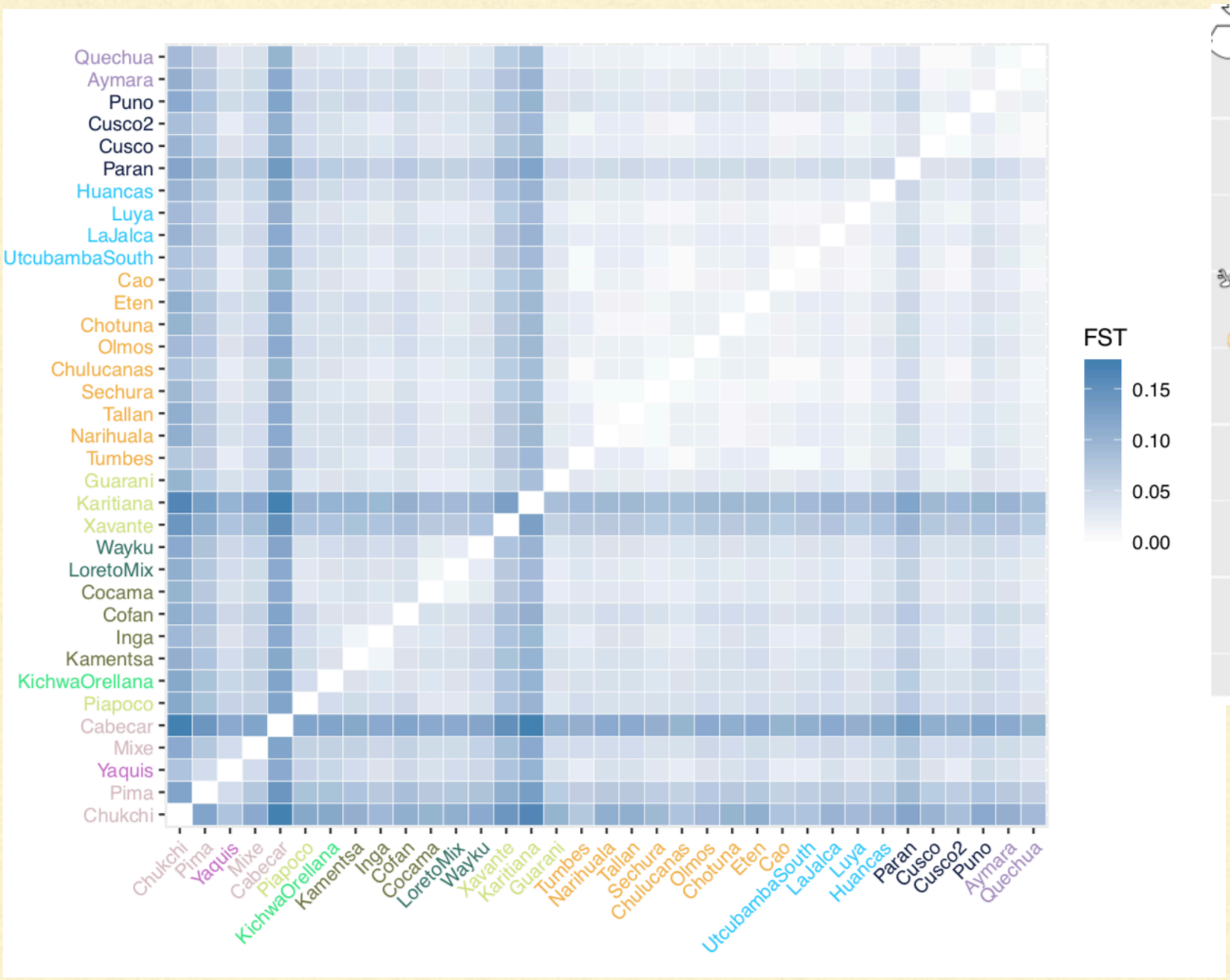


Fig. 2. Multidimensional Scaling plot based on Φ_{ST} distances. Population symbols indicate their linguistic affiliation, as shown in Figure 1. Stress value: 7.97%. [Color figure can be

Barbieri et al. AJPA 2014

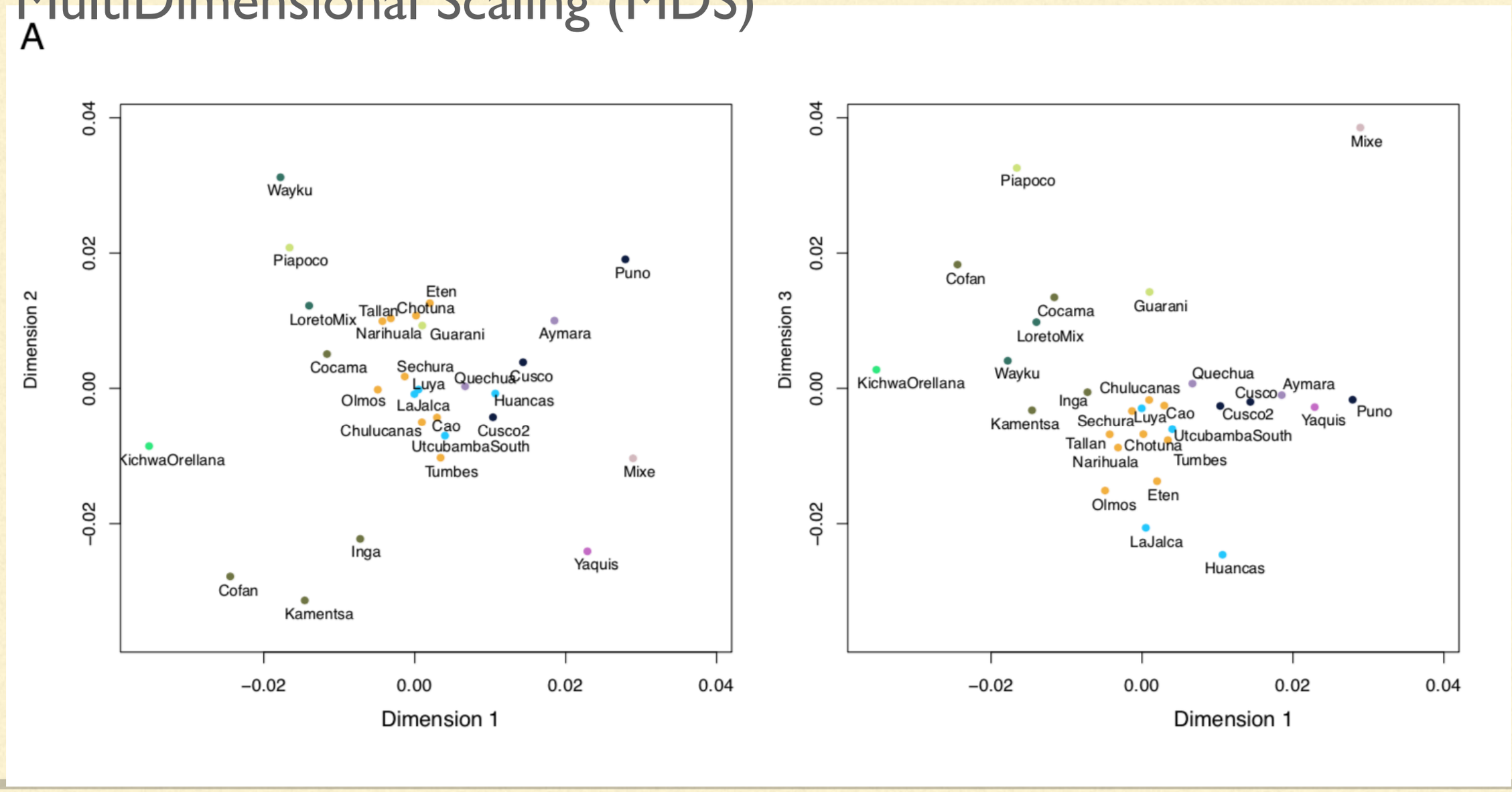
FST DISTANCE MATRIX

RELATIONSHIP BETWEEN POPULATIONS



RELATIONSHIP BETWEEN POPULATIONS

- MultiDimensional Scaling (MDS)



Barbieri et al. 2019 Mol Biol Evol

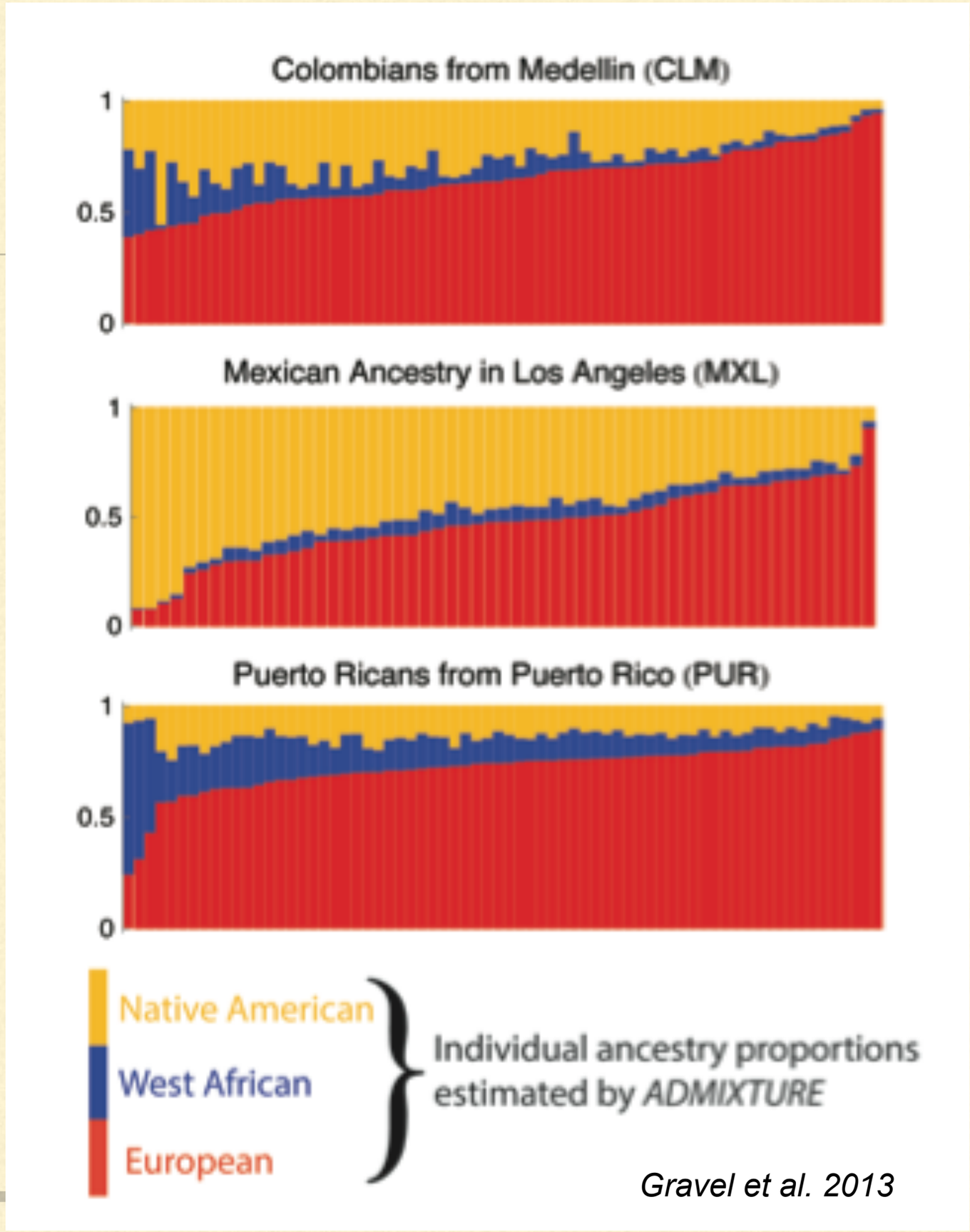
ADMIXTURE

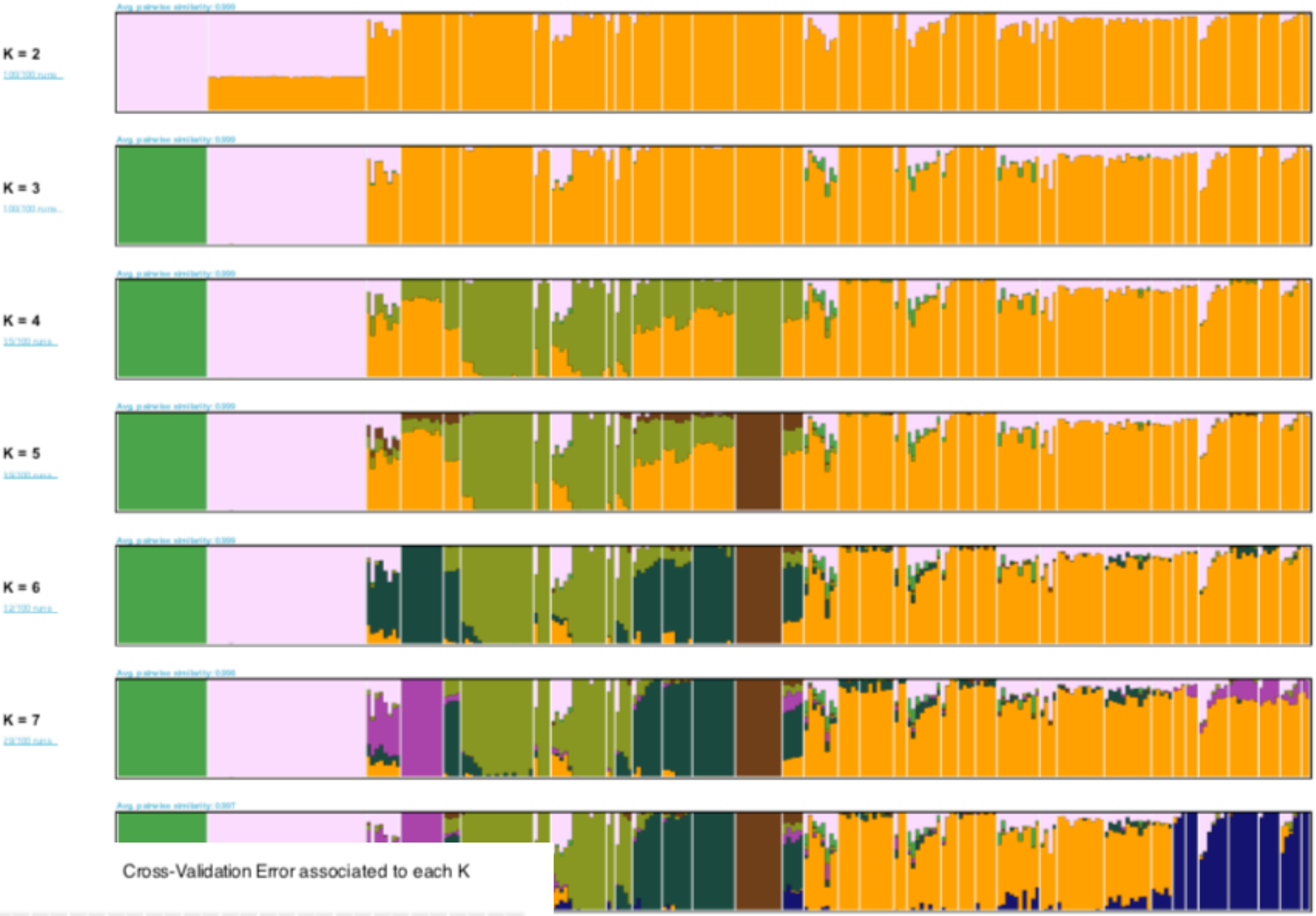
Admixture is a very useful and popular tool to analyse SNP data. It performs an unsupervised clustering of large numbers of samples, and allows each individual to be a mixture of clusters.

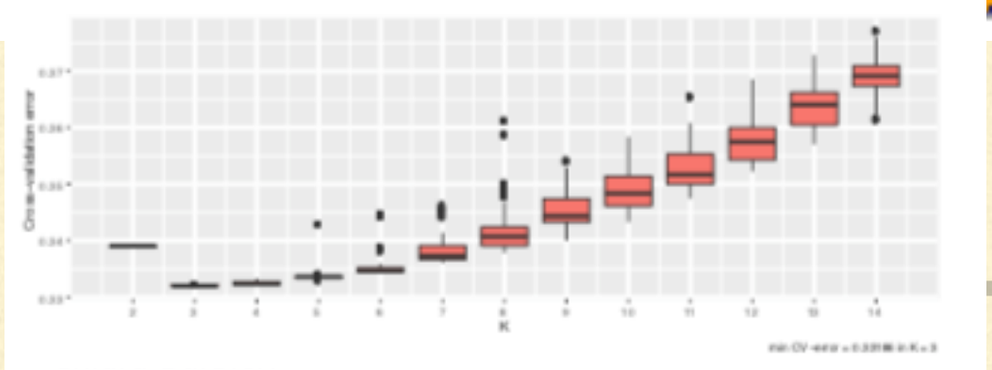
STRUCTURE barplot has become a de-facto standard used as a non-parametric description of genetic data[7] alongside a Principle Components Analysis.

ADMIXTURE

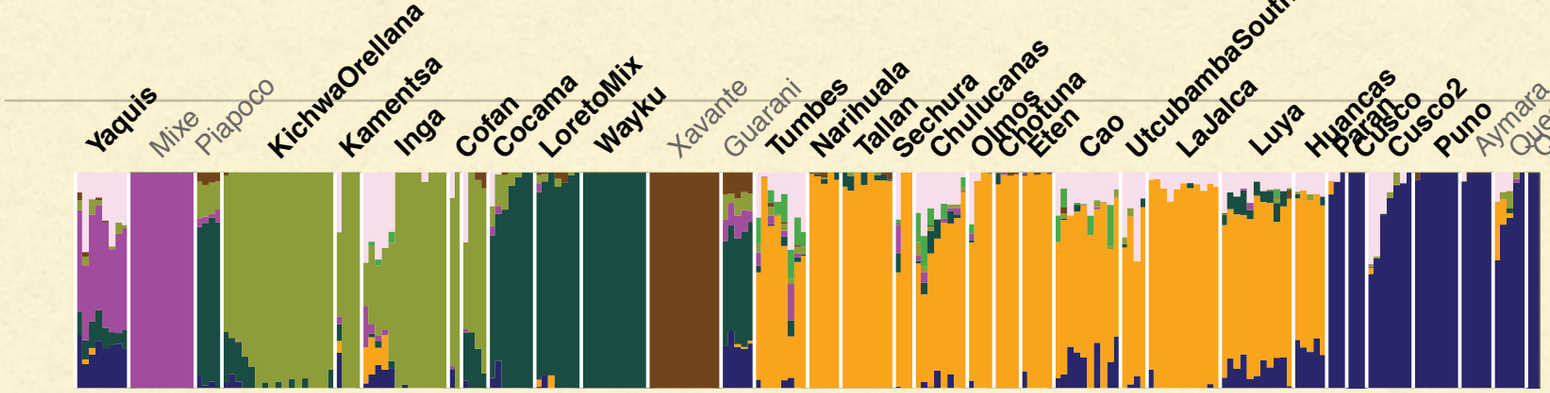
- Explain genetic diversity in terms of ancestry blocks

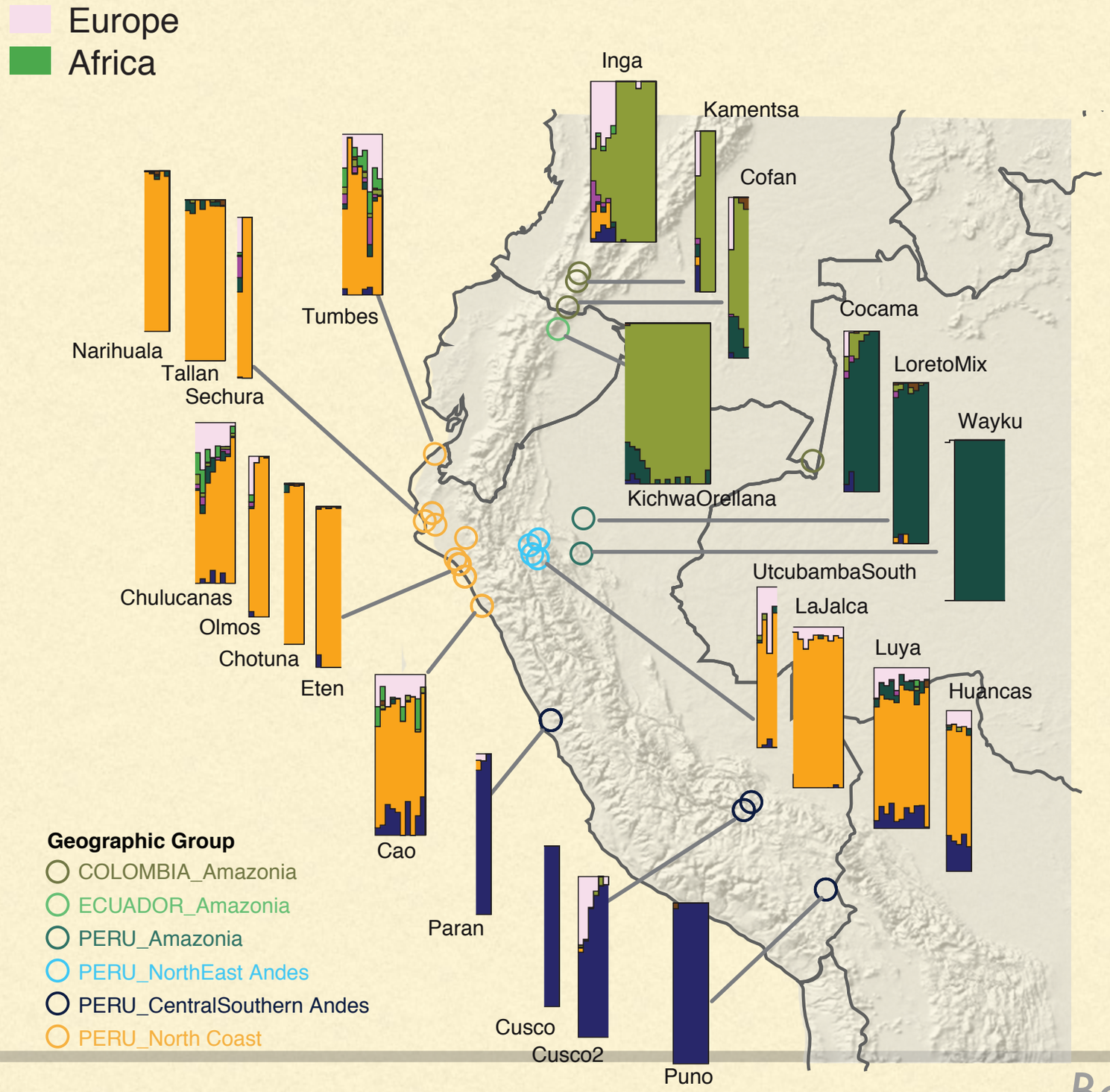






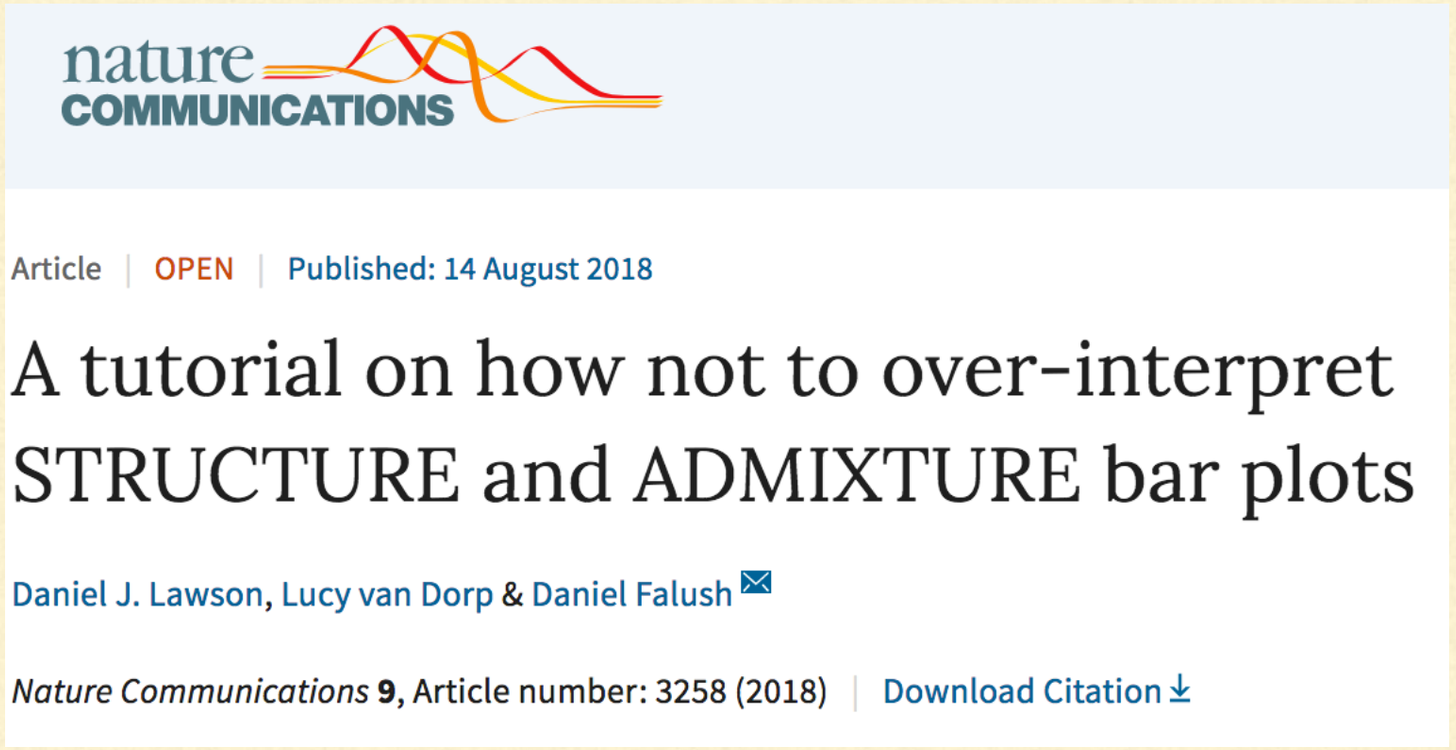
Barbieri et al. 2019 Mol Biol Evol


Yaquis
Mixe
Piapoco
KichwaOrellana
Kamentsa
Inga
Cofan
Cocama
LoretoMix
Wayku
Xavante
Guarani
Tumbes
Narihuala
Tallan
Sechura
Chulucanas
Olmos
Chotuna
Eten
Cao
UtcubambaSouth
LaJalca
Luya
Huancas
Paran
Cusco
Cusco2
Puno
Aymara
Quechua
Quechua2


Europe
Africa
Inga
Kamentsa
Cofan
Tumbes
Cocama
LoretoMix
Wayku
KichwaOrellana
Narihuala
Tallan
Sechura
Chulucanas
Olmos
Chotuna
Eten
UtcubambaSouth
LaJalca
Luya
Huancas
Cao
Paran
Cusco
Cusco2
Puno
Geographic Group
COLOMBIA_Amazonia
ECUADOR_Amazonia
PERU_Amazonia
PERU_NorthEast Andes
PERU_CentralSouthern Andes
PERU_North Coast

LIMITS OF ADMIXTURE

Histories can also be reconstructed using the same procedure for groups that do not have admixture in their recent history, where recent genetic drift is strong or that deviate in other ways from the underlying inference model. Unfortunately, such histories can be misleading.



nature COMMUNICATIONS

Article | OPEN | Published: 14 August 2018

A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE bar plots

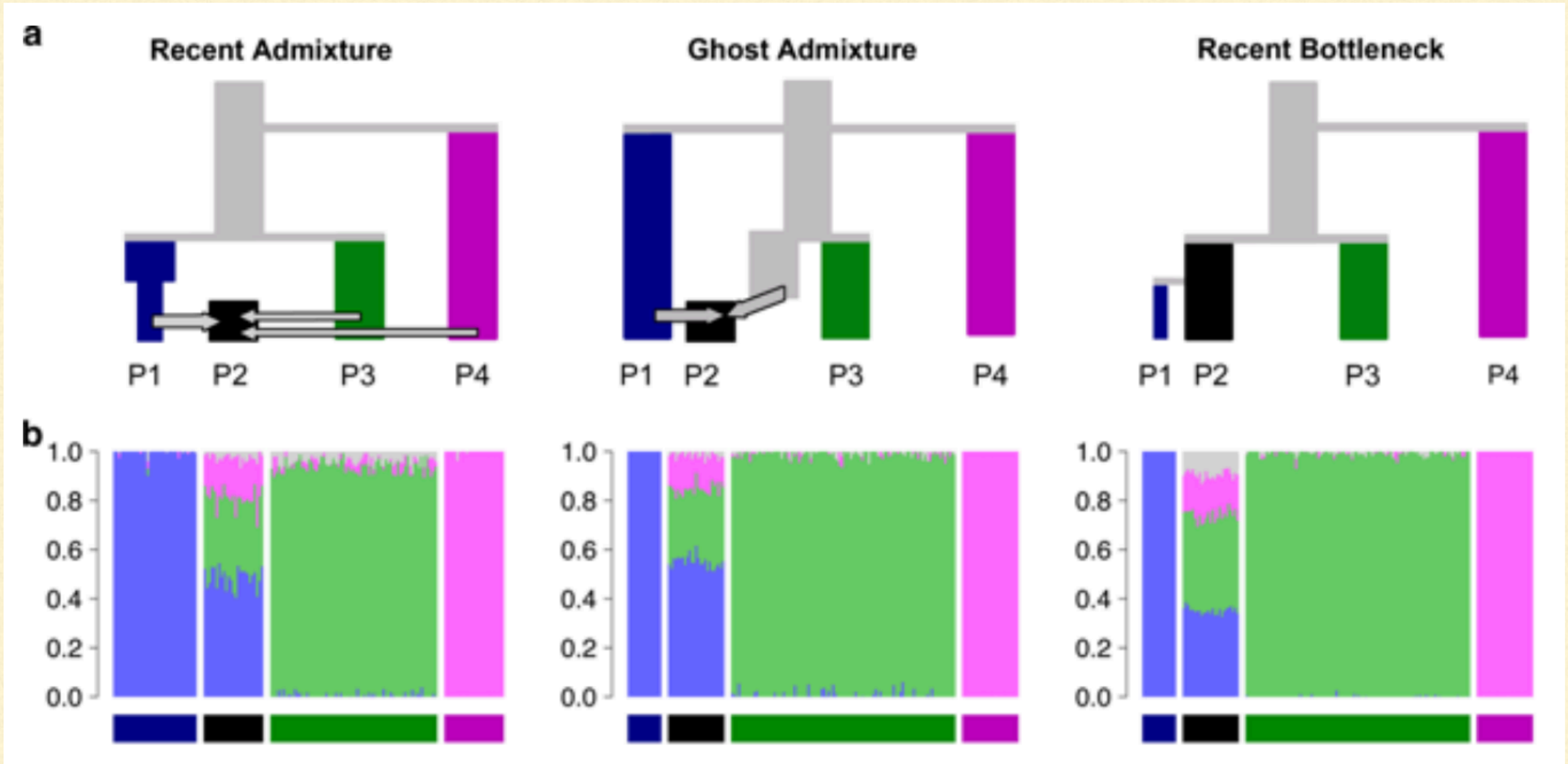
Daniel J. Lawson, Lucy van Dorp & Daniel Falush

Nature Communications **9**, Article number: 3258 (2018) | Download Citation

A procotol for (mis)interpreting admixture model results

Algorithm input	Core protocol	Interpretation
(0) Over-sample particular populations, relative to either current or historical sample sizes.	**(1) Estimate *K* using a refined statistical procedure.**	
	(2) Assume that the chosen value of *K* is the true value of *K*.	(3a) Choose the value of *K* that gives results closest to your prior expectation.
	(3) Assume each of the *K* ancestral population existed at some point in the past.	(3b) Label ancestral populations based on the locations they are currently most frequent in.
	4) Assume that modern individuals were produced by recent mixing of these ancestral populations.	(3c) Assume negligible substructure within the inferred ancestral populations.
Features of the data and algorithm input	*Assumptions of the modelling procedure*	*Features of the interpretation of ancestral populations*

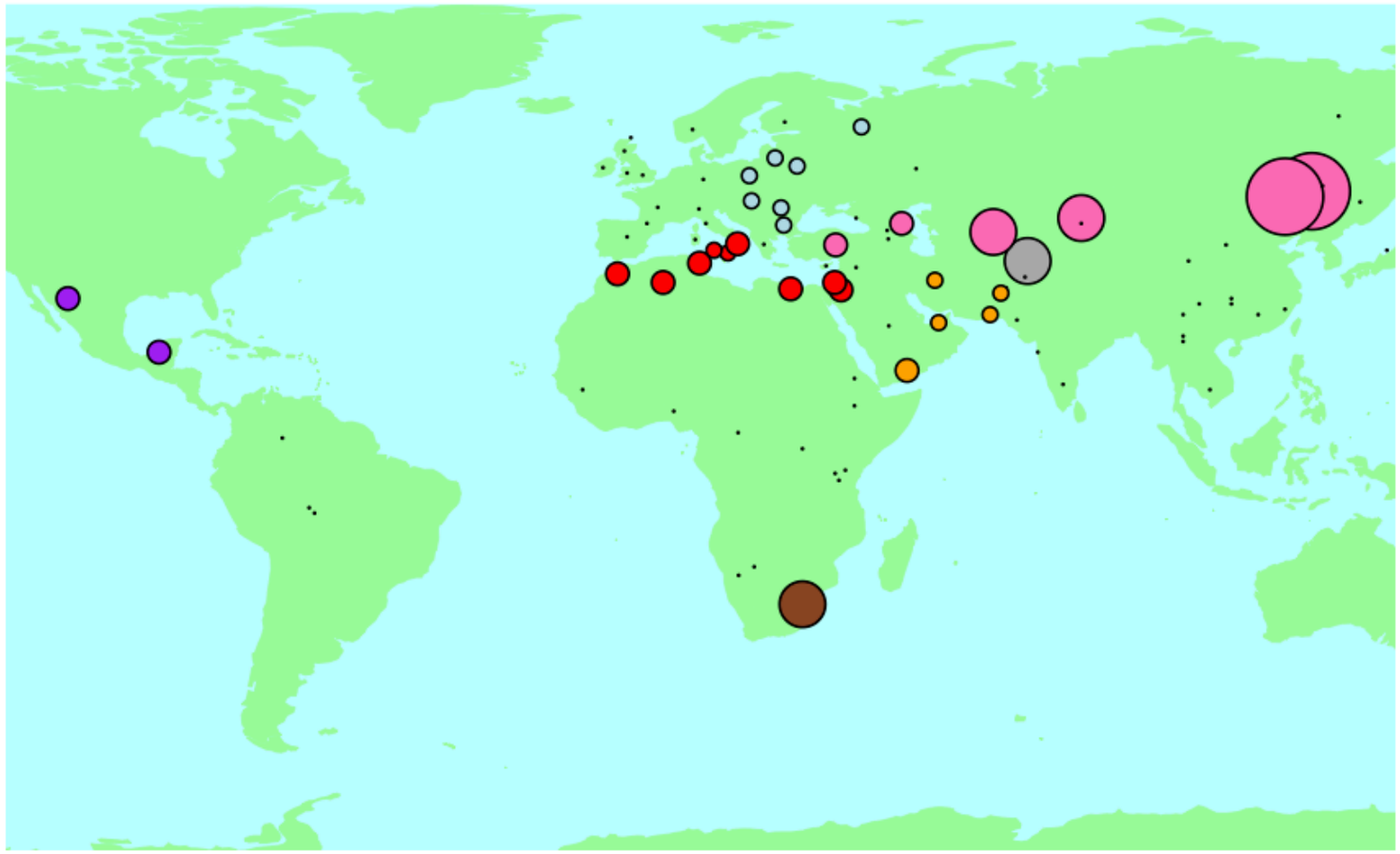
3 different population histories giving the same ADMIXTURE results



FINDING ADMIXTURE SOURCES AND DATING ADMIXTURE EVENTS

A Genetic Atlas of Human Admixture History

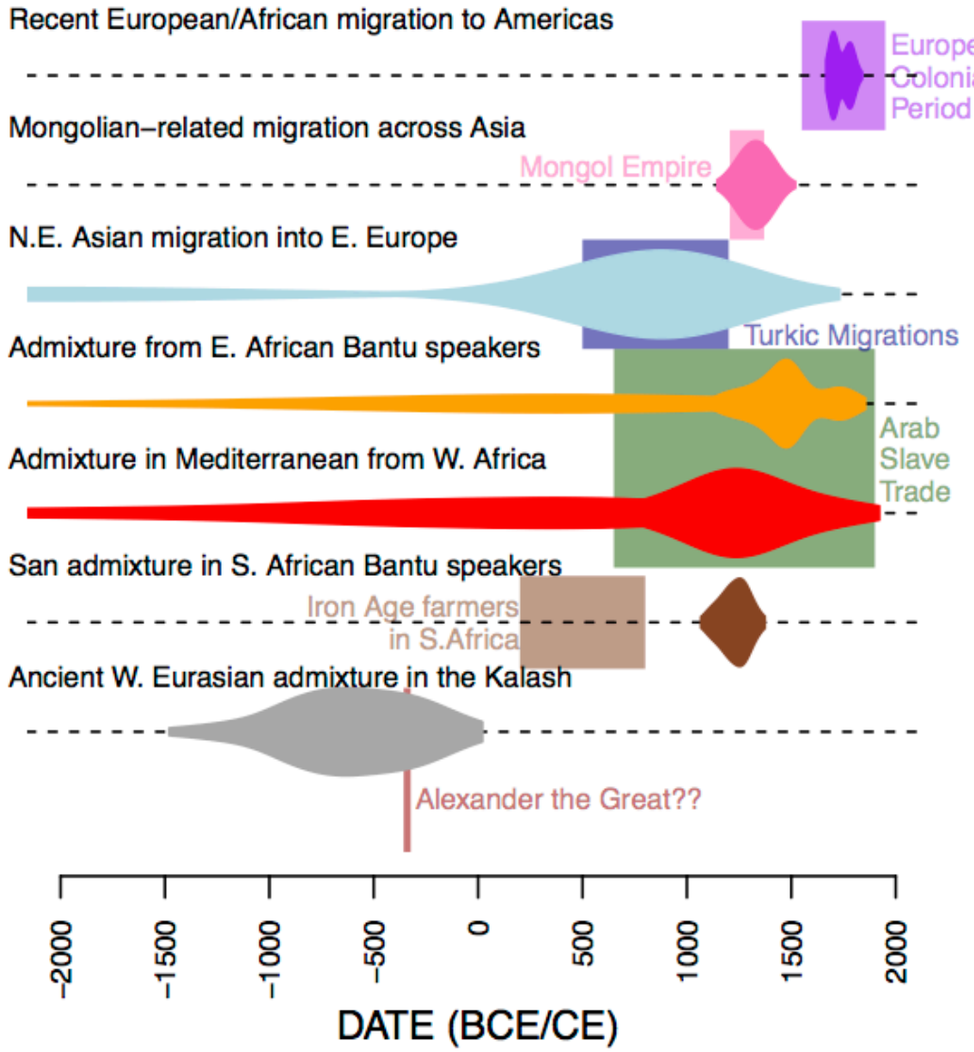
A summary of a selection of human migration events inferred only from genetic data using the new GLOBETROTTER method reported in: G. Hellenthal, G.B.J. Busby, G. Band, J.F. Wilson, C. Capelli, D. Falush & S. Myers **Science** *(14th Feb 2014)*



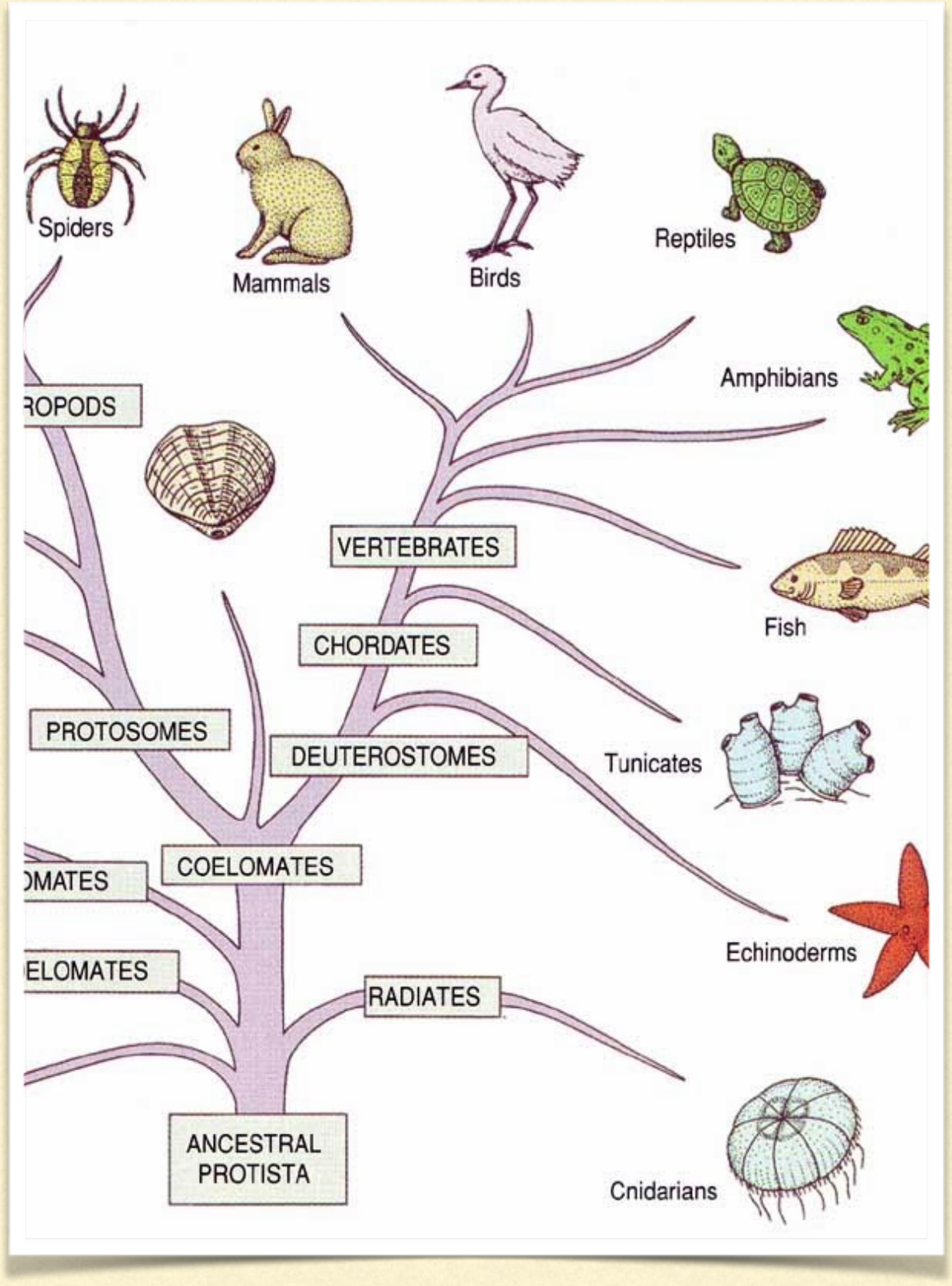
Proportion of DNA in a population from an incoming admixture event

NA 1–5% 5–20% 20–50% 50%+

Timeline of events inferred by GLOBETROTTER



PHYLOGENETIC RECONSTRUCTIONS

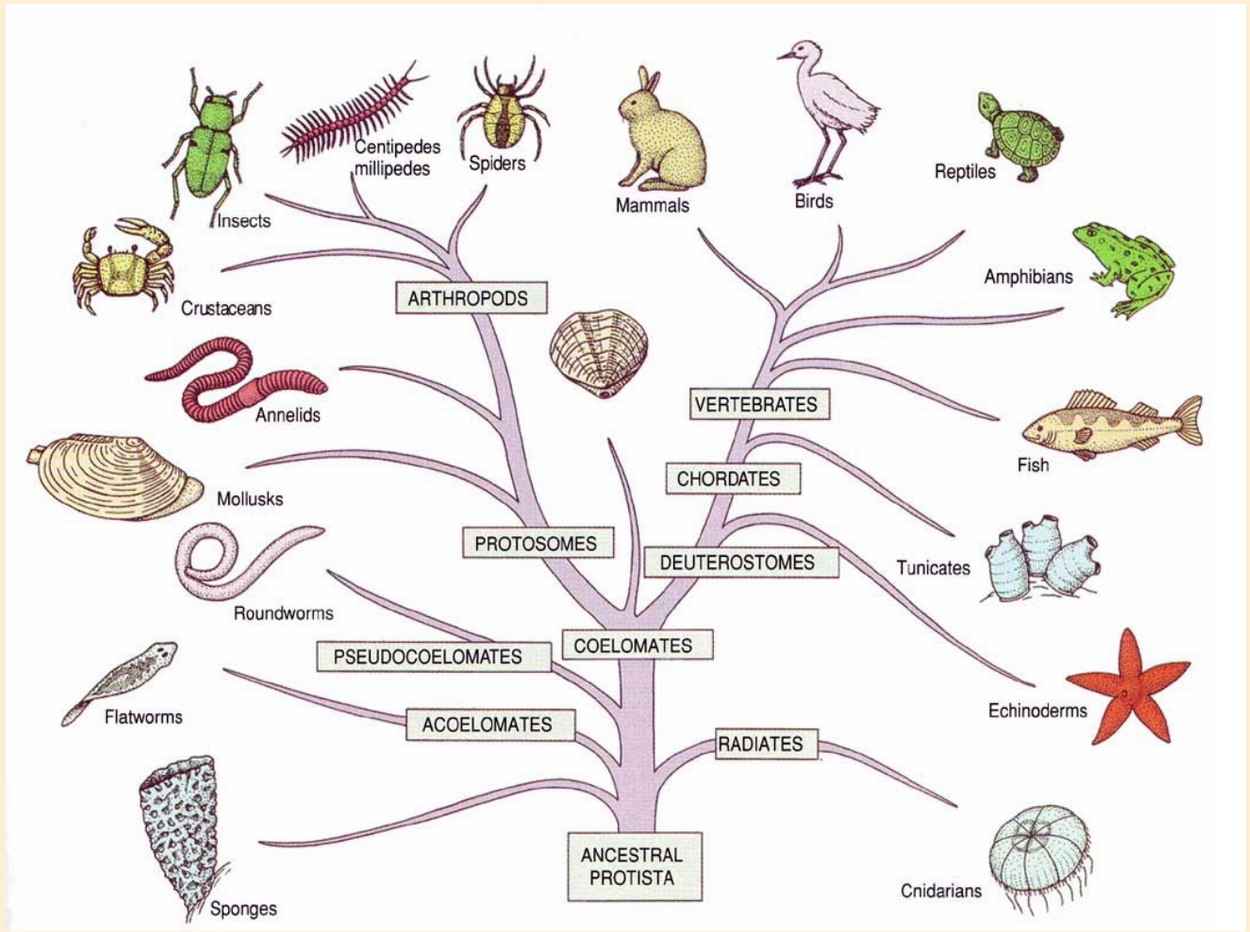


Chiara Barbieri

NOTHING IN BIOLOGY MAKES SENSE EXCEPT IN THE LIGHT OF EVOLUTION

T. Dobzhansky – 1973

EVERYTHING IN BIOLOGY MAKES MORE SENSE WITH A PHYLOGENETIC TREE



(muchos biólogos)

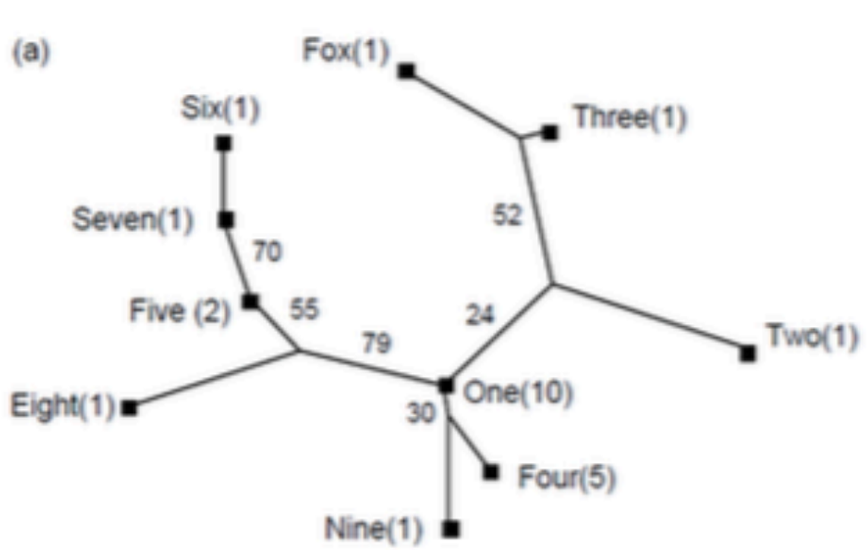
PHYLOGENETIC NETWORKS

- “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 TREES AND NETWORKS

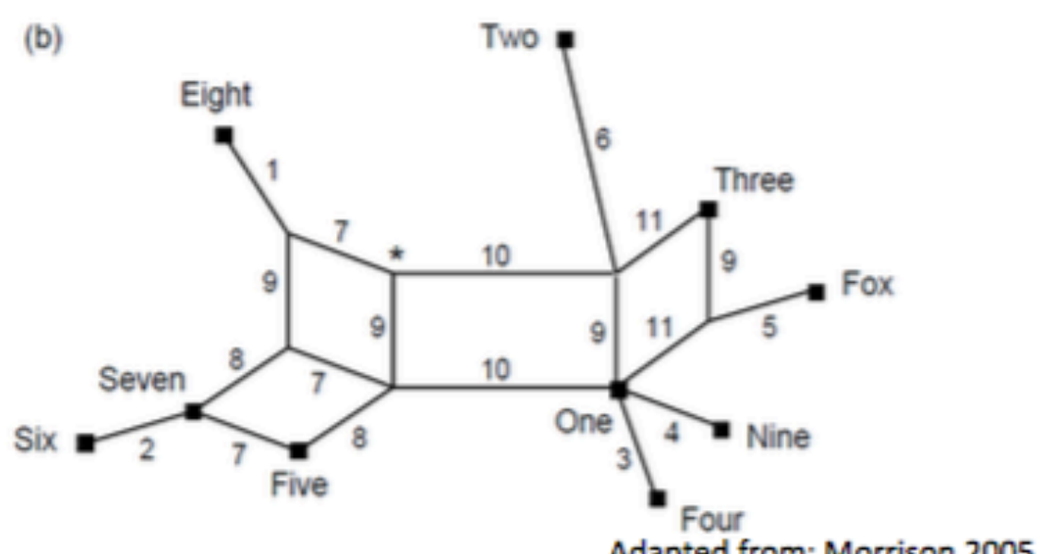
- Trees impose bifurcations
- Networks allow reticulations

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.

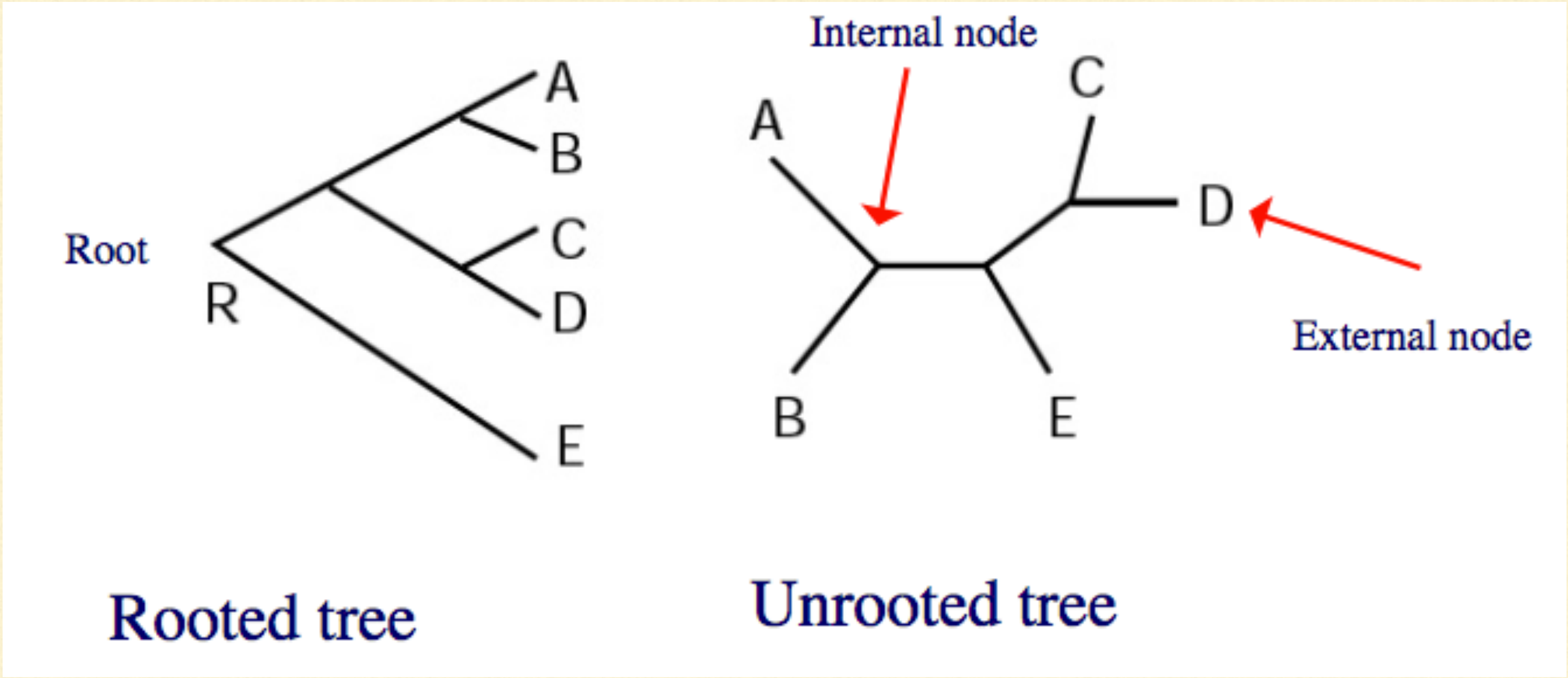
Phylogenetic network



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



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

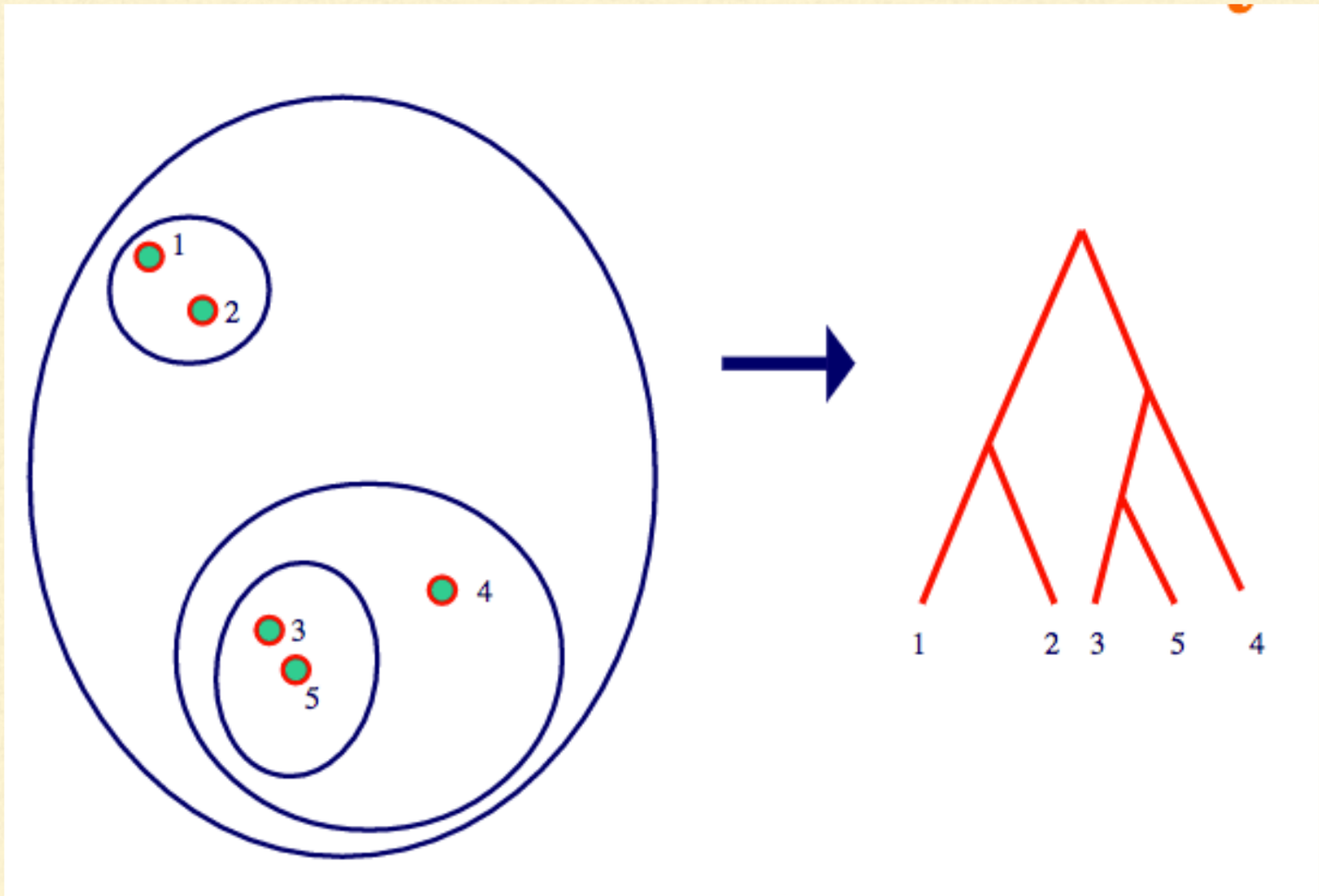
DISTANCE-BASED TREES

- 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

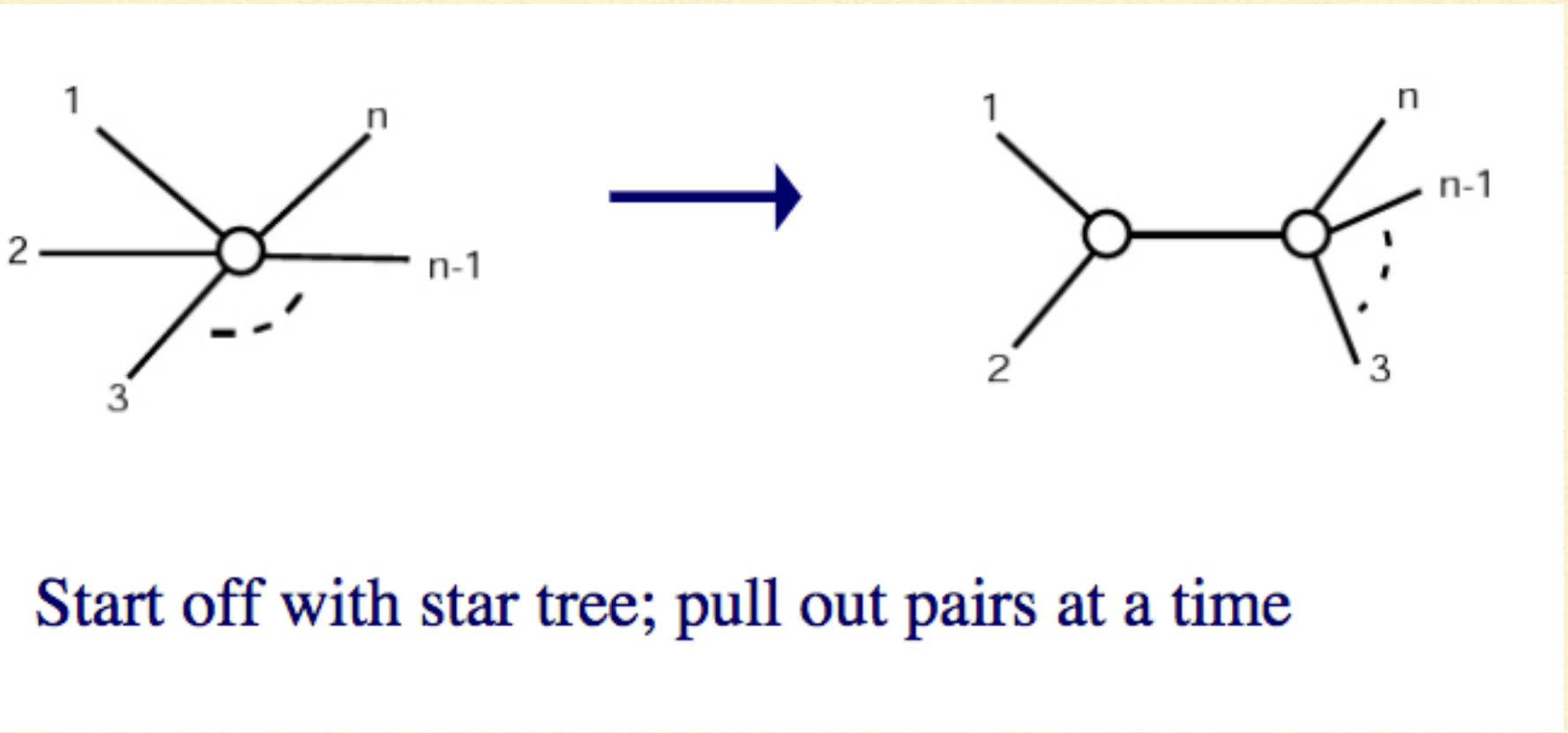
- **Neighbor-Joining NJ** (Saitou and Nei 1987)
  - **Unrooted** tree
  - Does not assume a **molecular clock**.

- Local search strategy using a Minimum Evolution (ME) optimality criteria

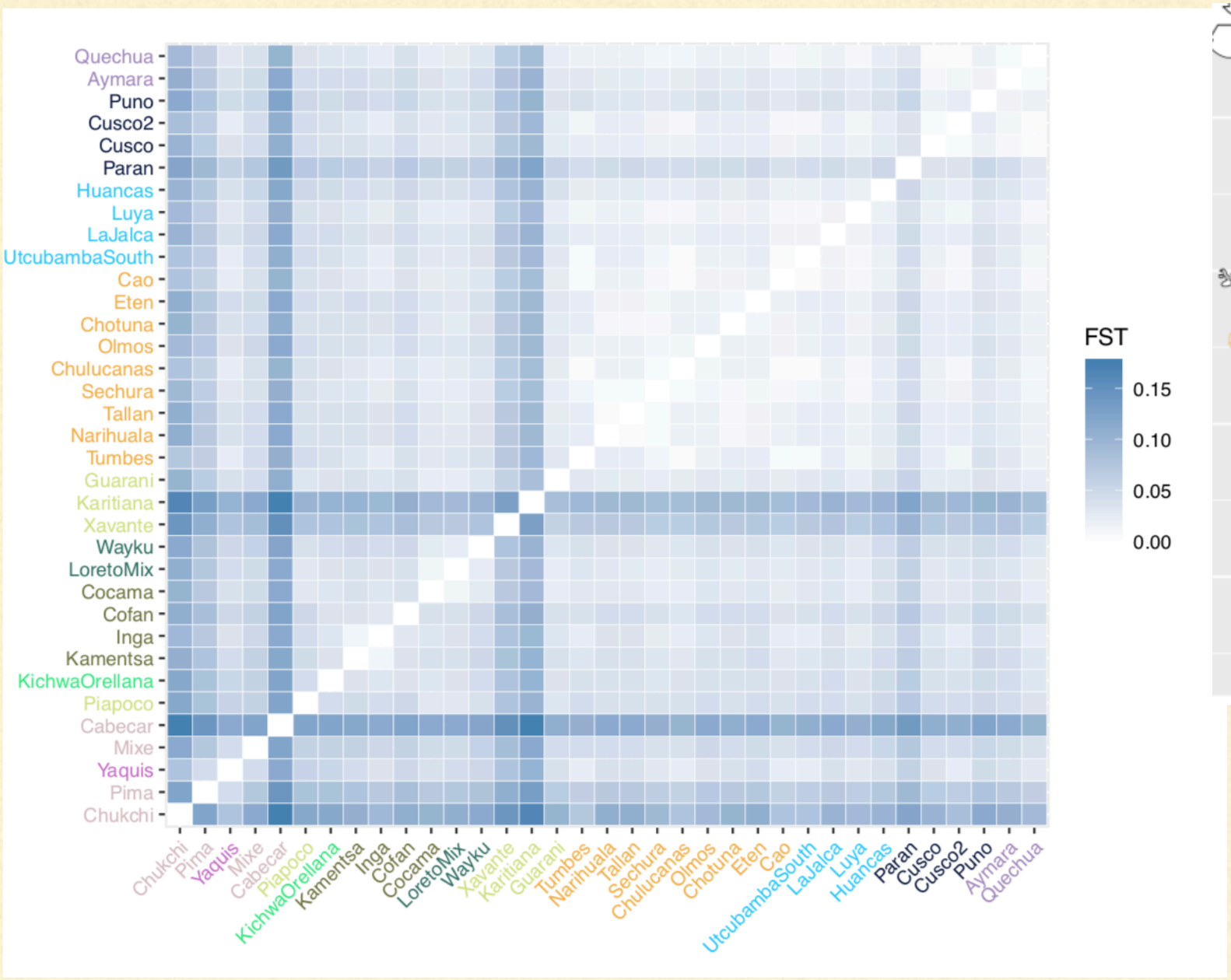
- Starts with an unresolved star-like tree, calculate the sum of branch length. Joins the pair with the closest branch length. And so on

DISTANCE-BASED TREES

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

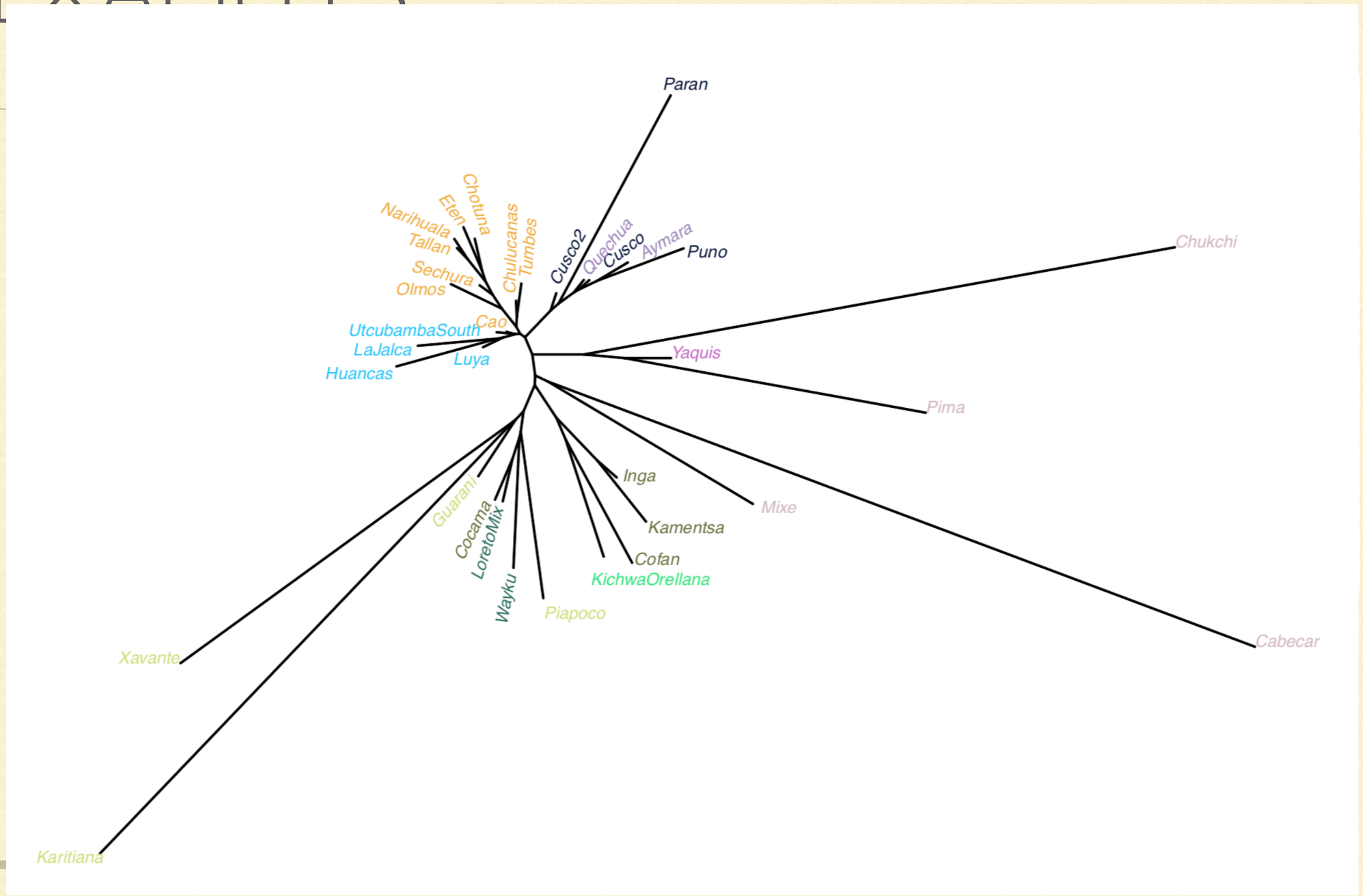


EXAMPLES


Quechua
Aymara
Puno
Cusco2
Cusco
Paran
Huancas
Luya
LaJalca
UtcubambaSouth
Cao
Eten
Chotuna
Olmos
Chulucanas
Sechura
Tallan
Narihuala
Tumbes
Guarani
Karitiana
Xavante
Wayku
LoretoMix
Cocama
Cofan
Inga
Kamentsa
KichwaOrellana
Piapoco
Cabecar
Mixe
Yaquis
Pima
Chukchi
FST
0.15
0.10
0.05
0.00


Cabecar_6
Piapoco_4
Cofan_4
Inga_13
KichwaOrellana_17
Kamentsa_4
Narihuala_5
Tumbes_7
Luya_11
LoretoMix_7
Sechura_3
Tallan_8
Olmos_4
Huancas_5
Cocama_7
Chulucanas_7
UtcubambaSouth_4
Wayku_10
Chotuna_4
Eten_5
LaJalca_11
Cao_9
Karitiana_16
Quechua_7
Cusco2_7
Xavante_11
Paran_3
Cusco_3
Puno_7
Aymara_5
Guarani_5

EXAMPLES


Paran
Chotuna
Eten
Narihuala
Tallan
Chulucanas
Tumbes
Cusco2
Quechua
Cusco
Aymara
Puno
Chukchi
Sechura
Olmos
Cao
UtcubambaSouth
LaJalca
Huancas
Luya
Yaquis
Pima
Inga
Mixe
Guarani
Cocama
LoretoMix
Kamentsa
Cofan
KichwaOrellana
Wayku
Piapoco
Cabecar
Xavante
Karitiana

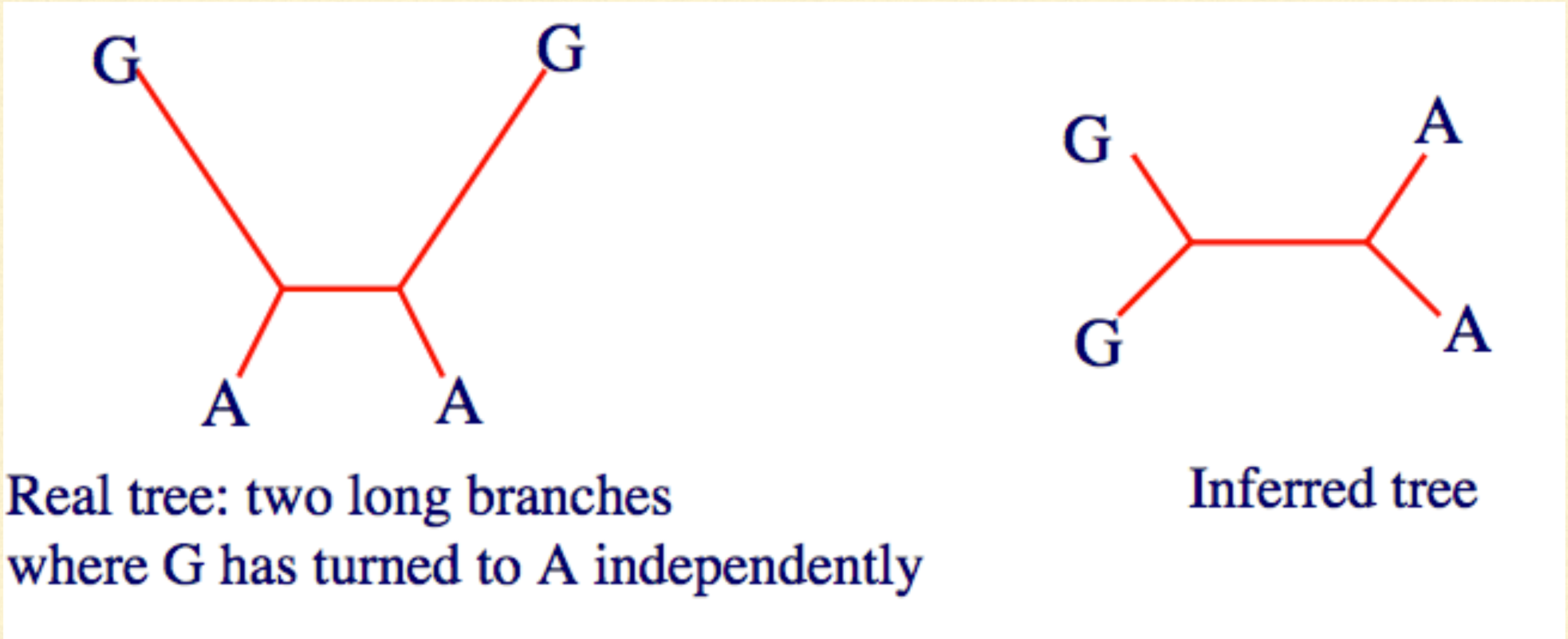
CHARACTER BASED TREES

- 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



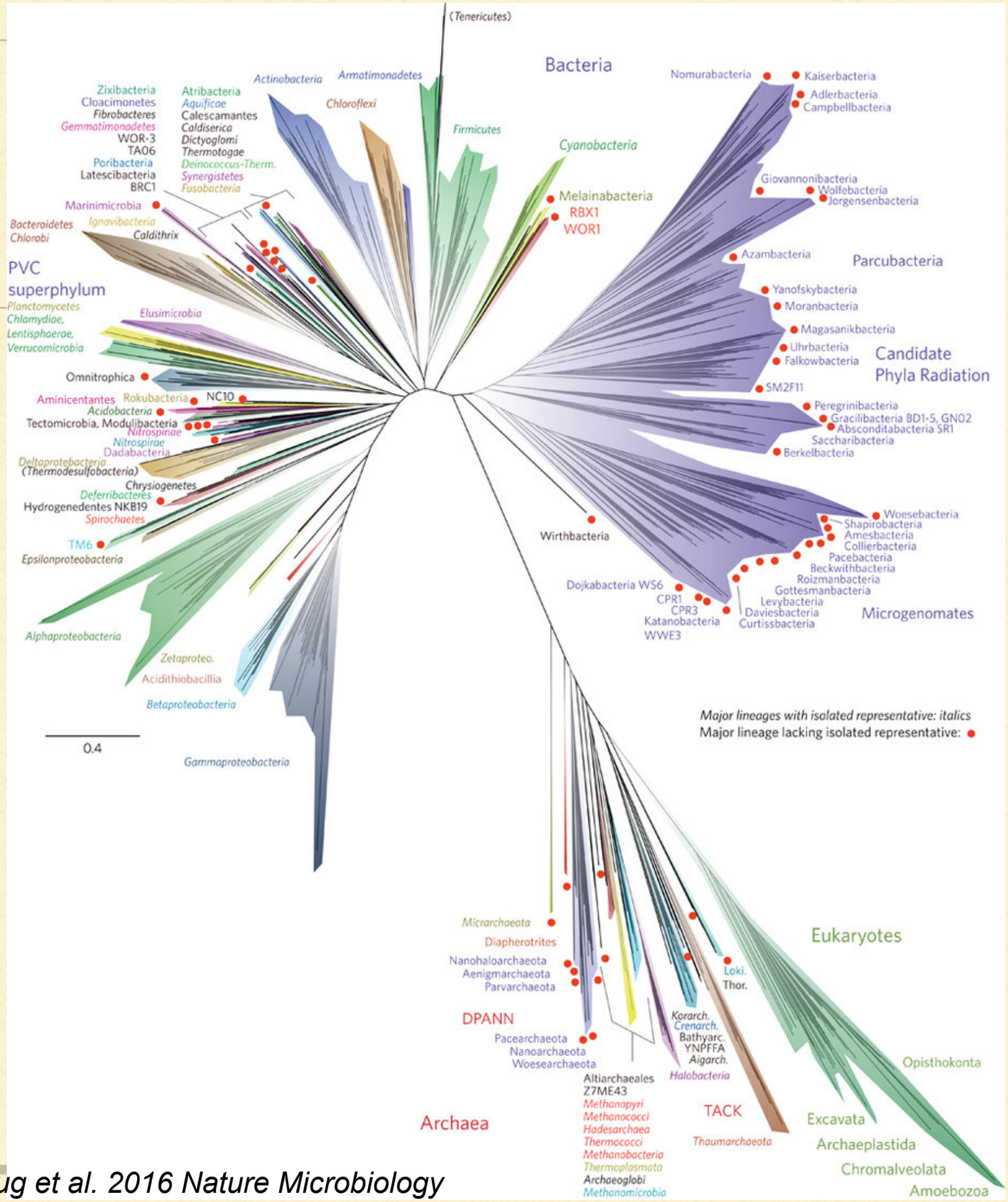
SUMMARY: TREES

- 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:

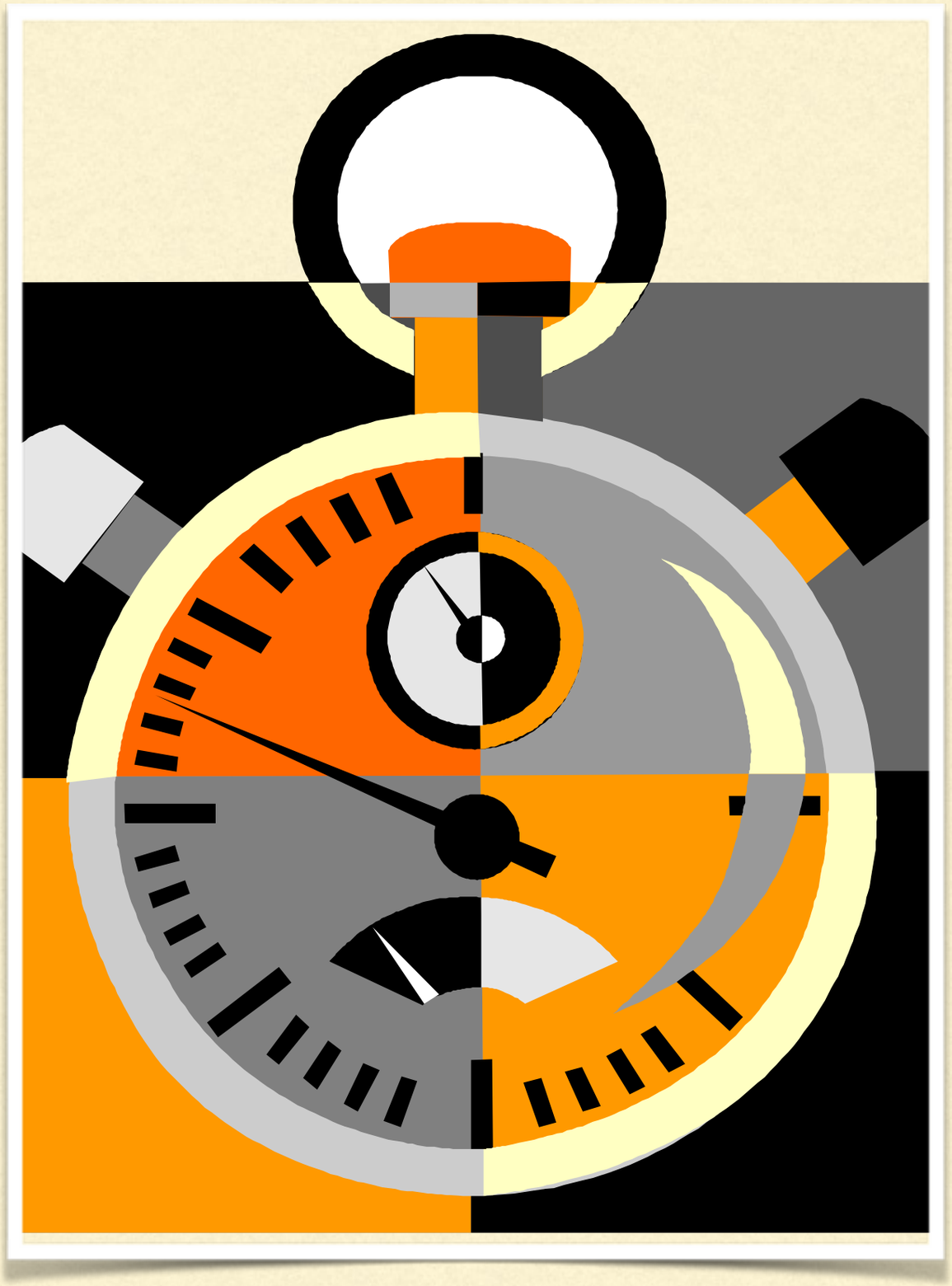
- Many people like **Max Likelihood** based methods:
  - sensitive at large evolutionary distances
- Often a **BEAST** tree is the answer
  - But takes computational time
  - Advantage of including complex models with priori assumptions

ML TREE OF LIFE



Hug et al. 2016 Nature Microbiology

MOLECULAR CLOCK



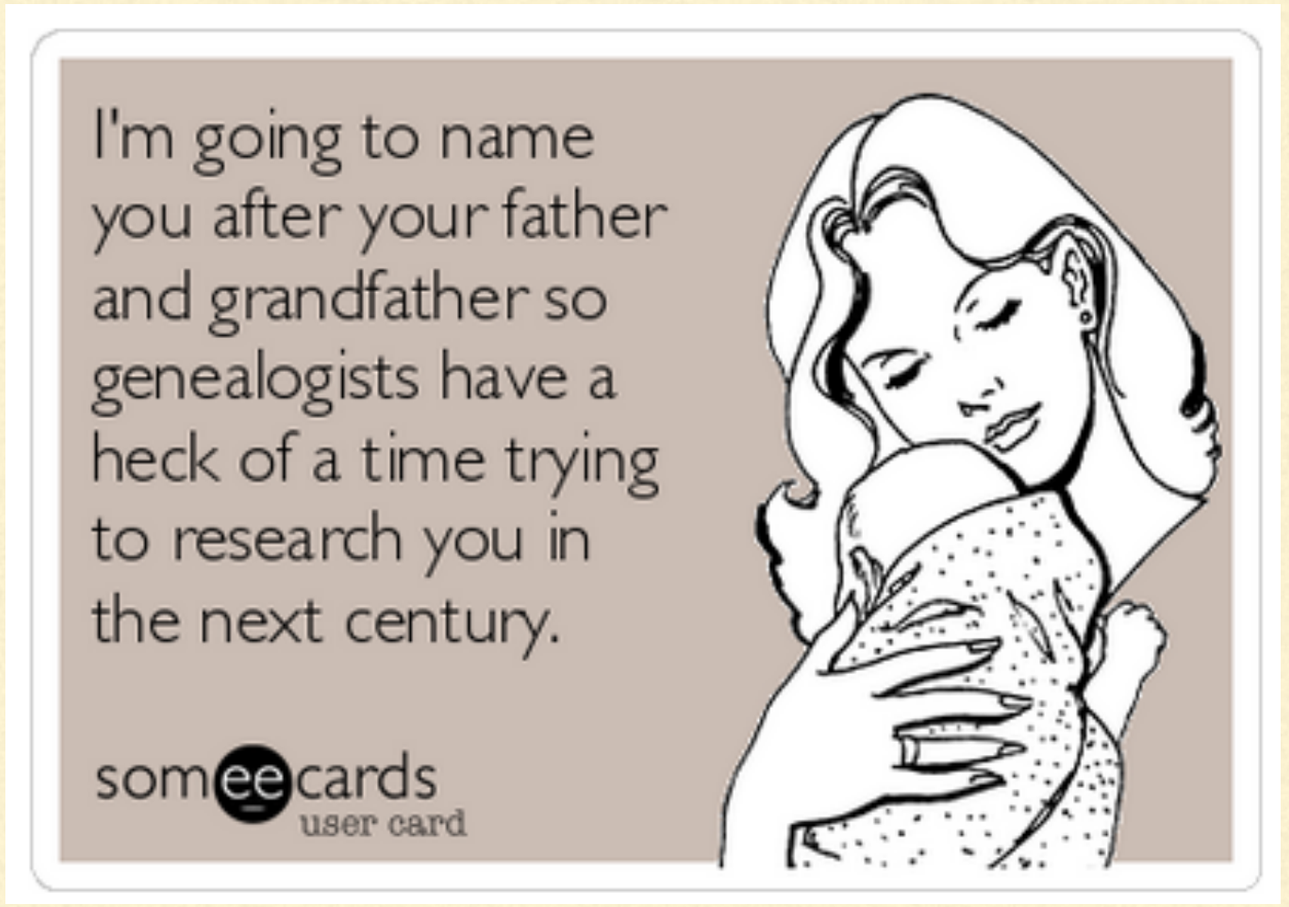
THE MOLECULAR CLOCK HYPOTHESIS

- 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

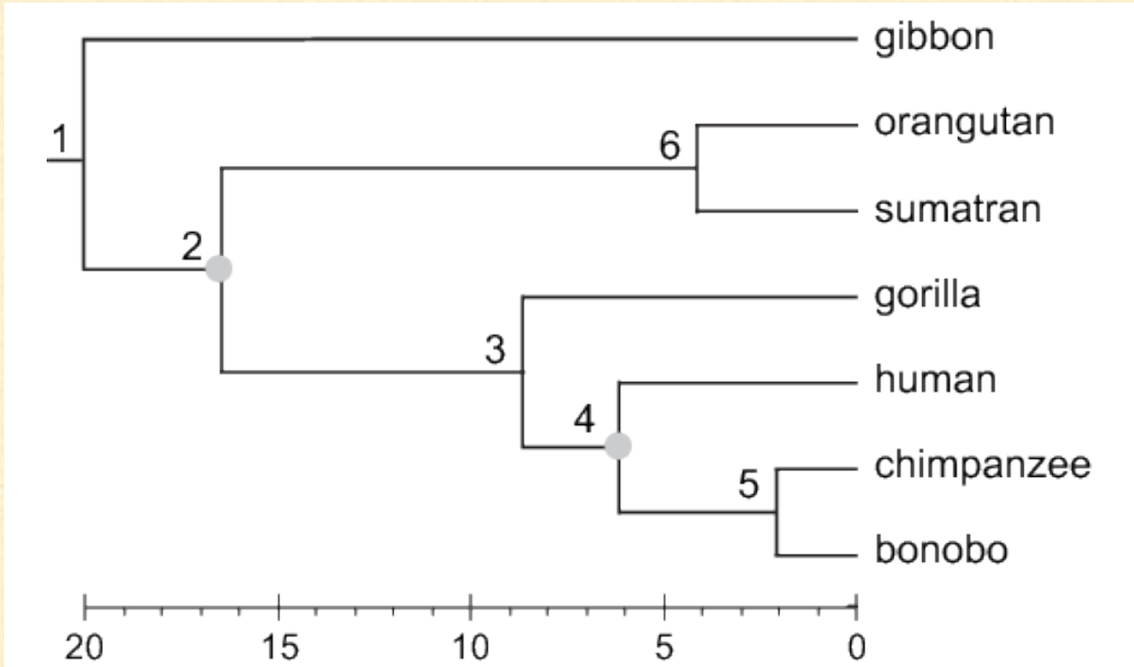
HOW TO CALIBRATE HUMAN GENETICS

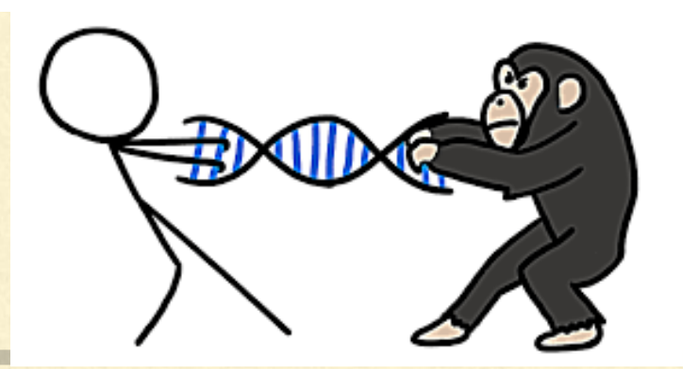
- Deep pedigree data
  - Count the mutations between nth grade cousins



HOW TO CALIBRATE HUMAN GENETICS

- ARCHAEOLOGY
  - Species divergence - E.g Human-chimp split
  - Historical events E.g Colonization of the pacific





HOW TO CALIBRATE HUMAN GENETICS

- Tip fossils
- aDNA from dated fossils

GENETIC MARKERS

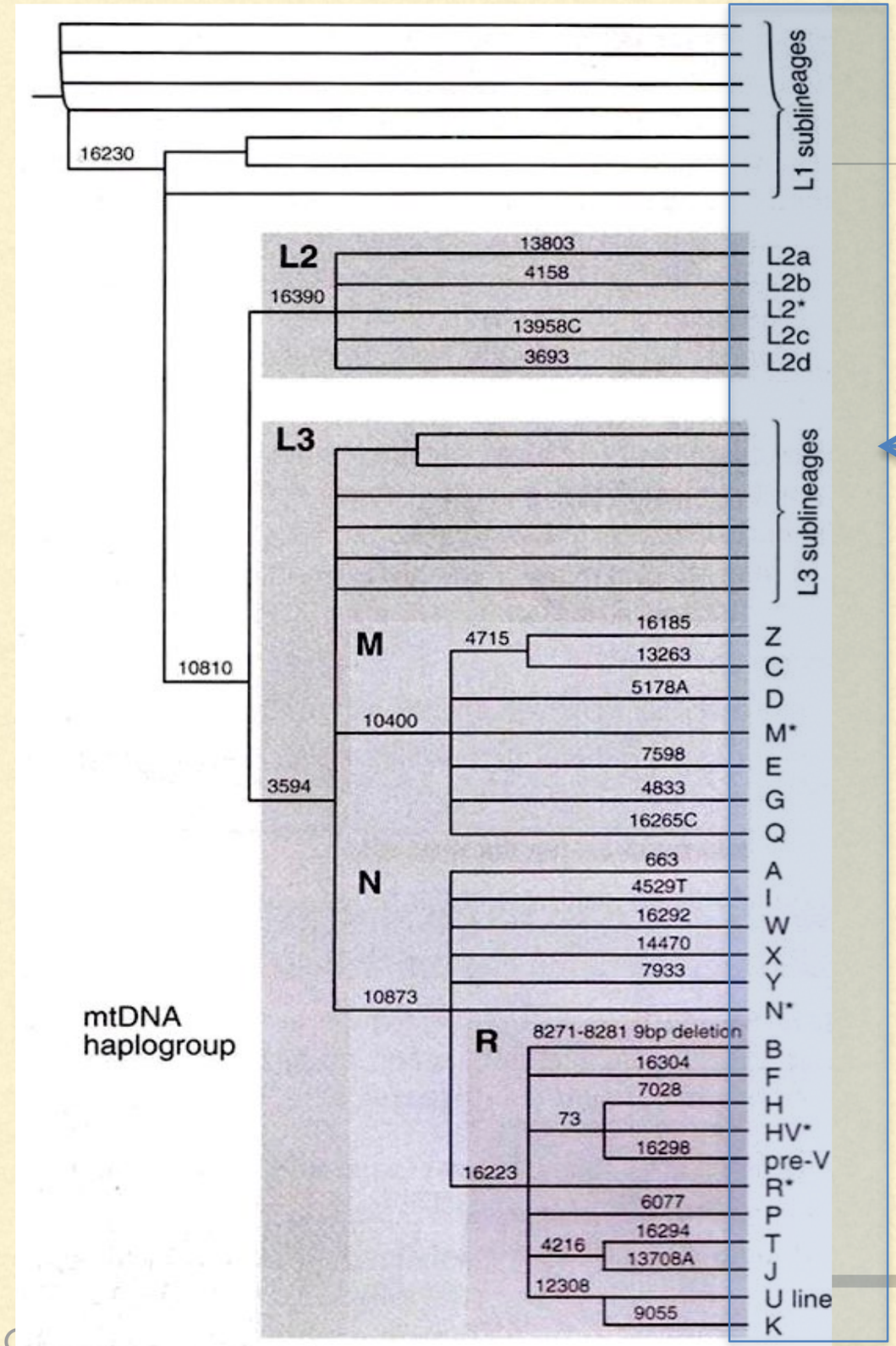
Uniparental markers:

- No recombination (HAPLOID)
- Mutations accumulate with time alone
- Shared mutations indicate shared ancestry (if mutations are rare)

LIMITS

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

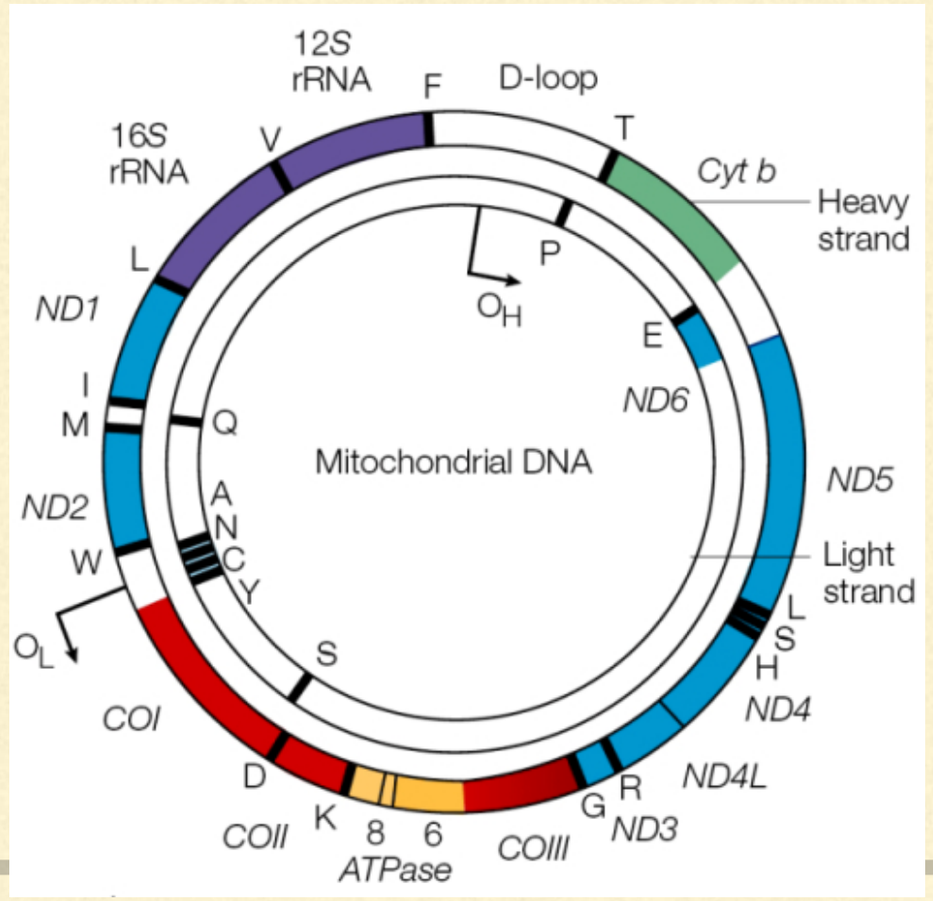
HAPLOGROUP PHYLOGENY



Phylogeny

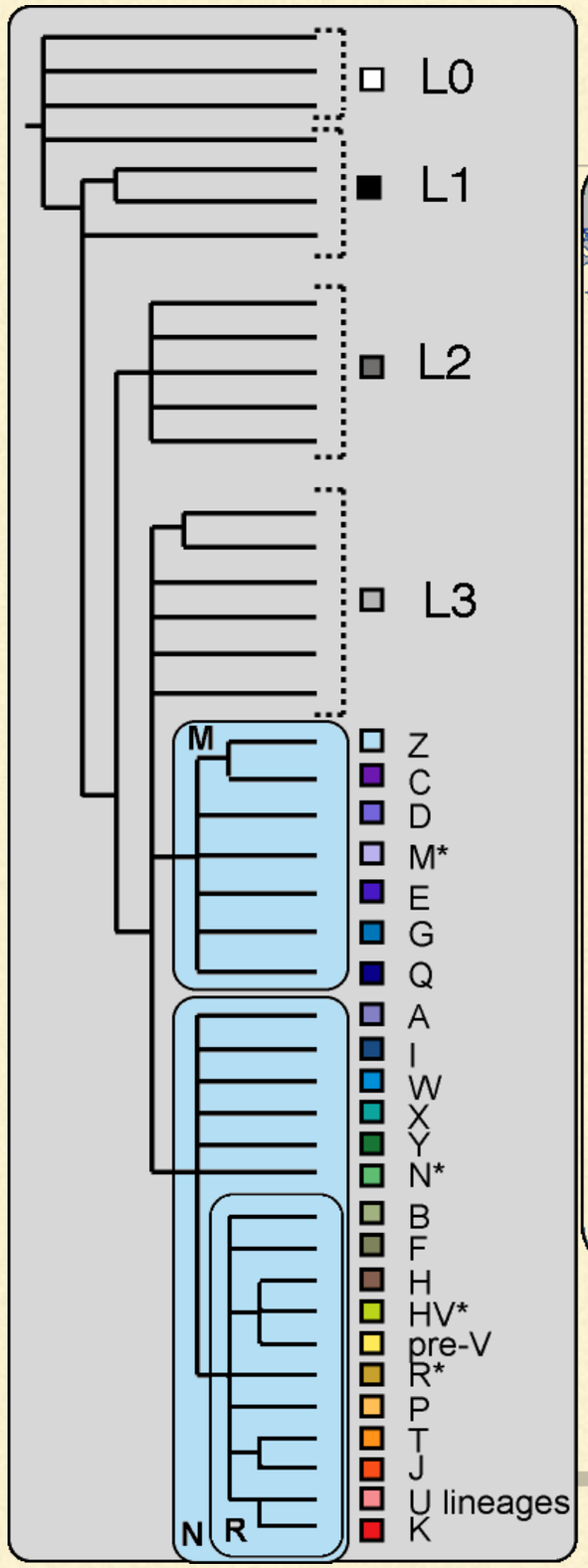
Tree structure representing evolutionary relationships between clades

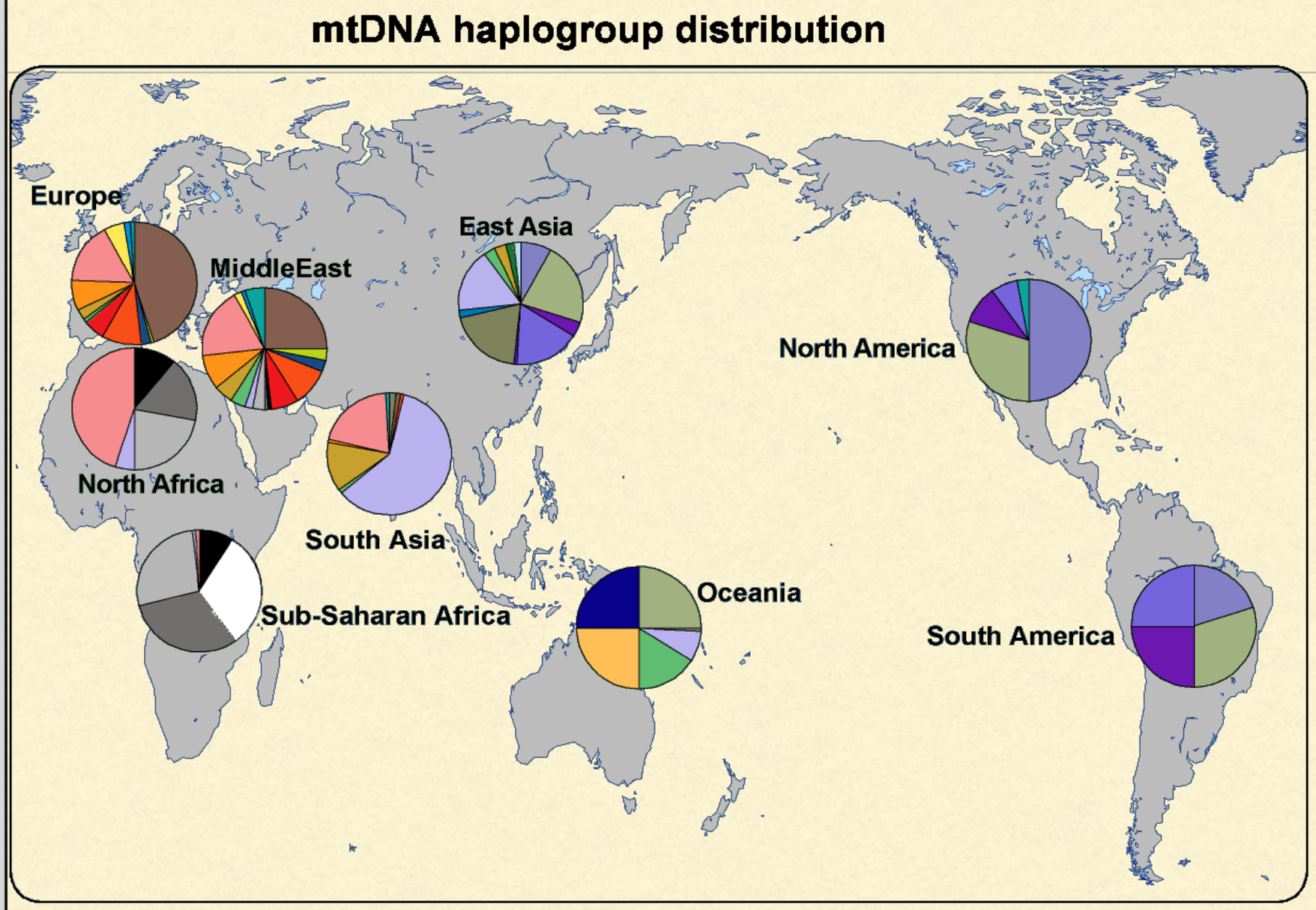
Haplogroups indicated with a capital letter



From nature.com

WORLD MTDNA PHYLOGEOGRAPHY

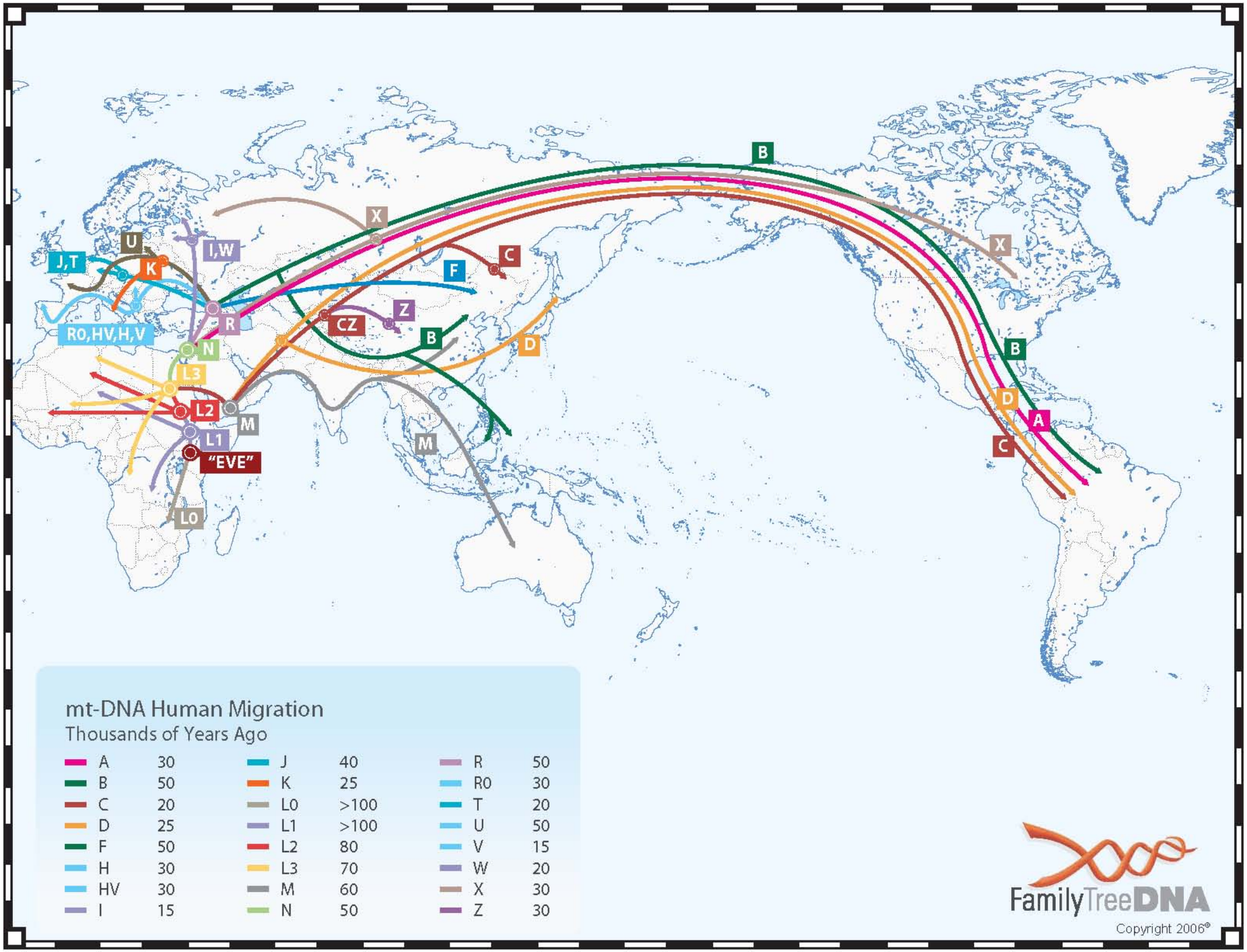




It combines temporal and evolutionary dimension of phylogeny with geographic distribution of haplogroups

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

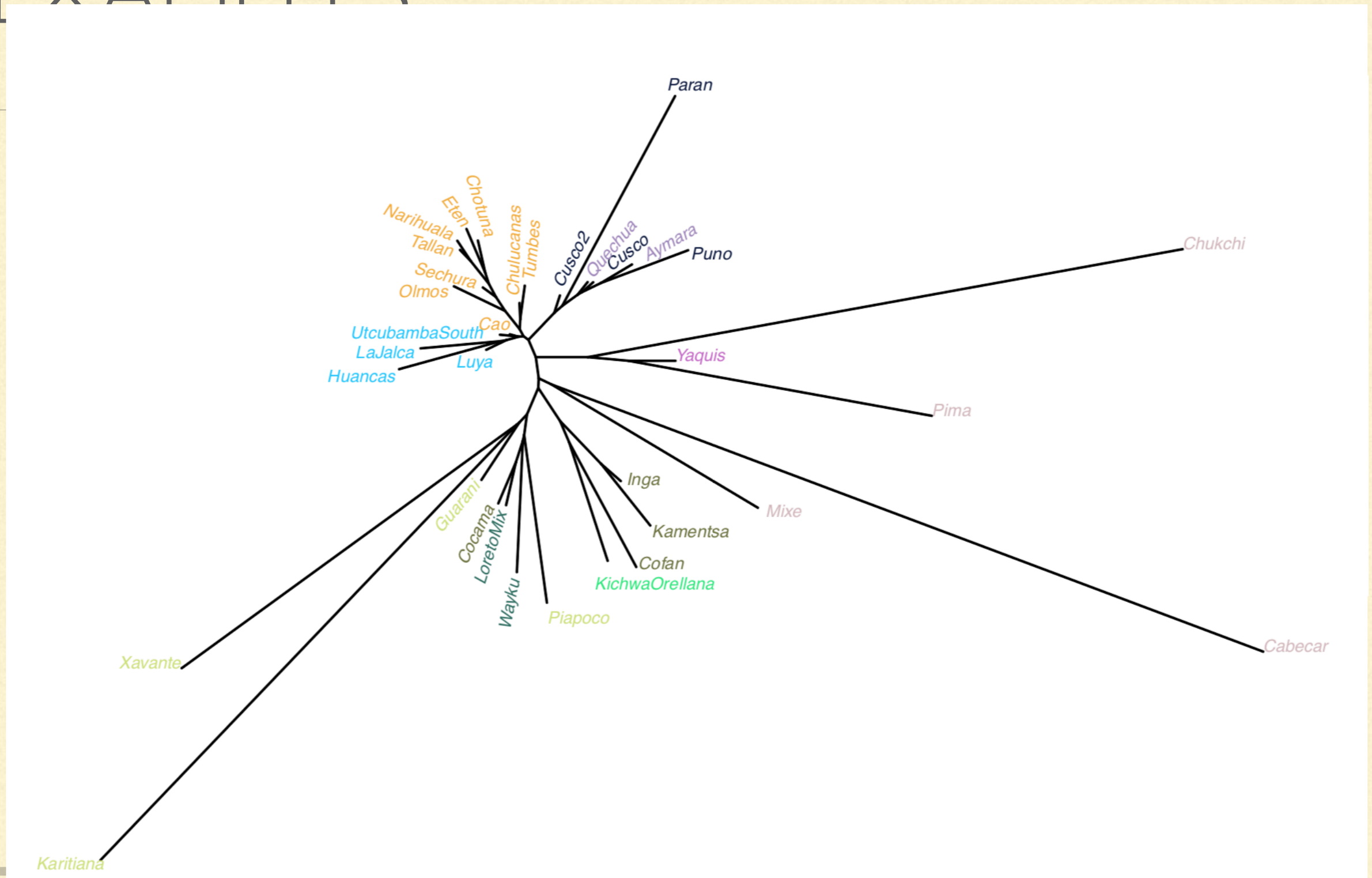
WORLD MTDNA PHYLOGEOGRAPHY



COMO RECONSTRUIR LA HISTORIA FILOGENETICA DE POBLACIONES?

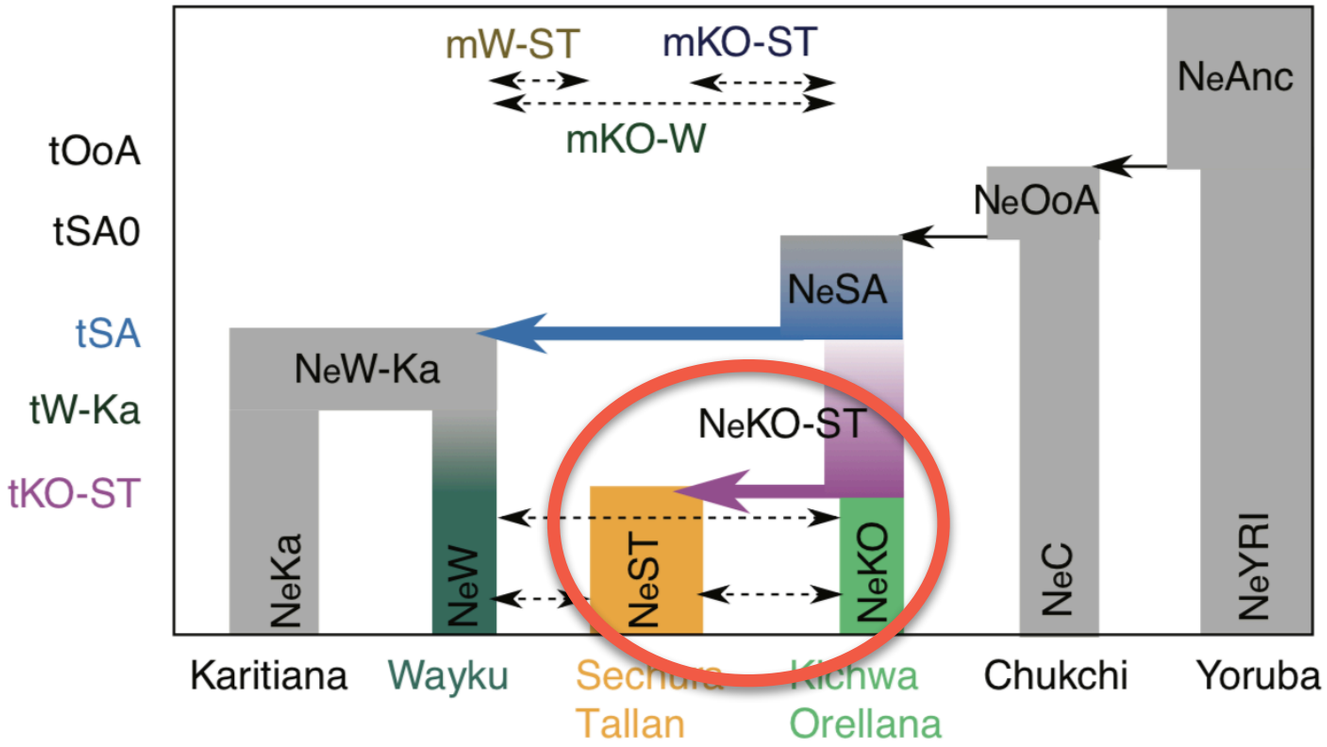
- Distancia genetica entre poblaciones (FST)
- Simulaciones de escenarios demográficos
- Admixture graphs, TreeMix

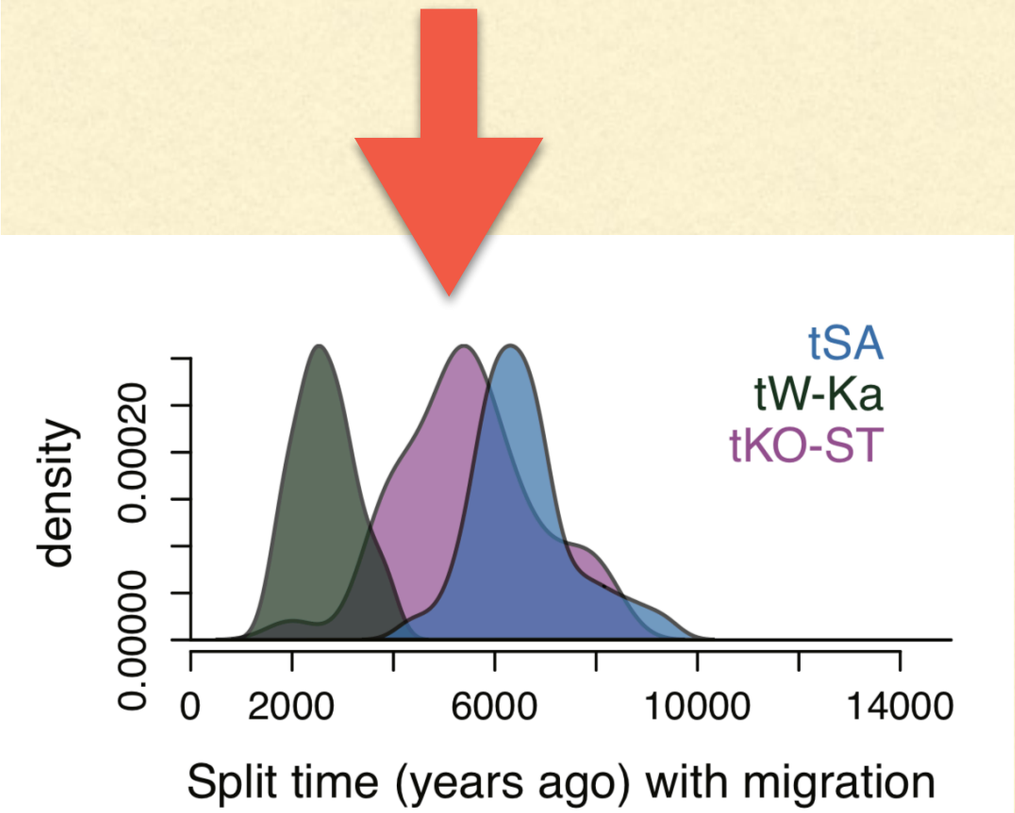
EXAMPLES



Barbieri et al. 2019 Mol Biol Evol

ANCESTRIES IN SOUTH AMERICA: DIVERGENCE TIMES

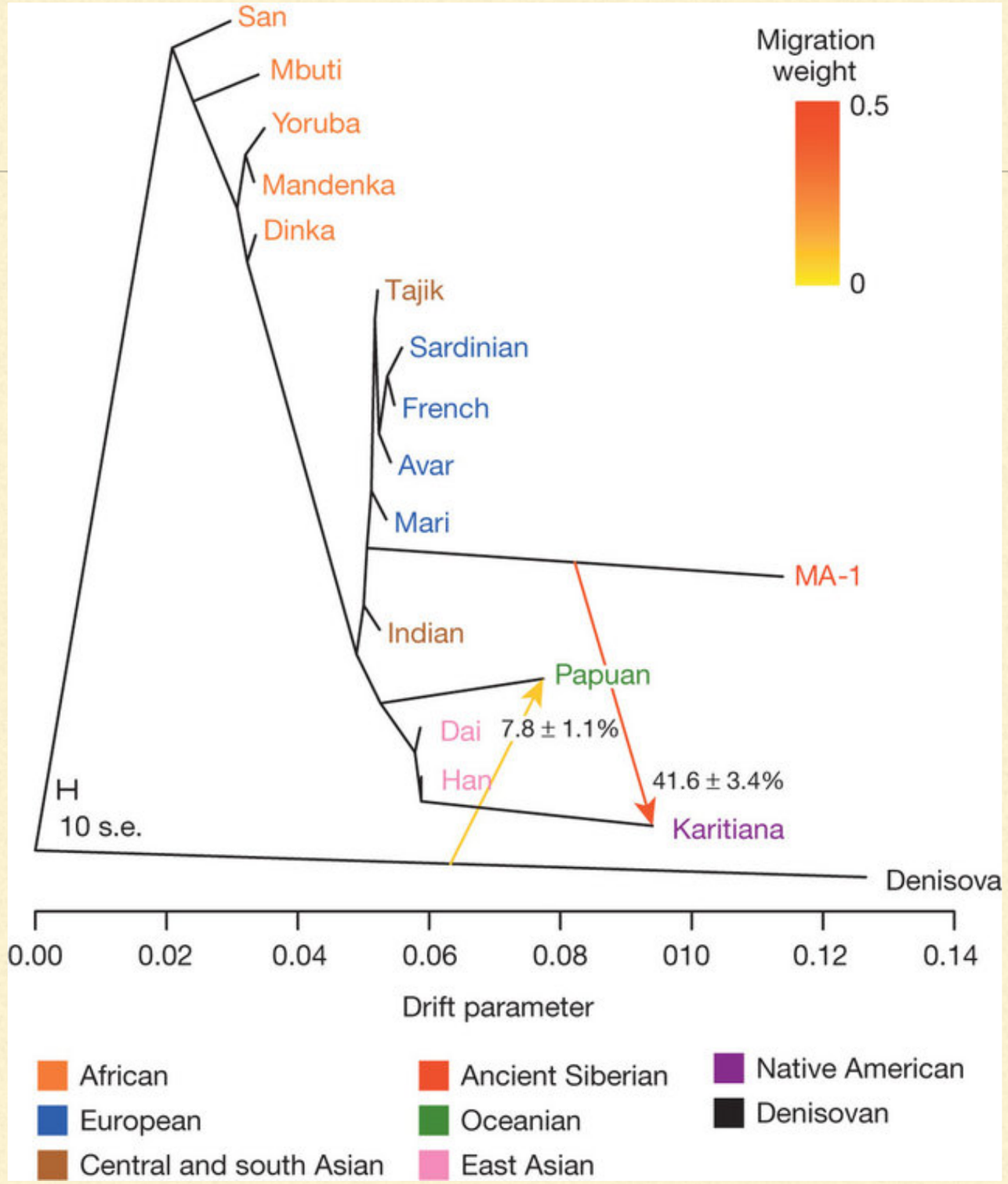




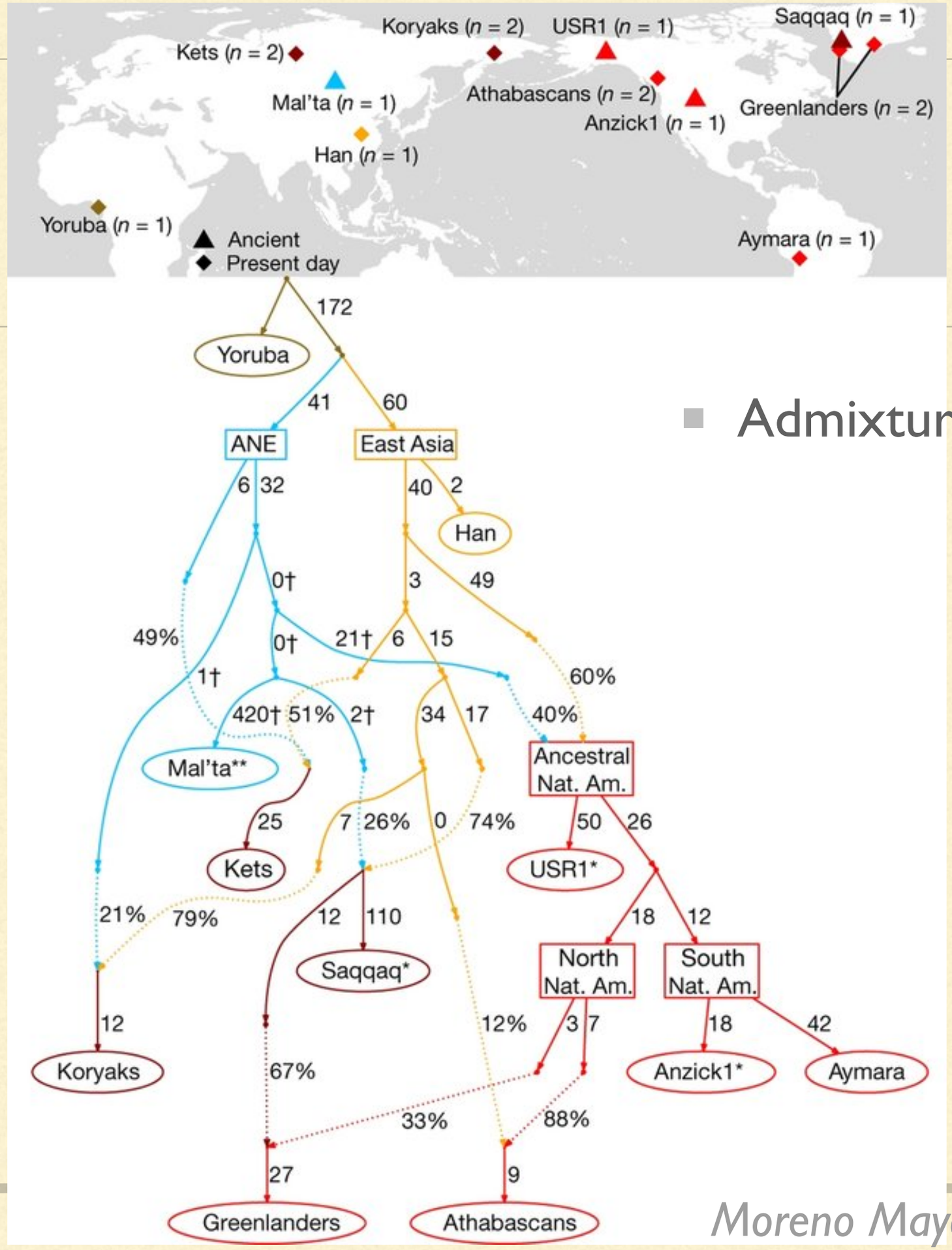
Simulating demographic scenarios, estimating parameters with ABC framework

Barbieri et al. 2019 Mol Biol Evol

- TreeMix



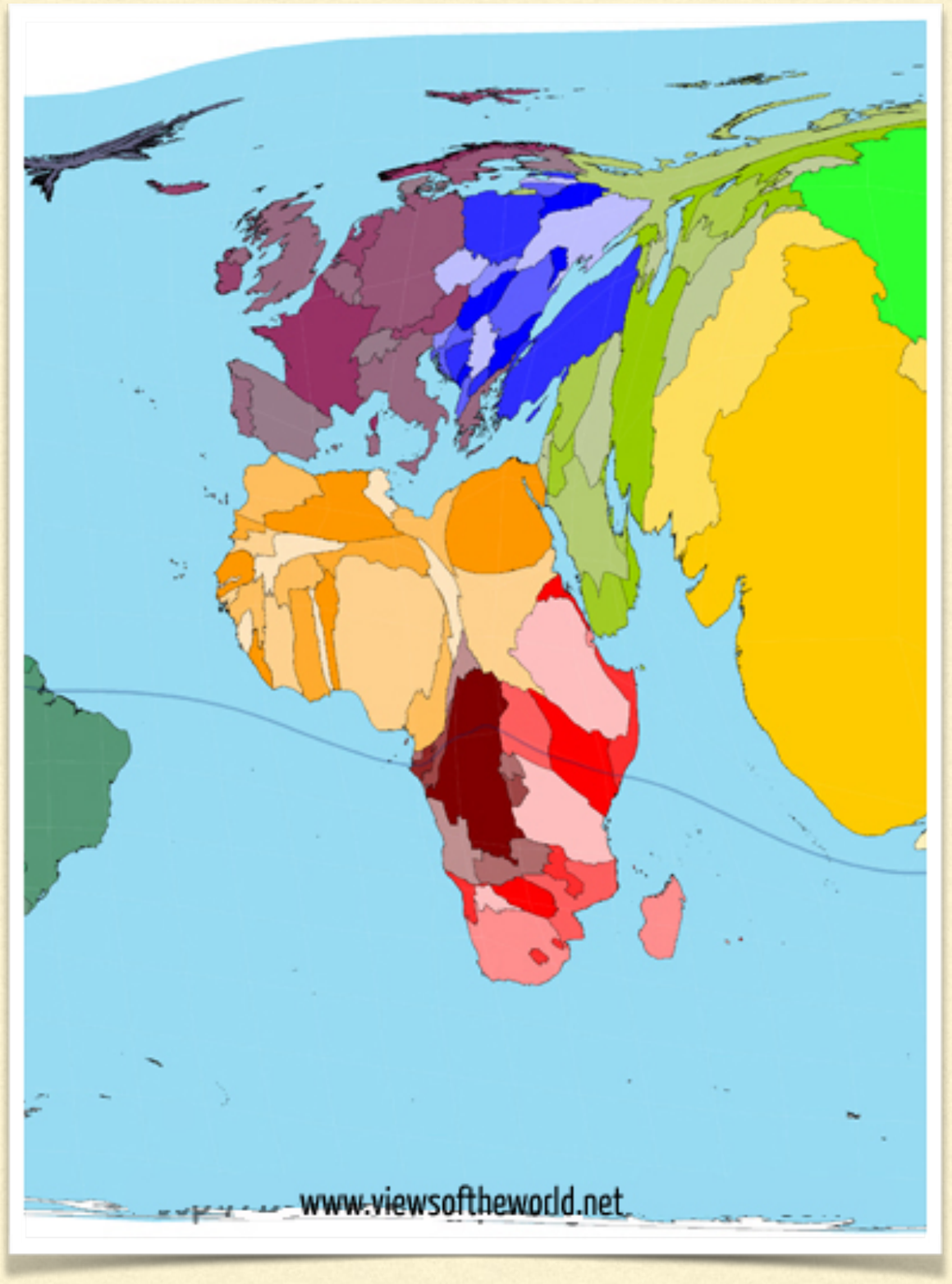
Raghavan et al. Nature, 2014



- Admixture graph

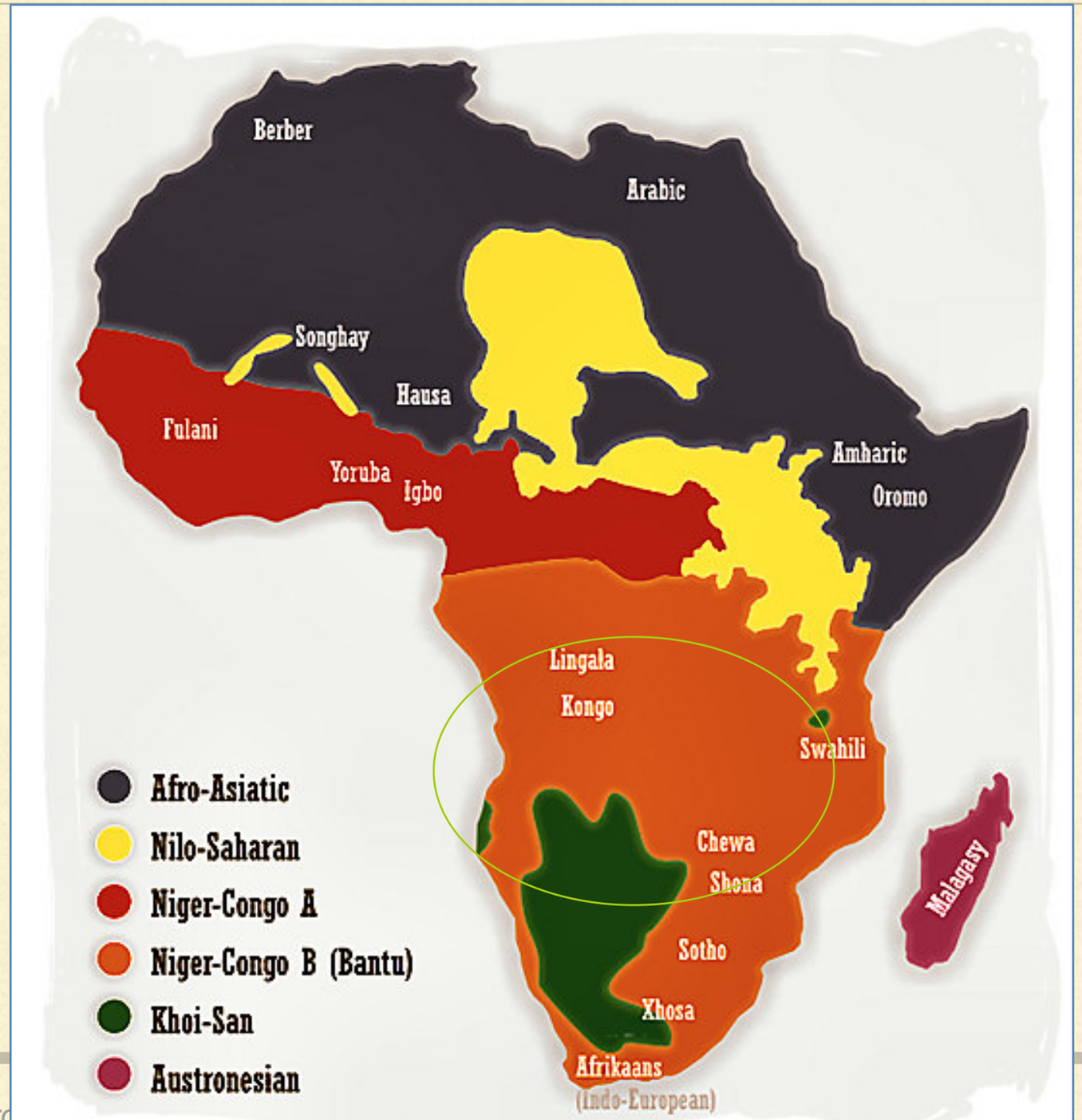
Moreno Mayar et al. Nature, 2018

HUMAN POPULATION HISTORY AND RELEVANCE FOR GENOMIC STUDIES



Chiara Barbieri

LINGUISTIC DIVERSITY IN AFRICA



- 4 linguistic stocks
  - **Bantu** is the most diffused language family

BANTU-SPEAKING PEOPLE

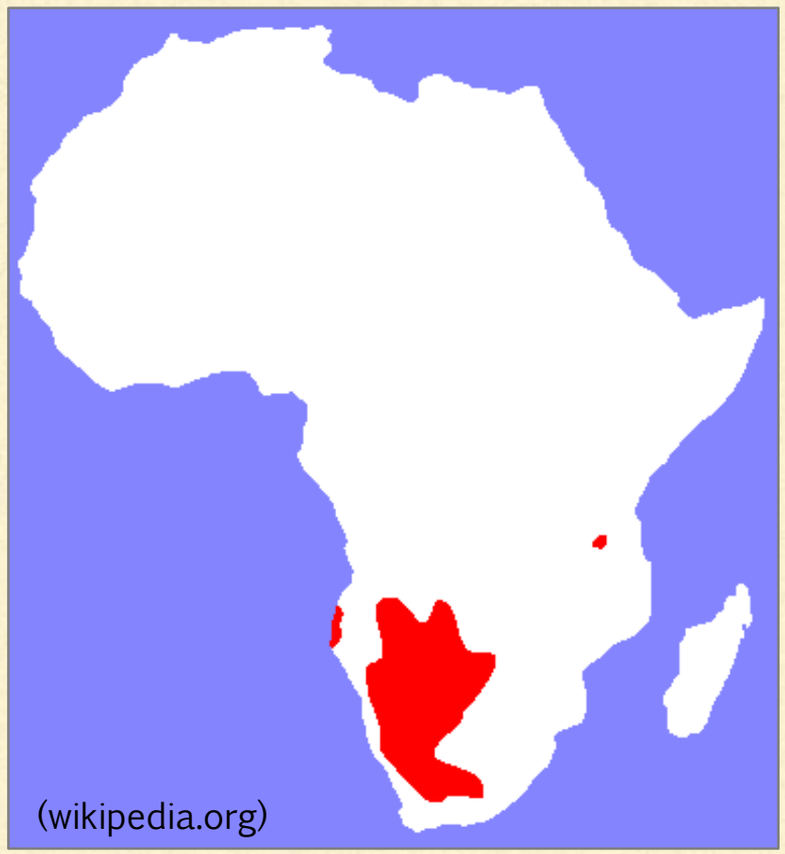
- Language spread with human migration
  - Starting ~5.000 kya from Cameroon
  - Diffusion of agriculture

From umanitoba.ca

KHOISAN PEOPLE

(Credits: Kure B)

- Non-Bantu groups
- click consonant as prominent linguistic feature
- pastoralist Khoi and hunter-gatherer San



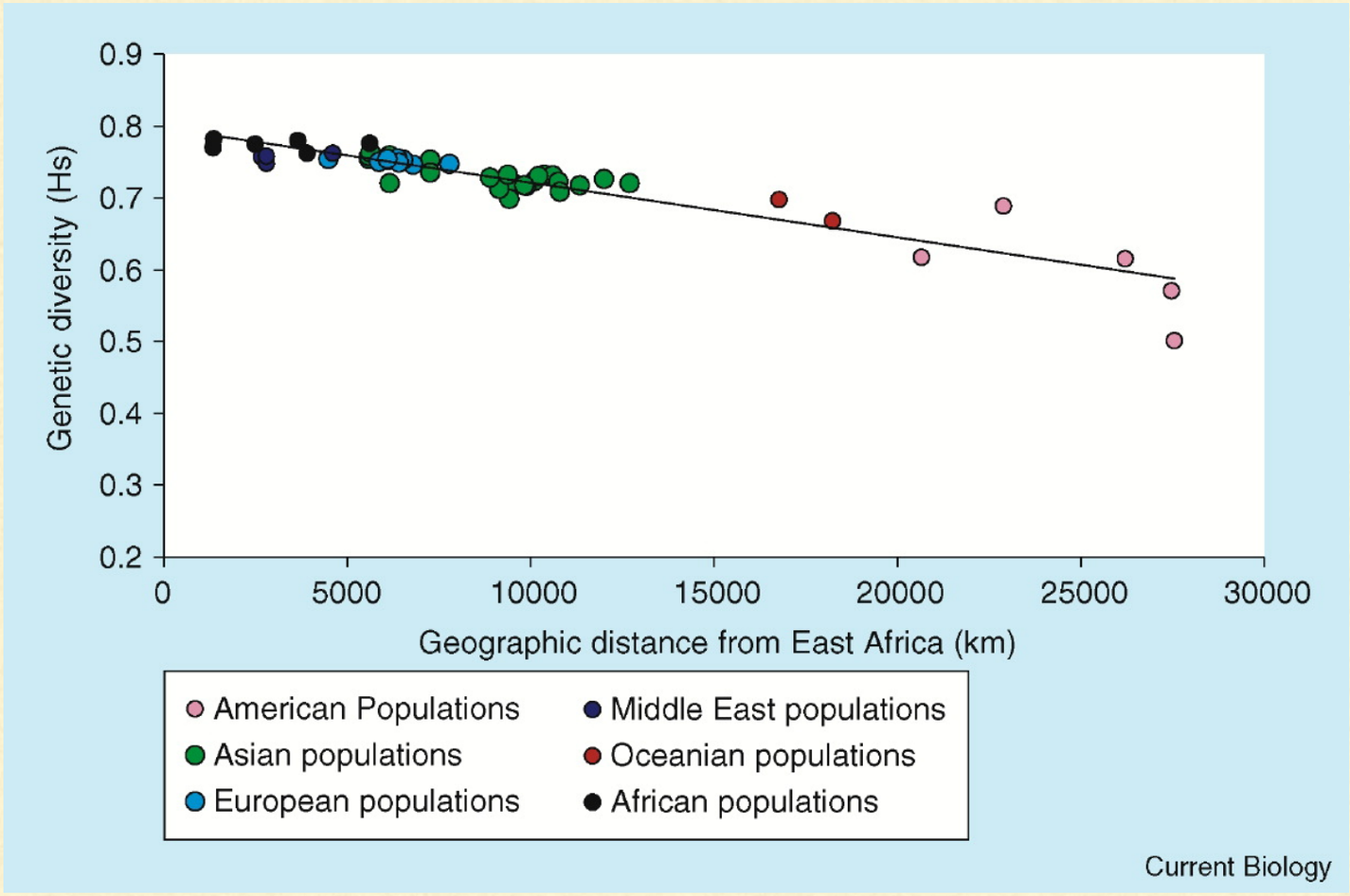
(wikipedia.org)

KHOISAN

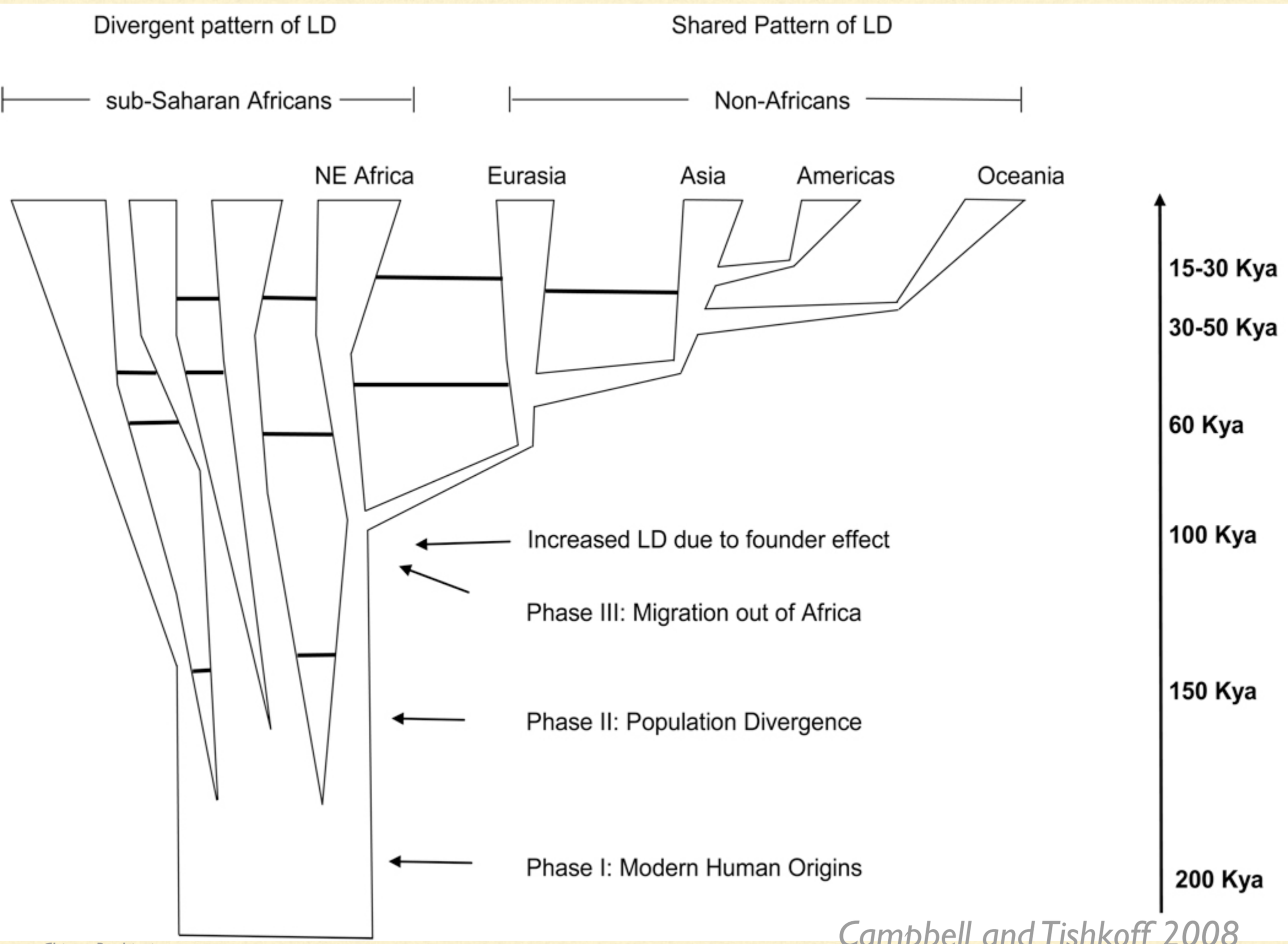
- Three major lineages:
  - **Tuu**
  - **Kx’aa**
  - **Khoe-Kwadi**
- Shared linguistic features (clicks) might be result of contact

HUMAN ORIGINS IN AFRICA

- Africa harbors the highest genetic diversity
- Out of Africa: Major bottleneck in the migration which colonised the rest of the continents ~70 cya



Prugnolle et al. Current Biology 2005

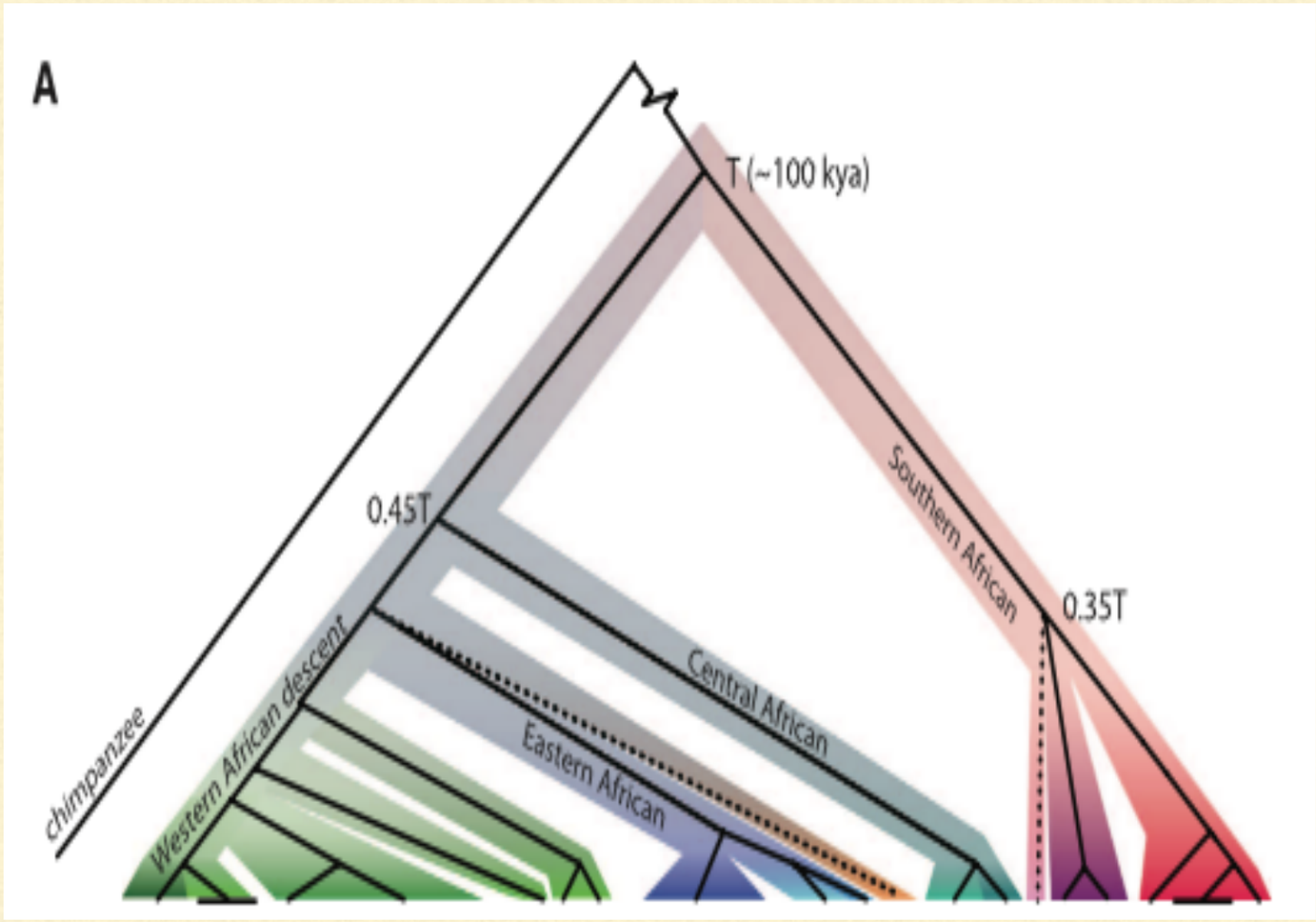


Campbell and Tishkoff 2008

WHERE IN AFRICA?

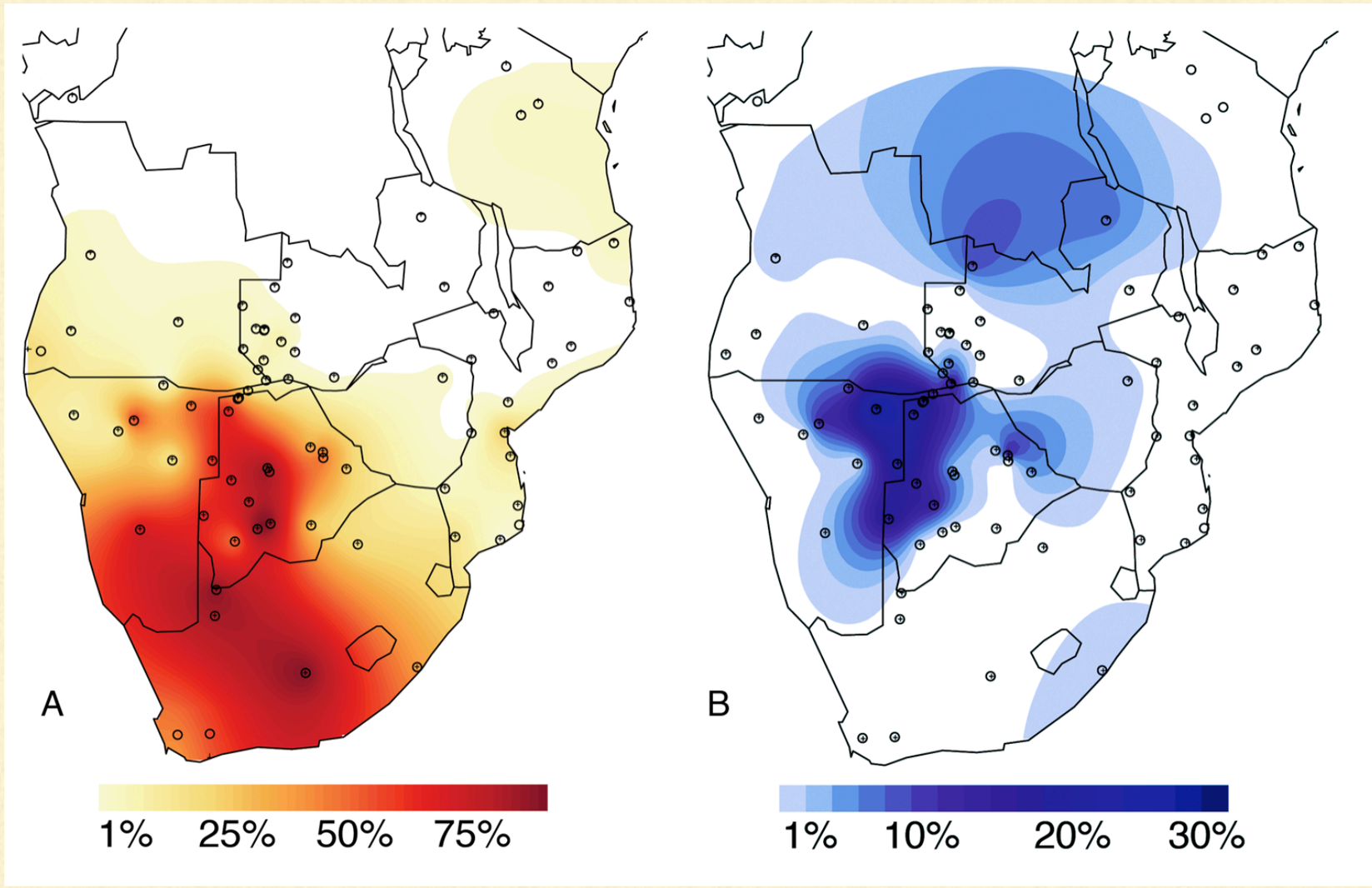
- San hunter gatherer of Southern Africa harbour the most divergent genetic lineages and early structure

Photo: C. Barbieri



Schlebusch et al. Science 2012

EARLY DIVERGING MTDNA LINEAGES

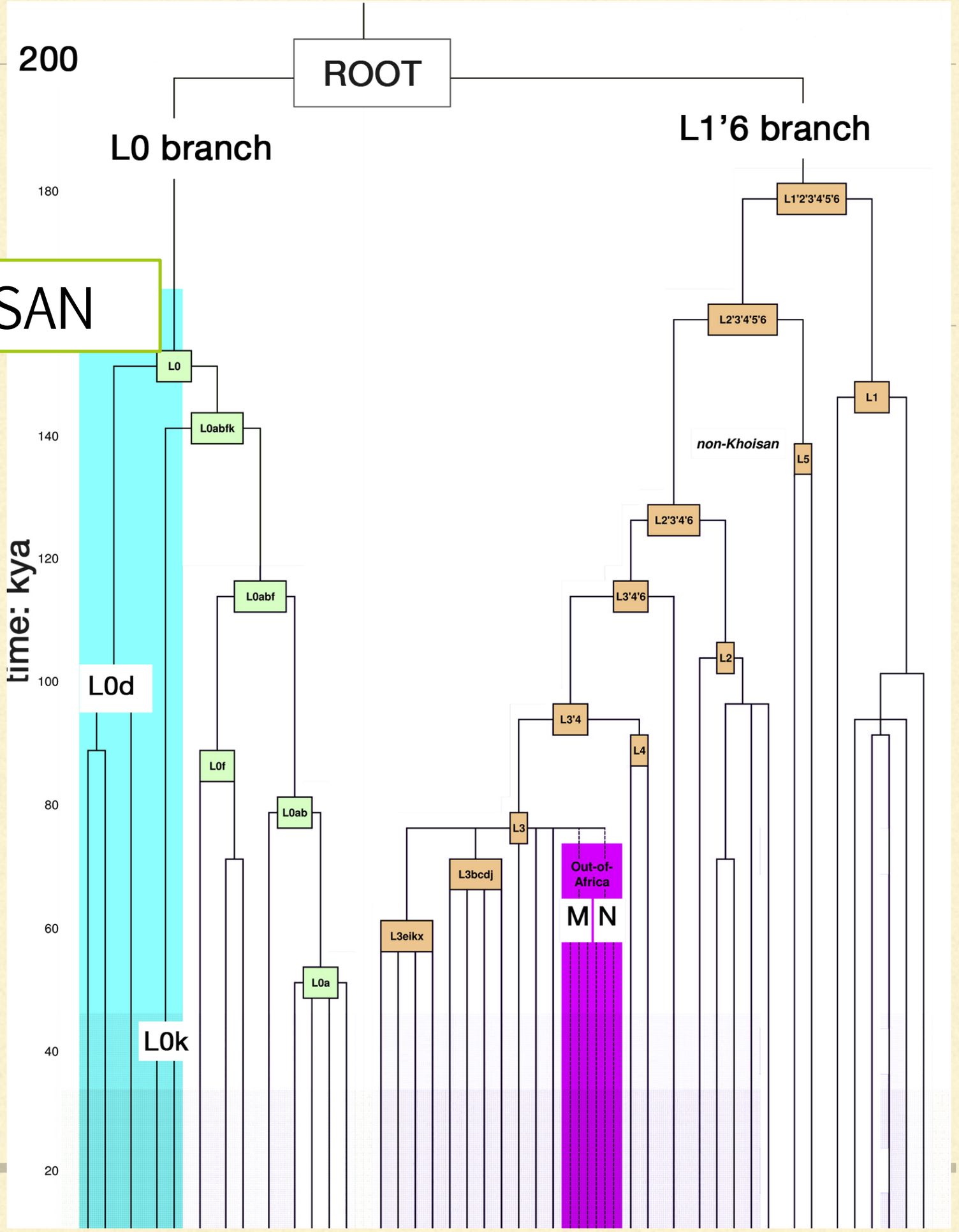


The first branching mtDNA lineages are found in southern Africa

Barbieri et al. Am J Hum Genet 2013

mtDNA tree

KHOISAN



Modified from

Behar et al. 2008, Am J Hum Genet

EARLY DIVERGING Y CHROMOSOME

But the earliest branch of the Y chromosome phylogeny is found in Cameroon

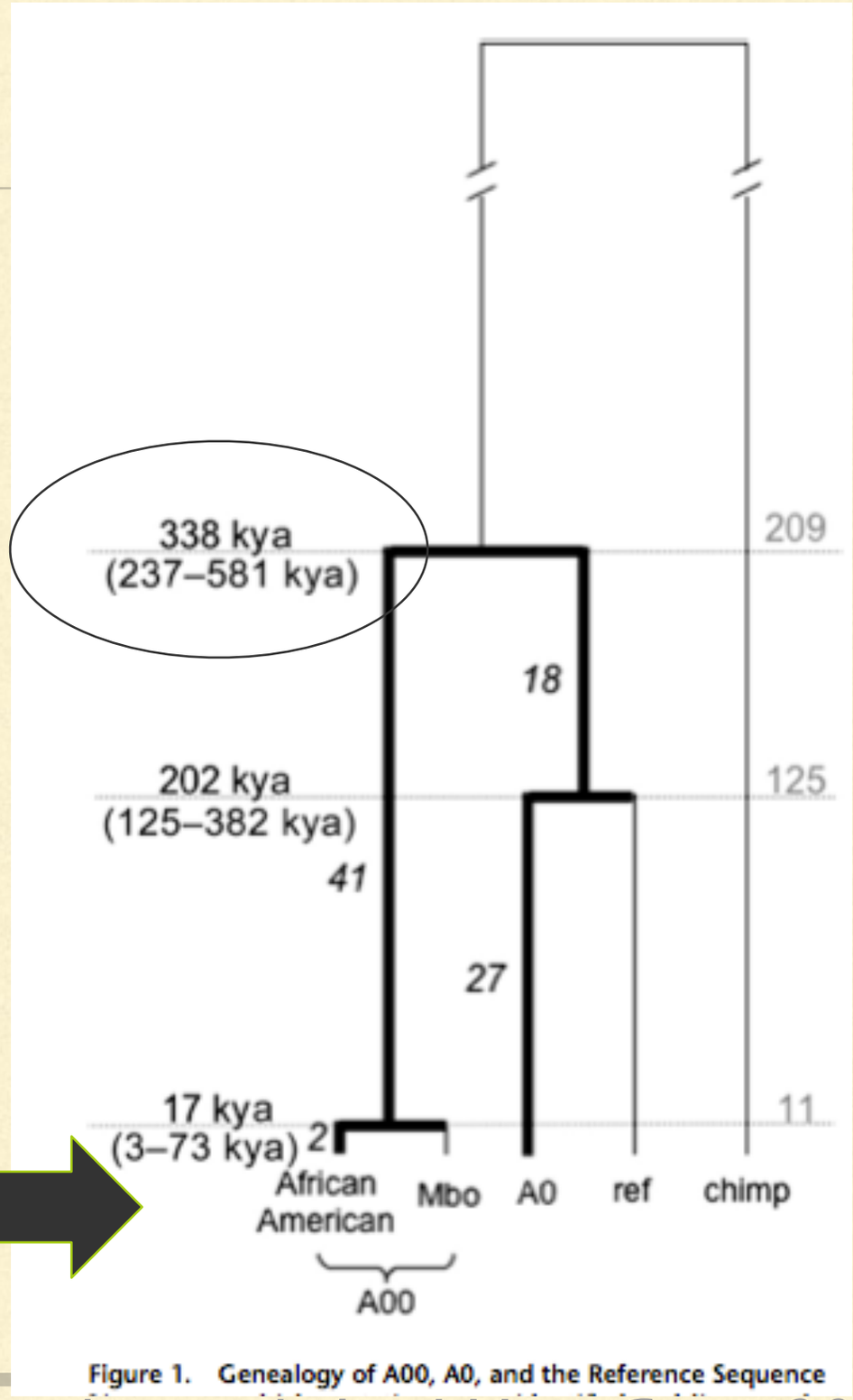
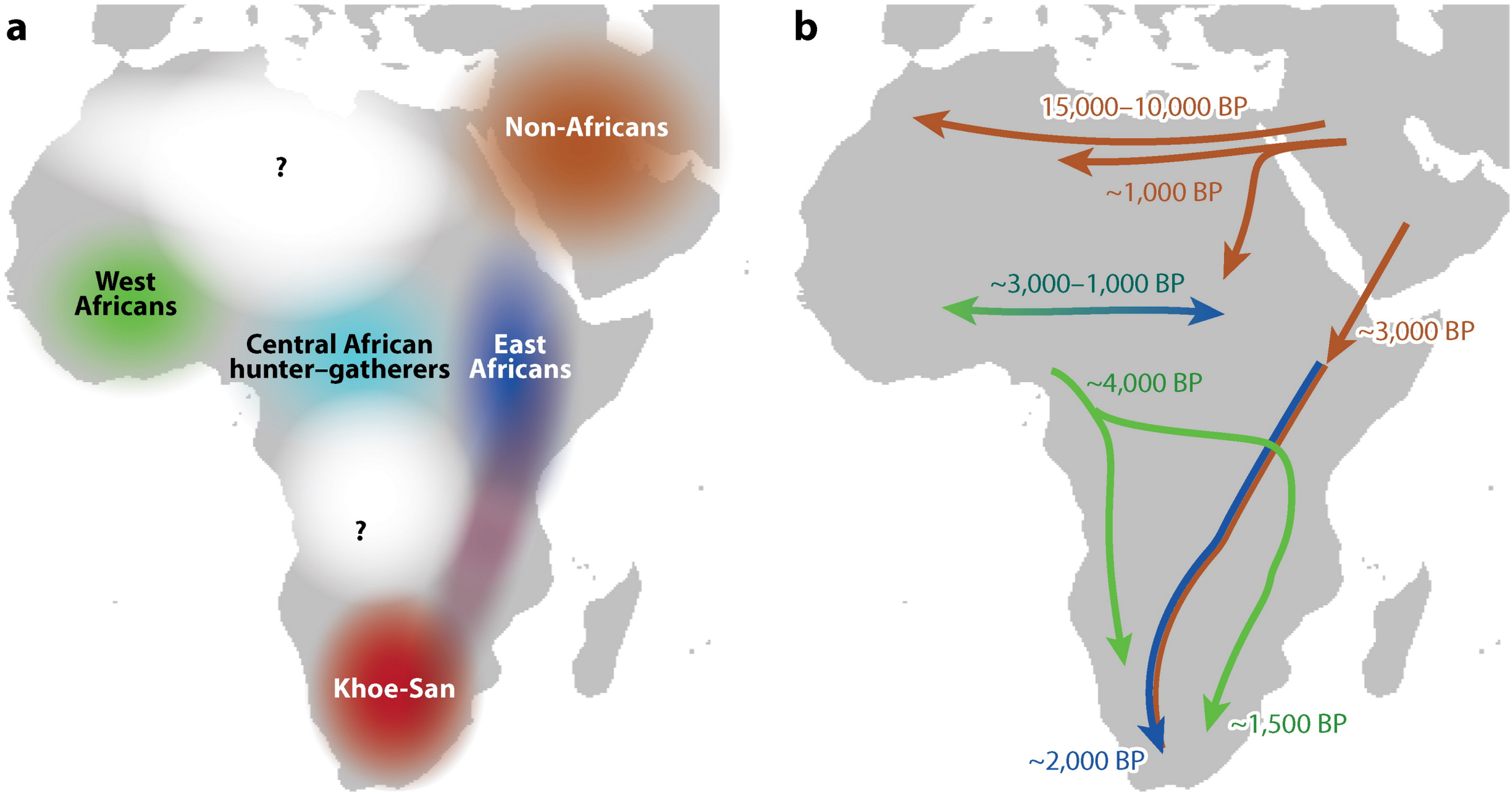


Figure 1. Genealogy of A00, A0, and the Reference Sequence

Mendez et al. Am J Hum Genet 2013



Schlebusch CM, Jakobsson M. 2018.
Annu. Rev. Genom. Hum. Genet. 19:405–28

PHYLOGENY OF THE LACTASE ALLELE IN SOUTHERN AFRICA

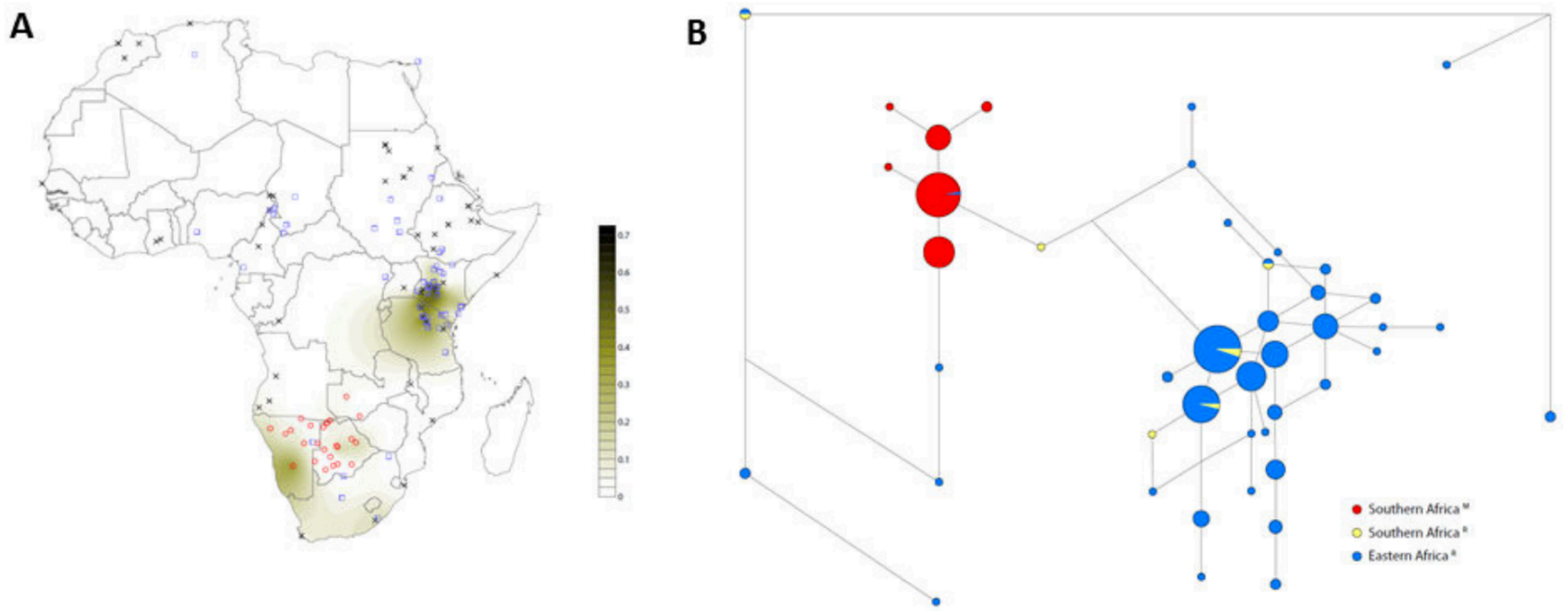


Fig. 1. Analyses of the C-14010 LP variant and associated STR haplotypes in Eastern and Southern African populations. **A**: Surfer map of the C-14010 allele frequency. Red circles denote sampling locations by Macholdt et al.; blue squares denote sampling locations by Ranciaro et al.; black crosses denote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

CULTURE: TRANSMISSION WITH MODIFICATION

- Human cultural traits (behaviors, ideas, and technologies that can be learned from other individuals) can exhibit complex patterns of transmission and evolution
- Cultural and genetic evolution can interact with one another and influence both transmission and selection

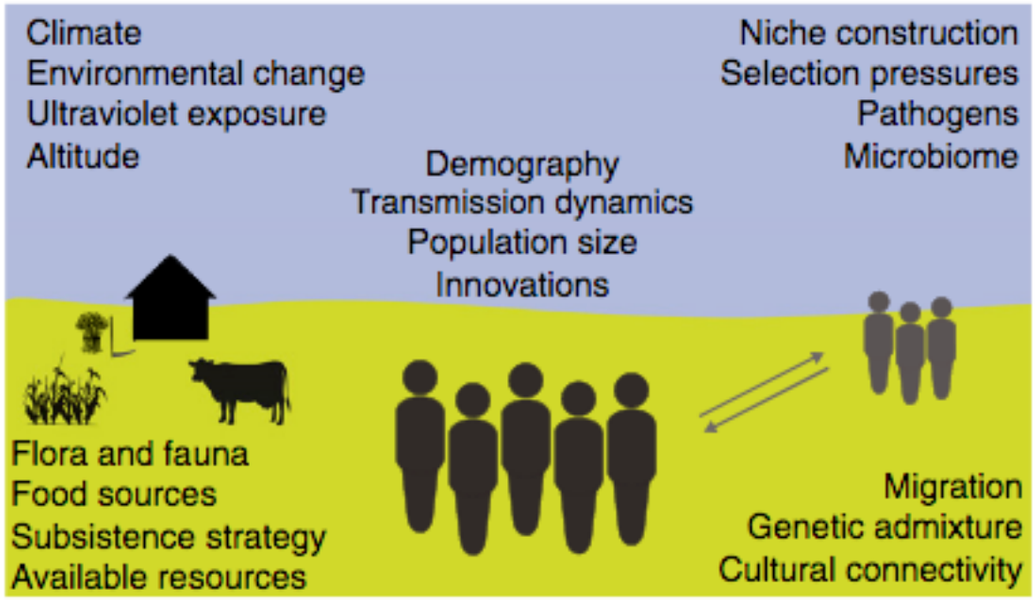


Fig. 2. Cultural, genetic, and environmental factors influencing evolution.

Creanza, Kolodny and Feldman - PNAS 2017

BIOLOGY + CULTURE = MULTIDISCIPLINARY STUDIES

- Evolving under similar demographic processes
- Addressing questions about human past and present diversity, ecology and evolution
- The whole study is greater than the sum of its discipline parts

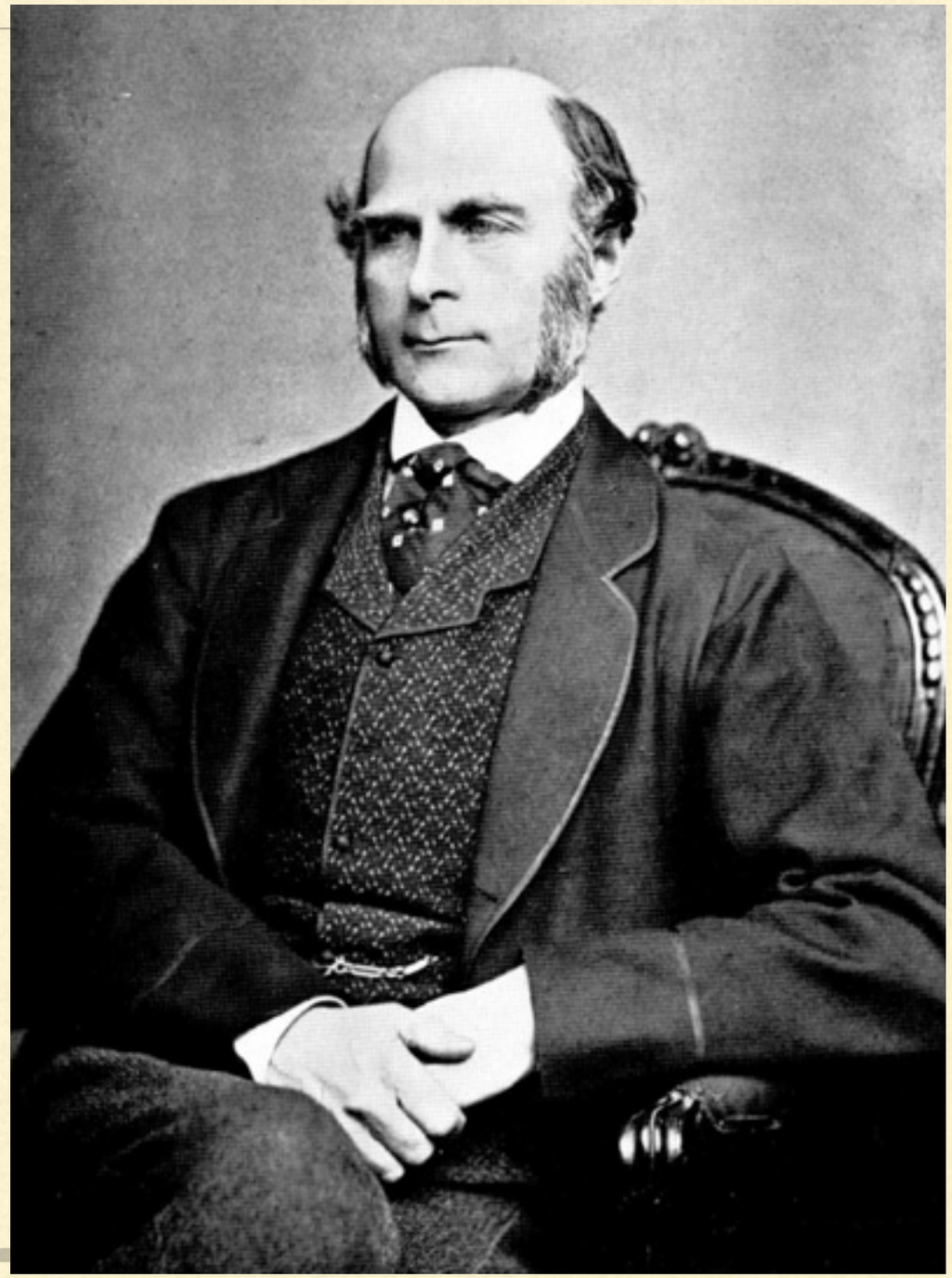
Picture: the Open University

Image: Minna Sundberg

EVOLUTIONARY PARALLELS

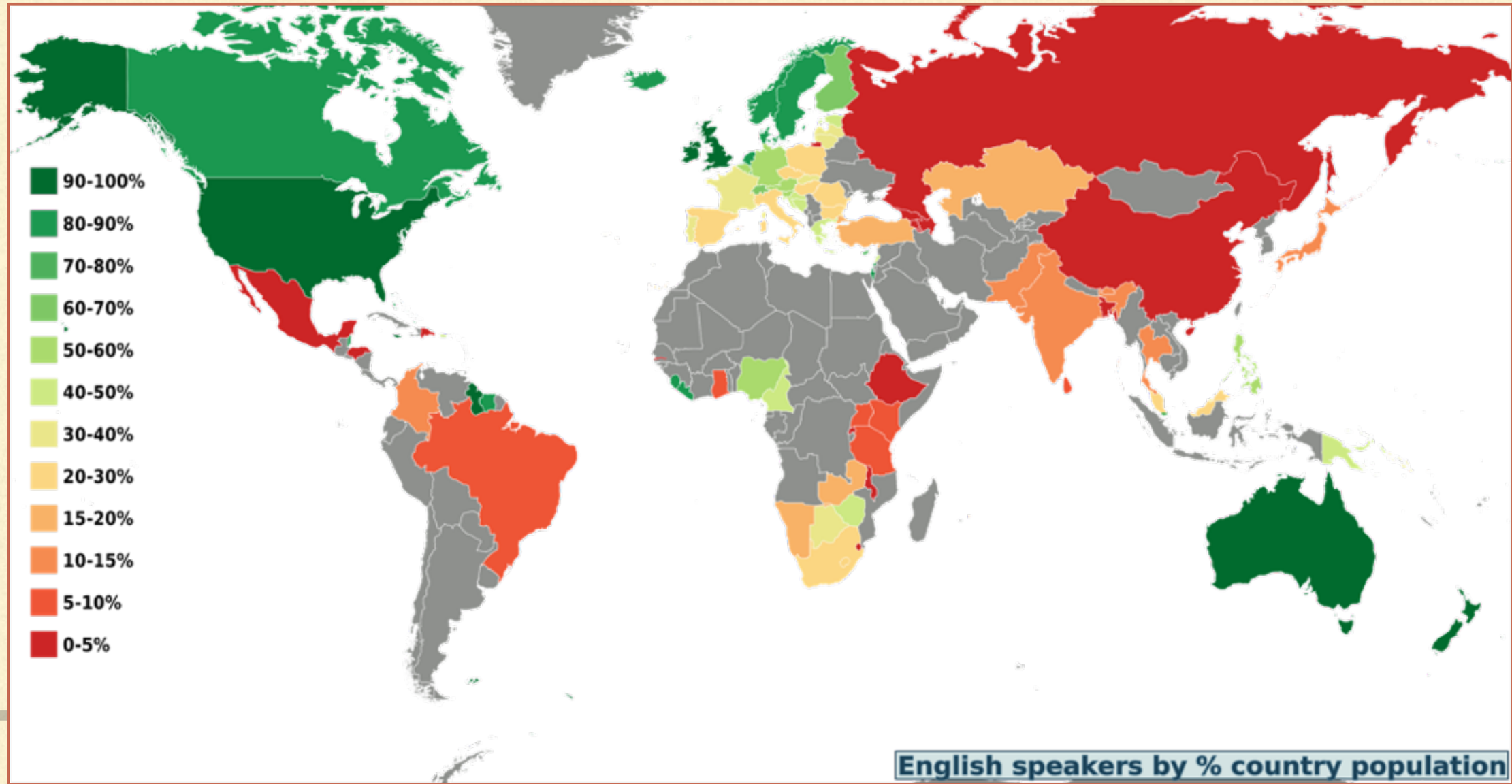
TREES AND GALTON'S PROBLEM

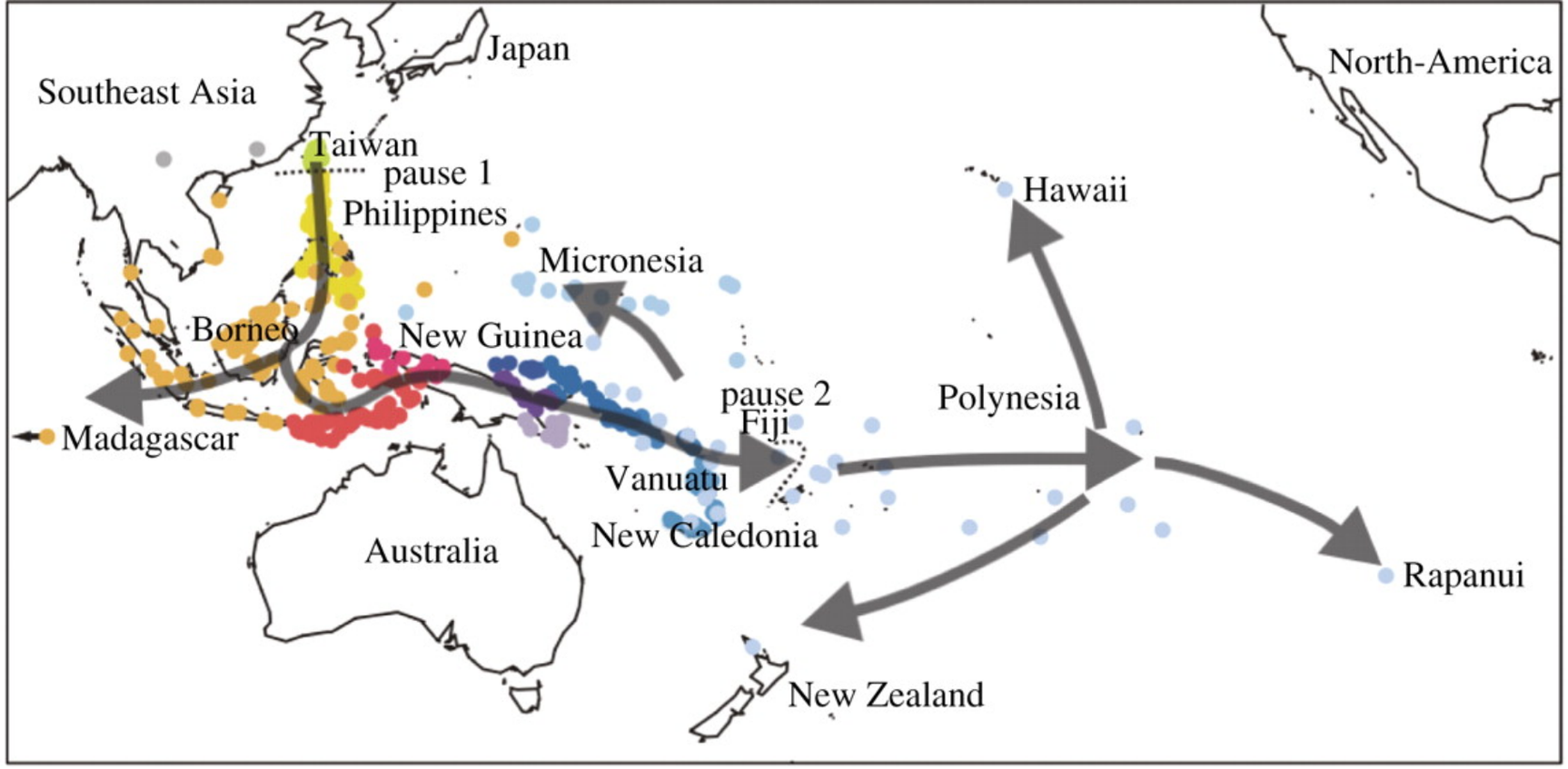
- Spurious associations:
- Relatedness by descent explains correlation between variables in two populations
-

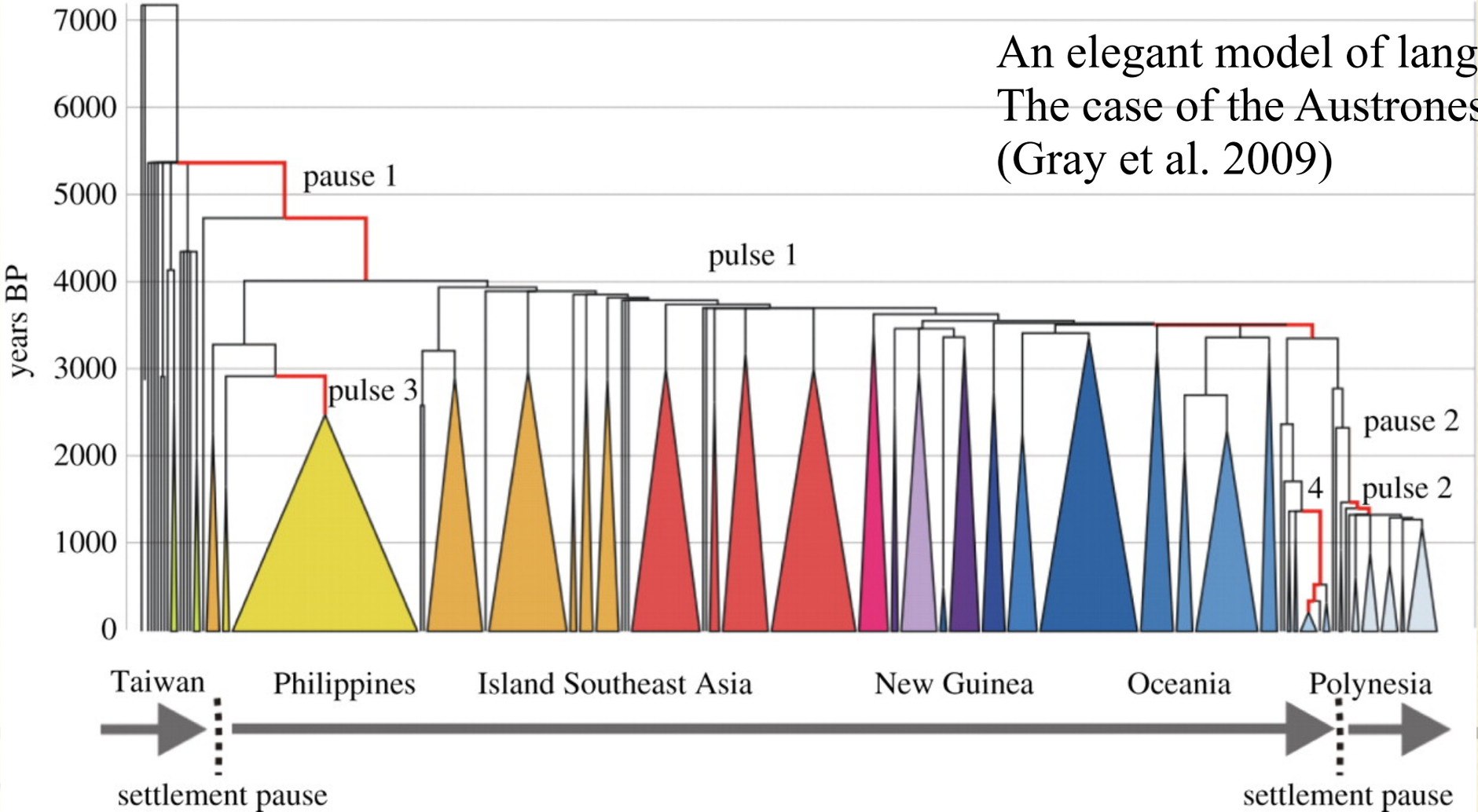


LIMITATIONS OF THE LINGUISTIC AND CULTURAL COMPARISON

- Horizontal transfer




Japan
Southeast Asia
North-America
Taiwan
pause 1
Philippines
Hawaii
Micronesia
Borneo
New Guinea
pause 2
Polynesia
Fiji
Madagascar
Vanuatu
New Caledonia
Australia
Rapanui
New Zealand


An elegant model of language spread:
The case of the Austronesian
(Gray et al. 2009)
7000
6000
5000
4000
3000
2000
1000
0
years BP
pause 1
pulse 1
pulse 3
pause 2
4
pulse 2
Taiwan
Philippines
Island Southeast Asia
New Guinea
Oceania
Polynesia
settlement pause
settlement pause