

# Introducción a la Genómica

## UNAL nov 2017

Alejandro Cáceres  
ISGlobal, Barcelona

# Genómica

## Estudio de la biología del genoma

- ▶ Estructura
  - ▶ Variantes estructurales: SNPs, deletions, translocaciones, CNVs, inversiones, etc
  - ▶ Topología: Regulación de la cromatina, Topological association domains (TADs), etc
- ▶ Función
  - ▶ Productos moleculares del genoma
  - ▶ Transcritos, microRNAs
  - ▶ Regulación de la expresión génica
- ▶ Efectos sobre otros niveles biológicos de interés
  - ▶ Rol del genoma en funciones fisiológicas
  - ▶ Rol del genoma en diferencias (heredables) entre individuos
  - ▶ Rol del genoma en la adaptación de individuos a su ambiente
  - ▶ Rol del genoma en la evolución

# Genómica

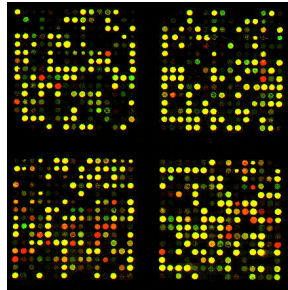
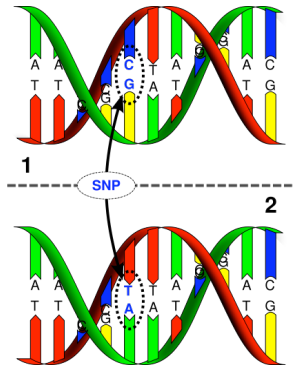
La biología del genoma también incluye interacciones

- ▶ Cómo la estructura determina función
- ▶ Cómo la estructura determina las características heredables
- ▶ Cómo la estructura influye en la adaptación
  
- ▶ Cómo se interaccionan diferentes estructuras para ...
- ▶ Cómo se interaccionan diferentes funciones para ...

# SNPs

- ▶ Los SNPs son una variante estructural a nivel nucleotídico.
- ▶ La tecnología de microarrays permite genotipar miles de individuos en millones de SNPs

# SNP (Single Nucleotide Polymorphism)



# SNPs

- ▶ Son variantes bi-alélicas
- ▶ Individuos son BB (homocigotos), Bb (heterocigotos), bb (homocigotos variantes)
- ▶ B y b toman 4 posibles valores: adenina (a), timina (t), citocina (c), guanina (g)
- ▶ En una población la frecuencia de alélica b es  $> 0.1\%$

# SNPs

- ▶ cubren el genoma (1 SNP/Kb se distribuyen uniformemente)
- ▶ son la variante estructural mas común
- ▶ en total acumulan gran cantidad de variabilidad genética

Su estudio a nivel individual, local y global es esencial para entender procesos genómicos

# SNPs

Como estructura:

- ▶ si afectan función: eQTLs (expresión), mQTLs (metilación)
- ▶ si estan relacionados con enfermedades heredables: SNPs de riesgo
- ▶ o pueden estar bajo presiones evolutivas
- ▶ o pueden no hacer nada ... (mendelian randomization)



# SNPs

Propiedades:

- ▶ Mutaciones Mendelianas
- ▶ Taza de mutación  $10^8$  generaciones
- ▶ Sus frecuencias alélicas están sometidas a deriva genética, selección o migración

# SNPs

- ▶ SNPs cercanos están altamente correlacionados
- ▶ Cada par de SNPs está sometido a recombinación
- ▶ La probabilidad de recombinación entre SNPs incrementa con su distancia
- ▶ Por lo tanto SNPs lejanos no se correacionan

## LD: Cromosomas

Correlación o Linkage Disequilibrium (LD) entre dos SNPs (por ejemplo con alelos: A/T y C/A) se mide con lo que se desvían las casillas de la tabla de contingencia,

	A	T	Total
C	$x_{CA}$	$x_{CT}$	$q_C$
A	$x_{AA}$	$x_{AT}$	$q_A$
Total	$p_A$	$p_T$	1

de la asociación por azar

$$D = p_A * q_C - x_{CA} \text{ o } D = -p_C * q_T + x_{CT}, \dots$$

## LD: genotipos

En los microarrays los SNPs se genotipan individualmente por lo que la fase se pierde!

	A/A=0	A/T=1	T/T=2	Total
C/C=0	$y_{00}$	$y_{01}$	$y_{02}$	$q_0^2$
C/A=1	$y_{10}$	$y_{11}$	$y_{12}$	$2 * q_0 * q_1$
A/A=2	$y_{20}$	$y_{21}$	$y_{22}$	$q_1^2$
Total	$p_0^2$	$2 * p_0 * p_1$	$p_1^2$	1

$$x_{CA} = 2 * y_{00} + y_{01} + y_{10} + n * y_{11}$$

pero  $n = ?$

# Fase

Un sujetos en  $y_{11}$  puede ser

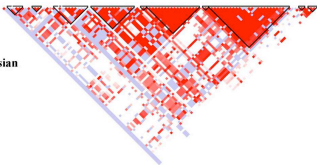
	SNP1		SNP2	
chr1:	C	-	A	(Sí está en $x_{CA}$ )
chr2:	A	-	T	o
<hr/>				
chr1:	C	-	T	(No está en $x_{CA}$ )
chr2:	A	-	A	

Si se conoce la fase se conocen los haplotipos, regiones donde el LD es alto

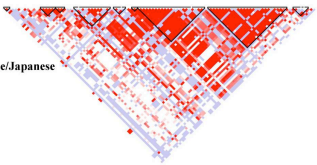
## Regiones con LD alto

Las regiones con LD alto dependen de la ancestría.  $r^2$  es una medida normalizada de  $D$ .  $r^2 = D^2 / (p_A * p_T * q_C * q_A)$

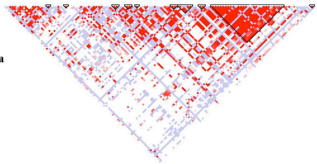
Caucasian



Chinese/Japanese

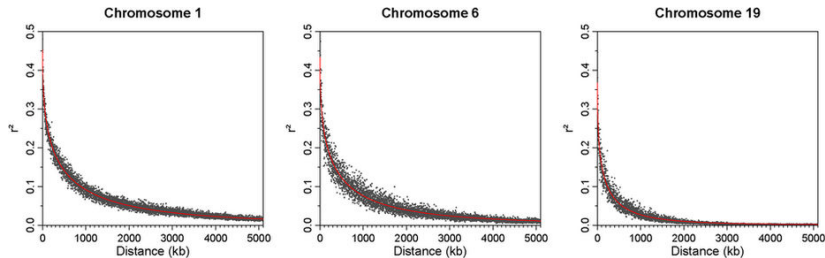


Yoruba



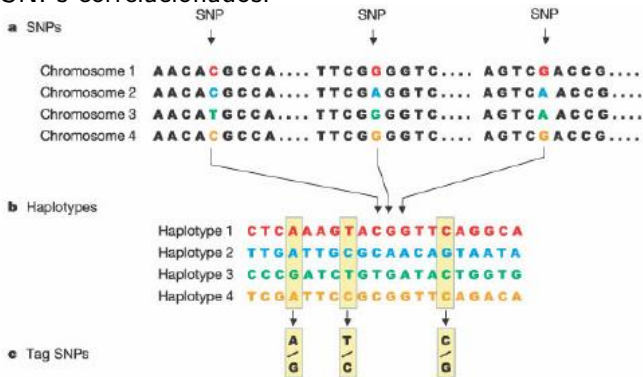
# LD con distancia

La recombinación disminuye el LD y aumenta con la distancia entre SNPs. Hay regiones con muchos SNPs que tienen LD alto entre ellos



# Haplotipos

Esas regiones forman haplotipos frecuentes: macroestructuras de SNPs correlacionados.





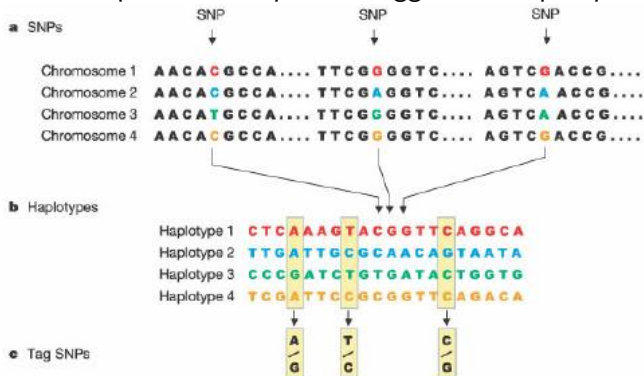
# Haplotipos

Los haplotipos son la principal estructura que generan los SNPs  
Son indicativos de

- ▶ historia evolutiva: haplotipos en EU son mas grandes que en AF
- ▶ patrones específicos de recombinación: Diferencias en hotspots
- ▶ presencia de otros variantes estructurales que suprimen la recombinación.

# Haplotipos

“Si se conocen los haplotipos mas frecuentes de una población entonces pocos SNPs pueden taggear los haplotipos”



# Haplotipos

Si tenemos los haplotipos de referencia de una población

- ▶ Podemos predecir los SNPs no genotipados (imputar)
- ▶ Podemos encontrar una asociación funcional en regiones alrededor de un SNP causal
- ▶ No hay que genotiparlos todos los SNPs

sin embargo

- ▶ No tendremos información sobre variantes mas raras o específicas
- ▶ Cada población tiene una historia evolutiva diferente

# Técnicas de genotipación

## ► Secuenciación

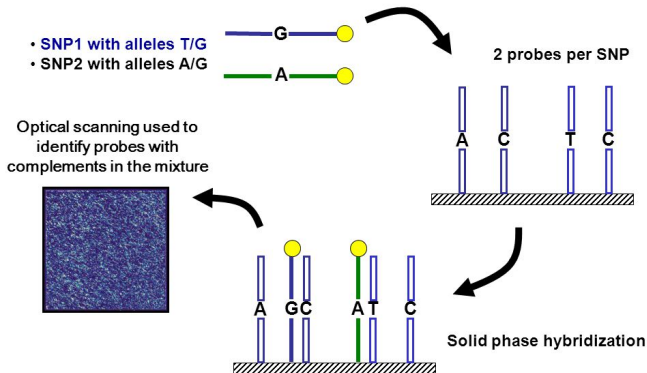
- los SNPs son variantes estructurales por lo que la secuenciación es método ideal
- es cara y estudios de cohortes son hoy en día impensables

## ► Hibridación

- se pueden sintetizar millones de sondas fluorescentes y ponerlas en un chip (microarrays)
- es barato y se puede hacer en miles de individuos
- hay que conocer que sondas hay que poner
- las sondas dependen de una población de referencia
- no es un método para descubrir variantes desconocidas

# microarrays

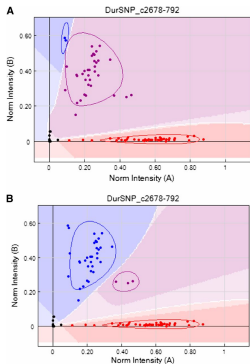
## SNP genotyping via direct hybridization



cada punto es una sonda y el color e intensidad de la luz emitida determina si una muestra hibridizó

## microarrays

cada sujeto tiene una intensidad for cada alelo(A y B)

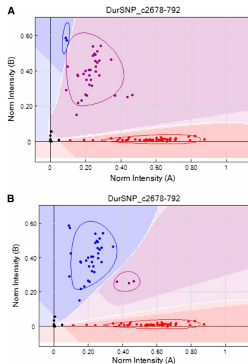


entre menos varianza en angulo entre los grupos mejor es la genotipacion del SNP

- ▶ B-allelefreq: la intensidad relativa del alelo B respecto al A (angulo)
- ▶ log2ratio: la intensidad de la observación respecto al grupo (magnitud)

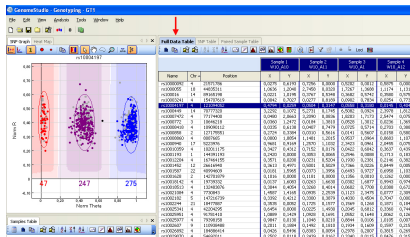
# microarrays

## SNP arrays como medida indirecta de otros variantes estructurales



- ▶ B-allelefreq: si un individuo esta entre dos clusters puede que tenga dos genotipos en sus celulas (Mosaicismos)
- ▶ log2ratio: mas intensidad es indicativo de mas copias de un alelo (CNVs)

Es el software de Illumina para hacer el genotipado de los SNPs (clustering)



- ▶ produce los genotipos en formato PLINK
- ▶ datos de B-allele freq y log2ratio son útiles para el genotipado de mosaicismos y CNVs (no lo hace GenomeStudio)



# Genome-wide SNPs

SNPs genome-wide: barrido genómico con alguna resolución  
objetivo: determinar a que cromosoma de un individuo corresponde cada alelo

- ▶ secuencia: ensamblar los reads
- ▶ microarrays: fasearlos

# Trios

Los datos de trios ayudan a resolver la fase

	SNP1		SNP2		SNP3		SNP4
hijo chr padre:	C	-	A/T	-	G	-	G/A
hijo chr madre:	C	-	A/T	-	G	-	G/A
padre chr1:	C/A	-	T	-	G/C	-	G/A
padre chr2:	C/A	-	T	-	G/C	-	G/A
madre chr1:	C	-	A/T	-	G	-	G/A
madre chr2:	C	-	A/T	-	G	-	G/A
común chr:	C	-	T	-	G	-	G
menos común chr	C	-	T	-	G	-	A

# Haplotipos

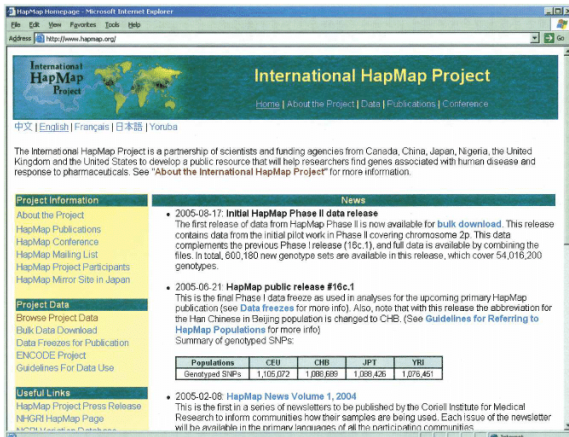
Recordemos: Si tenemos los haplotipos de referencia de una población

- ▶ Podemos predecir los SNPs no genotipados (imputar)
- ▶ Podemos encontrar una asociación funcional en regiones alrededor de un SNP causal
- ▶ No hay que genotipar todos los SNPs

es una estimación de la variabilidad genética de la población

# HAPMAP

## Un proyecto de libre acceso a los datos



International HapMap Project

Home | About the Project | Data | Publications | Conference

中文 | English | Français | 日本語 | Yoruba

The International HapMap Project is a partnership of scientists and funding agencies from Canada, China, Japan, Nigeria, the United Kingdom and the United States to develop a public resource that will help researchers find genes associated with human disease and response to pharmaceuticals. See "About the International HapMap Project" for more information.

**Project Information**

- About the Project
- HapMap Publications
- HapMap Conference
- HapMap Mailing List
- HapMap Project Participants
- HapMap Mirror Site in Japan

**Project Data**

- Browse Project Data
- Bulk Data Download
- Data Freezes for Publication
- ENCODE Project
- Guidelines For Data Use

**Useful Links**

- HapMap Project Press Release
- NIH/RI HapMap Page
- NCBI Variation Database

**News**

- 2005-08-17: Initial HapMap Phase II data release**  
The first release of data from HapMap Phase II is now available for [bulk download](#). This release contains data from the initial pilot work in Phase II covering chromosome 2p. This data complements the previous Phase I release (16c.1), and full data is available by combining the files. In total, 600,180 new genotype sets are available in this release, which cover 54,016,200 genotypes.
- 2005-06-21 HapMap public release #16c.1**  
This is the final Phase I data freeze as used in analyses for the upcoming primary HapMap publication (see [Data freezes](#) for more info). Also, note that with this release the abbreviation for the Han Chinese in Beijing population is changed to CHB. (See [Guidelines for Referring to HapMap Populations](#) for more info)  
Summary of genotyped SNPs:

Populations	CEU	CHB	JPT	YRI
Genotyped SNPs	1,105,072	1,086,699	1,086,426	1,076,451
- 2005-02-08: HapMap News Volume 1, 2004**  
This is the first in a series of newsletters to be published by the Coriell Institute for Medical Research to inform communities how their samples are being used. Each issue of the newsletter will be available in the primary languages of all the participating communities.

publicó los últimos datos (fase III) en el 2009. Los datos están accesibles en formato PLINK.

# HAPMAP

## HapMap II

- ▶ 30 tríos (padres e hijo) de Nigeria.
- ▶ 30 tríos de Estados Unidos de origen europeo.
- ▶ 44 individuos sin relación genética de Japón (Tokio).
- ▶ 45 individuos sin parentesco de China (Peking).

## HapMap III

- ▶ se extendio a 11 poblaciones diferentes incluyendo mas africanos, indios y mexicanos.
- ▶ Abarca unos 2 Millones de SNPs

# 1000 genomes

The screenshot shows a Windows Internet Explorer browser window displaying the 1000 Genomes Project website. The address bar shows the URL <http://www.1000genomes.org/page.php>. The website has a dark blue header with the text "1000 Genomes" in large yellow font and "A Deep Catalog of Human Genetic Variation" in white. To the right of the text is a graphic of human chromosomes. Below the header is a navigation menu with links: Home, About, Partners, Data, Contact, and Wiki. The main content area is divided into two columns. The left column features a section titled "1000 GENOMES PROJECT DATA RELEASE" with a sub-header "SNP data downloads and genome browser representing four high coverage individuals". The text below states that the first set of SNP calls representing the preliminary analysis of four genome sequences are now available for download through the [EBI FTP site](#) and the [NCBI FTP site](#). It also mentions that the data can be viewed directly through the 1000 Genomes browser at <http://browser.1000genomes.org>. The right column contains a "LOG IN" section with fields for Username and Password, and a "Log in" button with a link to "[forgot my password](#)". Below this is a "LINKS" section with two buttons: "Download the meeting report" and "View the participants". The Windows taskbar at the bottom shows the Start button and several open applications: Missing Heritability NG..., EpiSlides [Compatible...], 1000 Genomes - Home..., and Boulder 2009.

1000 Genomes - Home - Windows Internet Explorer

<http://www.1000genomes.org/page.php>

File Edit View Favorites Tools Help

Google [thousand genomes project](#) Go Bookmarks 2 blocked AutoLink one

1000 Genomes - Home

## 1000 Genomes

A Deep Catalog of Human Genetic Variation

Home About Partners Data Contact Wiki

### 1000 GENOMES PROJECT DATA RELEASE

#### SNP data downloads and genome browser representing four high coverage individuals

The first set of SNP calls representing the preliminary analysis of four genome sequences are now available to download through the [EBI FTP site](#) and the [NCBI FTP site](#). The README file dealing with the FTP structure will help you find the data you are looking for.

The data can also be viewed directly through the 1000 Genomes browser at <http://browser.1000genomes.org>. Launch the browser and view a sample region [here](#).

More information about the data release can be found in the [data section](#) of this web site.

[Download the 1000 Genomes Browser Quick Start Guide](#)

[Click about data](#)

#### LOG IN

Username:

Password:

([forgot my password](#))

#### LINKS

[Download the meeting report](#)

[View the participants](#)

Done

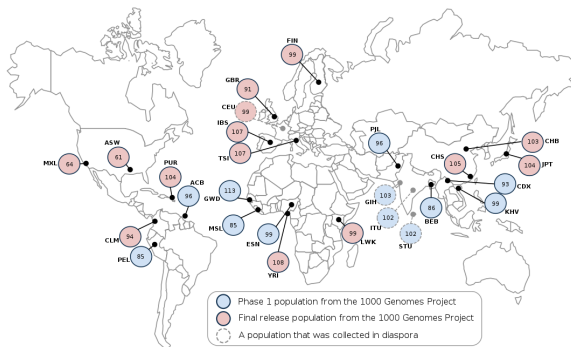
start Missing Heritability NG... EpiSlides [Compatible...] 1000 Genomes - Home... Boulder 2009

# 1000 Genomas

Rápidamente HapMap fue remplazado por los 1000 Genomas

- ▶ secuenciación de 1092 individuos
- ▶ 26 poblaciones
- ▶ 36 Millones de SNPs
- ▶ los datos están accesibles en formato VCF
- ▶ están integrados con el Genome Browser

# 1000 Genomes



Datos publicados el 2012



# The haplotye reference consortium

- ▶ 100,000 Genomes lanzado el 2014 (UK)
- ▶ Simons genome diversity proyect 300 individuos de 142 poblaciones lanzado 2016
- ▶ The haplotye reference consortium 64,976 individuos 39 millones de SNPs

# The haplotype reference consortium

Participating cohorts - The Haplotype Reference Consortium - Mozilla Firefox

www.haplotype-reference-consortium.org/participating-cohorts

## The Haplotype Reference Consortium

OVERVIEW PARTICIPATING COHORTS USING THE RESOURCE CONTACT SITE LIST

### Participating cohorts

A growing list of cohorts/groups that are contributing to the consortium is as follows

HRC COHORTS					
Cohort	# samples in Release 1	Total # samples	Depth	Website	Principal Investigators
1 UK10K	3715	3781	6.5x	<a href="http://www.uk10k.org/">http://www.uk10k.org/</a>	Richard Durbin, Nicole Segre
2 Sardinia	3445	3514	4x	<a href="http://sardinia.rn.rna.nih.gov/">http://sardinia.rn.rna.nih.gov/</a>	Francesco Cuccia, Serena
3 IBD	4478	4478	4x + 2x	<a href="http://www.broadresearch.co.uk/">http://www.broadresearch.co.uk/</a>	UK IBD Genetics Consor
4 GoT2D	2710	2974	4x/Exome	<a href="http://www.type2diabetesgenetics.org/information/got2d/">http://www.type2diabetesgenetics.org/information/got2d/</a>	Mike Boehnke, David Alt
5 BRIDGES	2487	4000	6-8x (12x)	<a href="http://www.1000genomes.org/">http://www.1000genomes.org/</a>	Mike Boehnke, Richard N
6 1000 Genomes	2495	2535	4x/Exome	<a href="http://www.1000genomes.org/">http://www.1000genomes.org/</a>	Richard Durbin, Goncalo
7 GoNL	748	748	12x	<a href="http://www.rnigenome.nl/">http://www.rnigenome.nl/</a>	Paul de Bakker
8 AMD	3305	3305	4x		Goncalo Abecasis, Aram
9 HUNT	1023	1254	4x		Cristen Wilke, Kristian H
10 ISG+ Kusumoto	1918	1918	4x		Richard Durbin, Aarno Pa
11 INGI-FVG	250	250	4-10x	<a href="http://www.rnigenome.it/ingifvg.asp">http://www.rnigenome.it/ingifvg.asp</a>	Pietro Gasparini, Nicole S
12 INGI-Val Borbera	225	225	6x	<a href="http://www.rnigenome.it/ingival.asp">http://www.rnigenome.it/ingival.asp</a>	Daniela Toniolo, Nicole S
13 MCTFR	1325	1339	10x	<a href="http://lmchb.psych.umn.edu/">http://lmchb.psych.umn.edu/</a>	Goncalo Abecasis, Scott
14 HELIC	247	2000	4x (1x)	<a href="http://www.helic.org/">http://www.helic.org/</a>	Eleonora Zeggini
15 ORCADES	398	399	4x	<a href="http://www.orcades.ed.ac.uk/orcades/">http://www.orcades.ed.ac.uk/orcades/</a>	Jim Wilson, Richard Dur
16 INCHANTI	676	680	7x	<a href="http://www.inchantedstudy.org/index.html">http://www.inchantedstudy.org/index.html</a>	Tim Frayling, Andrew Wo
17 GECCO	1131	3000	4-6x	<a href="http://www.hcc.org.uk/gecco/projects/cancer-prevention/projects/gecco.html">http://www.hcc.org.uk/gecco/projects/cancer-prevention/projects/gecco.html</a>	Ulrike Peters
18 GPC	697	768	30x		Carlos Pato, Michele Pat
19 Project MINE - NL	935	1250	45x	<a href="http://projectmine.com">http://projectmine.com</a>	Jan Veldink, Leonard van
20 NEPTUNE	403	403	4x	<a href="http://www.neptune-study.org/">http://www.neptune-study.org/</a>	Matthias Kretzsch, Matthe
Totals	32911	38821			

[https://www.google.com/url?https://sardinia.rn.rna.nih.gov/ksa+D&uc=1306591362137000&uq=AFQCNFR\\_uMFWW1dQOM-TH0Y85W](https://www.google.com/url?https://sardinia.rn.rna.nih.gov/ksa+D&uc=1306591362137000&uq=AFQCNFR_uMFWW1dQOM-TH0Y85W)

Datos publicados el 2012

# Haplotipos de referencia

Una vez identificados los haplotipos de referencia

- ▶ variabilidad genética de la población
- ▶ frecuencia de alelos funcionales
- ▶ frecuencia de alelos de riesgo

# Variabilidad genética

La variabilidad genética dada por los haplotipos de referencia contribuye a determinar:

- ▶ la especificidad y frecuencia de haplotipos
- ▶ medidas globales de heterocigosidad
- ▶ medidas globales de fijación ( $F_{ST}$ )
- ▶ patrones de recombinación
- ▶ la presencia de otras variantes genéticas (CNVs, mosaicismos o inversiones)

# Variabilidad genética

estudios comparativos permite estimar

- ▶ la distancia genética a otras poblaciones
- ▶ evidencia de selección en regiones particulares
- ▶ evidencia de patrones de migración y mestizaje

# Variabilidad genética

estudios comparativos permite estudiar

- ▶ la distancia genética a otras poblaciones
- ▶ evidencia de selección en regiones particulares
- ▶ evidencia de patrones de migración y mestizaje

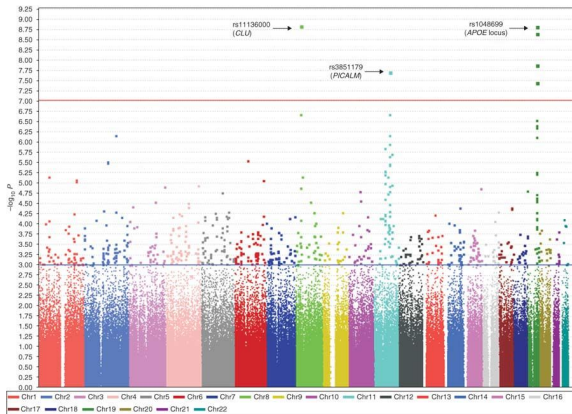
# Estudios de asociación genética

Tener un panel de referencia permite

- ▶ estimar las frecuencias que se esperan en la población general
- ▶ tener un grupo “control”
- ▶ estudiar el contexto genético de los efectos de un SNP de riesgo
- ▶ integrar estudios de asociación genética (imputación)

# Estudios de asociación genética

Los GWAS pretenden encontrar los SNPs de riesgo a una enfermedad (característica heredable)

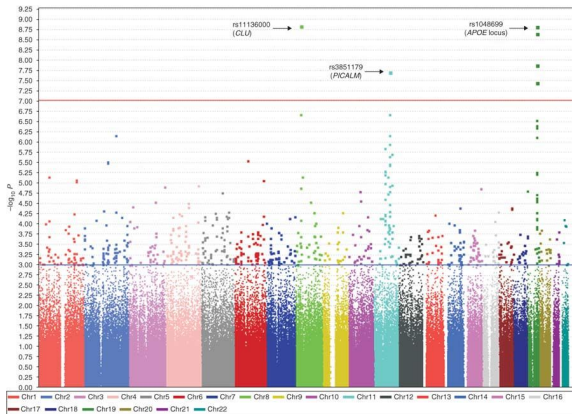


Se basan en la hipótesis de que enfermedades comunes pueden ser explicadas por variantes comunes



# Estudios de asociación genética

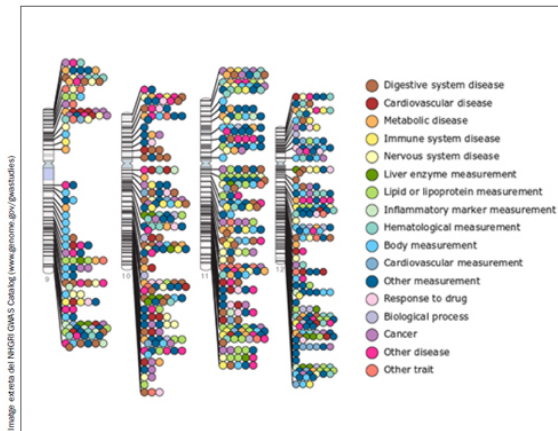
Los GWAS pretenden encontrar los SNPs de riesgo a una enfermedad (característica heredable)



Se basan en la hipótesis de que enfermedades comunes pueden ser explicadas por variantes comunes

# Estudios de asociación genética

Hay miles de estudios reportados

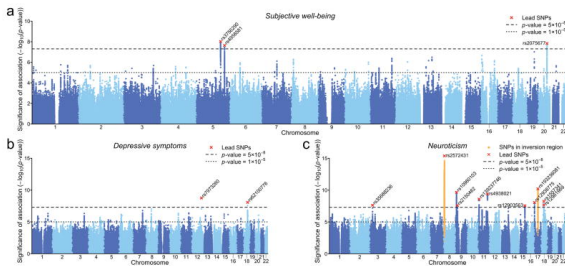


Cromosomes amb variants de risc per diverses patologies identificades per GWAS

Chromosomes with risk variants for several pathologies identified by GWAS

# Estudios de asociación genética

Se encuentran muchos efectos pequeños aditivos por lo que es necesario grandes estudios



- ▶ 298,420 sujetos para síntomas de bienestar
- ▶ 161,460 sujetos para síntomas de depresión
- ▶ 170,910 sujetos para síntomas de neuroticismo

# Estudios de asociación genética

Cada cohorte tiene unos datos de SNPs particulares, diferente densidad de SNPs.

- ▶ la imputación de los datos es necesaria para integrar las cohortes
- ▶ la imputación depende de los haplotipos de referencia

# Haplotipos de referencia Colombianos

## Variabilidad genética

- ▶ estructura de los haplotipos por regiones geográficas
- ▶ diferencias con haplotipos en otras poblaciones globales
- ▶ estructura de los haplotipos mestizos
- ▶ haplotipos indígenas
- ▶ los puntos de recombinación son fácilmente identificables
- ▶ efectos de migración
- ▶ senales de selección
- ▶ frecuencia de otros variantes estructurales

# Haplotipos de referencia Colombianos

## A nivel de SNPs

- ▶ frecuencia de SNPs de riesgo en población general
- ▶ efectos de SNPs en un contexto de mestizaje
- ▶ cambios funcionales
- ▶ cambios de riesgo a enfermedades
- ▶ efecto en la interacción entre SNPs (epistasia)

# Haplotipos de referencia Colombianos

## imputación

- ▶ como se efecta la imputación al tener haplotipos mestizos
- ▶ como afecta el faseado al tener una población mestiza
- ▶ los heplotipos del los 1000 genomas son representativos?