Mejoramiento Genético de la caña de azúcar mediante el uso de marcadores moleculares

CIDCA, A.C.

Noviembre, 2017.

Contenido

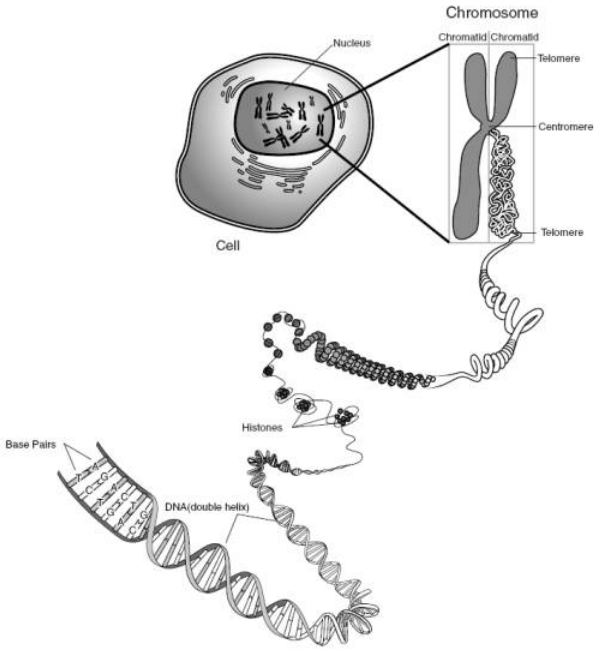
Datos genotípicos

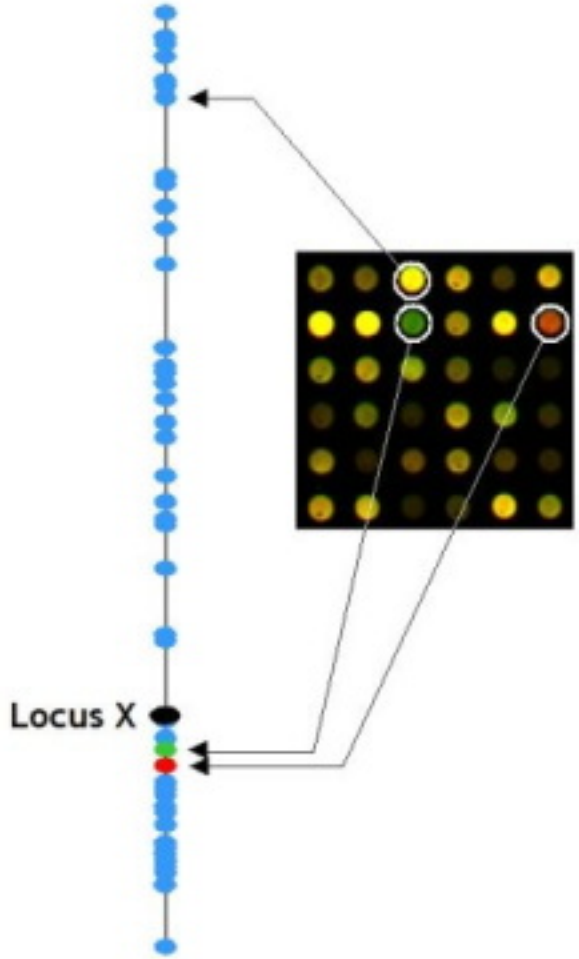
- Información a nivel molecular de 93 variedades de caña de azúcar.
- Información de marcadores obtenida por la compañía Diversity Arrays Technology, Canberra, Au. La compañía informó que las muestras pasaron sus controles de calidad.

Diversity Arrays Technology Pty Ltd
Bldg 3, Level D, University of Canberra (LPO Box 5067), Bruce, ACT 2617, Australia
www.diversityarrays.com<http://www.diversityarrays.com>
On-line ordering service logon<https://www.diversityarrays.com/cgi-bin/order/login.pl>

Dear Dr. Hilda Silva-Rojas,

Samples for the DArT service DSc16-2426 have passed Quality Control.





- Marcadores del tipo “presencia” (1) y ausencia (0).
- $52,828$ marcadores obtenidos.

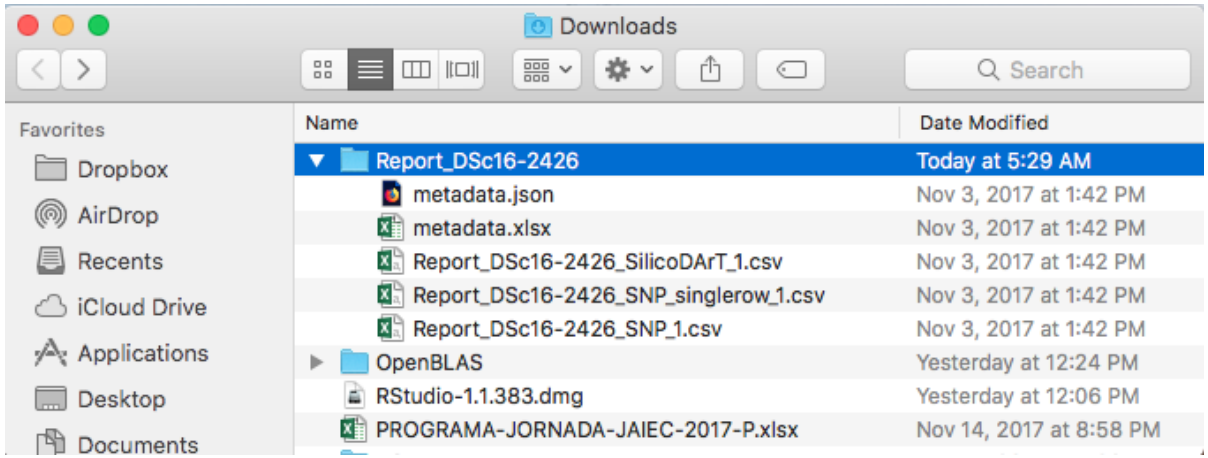
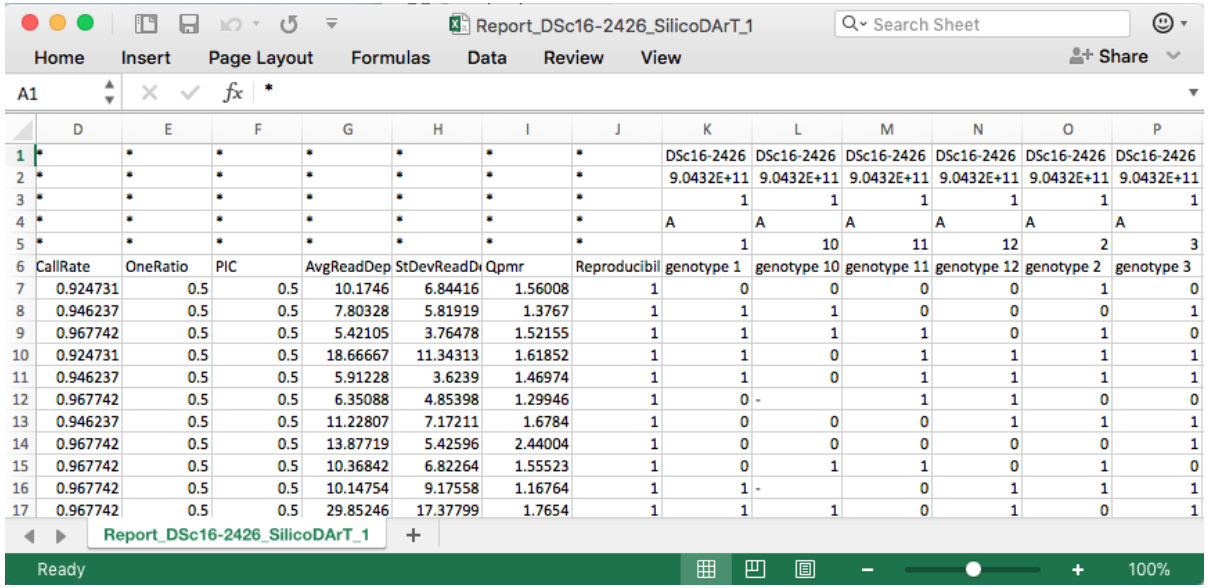


Figura 1: Estructura del directorios con resultados DArT.

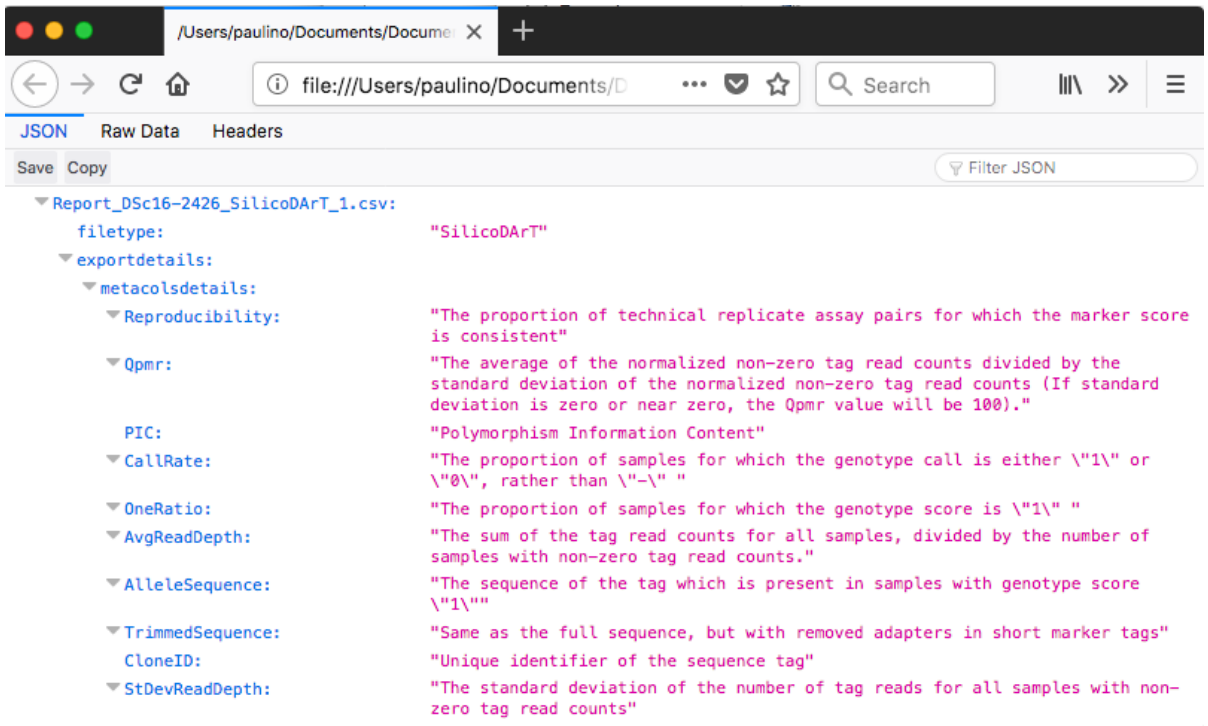


	D	E	F	G	H	I	J	K	L	M	N	O	P
1	*	*	*	*	*	*	*	DSc16-2426	DSc16-2426	DSc16-2426	DSc16-2426	DSc16-2426	DSc16-2426
2	*	*	*	*	*	*	*	9.0432E+11	9.0432E+11	9.0432E+11	9.0432E+11	9.0432E+11	9.0432E+11
3	*	*	*	*	*	*	*	1	1	1	1	1	1
4	*	*	*	*	*	*	*	A	A	A	A	A	A
5	*	*	*	*	*	*	*	1	10	11	12	2	3
6	CallRate	OneRatio	PIC	AvgReadDep	StDevReadD	Qpmr	Reproducibil	genotype 1	genotype 10	genotype 11	genotype 12	genotype 2	genotype 3
7	0.924731	0.5	0.5	10.1746	6.84416	1.56008	1	0	0	0	0	1	0
8	0.946237	0.5	0.5	7.80328	5.81919	1.3767	1	1	1	0	0	0	1
9	0.967742	0.5	0.5	5.42105	3.76478	1.52155	1	1	1	1	0	1	0
10	0.924731	0.5	0.5	18.66667	11.34313	1.61852	1	1	0	1	1	1	1
11	0.946237	0.5	0.5	5.91228	3.6239	1.46974	1	1	0	1	1	1	1
12	0.967742	0.5	0.5	6.35088	4.85398	1.29946	1	0	-	1	1	0	0
13	0.946237	0.5	0.5	11.22807	7.17211	1.6784	1	0	0	0	1	1	1
14	0.967742	0.5	0.5	13.87719	5.42596	2.44004	1	0	0	0	0	0	1
15	0.967742	0.5	0.5	10.36842	6.82264	1.55523	1	0	1	1	0	1	0
16	0.967742	0.5	0.5	10.14754	9.17558	1.16764	1	1	-	0	1	1	1
17	0.967742	0.5	0.5	29.85246	17.37799	1.7654	1	1	1	0	1	0	1

Figura 2: Archivo Report_DSc16-2426_SilicoDArT_1.csv

Continuar...

Las primeras 10 columnas del archivo contienen información de los marcadores,...



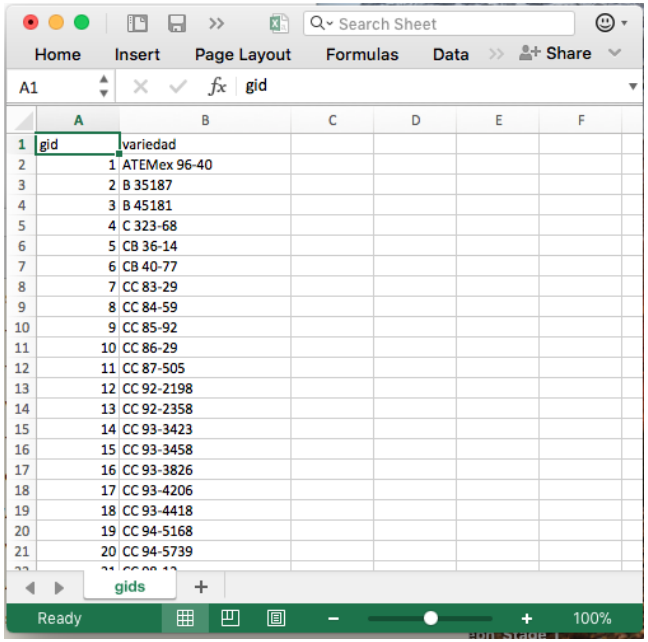
Continuar...

Los datos pueden ser analizados utilizando diferentes programas, en este caso se utilizó R, http://www.r-project.org,

```
#Clean workspace
rm(list=ls())

library(cluster)

setwd("~/Regression models")
source("rCode/Impute.R")

Geno=read.csv(file="data/Report_DSc16-2426_SilicoDArT_1.csv",skip=5,
              header=TRUE,na.strings="-")
```

La compañía no conoce el nombre de los materiales, la información se puede incluir posteriormente al momento de realizar los análisis.

Continuar...

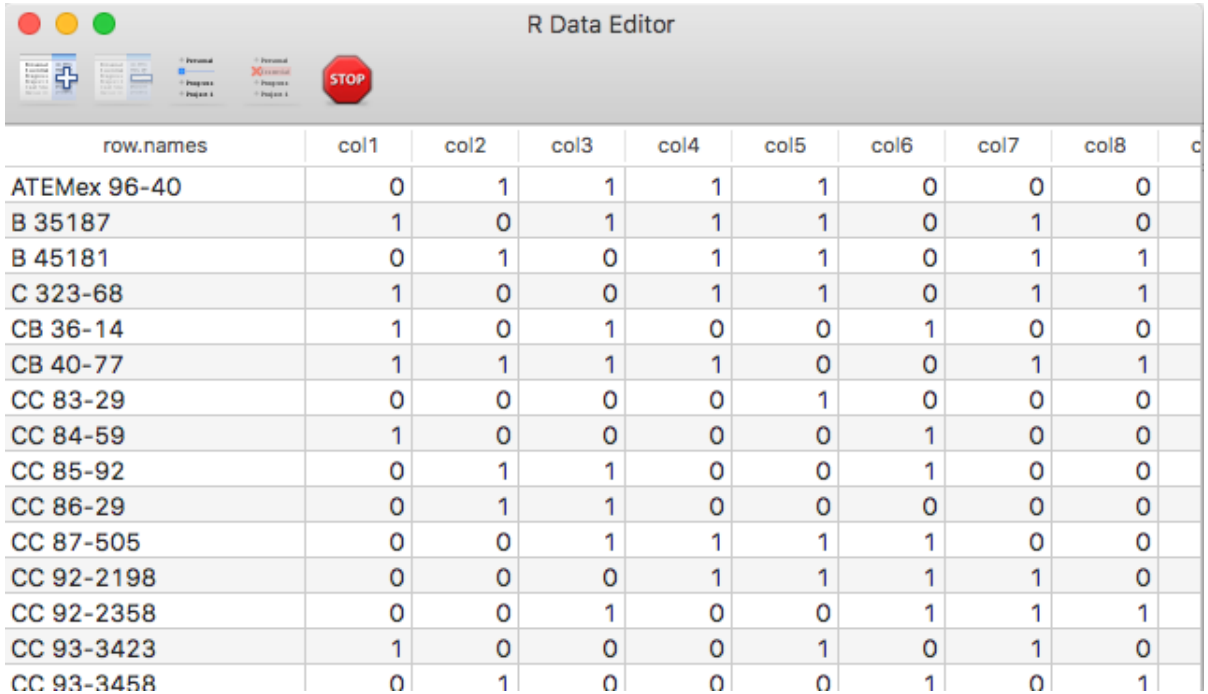


	A	B
1	gid	variedad
2	1	ATEMex 96-40
3	2	B 35187
4	3	B 45181
5	4	C 323-68
6	5	CB 36-14
7	6	CB 40-77
8	7	CC 83-29
9	8	CC 84-59
10	9	CC 85-92
11	10	CC 86-29
12	11	CC 87-505
13	12	CC 92-2198
14	13	CC 92-2358
15	14	CC 93-3423
16	15	CC 93-3458
17	16	CC 93-3826
18	17	CC 93-4206
19	18	CC 93-4418
20	19	CC 94-5168
21	20	CC 94-5739

Figura 3: Relación entre GID y variedad.

En R:

```
gids=read.csv(file="data/gids.csv",header=TRUE)
CloneID=Geno[,1]
Geno=Geno[,-c(1:10)]
#Take the transpose of Genotypes
Geno=as.matrix(Geno)
```



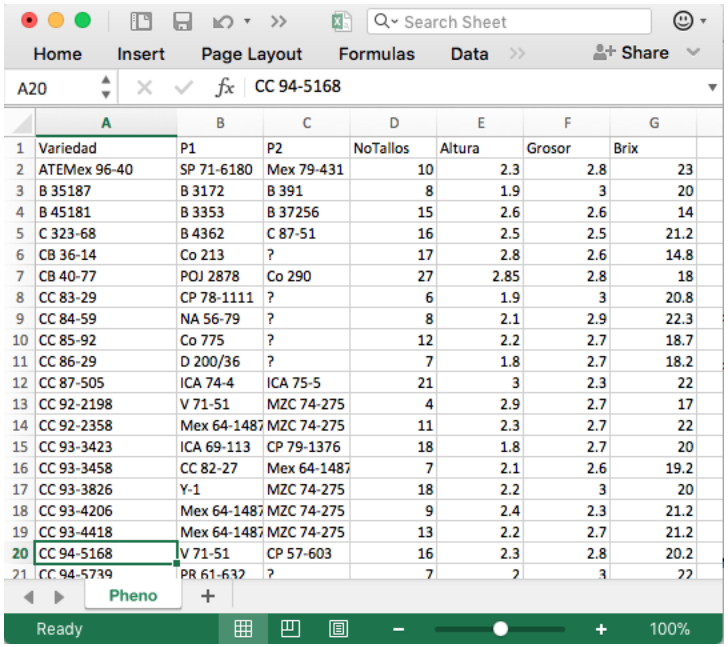
row.names	col1	col2	col3	col4	col5	col6	col7	col8
ATEMex 96-40	0	1	1	1	1	0	0	0
B 35187	1	0	1	1	1	0	1	0
B 45181	0	1	0	1	1	0	1	1
C 323-68	1	0	0	1	1	0	1	1
CB 36-14	1	0	1	0	0	1	0	0
CB 40-77	1	1	1	1	0	0	1	1
CC 83-29	0	0	0	0	1	0	0	0
CC 84-59	1	0	0	0	0	1	0	0
CC 85-92	0	1	1	0	0	1	0	0
CC 86-29	0	1	1	0	0	0	0	0
CC 87-505	0	0	1	1	1	1	0	0
CC 92-2198	0	0	0	1	1	1	1	0
CC 92-2358	0	0	1	0	0	1	1	1
CC 93-3423	1	0	0	0	1	0	1	0
CC 93-3458	0	1	0	0	0	1	0	1

Figura 4: Sub conjunto de marcadores.

Datos fenotípicos

- Información fenotípica de 137 variedades de caña.
- Características de interés: Número de tallos molederos, Altura (m) y Grosor (cm), grados Brix.

Continuar...



	A	B	C	D	E	F	G
1	Variedad	P1	P2	NoTallos	Altura	Grosor	Brix
2	ATEMex 96-40	SP 71-6180	Mex 79-431	10	2.3	2.8	23
3	B 35187	B 3172	B 391	8	1.9	3	20
4	B 45181	B 3353	B 37256	15	2.6	2.6	14
5	C 323-68	B 4362	C 87-51	16	2.5	2.5	21.2
6	CB 36-14	Co 213	?	17	2.8	2.6	14.8
7	CB 40-77	POJ 2878	Co 290	27	2.85	2.8	18
8	CC 83-29	CP 78-1111	?	6	1.9	3	20.8
9	CC 84-59	NA 56-79	?	8	2.1	2.9	22.3
10	CC 85-92	Co 775	?	12	2.2	2.7	18.7
11	CC 86-29	D 200/36	?	7	1.8	2.7	18.2
12	CC 87-505	ICA 74-4	ICA 75-5	21	3	2.3	22
13	CC 92-2198	V 71-51	MZC 74-275	4	2.9	2.7	17
14	CC 92-2358	Mex 64-1487	MZC 74-275	11	2.3	2.7	22
15	CC 93-3423	ICA 69-113	CP 79-1376	18	1.8	2.7	20
16	CC 93-3458	CC 82-27	Mex 64-1487	7	2.1	2.6	19.2
17	CC 93-3826	Y-1	MZC 74-275	18	2.2	3	20
18	CC 93-4206	Mex 64-1487	MZC 74-275	9	2.4	2.3	21.2
19	CC 93-4418	Mex 64-1487	MZC 74-275	13	2.2	2.7	21.2
20	CC 94-5168	V 71-51	CP 57-603	16	2.3	2.8	20.2
21	CC 94-5739	PR 61-632	?	7	2	3	22

Figura 5: Datos fenotípicos.

En R:

```
#Phenotypes
Pheno=read.csv(file="data/Pheno.csv",header=TRUE,na.strings="?")
Pheno$Variedad=as.character(Pheno$Variedad)
```

Contenido

Control de calidad

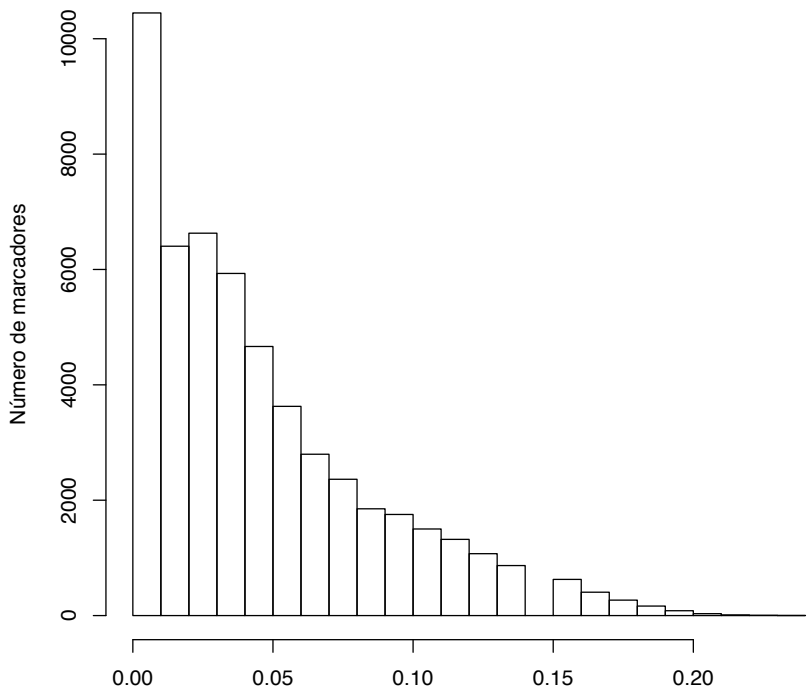


Figura 6: Distribución de valores faltantes.

```
   Min.     1st Qu.  Median  Mean   3rd Qu.    Max.
0.00000 0.01075 0.03226 0.04403 0.06452 0.23660
```

En R:

```
hist(colSums(is.na(Geno))/93,xlab="Proporción de faltantes",
             ylab="Número de marcadores")

summary(colSums(is.na(Geno))/93)
```

Para cada marcador existe una pequeña proporción de valores faltantes, se desarrolló una rutina (“Impute.R”) para imputar estos valores faltantes utilizando las frecuencias observadas. En R:

```
set.seed(123)
out=Impute(Geno)
#Markers already imputed
X=out$X
```

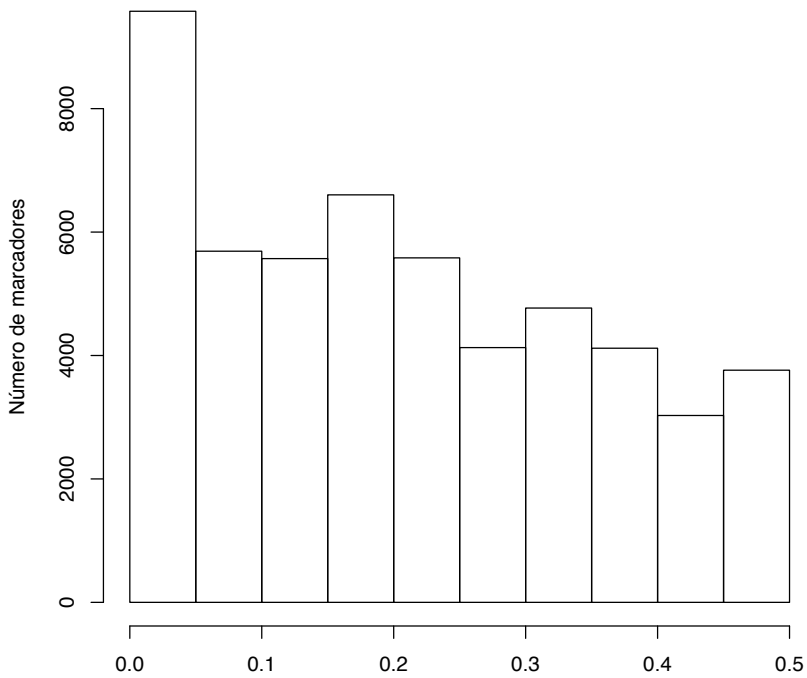


Figura 7: Distribución de frecuencias del alelo menor (MAF).

$\hat{p}_j$, proporción de $1's$ para marcador $j = 1, 2, ...,$

$$maf_j = \begin{cases} \hat{p}_j \text{ si } \hat{p}_j < 0.5 \\ 1 - \hat{p}_j \text{ en caso contrario.} \end{cases}$$

Continuar...

en R:

```
#Allelic frequency
p=colMeans(X)
quartz()
hist(p)

#Minor allele frequency
maf=ifelse(p<0.5,p,1-p)

hist(maf,xlab="MAF",ylab="Número de marcadores")

#Drop markers with maf<0.05
index=maf>0.05
sum(index)

X=X[,index]
p=p[index]
maf=maf[index]
```

Contenido

Matriz de relaciones genómicas

La matriz de relaciones genómicas(***G***) aparece de forma natural en varios modelos usados en selección genómica. VanRaden (2008) estudió algunos métodos eficientes para obtener predicciones basados en esta matriz.

J. Dairy Sci. 91:4414–4423
doi:10.3168/jds.2007-0980

Efficient Methods to Compute Genomic Predictions

P. M. VanRaden[1]
Animal Improvement Programs Laboratory, Agricultural Research Service, USDA, Beltsville, MD 20705-2350

Existen varias maneras de calcular la matriz ***G***,

1. $$\boldsymbol{G} = \boldsymbol{X}\boldsymbol{X}',$$

   donde $\boldsymbol{X}$ es la matriz de marcadores de dimensiones $n \times p$. Para DArT $x_{ij} \in \{0, 1\}$.

2. $$\boldsymbol{G} = \frac{(\boldsymbol{X} - \boldsymbol{E})(\boldsymbol{X} - \boldsymbol{E})'}{2\sum_{j=1}^{p} p_j(1 - p_j)},$$

   donde p_j is la frecuencia del alelo menor para el marcador $j = 1, ..., p$, y $\boldsymbol{E}$ es una matriz de frecuencias esperadas de x_{ij} bajo equilibrio de Hardy-Weinberg.

3. $$\boldsymbol{G} = \frac{\boldsymbol{Z}\boldsymbol{Z}'}{p},$$

   donde $\boldsymbol{Z}$ es la matriz de marcadores centrada y estandarizada y p es el número de marcadores, es decir, $z_{ij} = (x_{ij} - p_j)/\sqrt{p_j(1 - p_j)}$.

Continuar...

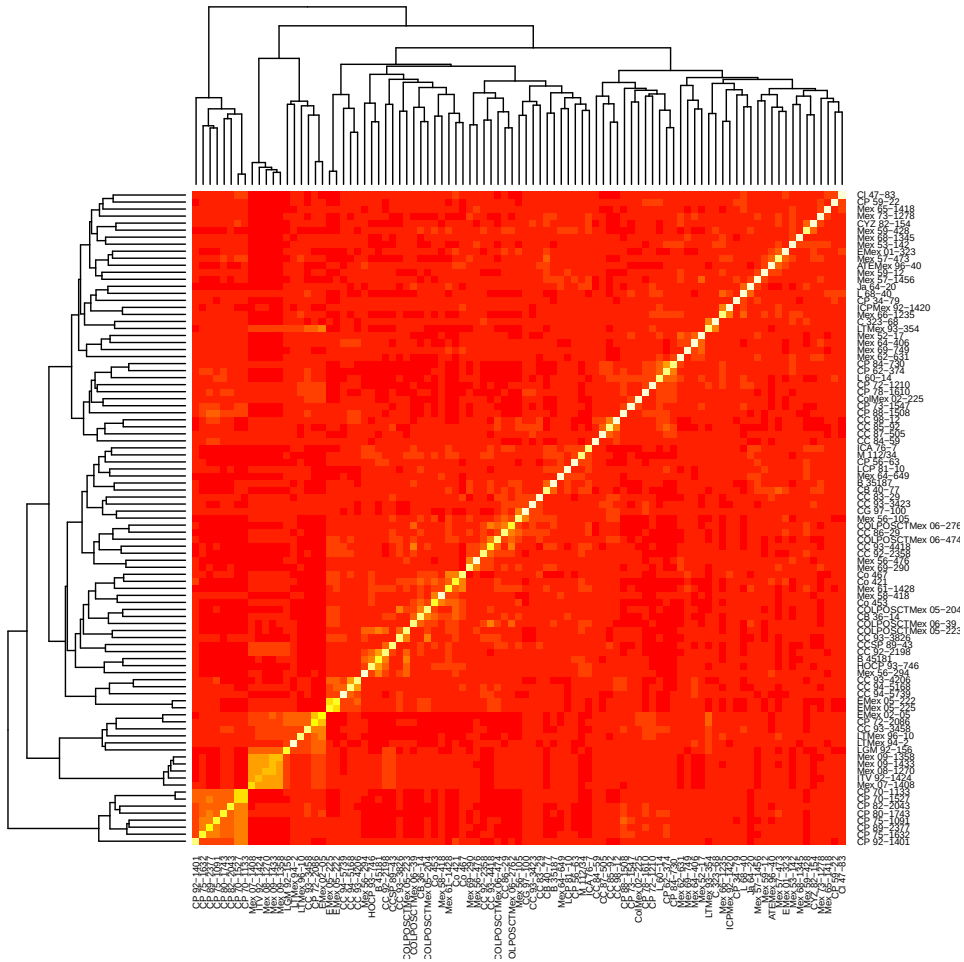


Figura 8: Mapa de calor de la matriz de relaciones genómicas.

Continuar...

En R:

```
Z=scale(X,center=TRUE,scale=TRUE)

G=tcrossprod(Z)/ncol(Z)

heatmap(G,cexCol=0.5,cexRow=0.5)
```

Distancias

Una vez que se tienen los marcadores, es posible calcular diferentes tipos de distancias, por ejemplo Euclideana, Rogers, etc.

En el caso de distancias Euclideanas para dos individuos cualesquiera,i y $j = 1, ..., 93$. con $x_{ij} \in \{0, 1\}$,

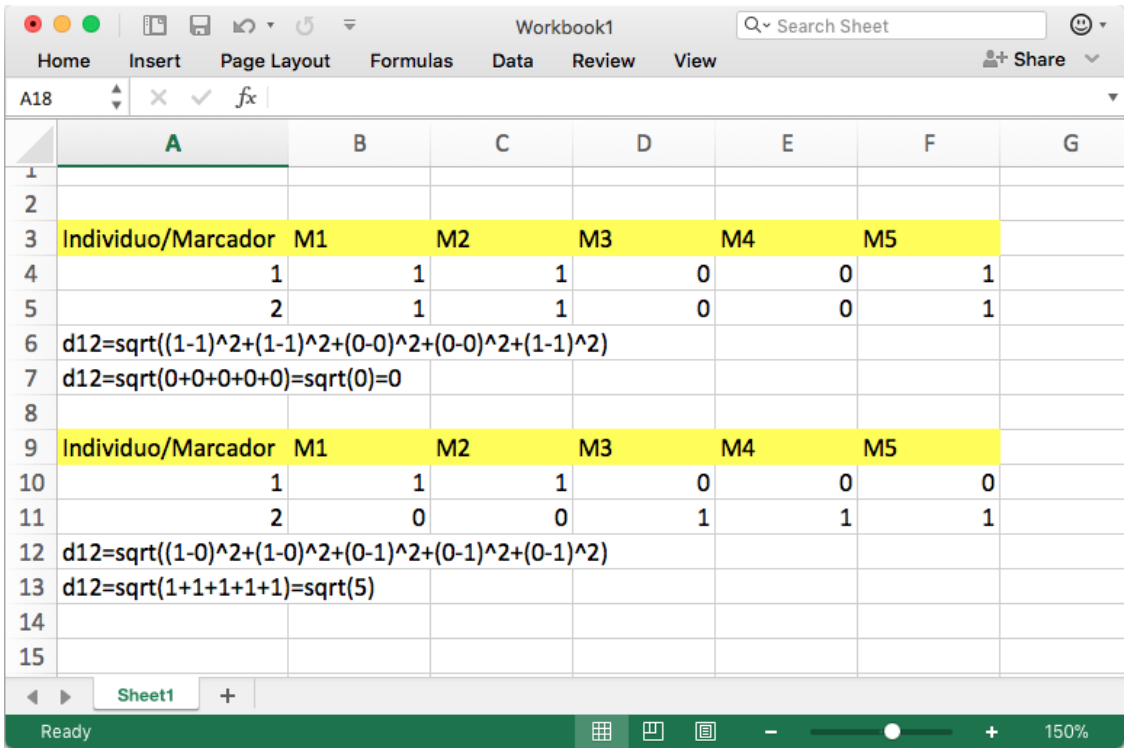
$$d_{ij} = \sqrt{||\boldsymbol{x}_i - \boldsymbol{x}_j||^2} = \sqrt{\sum_{k=1}^{p}(x_{ik} - x_{jk})^2}$$

para Rogers,

$$d_{ij} = \frac{1}{p}\sqrt{\sum_{k=1}^{p}(x_{ik} - x_{jk})^2/2}$$

Notas para distancia Euclideana

- El valor mínimo de d_{ij} es 0.
- El valor máximo de d_{ij} es $\sqrt{p}$, donde p es el número de marcadores.



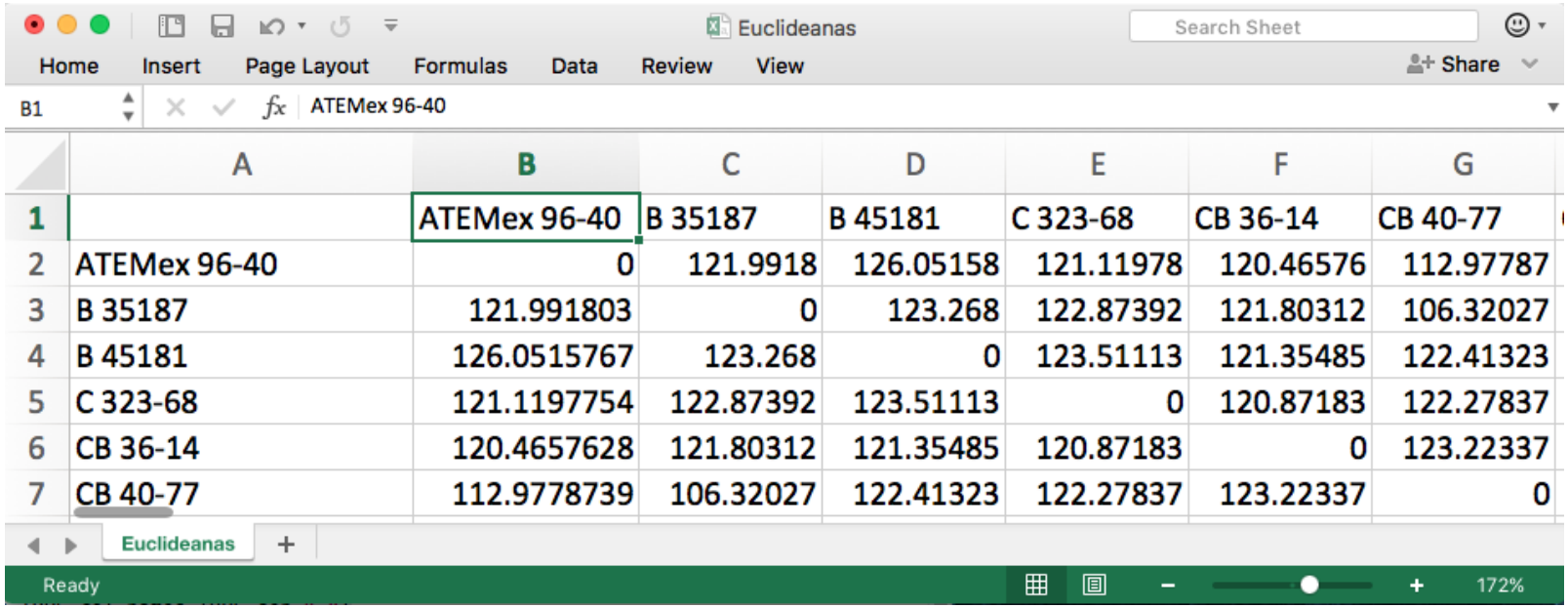
	A	B	C	D	E	F	G
3	Individuo/Marcador	M1	M2	M3	M4	M5	
4	1	1	1	0	0	1	
5	2	1	1	0	0	1	
6	d12=sqrt((1-1)^2+(1-1)^2+(0-0)^2+(0-0)^2+(1-1)^2)						
7	d12=sqrt(0+0+0+0+0)=sqrt(0)=0						
9	Individuo/Marcador	M1	M2	M3	M4	M5	
10	1	1	1	0	0	0	
11	2	0	0	1	1	1	
12	d12=sqrt((1-0)^2+(1-0)^2+(0-1)^2+(0-1)^2+(0-1)^2)						
13	d12=sqrt(1+1+1+1+1)=sqrt(5)						

Figura 9: Ejemplo de valores máximos y mínimos de d_{ij}.

Continuar...

En R:

```
D=dist(X,method = "euclidean")
D=as.matrix(D)
write.csv(D,file="Euclideanas.csv")
```



Euclideanas

Search Sheet

Home Insert Page Layout Formulas Data Review View Share

B1 ATEMex 96-40

	A	B	C	D	E	F	G
1		ATEMex 96-40	B 35187	B 45181	C 323-68	CB 36-14	CB 40-77
2	ATEMex 96-40	0	121.9918	126.05158	121.11978	120.46576	112.97787
3	B 35187	121.991803	0	123.268	122.87392	121.80312	106.32027
4	B 45181	126.0515767	123.268	0	123.51113	121.35485	122.41323
5	C 323-68	121.1197754	122.87392	123.51113	0	120.87183	122.27837
6	CB 36-14	120.4657628	121.80312	121.35485	120.87183	0	123.22337
7	CB 40-77	112.9778739	106.32027	122.41323	122.27837	123.22337	0

Euclideanas

Ready 172%

Figura 10: Distancias Euclideanas para 6 variedades.

Agrupamiento jerárquico

La idea es combinar las observaciones (es decir, la filas de X) en grupos relativamente homogéneos llamados conglomerados (clusters). Las observaciones del mismo grupo estarán, en algún sentido, “cerca”.

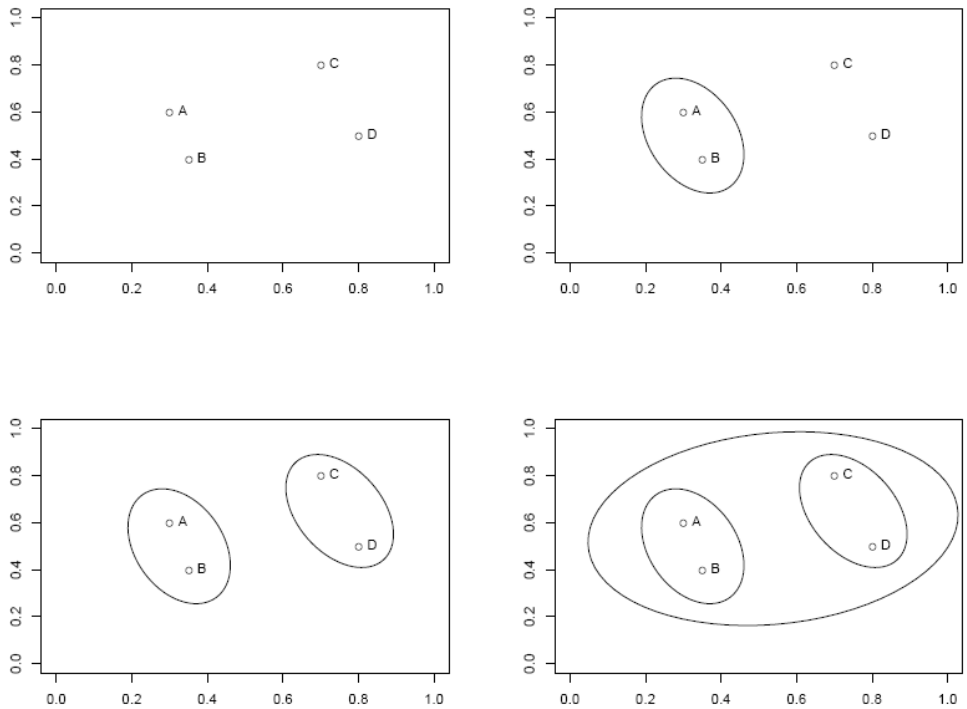


Figura 11: Ejemplo de cluster jerárquico.

Continuar...

En R:

```
D=dist(X)
out_hclust=hclust(D)
plot(out_hclust)
```

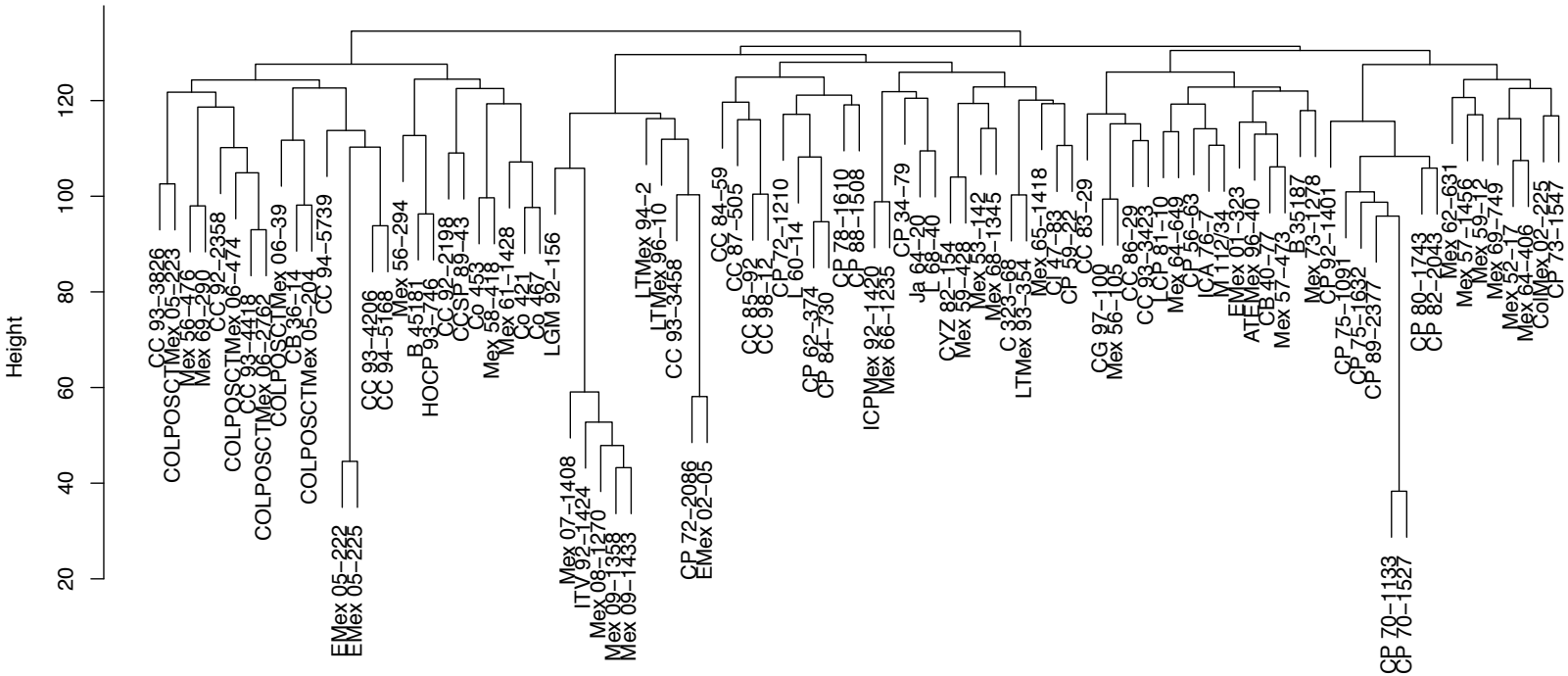


Figura 12: Dendograma basado en distancias Euclideanas con encadenamiento completo.

Contenido

Modelos de predicción basados en marcadores

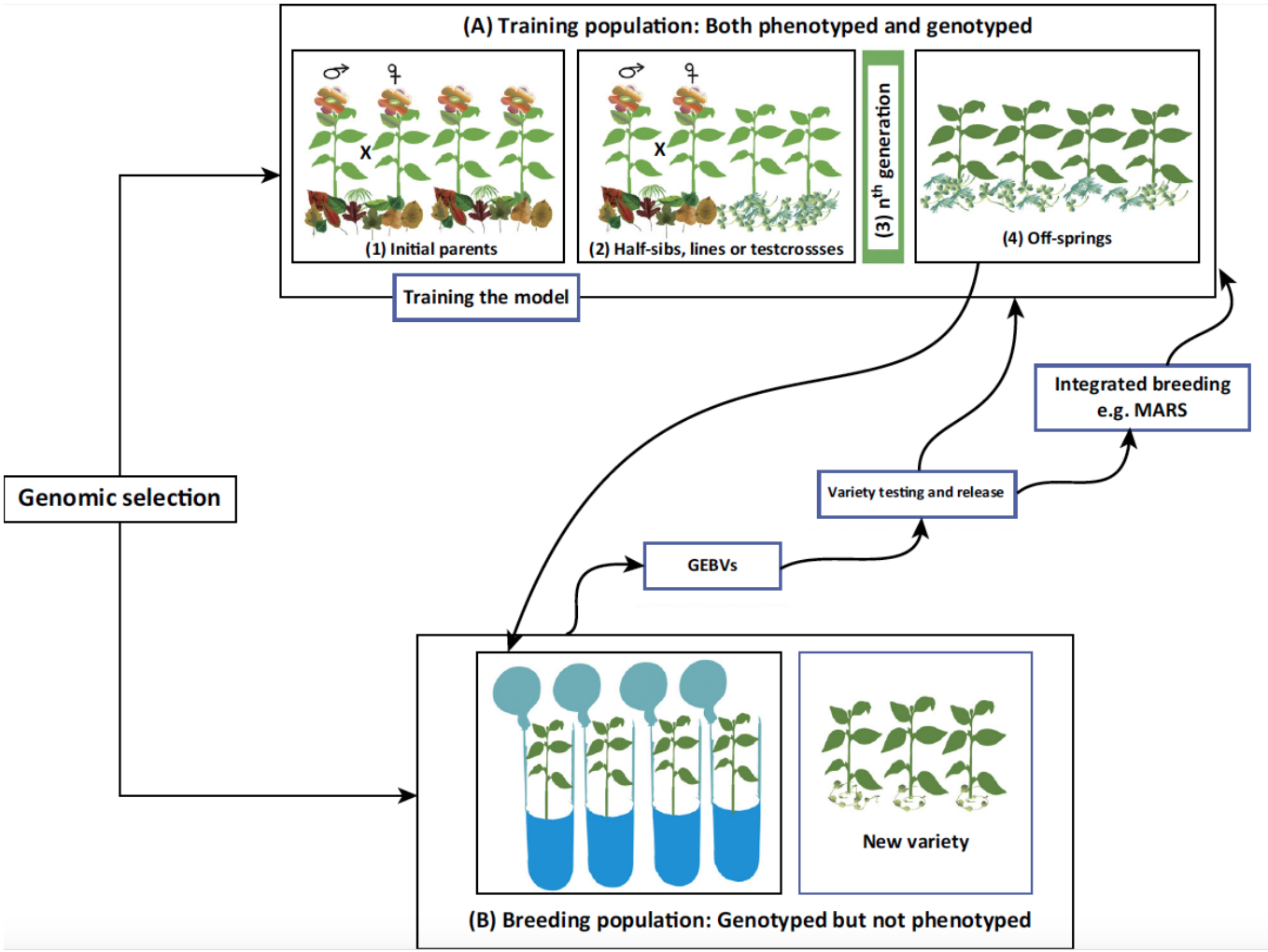
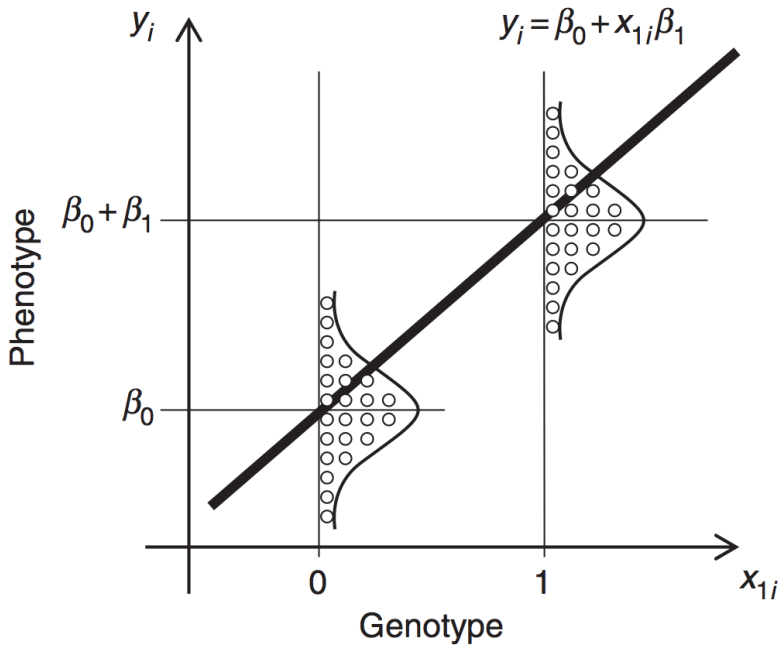


Figura 13: Esquema de selección genómica (Desta and Ortiz, 2014).

Continuar...

En Selección genómica, los valores genéticos son aproximados utilizando regresión lineal (Meuwissen et al., 2001), esto es:

$$y_i = g_i + e_i = \mu + \sum_{j=1}^{p} x_{ij}\beta_j + e_i \quad (1)$$



Relación entre marcadores (x_{1i}: 0 and 1) and fenotipos (y_i) para un conjunto de individuos (círculos sin relleno) en una población de entrenamiento (Nakaya e Isobe, 2012).

Continuar...

Los valores de cría (GEBVs) se obtienen como una combinación lineal de los marcadores con los pesos dados por los efectos de los mismos, es decir:

$$GEBV_i = \sum_{j=1}^{p} x_{ij}\hat{\beta}_j \qquad (2)$$

GBLUP-RR

El modelo más simple usado en selección genómica:

$$y_i = g_i + e_i = \mu + \sum_{j=1}^{p} x_{ij}\beta_j + e_i$$

Los efectos de marcadores pueden obtenerse resolviendo el siguiente problema de optimización,

$$min\ \beta, \lambda \left\{ (y - \sum X_j\beta_j)'(y - \sum X_j\beta_j) + \lambda \sum \beta_j^2 \right\}, \qquad (3)$$

donde $\lambda > 0$ es un parámetro de regularización.

Continuar...

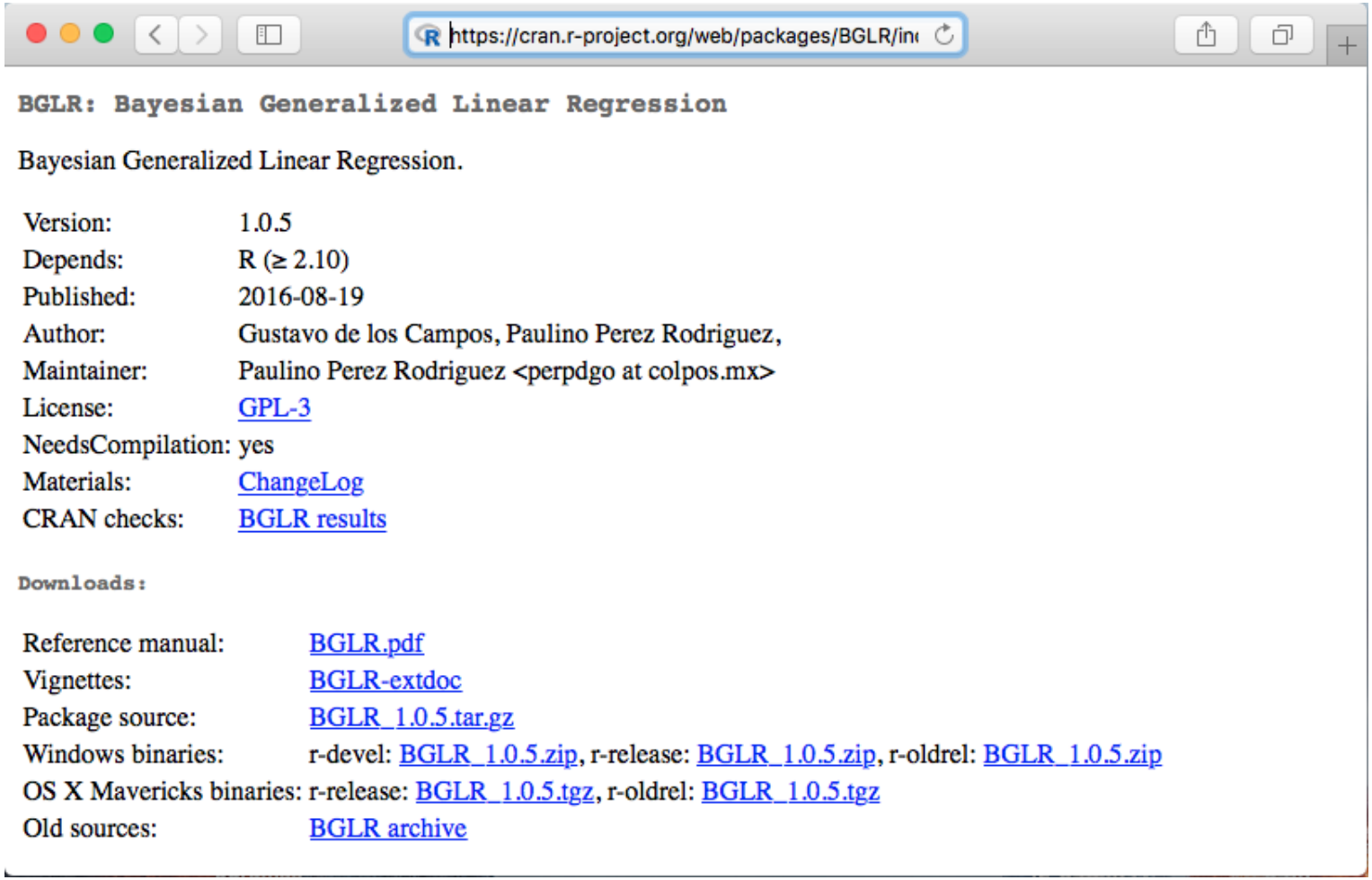
Theoretical and Applied Genetics
October 2013, Volume 126, Issue 10, pp 2575–2586

Experimental assessment of the accuracy of genomic selection in sugarcane

Authors | Authors and affiliations

M. Gouy, Y. Rousselle, D. Bastianelli, P. Lecomte, L. Bonnal, D. Roques, J.-C. Efile, S. Rocher, J. Daugrois, L. Toubi, S. Nabeneza, C. Hervouet, H. Telismart, M. Denis, A. Thong-Chane, J. C. Glaszmann, J.-Y Hoarau, S. Nibouche, L. Costet ✉, show less

Continue...



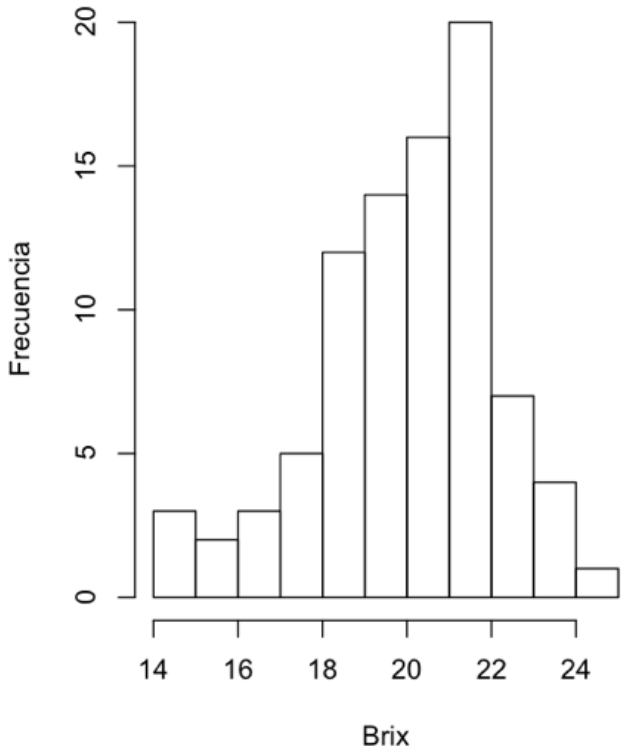
BGLR

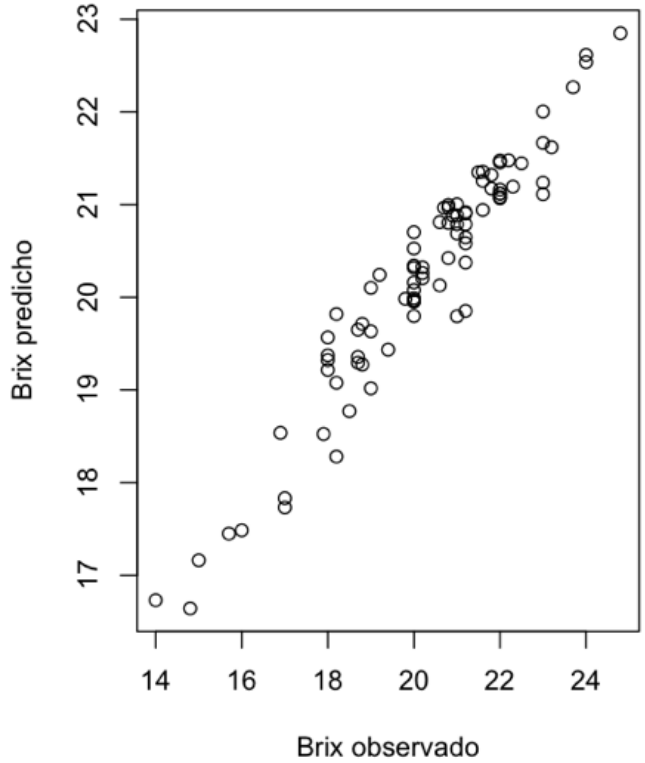
- A novel software for whole genomic regression an prediction for continuous, discrete traits, censored and uncensored.
- Suitable for big p and small n problems.
- Many non-parametric and parametric models implemented in a consistent manner.
- Large collection of Bayesian models included:
  - Bayesian ridge regression.
  - Bayesian LASSO.
  - BayesA, BayesB, BayesC-π.
  - Reproducing Kernel Hilbert Spaces.
  - Reproducing Kernel Hilbert Spaces with Kernel-Averaging.

En R


```
#Clean workspace
rm(list=ls())

library(BGLR)
#Load data
load("data/Pheno_Geno.RData")

#Compute the Genomic Relationship matrix
Z=scale(X,center=TRUE,scale=TRUE)
G=tcrossprod(X)/ncol(Z)

#Exploratory data analysis
hist(Pheno$Brix)

#Regression model
ETA0=list(list(X=Z,model="BRR"))
fm0=BGLR(y=Pheno$Brix,ETA=ETA0,nIter=10000)
```

Predicción de grados Brix





Heredabilidad genómica

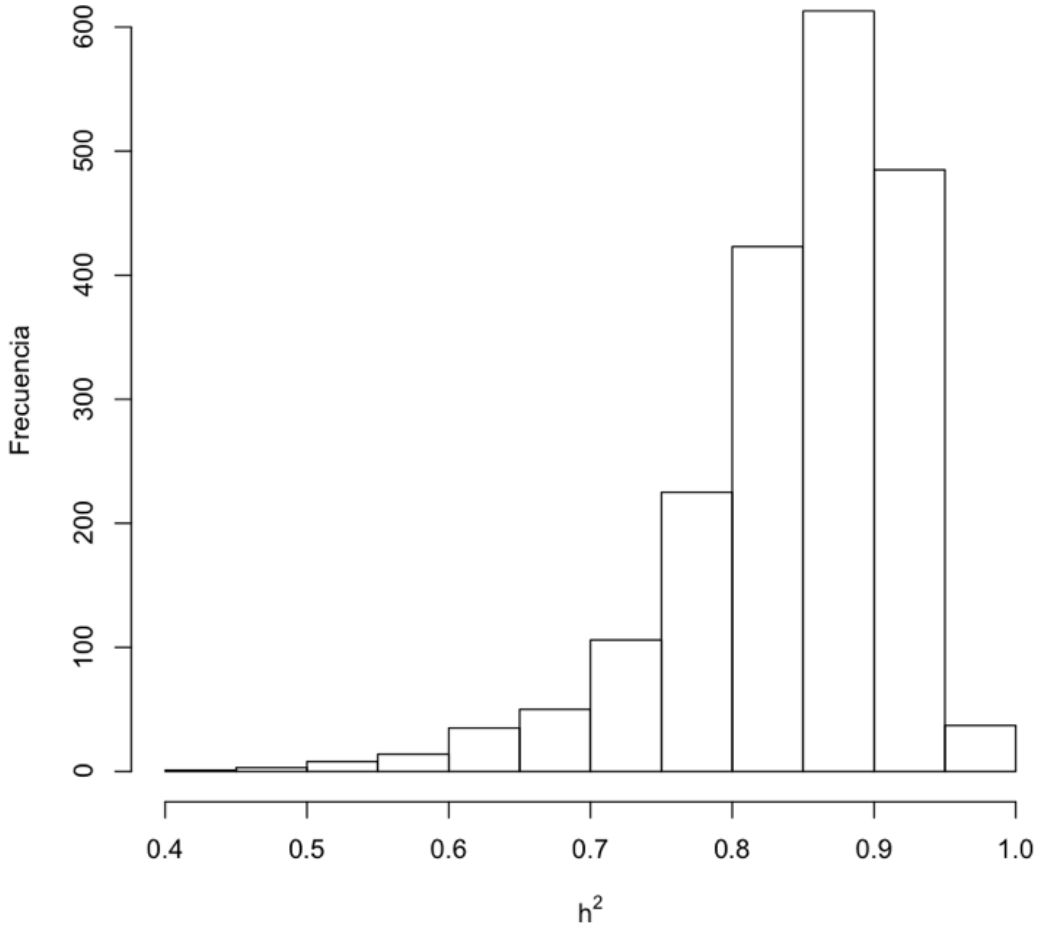


Figura 14: Distribución de $h^2 = \frac{\sigma_a^2}{\sigma_a^2+\sigma_e^2}$

Predicción valores fenotípicos de cruzas simples

- Predecir valores fenotípicos de cruzas usando información genotípica de padres.
- Datos:
  - Información fenotípica de algunas cruzas.
  - Información genotípica de padres y madres.
- El modelo debe predecir cruzas que no se han hecho y no se han probado en campo.

Lo que se hace en trigo y maíz...

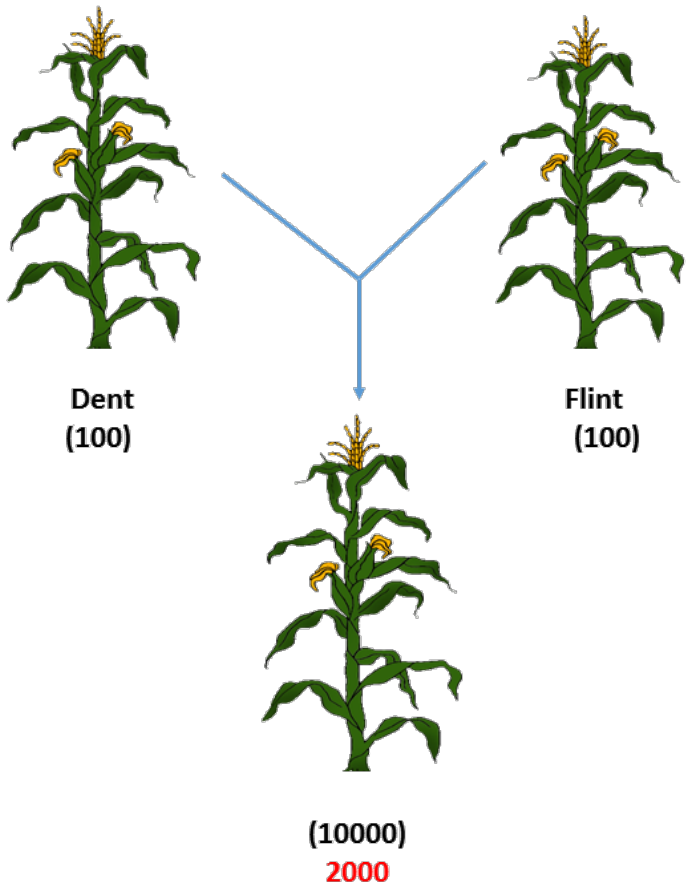


Figura 15: Cruza simple

Modelos

HIGHLIGHTED ARTICLE
GENOMIC SELECTION

Genome Properties and Prospects of Genomic Prediction of Hybrid Performance in a Breeding Program of Maize

Frank Technow,* Tobias A. Schrag,* Wolfgang Schipprack,* Eva Bauer,† Henner Simianer,‡ and Albrecht E. Melchinger*,1

*Institute of Plant Breeding, Seed Sciences, and Population Genetics, University of Hohenheim, 70599 Stuttgart, Germany, †Plant Breeding, Technische Universität München, 85354 Freising, Germany, and ‡Department of Animal Sciences, Georg-August-University Goettingen, 37075 Goettingen, Germany

Continuar...

$$\mathbf{y} = \mathbf{Z}_E\beta_E + \mathbf{Z}_D\mathbf{g}_D + \mathbf{Z}_F\mathbf{g}_F + \mathbf{Z}_H\mathbf{h} + \mathbf{e}$$

- $\mathbf{Z}_E$ matriz de incidencia para ambientes, β_E el efecto de ambientes.
- $\mathbf{Z}_D$ matriz de incidencia para machos.
- $\mathbf{g}_D$ vector de efectos aleatorios para la aptitud combinatoria general de machos, $\mathbf{g}_D \sim N(\mathbf{0}, \sigma^2_D\mathbf{G}_D)$.
- $\mathbf{Z}_F$ matriz de incidencias para hembras.
- $\mathbf{g}_F$ vector de efectos aleatorios para aptitud combinatoria general para hembras, $\mathbf{g}_F \sim N(\mathbf{0}, \sigma^2_F\mathbf{G}_F)$.

Continue...

- $\boldsymbol{Z}_H$ matriz de incidencias para híbridos.
- $\boldsymbol{h}_F$ vector de efectos aleatorios asociada a aptitud combinatoria específica de las cruzas, $\mathbf{h} \sim N(\mathbf{0}, \sigma^2_H \mathbf{H})$.
- $\boldsymbol{H} = \boldsymbol{G}_D \otimes \boldsymbol{G}_H$.

Otros análisis

- GWAS (Genome Wide Association Studies).
- Resistencia a enfermedades.

Continuar...

Hindawi Publishing Corporation
The Scientific World Journal
Volume 2013, Article ID 613062, 11 pages
http://dx.doi.org/10.1155/2013/613062

Hindawi

Research Article

Genetic Diversity Analysis of Sugarcane Parents in Chinese Breeding Programmes Using gSSR Markers

Qian You, Liping Xu, Yifeng Zheng, and Youxiong Que

Key Laboratory of Sugarcane Biology and Genetic Breeding, Ministry of Agriculture, Fujian Agriculture and Forestry University, Fuzhou 350002, China

Continuar...

Theoretical and Applied Genetics
August 2014, Volume 127, Issue 8, pp 1719–1732

Prospecting sugarcane resistance to Sugarcane yellow leaf virus by genome-wide association

Authors | Authors and affiliations

S. Debibakas, S. Rocher, O. Garsmeur, L. Toubi, D. Roques, A. D'Hont, J.-Y. Hoarau ✉, J. H. Daugrois

Open Access | Original Paper

First Online: 12 June 2014

DOI: 10.1007/s00122-014-2334-7

Cite this article as:
Debibakas, S., Rocher, S., Garsmeur, O. et al. Theor Appl Genet (2014) 127: 1719. doi:10.1007/s00122-014-2334-7

Modelo de umbrales para resistencia...

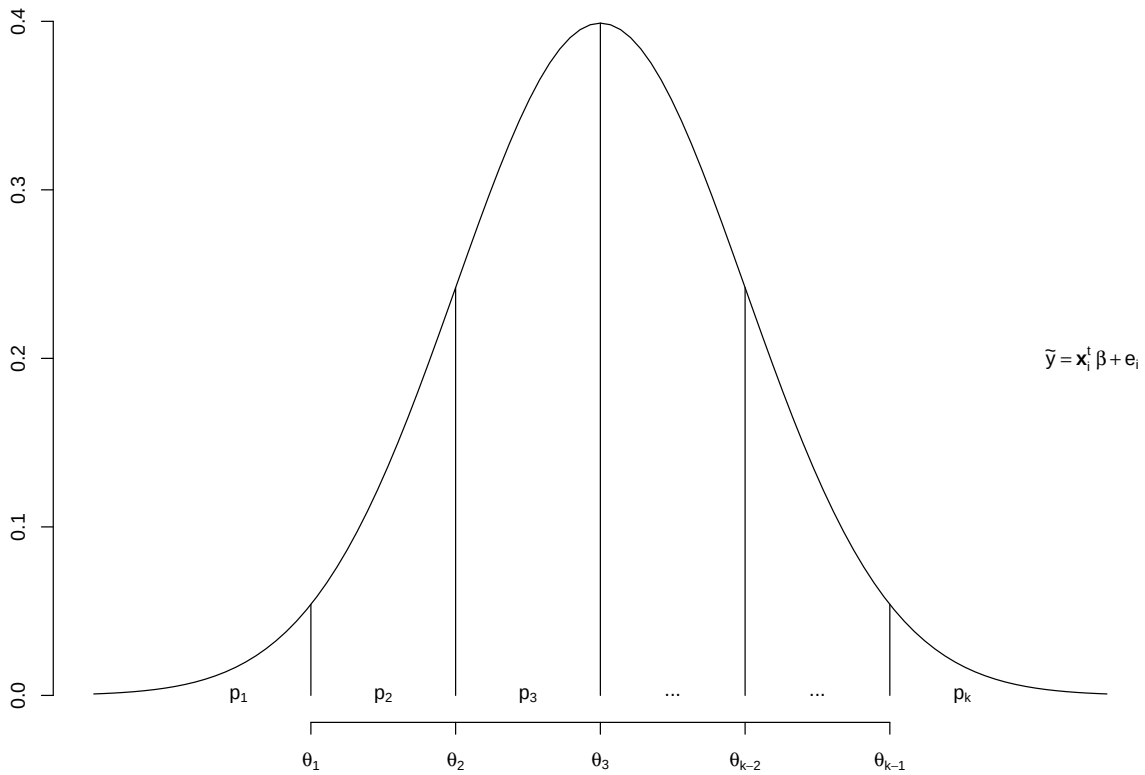
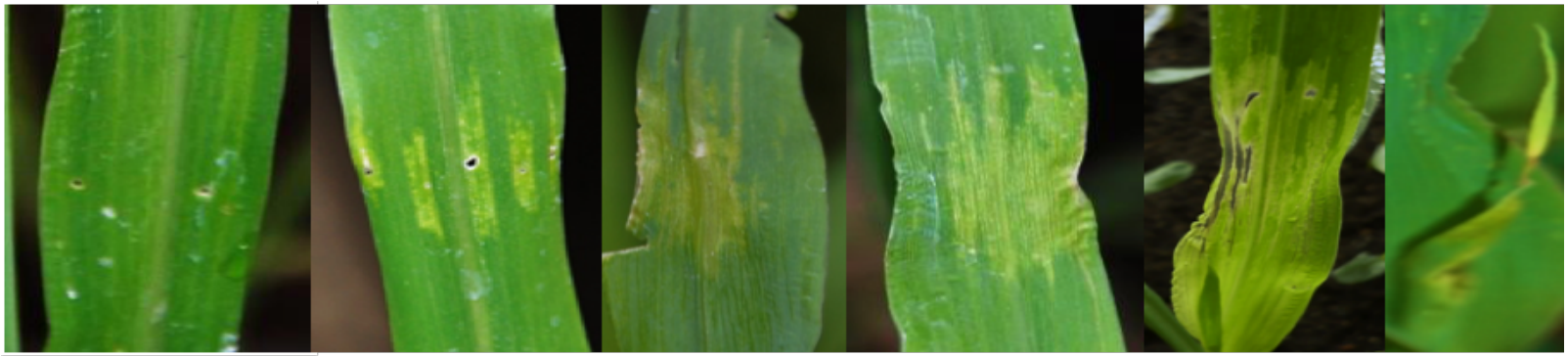


Figura 16: Función de densidad de la variable latente

Continuar...



0 1 2 3 4 5

Defina:

$$Y_i = \begin{cases} 0 & \tilde{Y}_i < \theta_1 \\ 1 & \tilde{\theta}_1 < \tilde{Y}_i < \theta_2 \\ \vdots & \\ k & \tilde{Y}_i > \theta_{k-1} \end{cases}$$

Continuar...

GENOMIC SELECTION

Threshold Models for Genome-Enabled Prediction of Ordinal Categorical Traits in Plant Breeding

Osval A. Montesinos-López,* Abelardo Montesinos-López,† Paulino Pérez-Rodríguez,‡ Gustavo de los Campos,§ Kent Eskridge, and José Crossa††,1**

*Facultad de Telemática, Universidad de Colima, Avenida Universidad 333, Colima, Colima, 28040, México, †Departamento de Estadística, Centro de Investigación en Matemáticas (CIMAT), Guanajuato, Guanajuato, 36240, México, ‡Colegio de Postgraduados, Edo. de México, 56230, México, §Department of Biostatistics, University of Alabama at Birmingham, Birmingham, Alabama 35294, **Department of Statistics, University of Nebraska, Lincoln, Nebraska 68583-0963, and ††Biometrics and Statistics Unit, International Maize and Wheat Improvement Center (CIMMYT), México, D.F., 06600, México

Referencias

Endelman, J. B., 2011.
Ridge Regression and Other Kernels for Genomic Selection with R package rrBLUP,
The Plant Genome 4(3): 250-255.