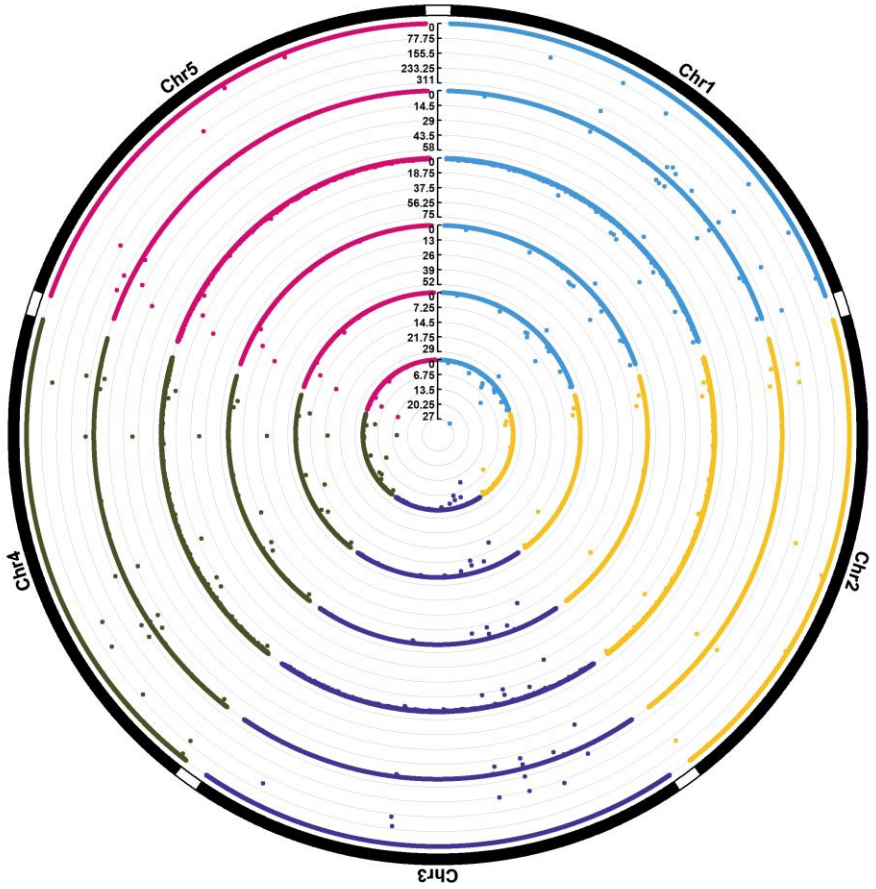


Mapeo por asociación

Dr. Martín N. García

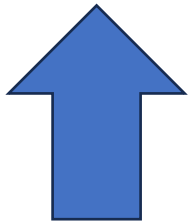
garcia.martin@inta.gob.ar

Lic. Catalina Molina

molina.catalina@inta.gob.ar

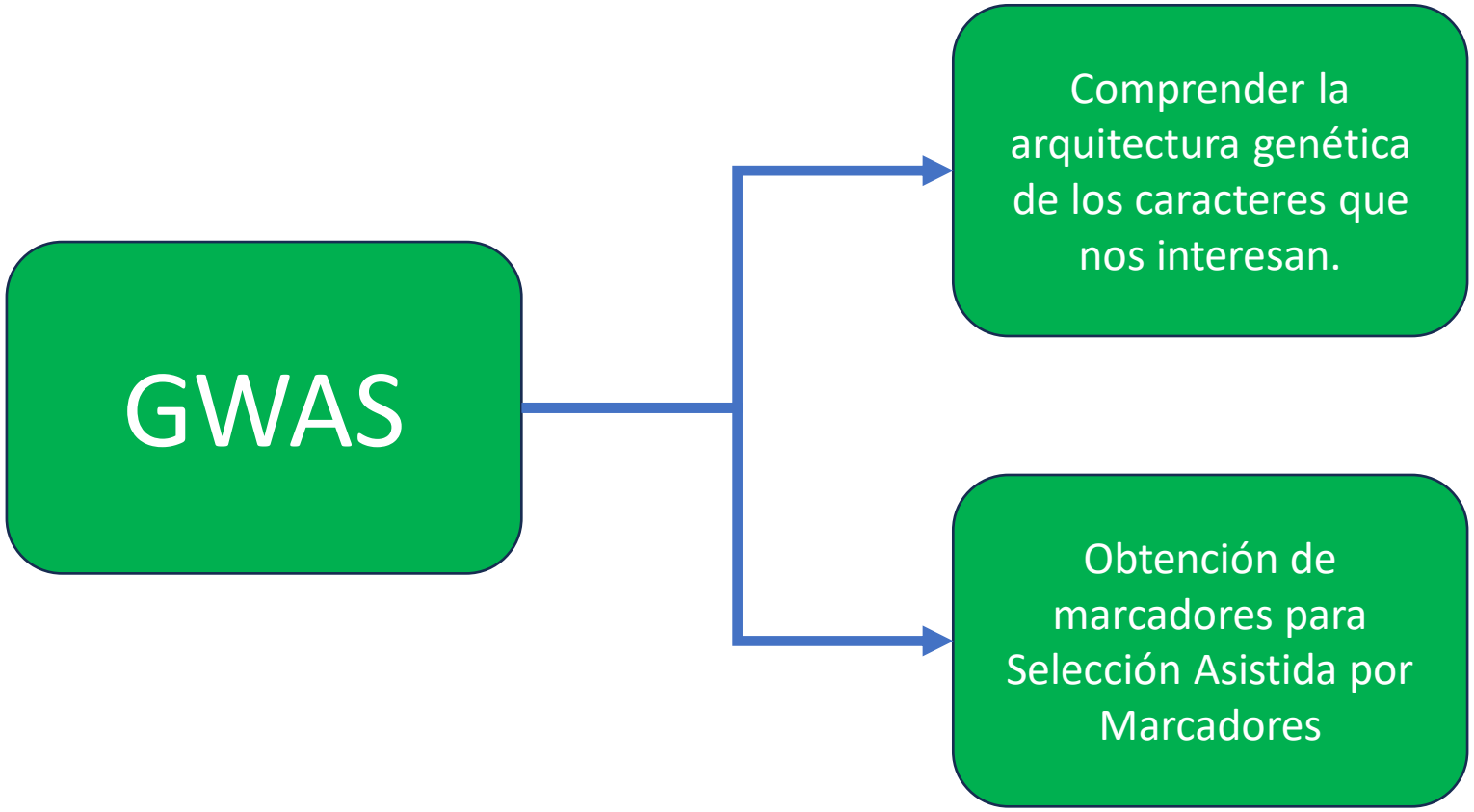
La mayor parte de los caracteres de interés agronómico en plantas cultivadas **son cuantitativos**, debiendo su patron de variación al efecto combinado de numerosos genes y del ambiente.

F = G + A



GWAS: Estudios de Asociación de Genoma Amplio

Son metodologías estadísticas que realizan la búsqueda de marcadores ampliamente distribuidos en el genoma sobre una población no relacionada para hallar variantes genéticas asociadas a un caracter particular.

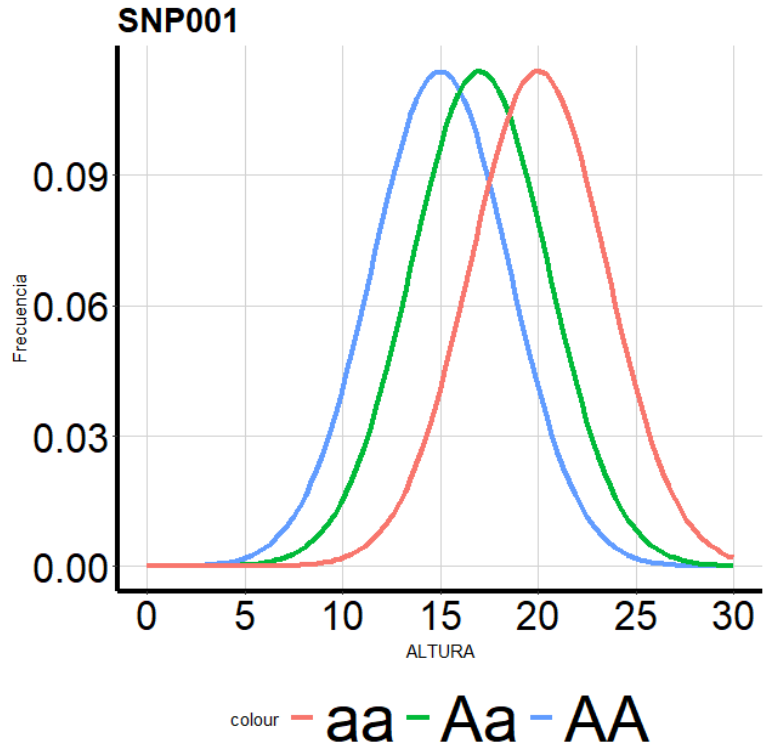

GWAS
Comprender la arquitectura genética de los caracteres que nos interesan.
Obtención de marcadores para Selección Asistida por Marcadores

Conceptualmente…

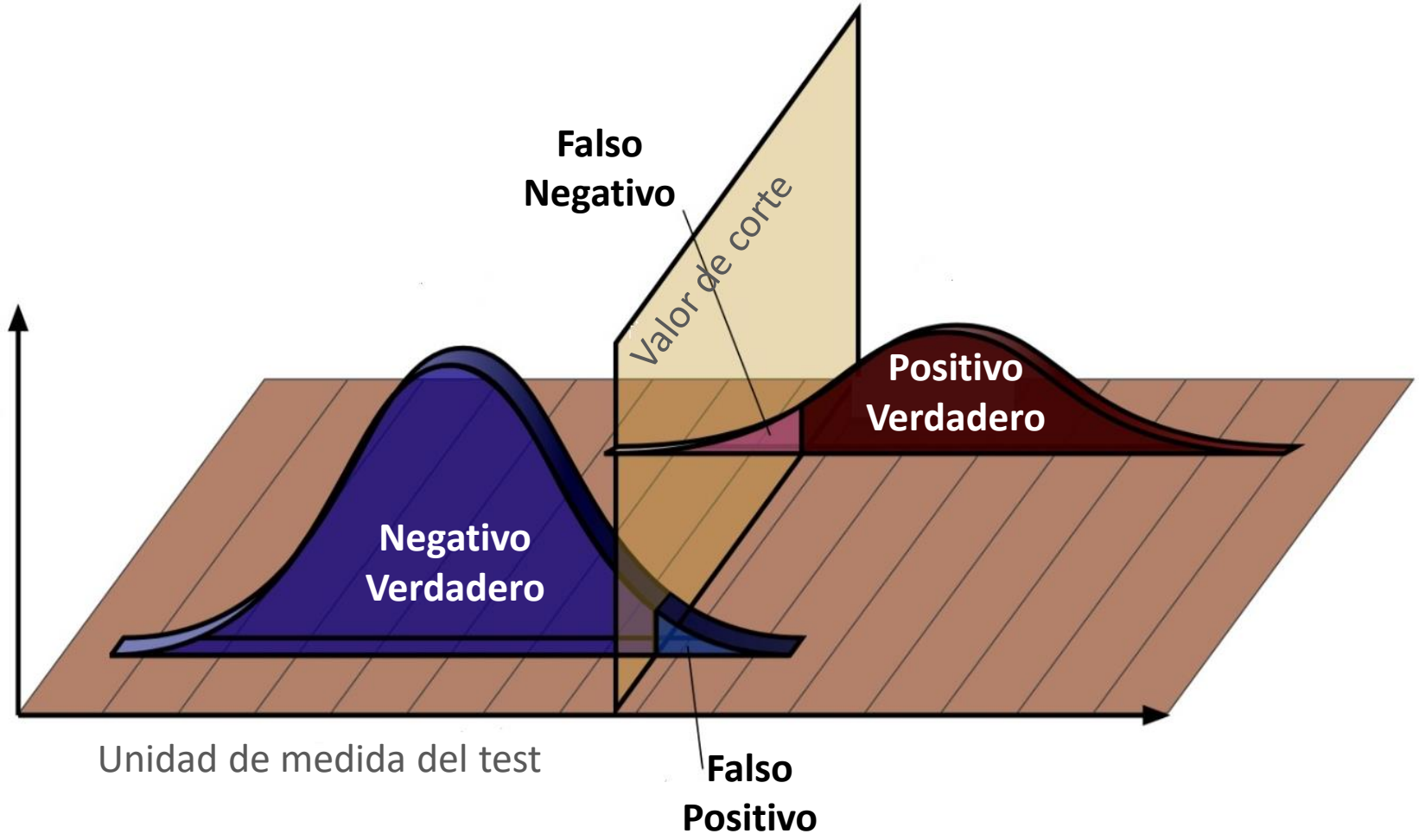
El objetivo de GWAS es hallar aquellos marcadores en los cuales la variación en el genotipo está significativamente asociada con la variación en el fenotipo.

Esto, en su forma más simple, podría ser realizado mediante una prueba estadística, como un **ANOVA**, sobre cada SNP individual. La hipótesis nula sería que no hay diferencia entre la media del caracter para cada grupo genotípico (AA, Aa, aa) y puede ser testeada para cada SNP (Bush y Moore, 2012).

Un gran problema de esta metodología naive es su alta tasa de falsos positivos, lo cual ocurre cuando un hallazgo es declarado significativo a pesar de no ser verdadero. Esto en parte se debe a que testear SNP en el conjunto de datos resulta en miles o millones de tests estadísticos.



Para comprender el problema con esto, **consideremos un umbral de significancia de 0.05**, comunmente utilizado en pruebas estadísticas. Lo que **significa que el investigador esta aceptando una tasa de falsos positivos de hasta el 5%**. Para un solo test esto es un riesgo aceptable. Sin embargo, cuanto mas marcadores son testeados la probabilidad de que alguno de ellos sea un falso positivo se incrementa.



Adaptado de imagen original de Luigi Albert Maria; Wikipedia

Métodos comunes de corrección de testeos múltiples incluyen limitar la tasa de descubrimiento de falsos positivos (**FDR**), la cual es la proporción de todos los resultados positivos que se espera que sean falsos positivos. O usando la **corrección de Bonferroni**, la cual divide el umbral de significancia deseado por el número total de pruebas conduciendo a determinar el umbral de significancia correcto.

$$\alpha_{\text{Bonferroni}} = \alpha_{\text{deseado}} / N_{\text{pruebas}}$$

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, *57*(1), 289-300.

Comparación con el mapeo biparental

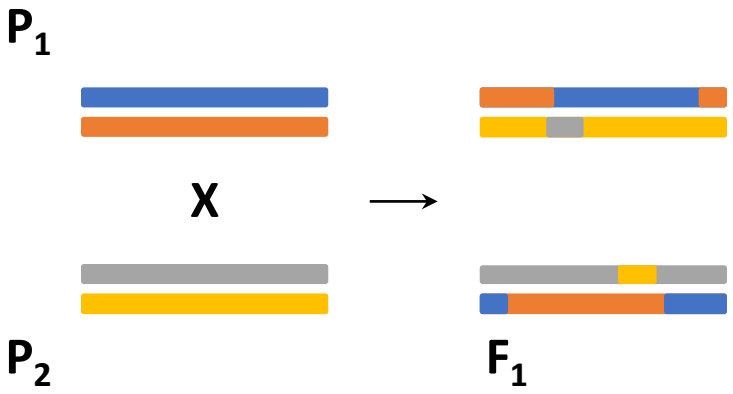
El analisis de ligamiento para mapeo de QTL fue el precursor directo de los estudios de asociacion incluyendo GWAS.

El mapeo de ligamiento estudia individuos con relaciones conocidas. Por ejemplo, los analisis de mapeo de ligamiento en especies cultivadas utilizan progenies generadas para este proposito desde cruzamientos biparentales, ya sean individuos F2 o RILs. Los marcadores asociados con el caracter de interes van a cosegregar con el fenotipo de interes mas de lo esperado por azar.

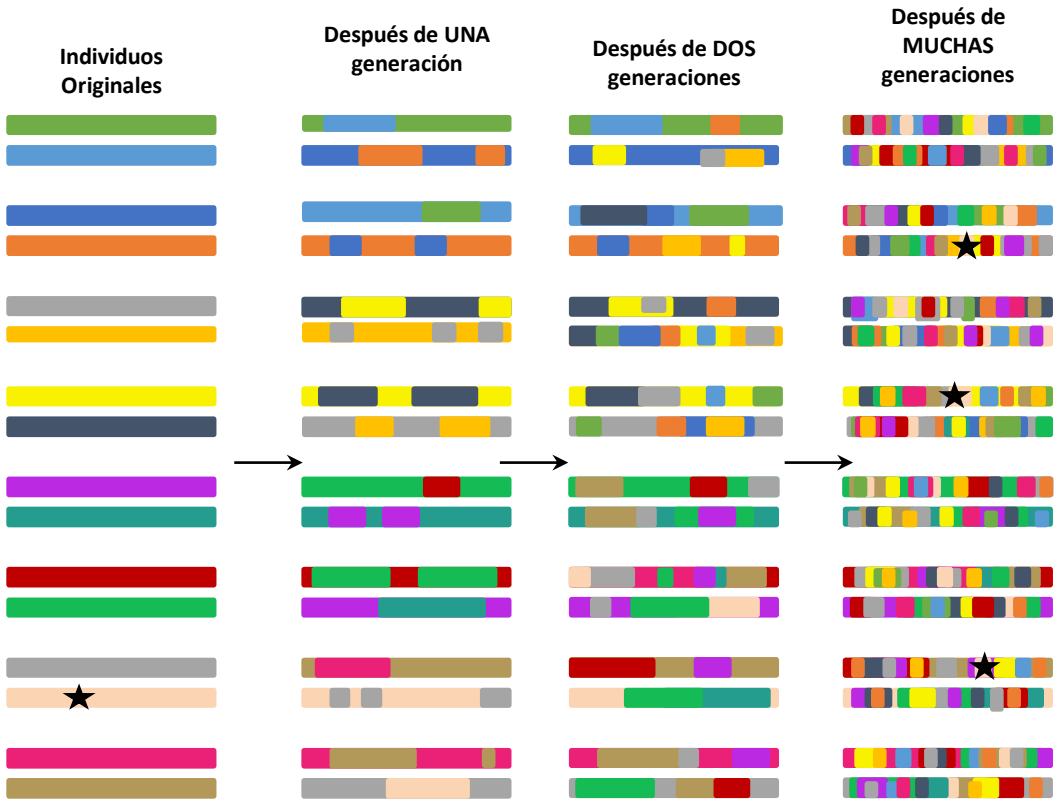
Debido a que los individuos estudiados son muy cercanos en pedigree **pocas rondas de recombinacion han ocurrido desde su ancestro común más cercano**, y por ende hay presentes grandes bloques de ligamiento. Esto significa que **los marcadores genéticos utilizados no tienen que ser tan densos como los utilizados en GWAS**.

Hoy el análisis de ligamiento y GWAS **son metodologías complementarias** que pueden ser usadas para comprender caracteres complejos en diferentes poblaciones.

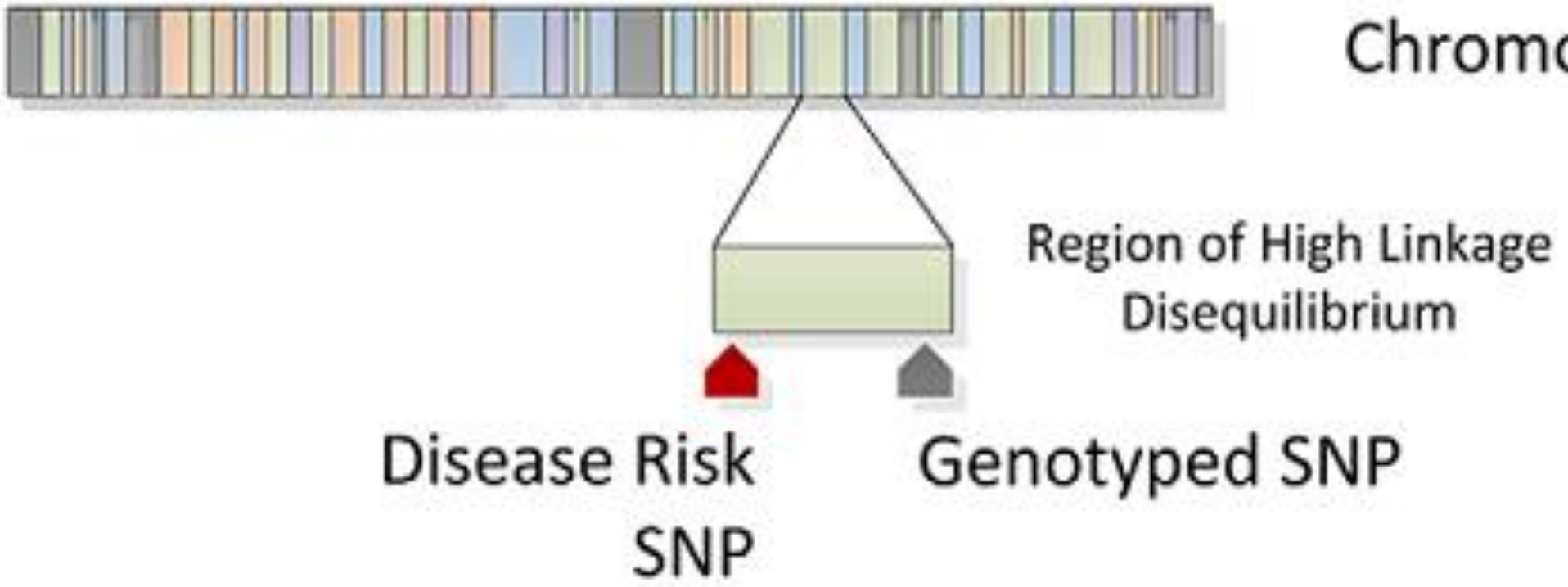
- Mapeo por Ligamiento



- Mapeo por Asociación



Indirect Association



doi:10.1371/journal.pcbi.1002822.g003

Ventajas de los estudios de asociación vs. estudios de ligamiento

- eliminan la necesidad de los cruzamientos experimentales

- incrementan el poder de detectar genes con efectos menores

- mejoran la resolución de pequeños bloques de LD

Línea de tiempo

Las ideas detras de GWAS fueron discutidas teóricamente al comienzo de la mitad de 1990 (Lander 1996, Lander y Kruglyak 1995, Lander y Schork 1994, Risch y Merikangas 1996). Estas publicaciones tempranas anticiparon problemas que aun estamos abordando, incluyendo altas tasas de falsos positivos como resultado de la estructura de la poblacion y el testeo múltiple.

Sin embargo, sus ideas tuvieron que esperar para ser puestas en práctica hasta la publicación del genoma humano y el desarrollo de los primeros conjuntos de datos de SNP tales como dbSNP y HapMap.

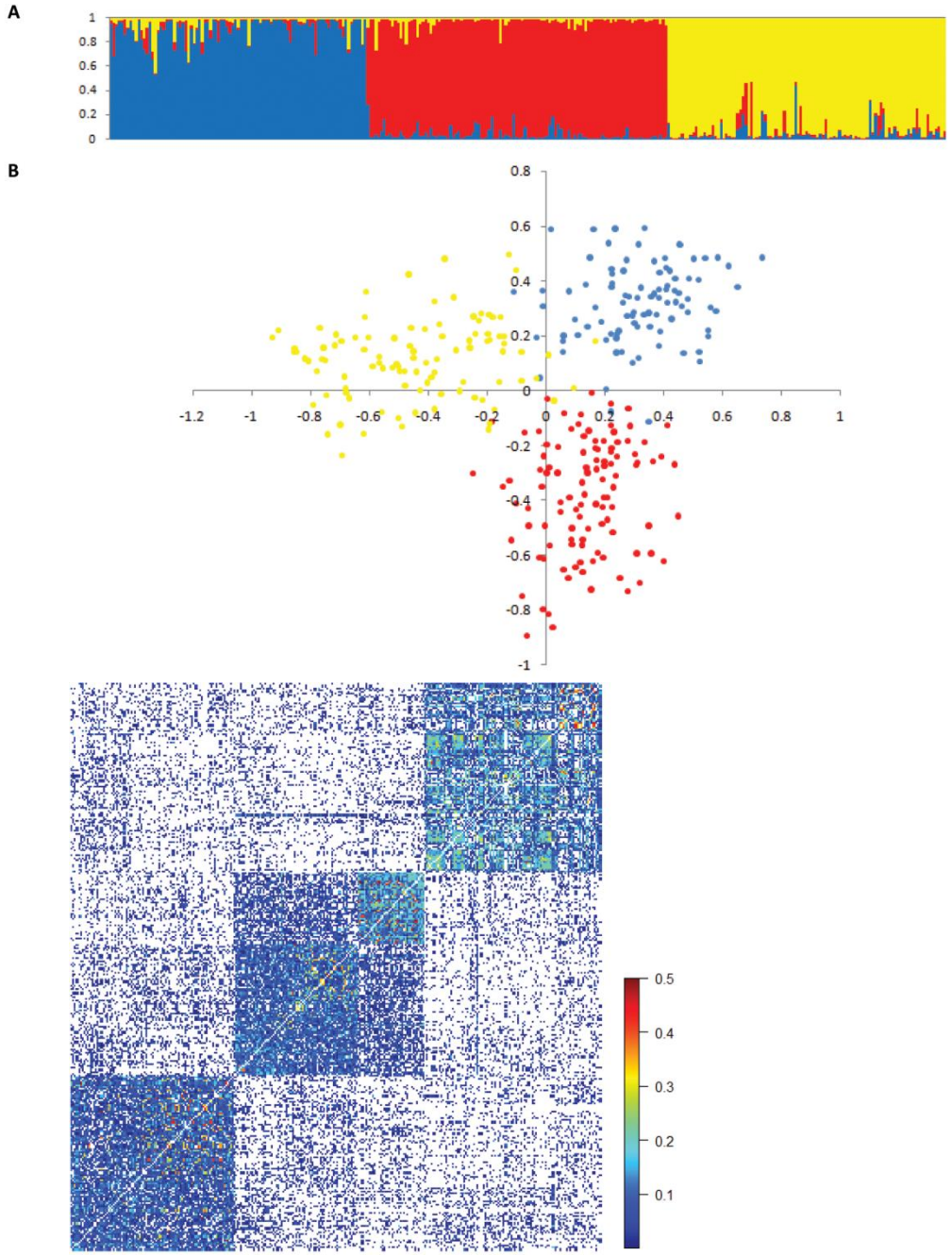
En 2002 primera publicación de GWAS. Este estudio basado en 65000 SNP y 94 individuos, reporto SNPs asociados al riesgo de infarto de miocardio (Ozaki et al, 2002).

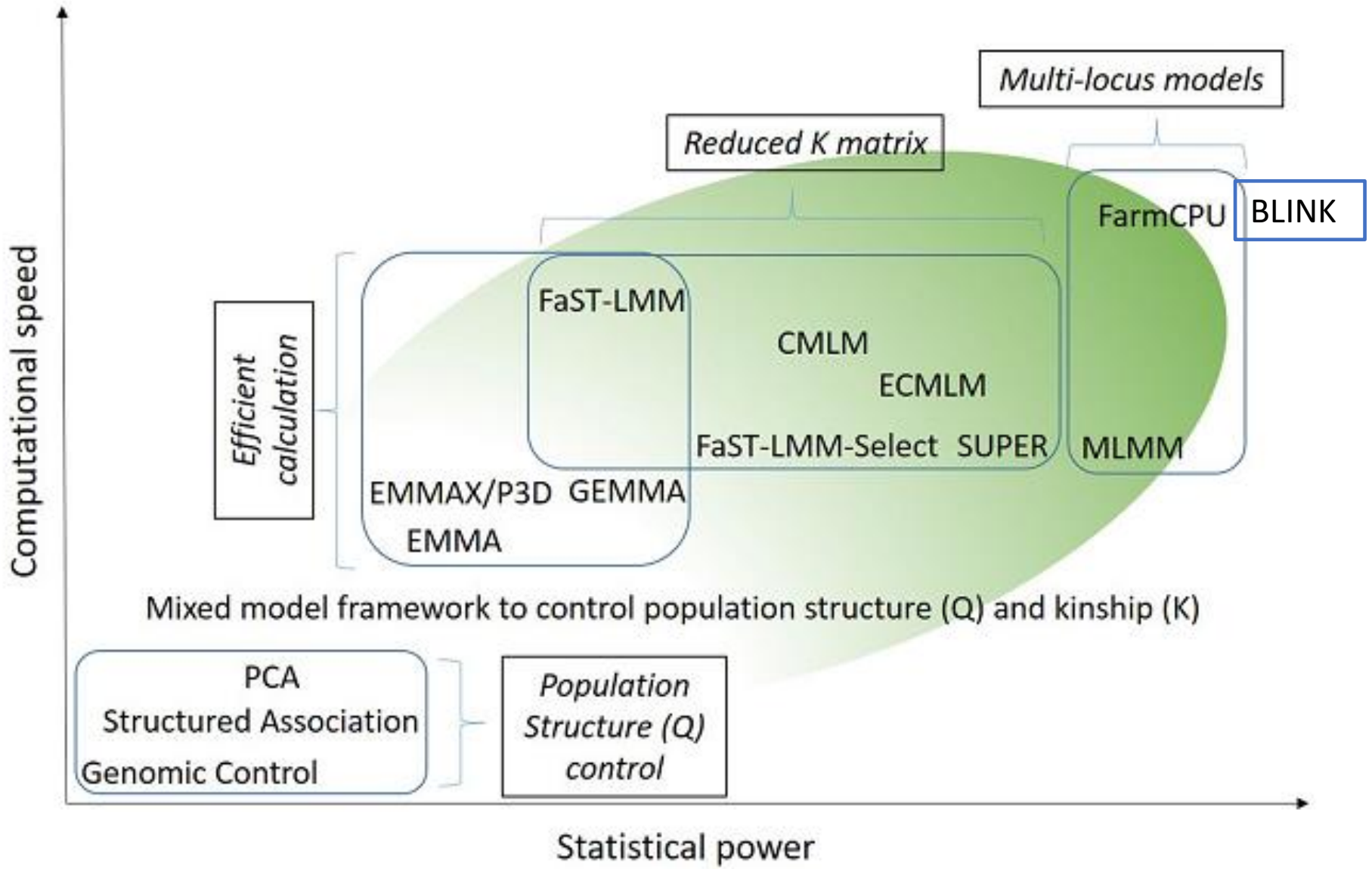
3 años despues, en 2005, el primer estudio fuera del área de la genética médica humana fue publicado, en Arabidopsis (Aranzana et al. 2005). Enseguida fueron publicados estudios en ganado (Abasht y Lamont, 2007) y especies cultivadas (Beló et al, 2008)

To test the association of the oleic acid content with the markers the following ANOVA model was used: $Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$, where Y is the oleic acid content, μ is the overall mean, α is the genotype effect, and ε is the experimental error for the genotype i and plant j.

Modelos Lineales Mixtos

- **base** $Y = 1\mu + X\beta + \varepsilon$
- **Q** $Y = 1\mu + X\beta + Q c + \varepsilon$
- **G** $Y = 1\mu + X\beta + G d + \varepsilon$
- **Q+G** $Y = 1\mu + X\beta + Q c + G d + \varepsilon$

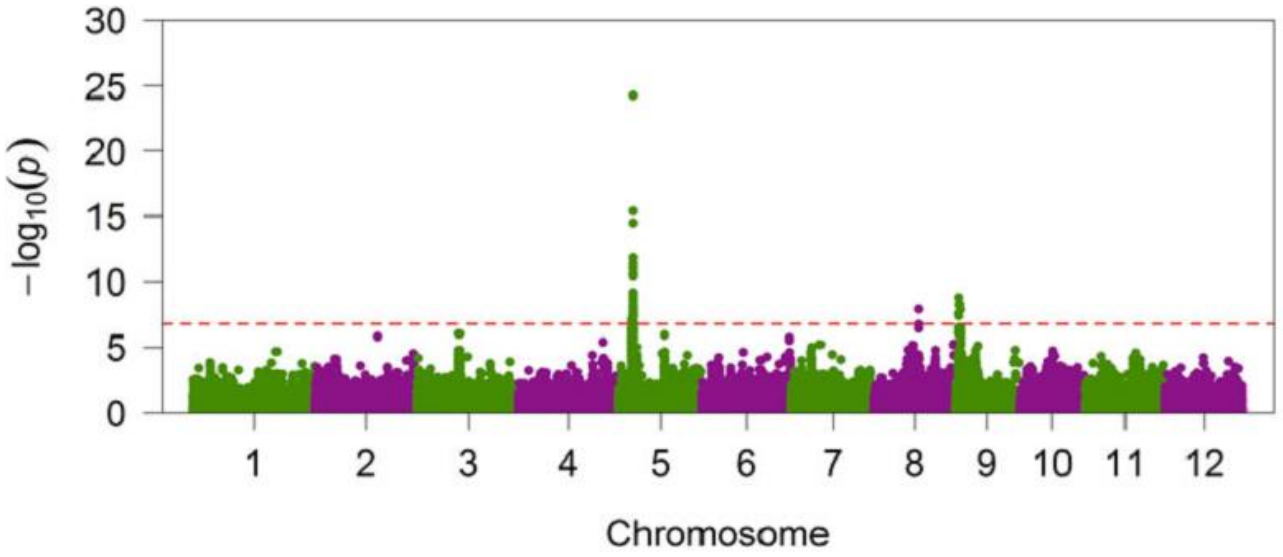




Doi: 0.1002/tpg2.20077

Métodos multi-locus

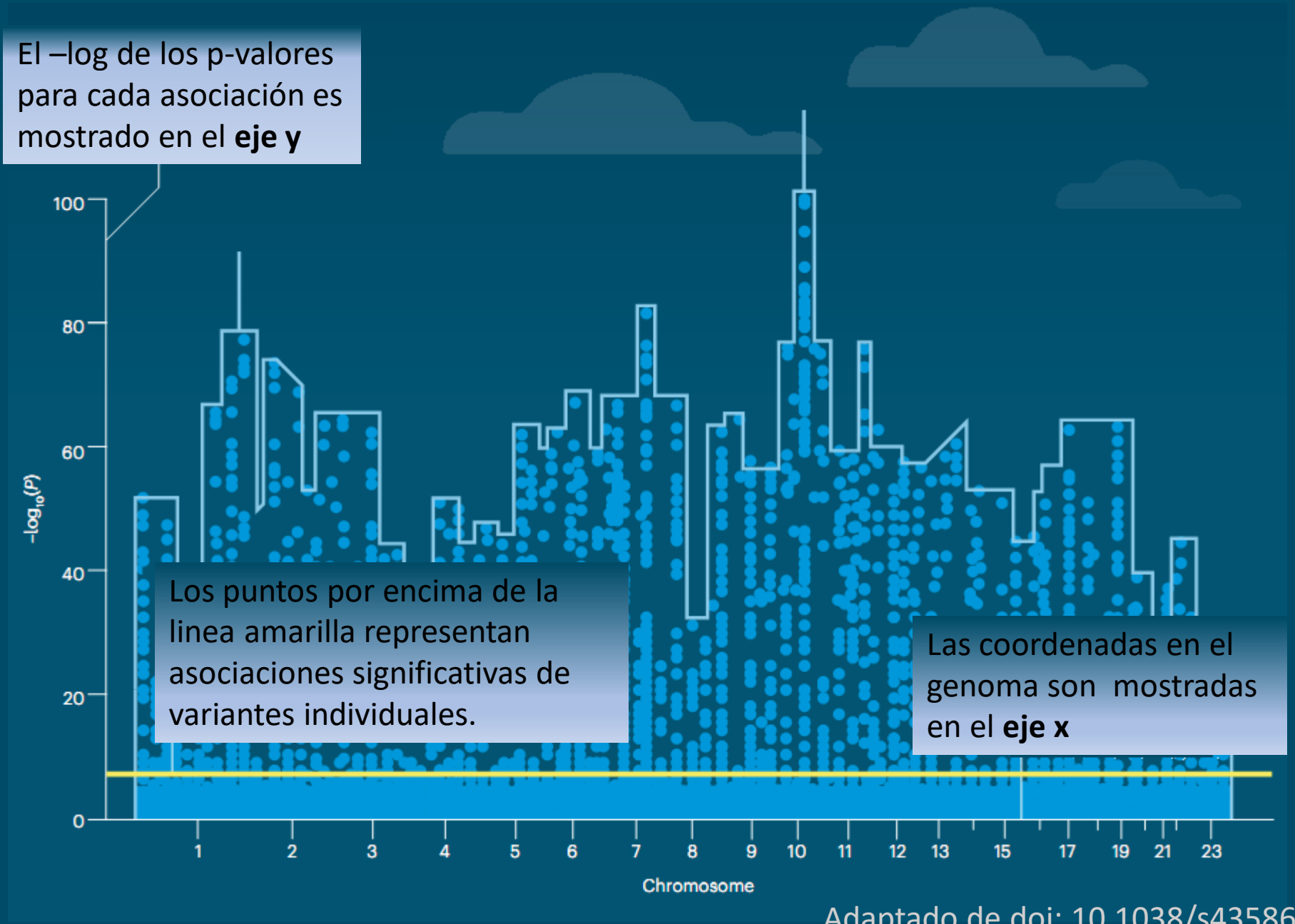
- Los métodos GWAS de locus múltiples mejoran el poder estadístico con respecto a los métodos de locus único al incorporar múltiples marcadores en el modelo simultáneamente como covariables. Este enfoque se implementó por primera vez en el modelo mixto de locus múltiples (MLMM) (Segura et al., 2012). El MLMM es un enfoque iterativo; en cada paso, se estiman los componentes de la varianza genética y del error y luego se utilizan para calcular los valores de p para la asociación de cada SNP con el rasgo de interés.



Manhattan plots

Manhattan plot

Los resultados son usalmente presentados en un Manhattan plot; cada punto representa una variante individual.



Adaptado de doi: 10.1038/s43586-021-00061-y

GWAS (generalidades)

Fenotipos y genotipos son tomados de una población de gran tamaño.

El genotipo usualmente consiste en SNPs (obtenidos mediante GBS o microarreglos).

Los marcadores asociados al caracter de interés son hallados utilizando métodos estadísticos.

Mientras que es posible que el marcador asociado se encuentre dentro de un gen causal, no siempre es así.

GWAS (generalidades)

Fenotipos y genotipos son tomados de una población de gran tamaño.

El genotipo usualmente consiste en SNPs (obtenidos mediante GBS o microarreglos).

Los marcadores asociados al caracter de interés son hallados utilizando métodos estadísticos.

Mientras que es posible que el marcador asociado se encuentre dentro de un gen causal, no siempre es así.

Tamaño de la población y la maldición del ganador

- En poblaciones pequeñas se agrava un sesgo de verificación que se conoce como “Beavis effect” o “Winner’s curse”, donde el efecto de los marcadores asociados están inflados. Esta sobreestimación del efecto genético puede hacer que los estudios de replicación fracasen porque se subestima el tamaño de muestra necesario.

GWAS (generalidades)

Fenotipos y genotipos son tomados de una población de gran tamaño.

El genotipo usualmente consiste en SNPs (obtenidos mediante GBS o microarreglos).

Los marcadores asociados al caracter de interés son hallados utilizando métodos estadísticos.

Mientras que es posible que el marcador asociado se encuentre dentro de un gen causal, no siempre es así.

Selección de la metodología de genotipificación

- Disponibilidad?
- Cobertura?
- Tasa de error de genotipificación?
- Tasa de datos perdidos?
- Dispondremos de esa plataforma en el futuro?

GWAS (generalidades)

Fenotipos y genotipos son tomados de una población de gran tamaño.

El genotipo usualmente consiste en SNPs (obtenidos mediante GBS o microarreglos).

Los marcadores asociados al caracter de interés son hallados utilizando métodos estadísticos.

Mientras que es posible que el marcador asociado se encuentre dentro de un gen causal, no siempre es así.

¿Qué metodología utilizar?

¡La de mayor poder estadístico!

GWAS (generalidades)

Fenotipos y genotipos son tomados de una población de gran tamaño.

El genotipo usualmente consiste en SNPs (obtenidos mediante GBS o microarreglos).

Los marcadores asociados al caracter de interés son hallados utilizando métodos estadísticos.

Mientras que es posible que el marcador asociado se encuentre dentro de un gen causal, no siempre es así.

Post-GWAS

- Comparación con otros estudios
- Búsqueda bioinformática de genes cercanos a marcadores asociados
- Generación de modelo predictivo
- Validación en población externa

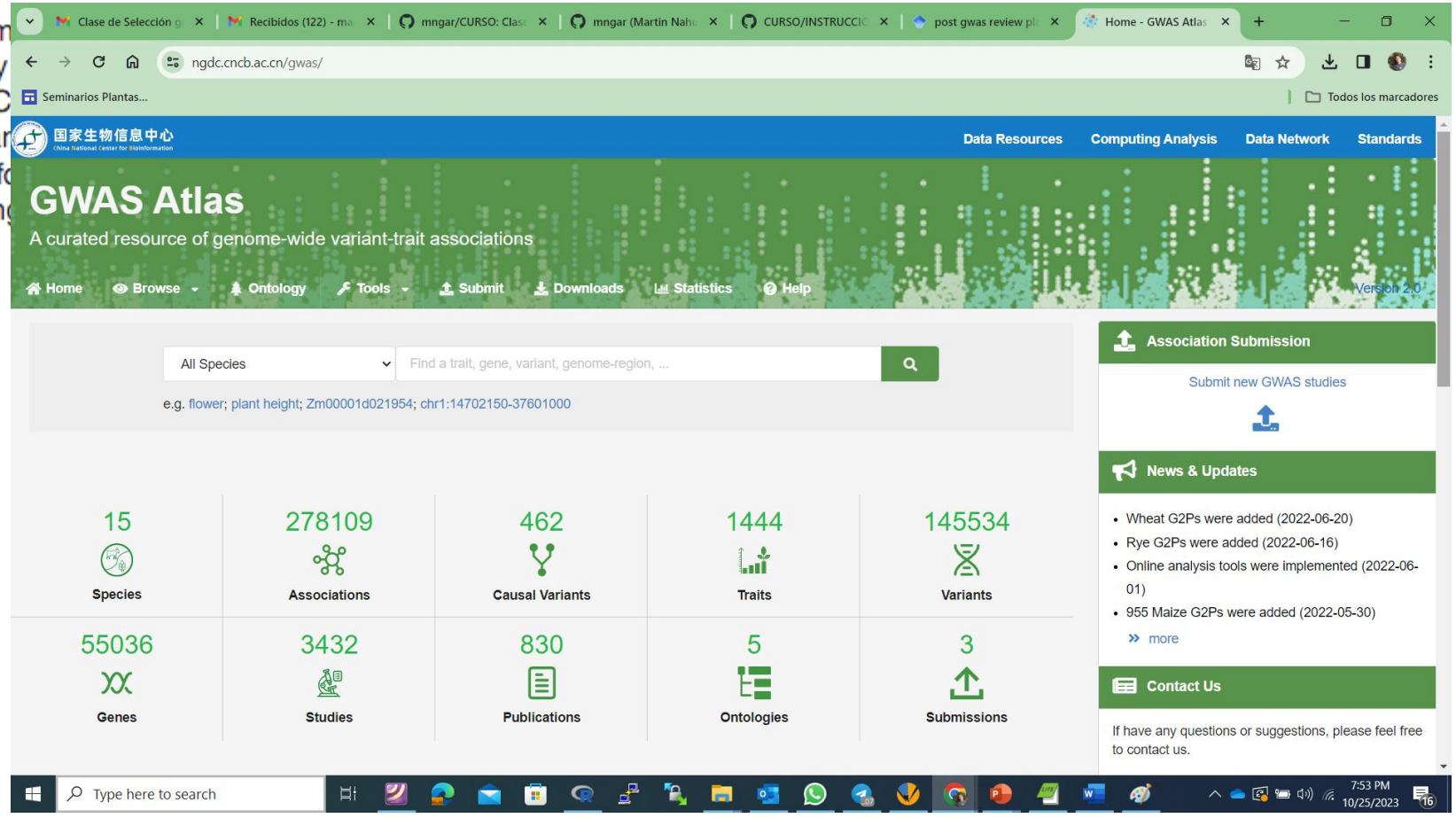
Published online 20 October 2022 *Nucleic Acids Research, 2023, Vol. 51, Database issue* **D969–D976**
https://doi.org/10.1093/nar/gkac924

GWAS Atlas: an updated knowledgebase integrating more curated associations in plants and animals

Xiaonan Liu [1,2,5,†], Dongmei Tian [1,†], Cuiping Li [1,†], Bixia Tang [1], Zhonghuang Wang [1,2,5], Rongqin Zhang [1,2,4,5], Yitong Pan [1,3,5], Yi Wang [1,2,5], Dong Zou [1], Zhang Zhang [1,2,4,5,*] and Shuhui Song [1,2,4,5,*]

[1]National Genomics Data Center, Beijing Institute of Genom
Center for Bioinformation, Beijing 100101, China, [2]CAS Key
Institute of Genomics, Chinese Academy of Sciences and C
Beijing 100101, China, [3]CAS Key Laboratory of Genomic ar
Chinese Academy of Sciences and China National Center fo
College, University of Chinese Academy of Sciences, Beijin
Sciences, Beijing 100049, China

https://ngdc.cncb.ac.cn/gwas/



Aspectos más destacados de GWAS

Investigación de caracteres de importancia agronómica en las especies cultivadas de mayor relevancia: maíz, trigo, soja, sorgo, cebada, algodón y muchas otras.

Identificación de regiones genómicas asociadas con muchos caracteres agronómicos, fisiológicos, tiempo de floración, tolerancia a estrés y rendimiento de granos.

Y a otros caracteres: Identificación de genes asociados con divergencia geográfica y adaptación durante la domesticación en arroz (Chen et al, 2019); así como con fenotipos bioquímicos y moleculares incluyendo flavonoides, ácidos grasos, amino ácidos, etc (Chen et al, 2016).

Datos generados mediante fenotipado de alta procesividad también han sido analizados mediante GWAS, por ejemplo: asociaciones para arquitectura de la panicula en sorgo utilizando extracción automática de caracteristicas desde imagenes y para biomasa utilizando drones.

DOI: 10.1002/tpg2.20077

Aspectos más destacados de GWAS

GWAS puede ser conducida como una investigación en sí, para detectar nuevas asociaciones, así como un componente de un estudio de clonado de genes, o como el paso fundacional en MAS.

También ha sido usada para llevar adelante ingeniería genética, como en el caso del maíz transgénico tolerante a la sequía después de la detección del gen ZmVPP1 por GWAS (Wang et al, 2016).

Para identificar blancos de edición génica, particularmente los basados en CRISPR (Zhang et al, 2018)

Qué es R?

es un lenguaje y entorno para gráficos y estadística computacional. Es un proyecto GNU similar al lenguaje y al e**R**ntorno **S** desarrollado en Bell Laboratories (anteriormente AT&T, ahora Lucent Technologies) por John Chambers y sus colegas.

Paquete → un conjunto de funciones

```
suma <- function(x, y, z) {
        output <- x+y+z
        print(output)
        output <<- output}
```

```
 suma(1,2,3)
[1] 6
 output
[1] 6
```

Parámetros

Con "<<-" le pedimos que el objeto permanezca en el entorno de trabajo.

Repositorios

Es el repositorio oficial, coordinado por el R core team. El paquete debe pasar ciertas pruebas y seguir las políticas de CRAN.

Dirigido a software bioinformático. Al igual que CRAN hay que pasar por un proceso de revisión.

Es el más popular, simple de usar y cuenta con espacio ilimitado.

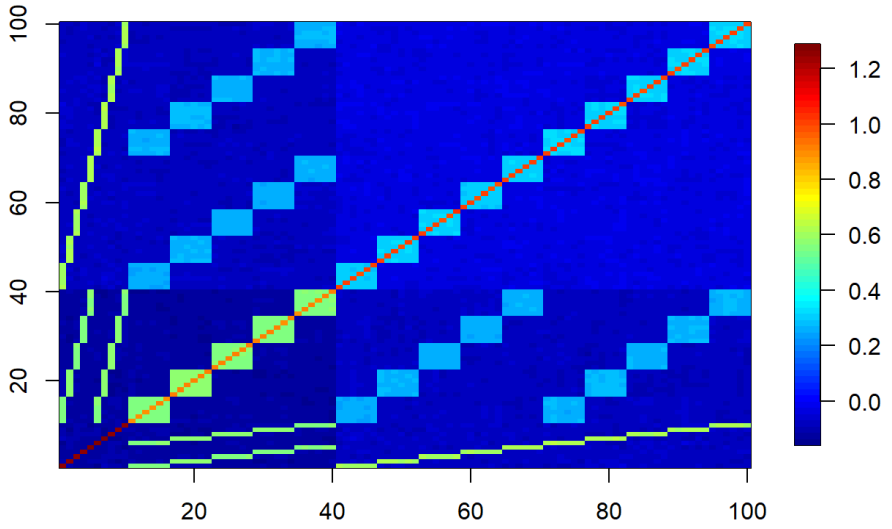
simulMGF

simGeno: simula genotipos en **matrices de SNP** como valores aleatorios desde una distribución uniforme, para organismos diploides (codificados por 0, 1 y 2). (VanRaden et al., 2009; Sikorska et al., 2013)

simPheno: simula un **fenotipo desde una matriz de SNP** con loci de caracteres cuantitativos (**QTLs**) con efectos muestreados desde una distribución Normal.

simulFS y simulHS: simula los **genotipos de progenies de hermanos completos** y de **medios hermanos** desde el genotipo de los parentales.

simulN y simulU : simulan una **matriz de SNP** y **caracteres controlados por determinado número de QTLs** y sus **efectos muestreados desde una distribución Normal** o una **distribución Uniforme** respectivamente.



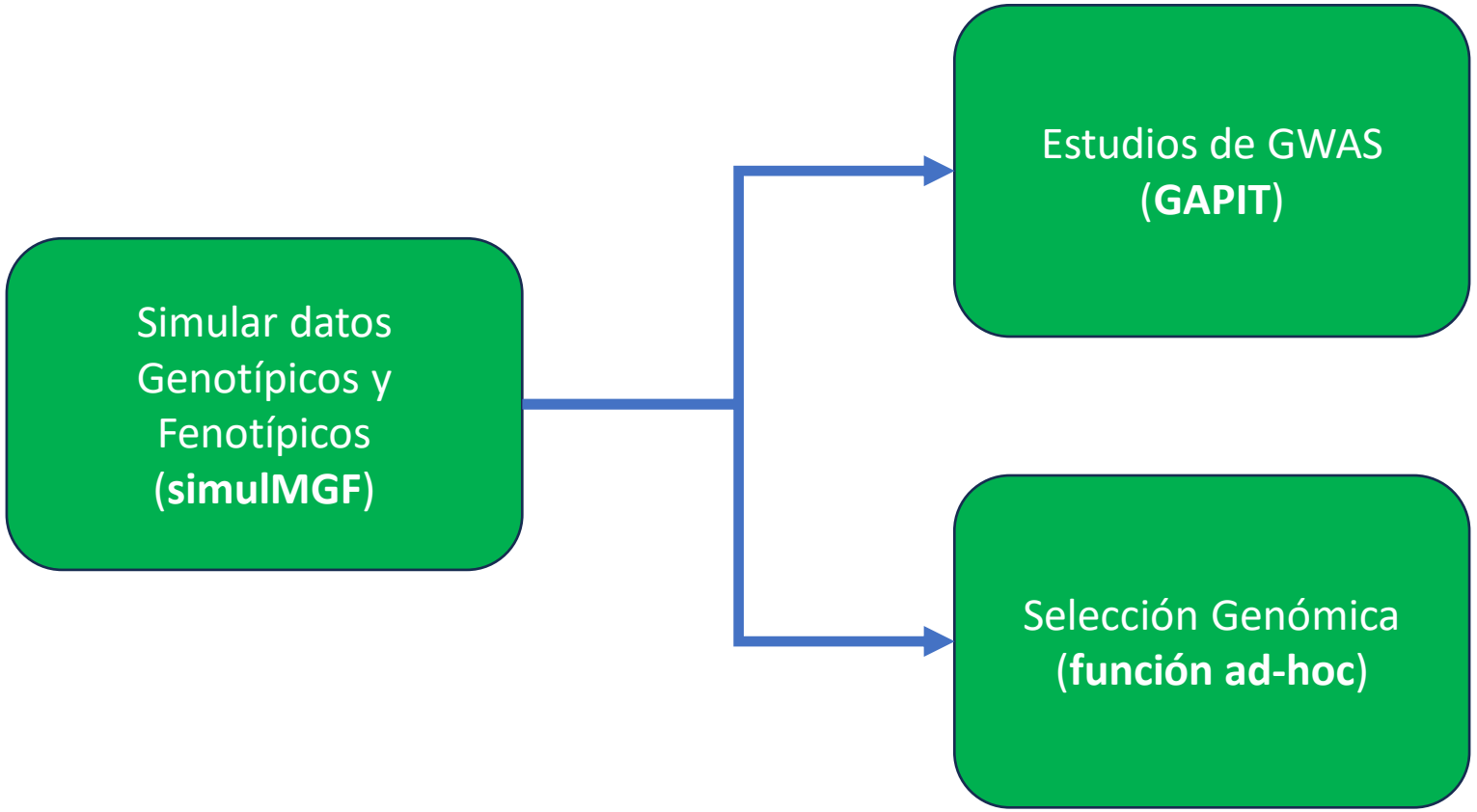
García, M.N. (2023). simulMGF: Simulate SNP Matrix, Phenotype and Genotypic Effects. R package version 0.1.1.
https://CRAN.R-project.org/package=simulMGF

simulMGF

simGeno(Nind, Nmarkers)	→	una matriz de dimensiones Nind x Nmarkers.
simPheno(x, Nqtl, Esigma, Pmean, Perror)	→	un objeto de clase “list” conteniendo el carácter, los marcadores asociados y sus efectos.
simulFS(x, y, Nprogeny)	→	una matriz de dimensiones (nrow(x)*Nprogeny) x ncol(x)
simulHS(x, Nprogeny)	→	una matriz de dimensiones (nrow(x)*Nprogeny) x ncol(x)
simulN(Nind, Nmarkers, Nqtl, Esigma, Pmean, Perror)	→	un objeto de clase “list” conteniendo una matriz de SNP, el caracter, los marcadores asociados y sus efectos.
simulU(Nind, Nmarkers, Nqtl, Pmean, Perror)	→	un objeto de clase “list” conteniendo una matriz de SNP, el caracter, los marcadores asociados y sus efectos.

García, M.N. (2023). simulMGF: Simulate SNP Matrix, Phenotype and Genotypic Effects. R package version 0.1.1.
https://CRAN.R-project.org/package=simulMGF

TPs de “Mapeo por Asociación” y “Selección Genómica”



Simulación de datos

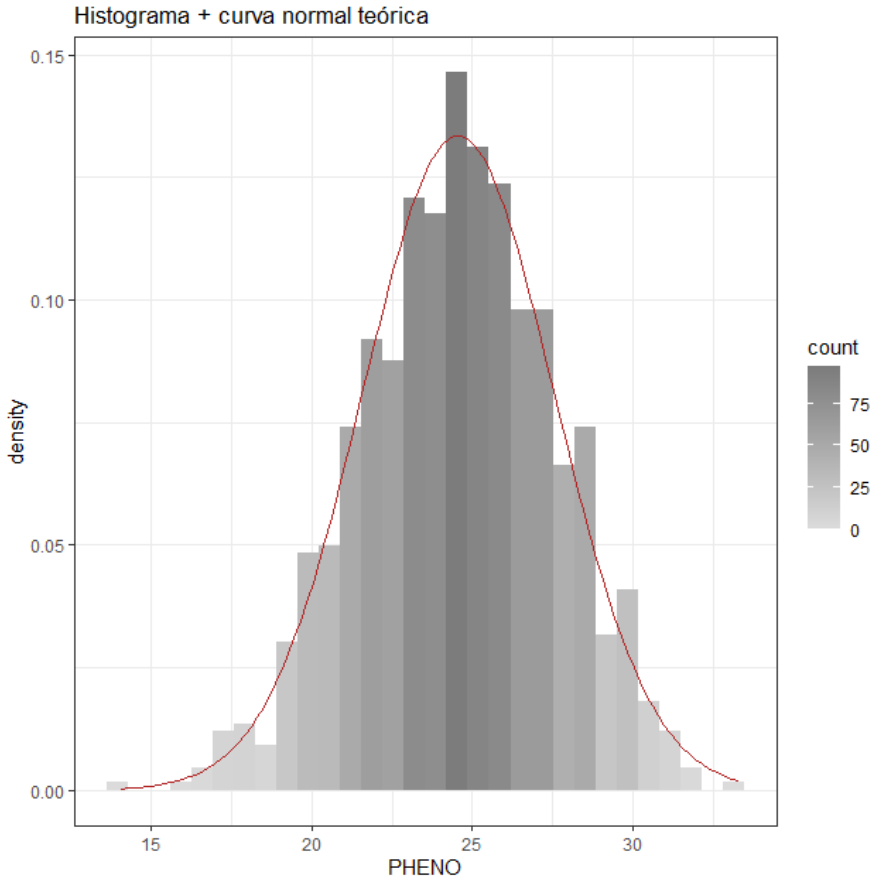
```
install.packages("simulMGF")

library(simulMGF)

set.seed(1234)

simulN(Nind = 1000, Nmarkers = 10000, Nqtl = 50,
Esigma = .5, Pmean = 25, Perror = .25)
```

Código disponible en: https://github.com/mngar/CURSO

OBJETO “nsimout”

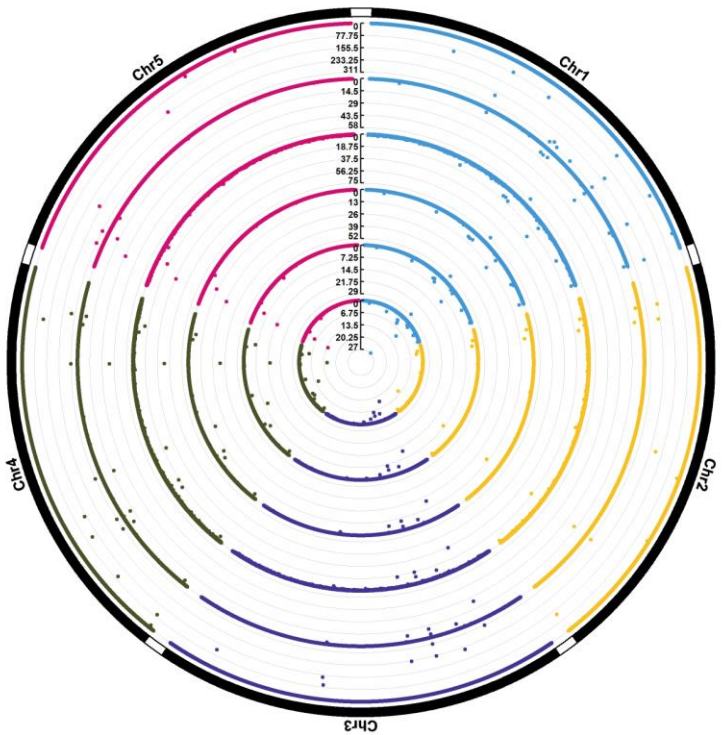
```
str(nsimout)
List of 4
 $ geno     : matriz de genotipos
 $ pheno    : vector de fenotipos
 $ QTN      : vector de marcadores asociados
 $ Meffects : vector de efectos
```

$$y = Pmean + \Sigma\ QTN\ x\ Meffects + \varepsilon$$



Código disponible en: https://github.com/mngar/CURSO

GWAS



Manhattan plot con marcadores asociados para el caracter simulado para cada metodología (de adentro hacia afuera: GLM, MLM, SUPER, MLMM, FarmCPU y Blink).

Poder estadístico estimado para cada metodología de asociación ordenadas de acuerdo a lo esperado (mayor poder estadístico arriba).

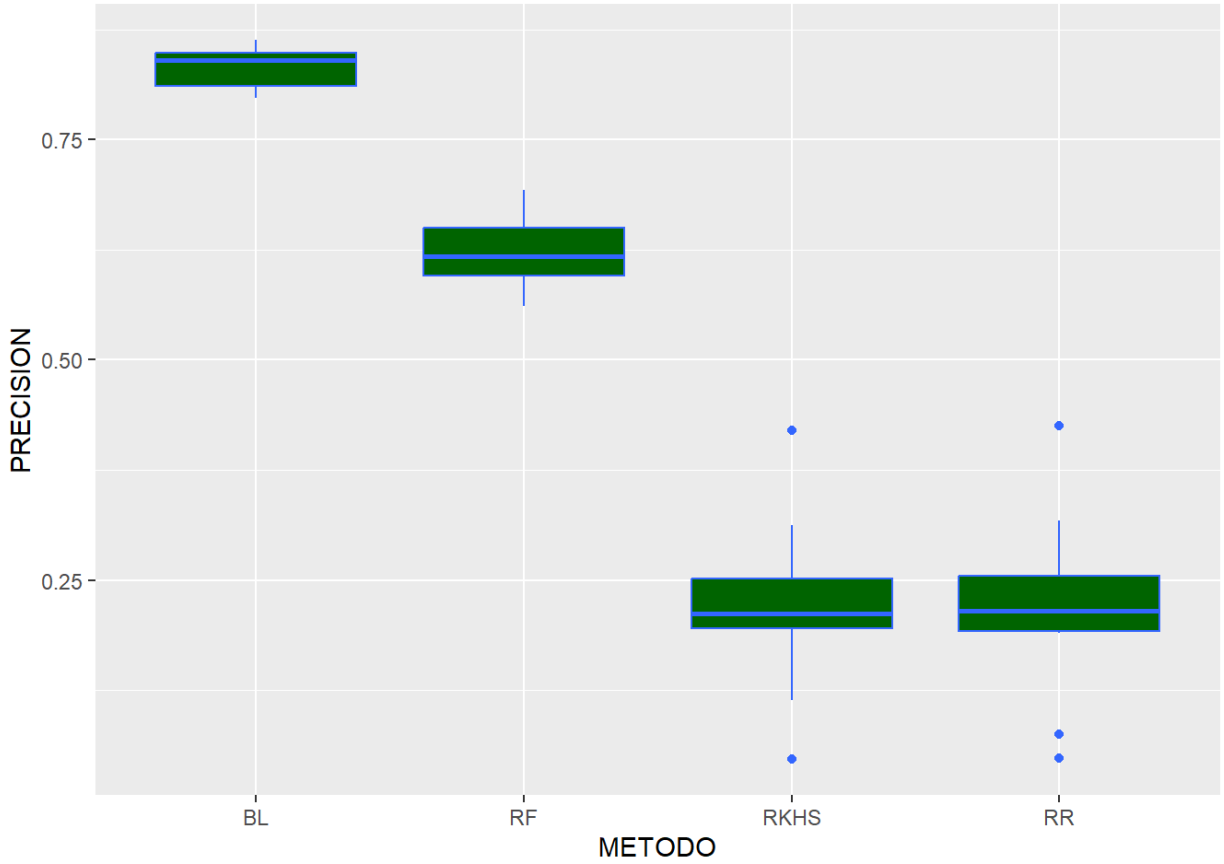
Método	SNP_asociados	SNP_realmente_asociados	Falso_positivos	PODER
blink	49	48	1	0.96
FarmCPU	31	28	3	0.56
MLMM	32	32	0	0.64
SUPER	30	28	2	0.56
MLM	26	24	2	0.48
GLM	30	26	4	0.52

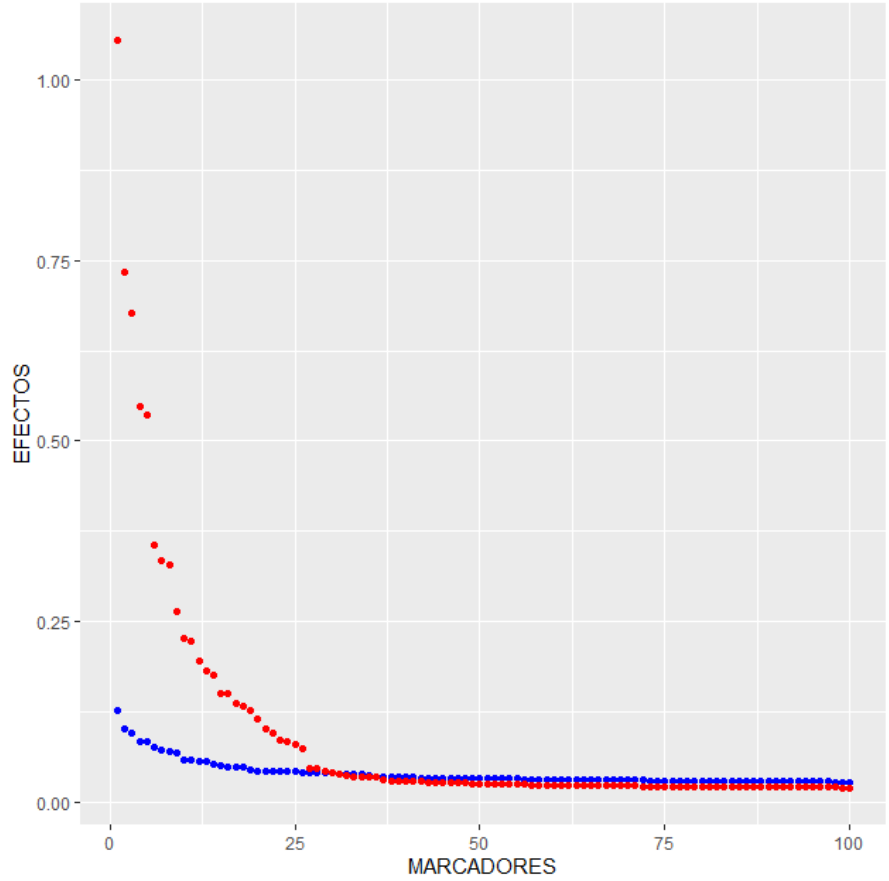
PODER = (SNP_realmente_asociados/#QTL)

Tibbs Cortes, L., Zhang, Z., & Yu, J. (2021). Status and prospects of genome-wide association studies in plants. The plant genome, 14(1), e20077.
Wang, J., Zhang, Z. (2021). GAPIT version 3: boosting power and accuracy for genomic association and prediction. Genomics, proteomics & bioinformatics, 19(4), 629-640.

SG

RR: Ridge Regression (regresión lineal)
BL: Bayesian LASSO (regresión lineal bayesiana)
RF: Random Forest (semiparamétrica)
RKHS: Reproducing Kernel Hilbert Spaces (no paramétrica)





BL

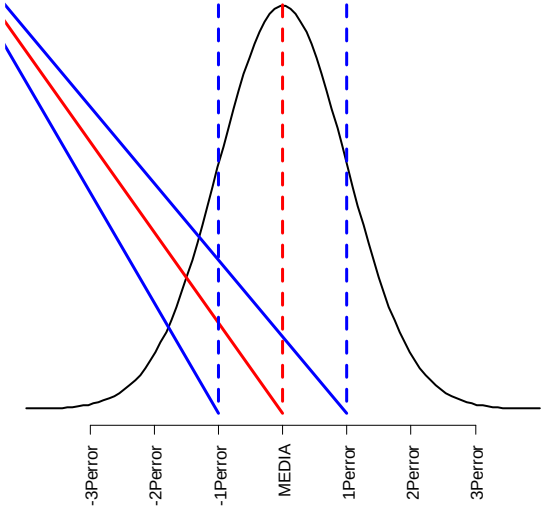
RR

- Endelman, J. B. (2011). Ridge regression and other kernels for genomic selection with R package rrBLUP. The plant genome, 4(3).
- de Los Campos, et al., (2013). Genome-enabled prediction using the BLR (Bayesian Linear Regression) R-package. Genome-wide association studies and genomic prediction, 299-320.
- Breiman, L. (2001). Random forests. Machine learning, 45, 5-32.
- Pérez, P., & de Los Campos, G. (2014). Genome-wide regression and prediction with the BGLR statistical package. Genetics, 198(2), 483-495.

Nuevas funciones (versión 0.1.2)

simClon <- function(x, y, Nclon, Pmean, Perror)

x	matriz de genotipos QTN
y	vector de efectos
Nclon	número de clones
Pmean	media fenotípica
Perror	desvío estándar del fenotipo



Estos dos terminos serán constantes para cada genotipo

$$y = Pmean + \Sigma\ QTN \times Meffects + \varepsilon$$

simEnv <- function(x, Nenv, Pmean, Perror, distr = c("normal", "uniform"), Esigma, Evar)

x	matriz de genotipos QTN
Nenv	número de ambientes
Pmean	media fenotípica en cada ambiente
Perror	desvío estándar del fenotipo en cada ambiente
distr	distribución
Esigma	desvío estándar efectos (distribución N)
Evar	desvío estándar efectos a través de ambientes

P = G + E + GxE

VP = VG + VE + V(GxE)