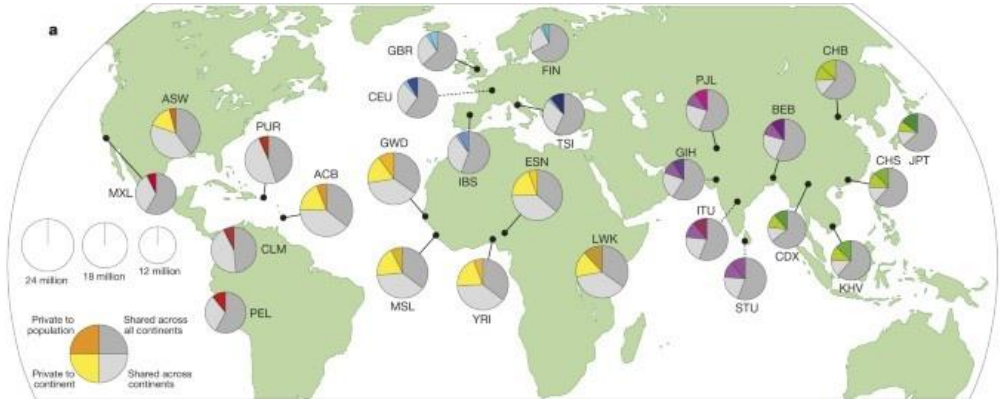
Genética de poblaciones 2

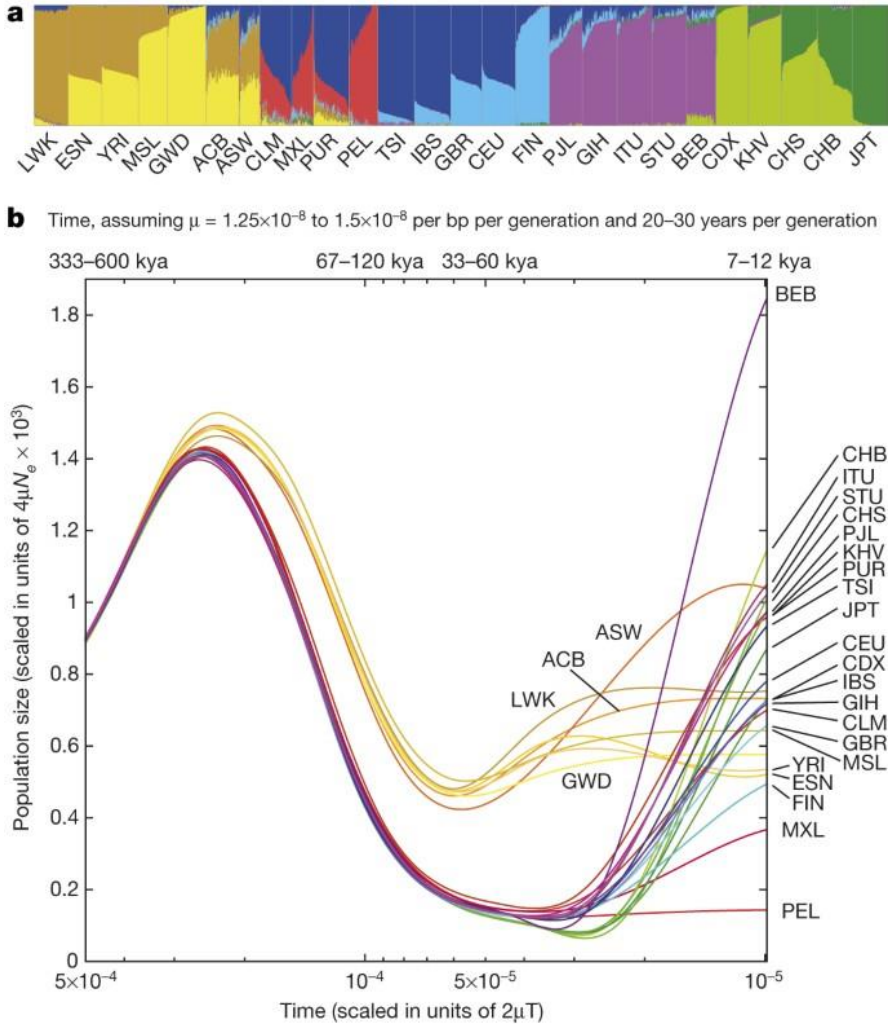
F-statistics: $f3$, $f4$, qpWave y qpAdm

Constanza de la Fuente Castro
Profesora Asistente ICBM
Facultad de Medicina
Universidad de Chile

Genómica de poblaciones

- Analizar patrones de variación genética a nivel genómico, intra y entre poblaciones.
- Estudiar los procesos evolutivos involucrados en estos patrones
- Bases de datos con múltiples variantes (array o secuenciación genómica).
- Enfocado en variación (variantes con frecuencia >= 1%)



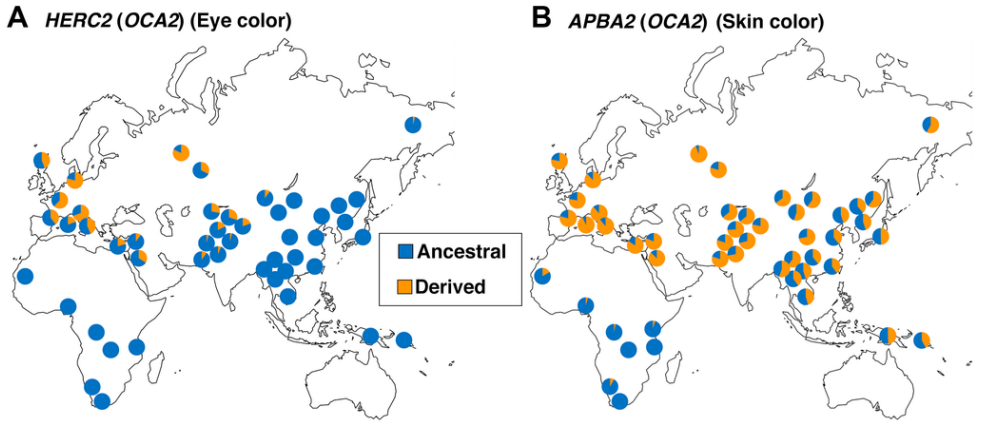


https://doi.org/10.1038/nature15393

Datos genómicos

ARRAY

SECUENCIACION GENOMICA

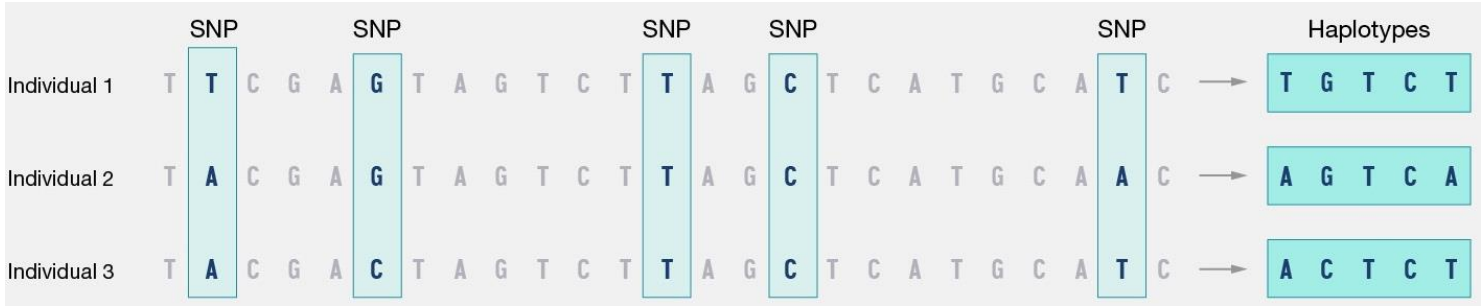


Datos genómicos

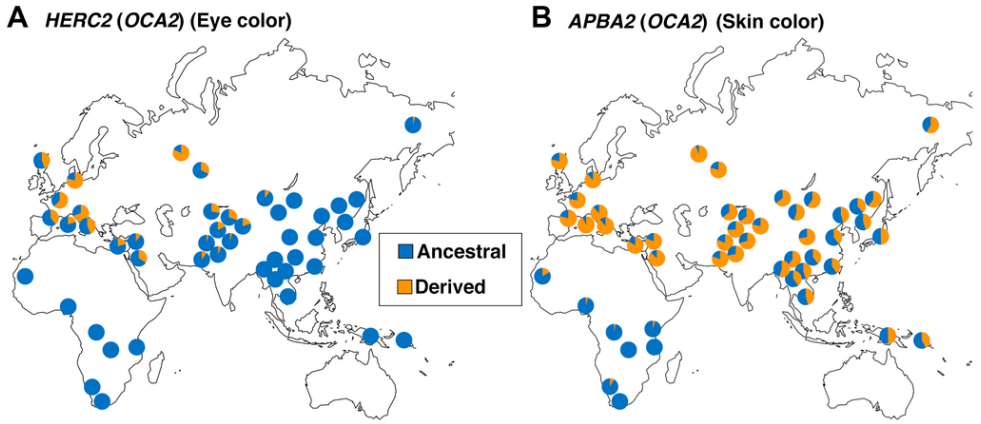
- Frecuencias alélicas
- Haplotipos

→ Patrones de variación

- Deriva genética
- Demografía
- Selección natural



https://doi.org/10.1371/journal.pgen.1003372
https://www.genome.gov/es/genetics-glossary/Haplotipo

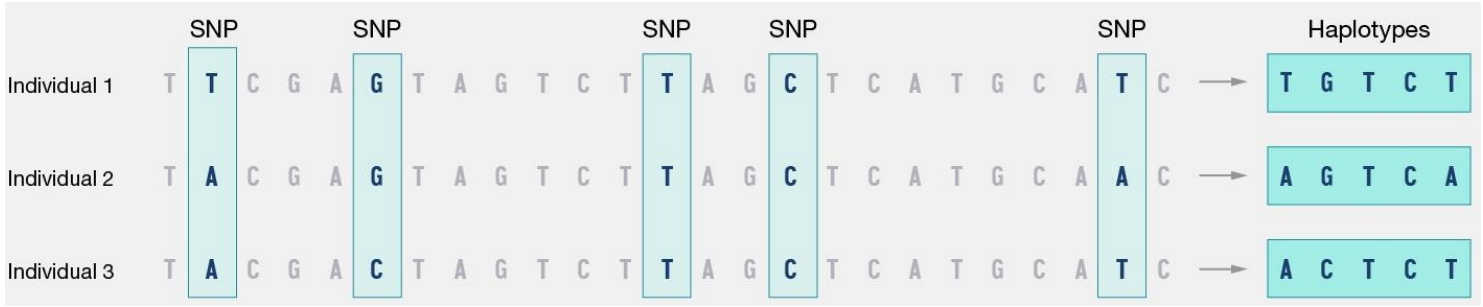


Datos genómicos

- Frecuencias alélicas
- Haplotipos

→ Patrones de variación

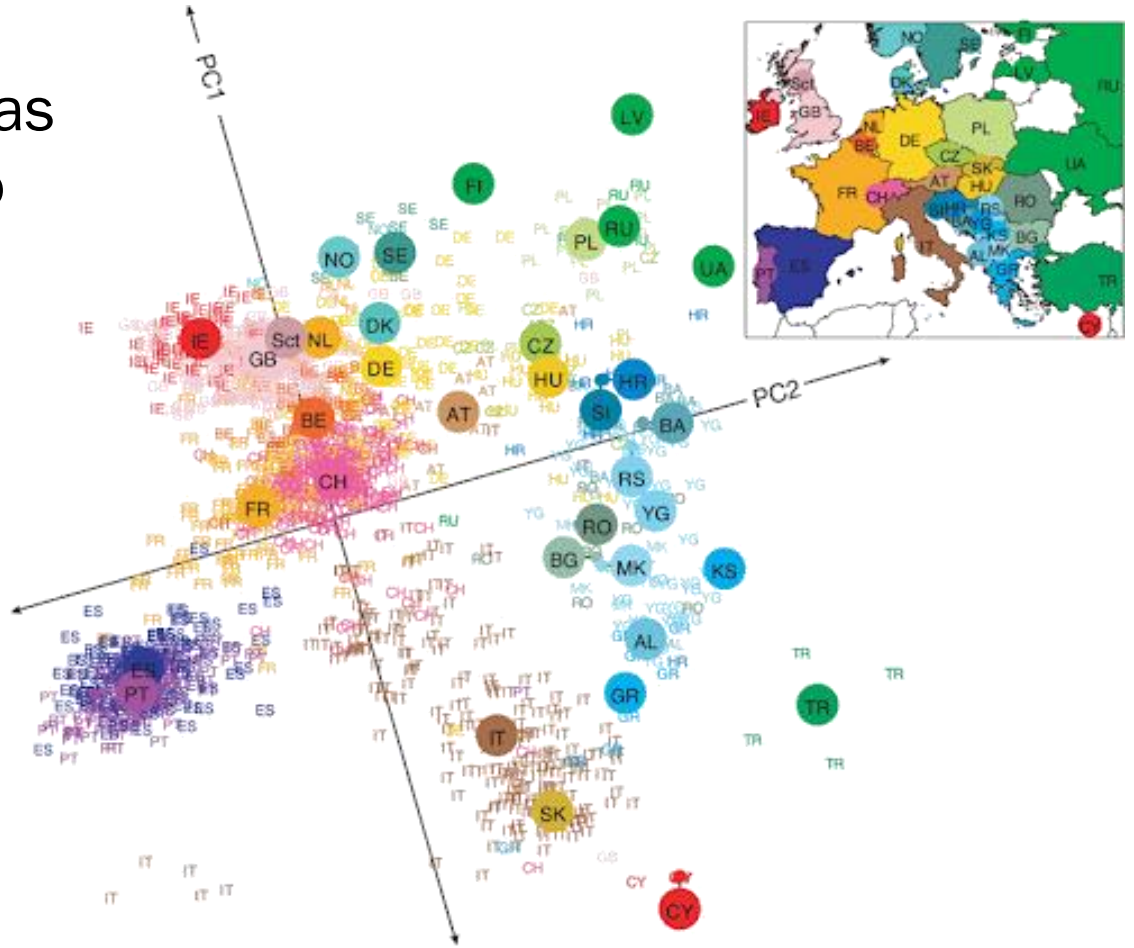
- Deriva genética
- Demografía
- Selección natural



https://doi.org/10.1371/journal.pgen.1003372
https://www.genome.gov/es/genetics-glossary/Haplotipo

Aproximaciones en Genómica de Poblaciones

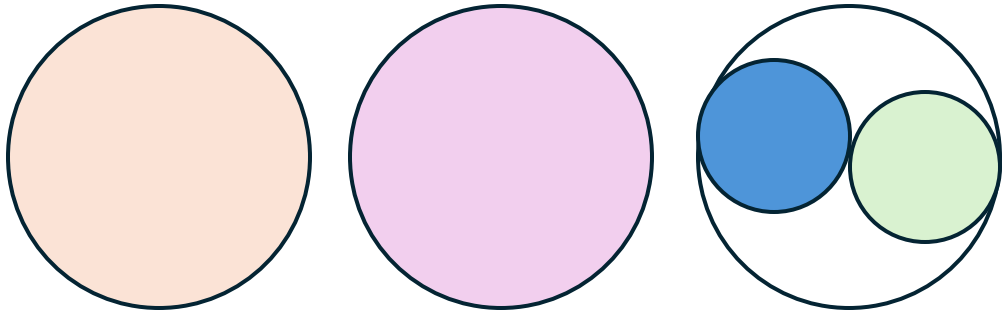
Estructura poblacional: diferencias en frecuencias alélicas o haplotípicas entre distintos grupos o poblaciones en una especie

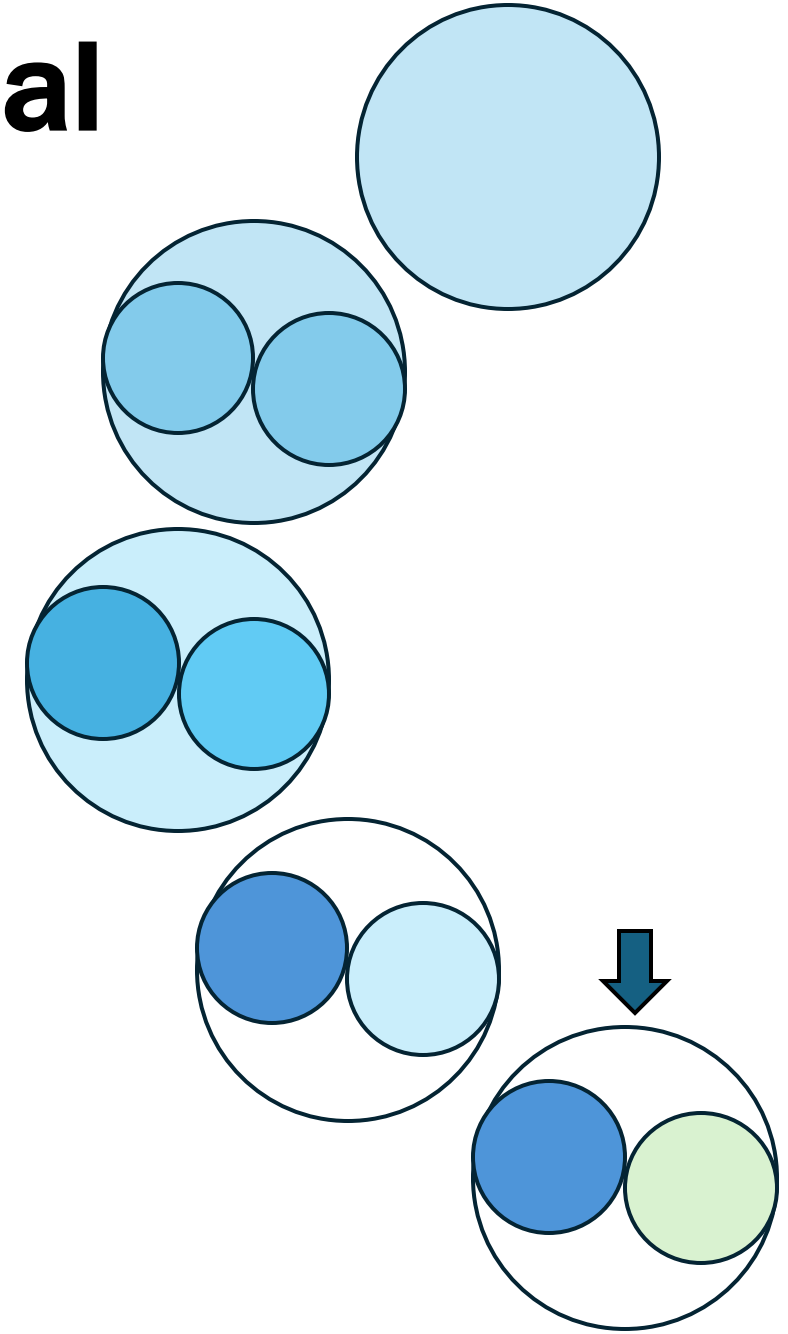


https://doi.org/10.1038/nature07331

Estructura poblacional

- Barreras al flujo génico (geográficas, culturales, etc) conducirán a variaciones en frecuencias alélicas entre grupos.
- Diversos análisis, la mayoría basados en técnicas de agrupamiento o clustering = agrupar objetos según similitud (en este caso, genética).
- No es fijo, ni temporal ni geográficamente. Puede variar según nuestra escala de análisis.





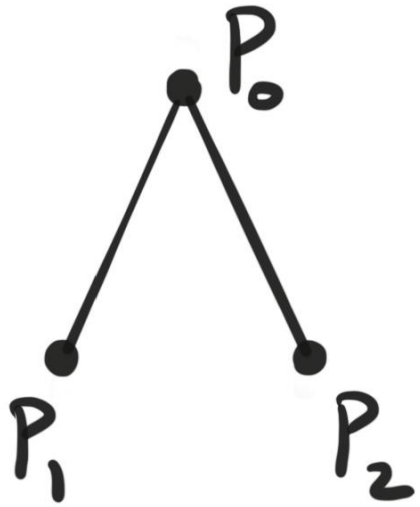
Estadísticos F: patrones de deriva genética compartida entre poblaciones

- Deriva genética compartida ~ historia evolutiva compartida
- Estadísticos basados en correlaciones de frecuencias alélicas entre poblaciones (2, 3 o 4)
- Variadas aplicaciones e interpretaciones
  - Evaluar relación genética entre poblaciones (2, 3, 4 o más)
  - Evaluar mestizaje genético (fuentes/origen y proporciones)
  - Modelar relación entre poblaciones (gráficos de mestizaje)

Filogenias  Admixture/Mestizaje

Estadísticos F: patrones de deriva genética compartida entre poblaciones

Estadísticos basados en correlaciones de frecuencias alélicas entre poblaciones (2, 3 o 4)

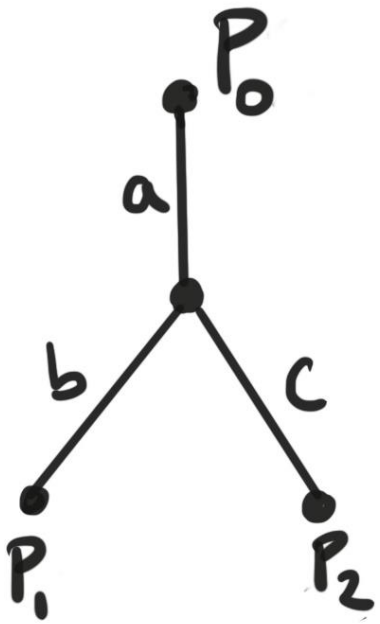


p0: frecuencia de un alelo en población ancestral
p1: frecuencia en Población 1
p2: frecuencia en Población 2

La deriva genética es independiente en ambas ramas
Por lo tanto, la covarianza es cero

Estadísticos F: patrones de deriva genética compartida entre poblaciones

Estadísticos basados en correlaciones de frecuencias alélicas entre poblaciones (2, 3 o 4)



p0: frecuencia de un alelo en población ancestral
p1: frecuencia en Población 1
p2: frecuencia en Población 2

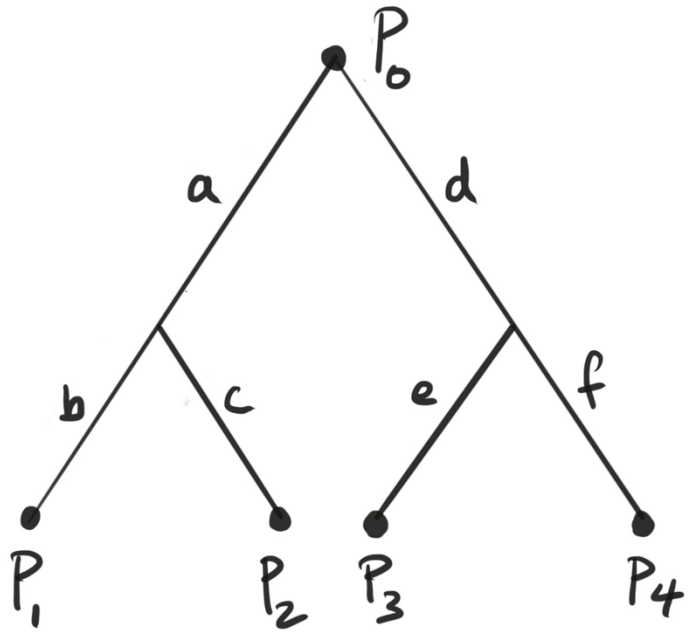
La deriva genética es independiente en ambas ramas
Por lo tanto, la covarianza es cero.

¿Qué ocurre con rama compartida?
p0-p1 = a + b
p0-p2 = a +c
p1-p2 = a

La covarianza estará dada por la suma del largo de las ramas compartidas (e.g., a entre P1 y P2)

Estadísticos F: patrones de deriva genética compartida entre poblaciones

Estadísticos basados en correlaciones de frecuencias alélicas entre poblaciones (2, 3 o 4)



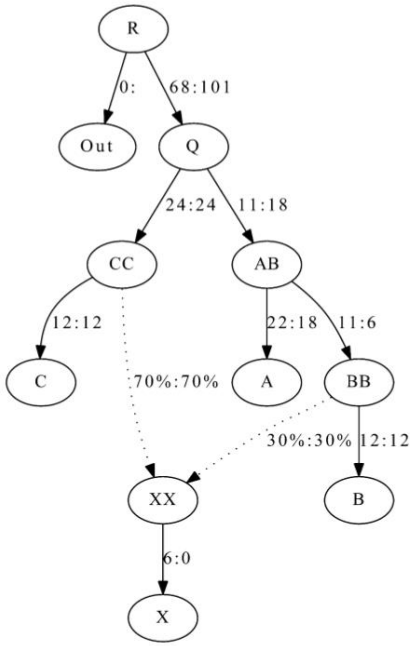
Matriz de covarianza de árbol

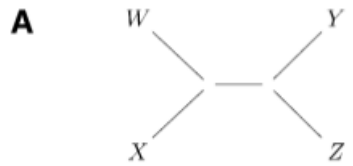
	P_1-P_0	P_2-P_0	P_3-P_0	P_4-P_0
P_1-P_0	$a+b$	a	0	0
P_2-P_0	a	$a+c$	0	0
P_3-P_0	0	0	$d+e$	d
P_4-P_0	0	0	d	$d+f$

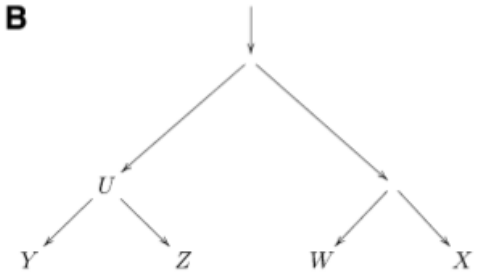
Estadísticos F: patrones de deriva genética compartida entre poblaciones

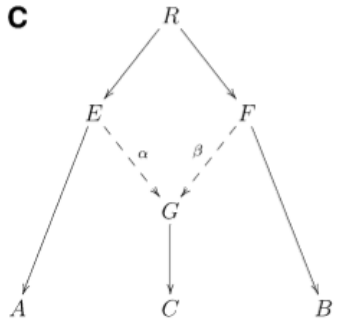
Métodos más comunes:

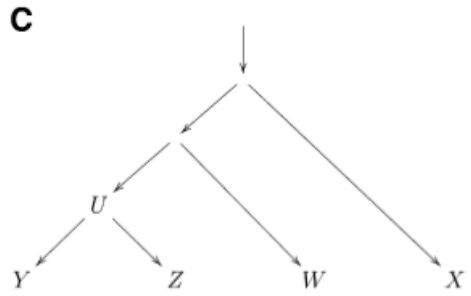
- $f3$
- $f4$ o D-statistic (o ABBA/BABA)
- qpWave – qpAdm
- qpGraph

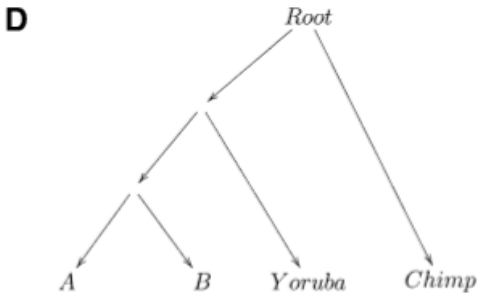








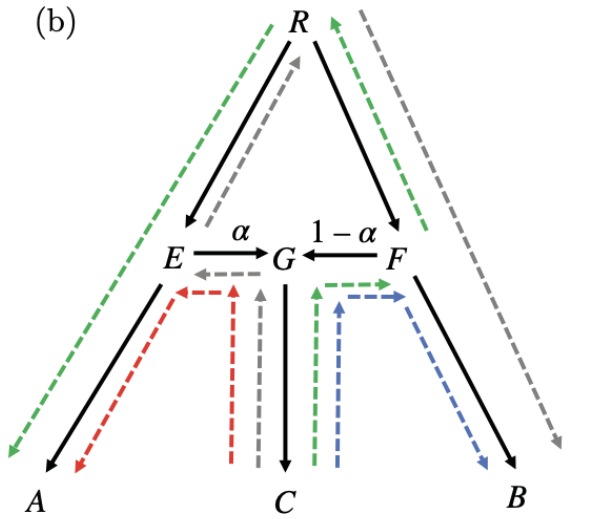




*f*3

admix

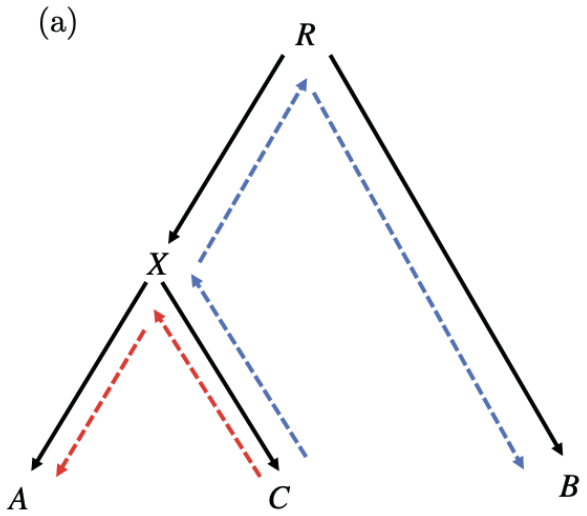
- C grupo target, A y B posibles fuentes de mestizaje (no directo)
- Valores positivos: no hay evidencias de mestizaje
- Valores negativos: evidencia de mestizaje



C is an admixed population related to *A* and *B*

outgroup

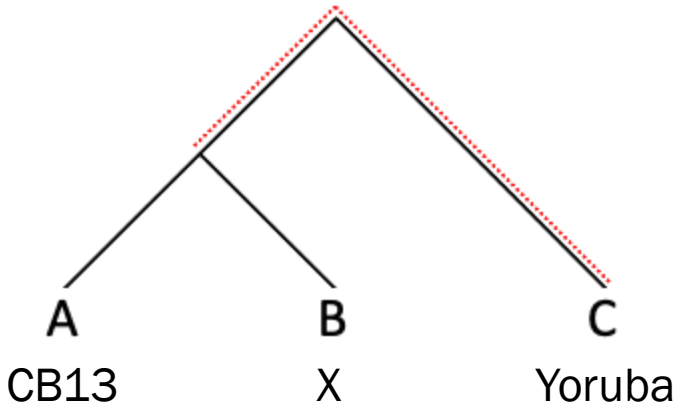
- B es outgroup a A y C
- Sólo valores positivos
- Deriva genética compartida entre A y C desde separación de B



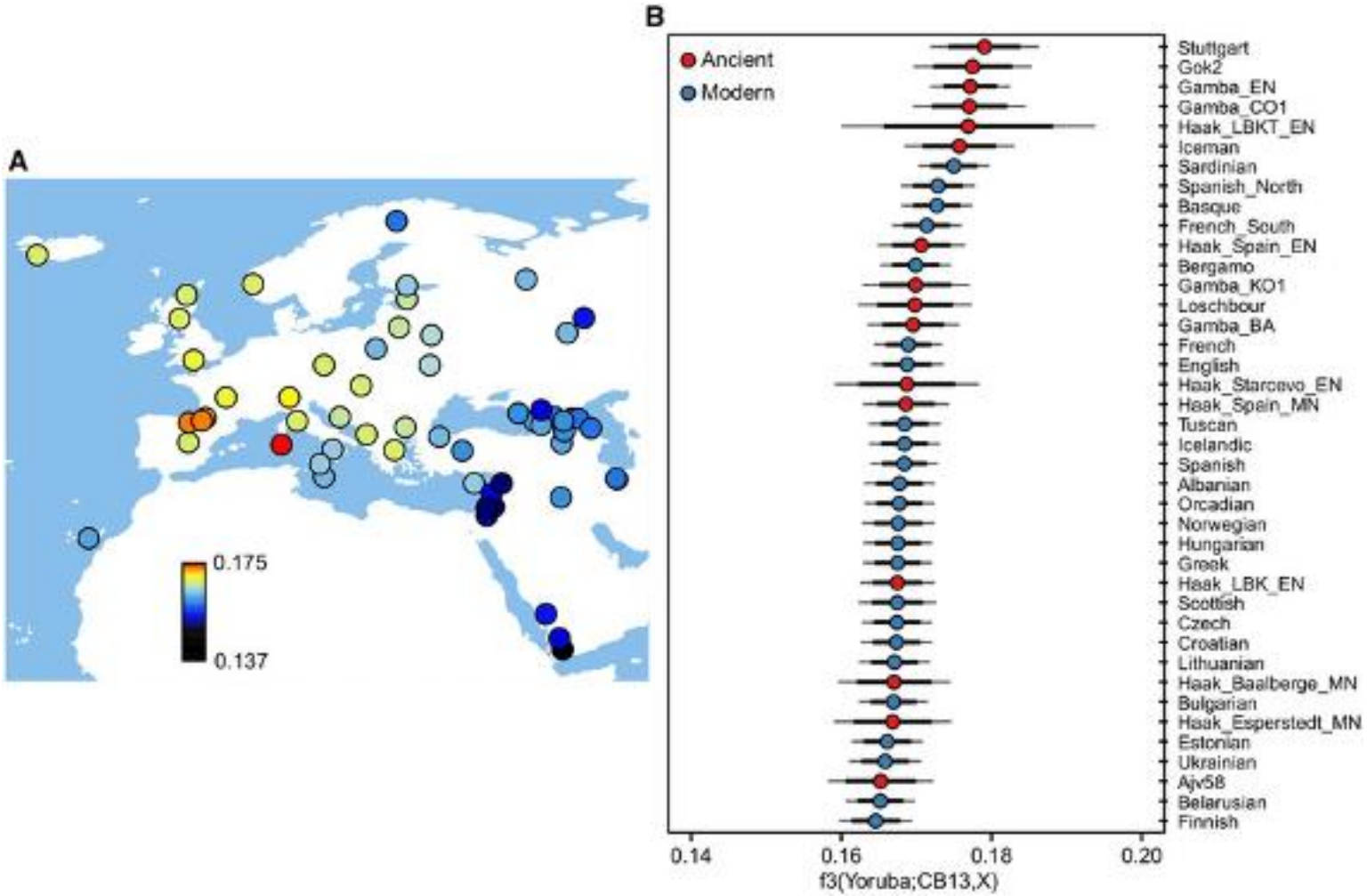
C is not an admixed population

Li and Akey, 2019

Visualización e interpretación (outgroup-$f3$)



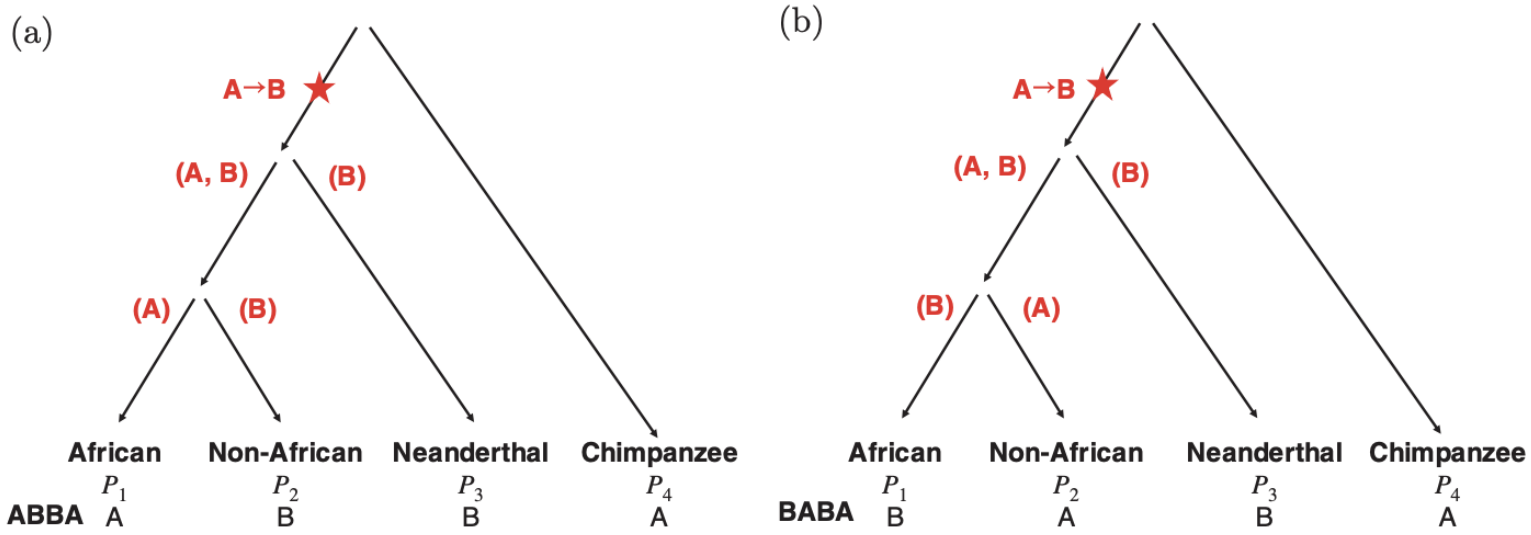
7,400 años, Cova Bonica, Vallirana (Barcelona)



(A) Shared genetic drift between CB13 and present-day Western Eurasian populations. (B) Top 40 populations/individuals (modern and ancient) showing the highest genetic drift with CB13. Olalde et al., 2015

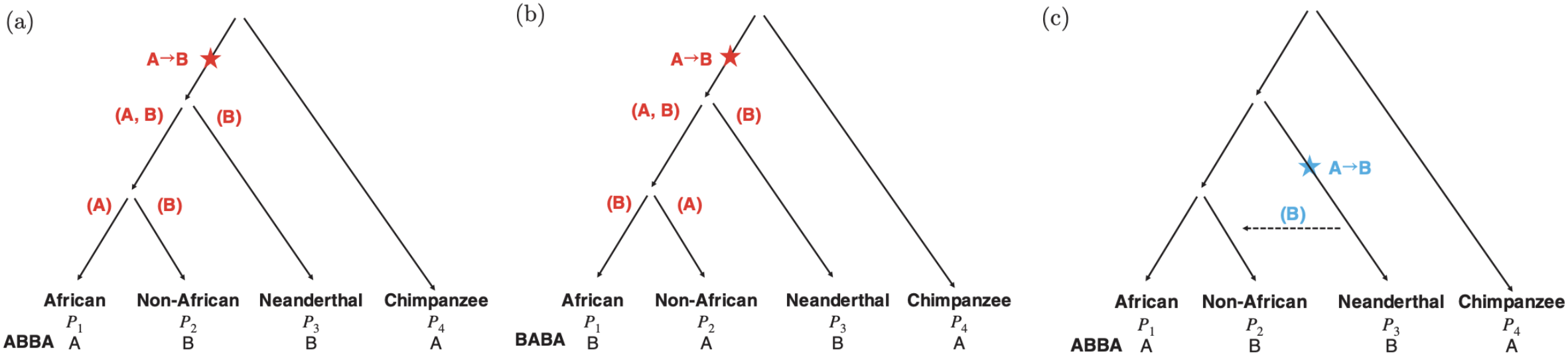
$f4$ o D-statistics

- D-statistic o test de 4 poblaciones o ABBA-BABA.
- Cuatro poblaciones.
- Requiere el uso de dos poblaciones hermanas, una con introgresión/mestizaje y un outgroup.
- Originalmente diseñado para evaluar introgresión de Neanderthal en poblaciones humanas (Reich et al., 2010; Durand et al., 2011)



*f*4 o D-statistics

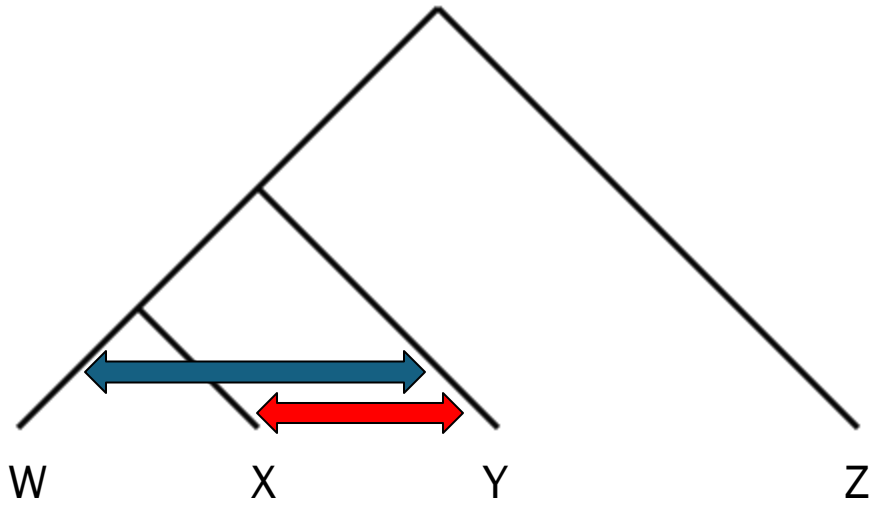
- D-statistic o test de 4 poblaciones o ABBA-BABA.
- Cuatro poblaciones.
- Requiere el uso de dos poblaciones hermanas, una con introgresión/mestizaje y un outgroup.
- Originalmente diseñado para evaluar introgresión de Neanderthal en poblaciones humanas (Reich et al., 2010; Durand et al., 2011)



*f*4 o D-statistics

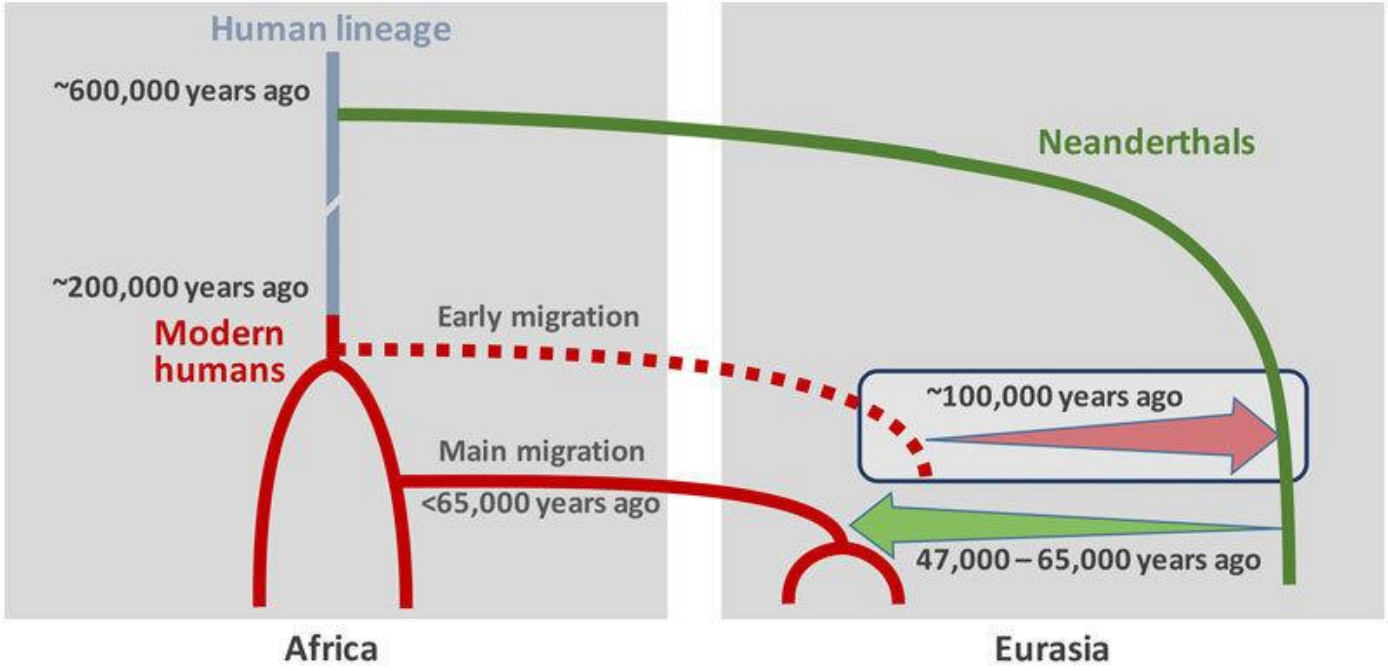
- ✦ Razón entre ABBA/BABA
- ✦ Y está igualmente relacionada a X y W
- ✦ X o W tienen flujo génico de Y

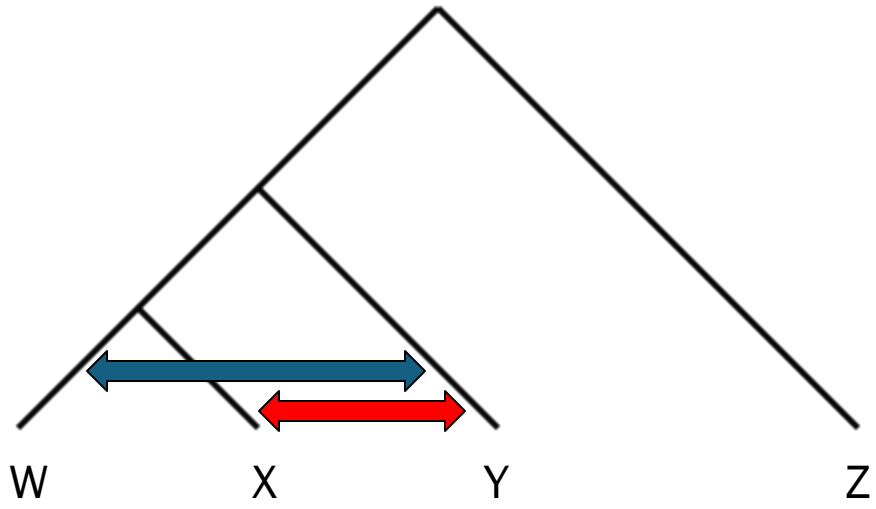
Valores ++ = flujo génico entre W e Y
Valores -- = flujo génico entre X e Y

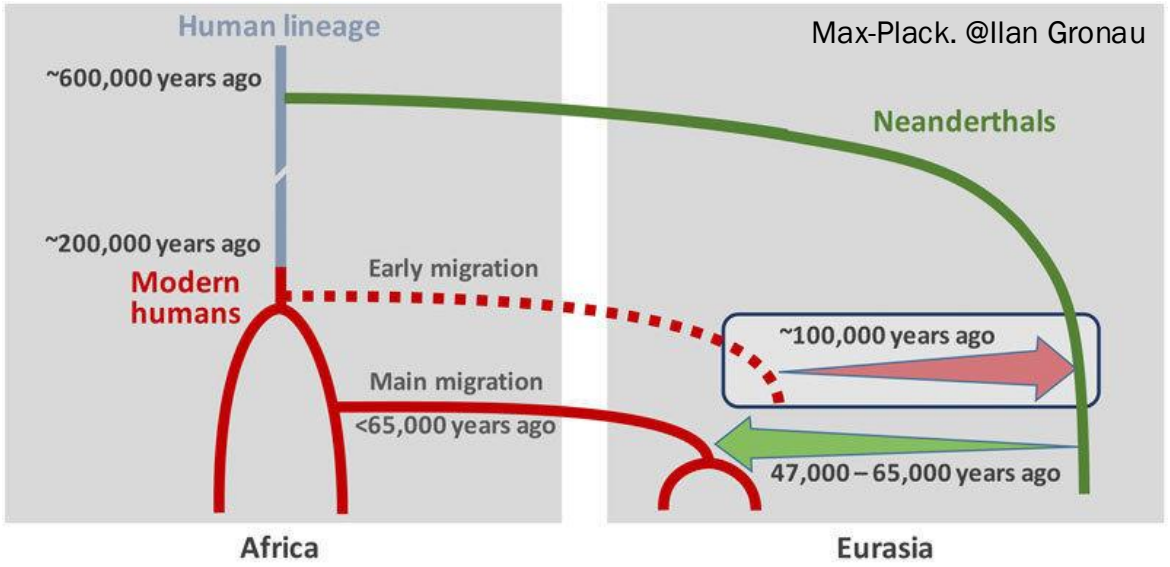


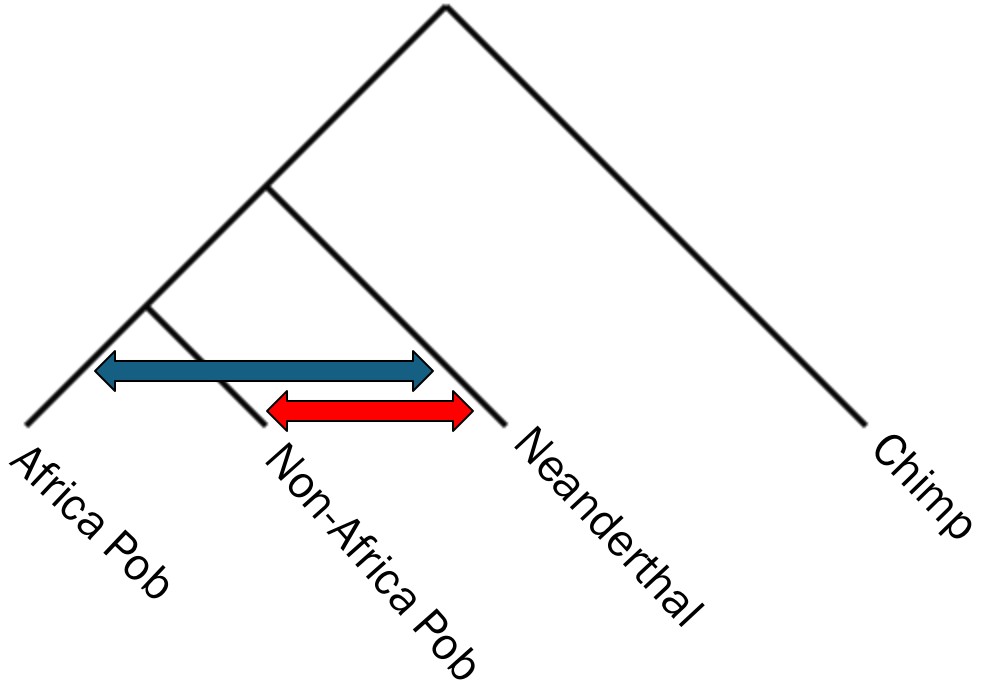
*f*4 o D-statistics

¿Qué poblaciones actuales presentan evidencias de mestizaje con Neanderthal?

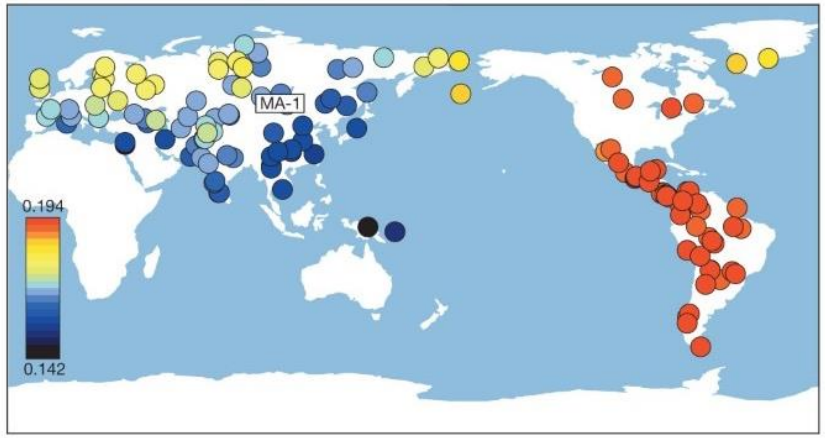


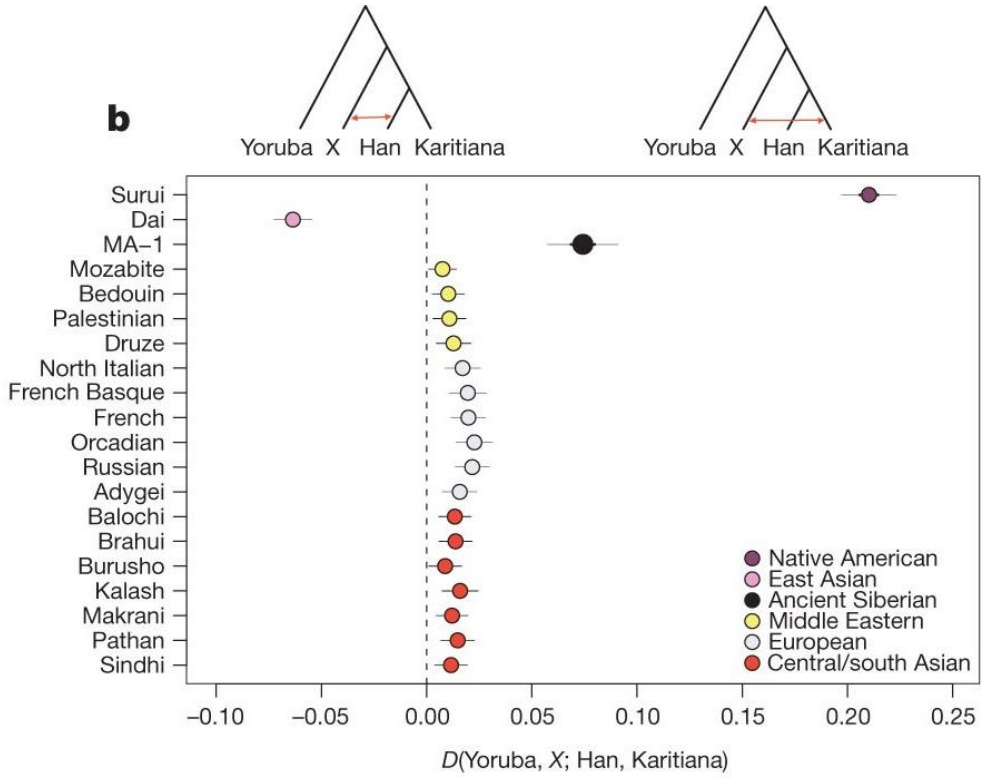


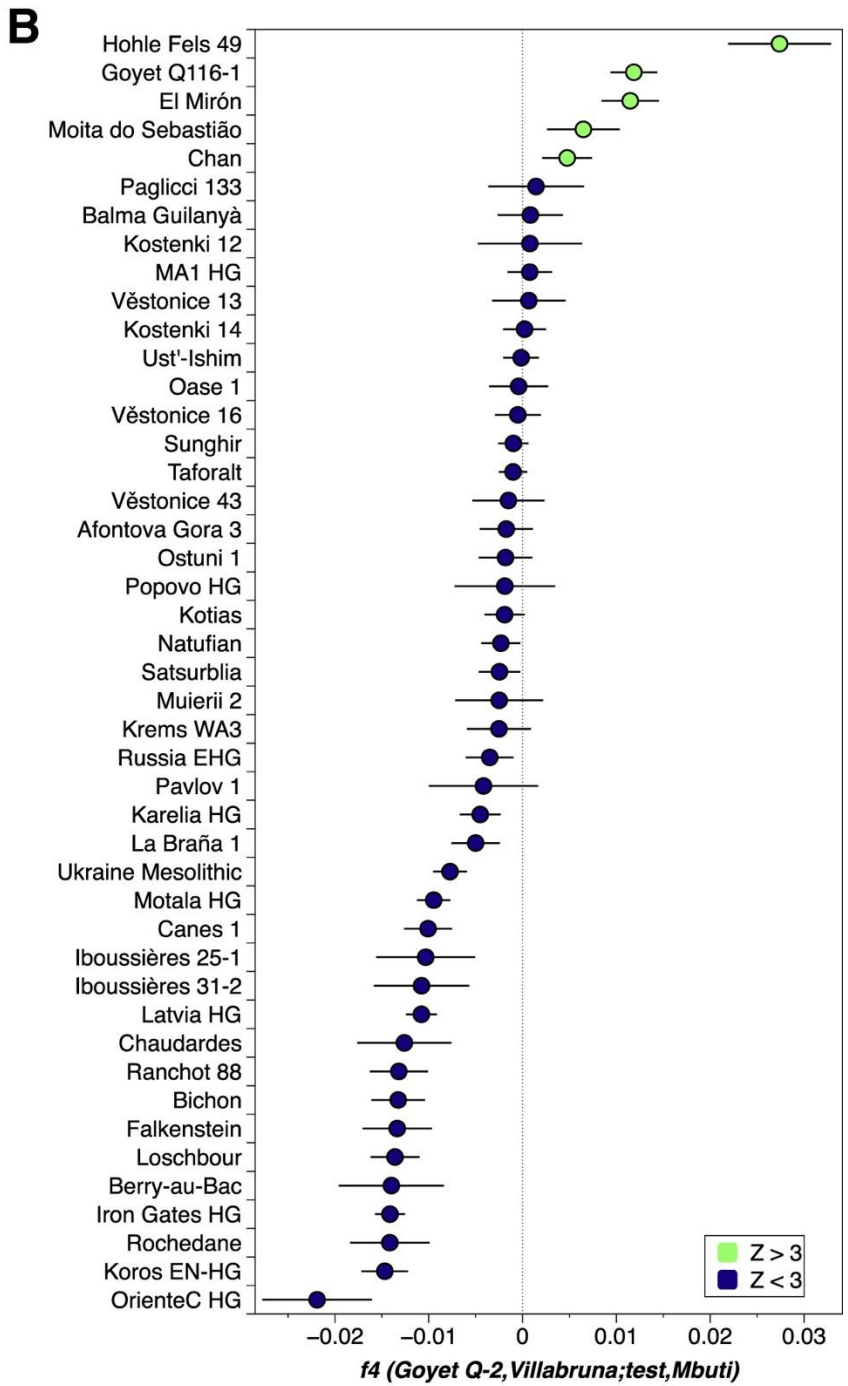




Dstat(África, NonAfrica; Neanderthal; Chimpancé)
D ~ 0 = no hay evidencias de mestizaje, poblaciones de África y fuera de África están igualmente relacionadas a Neanderthal
D << 0 = poblaciones fuera de África presentan evidencias de mestizaje con Neanderthal
D >> 0 = poblaciones en África presentan evidencias de mestizaje



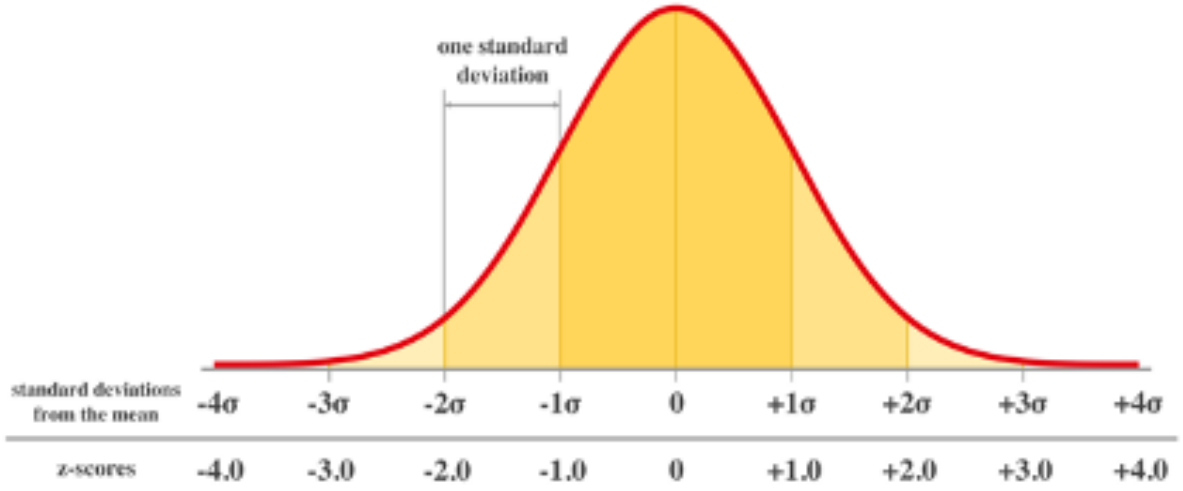




https://doi.org/10.1016/j.cub.2019.02.006
https://doi.org/10.1038/nature12736

Estadísticos F

- ✦ Valor estadístico ($f3$ o $f4$)
- ✦ Z-score: medida en términos de desviaciones estándares del promedio
- ✦ Z-score ~ 0: no hay diferencias con promedio
- ✦ Z-score >= abs(2) o abs(3): valores significativos



qpWave y qpAdm

- qpWave: Estimar el número mínimo de eventos de mezcla en una(s) población objetivo.
- qpAdm: Estimar la proporción de mestizaje de cada fuente.
- En ambos casos, estamos integrando f2, f3 y f4 de las distintas poblaciones o grupos analizados.
- Requerido:
  - Target/Objetivo
  - Posibles fuentes de mestizaje

  } "left populations"

  - Set de "right populations" o "outgroups" → contribuyen a diferenciar a las posibles fuentes de mestizaje

qpWave y qpAdm

“left and right populations” → por su posición en el f4:

f4(target, source; r1,r2)

target y source: left populations
r1 y r2: right populations. Esta lista debe ser mayor a “left” y las poblaciones tienen que estar diferencialmente relacionadas a las potenciales fuentes (source populations)

ADMIXTOOLS2

Overview

ADMIXTOOLS is a collection of programs which use genetic data to infer how populations are related to one another. It has been used in countless publications to test whether populations form clades (*qpDstat*, *qpWave*), to estimate ancestry proportions (*qpAdm*), and to fit admixture graphs (*qpGraph*).

ADMIXTOOLS 2 provides the same functionality as *ADMIXTOOLS* in a new look, and it's orders of magnitude faster. This is achieved through separating the computation of f_2-statistics from all other computations, and through a number of other optimizations.

Features

- Much faster than the original *ADMIXTOOLS* software
- Simple R command line interface
- Even simpler point-and-click interface
- Several new features and methodological innovations that make it easier to find robust models:
  - Automated and semi-automated admixture graph inference
  - Simultaneous exploration of many *qpAdm* models
  - Unbiased comparison of any two *qpGraph* models using out-of-sample scores
  - Jackknife and bootstrap standard errors and confidence intervals for any *qpAdm*, *qpWave*, and *qpGraph* parameters
  - Interface with *msprime* makes it easy to simulate genetic data under an admixture graph
- Full support for genotype data in *PACKEDANCESTRYMAP/EIGENSTRAT* format and *PLINK* format
- Wrapper functions around the original *ADMIXTOOLS* software (see also admixr)
- Extensive documentation
- New features available on request!

EIGENSTRAT/PACKEDANCESTRYMAP

EIGENSTRAT and *PACKEDANCESTRYMAP* use the same format for the metadata. The only difference is that the genotype data is stored in a text file in *EIGENSTRAT*, but in a binary file in *PACKEDANCESTRYMAP*.

The `.ind` file has 3 out of the 6 columns in the `.fam` file, in different order:

```
S_French-1.DG M French.DG
S_French-2.DG F French.DG
B_French-3.DG M French.DG
BR_Onge-2.DG F Onge.DG
BR_Onge-1.DG F Onge.DG
```

`.snp` files are the same as `.bim` files, except that the first two columns are swapped:

```
rs3094315 1 0.02013 752566  G   A
rs12124819 1    0.020242    776546  A   G
rs28765502 1    0.022137    832918  T   C
```

EIGENSTRAT `.geno` files look like this:

```
29101
10211
22901
```

One SNP per row, one sample per column, and missing data is denoted by `9`.