

Genética de poblaciones e inferencia de ancestría genética.



Grupo de Genética de Poblaciones y Bioinformática

LCG. Fernando Pérez Villatoro fperez@inmegen.edu.mx

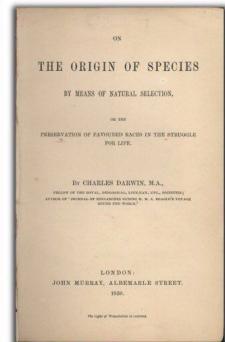
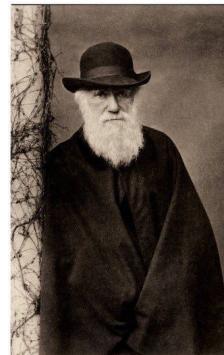
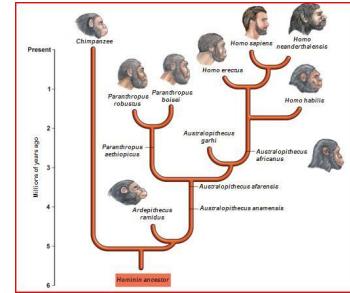
La Genética de Poblaciones: es el campo de la biología que estudia la composición genética de las poblaciones y los cambios en dicha composición que resultan de la intervención de *factores evolutivos*.



<https://www.nature.com/subjects/population-genetics>

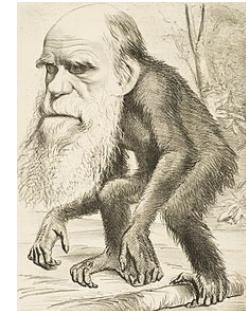
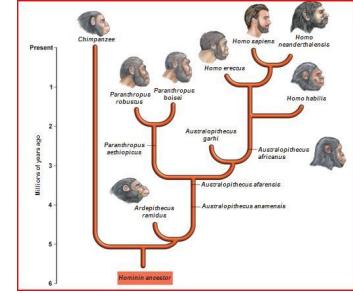
Factores involucrados en el proceso evolutivo

1. Variación
2. Herencia
3. Selección



Factores involucrados en el proceso evolutivo *by Darwin*

1. Variación
2. Herencia
3. Selección



Las mutaciones son cambios al azar en el DNA

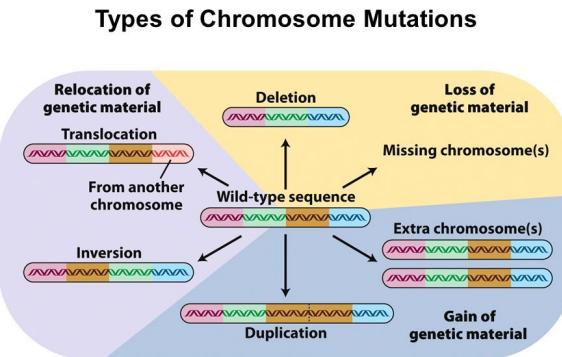
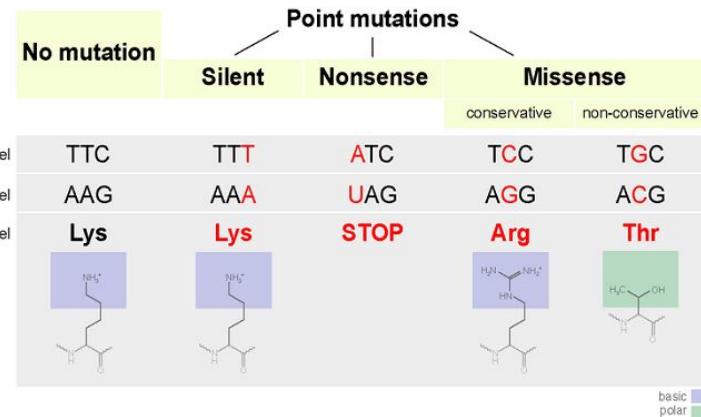


Figure 17-2
Introduction to Genetic Analysis, Tenth Edition
© 2012 W. H. Freeman and Company

| | | | |
|---------------|---------------------------|---|----------------------------|
| translocación | AGGTACCAT TCCATCCTA | ➔ | AACCGGTAT TTGGCCATA |
| inversión | ATGATT TCG TCA | ➔ | ATGAT <ins>G</ins> TCA |
| deletión | CACT AGG CATC | ➔ | CACT * ATC |
| inserción | GCATAACG | ➔ | GCAT <ins>TTCA</ins> TACCG |
| sustitución | GCATCCTA | ➔ | GT <u>CT</u> CCTA |



Las mutaciones son cambios al azar en el DNA

- La tasa de mutación en humanos se estima de:
 1.25×10^{-8} *mutaciones por generación**
- El genoma humano contiene alrededor de:
 3.2×10^9 *pares de bases*
- En total se esperaría que en cada generación ocurran al azar:
~40 mutaciones por individuo.

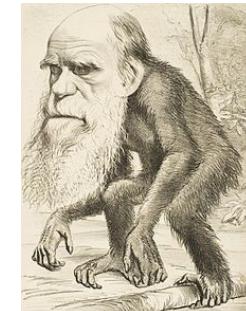
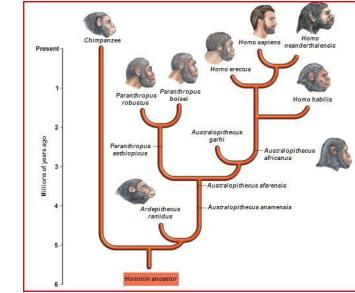
* Kong, A. et al (2012). *Nature*, 488(7412) / Scally, A., & Durbin, R. (2012). *Nature Reviews Genetics*, 13(10).

Si, todos somos mutantes

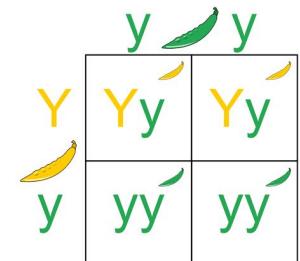
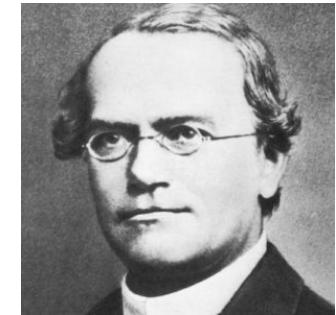
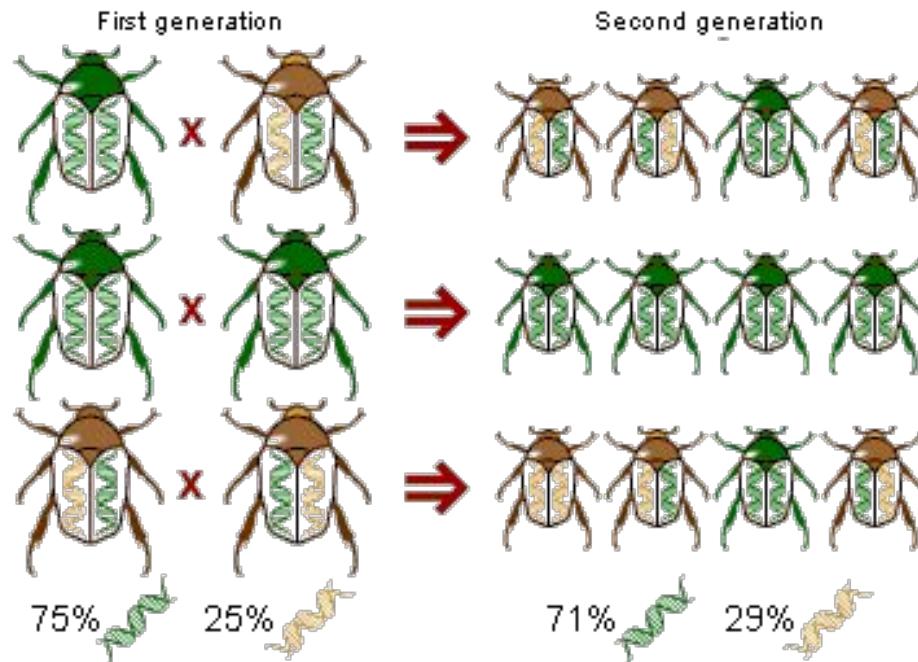


Factores involucrados en el proceso evolutivo *by Darwin*

1. Variación
2. Herencia
3. Selección

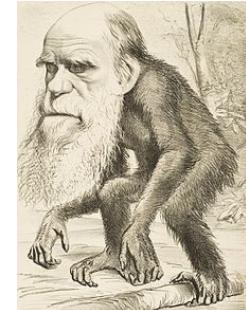
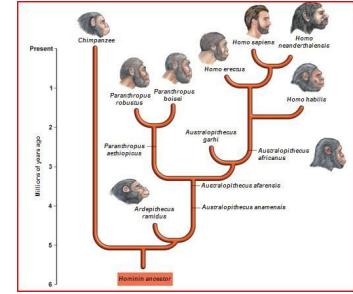


Características heredables siguen ciertas “leyes”



Factores involucrados en el proceso evolutivo *by Darwin*

1. Variación
2. Herencia
3. Selección

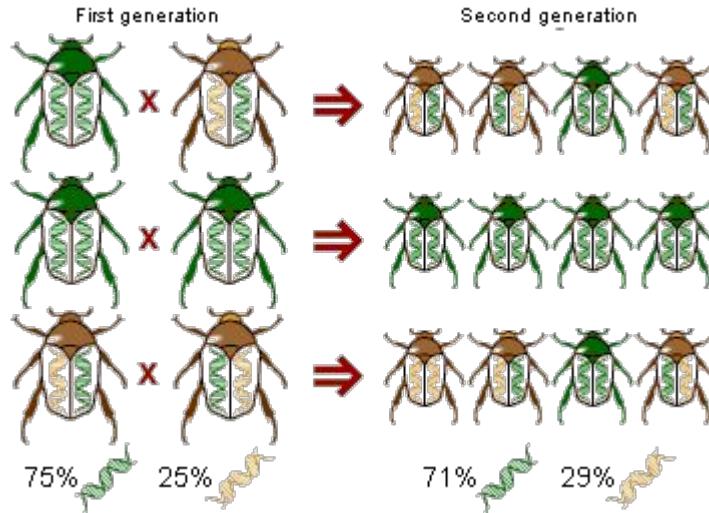


“Entonces aquellos miembros de la población con características menos adaptadas (según lo determine su medio ambiente) morirán con mayor probabilidad. Entonces aquellos miembros con características mejor adaptadas sobrevivirán más probablemente.”

Darwin, El origen de las especies

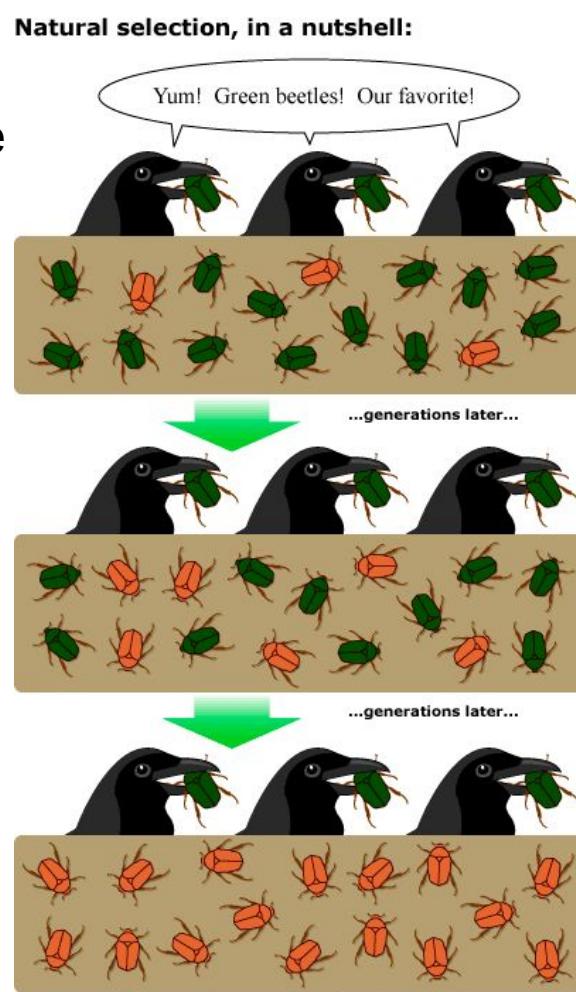
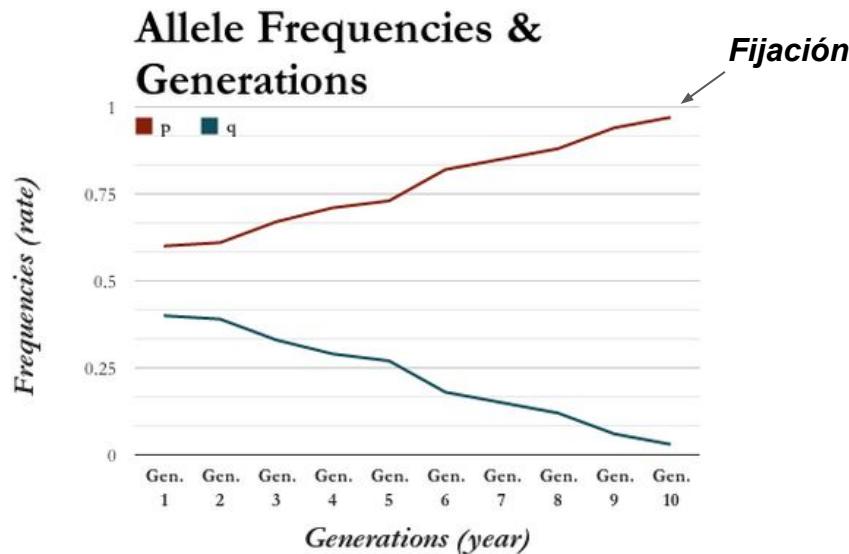


“Entonces aquellos miembros con características mejor adaptadas sobrevivirán más probablemente.”

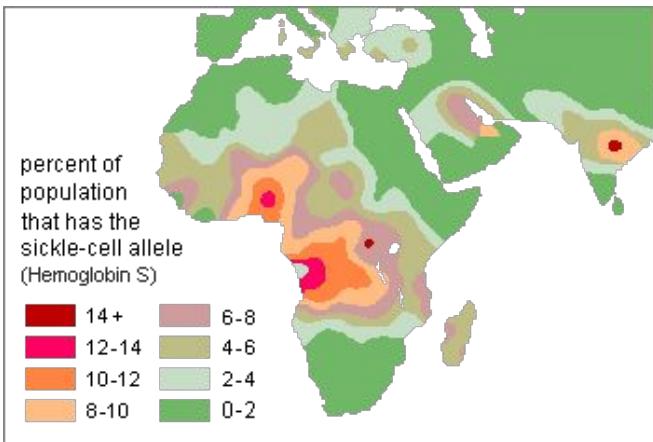


Natural selection, in a nutshell:

Si el ambiente cambia, las frecuencia alélicas de la población cambiarían.



En la población Africana, alelos (variantes genéticas) que protegen de malaria han aumentado su frecuencia.



nature
International journal of science

Letter | Published: 20 July 1995

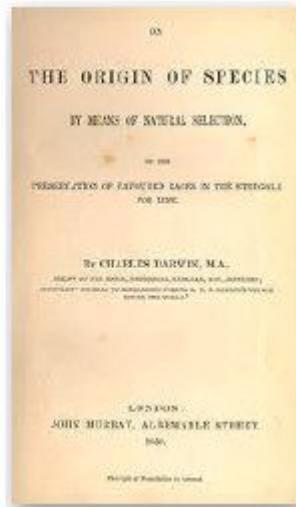
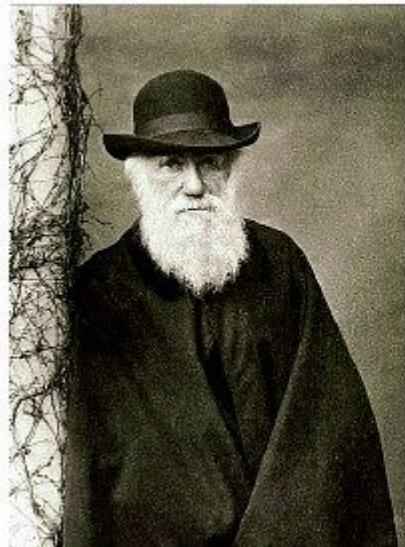
Natural selection of hemi- and heterozygotes for G6PD deficiency in Africa by resistance to severe malaria

C. Ruwende, S. C. Khoo, R. W. Snow, S. N. R. Yates, D. Kwiatkowski, S. Gupta, P. Warn, C. E. M. Allsopp, S. C. Gilbert, N. Pesch, C. I. Newbold, B. M. Greenwood, K. Marsh & A. V. S. Hill

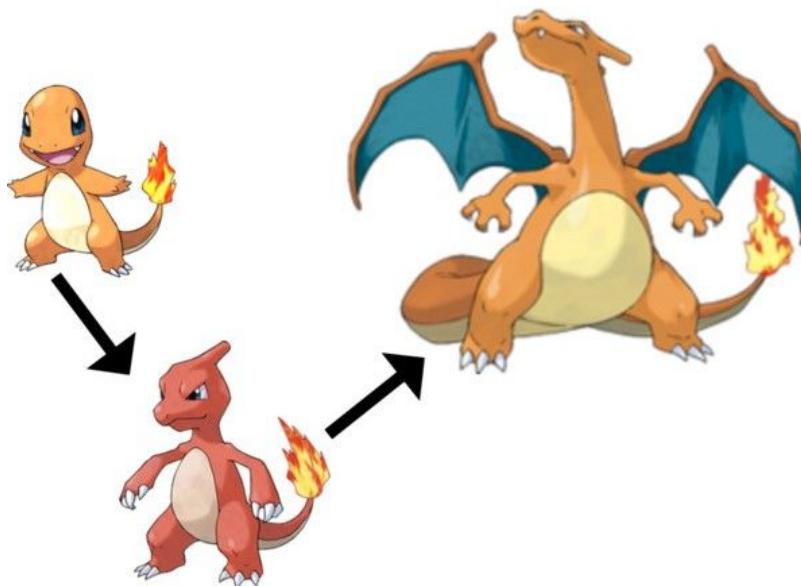
Nature **376**, 246–249 (20 July 1995) | Download Citation ↗

<http://sci.waikato.ac.nz/bioblog/2009/09/malaria-sicklecell-anaemia.shtml>

Darwin hizo una excelente aproximación sobre la evolución. Sin embargo no todo se explica bajo sus tres principios.



Por ejemplo....



Otros conceptos de genética de poblaciones

1. Deriva génica
2. Divergencia poblacional
3. Coalescencia
4. Flujo genético

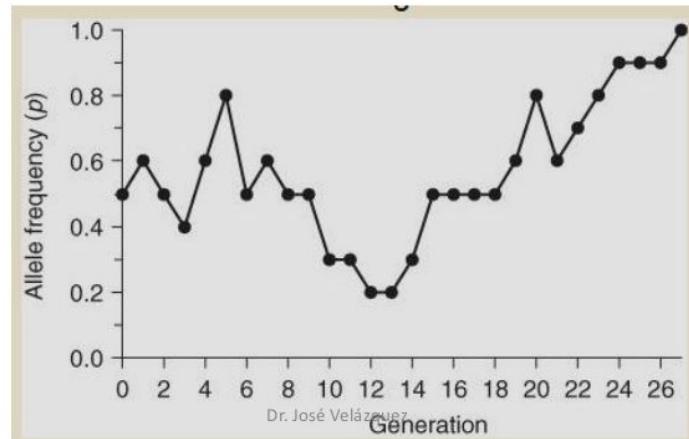
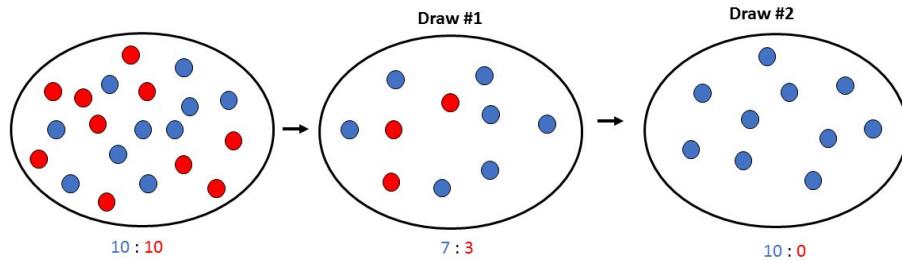


Deriva génica

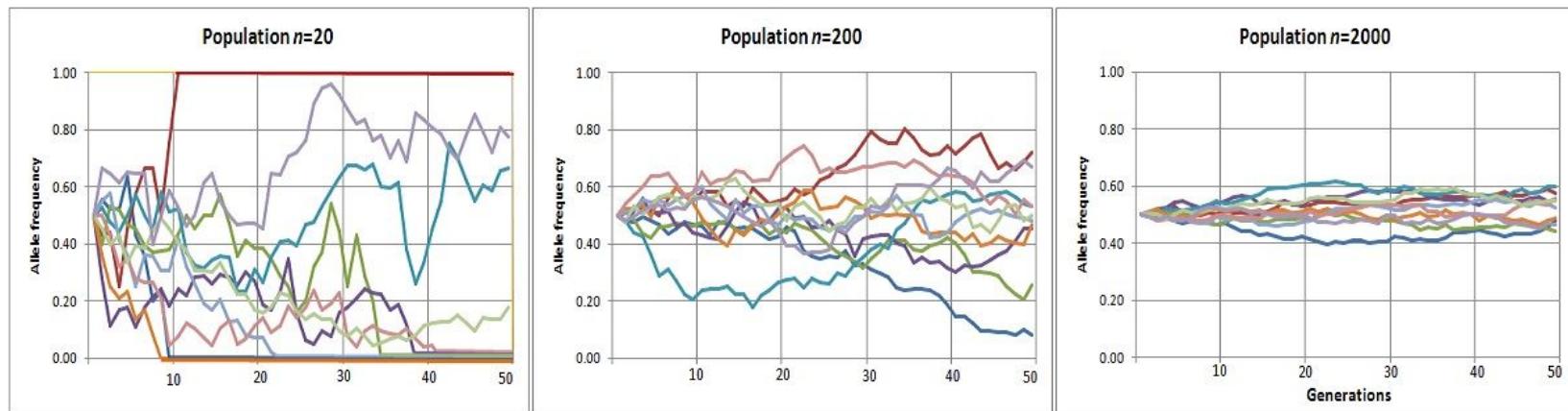
Las frecuencias alélicas de una población cambian a lo largo de varias generaciones debido al azar



La deriva genética a través de las generaciones produce cambios en frecuencias alélicas



El tamaño de población efectivo es uno de los factores principales con respecto al cambio de frecuencias alélicas por deriva genética.

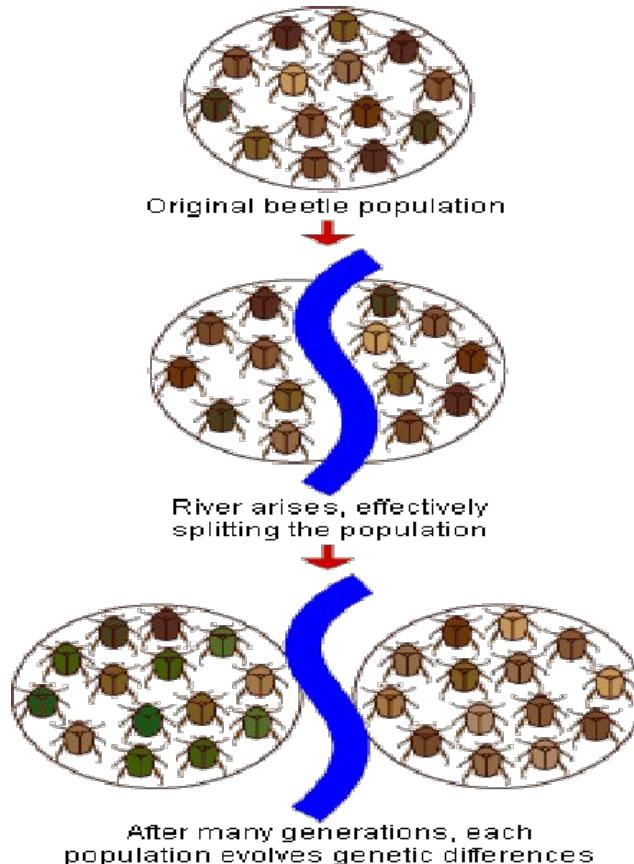


Conceptos de genética de poblaciones

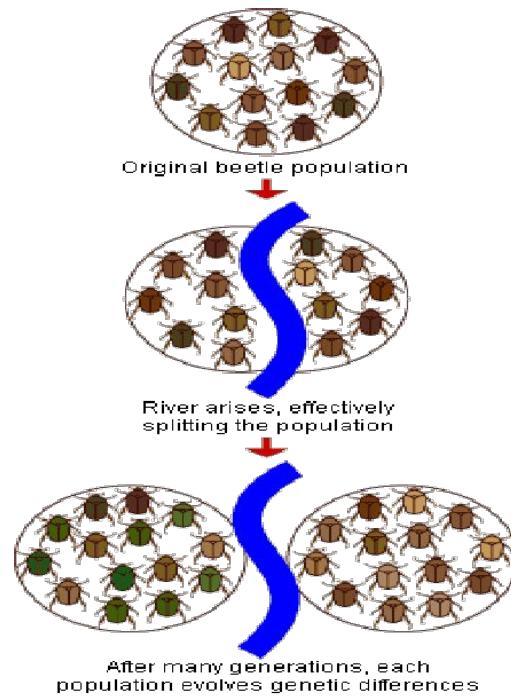
1. Deriva génica
2. Divergencia poblacional
3. Coalescencia
4. Flujo genético



Divergencia poblacional



Divergencia poblacional

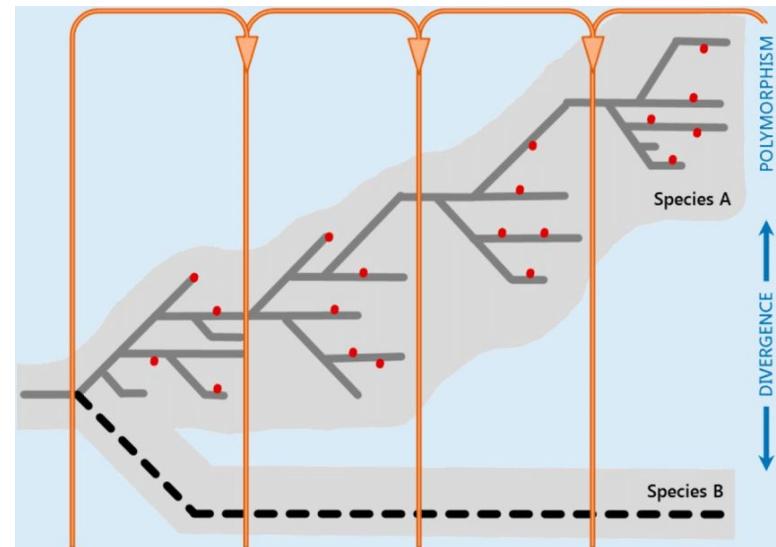


La deriva genética genera divergencias entre una especie de otra

La teoría neutralista de la evolución molecular



Motoo Kimura



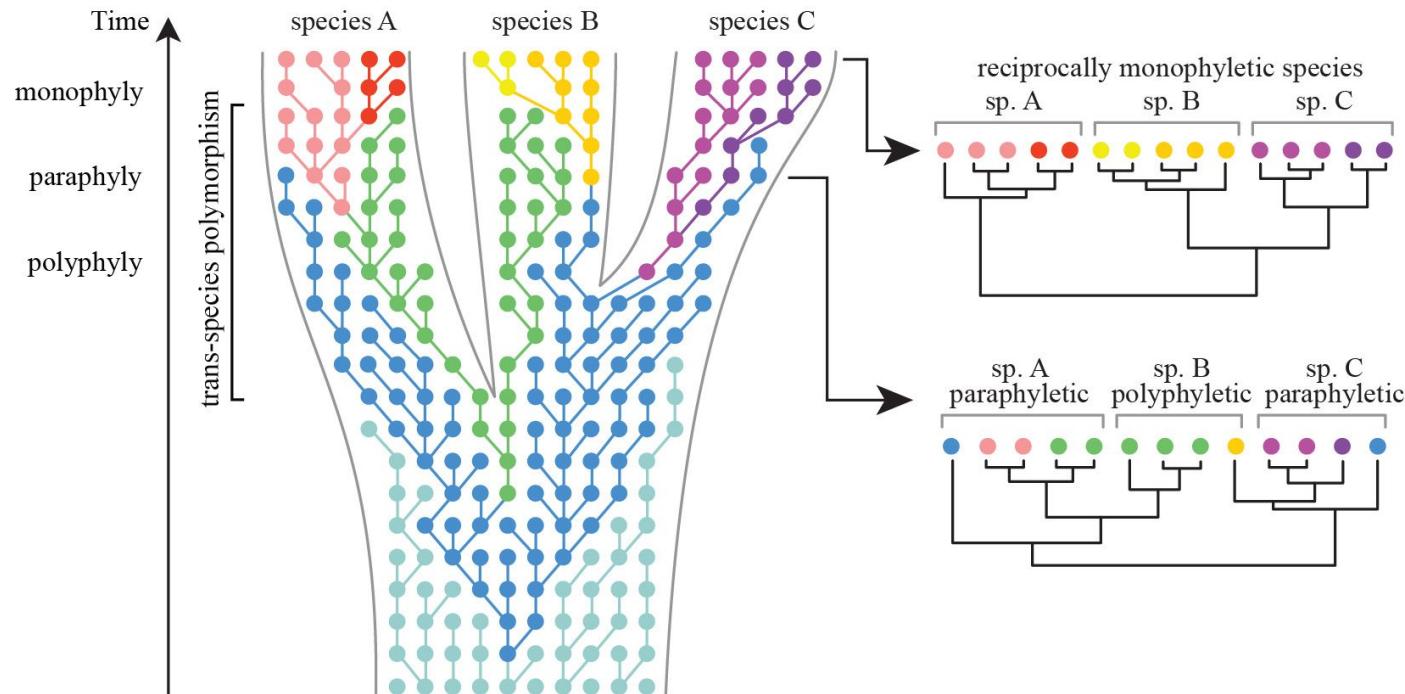
Casillas & Barbadilla (2017). Molecular Population Genetics. Genetics. 205. 1003.

Conceptos de genética de poblaciones

1. Deriva génica
2. Divergencia poblacional
3. Coalescencia
4. Flujo genético



Coalescencia

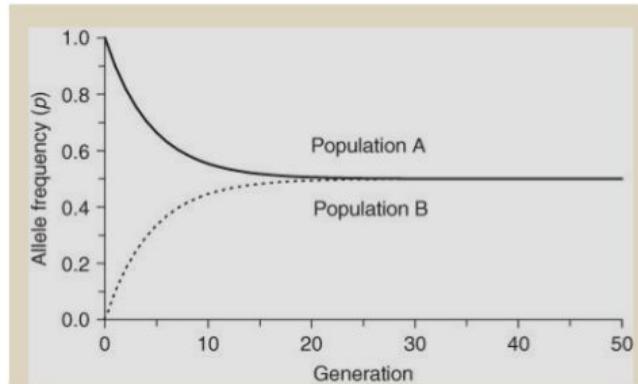
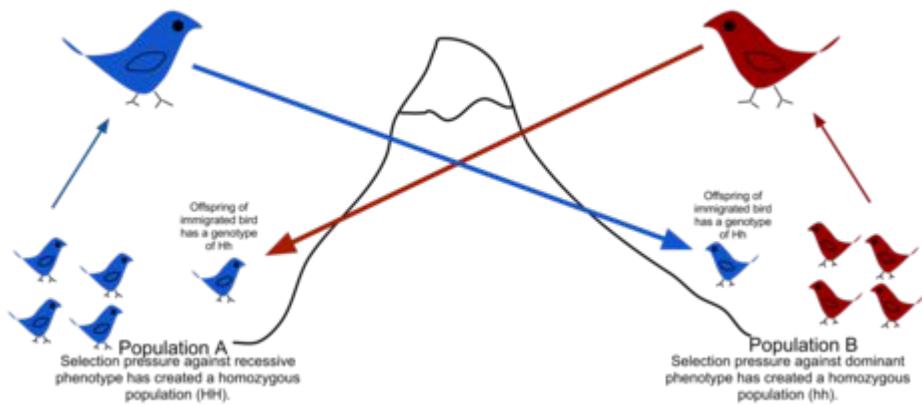


Conceptos de genética de poblaciones

1. Deriva génica
2. Divergencia poblacional
3. Coalescencia
4. Flujo genético



La introducción de individuos de otra población cambia las frecuencias alélicas.



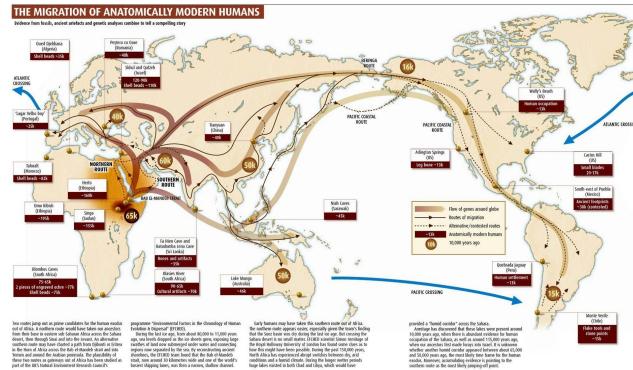
- Todos los procesos anteriores: su inferencia, cambios, reconstrucción son sujeto de estudio de la genética de poblaciones.

Las poblaciones humanas por miles de años se caracterizaron por:

Clanes pequeños



Migraciones y aislamiento



En los últimos miles de años han aumentado procesos de:

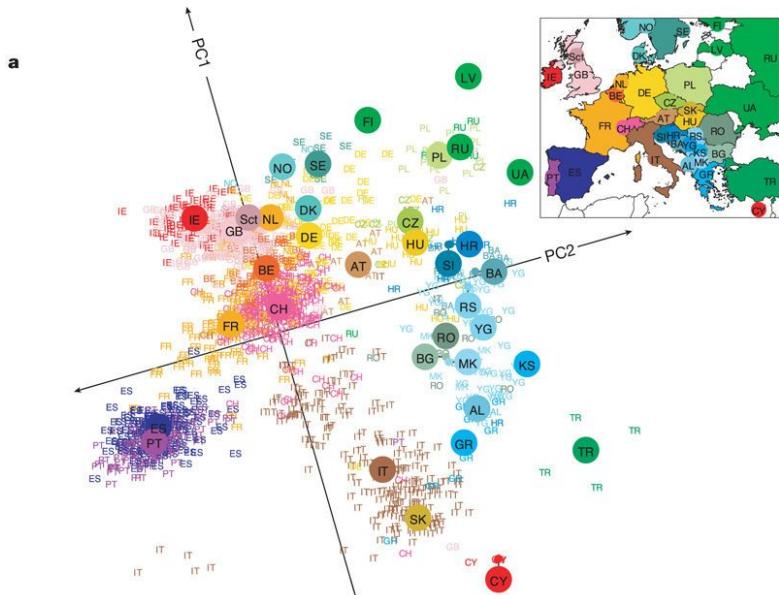
Fitness



Mestizaje



Las poblaciones humanas actuales pueden agruparse de acuerdo a sus frecuencias alélicas



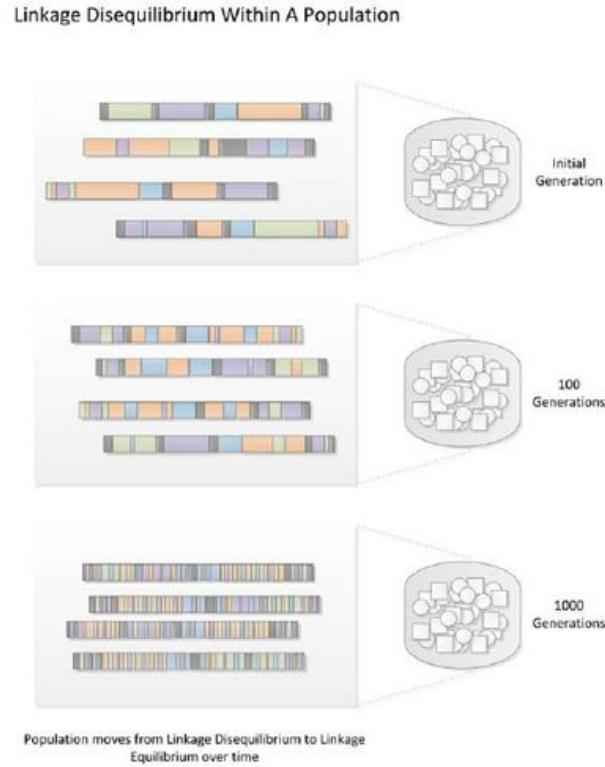
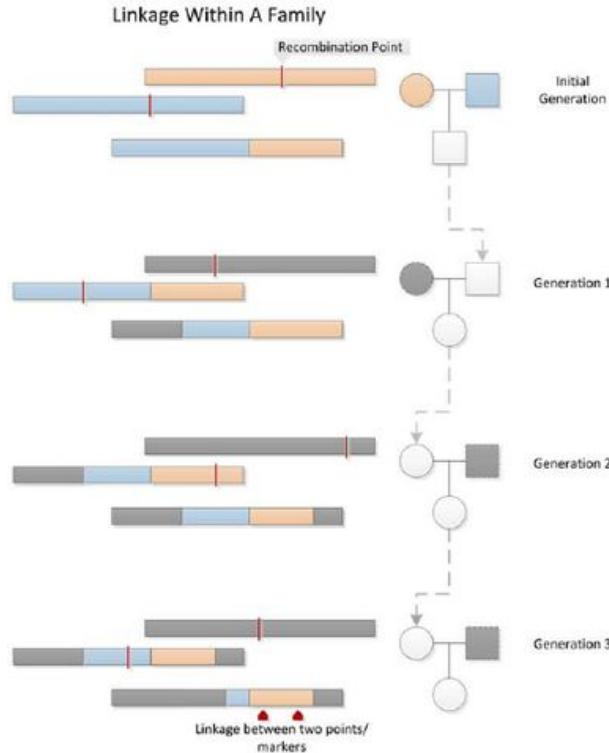
Novembre, et al, 2008

Haplotipos

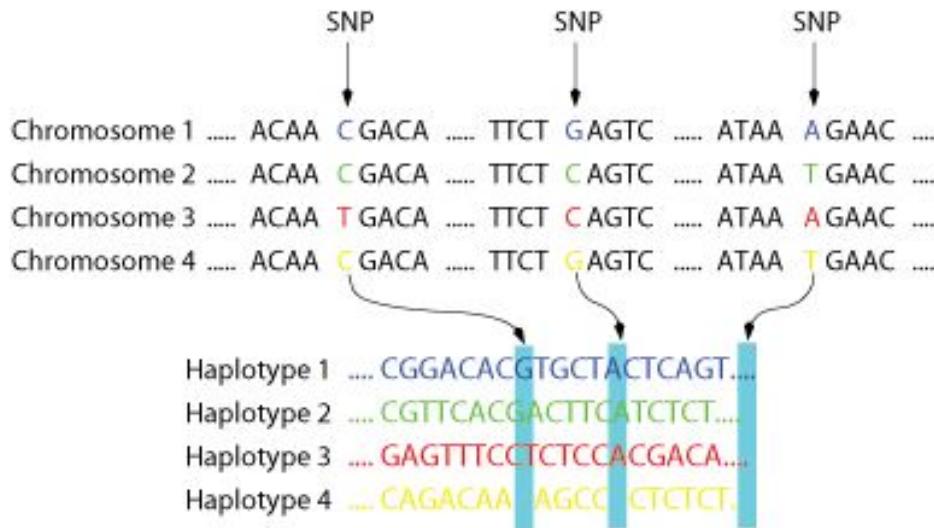
Otros conceptos importantes

1. Desequilibrio de ligamiento
2. Haplotipos

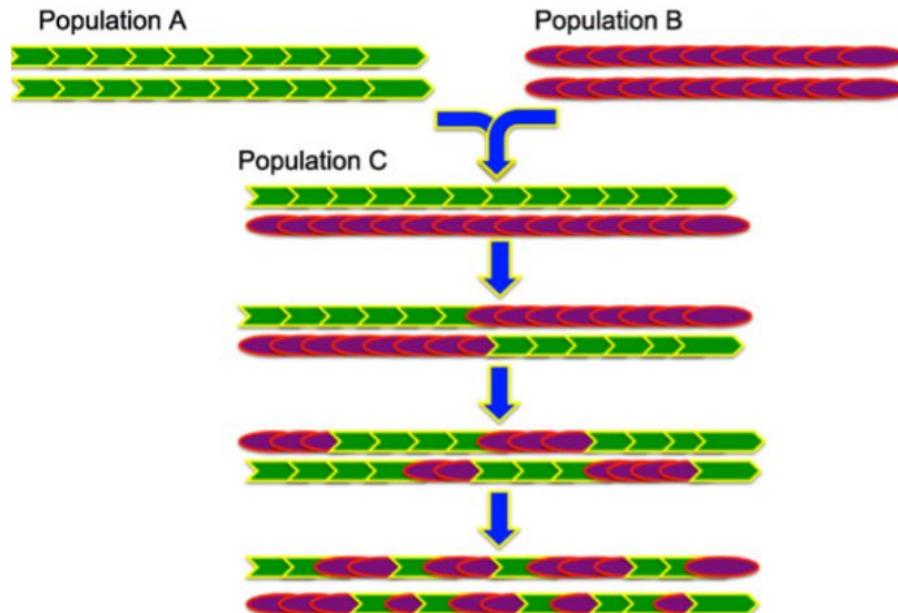
Desequilibrio de ligamiento



El desequilibrio de ligamiento origina **Haplótipos**

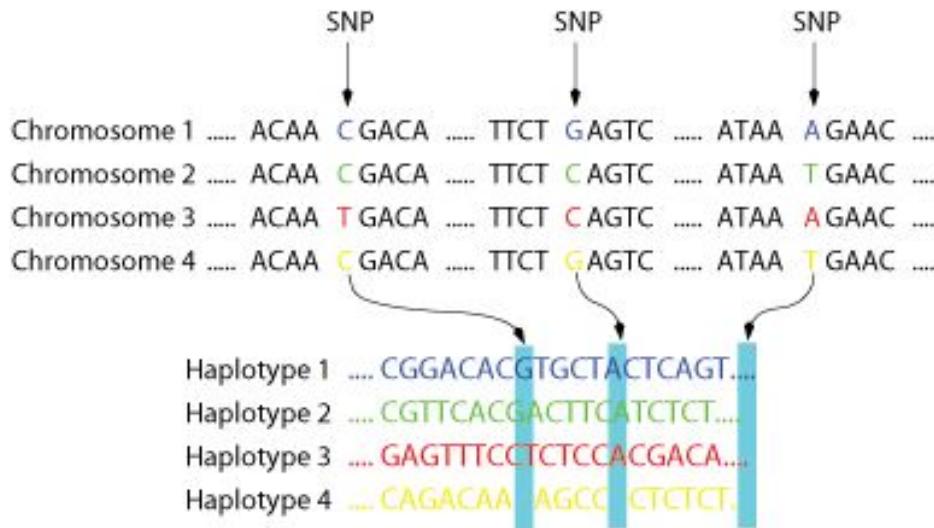


Mestizaje de individuos de dos poblaciones



A. Martin. ASHG 2015

¿Como podemos inferir si dos variantes se heredan en conjunto (presentan LD)?



¿Como podemos inferir si dos variantes se heredan en conjunto (presentan LD)?

| | |
|------------------------|---------------------|
| ACT G GTAT..... | GATCA A CCAG |
| ACT C GTAT..... | GATCA A CCAG |
| ACT C GTAT..... | GATCA T CCAG |

SNP1 **SNP2**

Showing only alleles for both SNPs

| Alleles | SNP1 | SNP2 |
|----------|----------|----------|
| Allele 1 | G | A |
| Allele 2 | C | T |

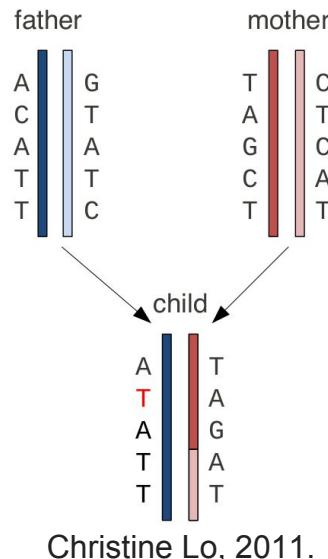
Se realiza una prueba estadística (Person generalmente) para demostrar si dos variantes genéticas se heredan en conjunto (mismo bloque LD).

| SNP1 | | SNP2 | |
|--------|-----------|--------|-----------|
| Allele | Frequency | Allele | Frequency |
| G | p1 | A | q1 |
| C | p2 | T | q2 |

| Haplotype | Frequency | Haplotype | Frequency |
|-----------|-----------|-----------|-----------|
| GA | p11 | GT | p12 |
| CA | p21 | CT | q22 |

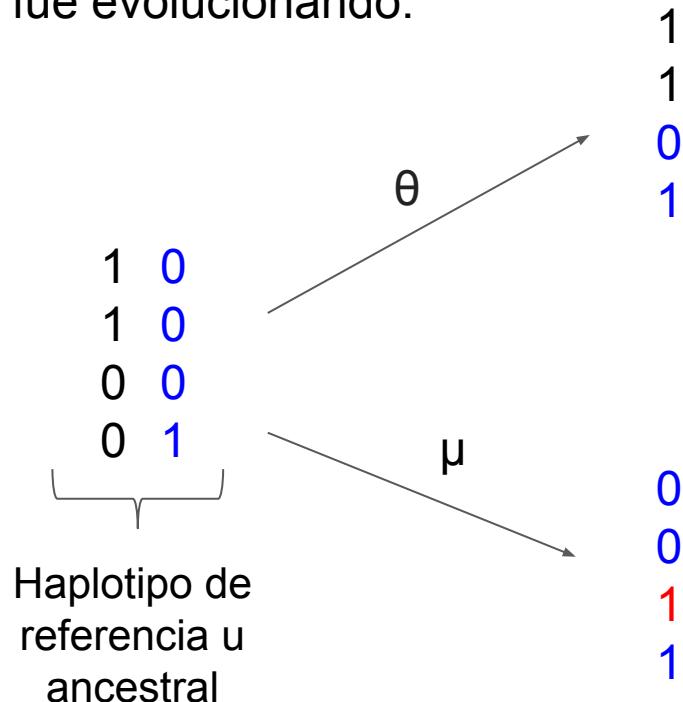
| Haplotype frequency | | Product of allelic frequency |
|---------------------|---|------------------------------|
| p11 | = | $p_1 q_1$ |
| p12 | = | $p_1 q_2$ |
| p21 | = | $p_2 q_1$ |
| p22 | = | $p_2 q_2$ |

- Se le llama Faseo a la reconstrucción de los haplotipos en el genoma.
- Un Faseo puede realizarse usando familiares, haplotipos de referencia o inferirlo de Novo.



Algoritmo de Stephens (PHASE)

Se basa en que los haplotipos en una población parten de un haplotipo ancestral que fue evolucionando.

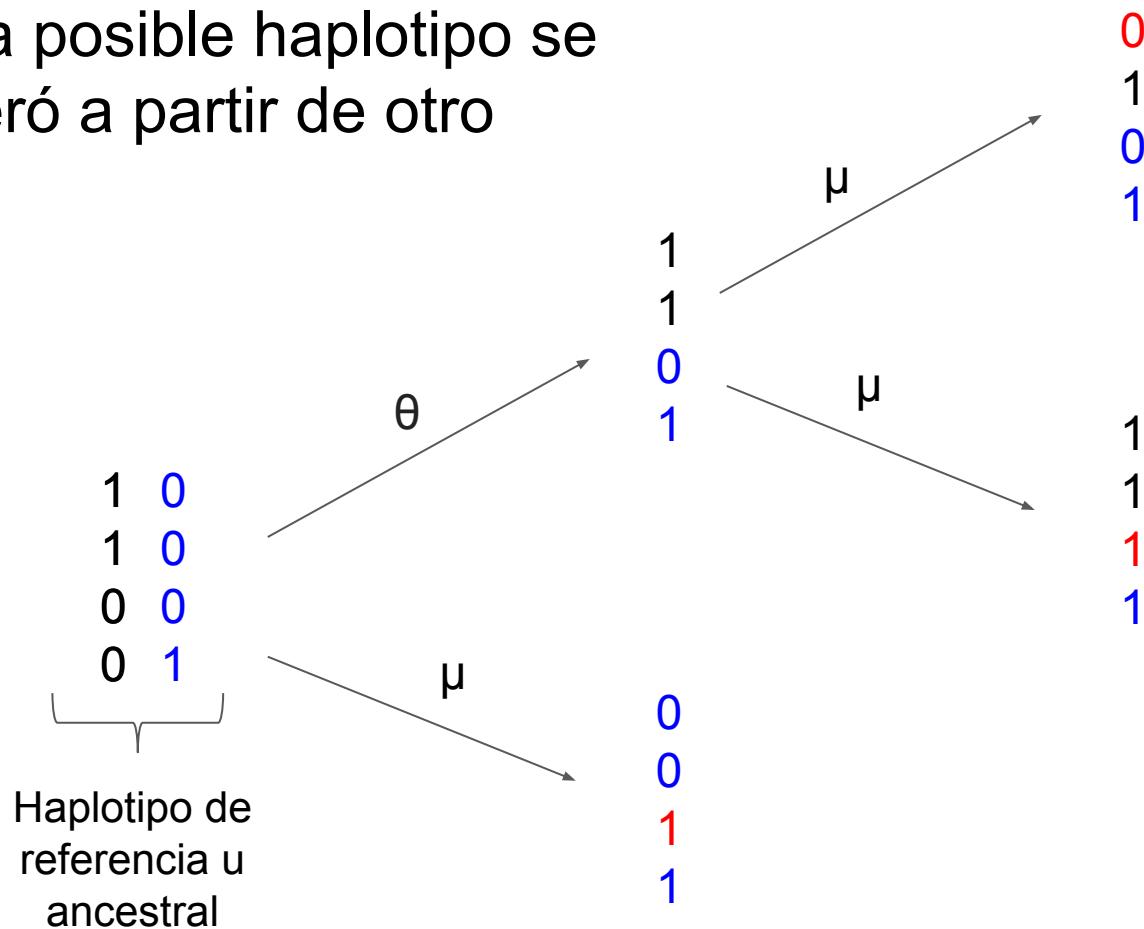


θ - Probabilidad de recombinación
 μ - Tasa de mutación

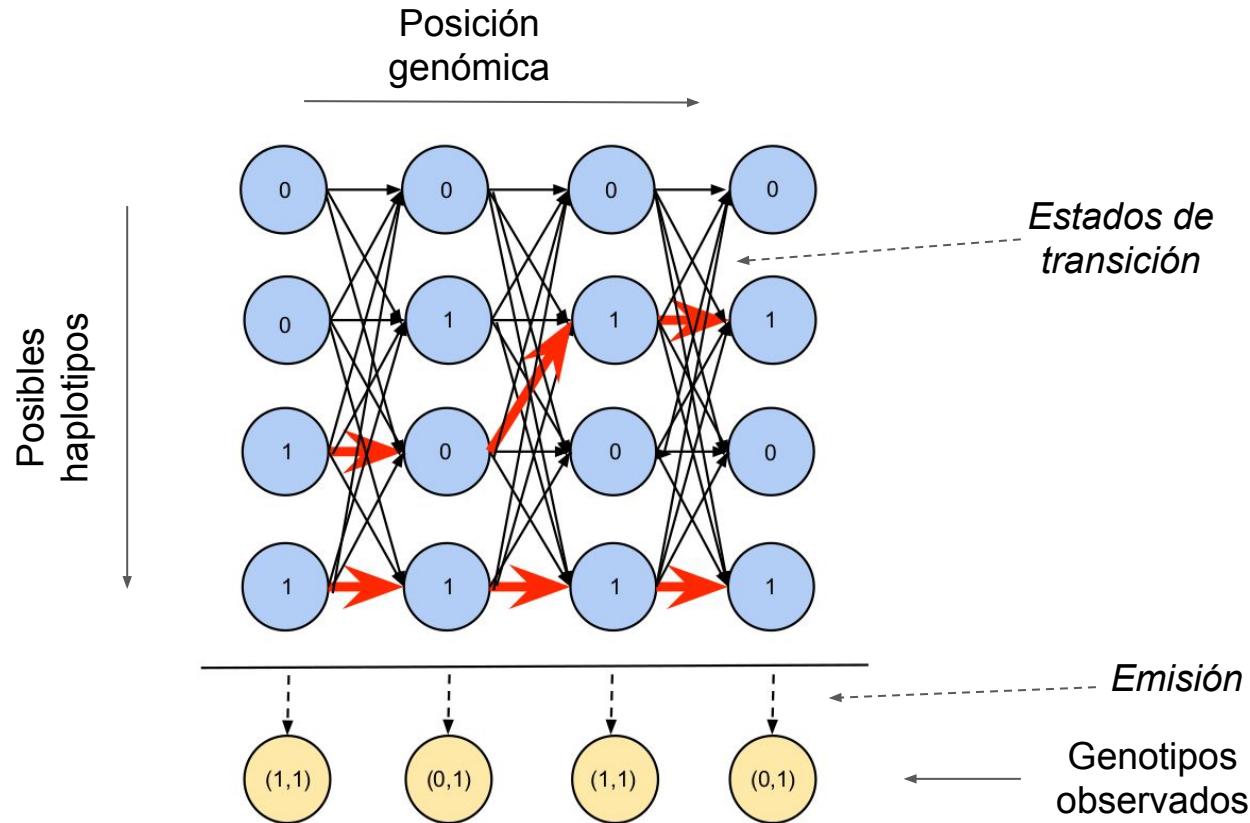
Algortimo de Li-Stephens

- *A new haplotype in the population is more likely to match an existing haplotype that has occurred more frequently in the population.*
- *The probability of seeing a novel haplotype increases as the sample size, n , increases.*
- *The probability of seeing a novel haplotype increases as mutation rate, μ , increases.*

Cada posible haplotipo se generó a partir de otro

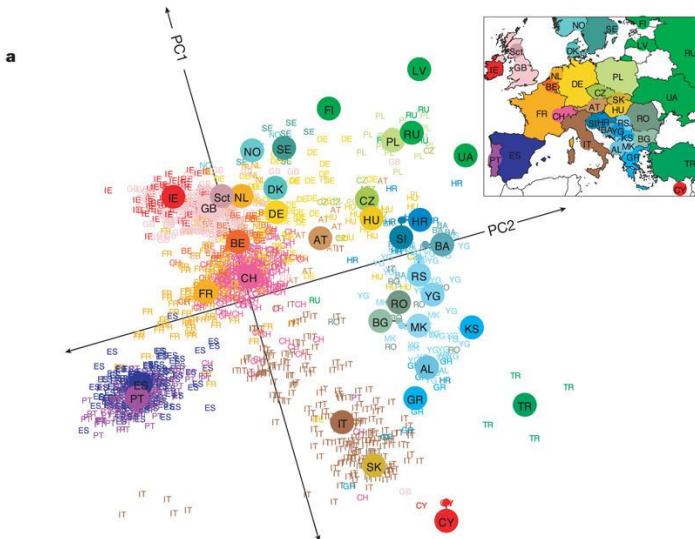


Algoritmo de Li-Stephens



Cálculos de Ancestría

Las poblaciones humanas actuales pueden agruparse de acuerdo a sus frecuencias alélicas



¿Para qué nos serviría conocer la “ancestría genética” de individuos de estudio?

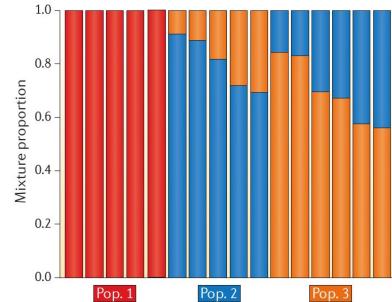


¿Para qué nos serviría conocer la “ancestría genética” de individuos de estudio?

- Estudios antropológicos
- Detectar efecto batch en poblaciones mestizas
- Otros...



- **Ancestría global:** Identificar el porcentaje de cada población ancestral en promedio a lo largo del genoma.

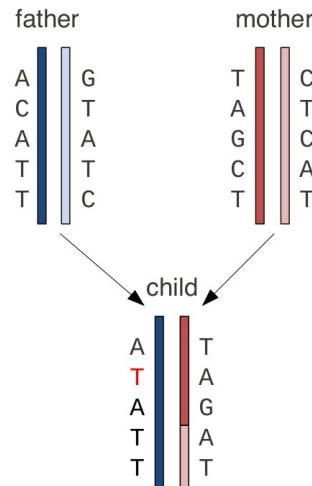


- **Ancestría local:** Identificar segmentos en el genoma correspondientes a las poblaciones ancestrales.



Ancestría Local

Los métodos para calcular requieren como *input* datos genéticos “faseados”, es decir que se conozcan los haplotipos a lo largo de los cromosomas.



Christine Lo, 2011.

Ancestría: Global y Local

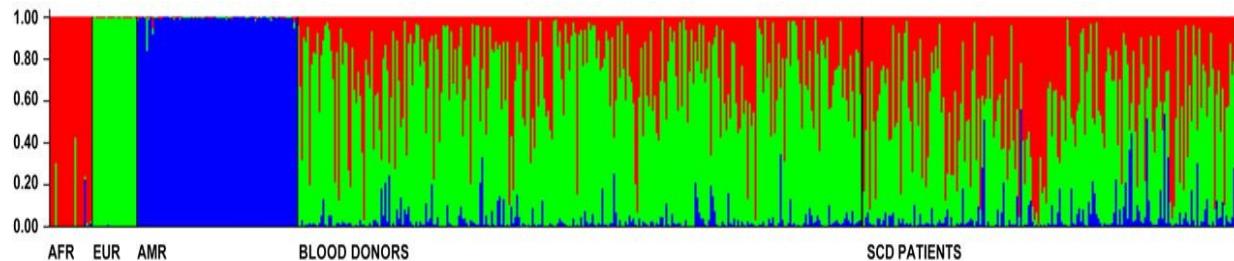
Existen dos tipos de algoritmos para calcular:

1. Algoritmos basados en modelos poblacionales
2. Basados en agrupamientos no supervisados

Ancestría Global

Métodos Basados en Modelos

1. Se basa en las frecuencias alélicas, agrupa poblaciones que estas sigan un modelo (*Hardy-Weinberg*).
2. El uso de AIMs genera mejores resultados
3. Uno de los software más utilizados es ADMIXTURE, basado en programa STRUCTURE.



da Silva, et al, 2011 *Blood*, 118

Ancestría Local

1. Algoritmos basados en modelos poblacionales:

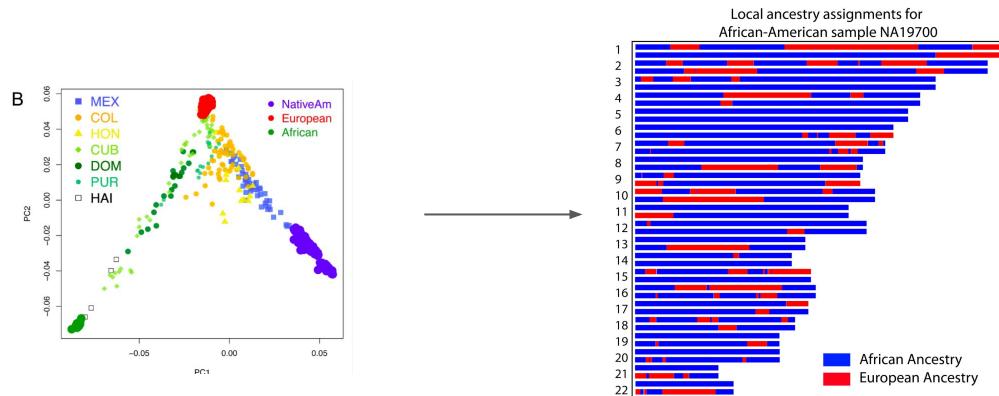
Se usan modelos para inferir la evolución de los segmentos de cromosomas (*haplotipos*) y calcular la **probabilidad que el haplotipo observado** corresponda a un determinado **haplotipo ancestral**.

Requieren como input haplotipos de poblaciones presuntamente ancestrales.

Ancestría Local

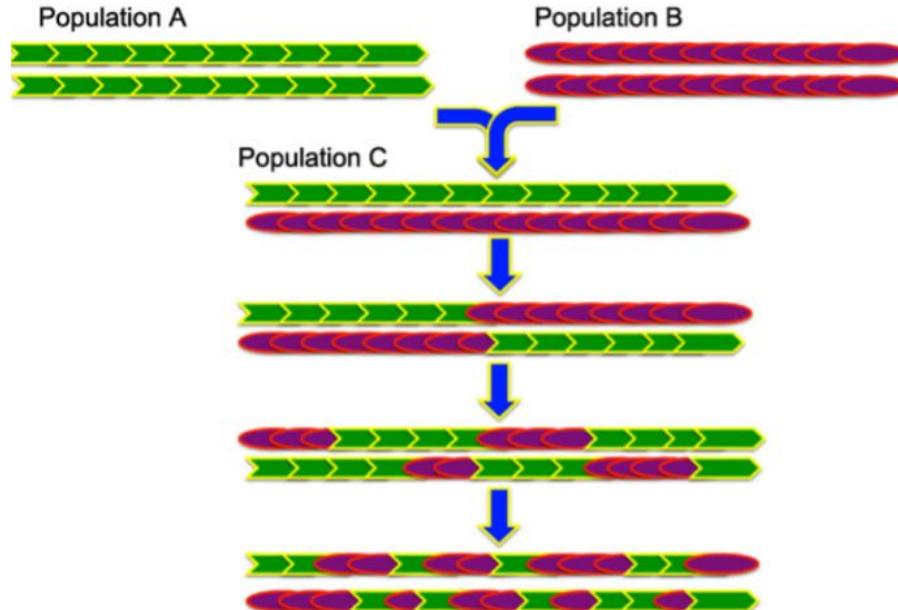
1. Algoritmos basados en agrupamiento:

Se basa en cortar los cromosomas en segmentos y a partir de algoritmos de agrupamiento (que tomen en cuenta recombinación) determinan la población ancestral de ese segmento.



Ancestría Local

2. Algoritmos basados en modelos poblacionales:



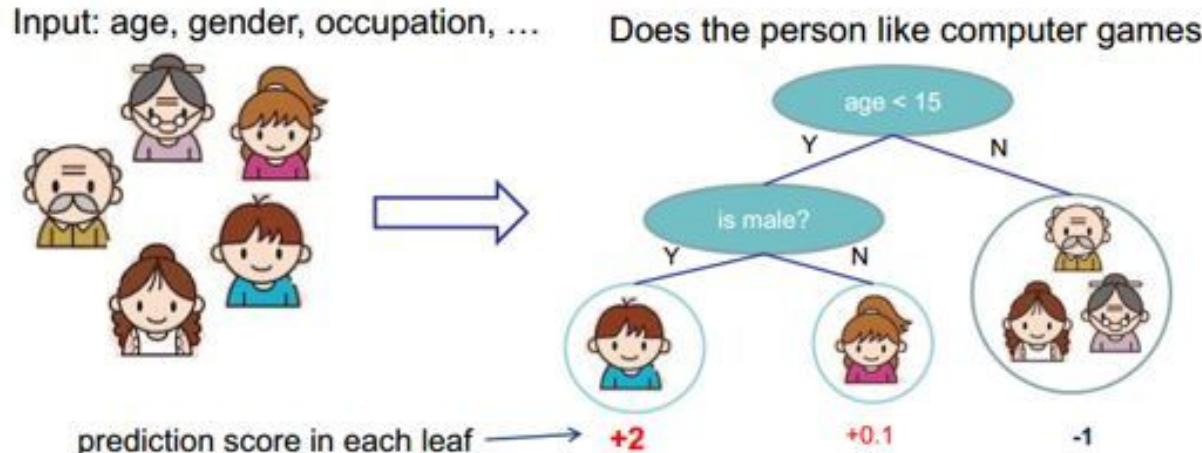
A. Martin. ASHG 2015

ARTICLE

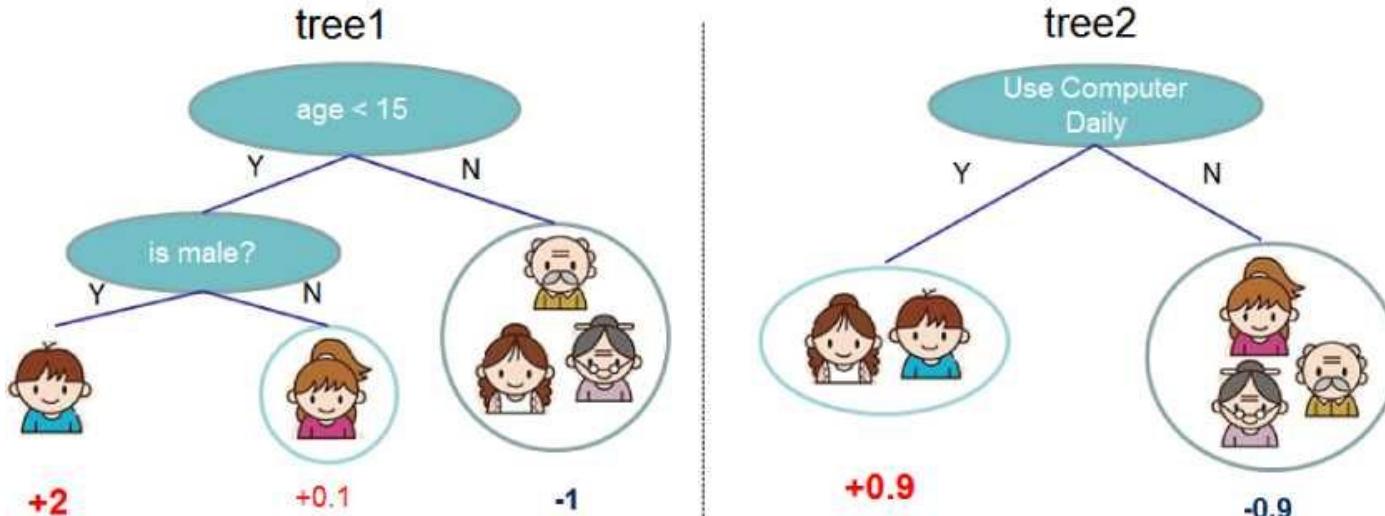
RFMix: A Discriminative Modeling Approach for Rapid and Robust Local-Ancestry Inference

Brian K. Maples,^{1,2} Simon Gravel,^{1,3} Eimear E. Kenny,^{1,4,5,6,7,8} and Carlos D. Bustamante^{1,8,*}

El algoritmo de **Random Forest** se basa en identificar cuáles son los principales atributos que dividen un grupo de otro.



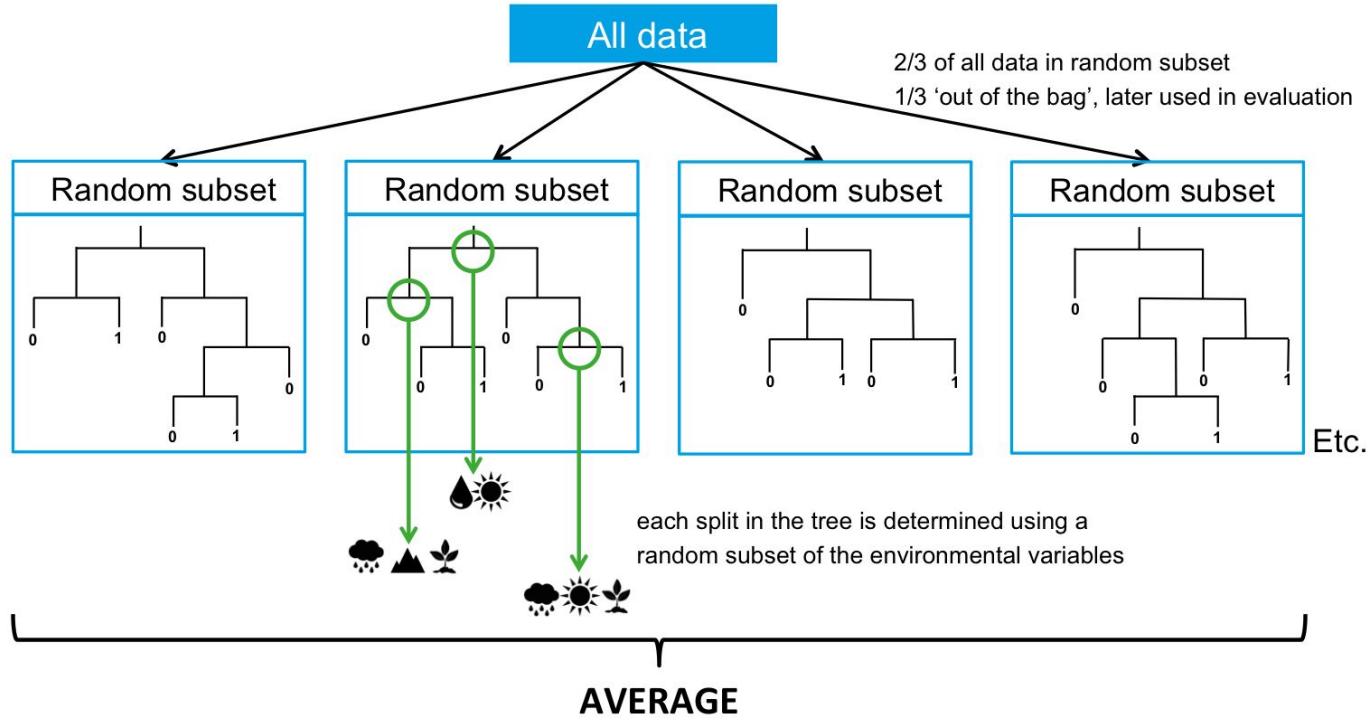
Para la identificación de los atributos principales, se construyen árboles de decisión al azar, tomando diferentes atributos y revisando su valor de predicción.



$$f(\text{boy}) = 2 + 0.9 = 2.9$$



$$f(\text{old man}) = -1 - 0.9 = -1.9$$

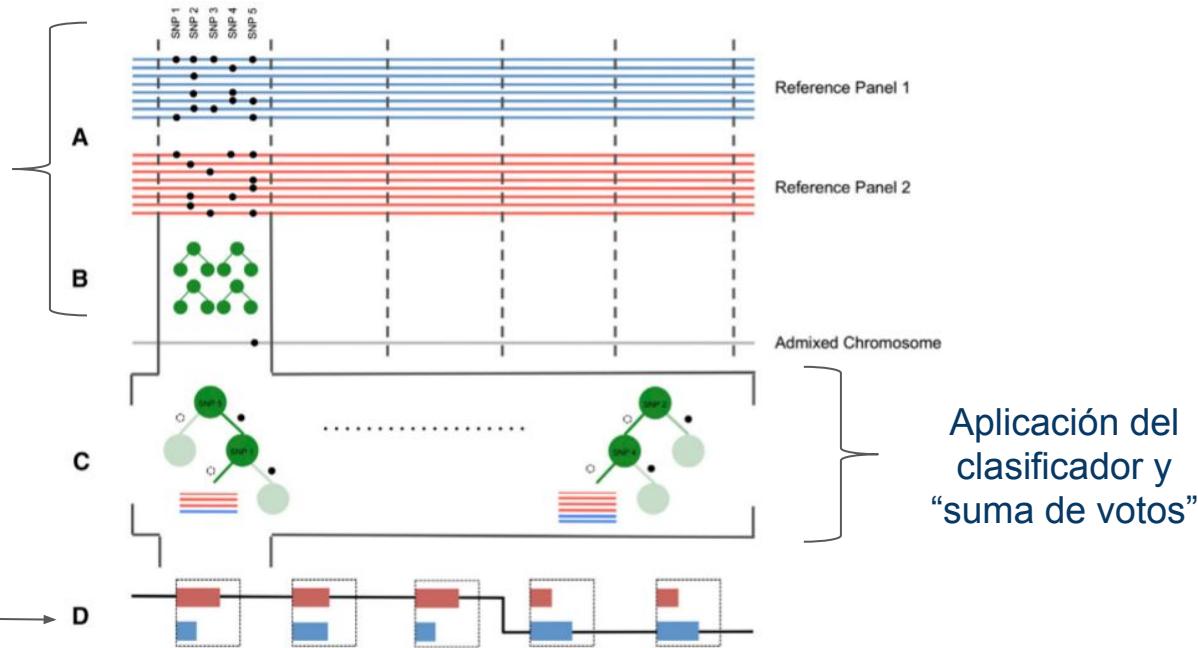


> find the set of predictor variables that produce the strongest classification model

RFMix basado en dividir por segmentos genomas de referencia y hacer un clasificador (usando **random forest**) con el que se clasifican los segmentos de genomas de los individuos de interés.

Entrenamiento y
creación del
clasificador

Asignación de la
ancestría por EM



1000 genomes project

- El 1000 genomes project cuenta con información de genoma completo para 2504 individuos de 26 poblaciones diferentes del mundo.
- Actualmente cuenta con información genómica usada como referencia para diferentes análisis en genética de poblaciones poblacionales.

IGSR: The International Genome Sample Resource

Providing ongoing support for the 1000 Genomes Project data

[Home](#)

[About](#)

[Data](#)

[Portal](#)

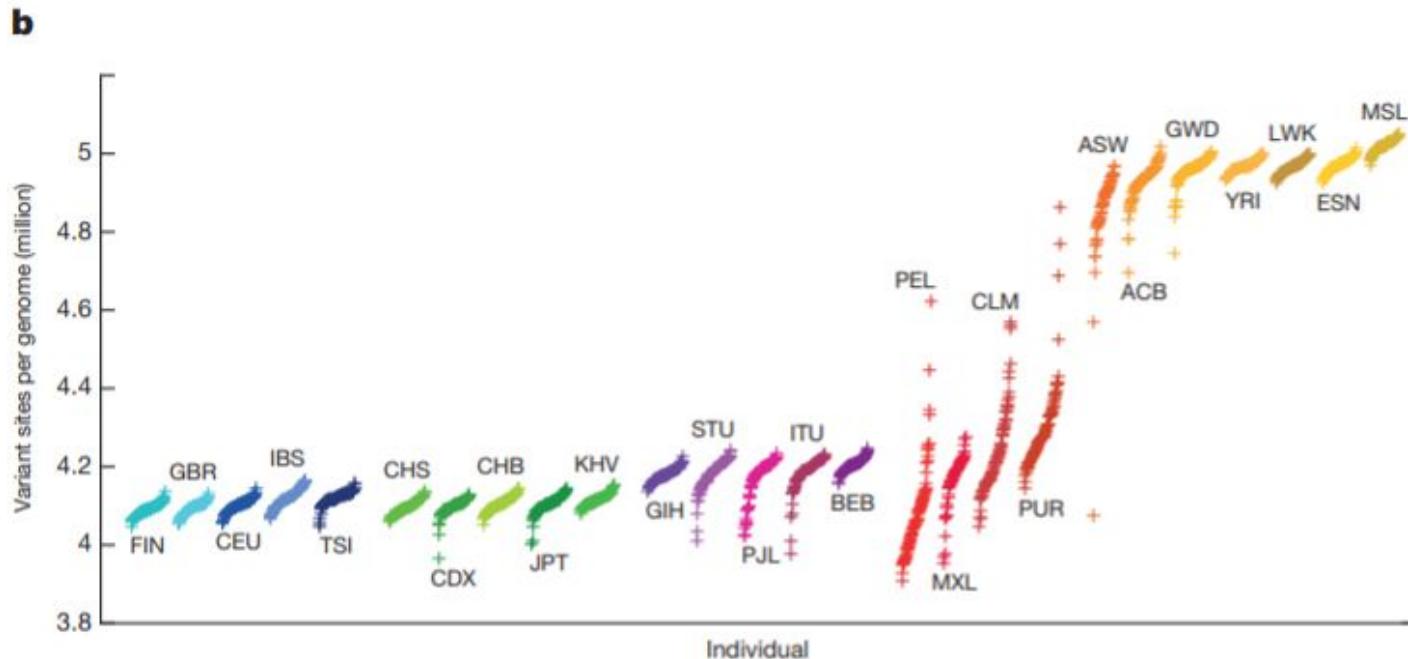
[Analysis](#)

[Contact](#)

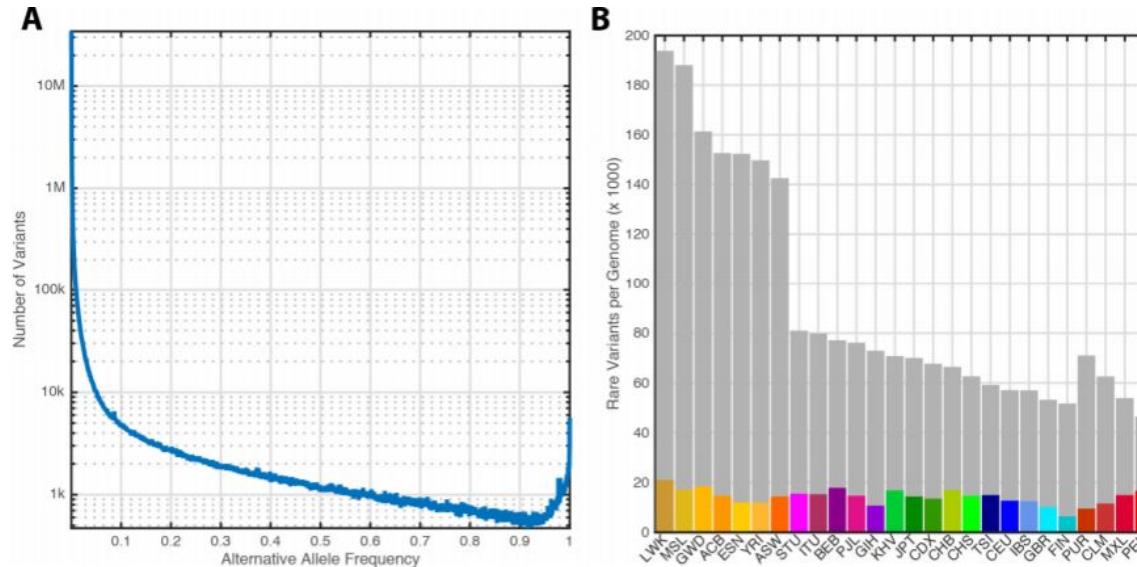
[Browser](#)

[FAQ](#)

- Por convención el alelo que porta es representado como variantes genéticas con respecto a un genoma de referencia.



Este gran catálogo de variación permitió caracterizar diferentes procesos pop-genéticos en las poblaciones humanas.



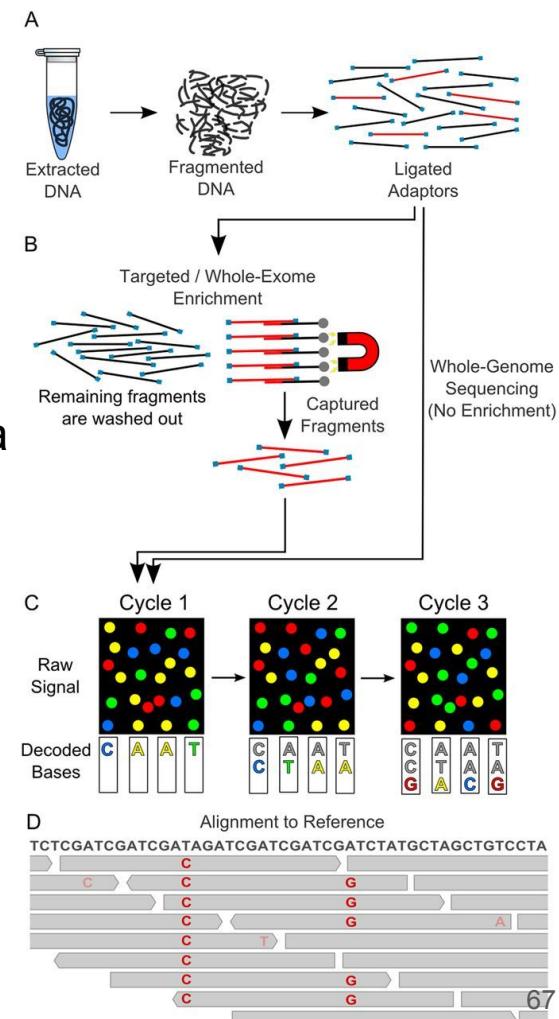
Extended Data Figure 3 | Variant counts. **a**, The number of variants within the phase 3 sample as a function of alternative allele frequency. **b**, The average number of detected variants per genome with whole-sample allele frequencies <0.5% (grey bars), with the average number of singletons indicated by colours.

Aplicaciones comunes de los datos del 1000 genomes:

- Referencias para imputar o fasear.
- Referencias para calcular ancestría.
- Consulta de frecuencias alélicas para genética médica (medicina de precisión).

En el portal del 1000 genomes project pueden ser localizados:

- Archivos que describen los alineamientos de lecturas contra el genoma de referencia (archivos BAM)
- Listado de variantes genéticas identificados con respecto a la referencia (archivos VCF).
- Otros archivos de soporte: Referencia genómica en formato FASTA, archivos resultantes de microarreglos.

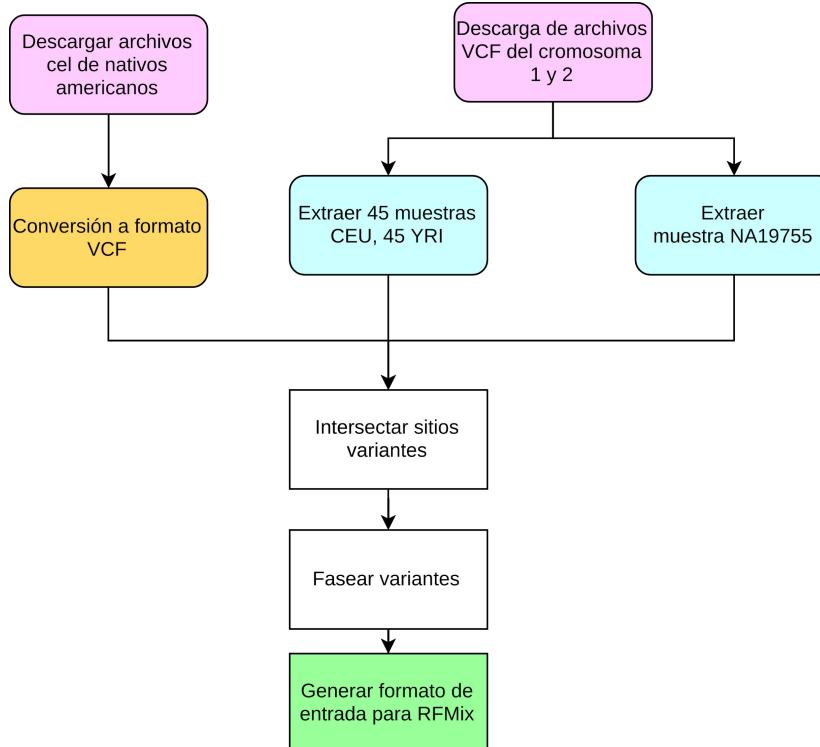


Práctica de RFMix

Para la práctica vamos a identificar ancestría local en una muestra del 1000 genomes project.

La muestra es un MXL con el ID: NA19755

Diagrama de flujo para la construcción del archivo de entrada de la práctica.



Archivos de entrada para **RFMix**:

1. **Haplotipos**: Matriz de 0 y 1, cada fila es una posición del genoma y cada columna el haplotipo de cada individuo.
2. **Markers**: Posiciones en centimorgans de cada variante en el archivo de haplotipos.
3. **Clases**: Archivo que describe cada la columna del archivo de haplotipos a cual población corresponde. El individuo de interés debe ser marcado con 0 0.

Realizar la práctica usando los comandos depositados en:

https://github.com/FRPV/Ancestry_class/tree/master/practica_RFMix