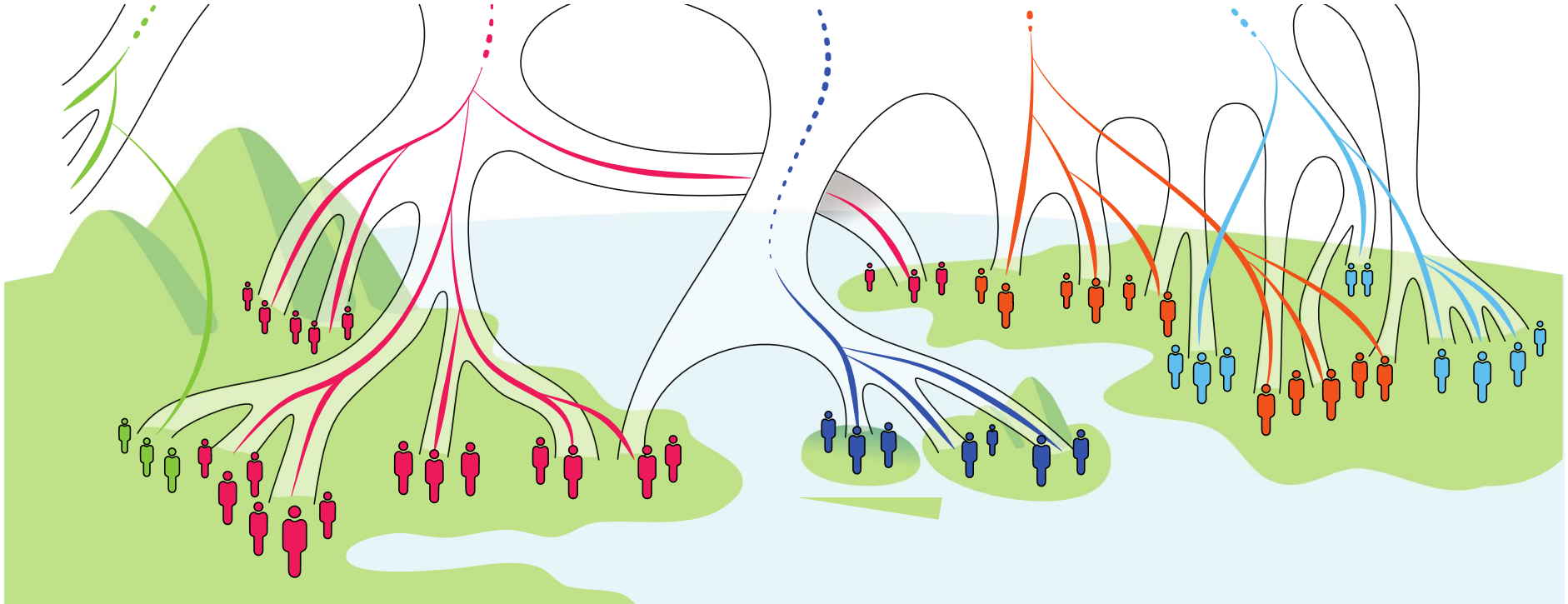
@chiarabarbieri.bsky.social
barbieri.chiara@gmail.com

Fundamentals of population genetics
to understand the links between micro and macro evolutionary processes

Chiara Barbieri

University of Cagliari, Italy

Research group “Human genetic diversity across languages and cultures”, University of Zurich, Switzerland

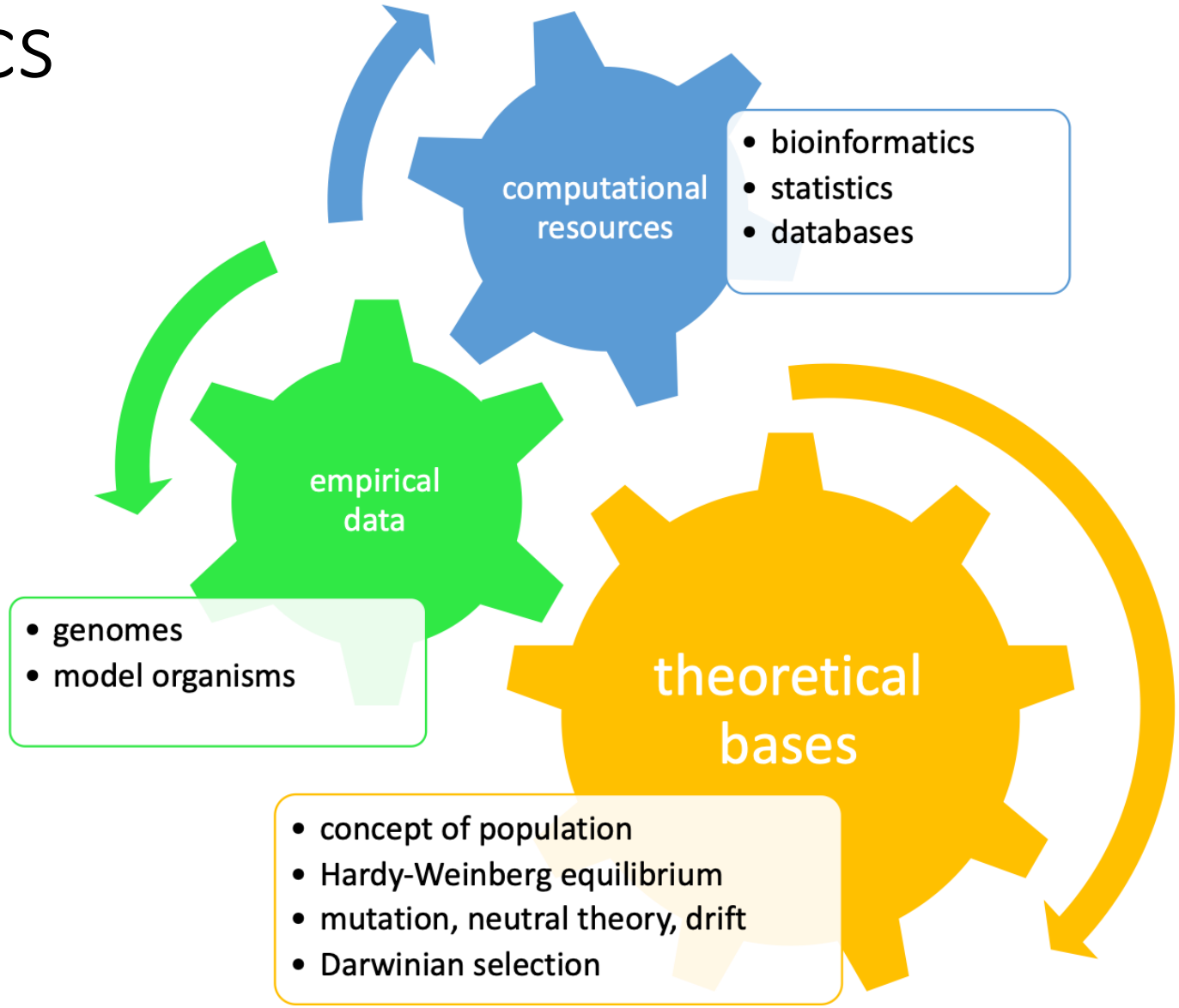


Population genetics

Population genetics

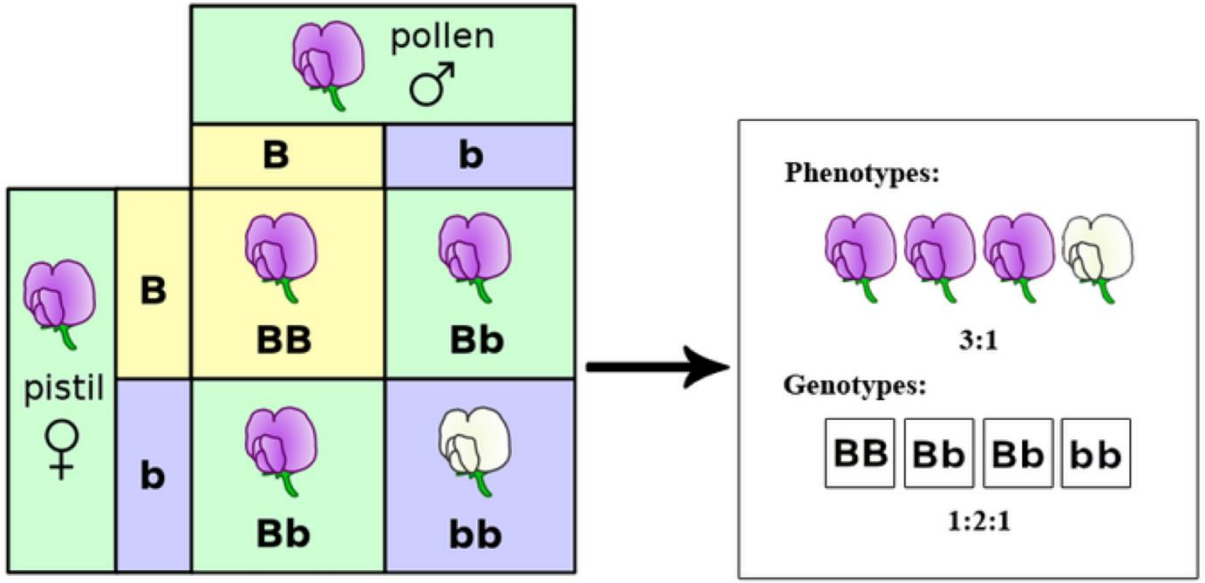
+ population ecology

→ population biology



Population genetics

- Mendel's law (heritability) applied to <u>populations</u> of organisms
- Understand the forces that produce evolutionary changes through time, and how variation is maintained
- Genotype + phenotype in the environment background



Darwin missed the inheritance mechanism

- Fleeming Jenkin (1867) **blending** inheritance: an offspring's phenotypic traits are a 'blend' of those of its parents. a sexually reproducing population will become phenotypically homogenous in a few generations.
- **Mendelian inheritance is particulate**: offspring inherit discrete particles (genes) from their parents, sexual reproduction does not diminish the heritable variation in the population.
- 1900: **Biometry** approach (Karl Pearson in London) involved statistical analysis of the phenotypic variation in natural populations. interested in **continuously** varying traits such as height. Following Darwinian **gradualism**.
- Opposed to the biometricians, the **Mendelians** (William Bateson) emphasized **discontinuous** variation. major adaptive change could be produced by single mutational steps, rather than by cumulative natural selection. Mendelian inheritance came to be associated with an anti-Darwinian view of evolution.

History of Popgen

- Reconcile gradual variation and mendelian law of heritability
- Fisher, Haldane, Wright (1920s)
- Kimura (1964)
- First theoretical discipline, then proved experimentally

wikipedia.org

Definition of “population”

- A spatial-temporal group of interbreeding individuals who share a common gene pool
- In the same time and space
  - Hardy-Weinberg equilibrium model
- … but, in the real world?

What is a population?

- Population as a unit of research
  - Vertical transmission of traits
  - Unit stable in space and time
- Geographically defined
- Or different ecological niches, or behaviours

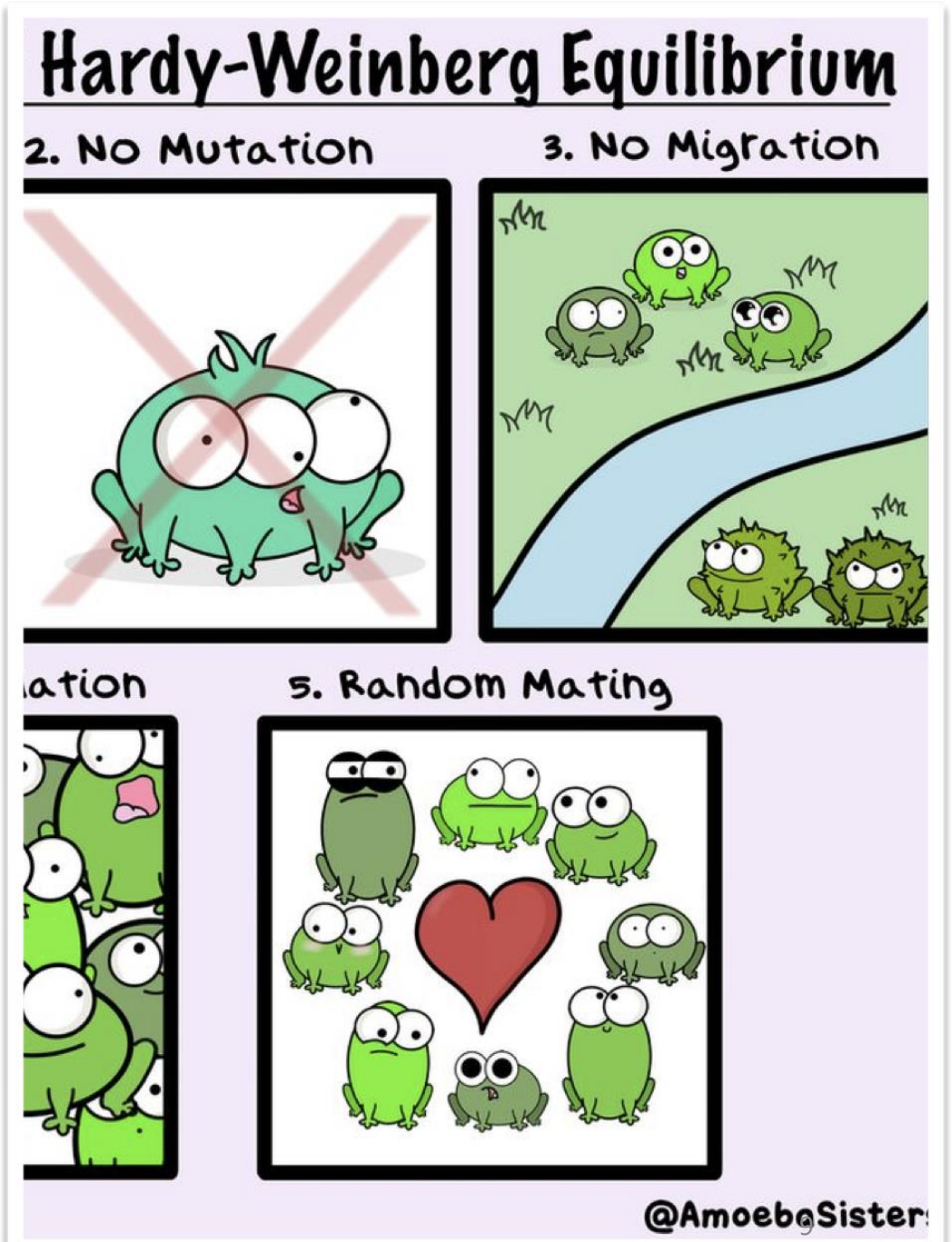
What is a population?

- 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

Individual vs. Population

- Genetic comparisons can be performed at the individual level or at the population level
- Individual level:
  - inbreeding coefficient, measurements of relatedness between pairs of individuals
- Population level:
  - Average of factors above
  - Ancestry detection
  - Admixture, split migration, variation in population size

Popgen Theory



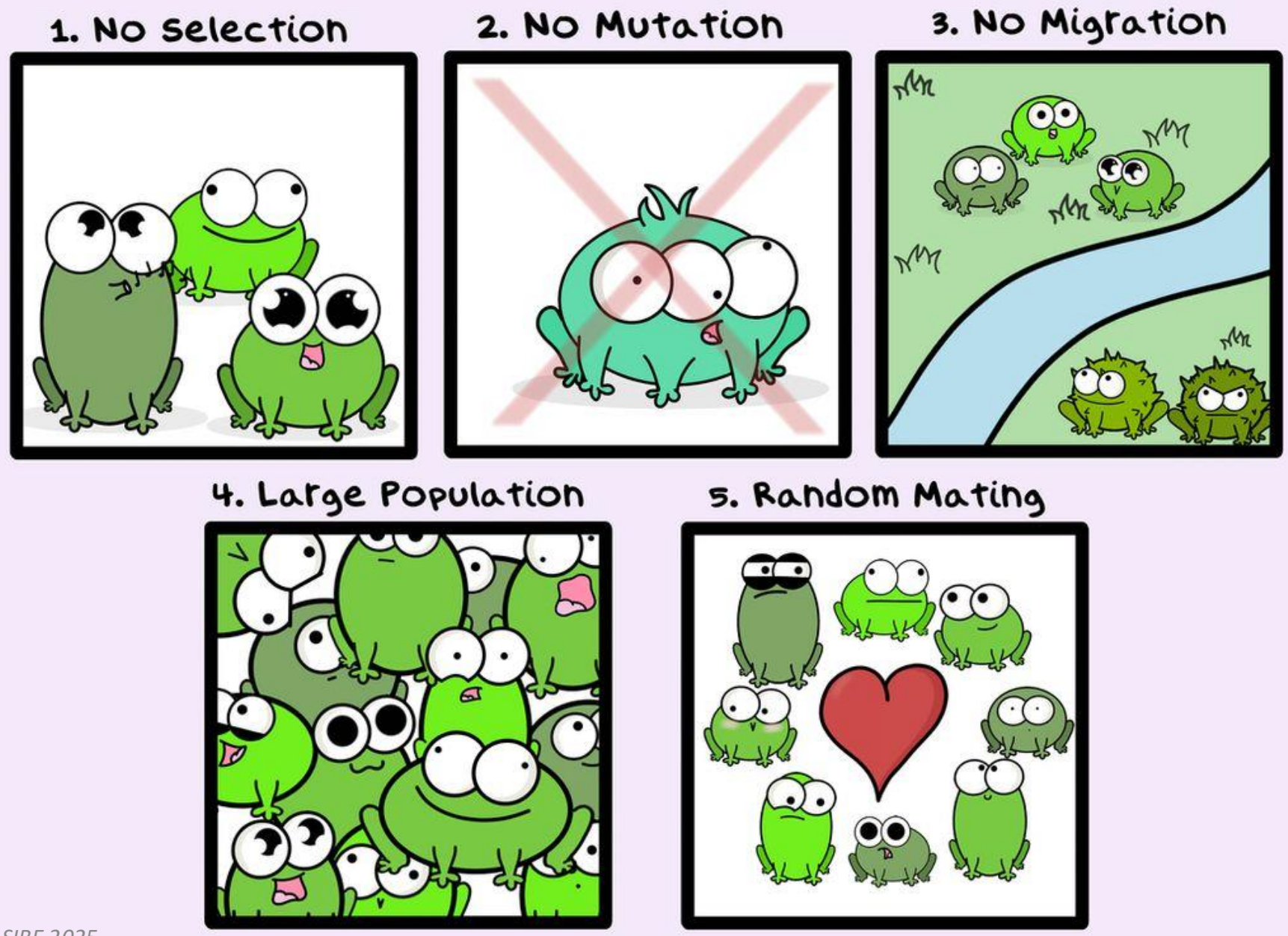
Hardy-weinberg equilibrium model

- Mathematical model to explain behavior of alleles in a population
  - the genetic variation in a population will remain constant from one generation to the next in the absence of disturbing factors
- IDEAL CIRCUMSTANCES
  - Sexual reproduction, equal number of females and males
  - Random mating
  - Large population size (in theory, infinite size)
  - No migration
  - No mutation
  - No natural selection

Hardy-weinberg equilibrium model

- Popgen theory: if a population is NOT under Hardy Weinberg equilibrium, some forces are in action!
  - Population structure
  - Drift, variation in population size
  - Selection
  - Gene-flow

Assumptions of Hardy-Weinberg Equilibrium



Population size

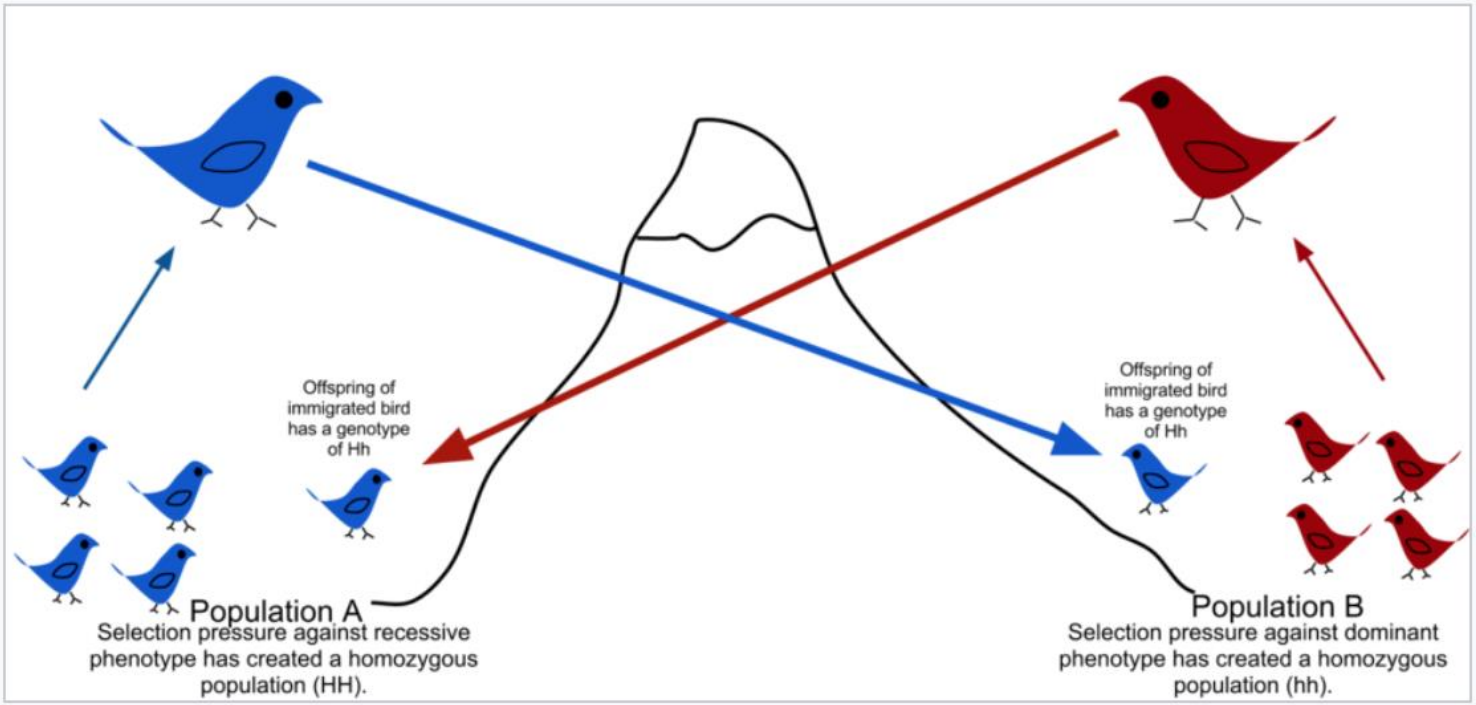
- **N_e: effective population size**
- Number of individuals who actually reproduce
- Smaller than the census size
- Depends on variation in sex ratio of breeding individuals, number of offspring, variation through generations

Small population size = Drift

Genetic bottleneck: reduction in population size corresponds to a loss of variation

Migration: gene flow

- Introducing new sources of variation from one population to another



Gene flow is the transfer of alleles from one population to another population through immigration of individuals.

wikipedia.org

Examples of genetic measurements

Genetic distance

Proportional to divergence time. Mutation accumulate through drift.

- Distance between two sequences
  - Count the number of loci where the sequences are different (pairwise distance)
  - Percentage difference

- Distance between two populations

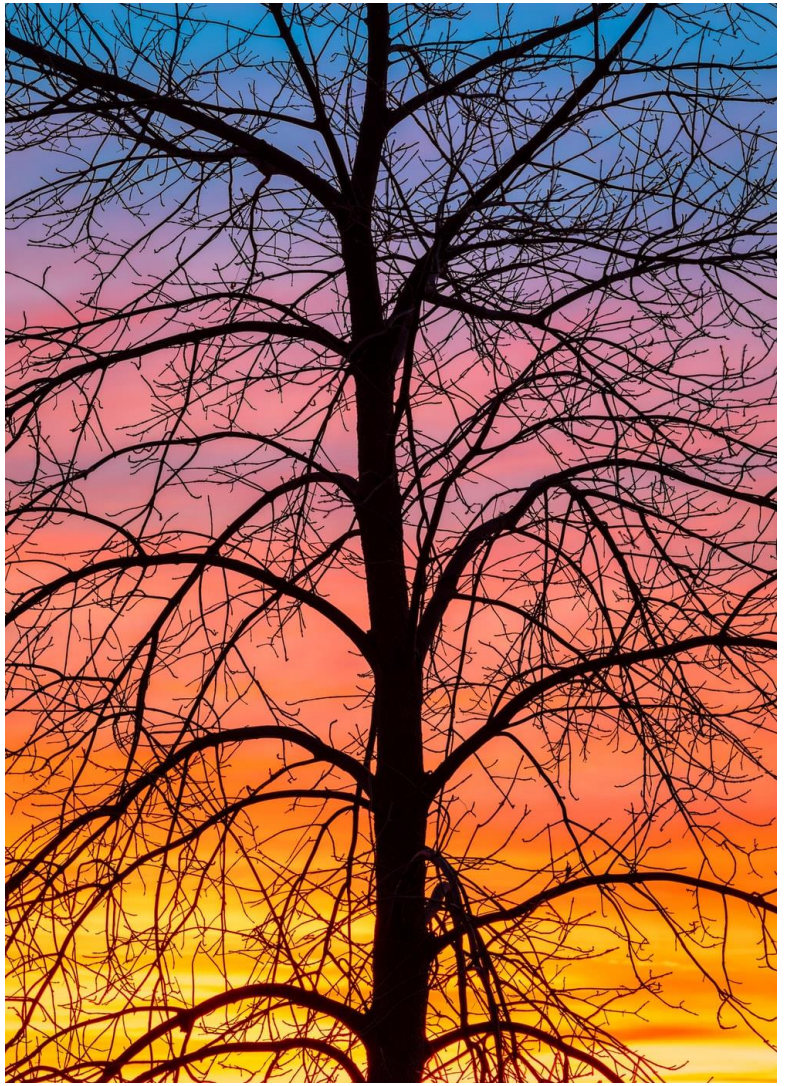
Genetic distance between populations: FST

- Distance between two <u>populations</u>
  - Fixation index (F_{ST}) is a measure of population differentiation due to genetic structure. 0=identical populations; 1= maximal diversity between populations.
  - Proportion between relatedness of individuals within each population and between populations
  - Compensate for differences in sample size

$$F_{ST} = \frac{\pi_{\text{Between}} - \pi_{\text{Within}}}{\pi_{\text{Between}}}$$

π: average number of nucleotide differences per site between two sequences

MOLECULAR PHYLOGENY

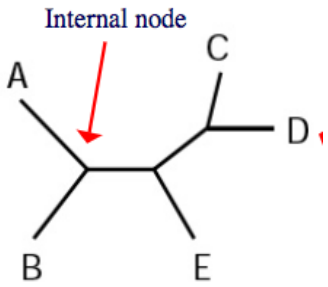


Molecular clock

- Dating split times and past events

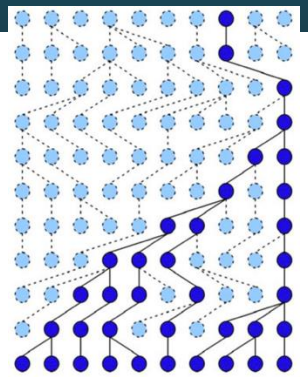
Trees and networks

- How to build a tree



Coalescent theory

- From present diversity going backwards in time



Building trees to understand evolution

Statistic methods for phylogenies

- Likelihood, Bayesian etc

Methods for building trees

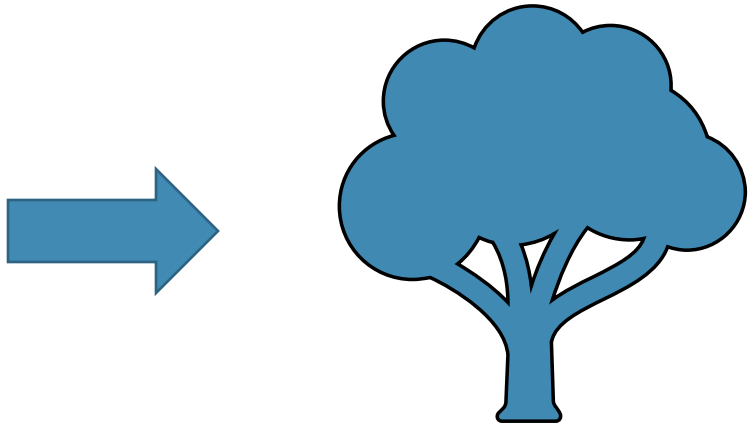
- Distance-based
  - UPGMA
  - Neighbour Joining (NJ)
- Character-based
  - Maximum Parsimony (MP)
  - Maximum Likelihood (ML)
  - Bayesian methods (Markov Chain Monte Carlo MCMC)

Distance-based trees

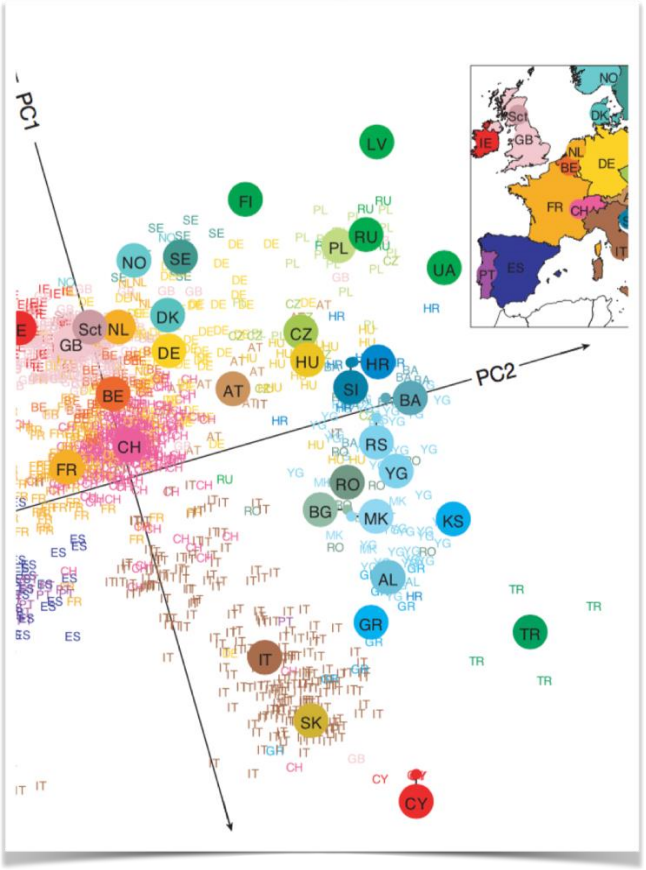
- First calculate distance matrix between pairs of sequences or populations
- Then build a tree

Distance Matrix

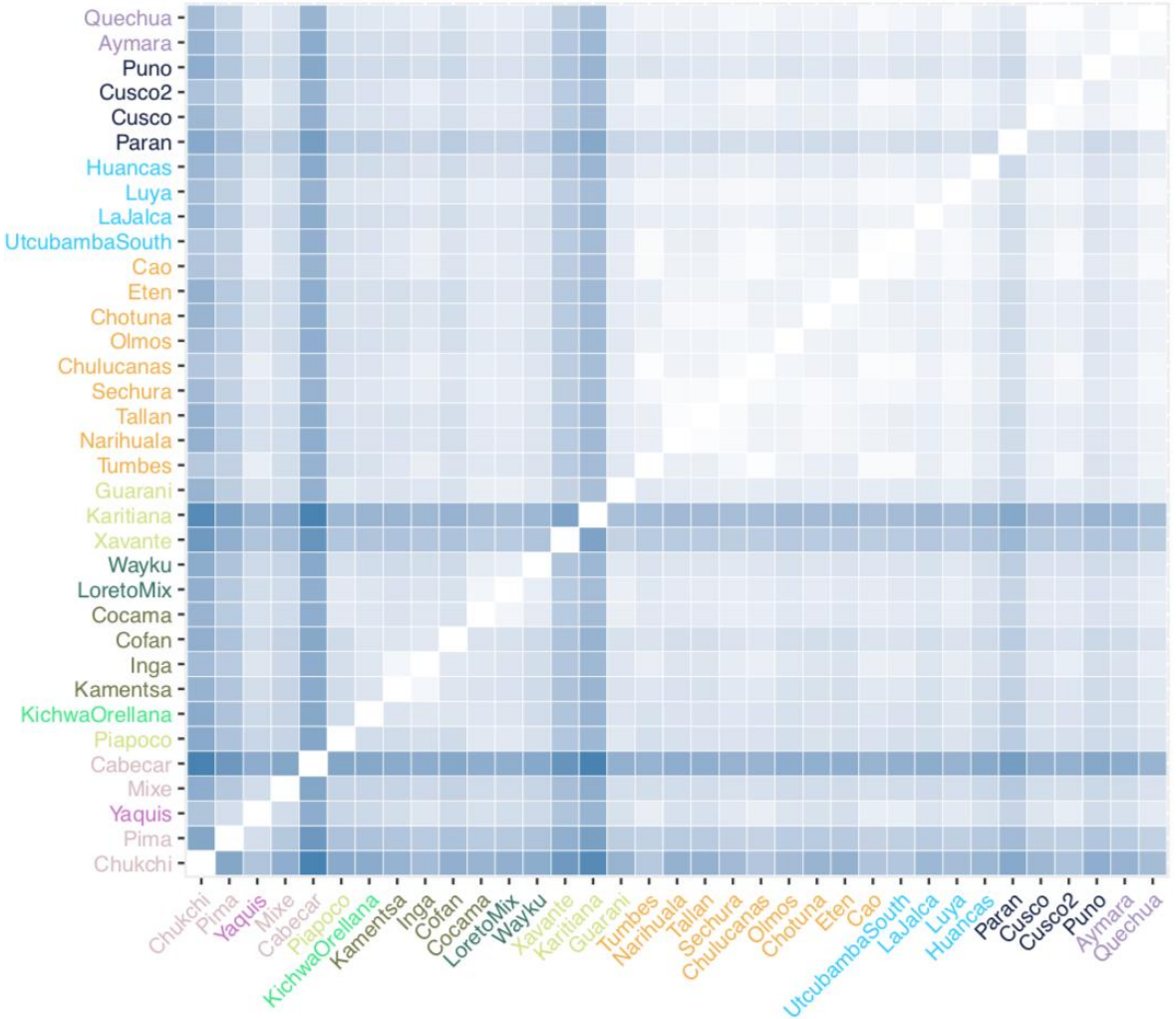
	A	B	C	D	E	F
A	0	16	47	72	77	79
B	16	0	37	57	65	66
C	47	37	0	40	30	35
D	72	57	40	0	31	23
E	77	65	30	31	0	10
F	79	66	35	23	10	0

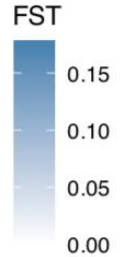


Visualize genetic distances



FST distance matrix: Relationship between populations





Barbieri et al. 2019 Mol Biol Evol

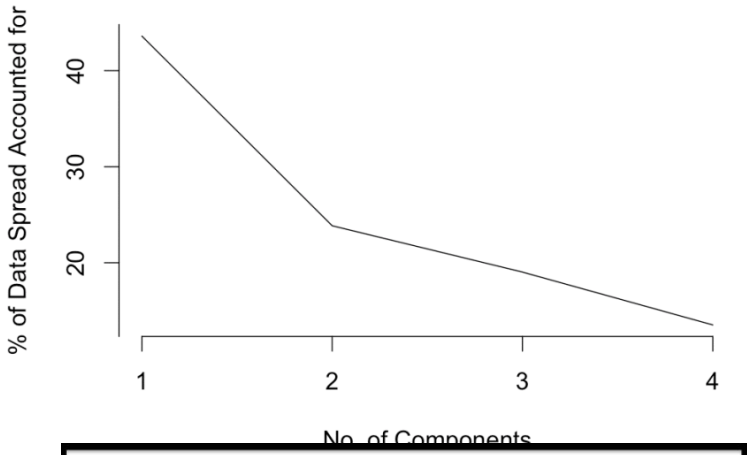
PCA - principal component analysis

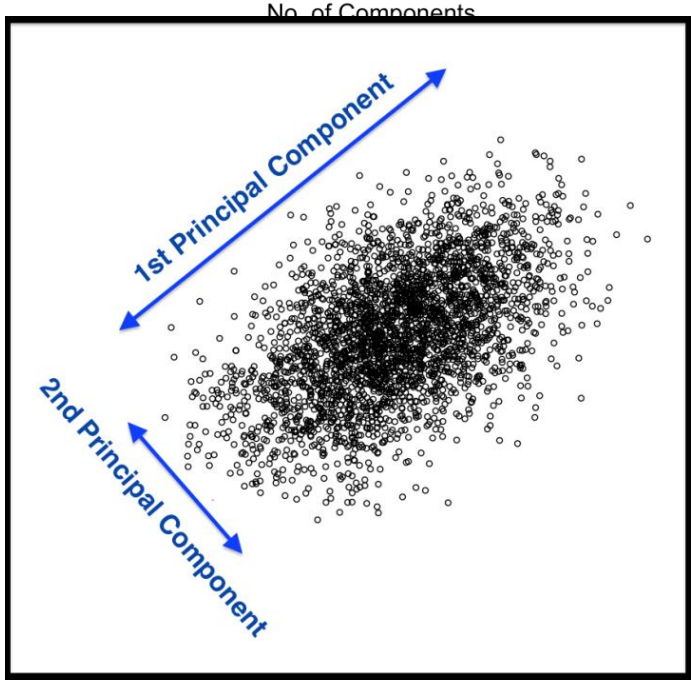
Principal component analysis (PCA) is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables (entities each of which takes on various numerical values) into a set of values of linearly uncorrelated variables called **principal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

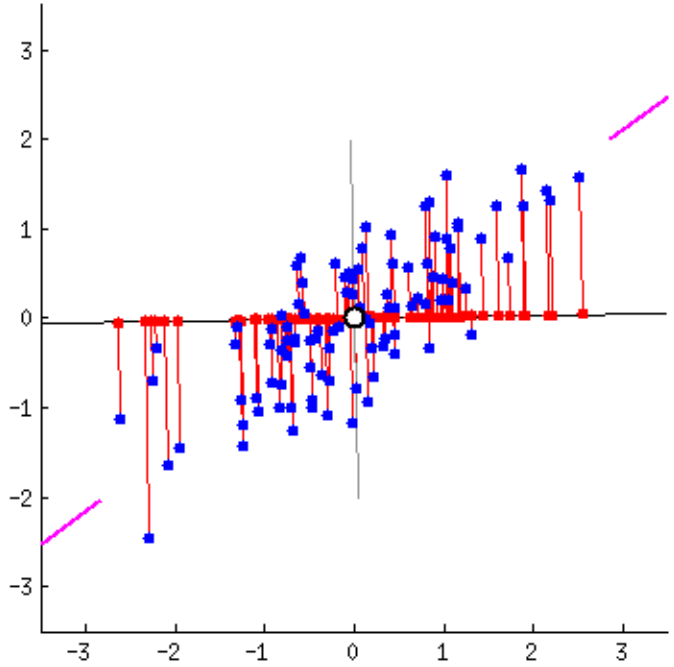
Eigenvalues decomposition of a covariance matrix from uncorrelated source of variation (genetic markers).

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





- PCA 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



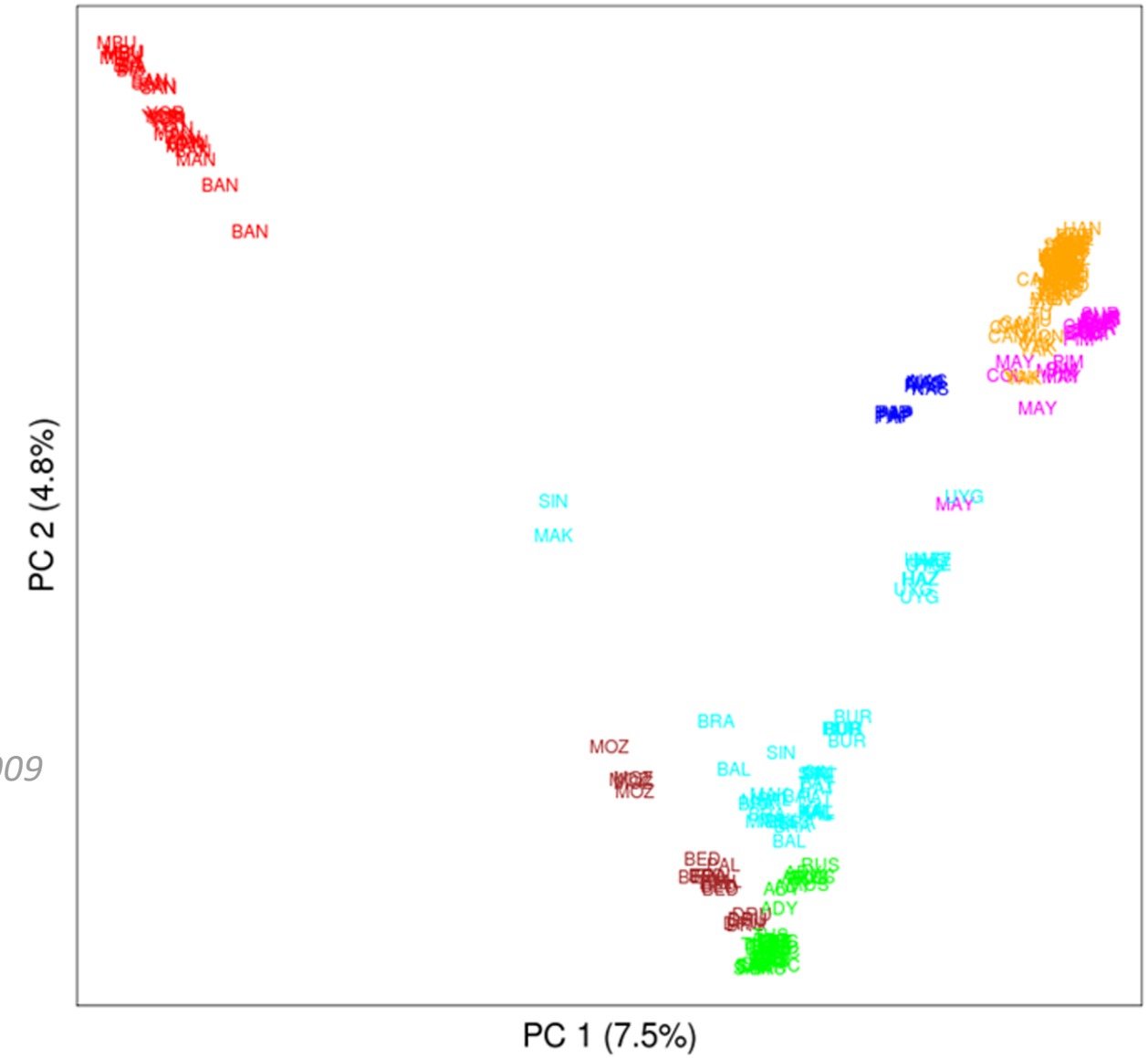
amoeba (https://stats.stackexchange.com/users/28666/amoeba), Making sense of principal component analysis, eigenvectors & eigenvalues

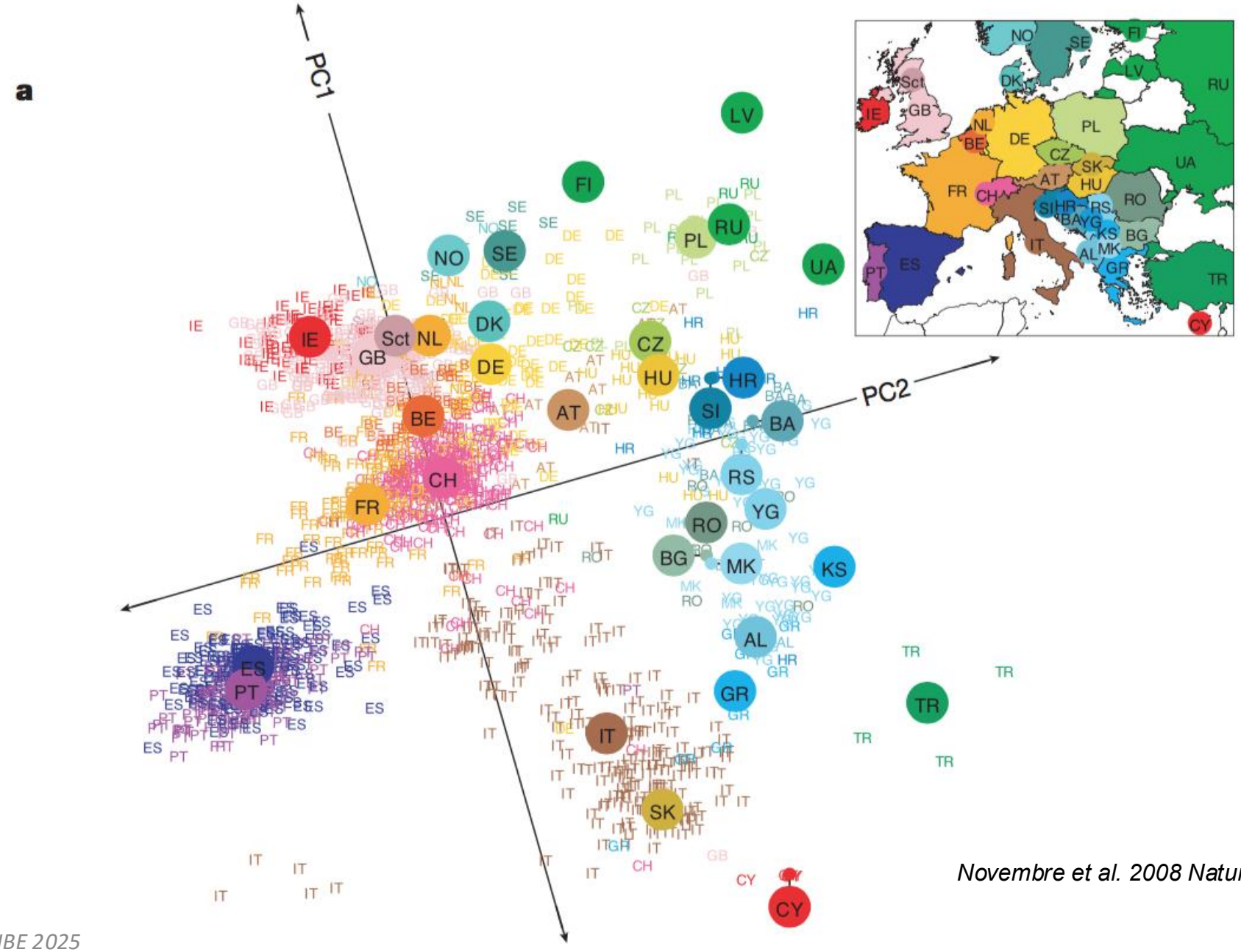
credits: algobeans.com

Relationship between individuals

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





Novembre et al. 2008 Nature

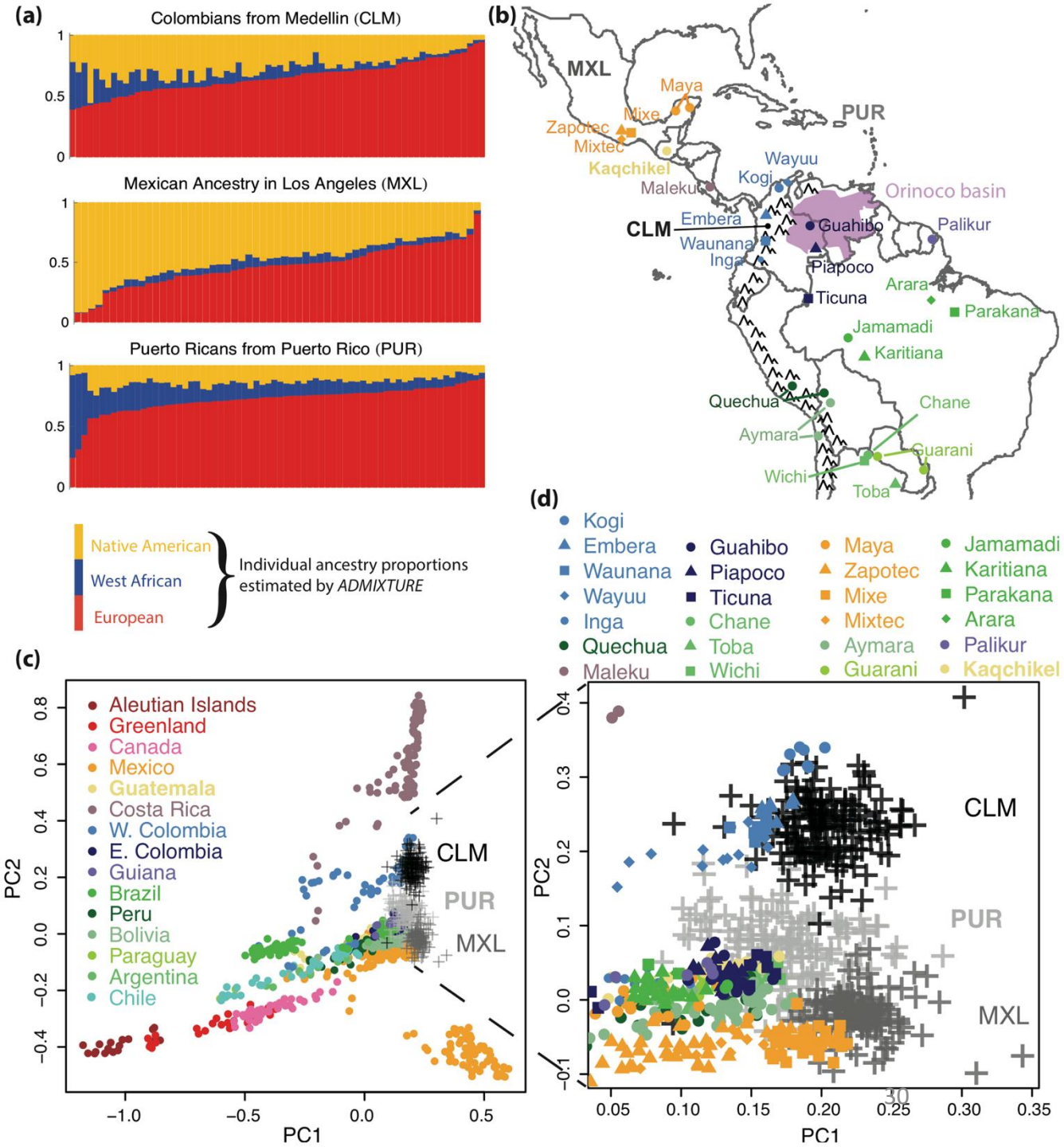
ADMIXTURE

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.

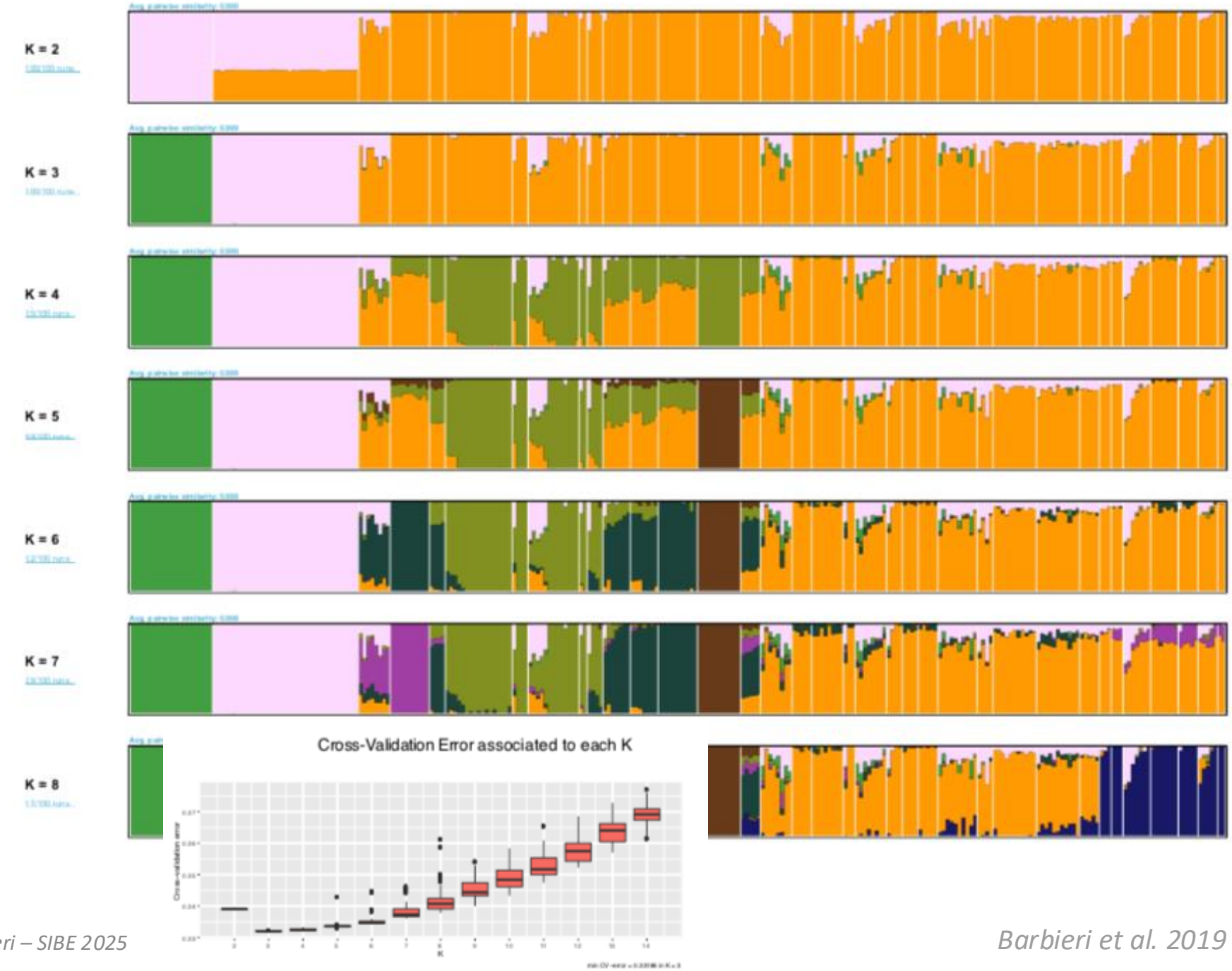
Admixture barplot has become a de-facto standard used as a non-parametric description of genetic data alongside a Principle Components Analysis.

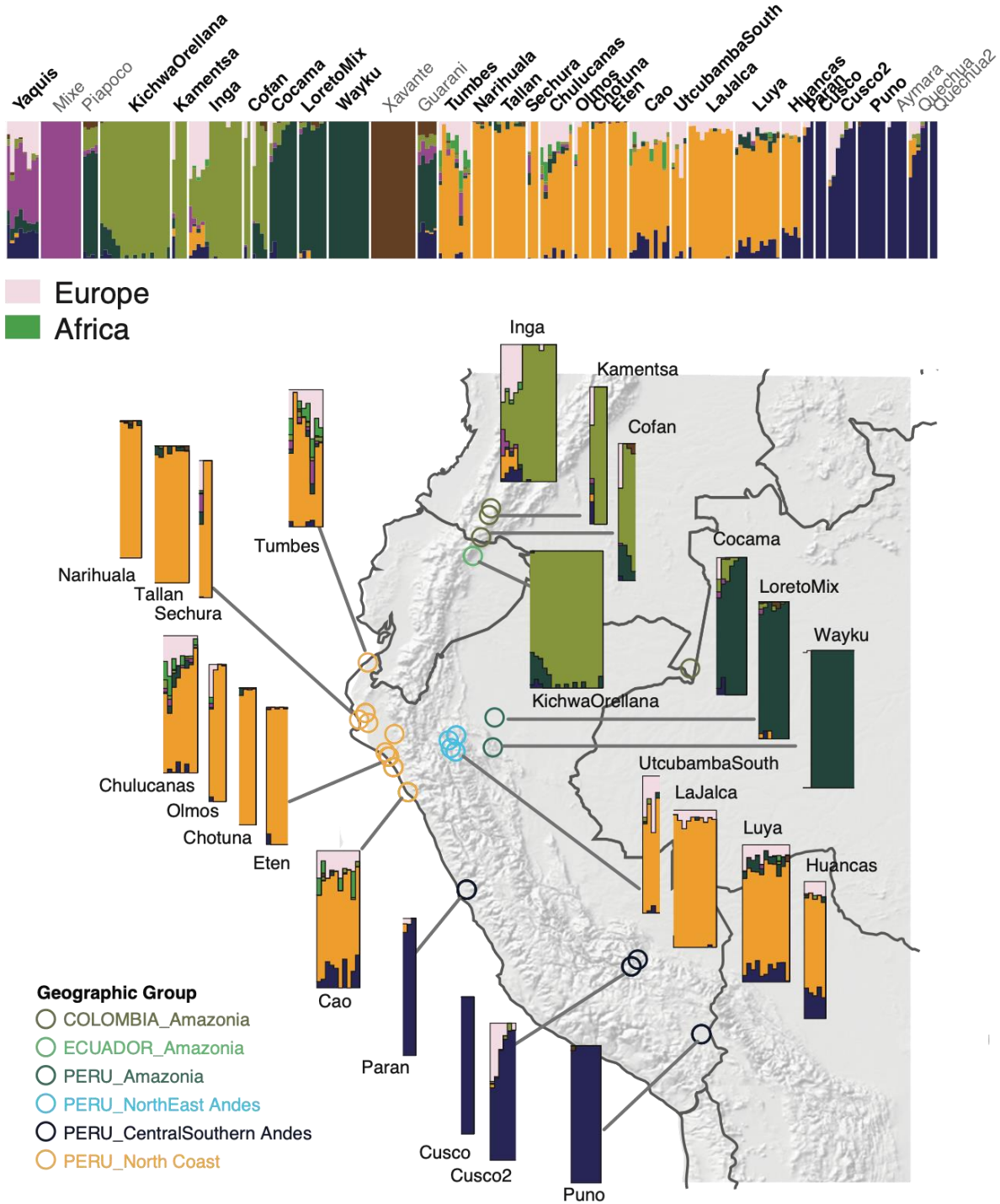
Admixture

Explain genetic diversity in terms of ancestry blocks



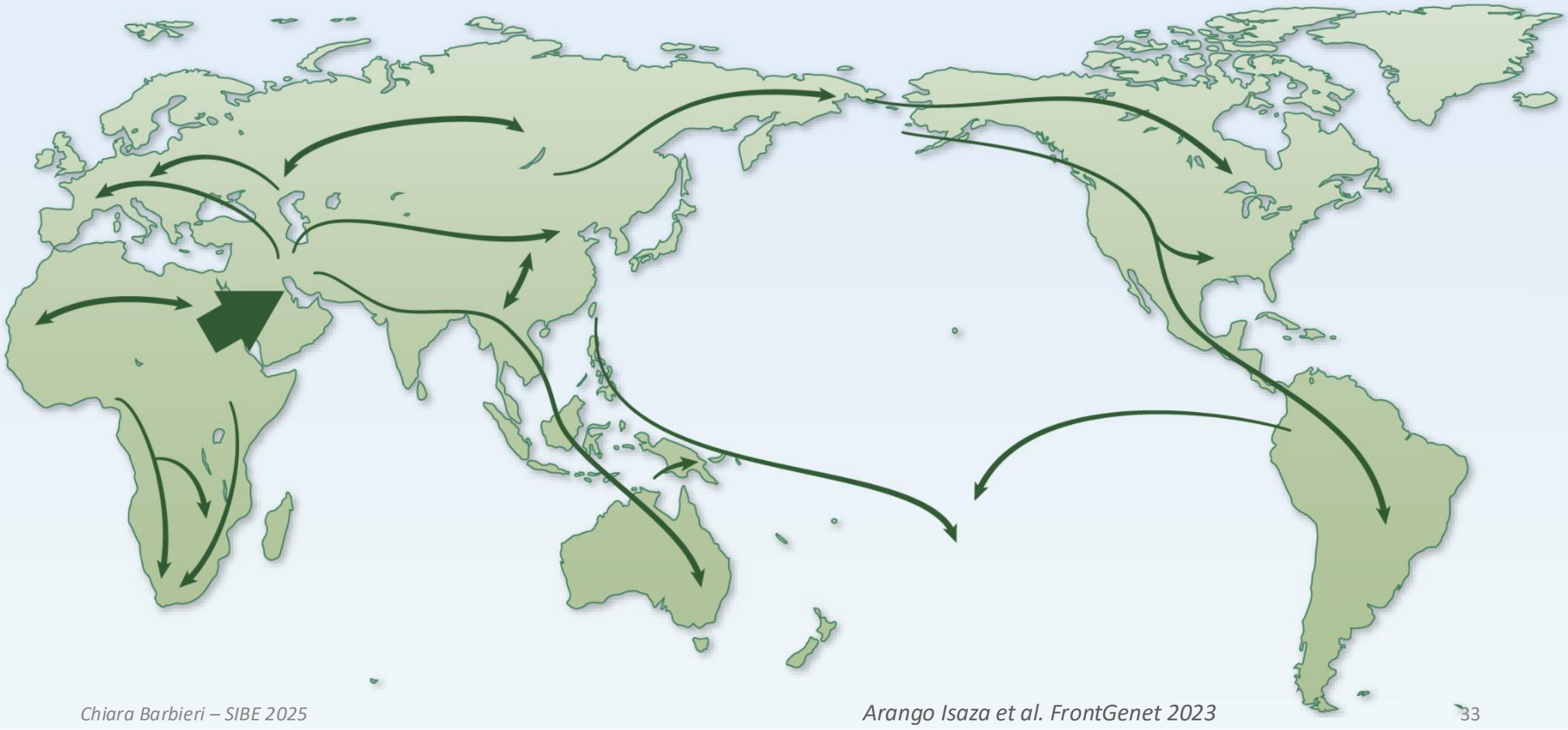
Gravel et al. Plos Genet 2013


K = 2
K = 3
K = 4
K = 5
K = 6
K = 7
K = 8
Cross-Validation Error associated to each K

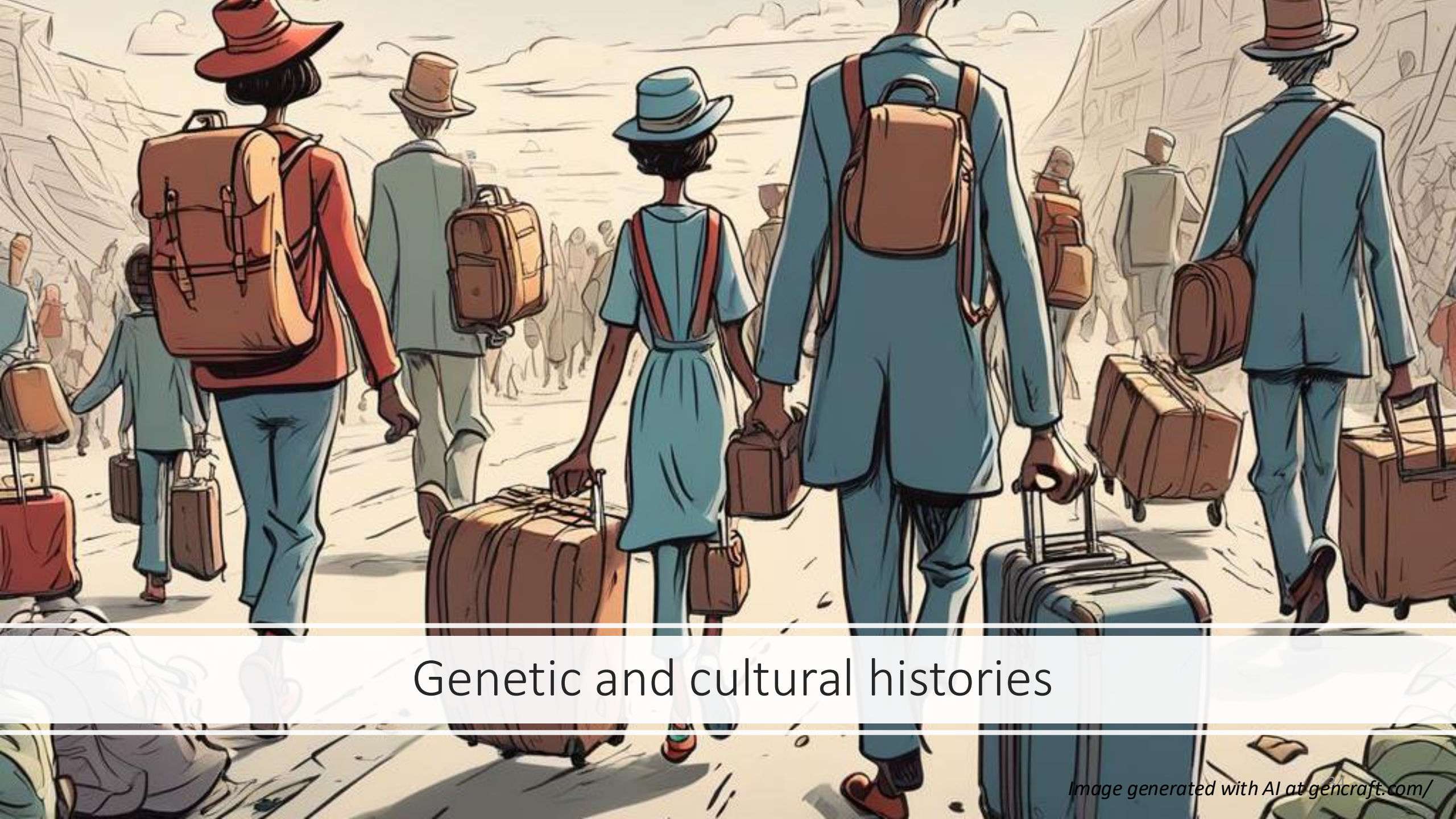


Barbieri et al. 2019 Mol Biol Evol

Genetics to reconstruct human migrations and demography



Arango Isaza et al. FrontGenet 2023



Genetic and cultural histories

Image generated with AI at gencraft.com/

Cultural evolution: Transmission with modification

- Evolutionary processes shaped by environment (society) and biology, influencing each others

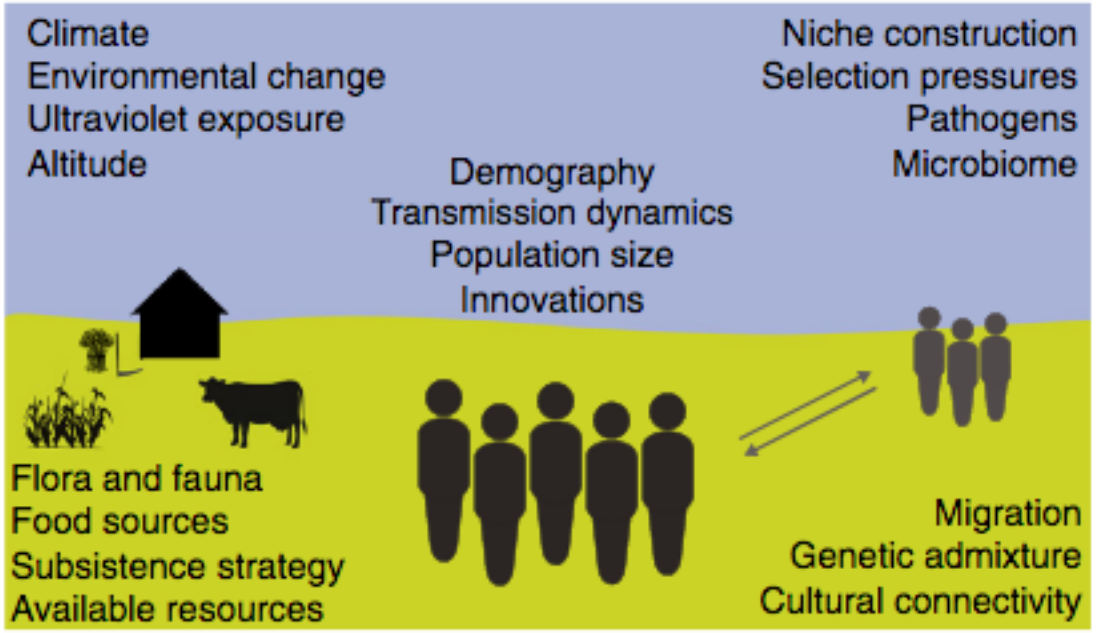
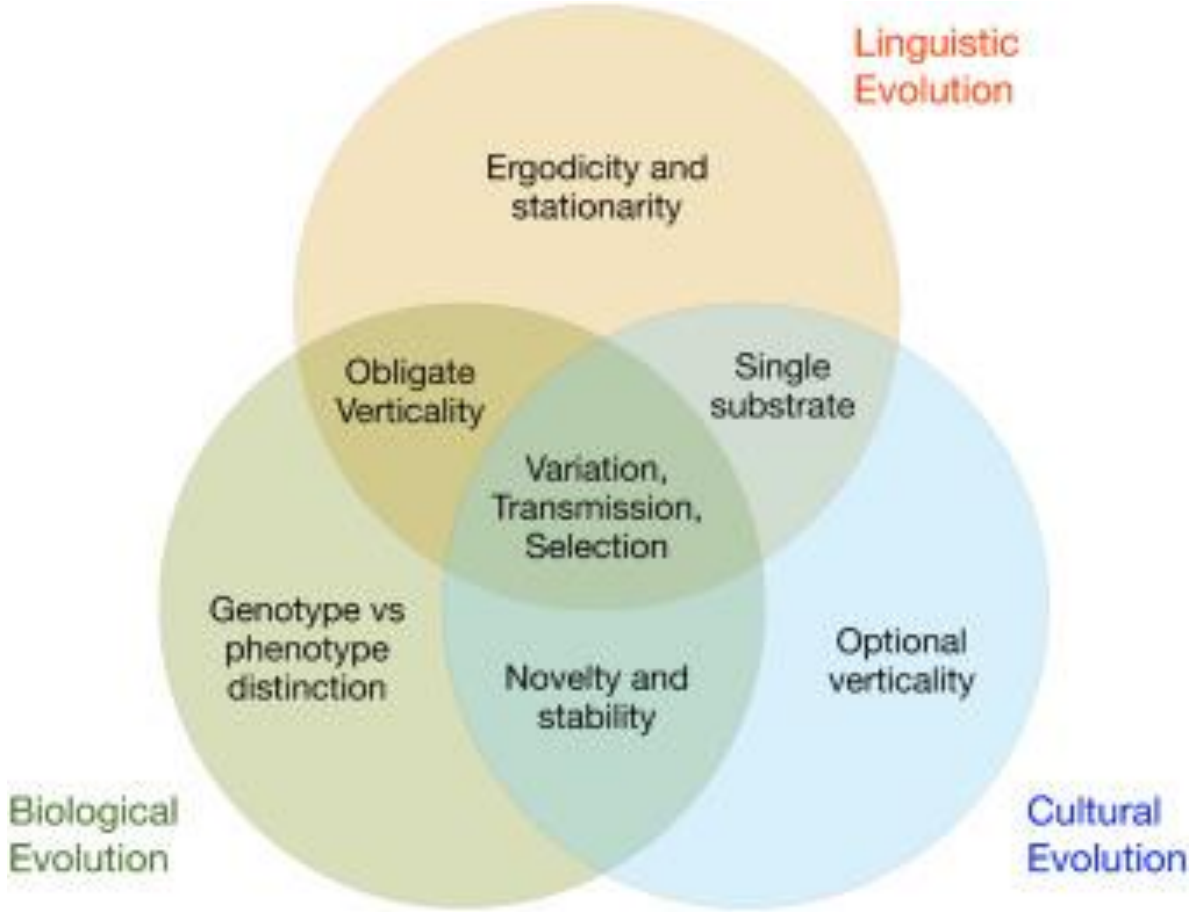


Fig. 2. Cultural, genetic, and environmental factors influencing evolution.

Creanza, Kolodny and Feldman - PNAS 2017

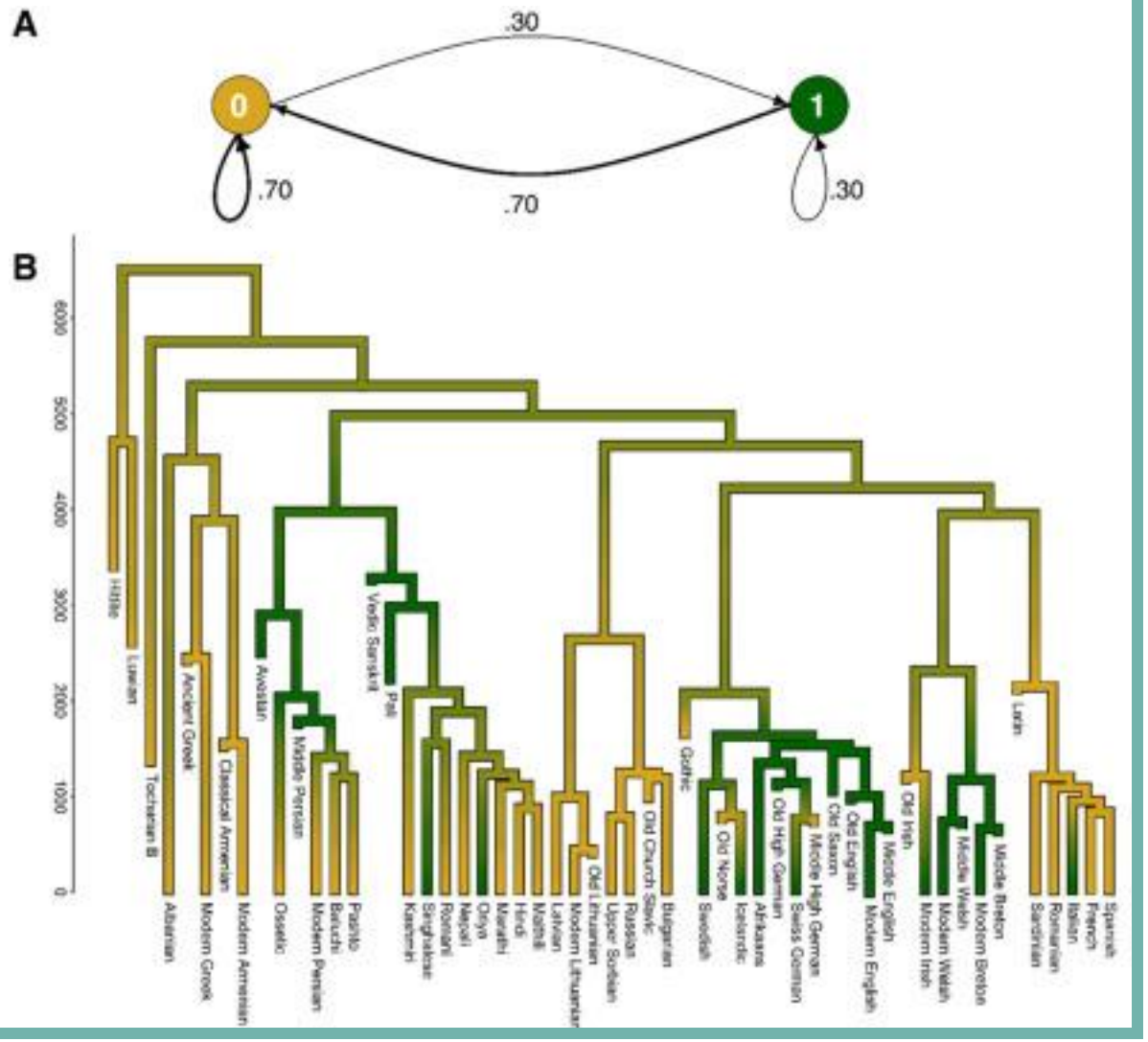
Linking biological, linguistic and cultural history

- Biological evolution: genotype vs phenotype
- Bio and cultural evolution: cumulative
- Linguistic evolution: cyclic oscillation between states

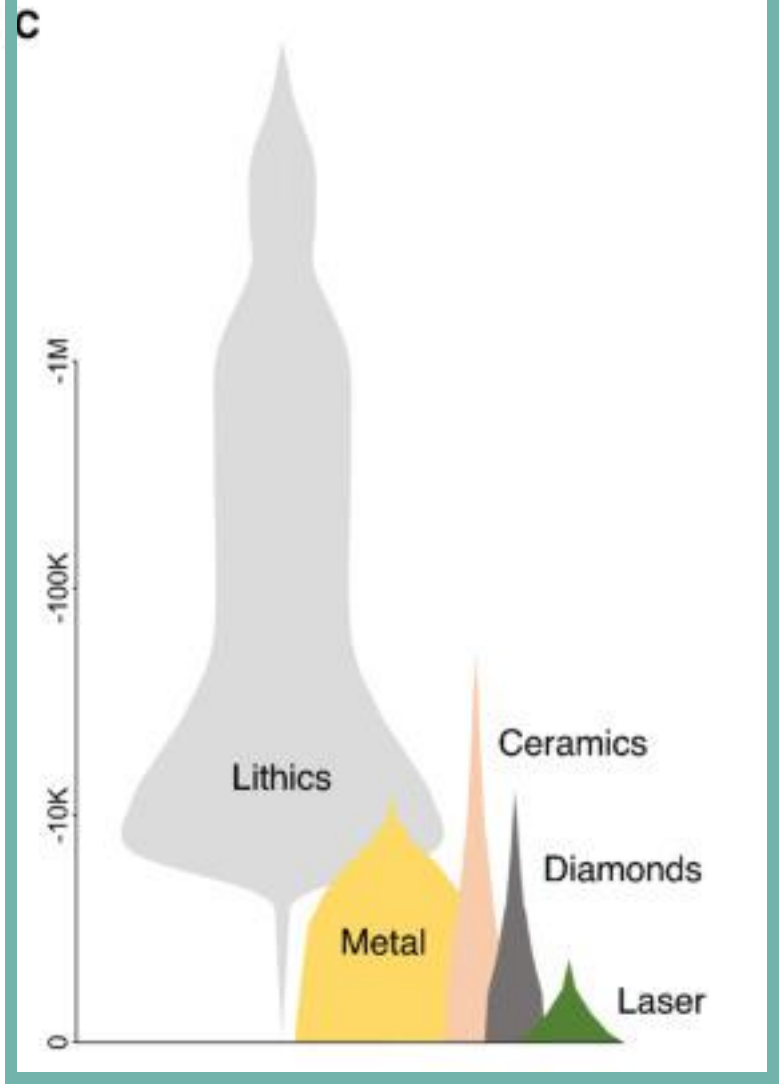


Bickel et al. 2024, Phys. Life Rev.

Linguistic evolution

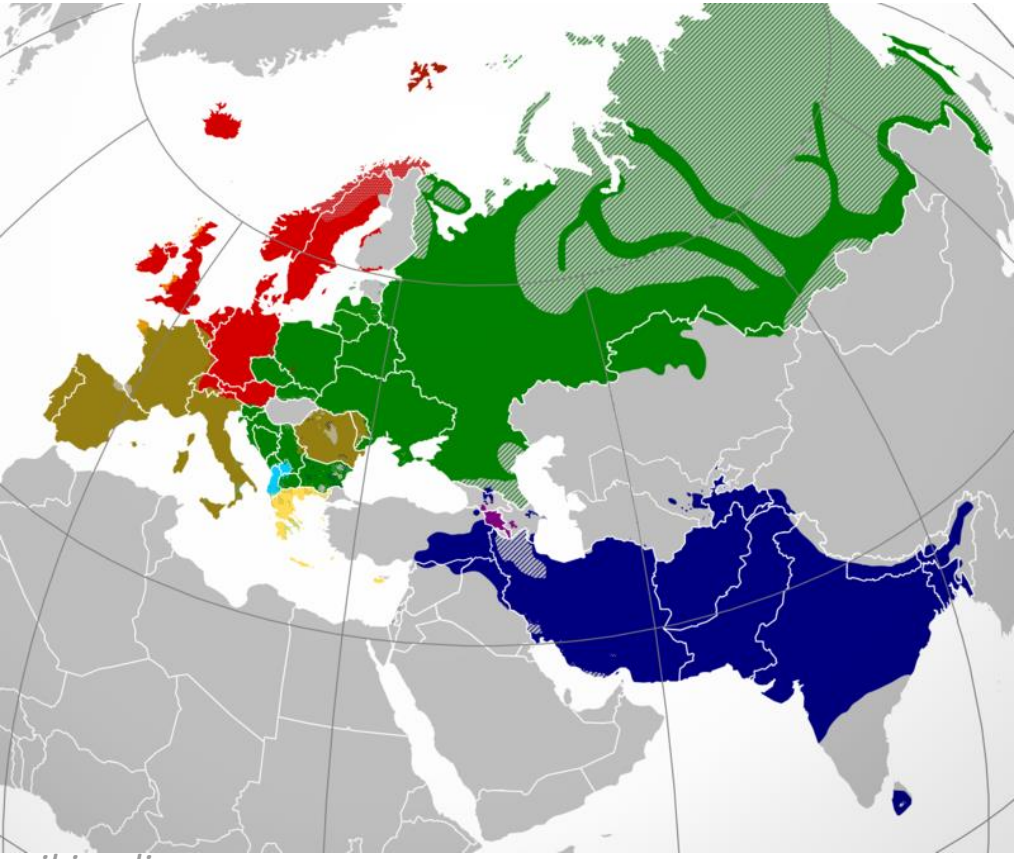


Cultural evolution



Bickel et al. 2024, Phys. Life Rev.

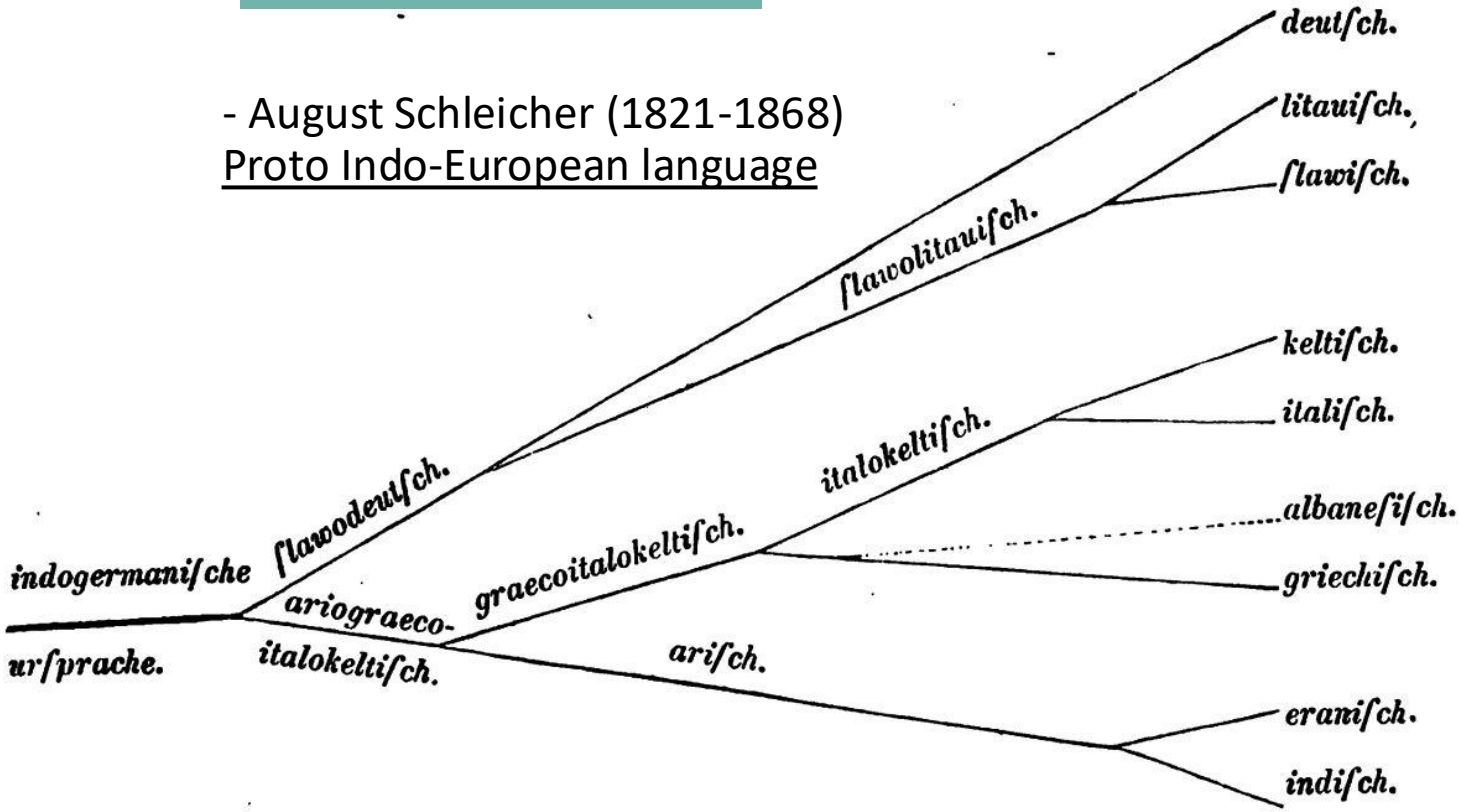
Classify languages and make trees

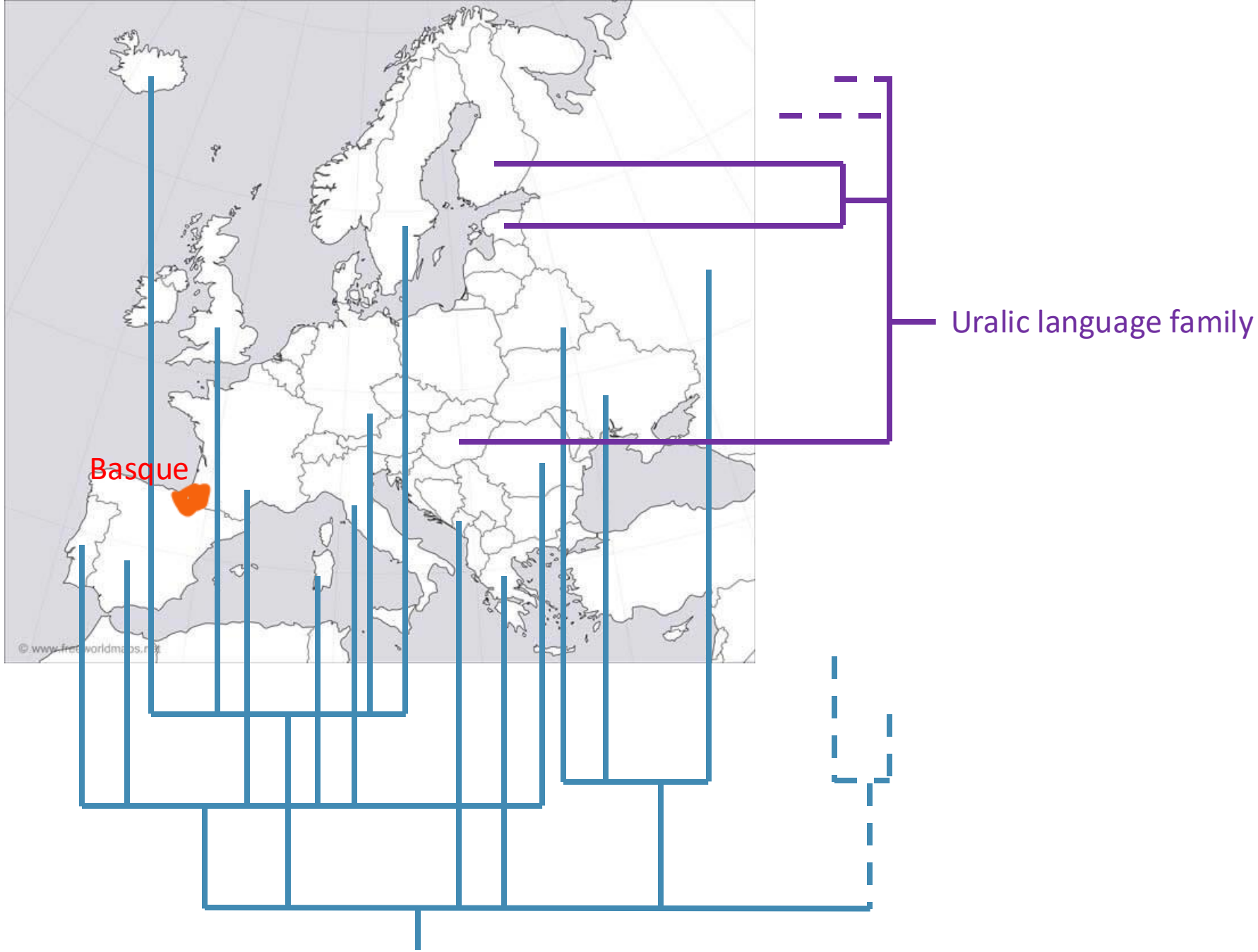


wikipedia.org

LANGUAGE FA

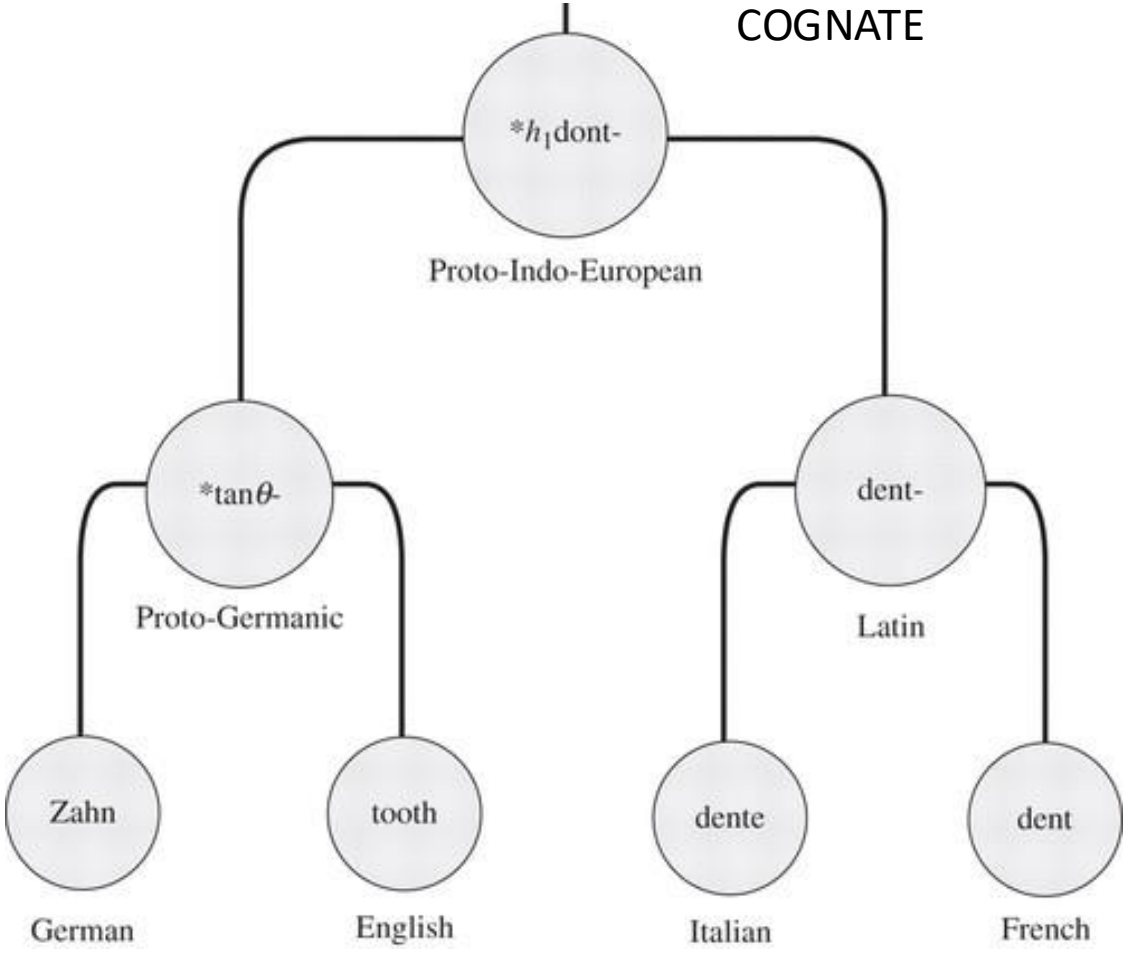
- August Schleicher (1821-1868)
Proto Indo-European language



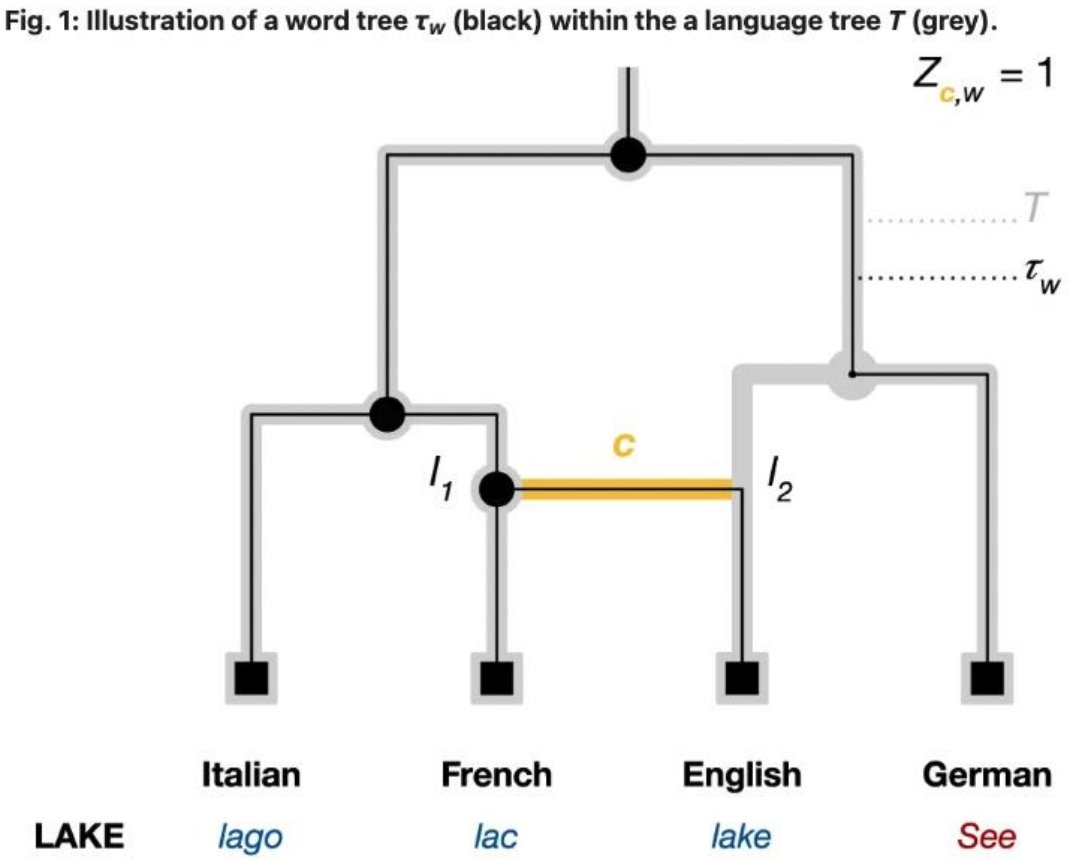

Basque
Uralic language family
Indo-European language family

Classify languages and make trees

COGNATE



BORROWING

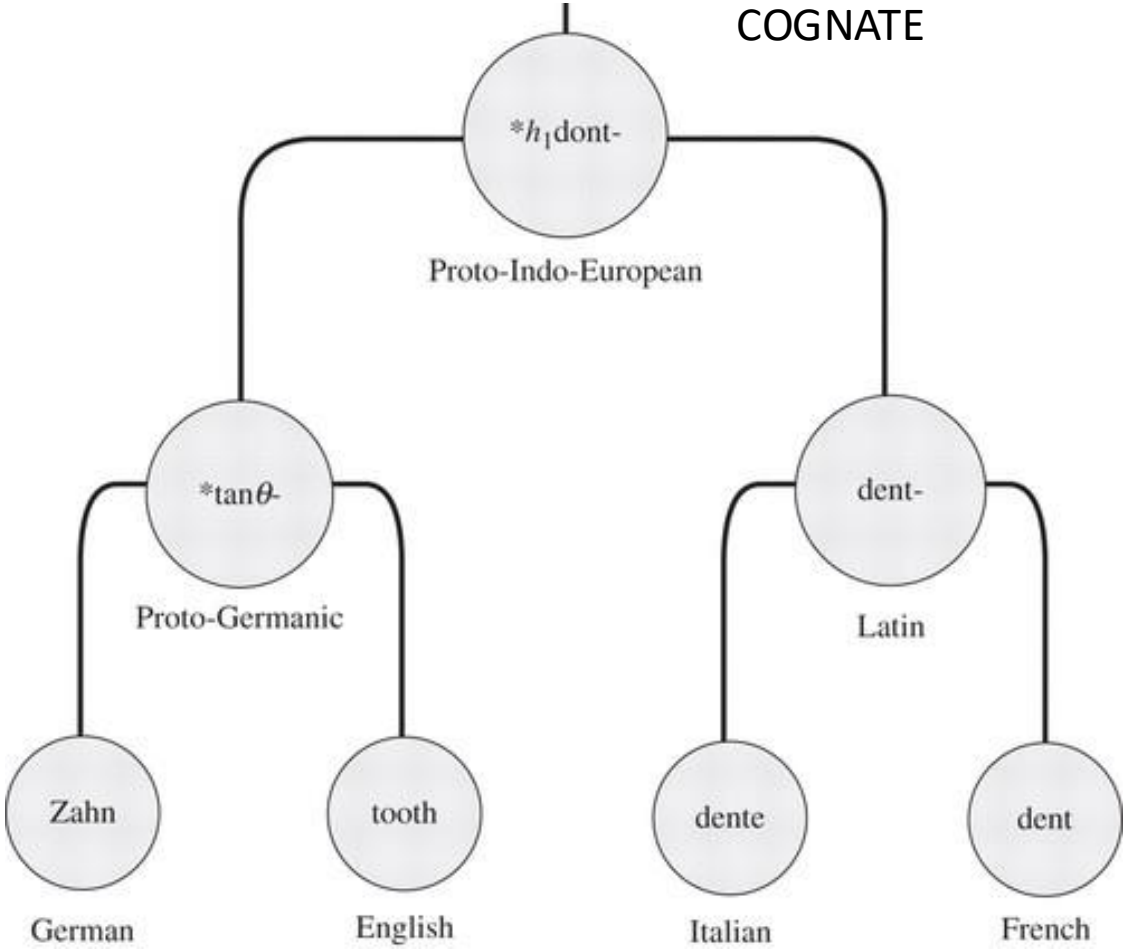


Nelson-Sathi et al. Proc. R. Soc. B. 2011

Neureiter et al. Hum Soc Science Comm 2022

Classify languages and make trees

COGNATE



BORROWING

Fig. 1: Illustration of a word tree τ_w (black) within the a language tree T (grey).

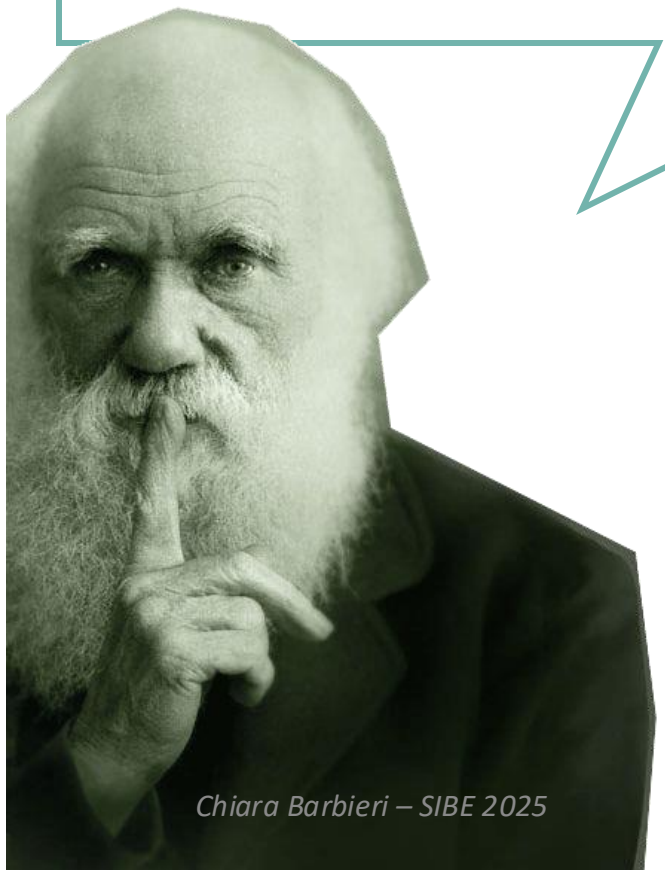
Nelson-Sathi et al. Proc. R. Soc. B. 2011

Neureiter et al. Hum Soc Science Comm 2022

Testing Darwin's idea

"If we possessed a perfect pedigree of mankind, a genealogical arrangement of the races of man would afford the best classification of the various languages now spoken throughout the world"

[The Origin of Species, 1859]



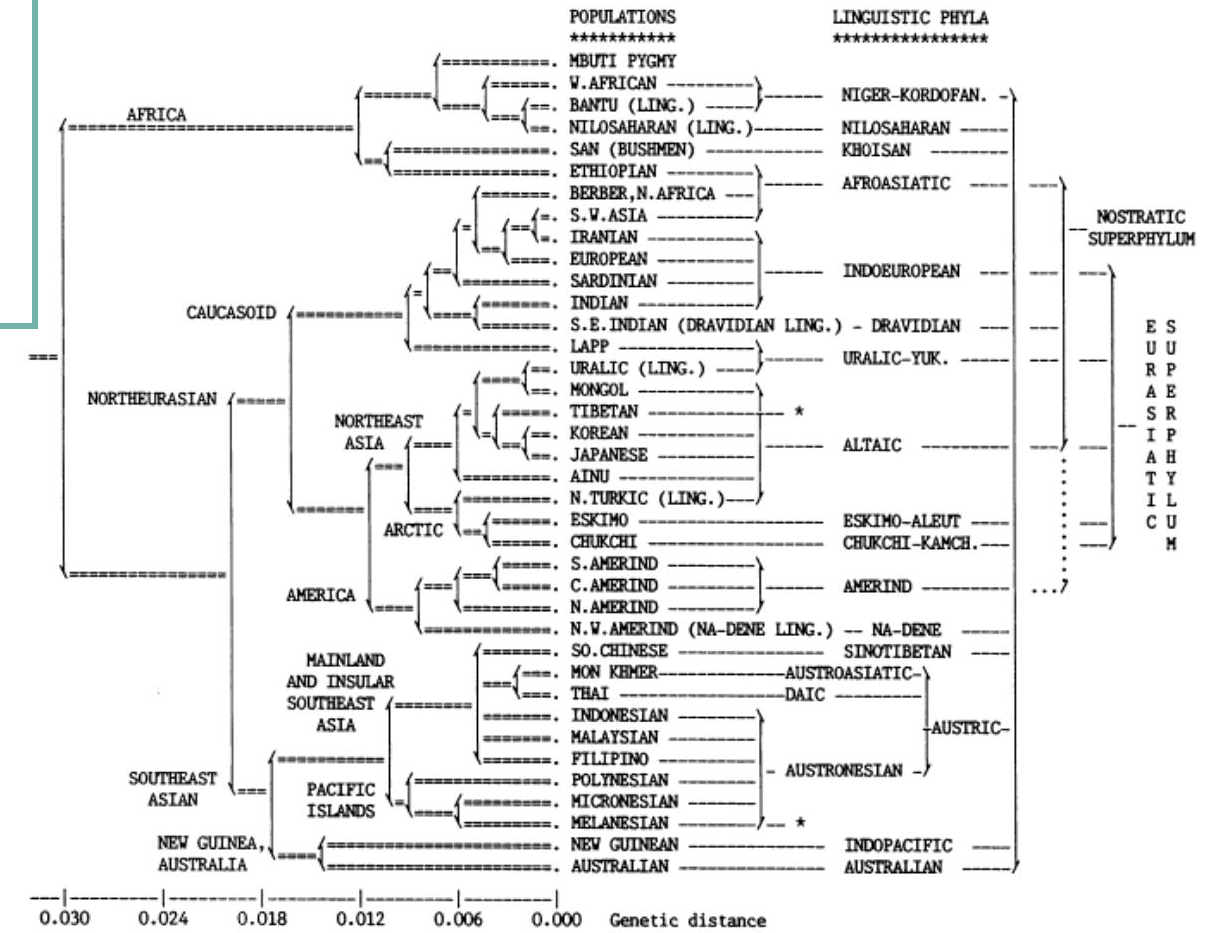
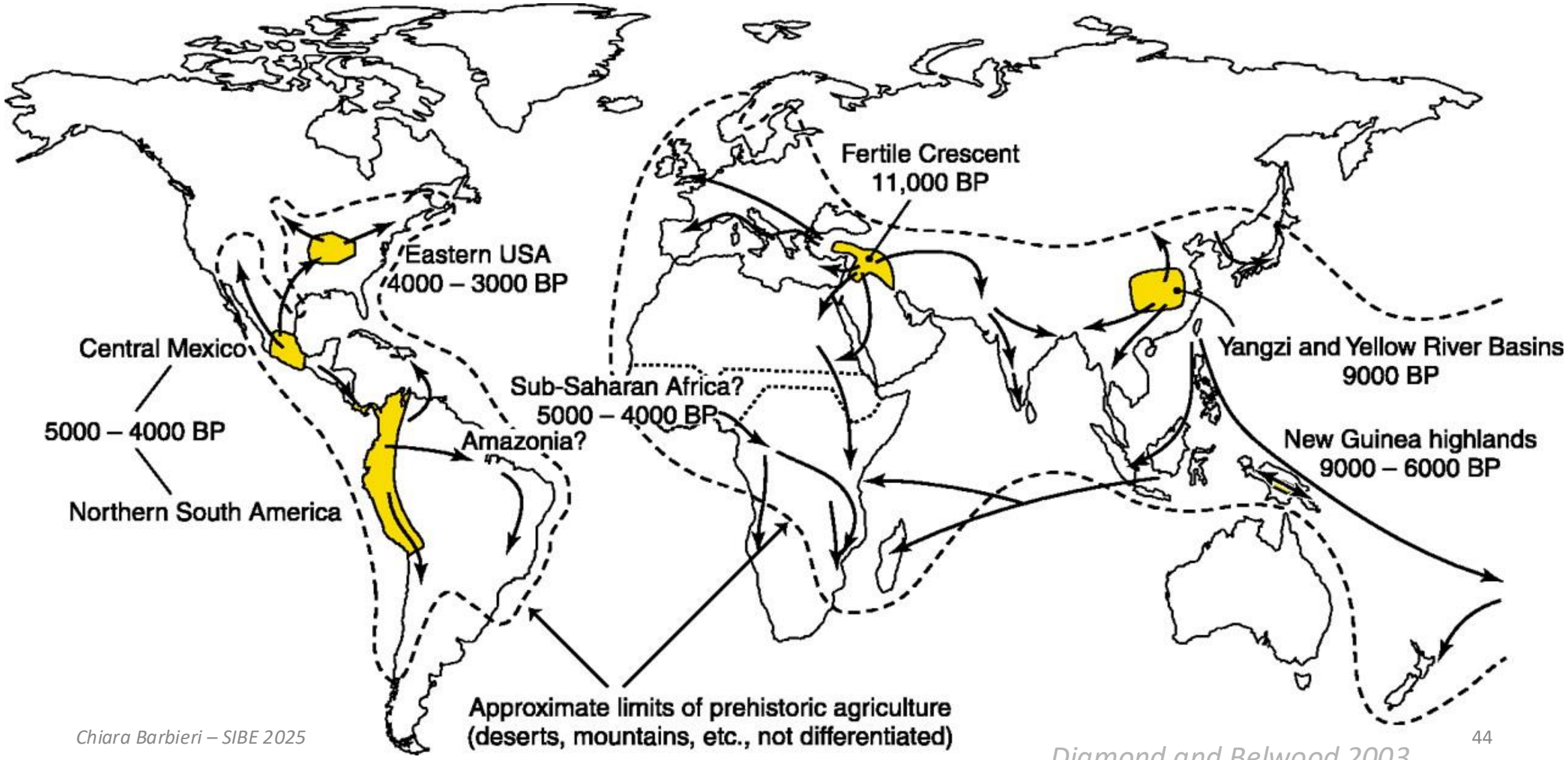


FIG. 1. Comparison of genetic tree and linguistic phyla. See text for details. (Ling.) indicates populations pooled on the basis of linguistic classification. The tree was constructed by average linkage analysis of Nei's genetic distances. Distances were calculated based on 120 allele frequencies from the following systems: *A1A2BO, MNS, RH, P, LU, K, FY, JK, DI, HP, TF, GC, LE, LP, PEPA, PEPB, PEPC, AG, HLAA* (12 alleles), *HLAB* (17 alleles), *PI, CP, ACP, PGD, PGM1, MDH, ADA, PTC, E1, SODA, GPT, PGK, C3, SE, ESD, GLO, KM, BF, LAD, E2, GM*, and *PG*.

Cavalli-Sforza et al. PNAS 1988

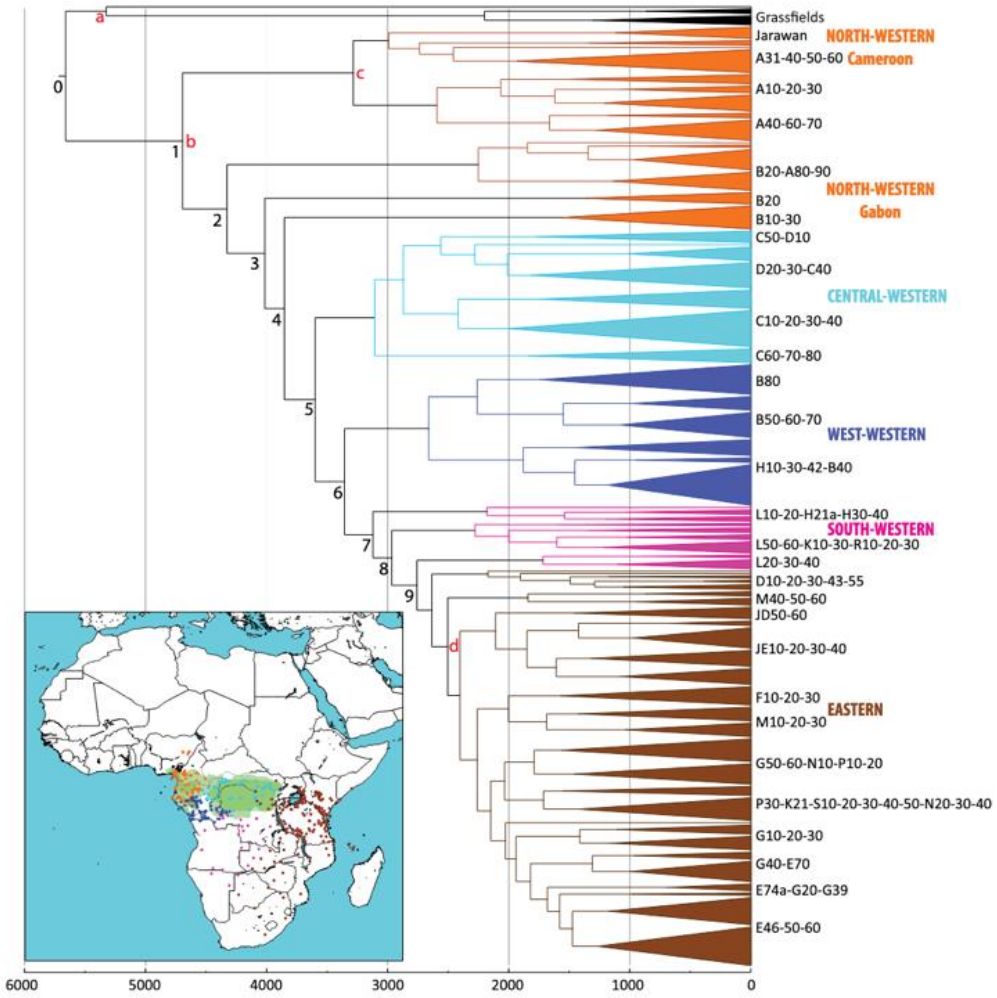
Language, farmers and expansions

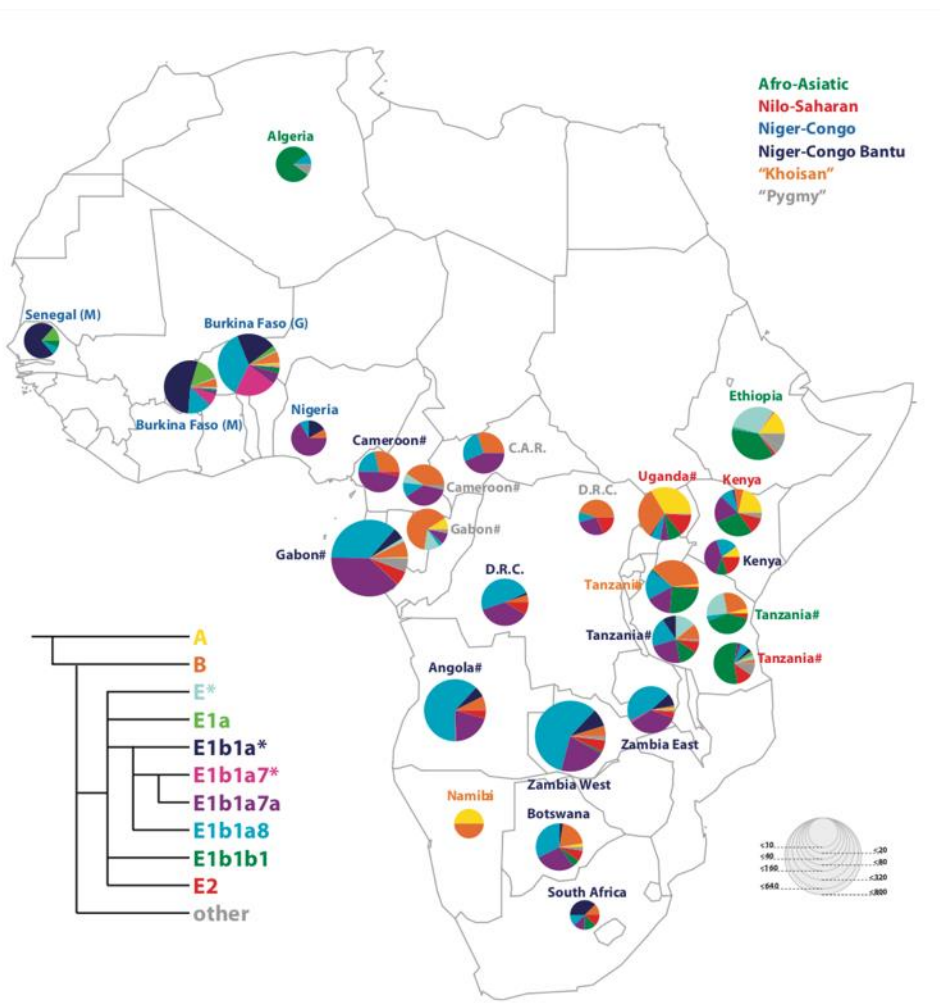
Large language families from centers of domestications



Diamond and Belwood 2003

Bantu language family spread with a demographic event



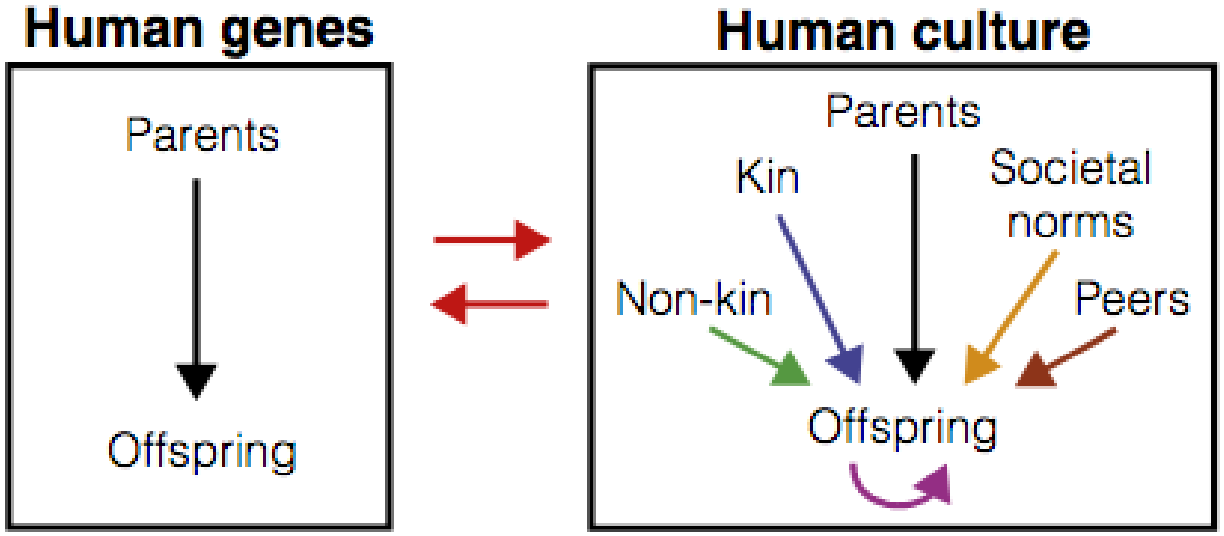


Grollemund et al. PNAS 2015

De Filippo, Barbieri* et al. MBE 2011*

Caveats for linguistic and genetic comparisons

- Spatial autocorrelation
- Evolutionary rates (languages evolve faster)
- Transmission modalities



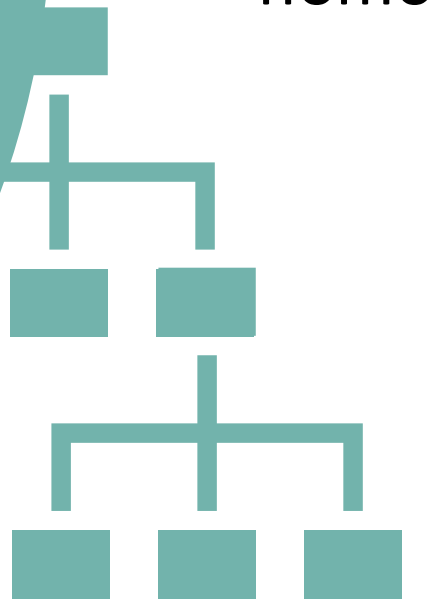
Creanza, Kolodny and Feldman - PNAS 2017

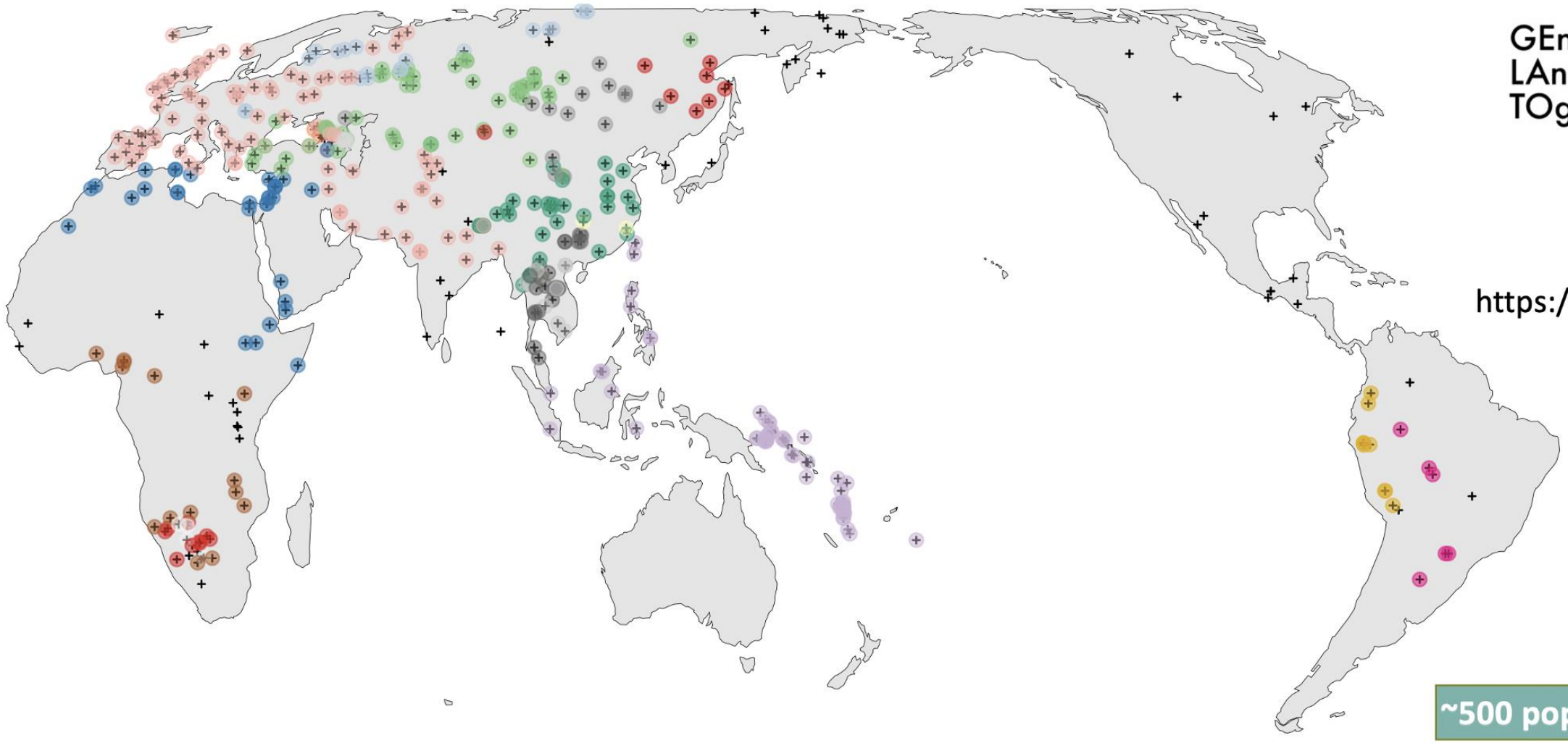
LANGUAGE SHIFT:
Adopt a new language for cultural exposure

Open questions

Systematic check for gene-language parallels

- Was Darwin right?
- How to define language shifts?
- Large language families are genetically homogeneous?





GEnes and LAnguages TOgether

https://gelato.clld.org/

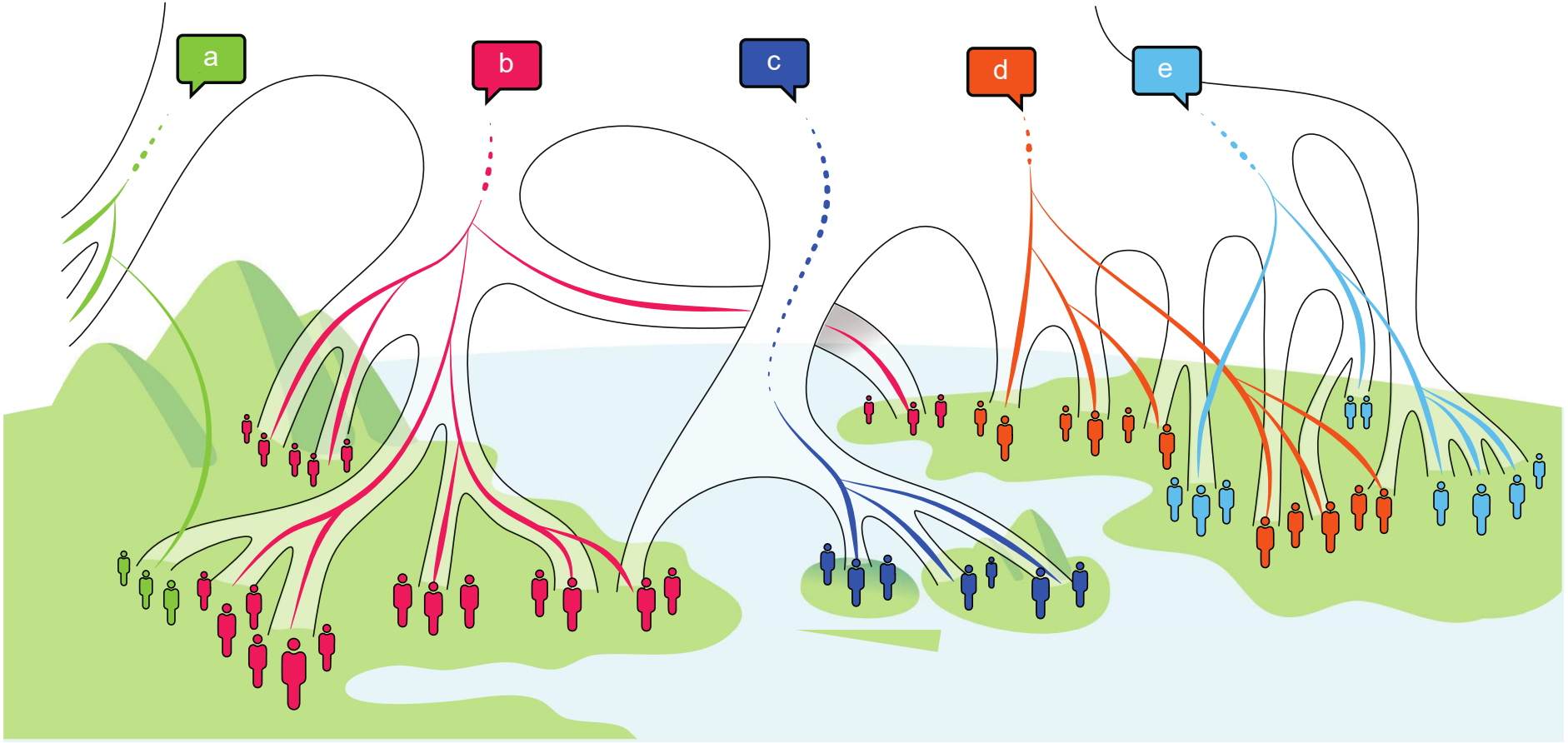
~500 populations

600,000 SNPs

Human Origins SNP chip

Abkhaz-Adyge (5)
Afro-Asiatic (33)
Atlantic-Congo (14)
Austroasiatic (14)
Austronesian (56)
Hmong-Mien (5)
Indo-European (93)
Khoe-Kwadi (9)
Kxa (5)
Mongolic-Khitan (16)
Nakh-Daghestanian (10)
Quechuan (9)
Sino-Tibetan (37)
Tai-Kadai (23)
Tungusic (8)
Tupian (6)
Turkic (56)
Uralic (20)

Testing Darwin's idea – with global data

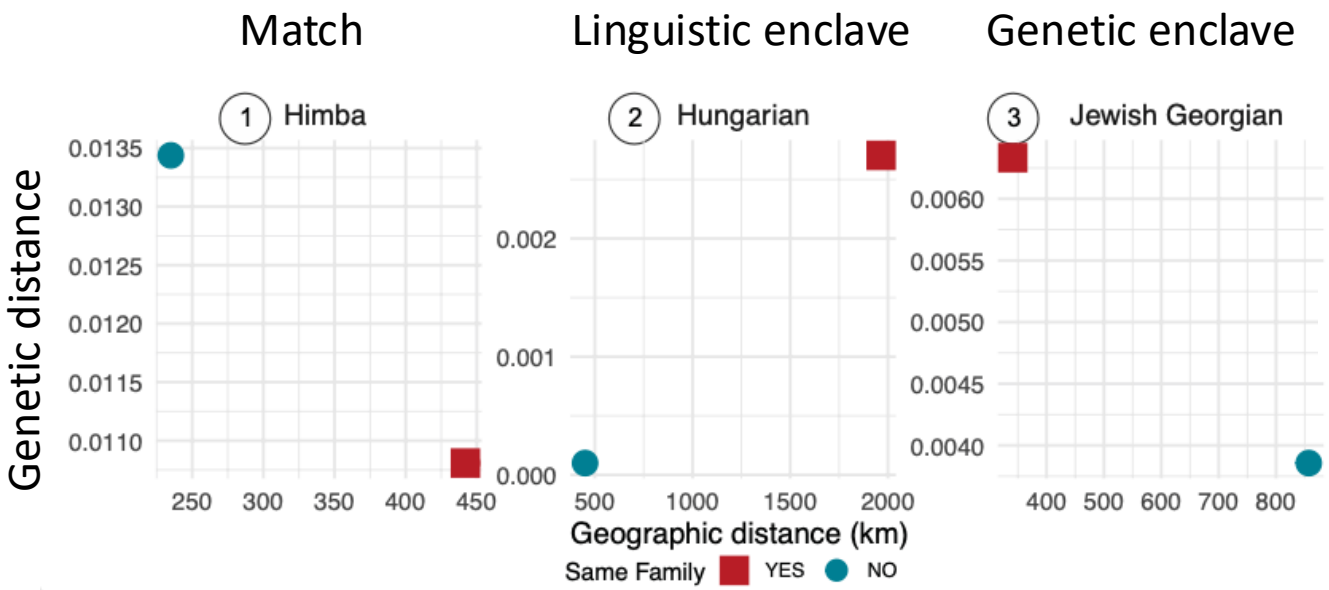


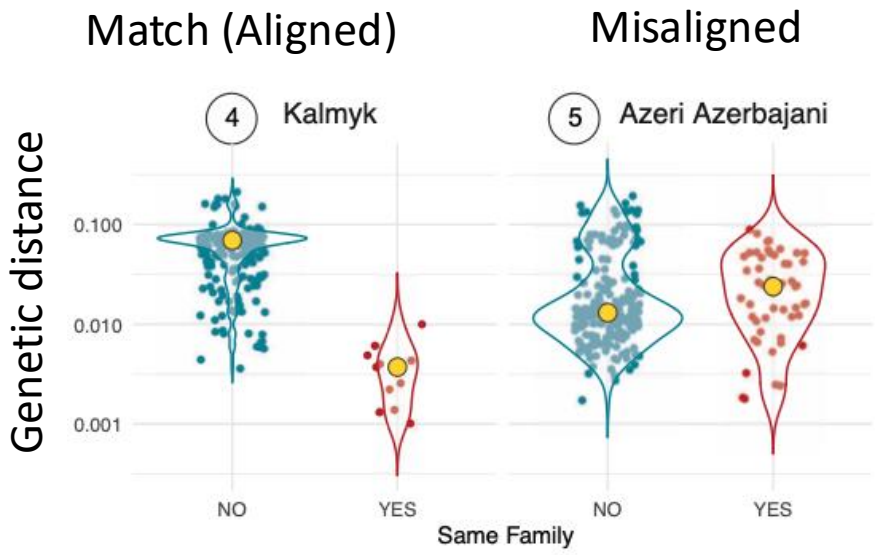
A global analysis of matches and mismatches between human genetic and linguistic histories

Chiara Barbieri[a,b,c,1,2], Damián E. Blasi[c,d,e,2], Epifanía Arango-Isaza[a,b], Alexandros G. Sotiropoulos[f], Harald Hammarström[g], Søren Wichmann[h], Simon J. Greenhill[c,i], Russell D. Gray[c], Robert Forkel[c,3], Balthasar Bickel[b,j,3], and Kentaro K. Shimizu[a,b,k,3]

PNAS 2022

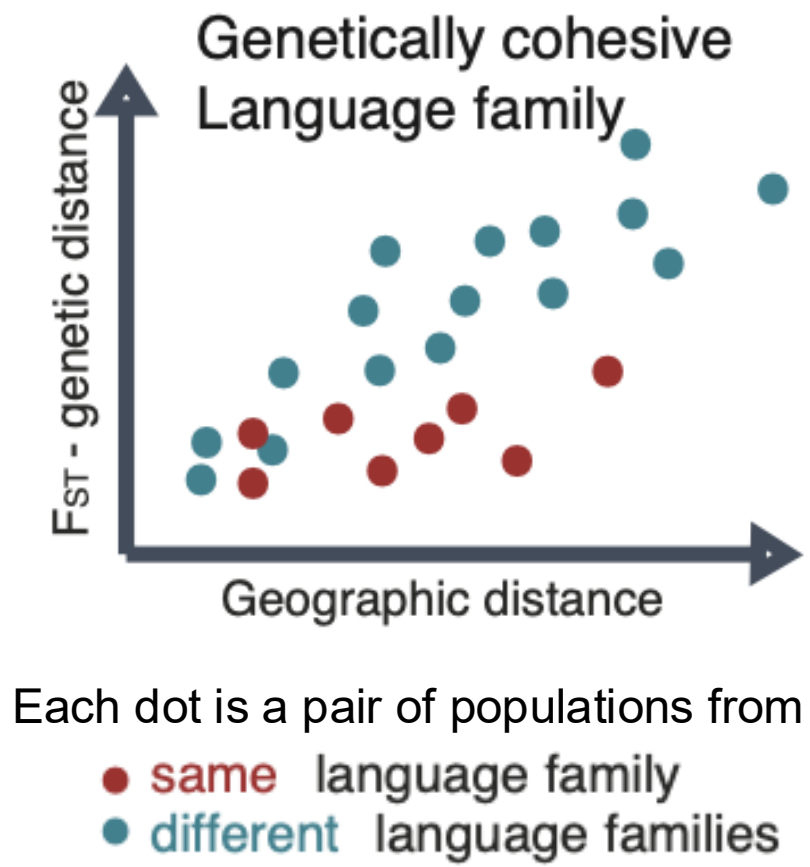
Systematic search for matches and mismatches





- Gene-language matches ✓
  - Trend for genetic cohesiveness (~50%) in speakers of related languages
- Language shift / mismatches ✓
  - 20% of populations in the dataset are genetically closer to speakers of distant languages

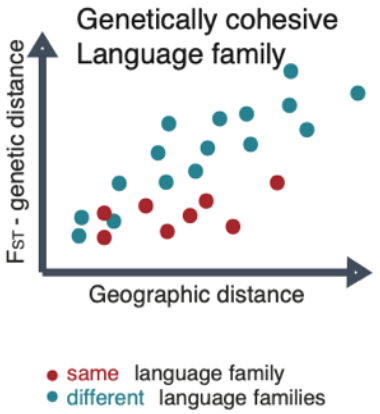
Correlation of pairwise distances using Isolation By Distance



For each language family, I take all the pairwise combinations with

- Two populations from the same target language family (red)
- One population from target language family and one population from a different language family, in geographic proximity (blue)

Genetic cohesiveness of language families



Genetically cohesive

Genetically non-cohesive

Genetic distance

Atlantic–Congo

Indo–European

Mongolic–Khitan

Sino–Tibetan

Afro–Asiatic

Nakh–Daghestanian

Tai–Kadai

Quechuan

Austroasiatic

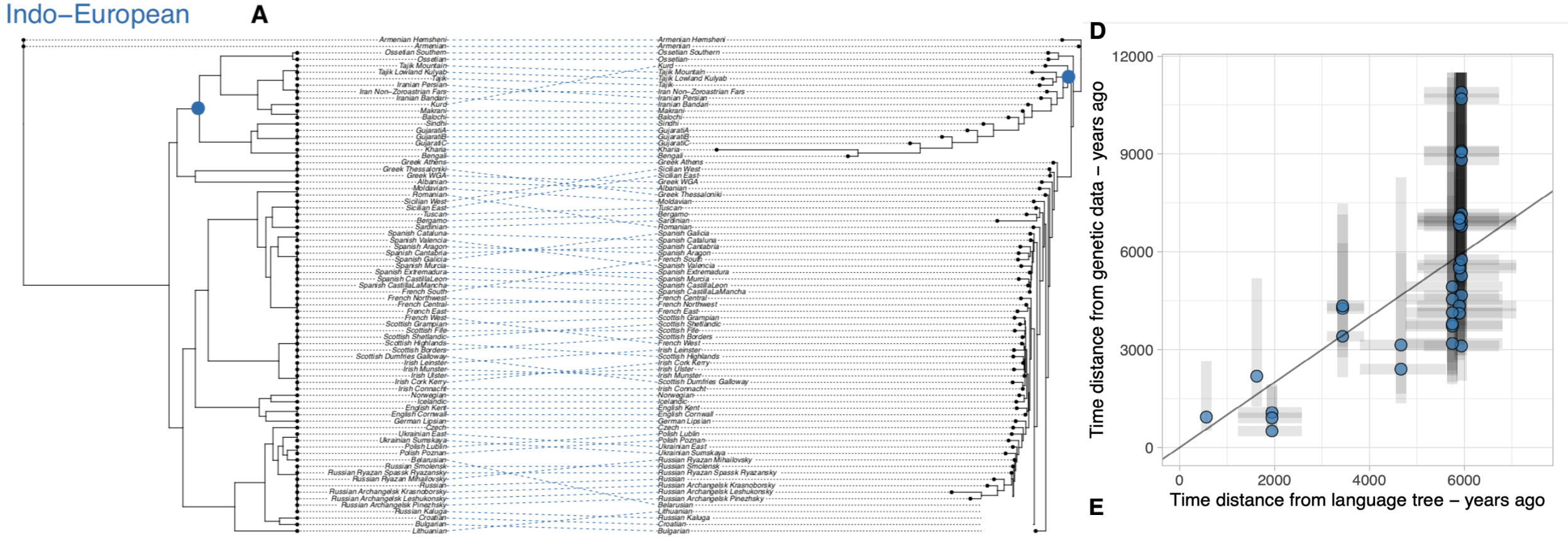
Turkic

Austronesian

Uralic

Geographic Distance – km

Divergence time: genetic and linguistic trees coincide also in time?



Language tree (Bouckaert et al. 2012)

Genetic tree

Testing gene-language correspondence systematically:

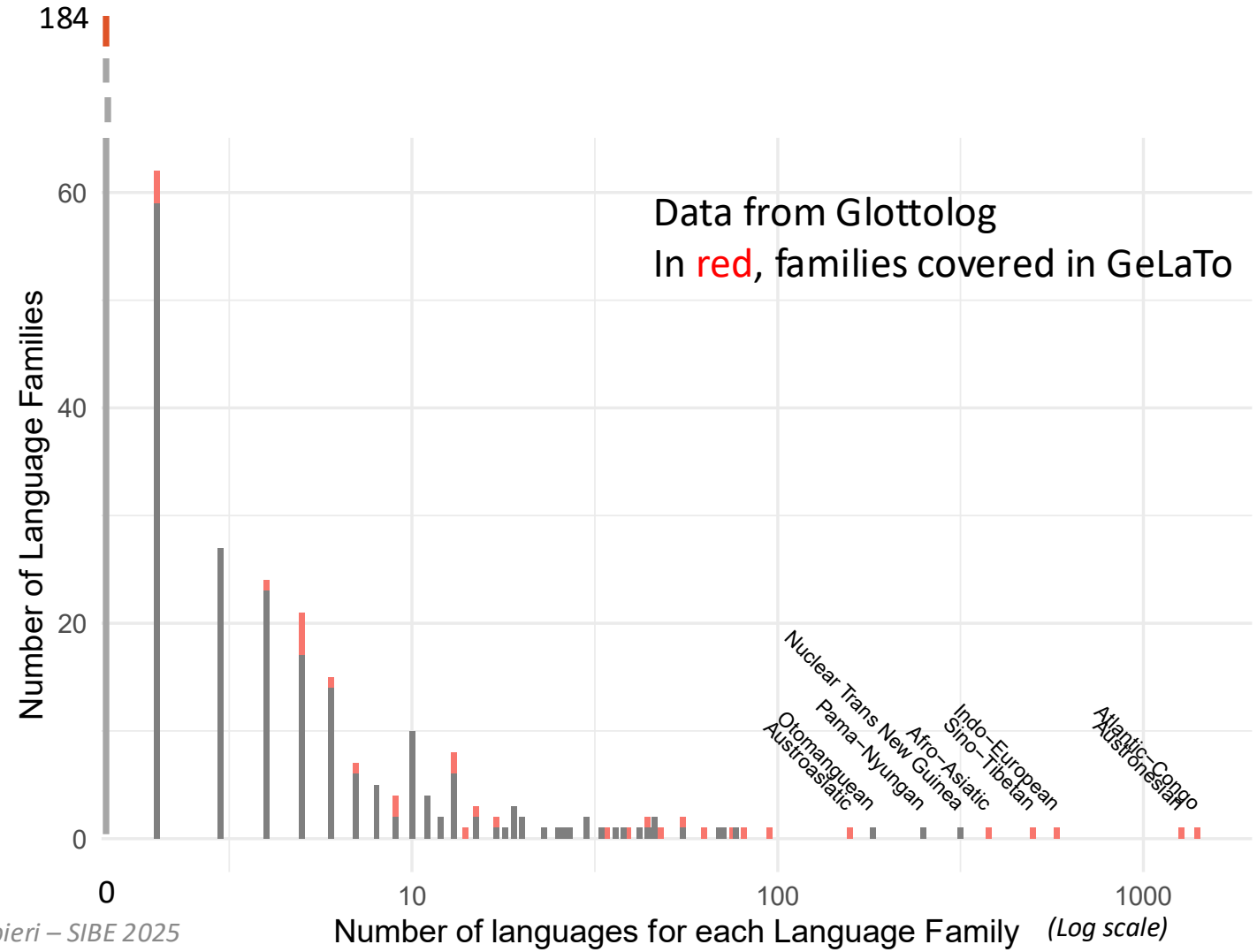
- Darwin was right? Most cases of gene-language match!
- Identification of language shifts, above spatial autocorrelation patterns

Language family size

Why some language families are large and others are small?

- Are language isolates spoken by genetically isolated populations?

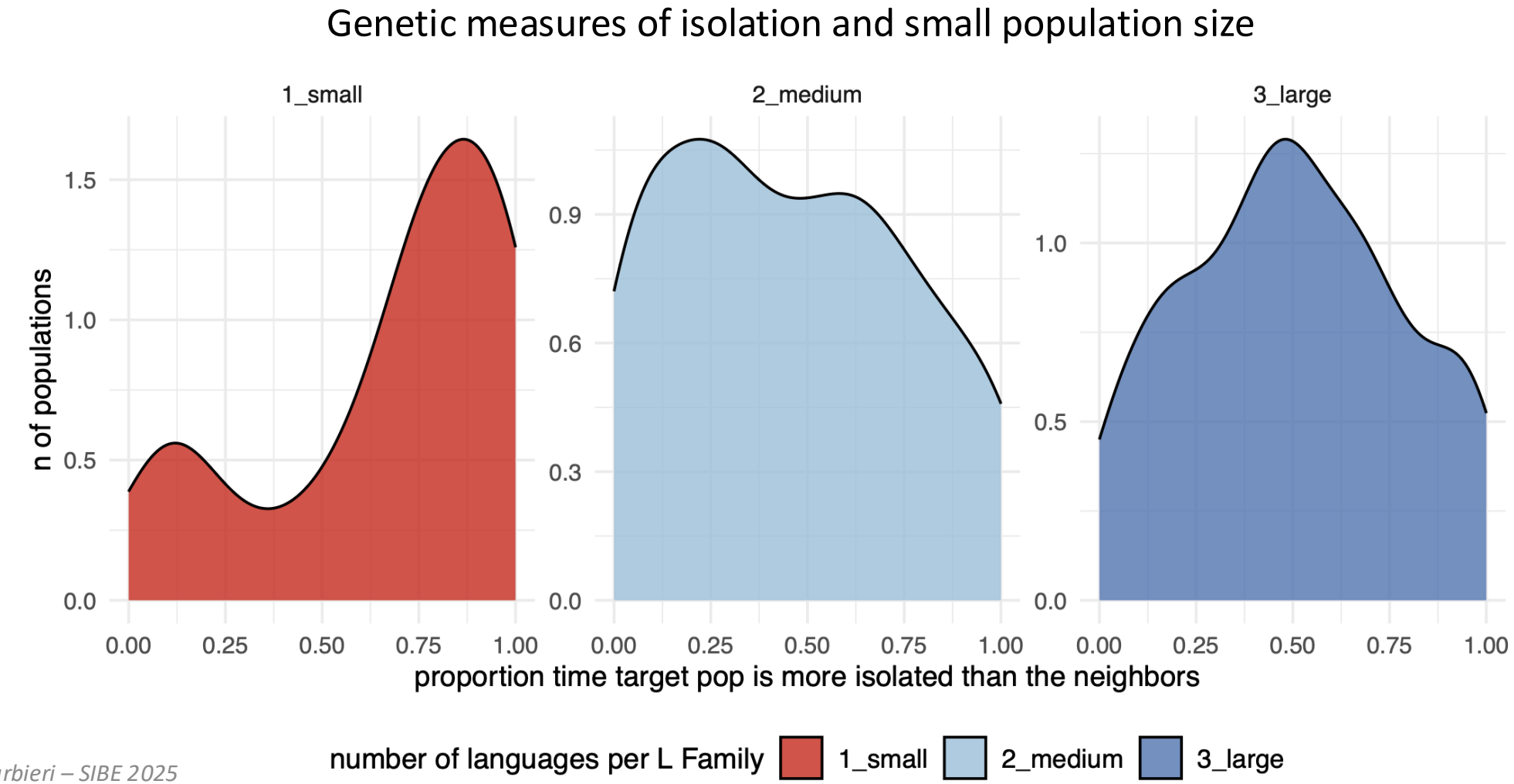
Large vs. small language families



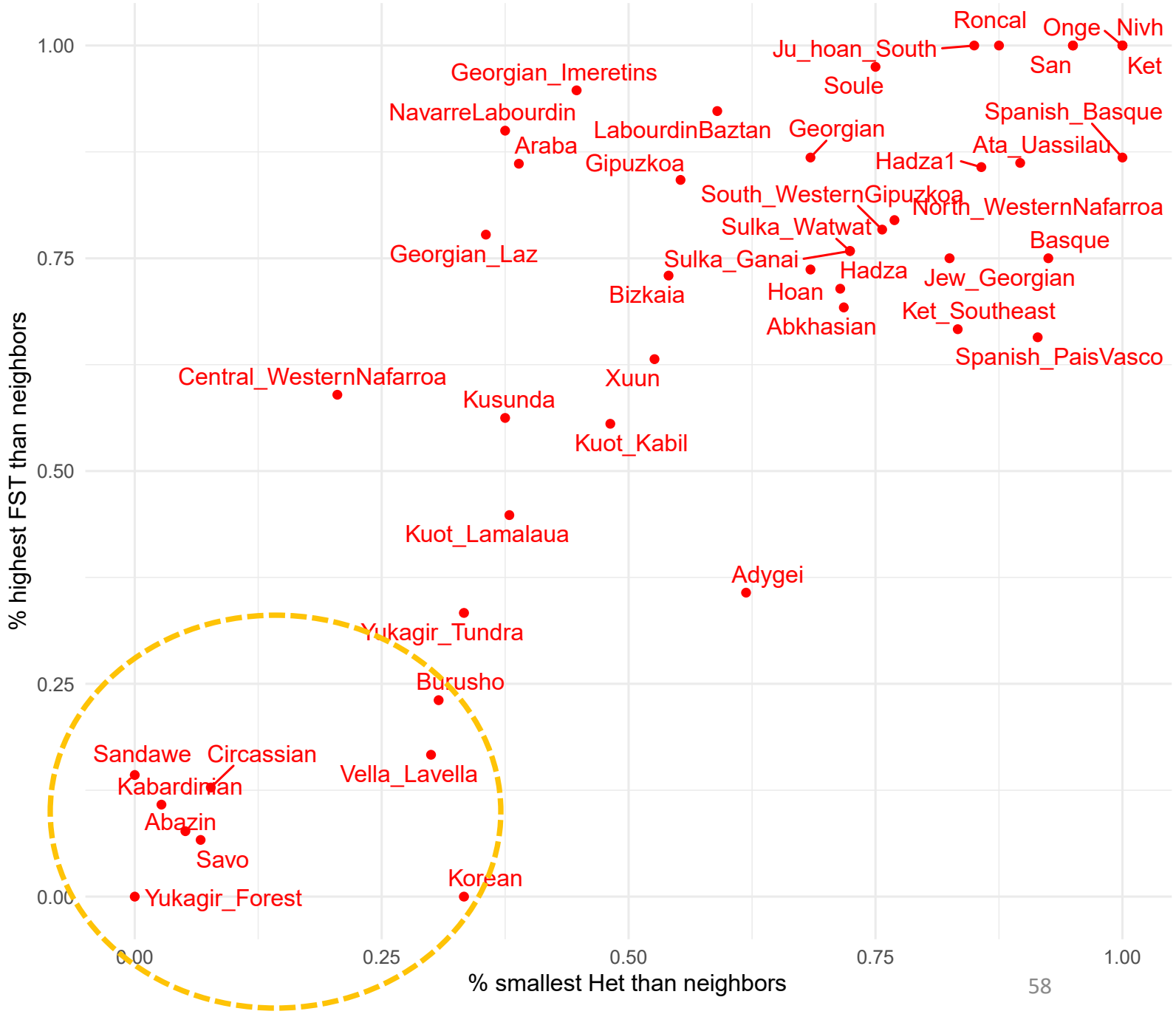
Which dynamics behind such large differences?

- Small language families like isolated relics?
- Large language families from population expansion, large population size?

Language isolates are spoken by genetically isolated populations



Why non-isolated populations speak isolated languages?



Linguistic and demographic contact

Contact as genetic admixture

- What happens to the languages? Do they become similar (borrowing) or diverge (schismogenesis)?

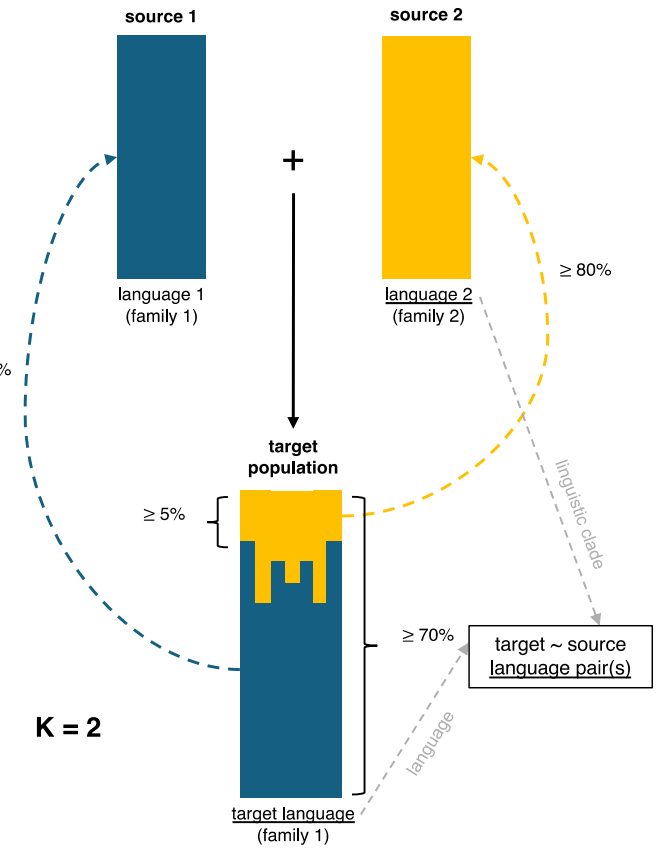
SCIENCE ADVANCES | RESEARCH ARTICLE

LINGUISTICS

Patterns of genetic admixture reveal similar rates of borrowing across diverse scenarios of language contact

Anna Graff[1,2]*, Damián E. Blasi[3,4], Erik J. Ringen[5], Vladimir Bajić[6], Daphné Bavelier[7], Kentaro K. Shimizu[1,2,8], Brigitte Pakendorf[9], Chiara Barbieri[2,10]*†, Balthasar Bickel[1]†

Genetic contact increases the probability of sharing linguistic features

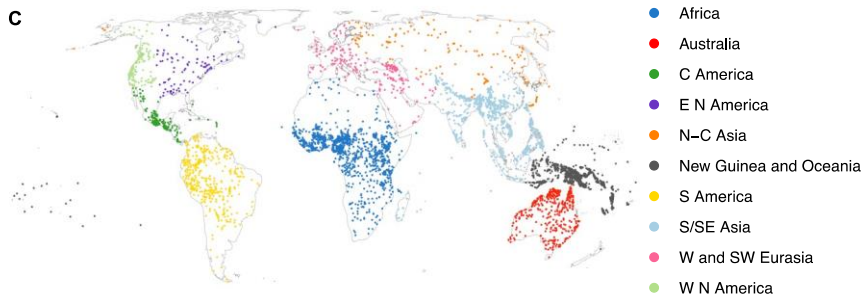


Genetic admixture

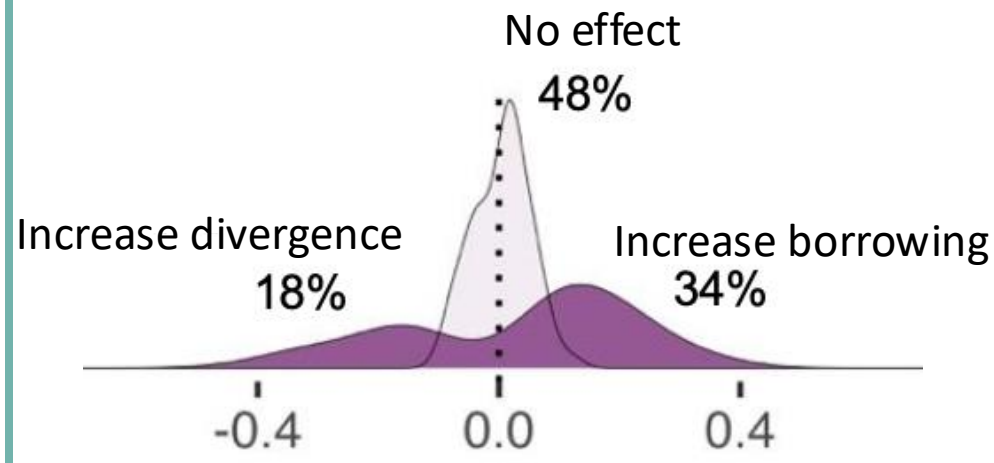
Genetic dataset

+

Linguistic dataset



Linguistic features breakdown in genetic contact



Posterior ΔP for the linguistic features

What's next?

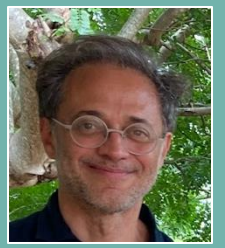
Population and individual relationships are linked to linguistic and cultural diversity

- Which role for environmental and climatic factors?
- Which scenario for linguistic/cultural loss of diversity?

Historical relationships

Structural diversity

Environment



THANK YOU FOR YOUR ATTENTION!

CREDITS:

University of Cagliari
Paolo Francalacci, Carla Caló, Laura Flore, Ignazio Pudzu, Nicoletta Puddu

University of Zurich
Kentaro Shimizu, Balthasar Bickel, Epifania Arango, Anna Graff, Simon Aeschbacher, Paul Widmer, Marcelo Sánchez, Gabriel Aguirre

Max Planck Inst. for Evolutionary Anthropology, Leipzig
Rodrigo Barquera, Russell Gray, Simon Greenhill, Robert Forkel, Mark Stoneking, Leonardo Arias

Pompeu Fabra University
Damián Blasi

University of Uppsala
Harald Hammarström

University of Kiel
Søren Wichmann

University of Lyon
Brigitte Pakendorf, Matthias Urban

University of São Paulo
Tábita Hünemeier

PUCP Lima
Paul Heggarty

Yale University
Serena Tucci

Swiss National Science Foundation

Latin-American Center Zurich

Wenner-Gren Foundation

nccr evolving language

URPP Evolution in Action


Saluti da Cagliari

Evolution of languages: Transmission with modification

- Evolutionary processes shaped by environment (society) and biology/genetics, influencing each others

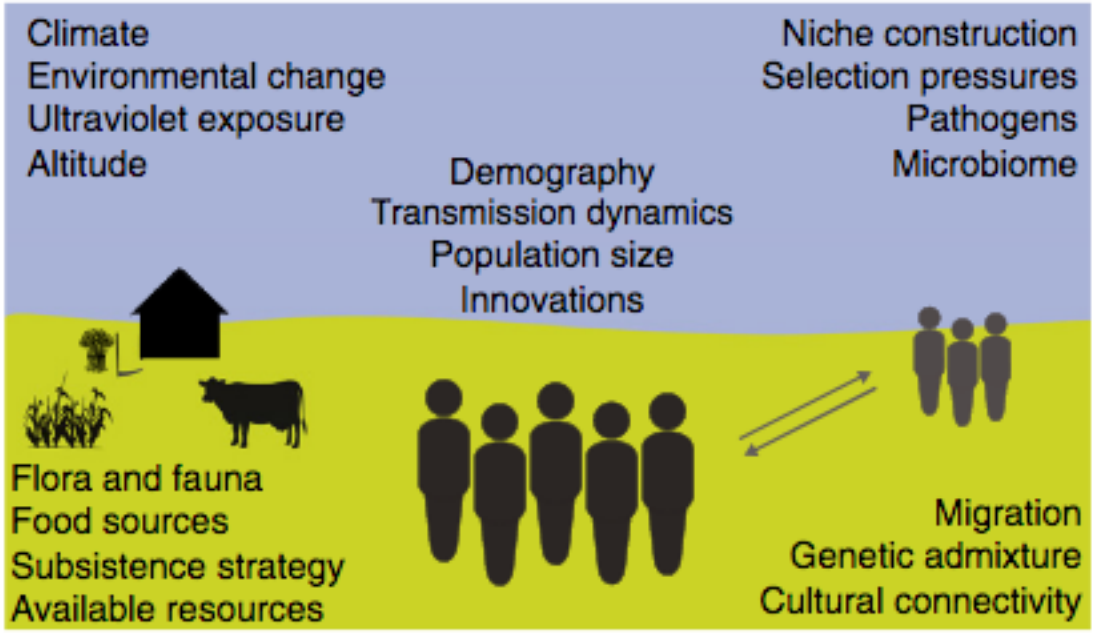
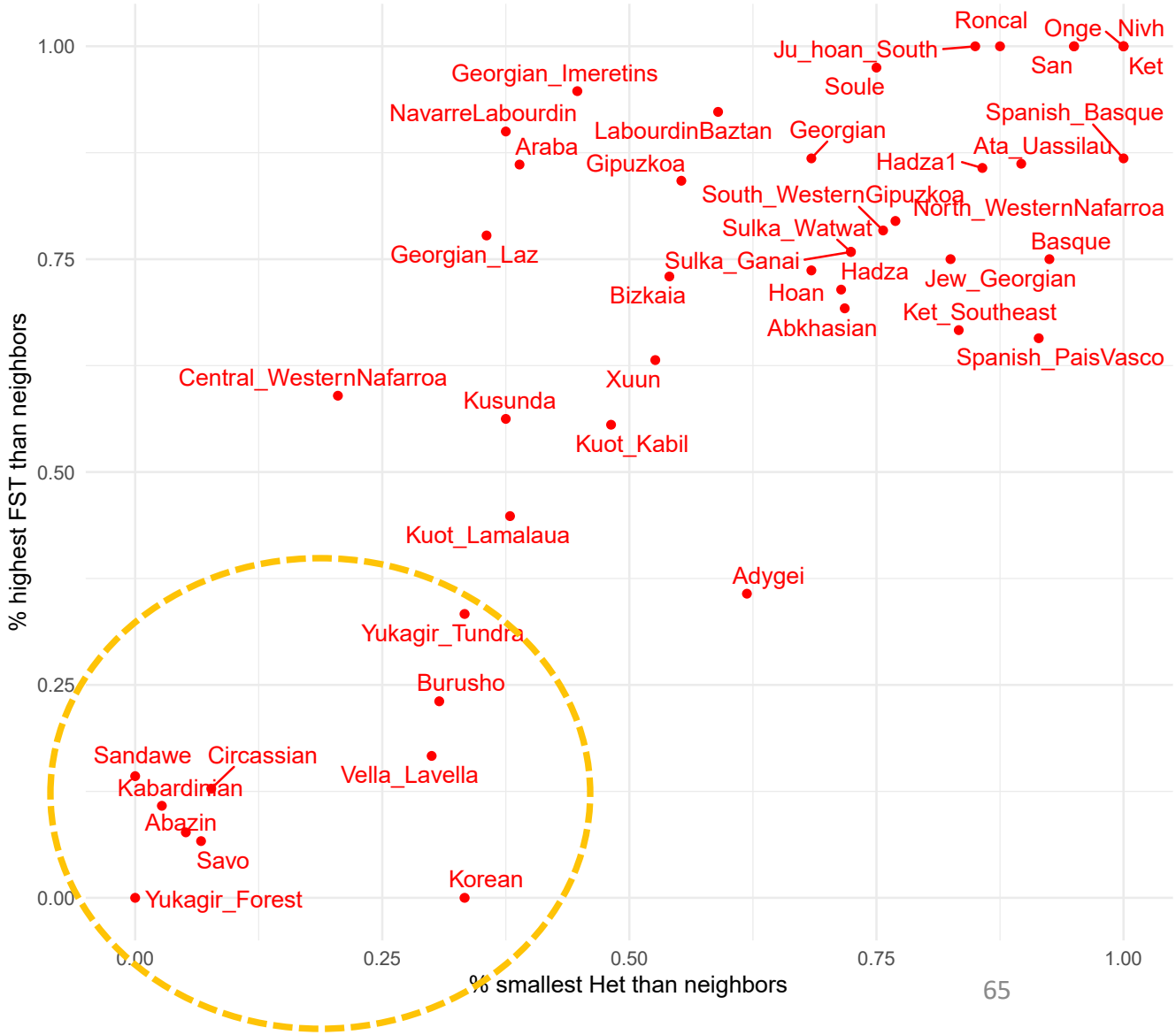


Fig. 2. Cultural, genetic, and environmental factors influencing evolution.

Creanza, Kolodny and Feldman - PNAS 2017

Why non-isolated populations speak isolated languages?

- Lack of representative neighbors to establish the baseline (*Siberia*)
- Isolated language spoken by large, high status groups (*Korean*)
- Diverse, structured population maintains isolated language (*Sandawe, Pacific*)



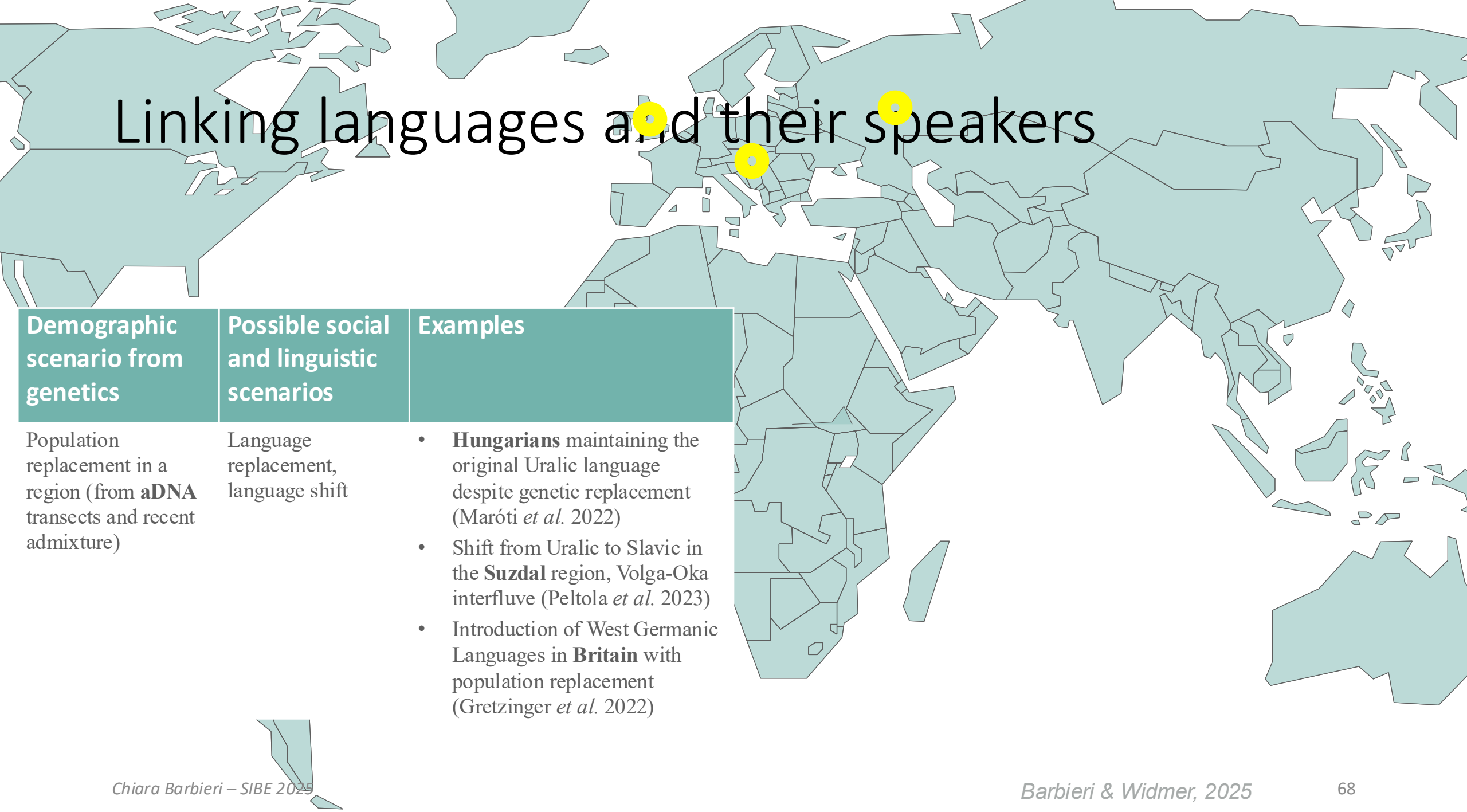
Linking languages and their speakers

Demographic scenario from genetics	Possible social and linguistic scenarios	Examples
Genetic homogeneity between populations	• Linguistic Areas • Phylogenetic relatedness within families	• Western Eurasia (Lao et al. 2008; Haspelmath 2001) • Central Andes (Urban & Barbieri 2021)

Linking languages and their speakers

Demographic scenario from genetics	Possible social and linguistic scenarios	Examples
Contact between populations	• Language contact • Lexical or Pattern borrowing	• Shared constructional patterns in the upper Northern Amazon (van Gijn *et al.* 2023) • Sharing of clicks between Khoisan and Bantu groups (Barbieri *et al.* 2013)

Linking languages and their speakers

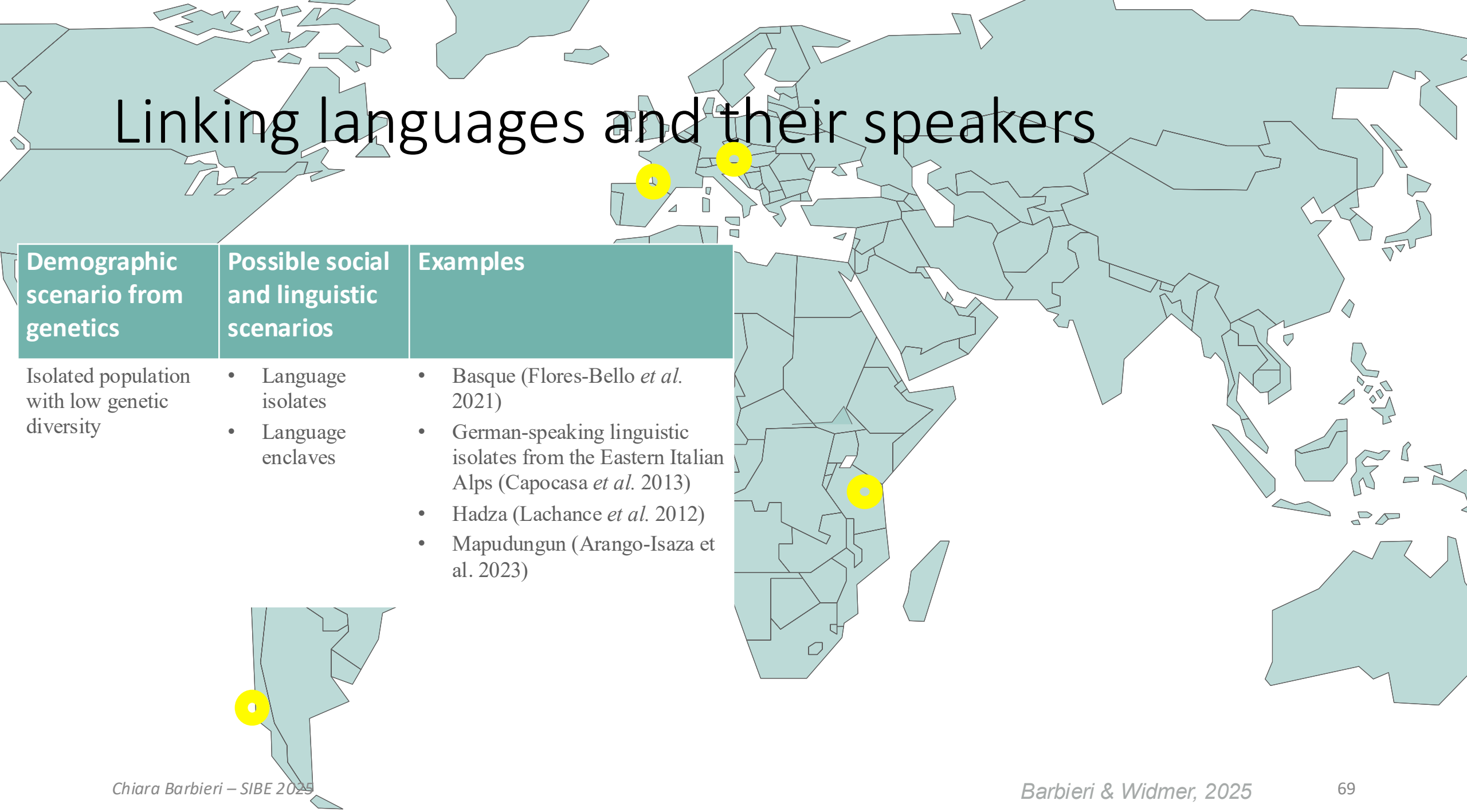


Demographic scenario from genetics	Possible social and linguistic scenarios	Examples
Population replacement in a region (from **aDNA** transects and recent admixture)	Language replacement, language shift	• **Hungarians** maintaining the original Uralic language despite genetic replacement (Maróti *et al.* 2022) • Shift from Uralic to Slavic in the **Suzdal** region, Volga-Oka interfluve (Peltola *et al.* 2023) • Introduction of West Germanic Languages in **Britain** with population replacement (Gretzinger *et al.* 2022)

Barbieri & Widmer, 2025

Linking languages and their speakers

Demographic scenario from genetics	Possible social and linguistic scenarios	Examples
Isolated population with low genetic diversity	• Language isolates • Language enclaves	• Basque (Flores-Bello *et al.* 2021) • German-speaking linguistic isolates from the Eastern Italian Alps (Capocasa *et al.* 2013) • Hadza (Lachance *et al.* 2012) • Mapudungun (Arango-Isaza et al. 2023)



Genes + Languages + Culture

Check for systematic culture-gene-language parallels

- Songs as case study for non-material culture

Cultural evolution of music: matching languages or genes?

- Global musical dataset: 5242 songs, 719 societies
- Musical style diversity has tree-like structure
- Musical patterns related to linguistic and genetic histories in Southeast Asia and sub-Saharan Africa

Article | Open access | Published: 10 May 2024

Global musical diversity is largely independent of linguistic and genetic histories

Sam Passmore, Anna L. C. Wood, Chiara Barbieri, Dor Shilton, Hideo Daikoku, Quentin D. Atkinson & Patrick E. Savage

Nature Communications 15, Article number: 3964 (2024) | Cite this article

GEnes and LAnguages TOgether