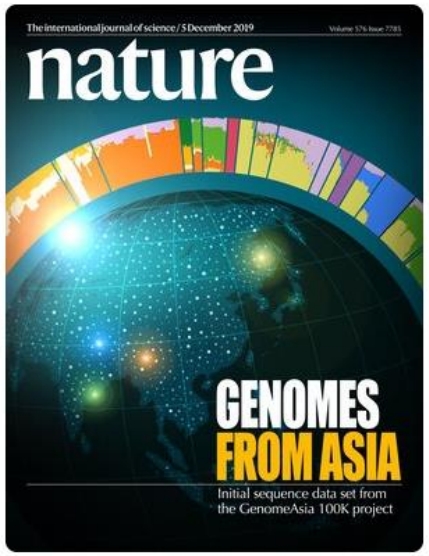
Análisis secuencias de ADN

Guillermo Barturen Briñas
(gbarturen@ugr.es)

¿Qué es un genoma de referencia?

- Un genoma de referencia es una representación del genoma de una especie que se usa por los investigadores como referencia para comparar las secuencias de ADN generadas durante sus estudios.
- En el caso del genoma humano, a parte del ensamblado T2T que proviene de una línea celular haploide (CHM13htert), el último ensamblado aceptado por la comunidad se denomina GRCh38 y casi en su totalidad (93%) proviene de 11 donantes anónimos.
- El donante RP11 (70% de la secuencia primaria) presenta una ascendencia mixta africana y europea, pero este ensamblado presenta un gran componente europeo.



Sequencing of African populations aims to make genomics more representative

By David Adam | **March 17, 2023**

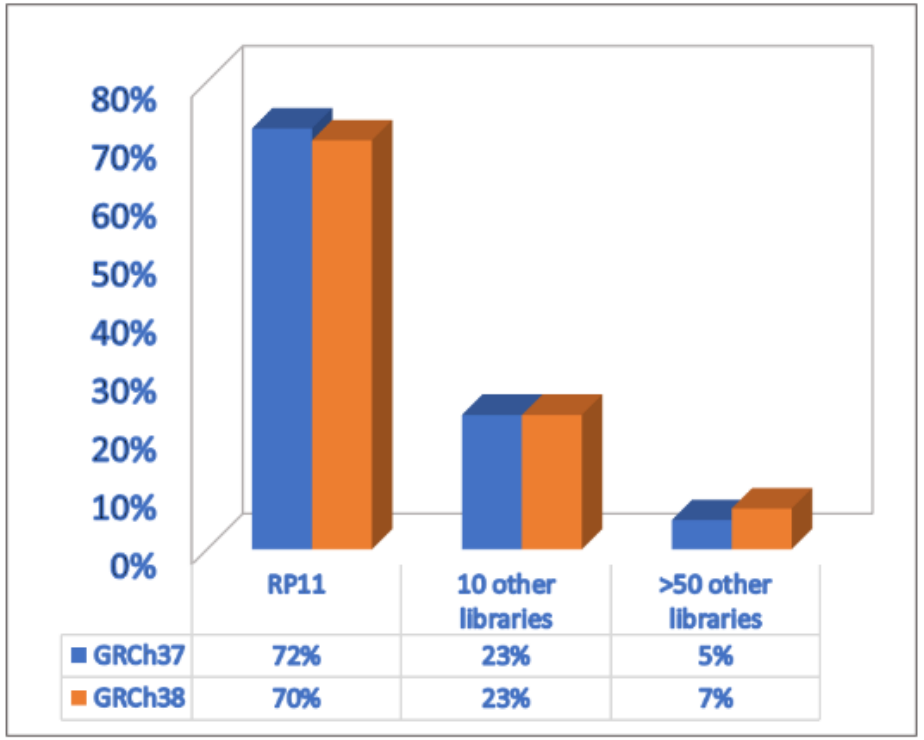


Figure 1. Contribution of genomic libraries to GRCh37 and GRCh38.

Tipos de variación de secuencia

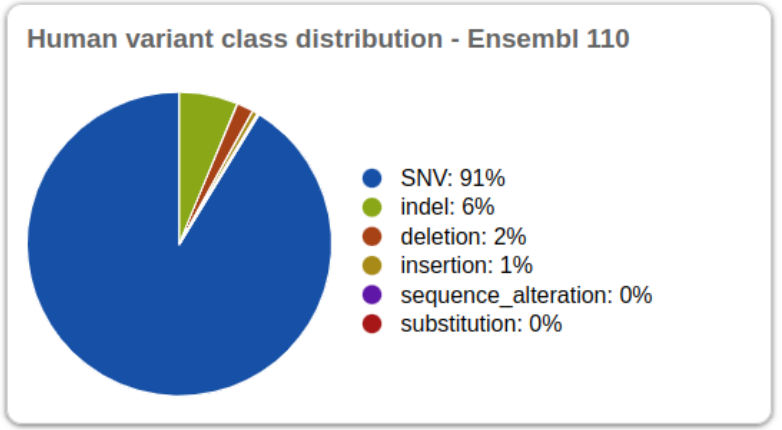
Ensembl Variation - Variant classification

Sequence variants

Type	Description	Example (Reference / Alternative)
SNP	Single Nucleotide Polymorphism	Ref: . . . TTG**A**CGTA . . . Alt: . . . TTG**G**CGTA . . .
Insertion	Insertion of one or several nucleotides	Ref: . . . TTGACGTA . . . Alt: . . . TTGA**TG**CGTA . . .
Deletion	Deletion of one or several nucleotides	Ref: . . . TTG**AC**GTA . . . Alt: . . . TTGGTA . . .
Indel	An insertion and a deletion, affecting 2 or more nucleotides	Ref: . . . TTG**A**CGTA . . . Alt: . . . TTG**GCT**CGTA . . .
Substitution	A sequence alteration where the length of the change in the variant is the same as that of the reference.	Ref: . . . TTG**AC**GTA . . . Alt: . . . TTG**TA**GTA . . .

Structural variants

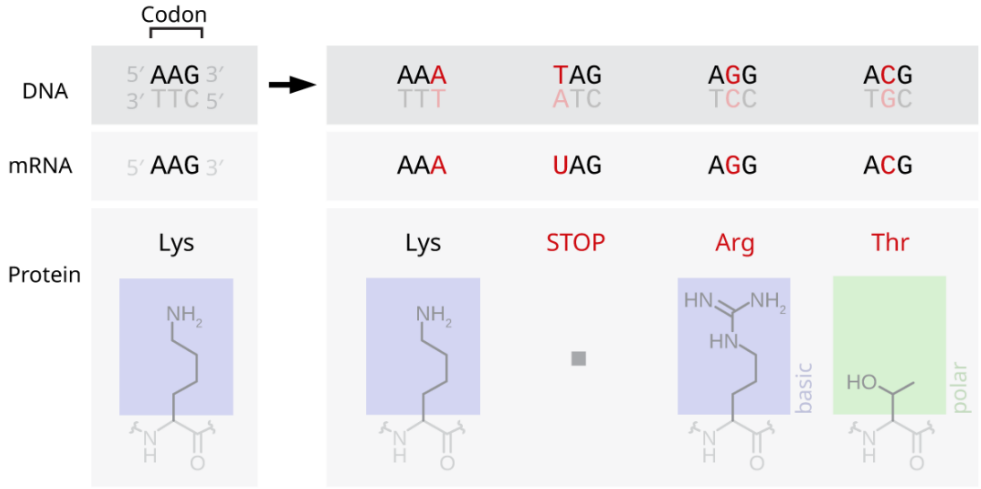
Type	Description	Example (Reference / Alternative)
CNV	Copy Number Variation: increases or decreases the copy number of a given region	Reference:  "Gain" of one copy:  "Loss" of one copy: 
Inversion	A continuous nucleotide sequence is inverted in the same position	Reference:  Alternative: 
Translocation	A region of nucleotide sequence that has translocated to a new position	Reference:  Alternative: 

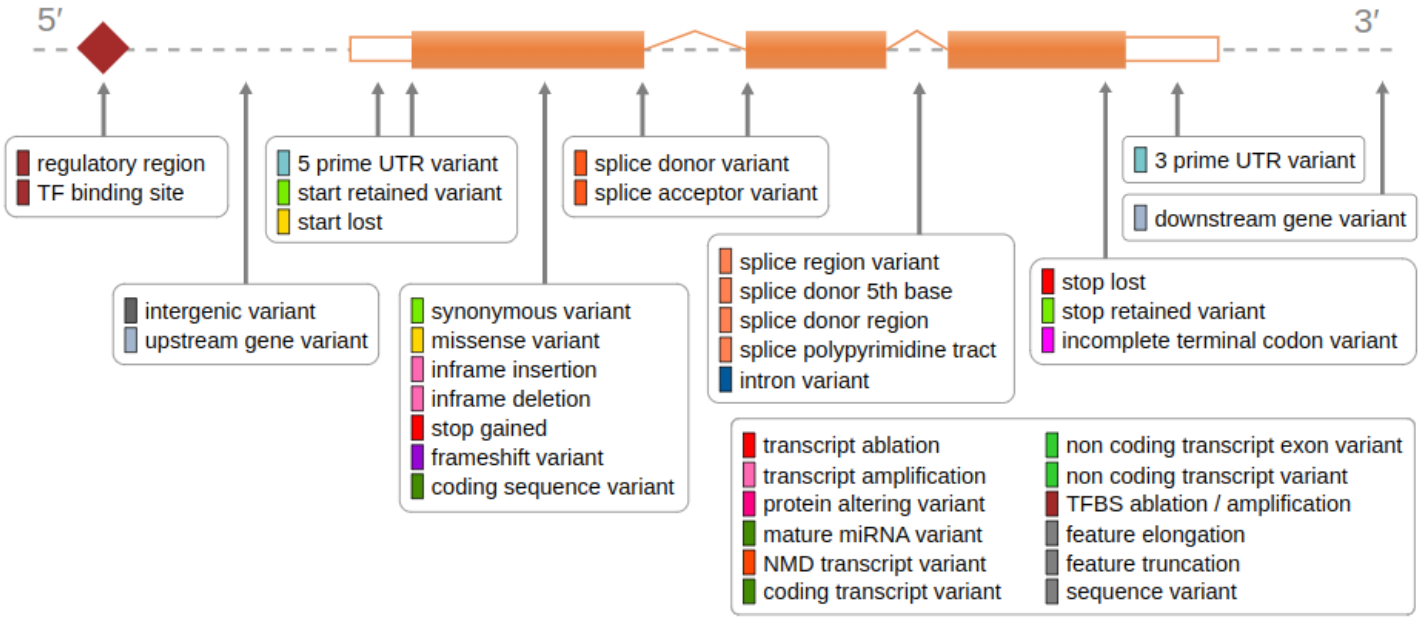


4-5 millones de variantes entre un individuo y el genoma de referencia (99.9% SNPs)

Consecuencias variantes de secuencia

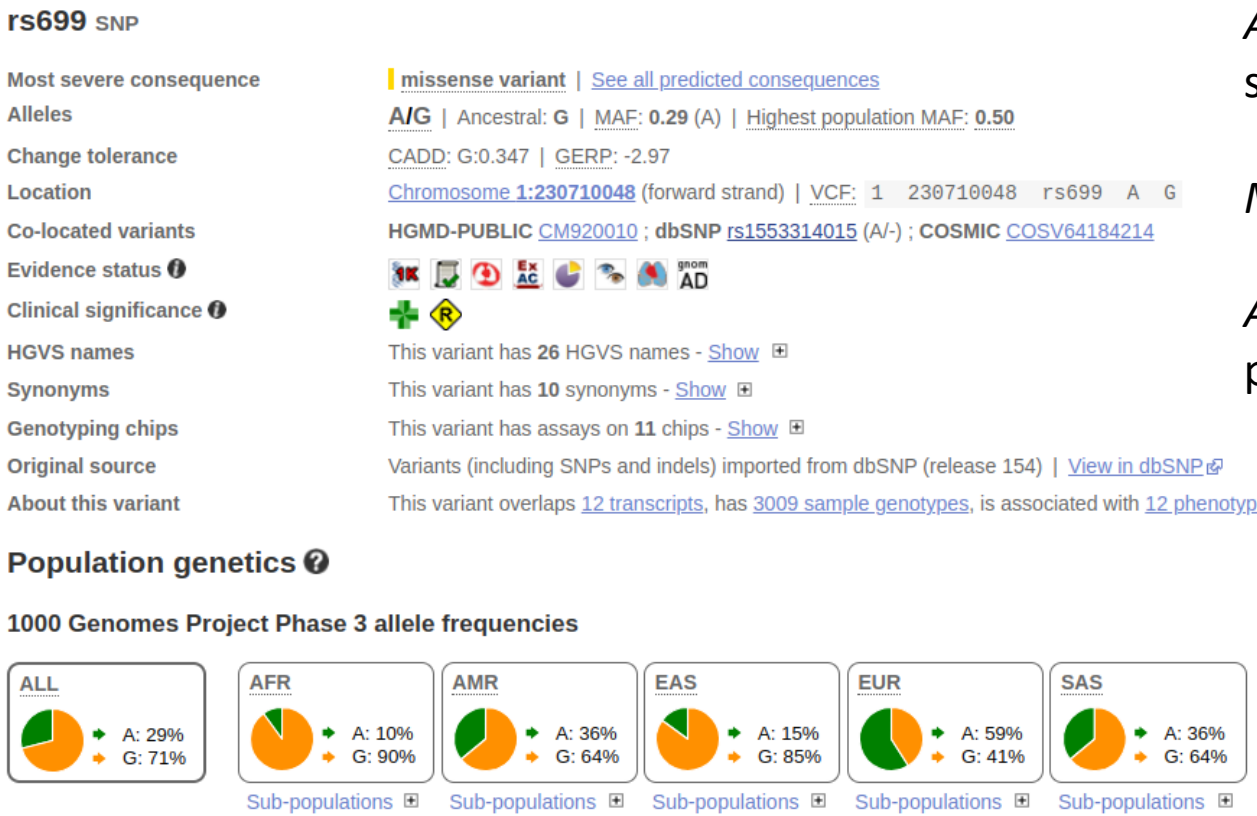
- Región no codificante: modificación reguladora
- Región codificante
  - Sinónima: sin cambio aminoacídico
  - No sinónima
    - Cambio de sentido (Missense): Cambio de aminoácido
      - Conservativa: características similares
      - No conservativa: características diferentes
    - Sin sentido (Nonsense): Codón de parada





Polimorfismos

- A diferencia de una variante de nucleótido único (SNV), que incluye cualquier tipo de variación de un solo nucleótido, un SNP (Polimorfismo de nucleótido único) es una variación que ocurre en la línea germinal del individuo y debe estar presente en un porcentaje suficiente en la población (generalmente >1%).
- Otros SNVs por debajo del 1% en la población suelen denominarse mutaciones.



Alelo: Una de las dos o más versiones de una secuencia genética en una región determinada.

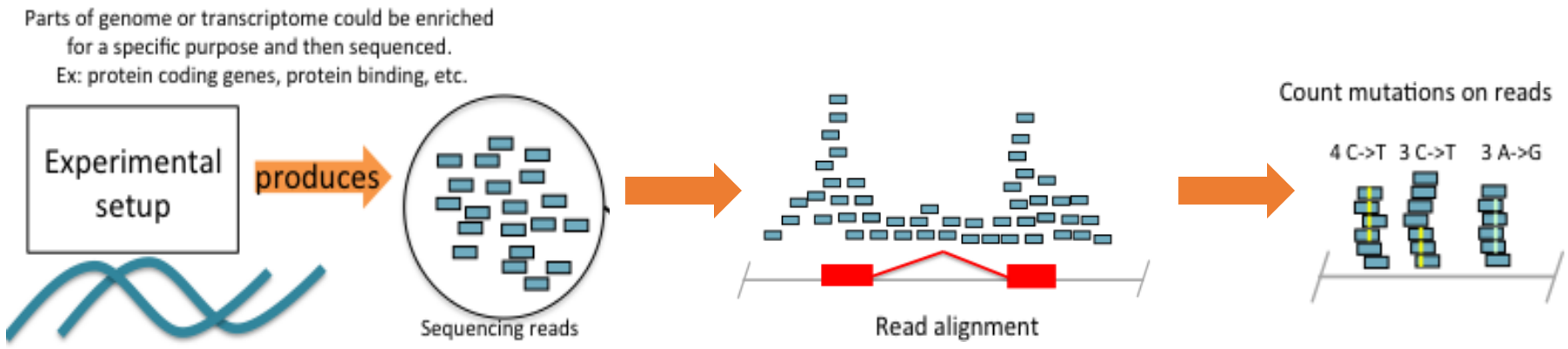
MAF: Frecuencia del alelo menor en la población.

Alelo ancestral: alelo presente en el ancestro común previo a la especiación de la especie en estudio.

https://www.ensembl.org/index.html

Re-secuenciación masiva de ADN

La secuenciación de ADN en **especies con un genoma de referencia** se basa en trocear aleatoriamente el genoma para secuenciar millones de fragmentos solapantes. Estos fragmentos llamados **lecturas** son reubicados en su posición original utilizando el genoma de referencia mediante un proceso denominado **alineamiento**. Una vez identificada la posición de cada lectura se cuantifican los nucleótidos presentes en cada posición del genoma secuenciado y se anota en función de las diferencias con el genoma de referencia.



Ventajas

- Información completa del genoma
- Todo tipo de variantes de secuencia
- Estudio de aberraciones cromosómicas en cáncer

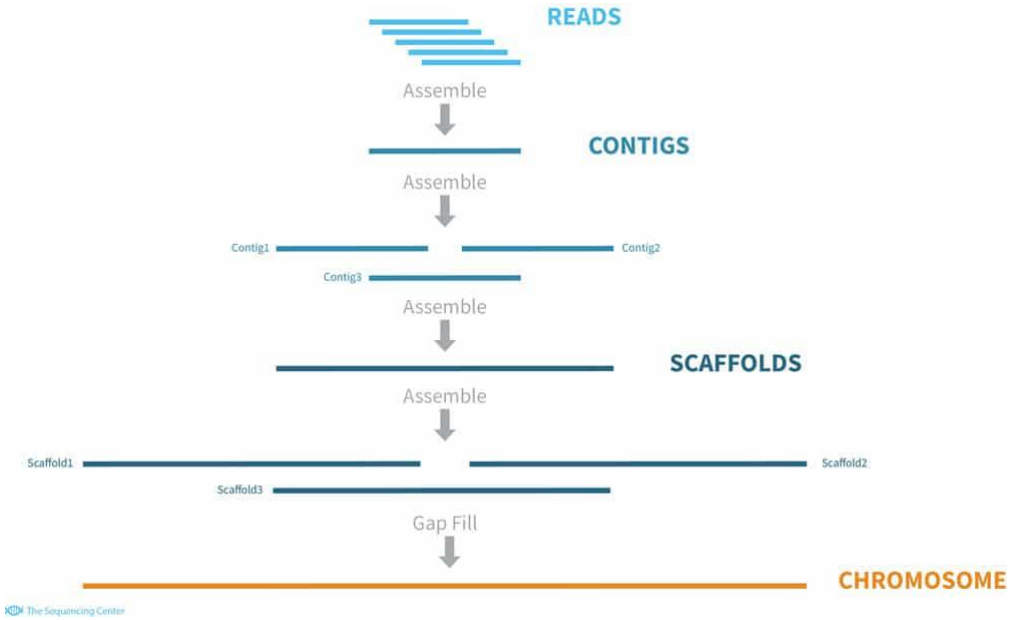
Retos

- Coste
- Genoma de referencia
- Interpretación de los resultados
- Analizar millones de lecturas

Otras aproximaciones a la secuenciación de ADN

Secuenciación de ADN *de novo* (Ensamblado *de novo*)

- Fundamental para el estudio sin genoma de referencia
- No fuerza el ensamblado sobre lo conocido, permite el descubrimiento de nuevas configuraciones genéticas (en humanos útil en muestras de cáncer): transferencia horizontal, inversiones, translocaciones...
- *Elevado coste computacional*
- *No es muy sensible para ver pequeñas diferencias en genomas bien anotados*



CONTIGS: fragmentos de ADN compuestos por al menos dos lecturas solapantes.

SCAFFOLDS: fragmentos de ADN compuestos por al menos dos contigs. Los scaffolds pueden contener huecos sin información, ya que los contigs pueden "unirse" con información de huecos conocidos (por ejemplo, secuenciación paired-end), es decir no necesariamente deben solapar para saber que uno va a continuación de otro.

Definir qué scaffolds forman un cromosoma, no es una tarea sencilla. Y es necesario combinar múltiples tecnologías para lograrlo, incluyendo lecturas largas, mapas ópticos...

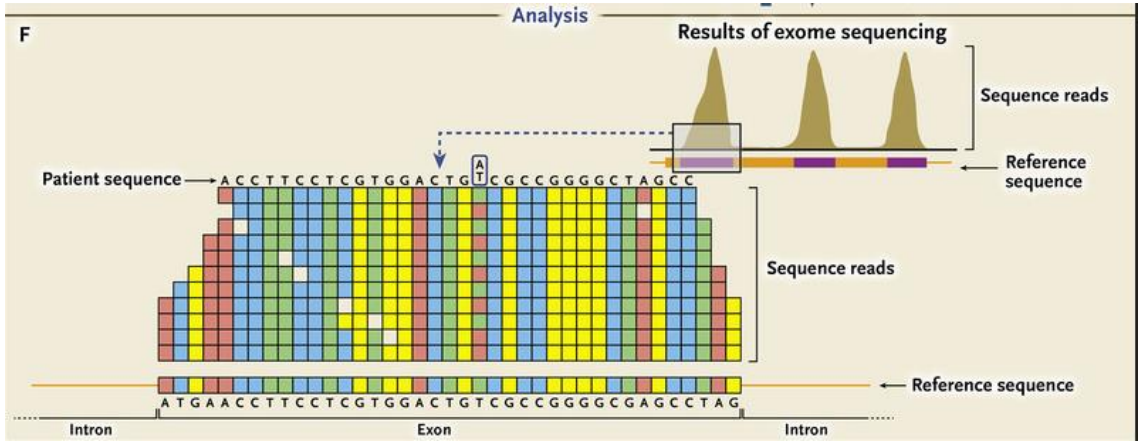
Otras aproximaciones a la secuenciación de ADN

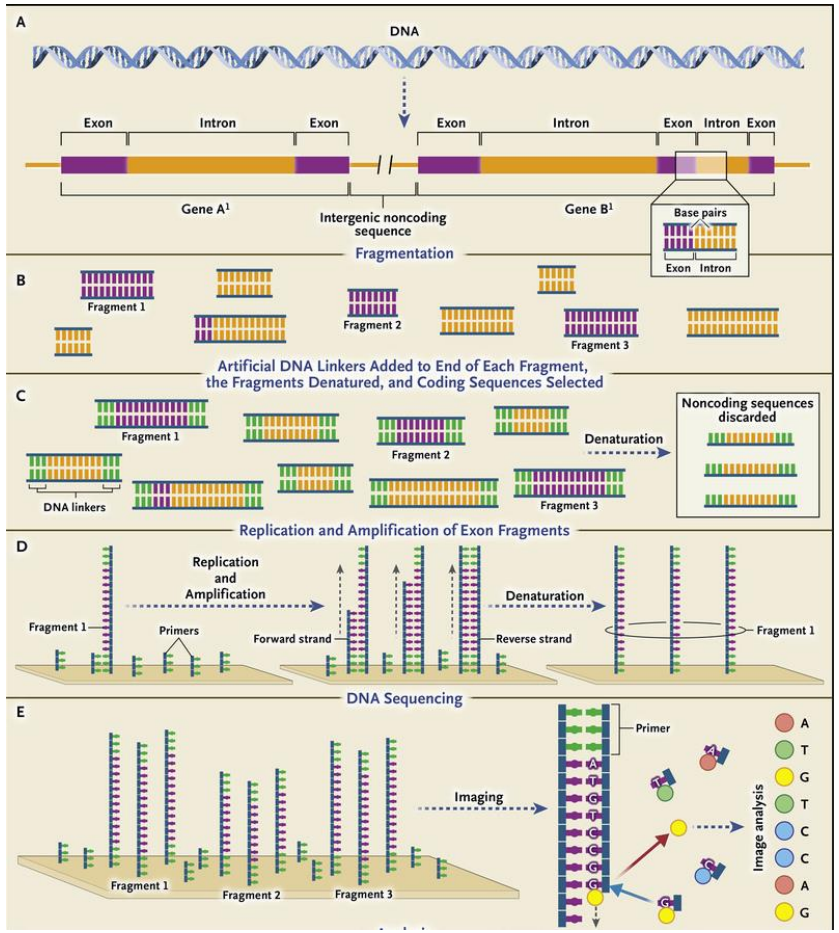
Secuenciación de exomas

- Se basa en capturar y secuenciar la parte codificante de los ~20.000 genes codificantes del genoma (1%)

- Menor coste de secuenciación
- Análisis funcional sencillo
- Enfermedades mendelianas

- No apto para variación en regiones reguladoras, ni intrónicas
- No apto para inversiones, translocaciones, CNVs...





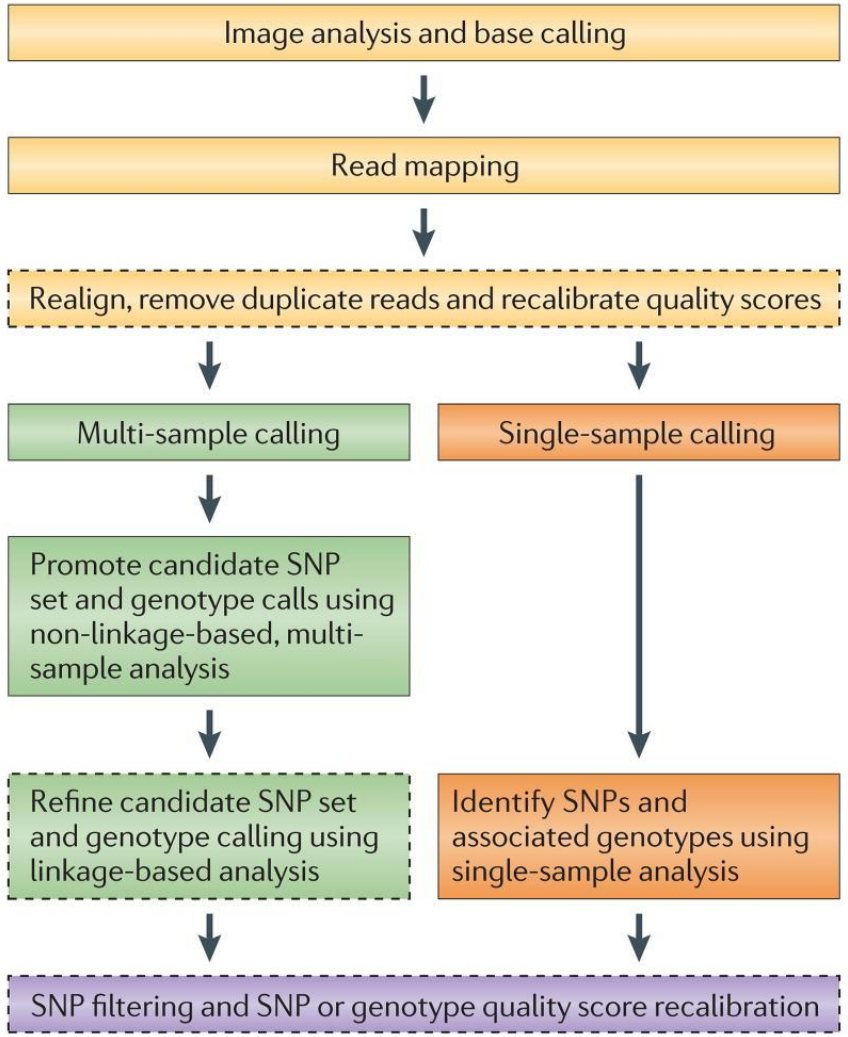
Estimación de genotipo a partir de HTS

- La interpretación de los resultados provenientes de secuenciación masiva de ADN requieren de **múltiples asunciones** y diferentes aproximaciones en función del número de muestras secuenciadas, la cobertura y la calidad de secuenciación.

- El proceso de **alineamiento** y la interpretación de los valores de **calidad de secuenciación** dados por el secuenciador son críticos. Realineamiento y recalibración de los primeros resultados son pasos muy utilizados.

- La **cobertura es fundamental**, genomas re-secuenciados a coberturas >20x devuelven buenos resultados usando algoritmos sencillos basados en umbrales. Sin embargo, alcanzar estas coberturas en estudios de cohortes es excesivamente caro por lo que se desarrollaron alternativas probabilísticas (estadística bayesiana).

- Recientemente, las estimaciones se están corrigiendo mediante información de bloques de ligamiento (**imputación**).

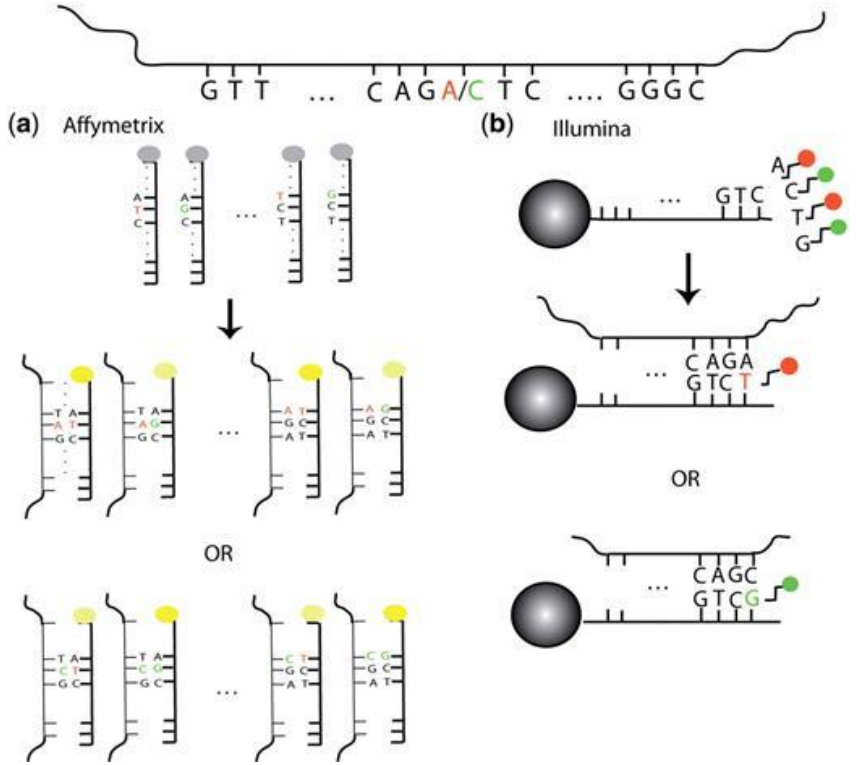


https://doi.org/10.1038%2Fnrg2986

Estudios de asociación de genoma completo (GWAS)

Genotipado mediante microarrays

Actualmente la tecnología que sigue siendo la más utilizada en estudios de asociación de genoma completo se basa en *microarrays*, ya que su **interpretación es relativamente más sencilla** que la secuenciación masiva, y sobre todo es **muchísimo más barata** lo que permite estudiar grandes cohortes. Sin embargo, esta tecnología sólo permite estudiar variantes conocidas y un número limitado de las mismas.



http://doi.org/10.1093/nar/gkp552

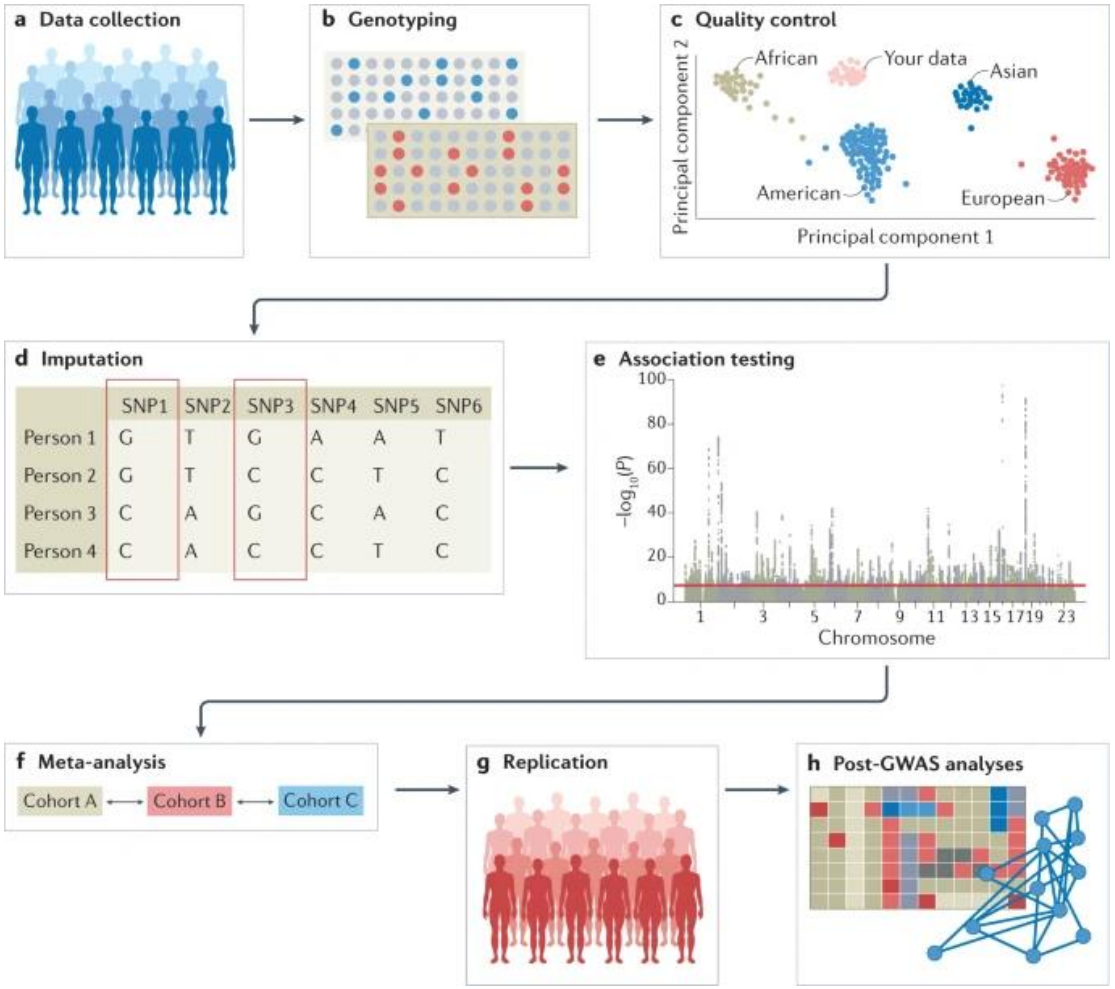
La tecnología de *microarrays* para la detección de variantes de secuencia se basa en la hibridación por complementariedad de hebra y en la detección mediante fluorescencia de la variable.

En algunos casos utilizan varios colores para identificar cada alelo posible del polimorfismo o diferentes intensidades de un mismo color.

En todos los casos la tecnología requiere de conocimiento previo y sólo puede analizar dos alelos por pocillo del array. Actualmente existen *microarrays* que permiten analizar millones de variantes en un solo experimento.

Estudios de asociación de genoma completo (GWAS)

Diseño y análisis de un GWAS



https://doi.org/10.1038/s43586-021-00056-9

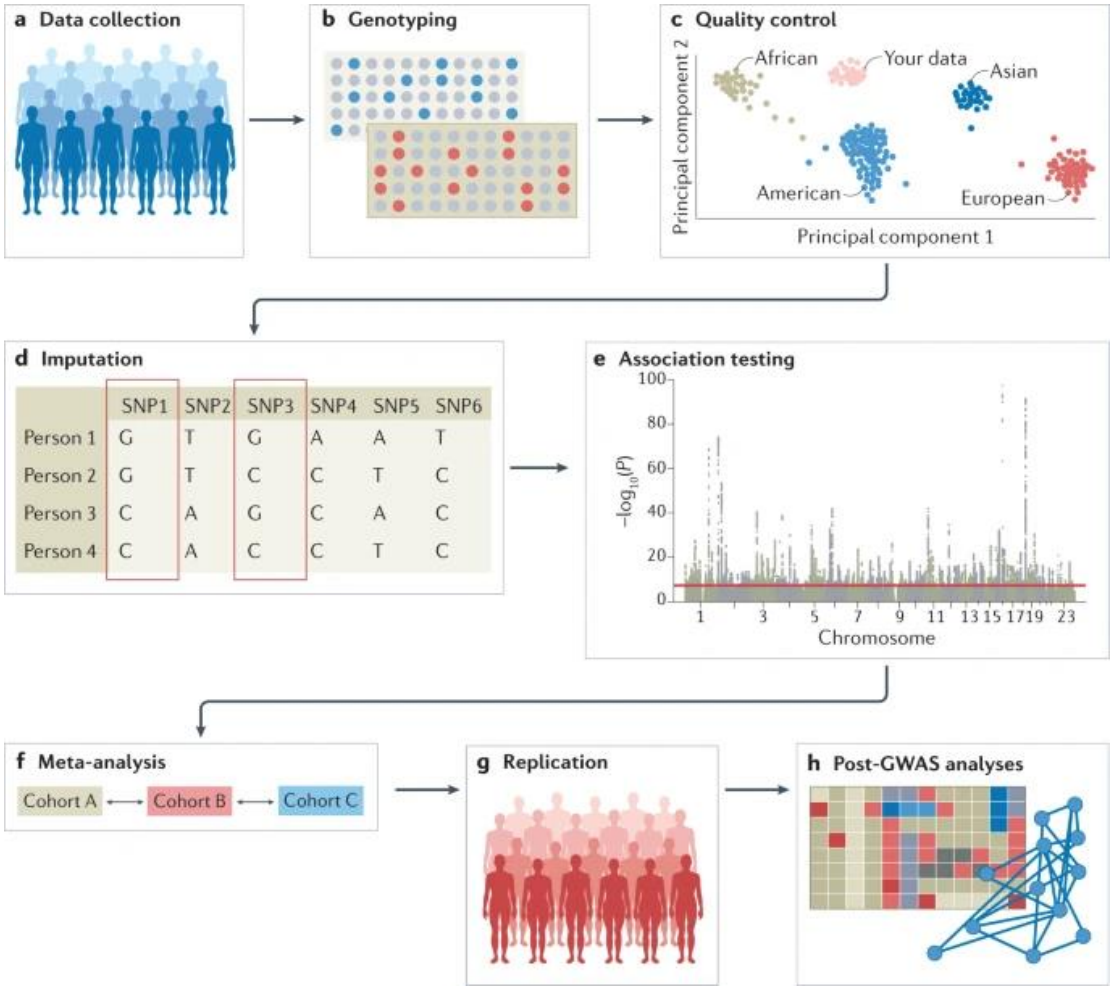
La gran mayoría de los pasos necesarios para controlar y diseñar un GWAS mediante genotipado aplican igual a un estudio de secuenciación.

- El número de individuo en estos estudios necesariamente debe ser **grande** (>1000) para tener el poder estadístico suficiente de detectar variantes asociado con el fenotipo de interés. Se debe tener especial cuidado en no generar **sesgos** en alguna de las **cohortes** reclutadas.

- Los **controles de calidad** incluyen eliminar variantes raras, variantes que no se encuentre en HWE, errores de genotipado e individuos con errores en la toma de datos entre otros. Pero también debe tenerse en cuenta la ancestría de los individuos de las cohortes a comparar, ya que diferencias de ancestría entre las cohortes puede resultar en diferencias debidas a las mismas y no a los fenotipos del estudio.

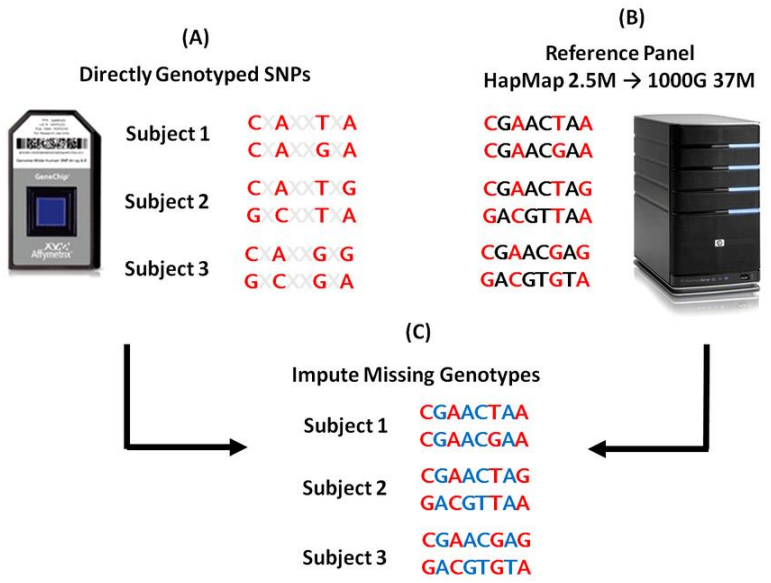
Estudios de asociación de genoma completo (GWAS)

Diseño y análisis de un GWAS



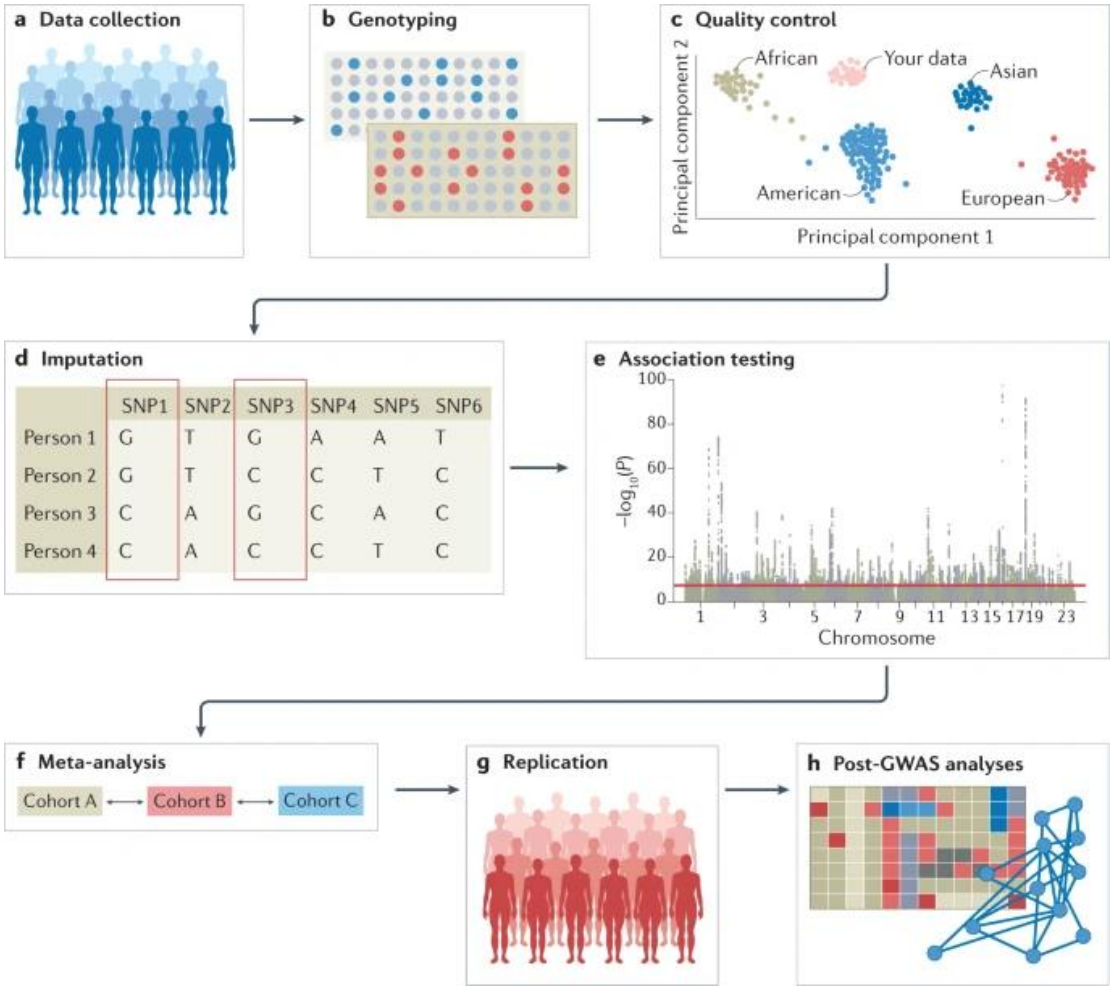
https://doi.org/10.1038/s43586-021-00056-9

- Actualmente, gracias a los miles de individuos genotipados y secuenciados para diferentes poblaciones podemos predecir con elevada precisión las variantes de secuencia en posiciones no genotipadas, mal secuenciadas o en general no cubiertas por el método de elección, a este proceso se le denomina **imputación**. La imputación se basa en el desequilibrio de ligamiento existente entre posiciones próximas en el genoma, posiciones que no segregan de manera independiente a las sucesivas generaciones ya que presentan una tasa de recombinación muy baja.



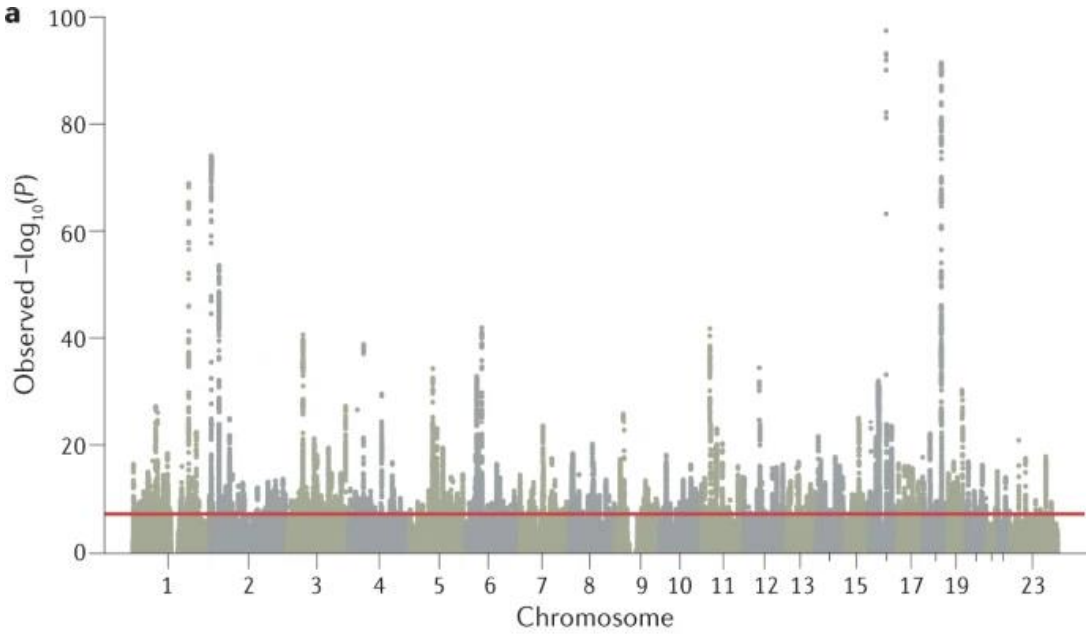
Estudios de asociación de genoma completo (GWAS)

Diseño y análisis de un GWAS



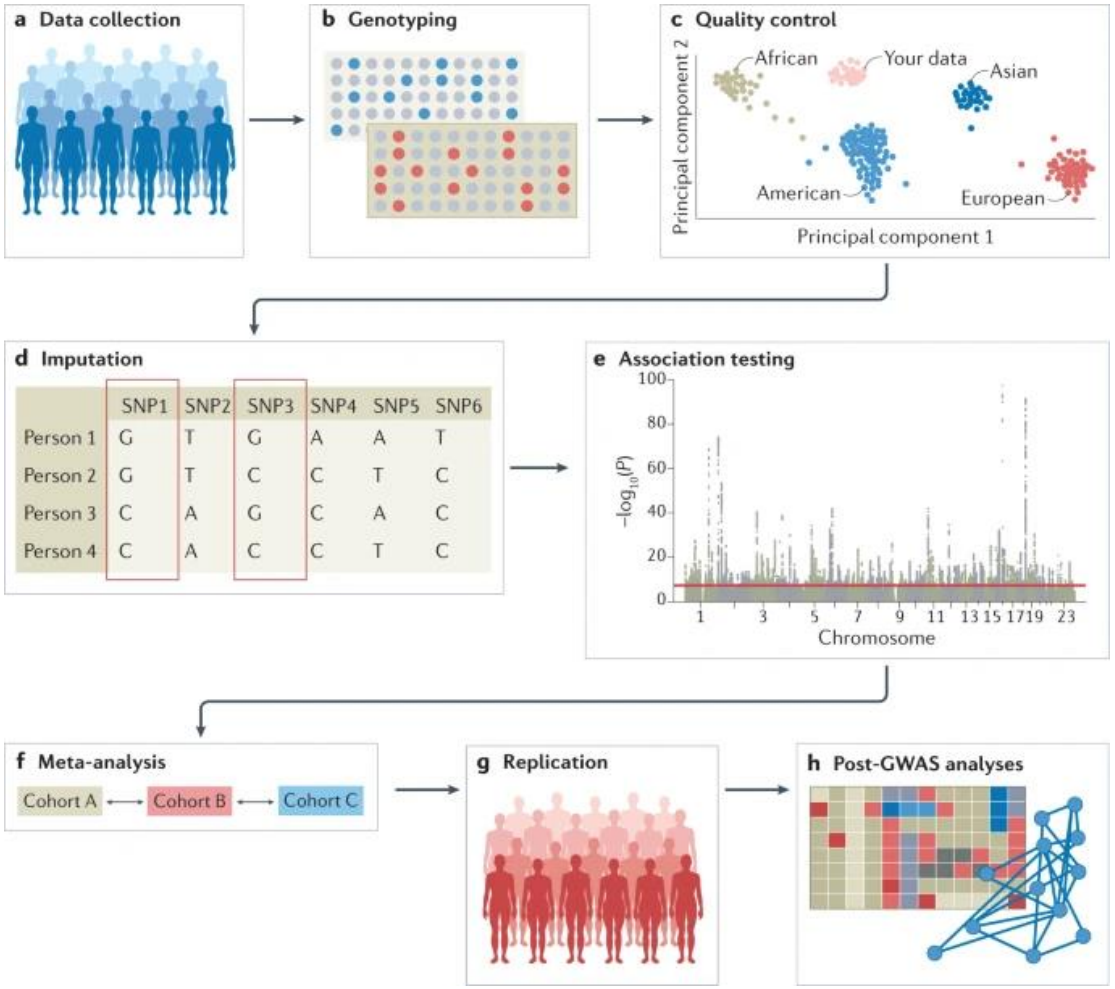
https://doi.org/10.1038/s43586-021-00056-9

- Una vez definidos los genotipos de los individuos de nuestras cohortes se utilizan diferentes **modelos de regresión** dependiendo del fenotipo que se esté analizando, lineales si son fenotipos continuos o logísticos si son binarios. Estos modelos de regresión pueden incluir covariables para limitar el efecto de posibles variables de confusión en el estudio. Los valores de probabilidad obtenidos suelen representarse en lo que se conoce como **gráficos manhattan**.



Estudios de asociación de genoma completo (GWAS)

Diseño y análisis de un GWAS

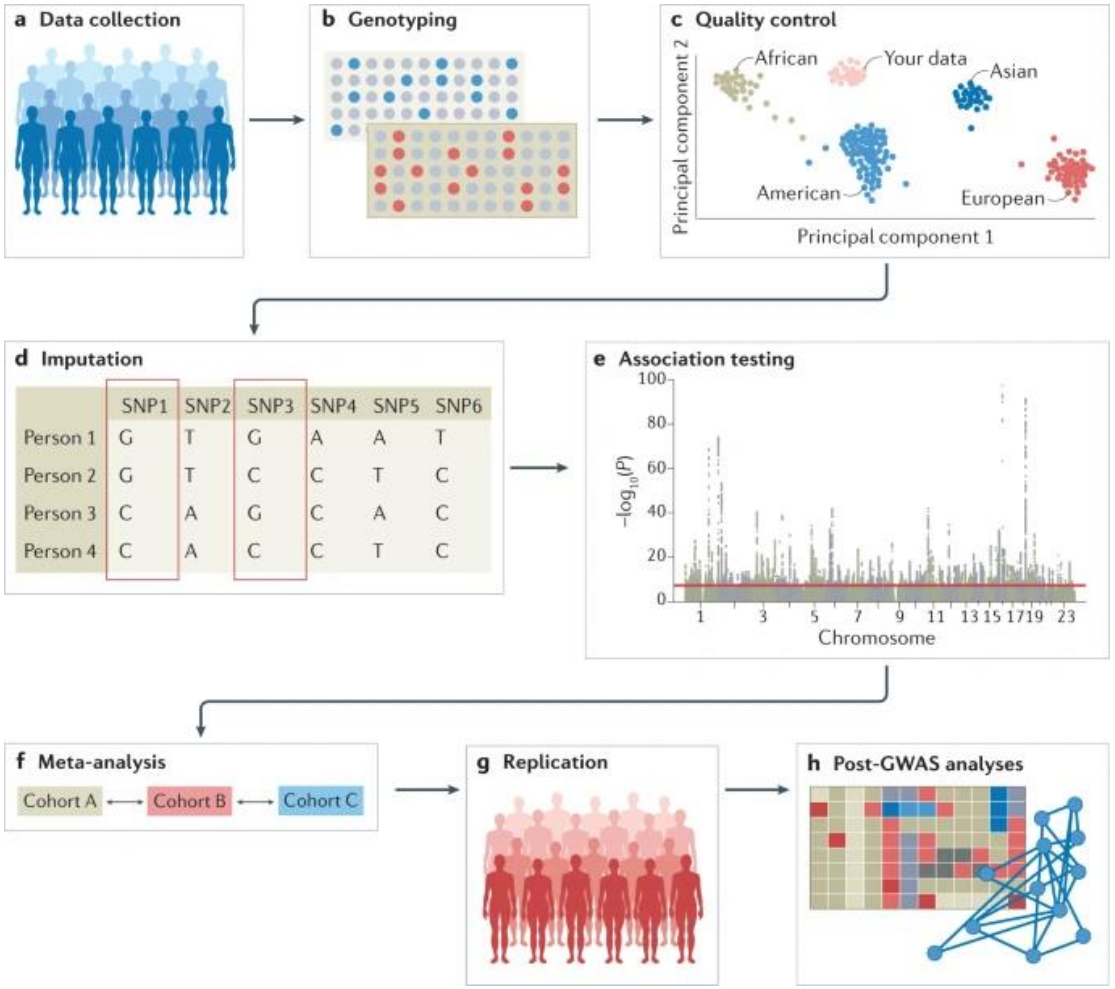


https://doi.org/10.1038/s43586-021-00056-9

- El análisis de datos múltiples conlleva una probabilidad asociada de encontrar asociaciones debido al azar, es decir aumenta el error de tipo I (aumento de falsos positivos). Para evitar esto, generalmente se usan **métodos de corrección múltiple**, el más sencillo de ellos es el método de *bonferroni* que supone dividir los valores de p por el número de análisis efectuados. Sin embargo, en genética los análisis efectuados como sabemos no son necesariamente independientes debido a desequilibro de ligamiento, por lo que el umbral de significación comúnmente aceptado es de $5e\text{-}10^8$ (1 millon de variantes comunes independientes en el genoma $0.05/10^6$).

Estudios de asociación de genoma completo (GWAS)

Diseño y análisis de un GWAS



- Generalmente este tipo de estudios suelen realizarse en grandes consorcios de investigación por que se generan múltiples cohortes de datos que son analizados de manera independiente para posteriormente analizar conjuntamente sus resultados, a este proceso se le denomina **meta-análisis**. Este proceso se realiza directamente con los estadísticos resultantes de los modelos de regresión para cada variante y suele conferir un gran poder estadístico para detectar variantes con frecuencias bajas en la población pero asociadas al fenotipo de estudio.

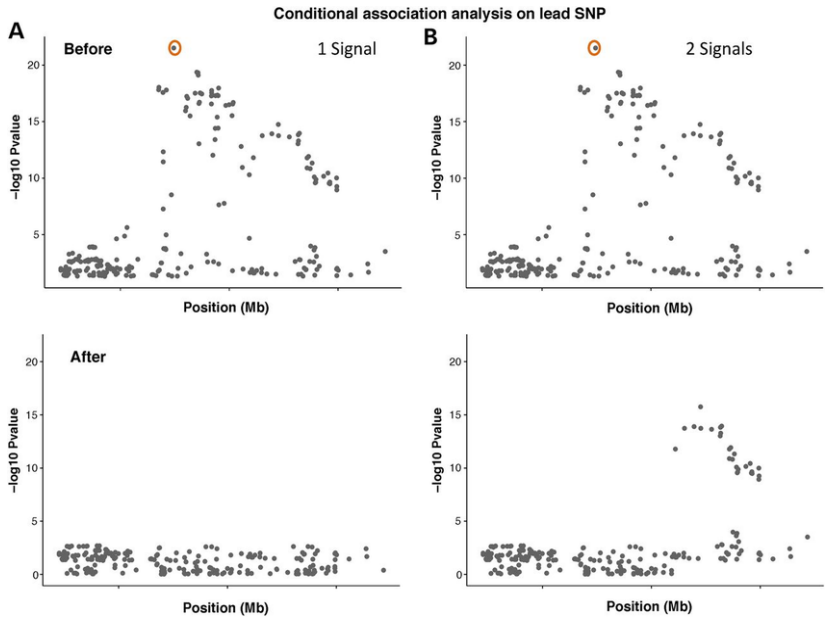
- Debido al gran número de sesgos que se pueden introducir y asunciones que se realizan en estos estudios suele ser necesario **replicar los resultados** en cohortes independientes.

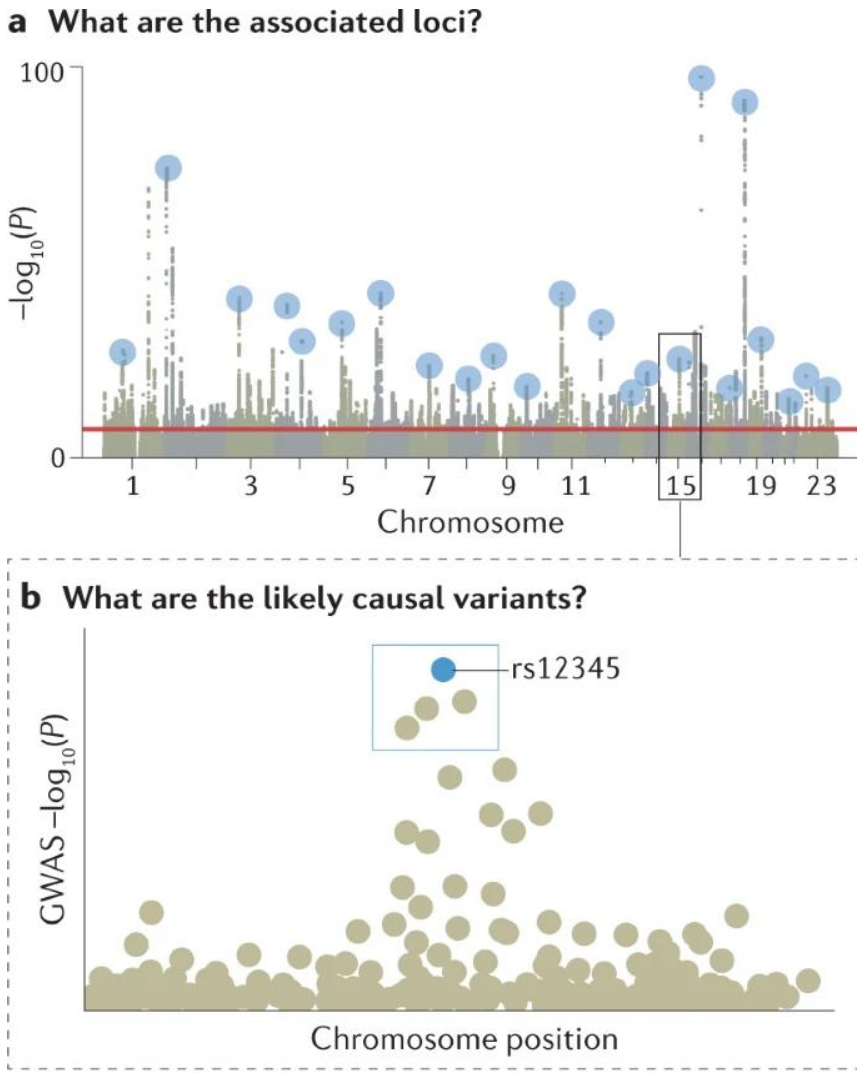
https://doi.org/10.1038/s43586-021-00056-9

Estudios de asociación de genoma completo (GWAS)

Variantes causales

- Múltiples loci agrupados en regiones relativamente pequeñas suelen verse asociados significativamente al fenotipo. Sin embargo, la mayoría serán variantes no-causales arrastradas debido al desequilibrio de ligamiento. La búsqueda de la(s) variante(s) causal(es) puede incluir tanto procesos *in-silico* como análisis de asociación condicionados o métodos bayesianos más elaborados (*fine-mapping*), pero también la necesidad de re-secuenciar la región de interés.





https://doi.org/10.1038/s43586-021-00056-9

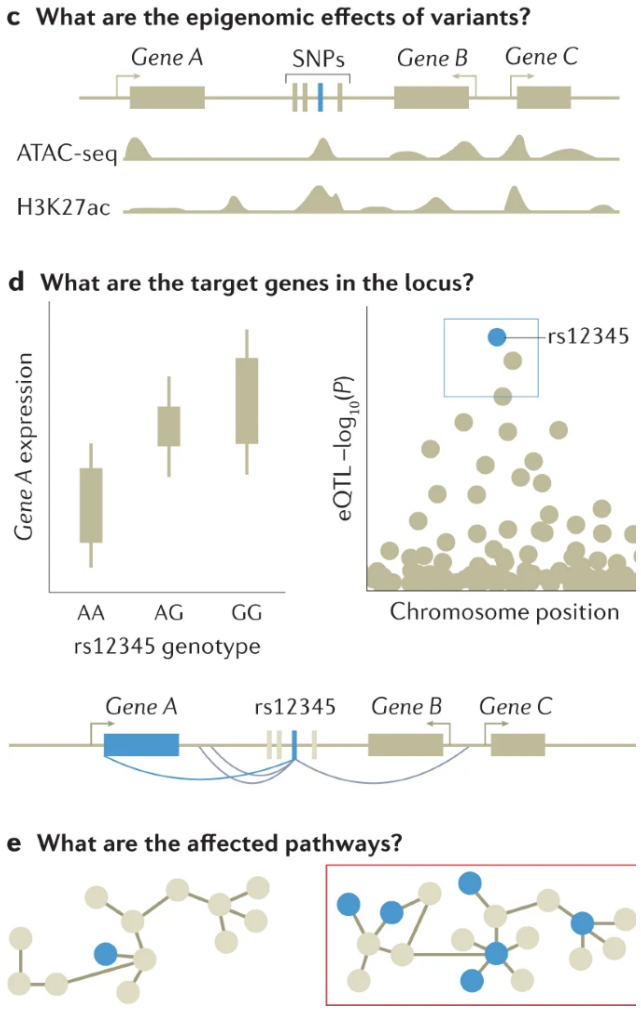
Estudios de asociación de genoma completo (GWAS)

Análisis funcional

- Si la variante causal se encuentra localizada en una región codificante su funcionalidad es relativamente fácil de determinar. Sin embargo, sólo el 2-3% de las variantes identificadas por GWAS se encuentran en regiones codificantes lo que hace su interpretación un poco más complicada.
- Aquellas variantes en regiones no codificantes deben ser interpretadas a información reguladora conocida del genoma, como modificaciones sitios de unión a factores de transcripción, disrupciones en marcas epigenéticas y/o análisis de asociación de los genotipos con valores moleculares (expresión de genes, metilación...)
- Lo ideal sería producir esos datos y realizar los análisis sobre las cohortes de estudio. Pero análisis de este tipo son costosos, por lo que se usa la información almacenada en bases de datos como GTEx.

GTEx Portal

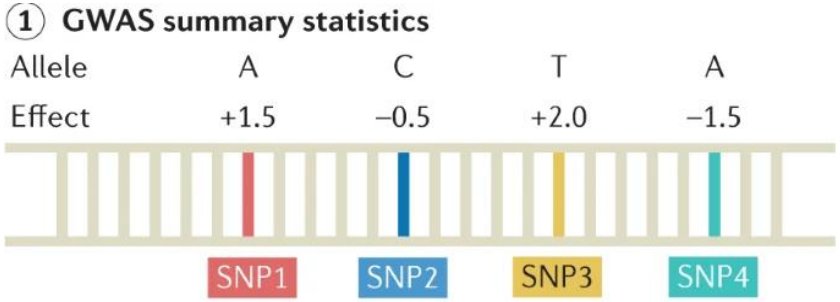
https://gtexportal.org/home/

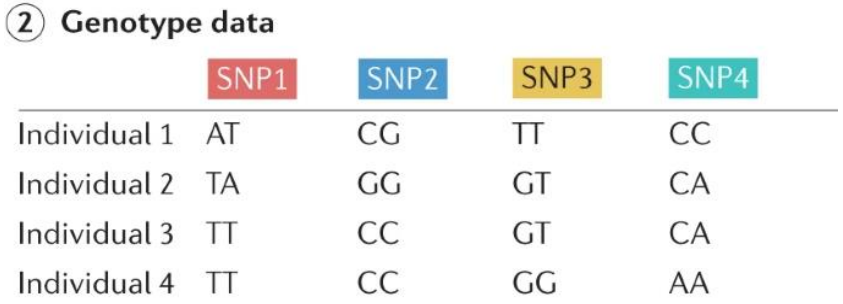


https://doi.org/10.1038/s43586-021-00056-9

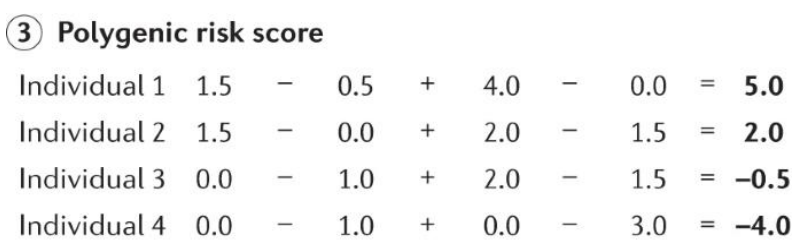
Estudios de asociación de genoma completo (GWAS)

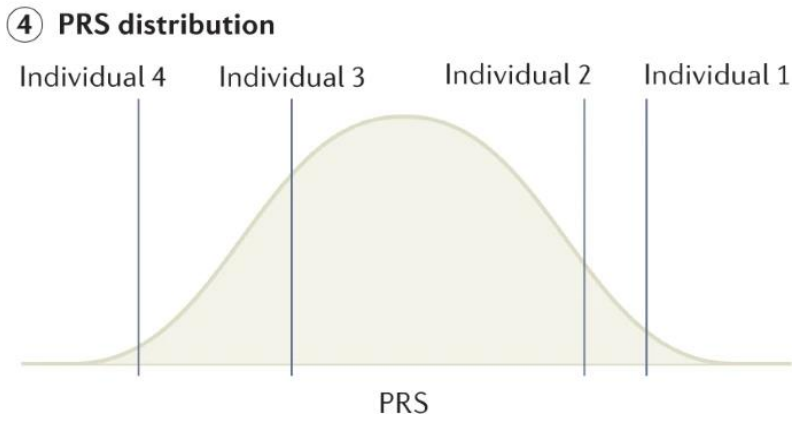
Predicción del riesgo genético en enfermedades complejas





	SNP1	SNP2	SNP3	SNP4
Individual 1	AT	CG	TT	CC
Individual 2	TA	GG	GT	CA
Individual 3	TT	CC	GT	CA
Individual 4	TT	CC	GG	AA





GWAS Catalog

https://www.ebi.ac.uk/gwas/

$$H_0 : \text{Phenotype} \sim \text{covariates} + e$$

$$H_1 : \text{Phenotype} \sim \text{PRS} + \text{covariates} + e$$

https://doi.org/10.1038/s43586-021-00056-9

- Los estadísticos resultantes de estudios de GWAS pueden utilizarse para definir puntuaciones de riesgo genético (PRS) en nuevas cohortes. Existen múltiples aproximaciones, la más sencilla implica seleccionar SNPs asociados al fenotipo y sumar sus efectos en los nuevos individuos.
- Por otro lado, también podemos calcular el porcentaje de la varianza explicada por la genética en el fenotipo de estudio mediante el análisis diferencial de la varianza explicada entre un modelo de regresión que no incluya el PRS (modelo nulo, H_0) y un modelo que incluya el PRS (H_1).