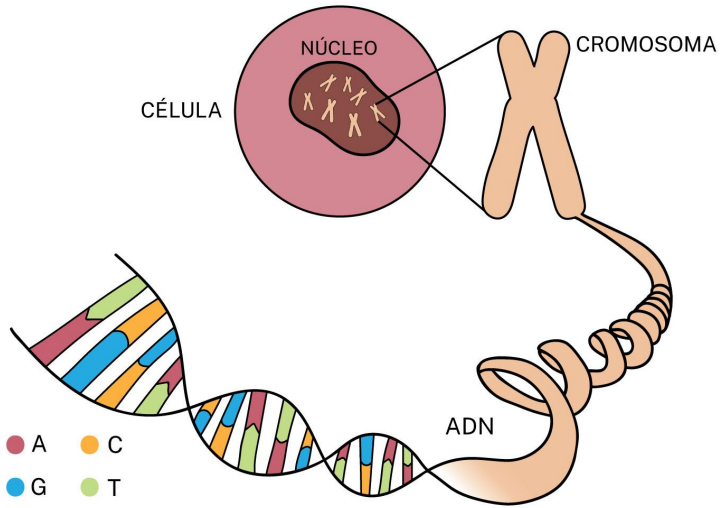
Curso MexVar
ALAB - Juan Comas 2025

Explorando la diversidad genética del Biobanco Mexicano

¿Qué es el ADN?

Las instrucciones que determinan todas las características y funciones de un organismo se encuentran en su material genético: **el ADN** (ácido desoxirribonucleico).

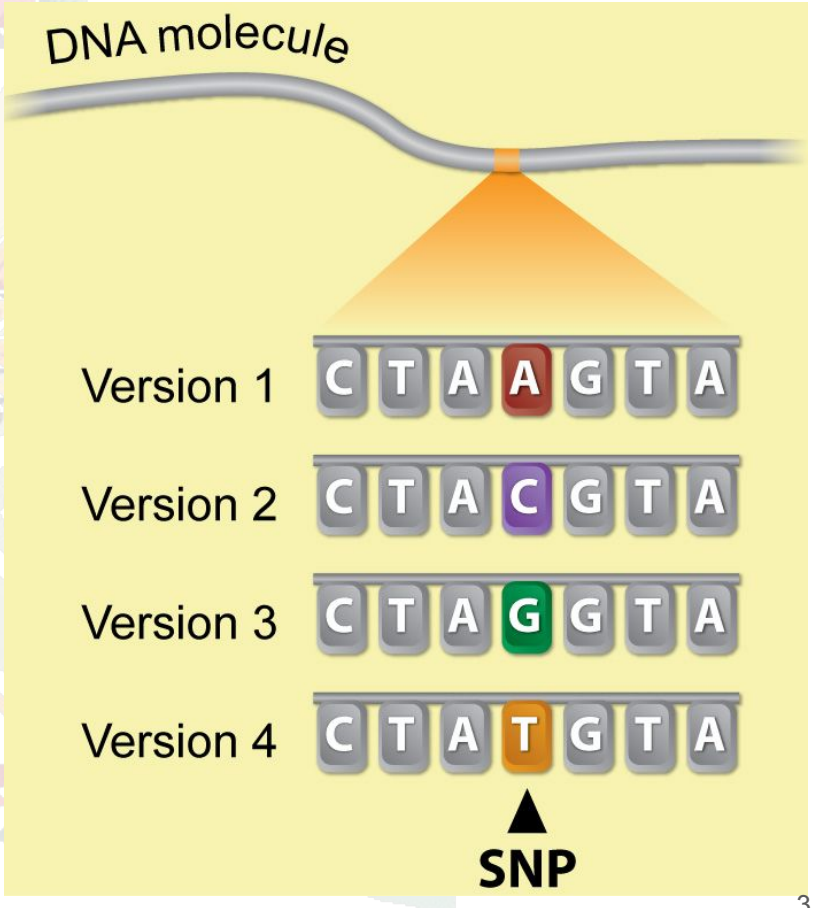


Función del ADN

- Guarda la información genética que determina la forma, características y funciones de los organismos
- Transmite esa información a la descendencia
- Presente en todas las células

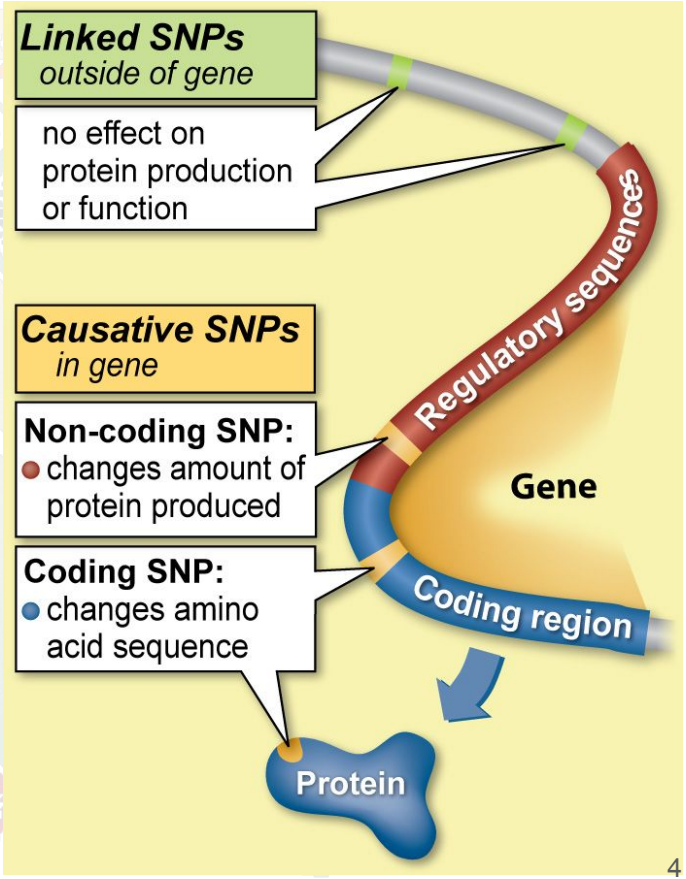
¿Qué es un SNP?

- SNP (“snip”) significa Polimorfismo de Nucleótido Único
- Es un cambio de una sola base en la secuencia del ADN (A, C, G o T)
- Para considerarse SNP, las variantes deben estar presentes en ≥1% de la población
- Son muy comunes: ~1 cada 300 pares de bases → ~10 millones en el genoma humano

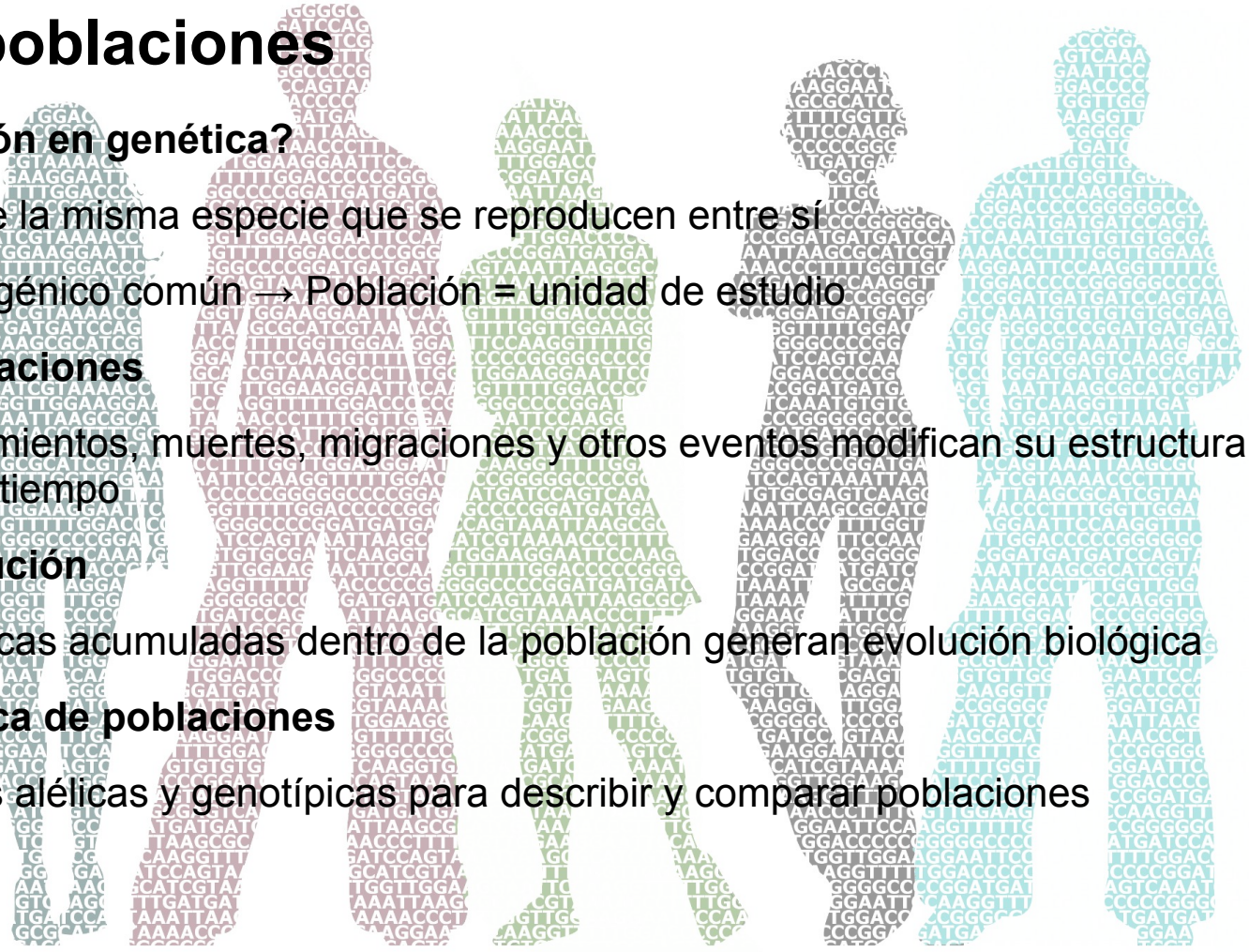


SNPs vs. Mutaciones que causan enfermedad

- Ambos son cambios de nucleótido, pero no son lo mismo
- SNP: presente en ≥1% de la población → no suelen ser tan comunes en mutaciones patogénicas
- Mutaciones causantes de enfermedad: ocurren en regiones codificantes o reguladoras y alteran la función génica o proteica
- SNPs pueden estar dentro o fuera de genes y no necesariamente afectan la función



Genética de poblaciones



¿Qué es una población en genética?

Grupo de individuos de la misma especie que se reproducen entre sí

Comparten un acervo génico común → Población = unidad de estudio

Dinámica de las poblaciones

No son estáticas: nacimientos, muertes, migraciones y otros eventos modifican su estructura genética a lo largo del tiempo

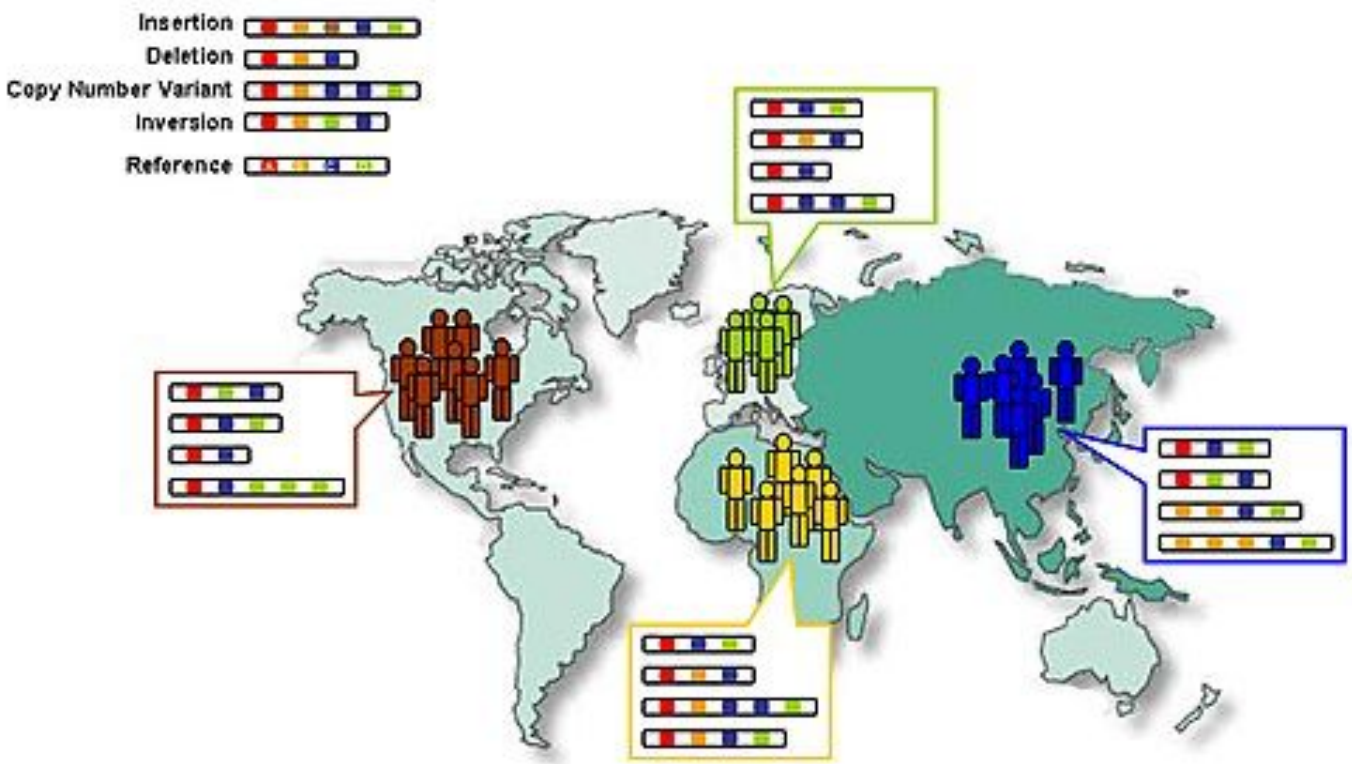
Relación con la evolución

Las variaciones genéticas acumuladas dentro de la población generan evolución biológica

Enfoque de la genética de poblaciones

Cuantificar frecuencias alélicas y genotípicas para describir y comparar poblaciones

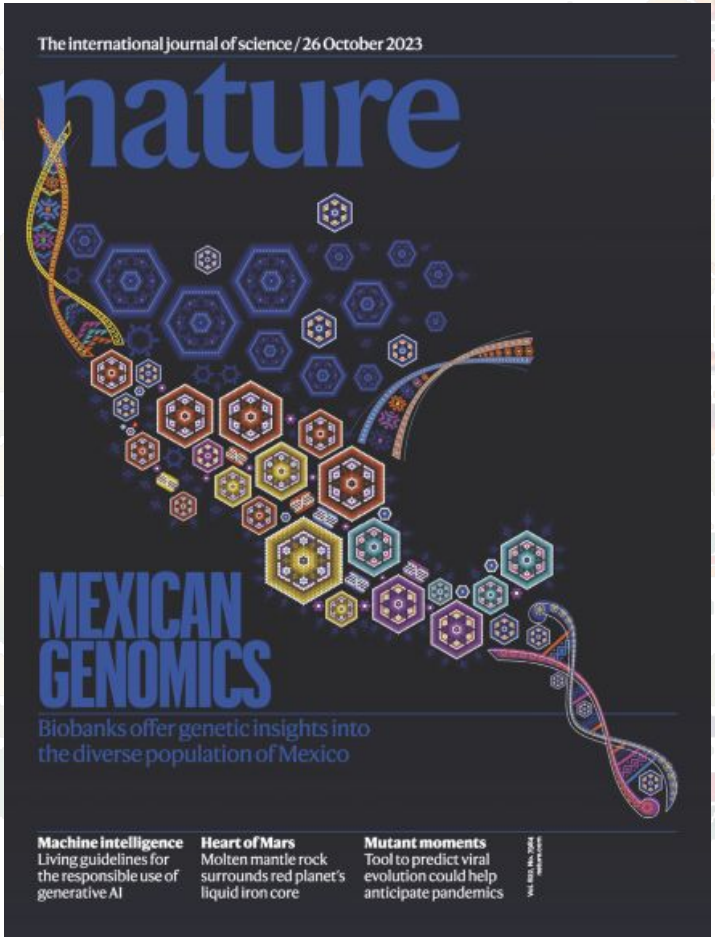
Genética de poblaciones humanas



Estudia cómo la variabilidad genética se distribuye y cambia en las poblaciones humanas a lo largo del tiempo

El Biobanco Mexicano: representación genómica de México

- Iniciativa pionera para comprender la **diversidad genética de la población mexicana** y su impacto en la salud y la evolución humana
- Analiza **más de 6,000 individuos** de regiones y contextos culturales diversos en todo el país
- Recurso genómico **único en América Latina**, desarrollado completamente en México
- Impulsa la investigación biomédica, apoya políticas de salud pública y promueve una visión global **más inclusiva de** la genética humana



El Biobanco Mexicano: Diversidad genética en México

Migraciones hacia Nueva España (S. XVI–XVIII)

- **Española** (voluntaria): Llegaron tras la conquista, se asentaron en el Bajío y centro; pocos en número (~1400 en S. XVII) pero con gran influencia política, económica y eclesiástica.
- **Africana** (forzada): Personas esclavizadas traídas para trabajo doméstico y agrícola (caña, algodón, tabaco), especialmente en Veracruz, Oaxaca, Guerrero, Yucatán y Campeche.
- **Asiática** (comercial): Llegó vía la Nao de China desde Filipinas (China, Malasia, Indonesia). Conocidos como “indios chinos” y sujetos a tributo como los pueblos indígenas.

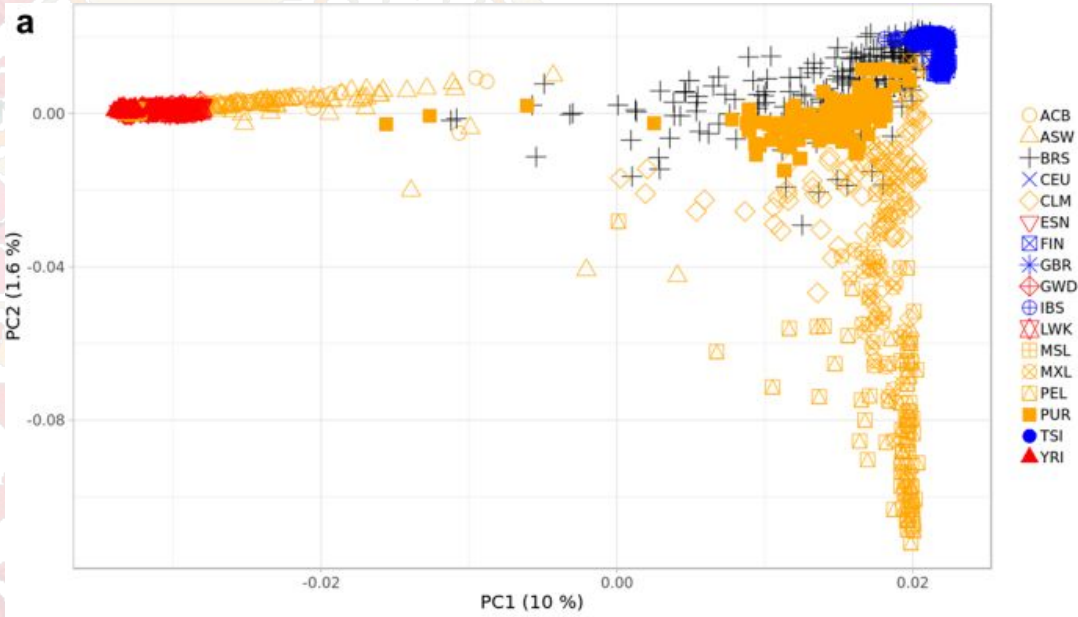
🌎 Análisis de ancestría Global

- Estima la proporción de origen genético de cada individuo a grandes escalas (continental o regional)
- Resume el genoma completo en porcentajes de contribución (p. ej., 40% Europea, 50% Indígena, 10% Africana)
- Útil para estudios poblacionales, historia demográfica y control de confusión en análisis biomédicos

📌Herramientas para estudiar la ancestría global

PCA (Análisis de Componentes Principales)

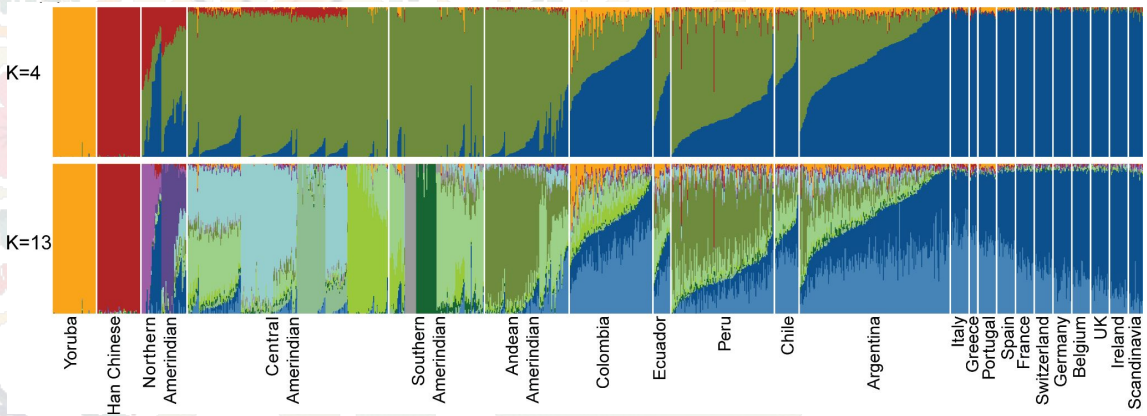
- Método estadístico que reduce dimensiones y resume la variación genética
- Representa individuos en un mapa genético continuo donde la distancia refleja similitud
- Captura estructura poblacional, mezcla, aislamiento por distancia y subpoblaciones



📌Herramientas para estudiar la ancestría global

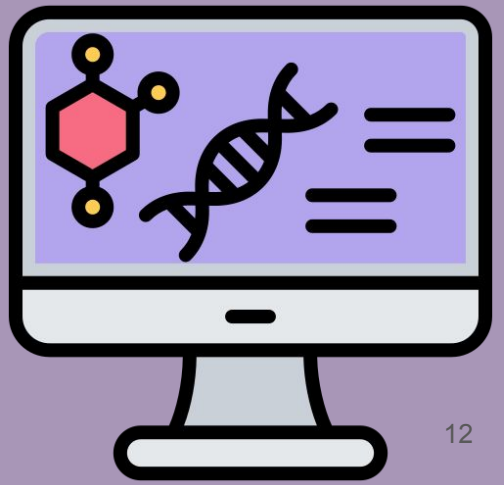
ADMIXTURE

- Modelo que estima la proporción de cada componente ancestral en individuos
- Asume un número K de poblaciones ancestrales y asigna proporciones de mezcla
- Útil para identificar patrones de mezcla, divergencia y migración en poblaciones humanas



Parte teórica - práctica

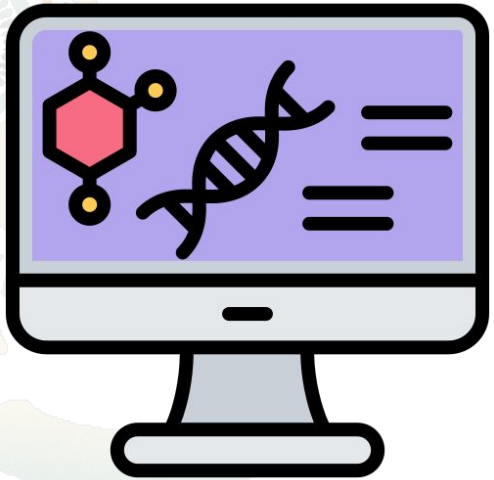
Explorando la diversidad genética del Biobanco Mexicano



📌 Objetivo 1: análisis de calidad a un dataset de genotipado

📌 PLINK: Herramienta para análisis genómicos

- Software para manejar y analizar datos genéticos a gran escala
- Usado en genética de poblaciones, GWAS y estudios biomédicos
- Permite filtrar, limpiar y formatear datos SNP
- Estándar en investigación genómica global



📌Archivos principales en PLINK (formato binario)

- Terminación **.bed** → matriz binaria de genotipos (0/1/2), datos SNP comprimidos
- Terminación **.bim** → información de marcadores (cromosoma, ID, posición, alelos)
- Terminación **.fam** → lista de individuos con IDs, sexo y fenotipo

📌 Cómo activar el ambiente de trabajo

```
#Abrir aplicación Terminal y teclear
conda activate popgen
```

📌 Explorar archivos plink

```
# Muestra las primeras 10 líneas del archivo .bim

# .bim contiene la información por SNP: CHR, ID, POS_GENÉTICA, POS_FÍSICA, A1, A2

head poptest.bim

# .fam tiene 6 columnas: FID IID PADRE MADRE SEXO FENOTIPO

head poptest.fam

# Cuenta cuántas filas tiene .fam = número de muestras

wc -l poptest.fam

# Cuenta filas en .bim = número de variantes

wc -l poptest.bim
```

Objetivo del QC: garantizar que los datos genotípicos utilizados en análisis poblacionales no estén sesgados por errores técnicos, baja calidad o individuos problemáticos (parientes, contaminación, baja cobertura).

📌 **Paso 1 QC:** quitar SNPs problemáticos como variantes duplicadas, variantes estructurales.

📌 **Paso 2 QC:** evaluar "missingnes" de los datos (por individuo y por variante)

Control de calidad (QC)

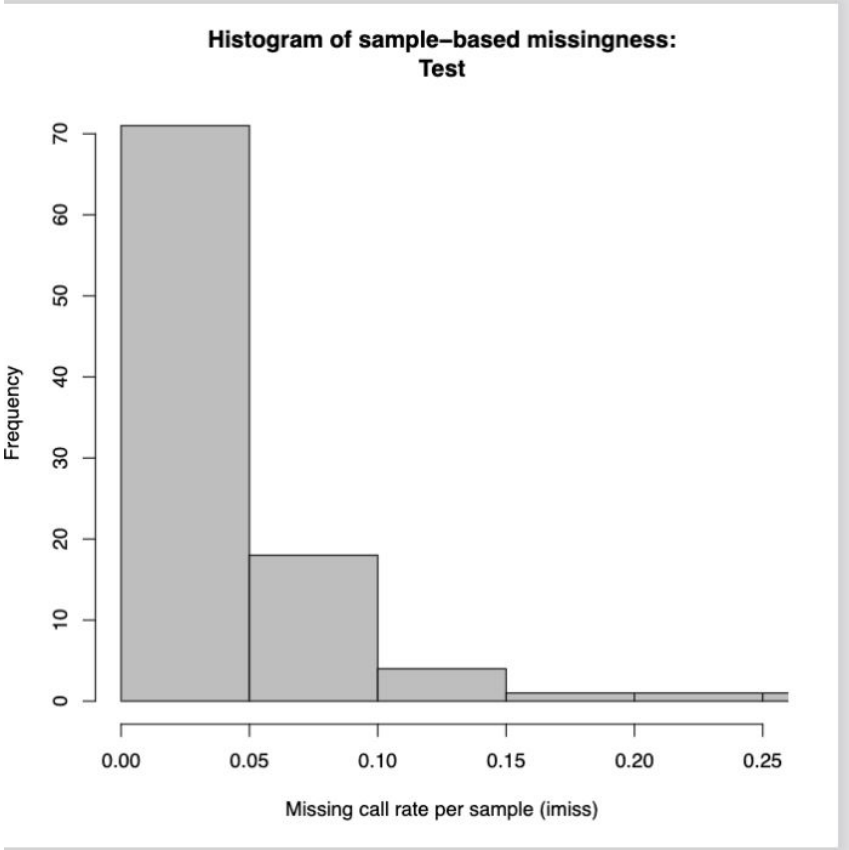
Histogram of sample-based missingness (imiss)

Este histograma muestra **qué tan incompletos están los genotipos por individuo**.

- Eje X → proporción de genotipos faltantes **por muestra**
- Eje Y → número de individuos con ese nivel de missingness

Interpretación:

- La mayoría de los individuos tienen **<5% missing**, lo cual es bueno.
- Hay algunos individuos con ~0.1 (10%) y pocos con >0.15 → estos suelen eliminarse.



Control de calidad (QC)

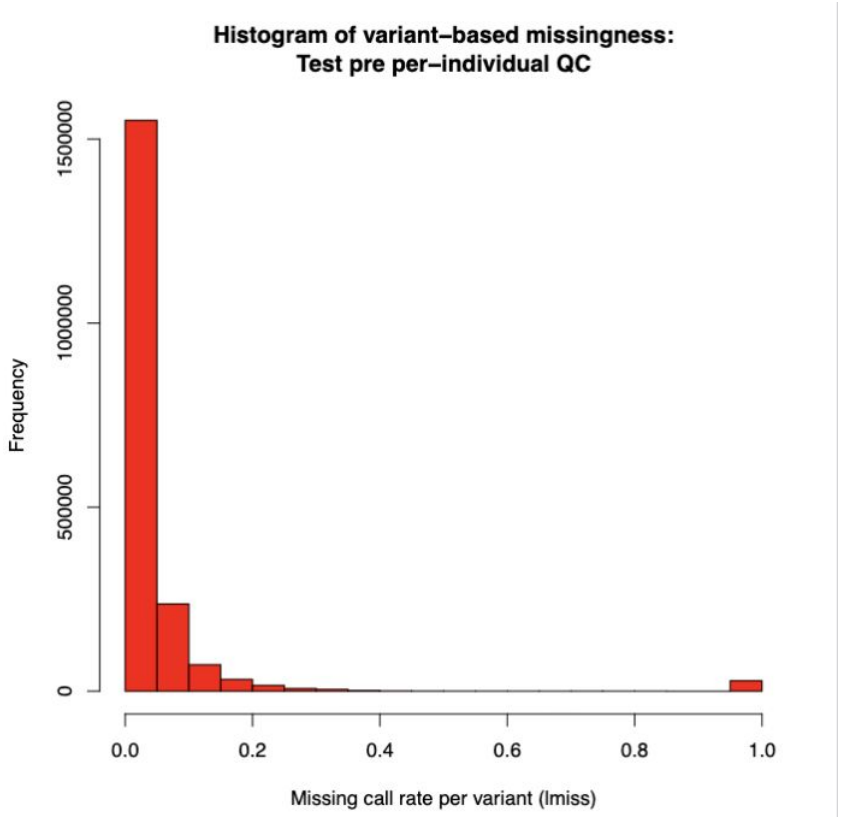
Histogram of variant-based missingness (per SNP)

Este gráfica **qué tan incompletos están los SNPs a lo largo de todas las muestras**.

- X → proporción de genotipos faltantes **por marcador**
- Y → frecuencia (cuántos SNPs tienen ese nivel)

Interpretación:

- La mayoría de los SNPs tienen missingness muy bajo (<0.05).
- Pero hay una cola larga hacia SNPs muy faltantes, incluso algunos con ~100% missing (no genotipados en nadie).



Control de calidad (QC)

Heterocigosidad vs. proporción de genotipos faltantes

Cada punto representa un individuo.

El eje X (Proportion of missing genotypes) muestra el porcentaje de genotipos no llamados: **Valores altos indican baja calidad o problemas de llamado.**

Líneas de referencia indican umbrales típicos de exclusión:
3% (morado) → QC estricto 5% (verde) → moderado 10% (naranja) → permisivo

El eje Y (Heterozygosity rate) indica la p**roporción de sitios heterocigotos.**
Líneas rojas marcan el rango esperado en muestras sanas.

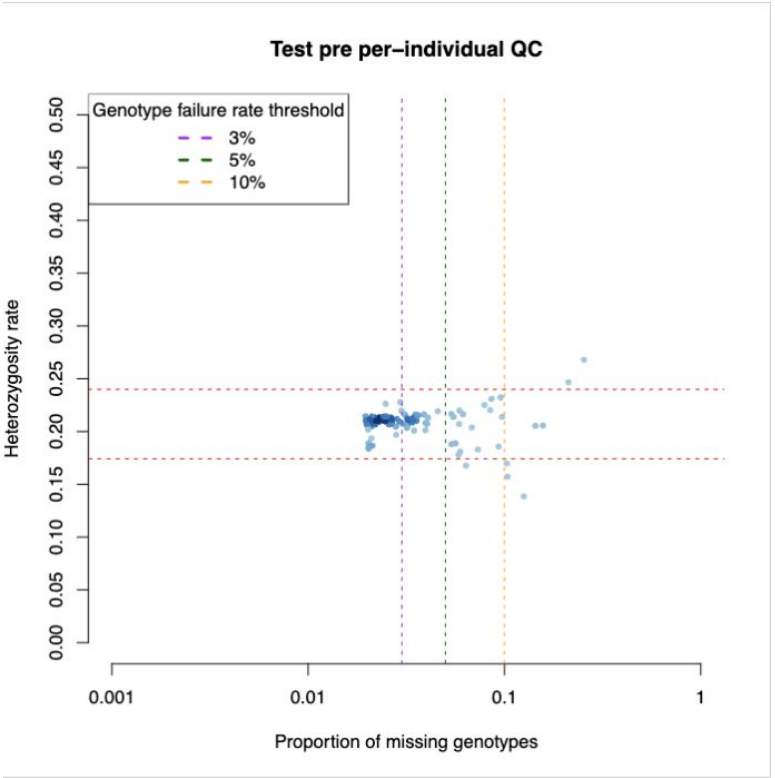
Desviaciones sugieren problemas:

↑ Alto → posible contaminación, mezcla de individuos, errores de llamado.
↓ Bajo → endogamia, baja cobertura

Interpretación general:
La mayoría de los individuos caen dentro del rango aceptable → **buena calidad**.

Individuos con:

>5–10% missing → candidatos a excluir por mala calidad.
Heterocigosidad >0.25 → sospecha de contaminación.

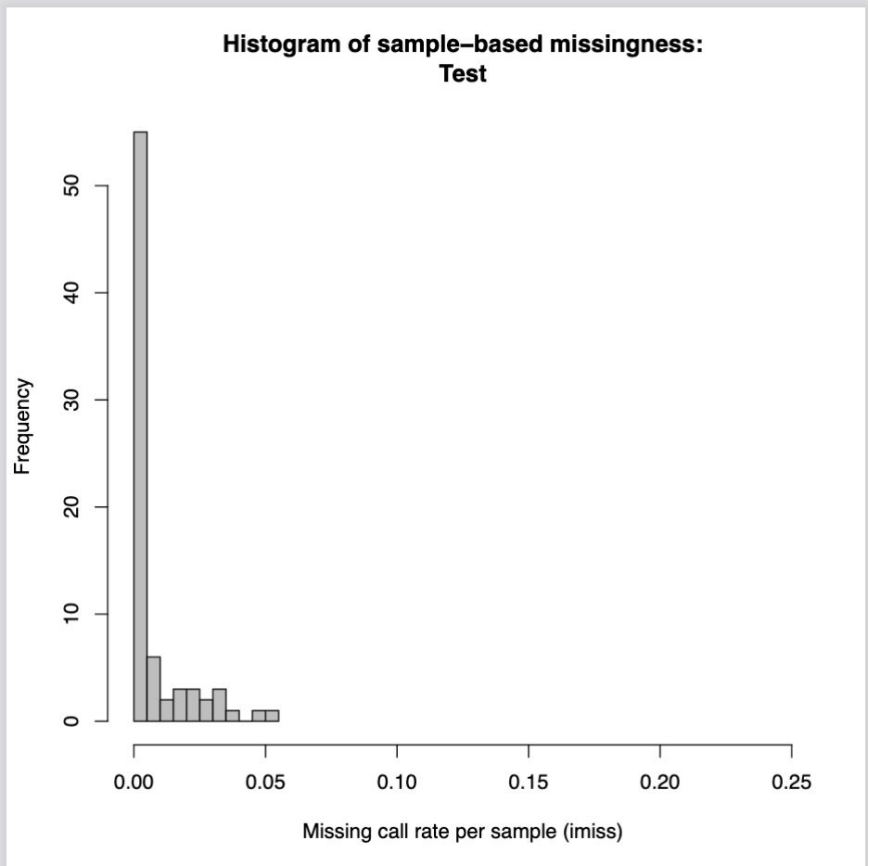


📌 Después de los filtros…

Control de calidad post filtros

Distribución de genotipos faltantes por individuo.

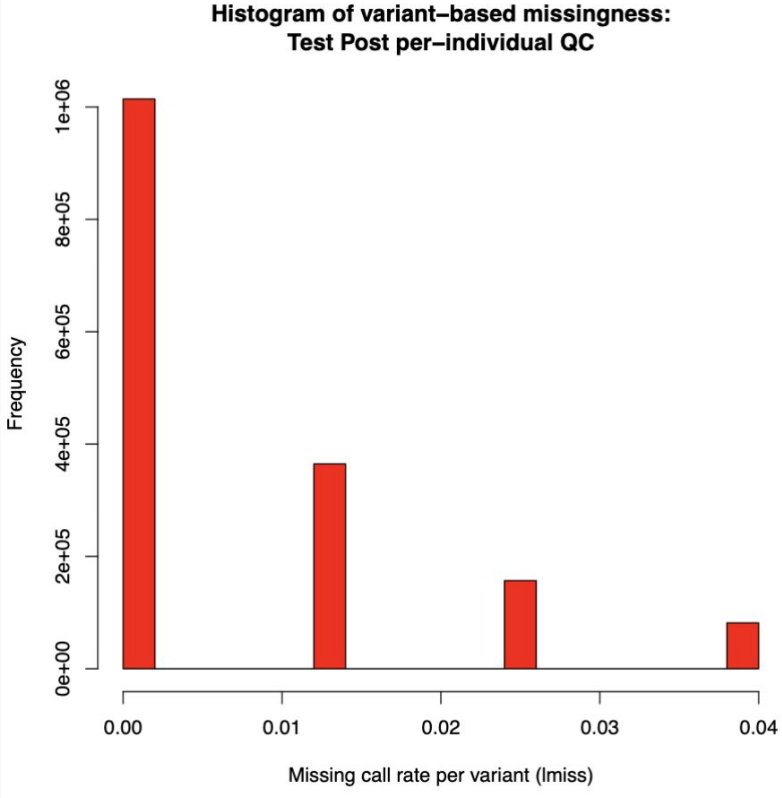
- La mayoría de las muestras tienen <1% missing → alta integridad genotípica.
- Pocas muestras en el rango 2–7%, ninguna >10% tras filtrado.
- Los filtros removieron individuos de baja calidad → dataset limpio para análisis poblacionales.



Control de calidad post filtros

Distribución de datos faltantes por SNP tras filtrar individuos.

- La mayoría de los sitios tienen ~0% missing → variante bien llamada en todas las muestras retenidas.
- Se observan pocos SNPs con 1–4% missing; no hay variantes con missingness alto.
- Confirma mejora en calidad global del dataset y aptitud para análisis poblacionales.



✅ **Paso 1 QC**: Quitar SNPs problemáticos como variantes duplicadas, variantes estructurales

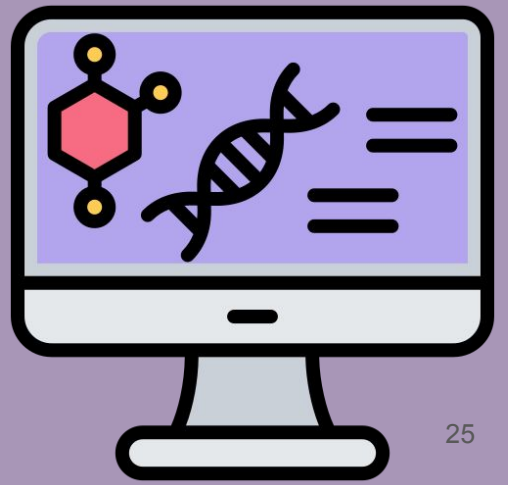
✅ **Paso 2 QC**: Evaluar "missingnes" de los datos (por individuo y por variante)

📌 **Paso 3 QC:** Identificar individuos relacionados

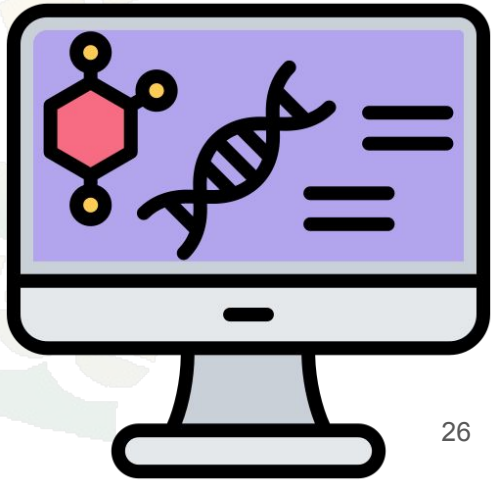
📌 **Paso 4 QC:** Aplicar filtros

Parte teórica práctica

BREAK



📌 Objetivo 2: análisis de ancestría global con PCA y ADMIXTURE



¿Qué es LD y por qué hacemos LD pruning?

- ¿Qué es Linkage Disequilibrium (LD)?

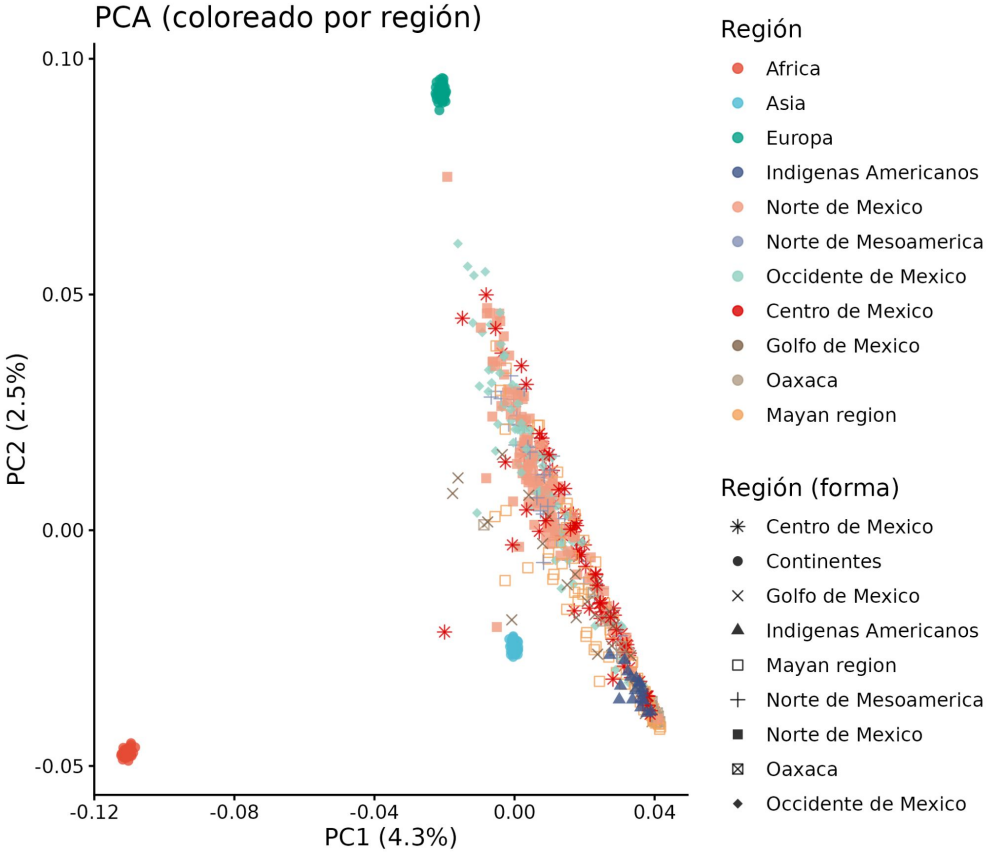
El LD ocurre cuando dos variantes genéticas no se heredan de manera independiente, sino que están correlacionadas porque están físicamente cercanas en el ADN o afectadas por selección.

Es decir, la frecuencia de un alelo depende del otro.

- **¿Por qué hacemos LD pruning?**
  - Para quedarnos con un subconjunto de SNPs independientes.
  - Reduce señales redundantes
  - Mejora interpretaciones en PCA y ADMIXTURE

Resultados de PCA

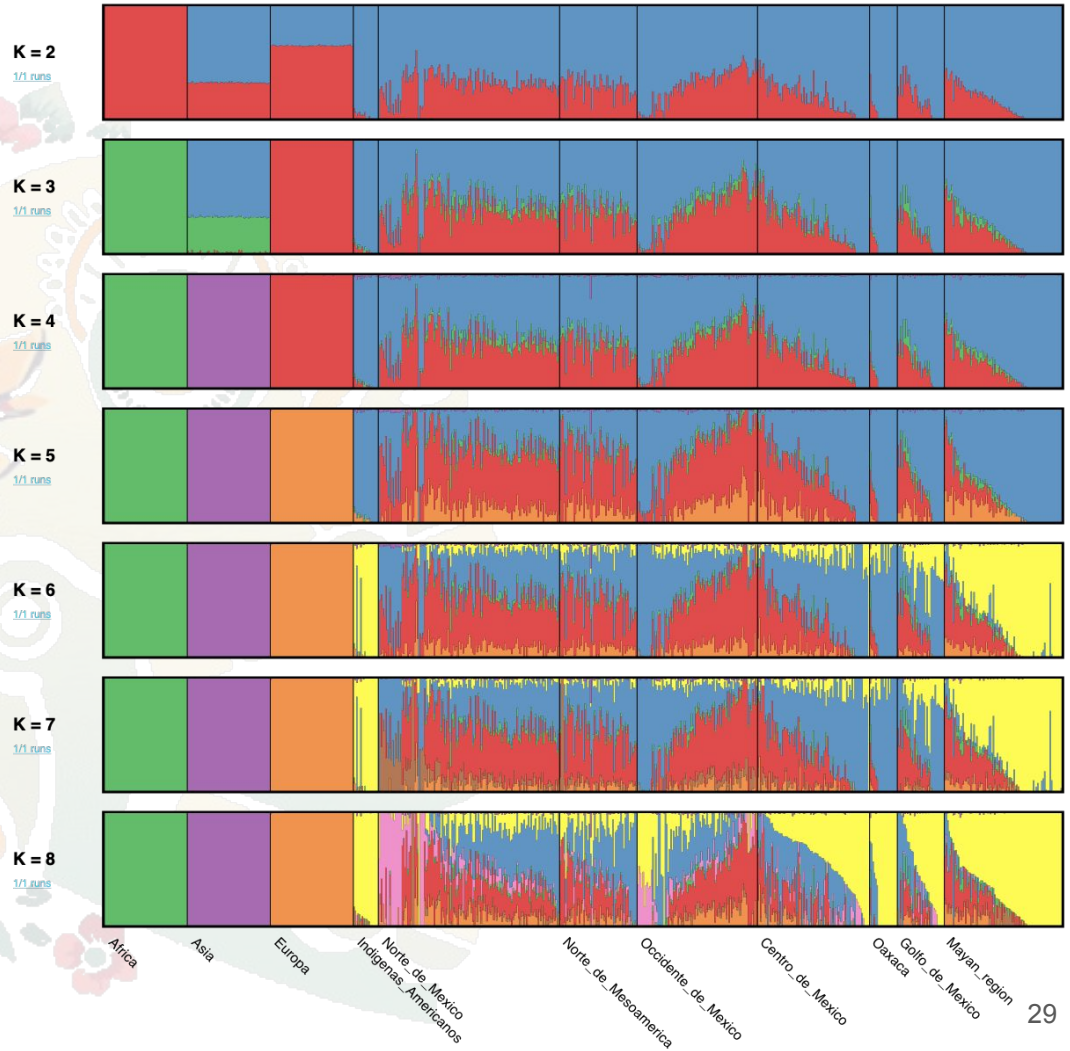
- Proyección de los genotipos en unas pocas componentes principales que resumen la variación genética (PC1 = 4.3%, PC2 = 2.5%).
- Cada punto es un individuo.
- El color indica la región geográfica y la forma distingue continentes vs distintas regiones de México.
- **Patrones principales**
  - África, Asia y Europa aparecen separados en los extremos del espacio de PCs.
  - Los individuos de México forman un gradiente continuo, donde se observa la variación en la mezcla entre ancestría indígena americana y ancestrías continentales.



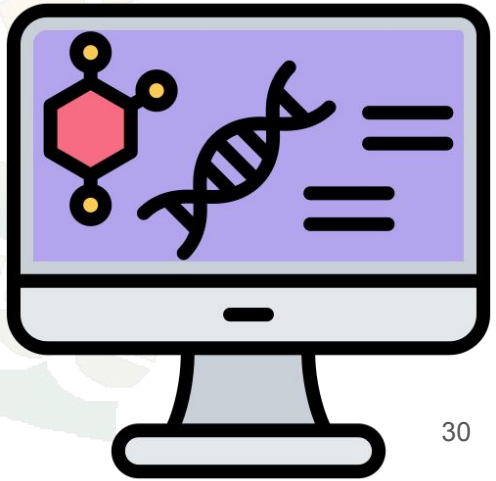
- El PCA captura tanto las diferencias a gran escala entre continentes como la estructura fina entre regiones mexicanas.

Resultados de ADMIXTURE

- Es un método de inferencia de estructura genética que asigna a cada individuo proporciones de ascendencia provenientes de K componentes inferidos de forma no supervisada.
- A valores bajos de K se observan grandes clústeres continentales (e.g., África, Asia, Europa, América), mientras que valores más altos permiten distinguir subdivisiones regionales dentro de México y Mesoamérica.
- Cada barra es un individuo, coloreado según su proporción de ancestría; se ordenaron según su región geográfica para facilitar la interpretación.



📌 Objetivo 2: análisis de ancestría global con PCA y ADMIXTURE



🔬 Ancestría global vs. ancestría local

📍 Ancestría global

Métodos como **ADMIXTURE o PCA** estiman la proporción total de ancestría a nivel del genoma completo.

- Resumen la mezcla histórica de un individuo como un promedio general.
- Útiles para ver patrones poblacionales amplios (e.g., % europeo vs. % indígena americana).
- No indican dónde en el genoma ocurre cada ancestría.

Ejemplo: 40% europeo, 60% indígena → estimado globalmente.

📍 Ancestría local

La ancestría se asigna segmento por segmento a lo largo del genoma, determinando de qué población proviene cada tramo heredado.

- Requiere información haplotípica

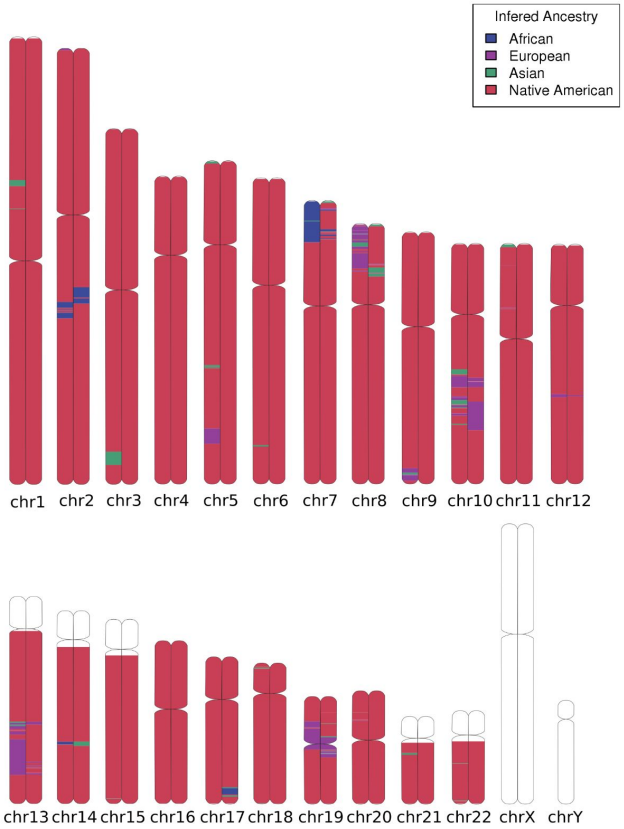
(un haplotipo = conjunto de variantes cercanas heredadas juntas como bloque)

- Permite analizar cromosomas

completos y recombinación.

- Métodos: RFMix, GNOMIX, etc

Ejemplo: región en chr6 es 100% indígena; brazo largo del chr2 es europeo.



Curso MexVar
ALAB - Juan Comas 2025

Explorando la diversidad genética del Biobanco Mexicano