Cecilia Bruno
Eugenia Videla
Andrea Peña
Mónica Balzarini

Guía para la construcción de modelos de asociación genómica

DISEMINACIÓN CIENTÍFICA Y TRASNFERENCIA DE RESULTADOS DE INVESTIGACIÓN, PROMOVIDAS POR EL MINISTERIO DE CIENCIA Y TECNOLOGÍA DE LA PROVINCIA DE CÓRDOBA

La cita bibliográfica para el presente documento

Bruno C, Videla E, Peña A, Balzarini M. 2019. Guía para la construcción de modelos de asociación genómica. Serie Estadística Aplicada. Com. Balzarini M. Brujas. Córdoba, Argentina.

Guía para la construcción de modelos de asociación genómica / Cecilia Bruno ... [et al.]. - 1a ed. - Córdoba : Brujas, 2019.
136 p. ; 23 x 15 cm.

ISBN 978-987-760-271-5

1. Estadísticas. 2. Análisis Estadístico. 3. Agronomía. I. Bruno, Cecilia
CDD 310

1 ° Edición
Primera Impresión
Impreso en Argentina
ISBN 978-987-760-271-5

Queda hecho el depósito que prevé la ley 11.723

A los exploradores del genoma

Índice general

Introducción

El mapeo asociativo (MA) o el estudio de asociación a través del genoma completo (GWAS por sus siglas en inglés, *Genome Wide Association Study*) es usado para identificar sitios específicos del genoma, marcados por un marcador molecular (MM), que deben interpretarse como asociados con la variación de un carácter fenotípico continuo. Este tipo de análisis de asociación genotipo-fenotipo puede ser implementado sobre colecciones o paneles de individuos o líneas diversas, *i.e.* con alta variabilidad genética. No es necesario generar poblaciones a partir de cruzamientos experimentales, pero sí conformar una población de mapeo (usualmente más de 120 individuos) de varios individuos genéticamente diferentes. Suelen usarse cientos o miles de marcadores moleculares para expresar la variabilidad genómica siendo la cantidad de MM generalmente mayor a la cantidad de líneas que conforman la población de mapeo. Las variantes genéticas identificadas mediante GWAS pueden explicar distinta proporción de la variación fenotípica y estar ligadas a uno o mas loci que determinan el carácter cuantitativo (QTL por sus siglas en inglés, *Quantitative Trait Loci*). El MA ofrece una oportunidad para identificar polimorfismos asociados con fenotipos de interés y para comprender la base genética de la variación cuantitativa de los caracteres fenotípicos. La alta dimensionalidad de la matriz de MM y las posibles estructuras o patrones en esta matriz, indicando estructura genética o falta de independencia entre los individuos de la población de mapeo, imprimen desafíos estadísticos para la

evaluación de la significancia de las asociaciones marcador-fenotipo. Los modelos estadísticos más usados en MA son los modelos lineales mixtos (MLM) (West *et al.*, 2014), desarrollados para estimar el efecto de cada marcador sobre el fenotipo. Sin embargo, otros efectos, principalmente aquellos relacionados a la estructura genética poblacional (EGP), deben ser considerados también en el modelo estadístico de MA. Si el fenotipado se ha realizado en varios ambientes, entonces los efectos de ambiente y de interacción entre genotipo y ambiente, también serán incluidos en los modelos que evalúan la asociación. Numerosos métodos estadísticos han sido desarrollados para contemplar la EGP en el modelo de MA (Malavera *et al.*, 2018). Estos tienen como objetivo controlar los errores que podrían provenir de posibles asociaciones espurias o infladas al usar un modelo de asociación que supone independencia cuando existe estructura poblacional. Se supone que la modelación estadística que incorpora información sobre correlaciones entre los datos hace más eficiente el MA.

En este texto se describen modelos de asociación que permiten realizar GWAS en paneles de líneas diversas, aún cuando éstas se encuentran estructuradas genéticamente. Para evitar falsos descubrimientos de asociaciones se trabaja primero identificando la estructura genética subyacente en la población de mapeo y luego incorporando la información de correlación entre líneas en los modelos. Adicionalmente, se controla la inflación de la tasa de falsos positivos debida a la inferencia simultánea o multiplicidad de prueba estadísticas que deben realizarse en estudios de asociación con muchos MM. Para facilitar el ajuste de estos modelos se describe el nuevo menú de MA embebido en el software para análisis de datos genéticos *Info-Gen* (Balzarini y Di Rienzo, 2018) y códigos para implementar cada uno de los procedimientos estadísticos descriptos, tanto para el análisis de EGP como para MA, usando el

software *R* (www.R-org.com). En la última parte de este documento se ilustra cómo trabajar directamente en *Info-Gen* y cómo ejecutar scripts de *R* desde el intérprete de *R* disponible en *Info-Gen*.

1

Análisis de estructura genética poblacional

1.1 Métodos de agrupamiento

Los datos de MM permiten evaluar la similitud u obtener distancias entre entidades biológicas para realizar estudios de variabilidad genética e identificar estructuras o subgrupos dentro de una población. Las investigaciones sobre tipo, abundancia y distribución de los organismos necesitan identificar la estructura subyacente en los datos, es decir el agrupamiento o conglomeración de las entidades en grupos (o *clusters*) relativamente homogéneos. Para identificar la EGP es común el uso de métodos de clasificación, tanto no supervisada (sin conocimiento *a priori* del análisis de los agrupamientos subyacentes) como supervisada (con conocimiento *a priori* de la existencia de agrupamientos de los datos). El objetivo del análisis de conglomerados, o de clasificación no supervisado, es formar grupos tal que los elementos de un grupo sean más parecidos entre sí que con los elementos de otro grupo.

Referidos a poblaciones de MA, se utiliza una matriz de datos n×m (entidades×marcadores), se calcula primero una matriz (n×n) que contiene las distancias entre todos los pares de casos y luego sobre esa matriz se aplica el procedimiento de agrupamiento que se haya seleccionado. Diferentes métodos multivariados para identificar grupos sur-

gen como alternativa para analizar EGP. Por ejemplo, los conglomerados jerárquicos son frecuentemente utilizados dado que se encuentran disponibles en gran cantidad de software y pueden ser aplicados directamente sobre datos moleculares seleccionando una métrica de distancia apropiada (Bruno y Balzarini 2010; Odong *et al.*, 2011). Los algoritmos no jerárquicos, *K-means*, también son utilizados para detectar EGP (Lee *et al.*, 2009). Pritchard *et al.* (2000) propusieron un método de agrupamiento Bayesiano que se encuentra implementado en el software *Structure* y en el paquete *LEA* del software *R* (Frichot y François, 2015) y que es ampliamente usado para estudiar EGP. Otras técnicas de agrupamiento basadas en redes neuronales, tales como los mapas auto-organizativos (SOM, de sus siglas en Inglés *Self Organizing Maps*), también han sido utilizadas sobre datos genéticos (Fernandez y Balzarini, 2007; Roux *et al.*, 2007; Nikolic *et al.*, 2009).

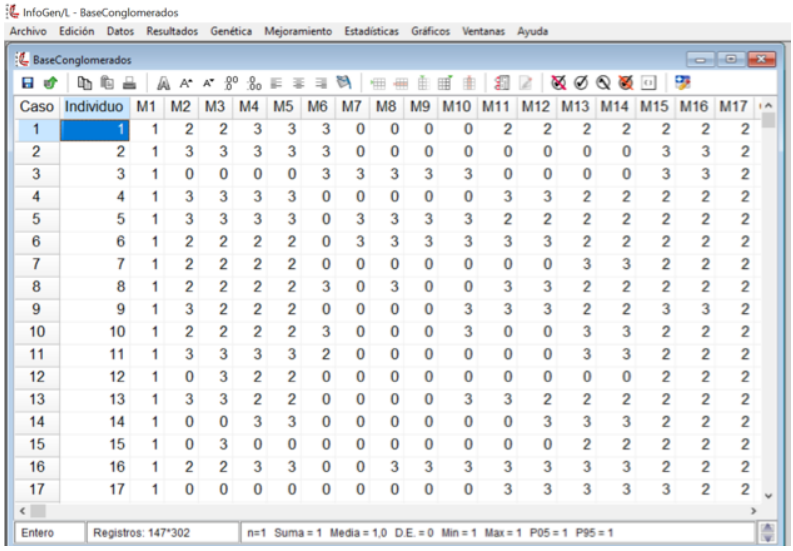
1.1.1 Conglomerados jerárquicos

Los análisis de conglomerados jerárquicos agrupan individuos o variables a partir de una matriz de distancias entre cada elemento y los restantes. Si el método es jerárquico aglomerativo, se comienza el análisis con tantos grupos como individuos y en sucesivos pasos se van uniendo hasta que al final del proceso todos los individuos están en un único conglomerado.

Ejemplo 1.1. Clasificación. Análisis de Conglomerados Jerárquico. UPGMA. Archivo Base Conglomerados

En esta sección se ejmplificará el uso del método de agrupamiento jerárquico UPGMA sobre la matriz de distancias conformada a partir la métrica de Jaccard (Jaccard, 1901) sobre una base de datos que posee 147 individuos y 300 MM. Los MM son co-dominantes, los genotipos AA fue-

ron codificados con 0, los BB con 2 y los AB o BA con 3 (Figura 1.1).



Caso	Individuo	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13	M14	M15	M16	M17
1	1	1	2	2	3	3	3	0	0	0	0	2	2	2	2	2	2	2
2	2	1	3	3	3	3	3	0	0	0	0	0	0	0	0	3	3	2
3	3	1	0	0	0	0	3	3	3	3	3	0	0	0	0	3	3	2
4	4	1	3	3	3	3	0	0	0	0	0	3	3	2	2	2	2	2
5	5	1	3	3	3	3	0	3	3	3	3	2	2	2	2	2	2	2
6	6	1	2	2	2	2	0	3	3	3	3	3	3	2	2	2	2	2
7	7	1	2	2	2	2	0	0	0	0	0	0	0	3	3	2	2	2
8	8	1	2	2	2	2	3	0	3	0	0	3	3	2	2	2	2	2
9	9	1	3	2	2	2	0	0	0	0	3	3	3	2	2	3	3	2
10	10	1	2	2	2	2	3	0	0	0	3	0	0	3	3	2	2	2
11	11	1	3	3	3	3	2	0	0	0	0	0	0	3	3	2	2	2
12	12	1	0	3	2	2	0	0	0	0	0	0	0	0	0	2	2	2
13	13	1	3	3	2	2	0	0	0	0	3	3	2	2	2	2	2	2
14	14	1	0	0	3	3	0	0	0	0	0	0	3	3	3	2	2	2
15	15	1	0	3	0	0	0	0	0	0	0	0	0	2	2	2	2	2
16	16	1	2	2	3	3	0	0	3	3	3	3	3	3	3	2	2	2
17	17	1	0	0	0	0	0	0	0	0	0	3	3	3	3	3	2	2

Figura 1.1: Base de datos de 147 individuos y 300 marcadores.

El primer paso es crear la matriz de distancia. En el software *R* se utiliza la función "vegdist" del paquete *vegan*.

```
library(vegan)
D<-vegdist(as.matrix(BaseConglomerados[,-c(1:2)],
                     method="jaccard",
                     binary=TRUE))
```

```
head(D)
```

```
#>         1      2      3      4      5
#> 1 0,000 0,310 0,303 0,267 0,289
#> 2 0,310 0,000 0,285 0,255 0,291
#> 3 0,303 0,285 0,000 0,277 0,285
#> 4 0,267 0,255 0,277 0,000 0,306
#> 5 0,289 0,291 0,285 0,306 0,000
```

A continuación realizamos el agrupamiento utilizando la función "hclust" y graficamos el dendrograma (Figura 1.2). Esta función tiene como primer argumento la matriz de distancia de los individuos a agrupar y en el segundo se indica el método de agrupamiento deseado. Para indicar que el método de agrupamiento a utilizar sea UPGMA debemos completar el argumento *method="average"*.

```
fit <- hclust(D, method="average")
plot(fit, ylab="Distancia",
     labels= BaseConglomerados$X, cex=0.3)
```

Figura 1.2: Dendrograma obtenido de la matriz de distancia de Jaccard para 147 individuos caracterizados moleculamente con 300 marcadores.

Interpretación

El perfil molecular del individuo 88 es el más distante del resto de los individuos. Los perfiles moleculares de los individuos 56 y 57 se unen a menor distancia que del resto de los individuos.

1.1.2 Conglomerados no-jerárquicos

Otros métodos usados para clasificación no supervisada son los métodos no-jerárquicos, como *K-means*, donde la clasificación de entidades depende de la relación entre sumas de cuadrados entre y dentro de un número determinado de grupos, que se hipotetiza *a priori*.

Ejemplo 1.2. Clasificación. Análisis de Conglomerados no-jerárquicos. *K-means*. Archivo Base Conglomerados

Para implementar este método se utiliza la función "kmeans" de *R*. En el primer argumento de la función se indica la base de datos en la que se agruparán los individuos y en el segundo la cantidad de grupos que se desea formar. En el ejemplo se utiliza k=3

```
fit<-kmeans(BaseConglomerados[,-c(1:2)], 3)
```

Con el comando *fit$cluster* podemos ver el vector de asignación de cada individuo a alguno de los tres grupos.

```
fit$cluster
#>   [1] 2 3 3 3 3 3 3 3 2 3 3 3 3 3 3 3 3
#>  [18] 3 3 3 3 3 3 3 3 3 2 3 3 3 2 3 2 3
#>  [35] 1 3 3 3 3 2 2 2 3 3 1 1 1 1 1 1 1
#>  [52] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
#>  [69] 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 3 1
#>  [86] 1 1 1 1 1 2 2 1 1 1 1 2 2 2 2 2 2
#> [103] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
#> [120] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
#> [137] 2 2 2 2 2 2 2 2 2 2 2
```

Una manera de visualizar el agrupamiento realizado por este método es a partir de un Análisis de Coordenadas Principales de la matriz de distancia entre individuos donde se indique el grupo de pertenencia de cada individuo. Los códigos que siguen permiten graficar a los individuos en el espacio de las dos primeras coordenadas principales y colorear a cada uno de ellos según la clasificación que se obtuvo con *K-means* (Figura 1.3).

```
# Calculamos la matriz de distancia
library(vegan)
D<-vegdist(as.matrix(BaseConglomerados
                     [,-c(1:2)]),
           method="jaccard", binary=TRUE)
# Realizamos el análisis de coordenadas
# principales
library(labdsv)
euc.pco <- pco(D,k=2)
```

```
# Graficamos
library(ggplot2)
ggplot(as.data.frame(euc.pco$points),
       aes(V1, V2, shape=
             as.factor(fit$cluster))) +
  geom_point(size =2) +  theme_minimal() +
      labs(x = "CPO1",  y = "CPO2",
           shape="Grupos")
```

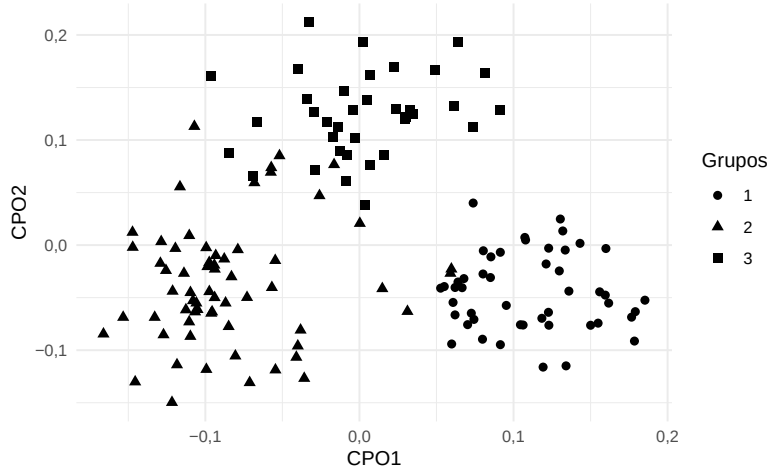


Figura 1.3: Diagrama de dispersión de las dos primeras coordenadas del Análisis de Coordenadas Principales de 147 individuos caracterizados molecularmente por 300 marcadores. Individuos coloreados según agrupamiento realizado por el método K-means.

Interpretación

A partir del método de conglomerado no jerárquico *K-means*, se agrupó la población de 147 de individuos genotipadas por 300 marcadores moleculares en tres subpoblaciones. La Figura 1.3 muestra que los grupos conformados son compactos y separables en el plano de las dos coordenadas principales.

1.1.3 Conglomerados bayesianos

El método de agrupamiento Bayesiano más usado para estudios de EGP conglomera los genotipos asignandolos de manera probabilística a uno de K grupos por similitud molecular o conjuntamente a dos o más conglomerados si el genotipo indica una mezcla de patrones moleculares. El software *Structure* (Pritchard *et al.*, 2000) ha sido ampliamente utilizado para implementar este algoritmo y producir una visualización de la estructura genética poblacional subyacente mediante una gráfica de barras, en donde cada individuo en el conjunto de datos está representado por una línea vertical, que está dividida en K segmentos coloreados según la probabilidad de pertenencia estimada de ese individuo en cada uno de los K clústeres inferidos. Es un tipo de agrupamiento difuso donde las probabilidades *a posteriori* indican la certidumbre de la asignación del genotipo al clúster.

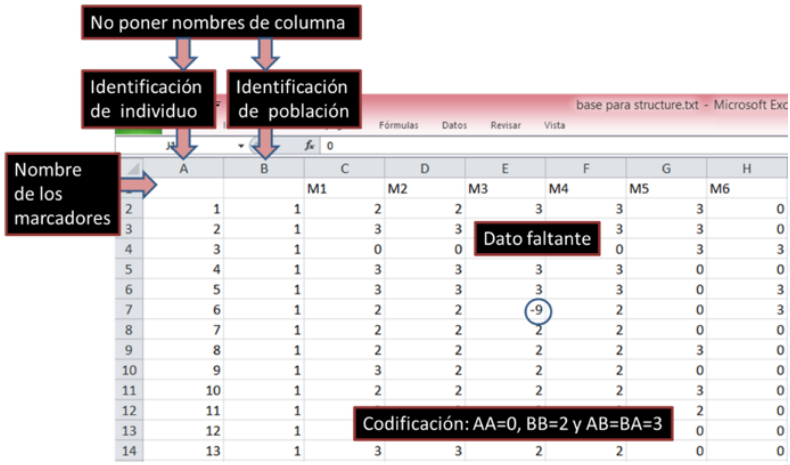
Ejemplo 1.3. Clasificación. Análisis de Conglomerados Bayesianos. *Structure*. Archivo Base Conglomerados

Formato de los datos para *Structure*:

- Matriz con los individuos en filas y los loci en columnas.
- Los individuos codificados con números enteros.
- Los valores faltantes deben ser indicados con un número diferente de manera tal que no sea confundido con la información que identifica el genotipo, por ejemplo con -1 o -9.
- El archivo de datos debe ser un archivo de texto con extensión *.txt*

El archivo de ejemplo posee 147 líneas, que representan los 147 individuos, con 302 columnas que corresponden a las columnas de los 300 marcadores más dos columnas

indicadoras: una columna con el nombre del individuo y otra columna con la población a la que pertenece dicho individuo (Figura 1.4).



	A	B	C	D	E	F	G	H
			M1	M2	M3	M4	M5	M6
2	1	1	2	2	3	3	3	0
3	2	1	3	3		3	3	0
4	3	1	0	0		0	3	3
5	4	1	3	3	3	3	0	0
6	5	1	3	3	3	3	0	3
7	6	1	2	2	-9	2	0	3
8	7	1	2	2	2	2	0	0
9	8	1	2	2	2	2	3	0
10	9	1	3	2	2	2	0	0
11	10	1	2	2	2	2	3	0
12	11	1					2	0
13	12	1					0	0
14	13	1	3	3	2	2	0	0

Figura 1.4: Formato del archivo de datos para análisis de conglomerados en Structure.

Para realizar un análisis de EGP en *Structure*, es necesario crear un nuevo proyecto . Para ello ir al menú FILE comando **New Proyect**. Una vez que se crea un nuevo proyecto y se selecciona el directorio donde se encuentra la base de datos (Figura 1.5), se debe indicar el número de individuos, ploidía, número de loci y dígito con el que se identifica a los valores faltantes (Figura 1.6). En el archivo de ejemplo, cada fila representa un individuo, por lo tanto, la ploidía es 1. Para colocar ploidía 2, el formato del archivo de datos debe tener dos filas por individuo, por ejemplo, en el caso de marcadores del tipo SSR. La casilla de **Missing data value** no debe dejarse sin completar, es necesario colocar un valor diferente a la codificación del genotipo haya o no haya datos faltantes en el archivo.

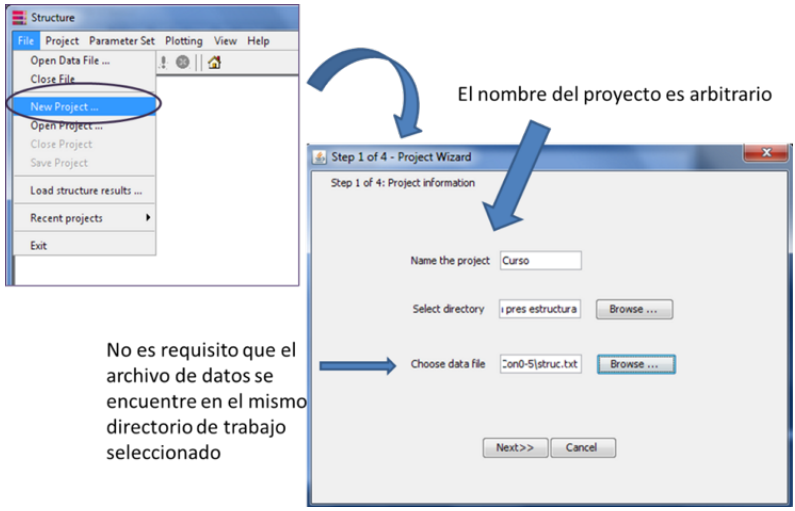


Figura 1.5: Menu FILE de Structure. Comando New Project. Ventana "Proyect Wizard" Paso 1.

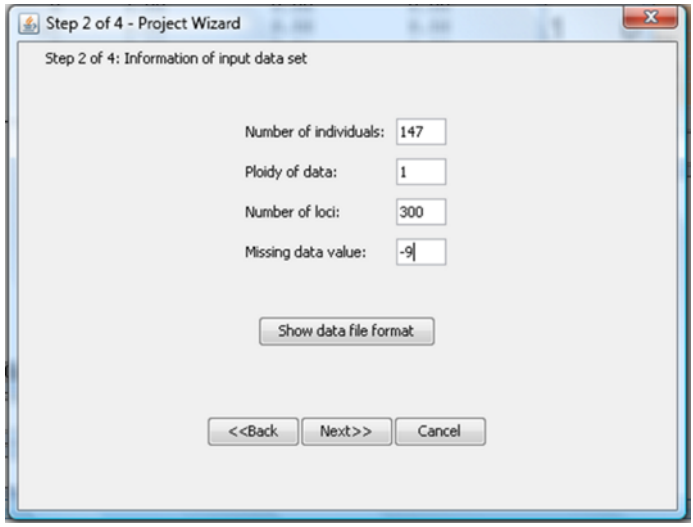


Figura 1.6: Menu FILE de Structure. Comando New Project. Ventana "Proyect Wizard" Paso 2.

En el paso 3, se puede tildar la opción **Row of marker names** si el archivo posee la primer fila con los nombres de los marcadores, como es el caso del ejemplo. En el archivo

de ejemplo el perfil molecular de cada individuos representa una fila del archivo de datos de ejemplo, por lo que también debe tildarse la opción **Data file stores data for individuals in a single line** (Figura 1.7).

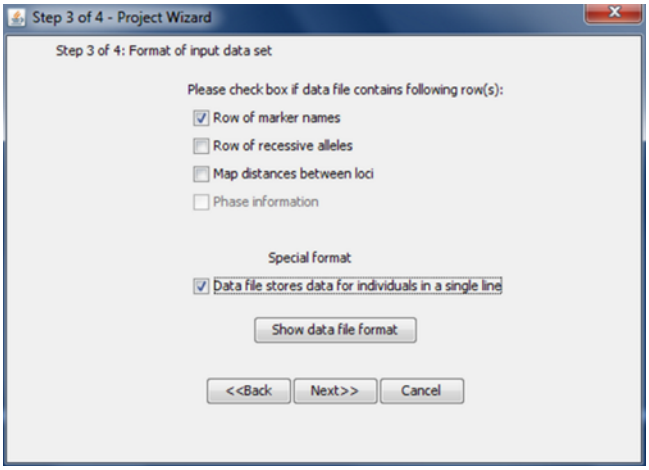


Figura 1.7: Menu FILE de Structure. Comando New Project. Ventana "Proyect Wizard" Paso 3.

En el último paso, indicamos si la primera columna del archivo posee el nombre de cada individuo, y la segunda columna la población de origen. Estas dos columnas no son obligatorias para el análisis de búsqueda de estructura poblacional (Figura 1.8).

Si se desean poner columnas con informacional adicional, que no esté contemplada las casillas, se debe indicar que existen con la opción **Other extra columns** y colocar la cantidad en **Number of extra Columns**.

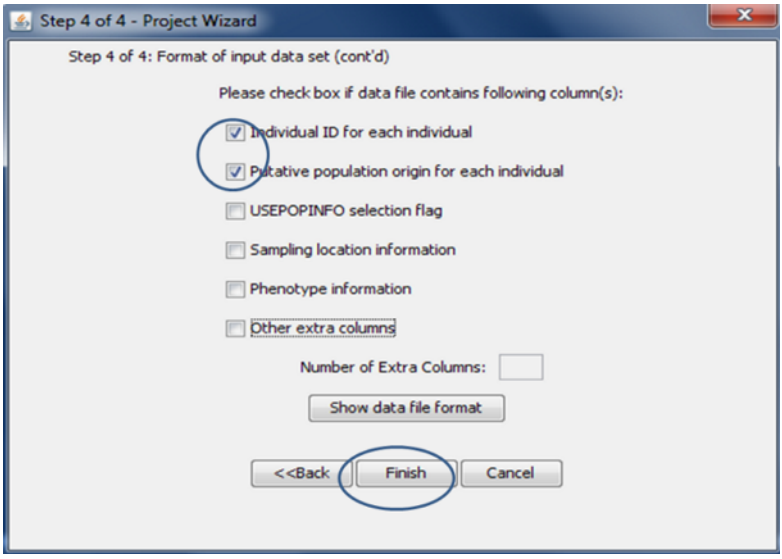


Figura 1.8: Menu FILE de Structure. Comando New Project. Ventana "Proyect Wizard" Paso 3.

Los datos generados en el proyecto conforman la base de datos con la que trabajará *Structure*, la misma se visualiza en el panel superior derecho.

A continuación, se debe crear un nuevo set de parámetros. Para ello, ir a Menú PARAMETER SET comando **New**. En la solapa **Run Length** se debe indicar la Duración del Período de Generación de parámetros, es decir la cantidad de ciclos y el número de MCMC después del periodo de generación, es decir la cantidad de simulaciones requeridas que serán usada para la estimación de los parámetros (Figura 1.9).

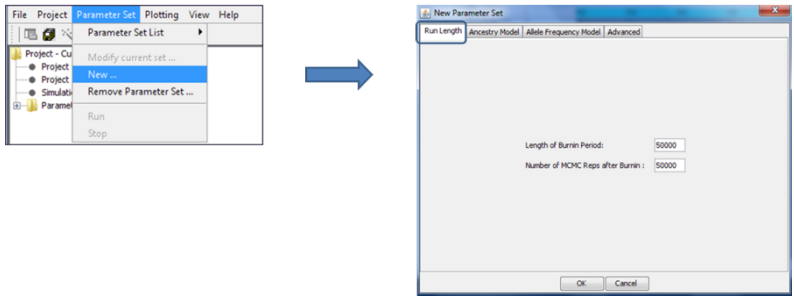


Figura 1.9: Opciones del Set de Parámetros que puede crearse en Structure en la solapa Run Length de la ventana "New Parameter Set".

En la solapa **Advanced** tildamos la opción **Compute probability of the data** para obtener las probabilidades de pertenencia de un individuo a un grupo (Figura 1.10). Luego seleccionamos **Ok** para indicar la creación del set de parámetros.

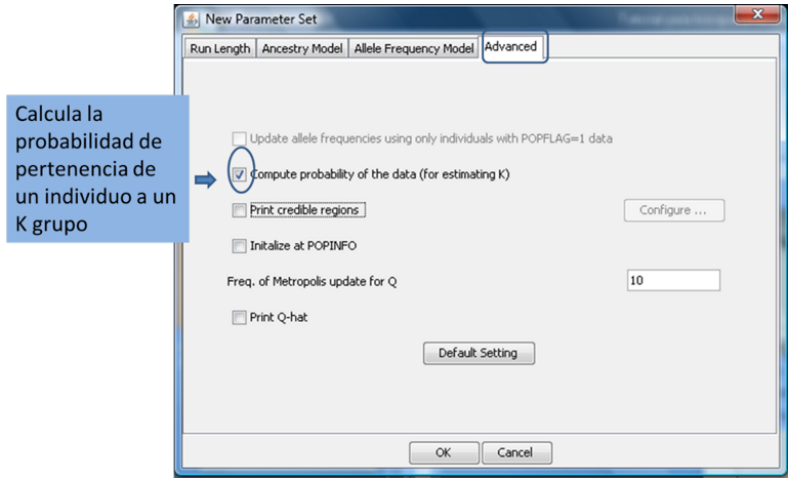


Figura 1.10: Opciones del Set de Parámetros que puede crearse en Structure en la solapa Advanced de la ventana "New Parameter Set".

Para buscar estructura genética poblacional en *Structure*, una opción es correr el set de parámetros creados con el Menú PARAMETER SET comando **Run**. Luego indicamos el número de poblaciones que se asumen, en el archivo de ejemplo son tres poblaciones (Figura 1.11). Esta opción permite una simulación para un único valor de poblaciones (K).

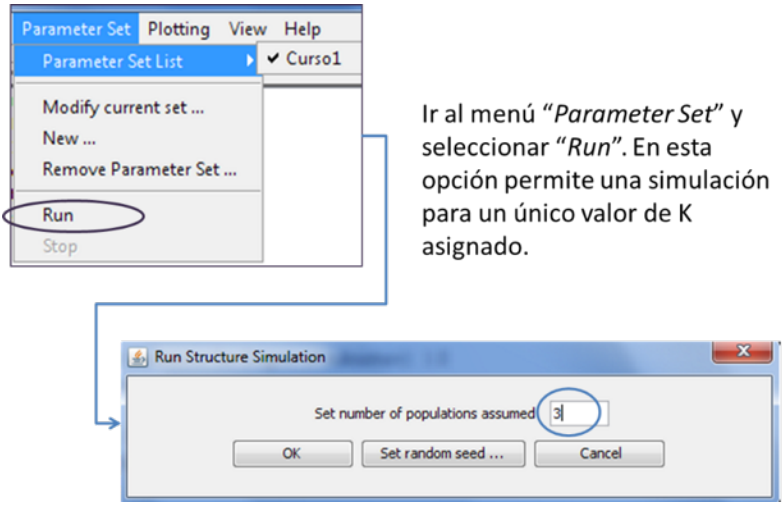


Figura 1.11: Opciones para búsqueda de Estructura Genética Poblacional con Structure en el Menú PARAMETER SET comando Run.

Otra opción es ir al menú PROJECT y seleccionar el comando **Start a Job**. Esta opción permite seleccionar un set de parámetros creado en los pasos anteriores de una lista de set de parámetros. También permite seleccionar un conjunto de valores K consecutivos, por ejemplo, desde K=2 hasta K=5. Además se puede seleccionar el número de iteraciones que se deseen realizar (Figura 1.12).

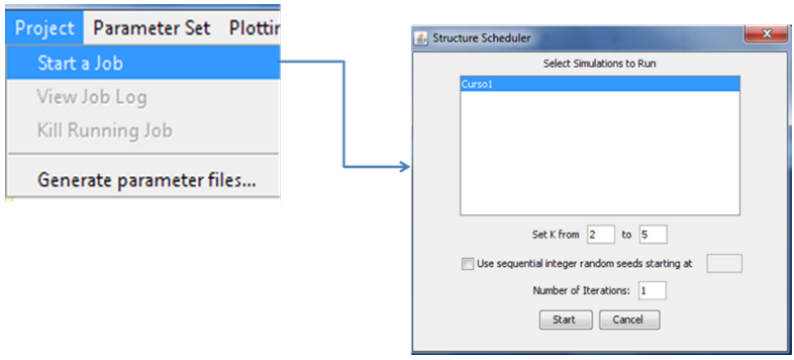


Figura 1.12: Opciones para búsqueda de Estructura Genética Poblacional con Structure en el Menú PROJECT comando Start a Job.

En la primer ventana de resultados se muestra el resumen del proyecto en donde se informa la probabilidad estimada Ln P(D) para cada número de grupo (K) (Figura 1.13). En la segunda ventana se muestra la matriz Q de asignación de probabilidad (Figura 1.14).

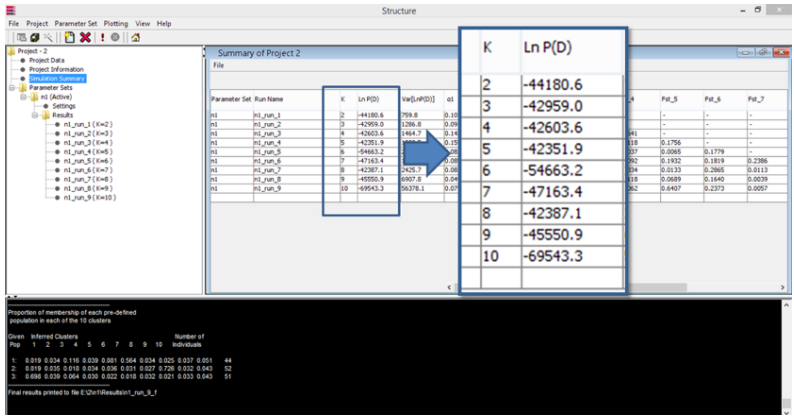


Figura 1.13: Resumen del proyecto de análisis de conglomerados de Structure.

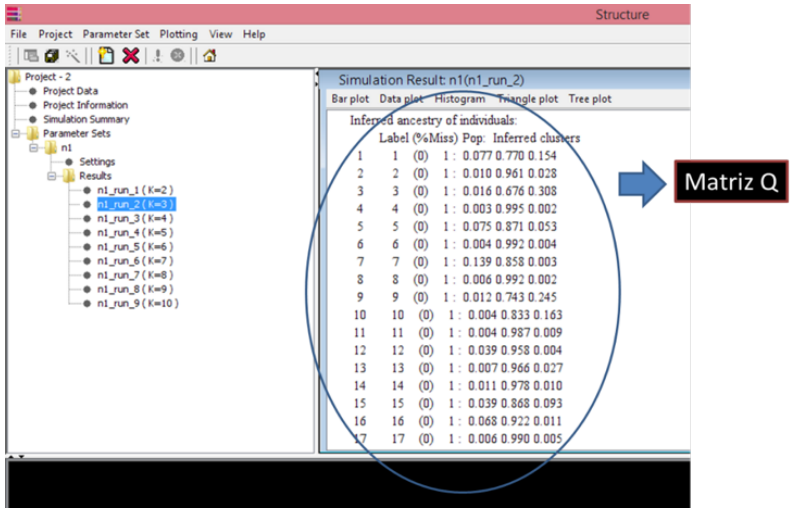


Figura 1.14: Matriz Q de probabilidad de asignación obtenida con Structure

Los resultados, para un K determinado, se visualizan a través de un barplot, donde cada grupo es representado con un color (Figura 1.15).

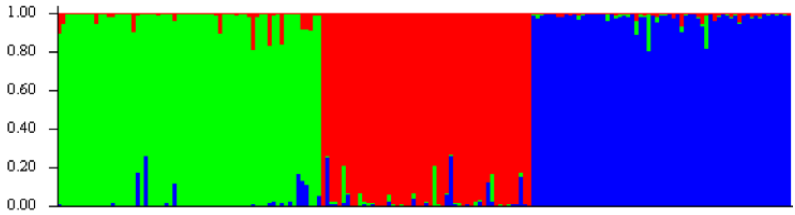


Figura 1.15: Barplot obtenido con Strucutre de base de datos de 147 individuos y 300 marcadores.

Interpretación

La población de 147 de individuos genotipadas por 300 marcadores moleculares está conformada por tres subpoblaciones donde los individuos se han clasificado con alta certidumbre (no hay un grupo de individuos mezcla que no se pueda asignar a alguno de los grupos que se muestran).

Ejemplo 1.4. Clasificación. Análisis de Conglomerados Bayesianos en R. Archivos: Base Conglomerados

El primer paso es instalar la librería *LEA* del repositorio *Bioconductor*. El paquete se instalará por única vez con las primeras dos líneas de comando y se invocará cada vez que se quiera utilizar el script con el comando *library(LEA)*.

```
install.packages("BiocManager")
BiocManager::install("LEA")
library(LEA)
```

Luego, se acondiciona la base de datos para utilizarla en la función "snmf" que generará la clasificación. El primer paso es eliminar la primer columna que contiene los nombres de los individuos y la segunda columna que indica la clasificación de cada individuo en cada grupo.

```
geno <- BaseConglomerados[,-c(1:2)]
```

Luego se recodifica la base de datos indicando con 0 los genotipos AA, con 1 los BB y con 2 los AB o BA. Los datos faltantes se indican con el número 9. Además eliminamos nombres de columnas y filas.

```
Codif <- function(vect) {
  vect <- replace(vect, vect == "0" , 0)
  vect <- replace(vect, vect == "2", 1)
  vect <- replace(vect, vect == "3", 2)
  vect
}
```

```
geno.lea<-t(apply(geno, 1, Codif))
geno.lea[is.na(geno.lea)] <- 9
colnames(geno.lea) <- NULL
rownames(geno.lea) <- NULL
```

Por último, para convertir la base de datos en un objeto de tipo *geno* se utiliza la función "write.geno" que colapasa todas las columnas de la base de datos en una única columna. De este modo la base de datos que se utilizará para la función "snmf" tendrá tantas filas por individuos y una columna con la cadena de caracteres de los marcadores.

```
write.geno(geno.lea, "base.geno")
#> [1] "base.geno"
```

Se utiliza la función "snmf" para estimar el número óptimo de grupos a través del criterio de entropía. El criterio de entropía cruzada se basa en la predicción de genotipos enmascarados para evaluar el ajuste de un modelo con K poblaciones. Este criterio ayuda a elegir el número de poblaciones ancestrales. Un valor menor de entropía cruzada significa una mejor ejecución en términos de capacidad de predicción. En este caso probaremos de k=2 a k=10 grupos.

```
best.k <- snmf(input.file = "base.geno", K = 1:10,
               project = "new", repetitions = 10,
               entropy = T)
plot(best.k, pch = 18, col = 1, lwd = 5)
```

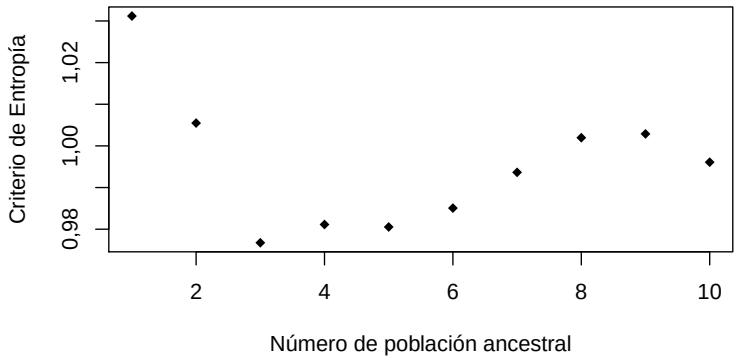


Figura 1.16: Criterio de entropía cruzada para k=1 a k=10 subpoblaciones obtenido con paquete LEA de R de una base de datos de 147 individuos y 300 marcadores.

En el ejemplo se observa que el menor valor del criterio de entropía se obtienen con k=3 subpoblaciones (Figura 1.16). A continuación, se genera la matriz Q de asignación de probabilidad para k=3.

```
# Obtención del valor de CE para K = 3
ce <- cross.entropy(best.k, K = 3)
# Seleccion de la corrida con la menor CE
best <- which.min(ce)
# Coeficientes de ancestria Q para el K
# seleccionado
q <- data.frame(Q(best.k, K = 3, run = best))
colnames(q) <- paste("k", c(1:3), sep = "")
rownames(q) <- rownames(geno)
head(q)
#>        k1     k2    k3
#> 1 0,1429 0,1993 0,658
```

```
#> 2 0,0001 0,1688 0,831
#> 3 0,0001 0,3599 0,640
#> 4 0,0001 0,0001 1,000
#> 5 0,0793 0,1468 0,774
#> 6 0,0001 0,0001 1,000
```

Se observa que el primer individuo tiene probabilidad de 0.14 de pertenecer al grupo 1, 0.2 de pertenecer al grupo 2 y 0.66 de pertenecer al grupo 3. Si se solicita un vector de clasificación el individuo 1 será clasificado en el grupo 3 por tener mayor probabilidad de pertenecer a este que a los demás conglomerados. Con la siguiente línea de comando se solicita el vector de asignación:

```
q.k3 <- apply(X = q, MARGIN = 1,
              FUN = function(x) which.max(x))
head(q.k3)
#> 1 2 3 4 5 6
#> 3 3 3 3 3 3
```

Por último, realizamos el gráfico barplot que mostrará la mixtura de cada uno de los grupos.

```
col.3 <- palette(rainbow(3))
q.ord <-  rbind(q[q.k3==1,][order(q[q.k3==1,
                                    "k1"],
                                  decreasing = T),],
                q[q.k3==2,][order(q[q.k3==2,
                                    "k2"],
                                  decreasing = T),],
                q[q.k3==3,][order(q[q.k3==3,
```

```
                                        "k3"],
                                decreasing = T),])
barplot(t(q.ord), col = 1:3, border = NA,
        space = 0,
        xlab = "Individuos",
        ylab = "Probabilidades de pertenencia")
```

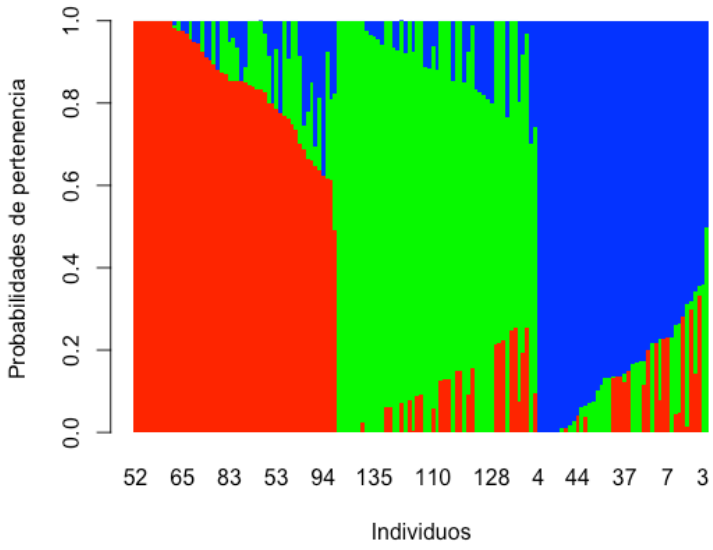


Figura 1.17: Barplot obtenido con paquete LEA de R de base de datos de 147 individuos y 300 marcadores.

Interpretación

La población de 147 de individuos genotipadas por 300 marcadores moleculares está conformada por tres subpoblaciones donde los individuos se han clasificado con alta certidumbre (no hay un grupo de individuos mezcla que no se pueda asignar a alguno de los grupos que se muestran).

1.1.4 Conglomerados basados en máquinas de aprendizaje automático

Con el advenimiento de mayores capacidades de cómputo se están usando técnicas computacionalmente intensivas para clasificación como son los mapas auto-organizativos (SOM del inglés *Self Organizing Maps*), que pertenecen a la familia de redes neuronales (Nikolic *et al.*, 2009). La idea central que soporta el algoritmo SOM es que objetos similares en un espacio de entrada multidimensional serán mapeados cerca uno del otro dentro del mapa resultante de un proceso de organización de información según parecido que se va dando automáticamente. La identificación de conglomerados en SOM puede ser realizada a través de métodos de visualización (Ultsch, 1999) o a través de estadísticos que relacionan la variabilidad entre y dentro de grupos (Fernandex y Balzarini, 2007).

Ejemplo 1.5. Clasificación. Análisis de Conglomerados. Mapas auto-organizativos en R. Archivo Base Conglomerados

Con el siguiente script de R se realiza un mapa auto-organizativo (SOM) para la base de datos que posee 147 individuos y 300 MM. El primer paso es invocar las librerías *class*, *Mass* y *kohonen*.

```
library(class)
library(MASS)
library(kohonen)
```

Luego, se acondiciona la base de datos para utilizarla en la función “som” eliminando la primer columna que contiene los nombres de los individuos.

```
datos<-BaseConglomerados[,-c(1,2)]
```

El inicio del entrenamiento de la red neuronal parte de un muestreo aleatorio sin reposición de los individuos de la base de datos, para ello se usa la función "sample". En este ejemplo se usan todos los individuos de la base de datos, no un subconjunto. Luego del muestreo se realiza un escalamiento de los datos muestreados para conformar el conjunto de datos de entrenamiento.

```
# Creamos el conjunto de entrenamiento
training1<-sample(nrow(datos),147)
Xtraining1<-scale(datos[training1,])
```

Una prueba de validación cruzada puede realizarse muestreando los datos que no fueron utilizados en el paso anterior de entrenamiento de la red. En este ejemplo, al usar la base de datos completa para el muestreo inicial, no tiene datos para una validación externa sino se usan los mismos datos del ajuste.

```
# Conjunto de validacion
Xtest<-scale(datos[-training1,],
      center=attr(Xtraining1,"scaled:center"),
      scale=attr(Xtraining1,"scaled:scale"))
```

La función "som" permite realizar un mapa autoorganizativo de Kohonen (Kohonen, 1997). Los argumentos de la función son: una matriz donde cada fila representa un objeto, por ejemplo *Xtraining1* y la grilla donde se representará el mapa, que al ser de forma hexagonal requiere

el número de filas y columnas. Por ejemplo, en este caso, las opciones son que la grilla tenga una dimensión 4×5 de topología "hexagonal".

```
# Grilla en donde se ubicarán los nodos de la
# red
som.datos<-som(Xtraining1,
               grid=somgrid(4,5,"hexagonal"))
```

La función "som" devuelve un objeto de clase *Kohonen* y los valores del objeto pueden visualizarse con la función *attributes(som.datos)*. Cada uno de esos valores pueden extraerse colocando el nombre del objeto, el signo *$* y el nombre del valor, por ejemplo, *som.datos$data*.

```
attributes(som.datos)
#> $names
#>  [1] "data"
#>  [2] "unit.classif"
#>  [3] "distances"
#>  [4] "grid"
#>  [5] "codes"
#>  [6] "changes"
#>  [7] "alpha"
#>  [8] "radius"
#>  [9] "user.weights"
#> [10] "distance.weights"
#> [11] "whatmap"
#> [12] "maxNA.fraction"
#> [13] "dist.fcts"
#>
#> $class
#> [1] "kohonen"
```

Para la visualización de la grilla guardada en el objeto *som.datos*, definiremos los colores que tendrán los nodos, los mismos se guardaran en un objeto que denominaremos *bgcols* y contiene un vector de tres elementos. Además, en el objeto *clases*, se guardará un vector con la cantidad de individuos que se espera que haya en cada grupo y luego graficamos la red.

```
# Visualizacion
bgcols = c ("cyan", "white", "yellow")
clases<-c(rep(1,44),rep(2,52),rep(3,51))
plot (som.datos, type = "mapping",
      labels = clases[training1],
      col = clases[training1] + 1,
      bgcol = bgcols[ as.integer
                    (classmat2classvec
                      (predict
                      (som.datos,
                      trainY = factor
                      (clases[training1]))
          $unit.predictions))],main="SOM")
```

En el gráfico de la red de nodos, cada círculo representa un nodo de la red que fue ordenado sobre la grilla generada de dimensión 4×5. Dentro de cada nodo se visualiza el grupo al que fue asignado cada individuo (Figura 1.18).

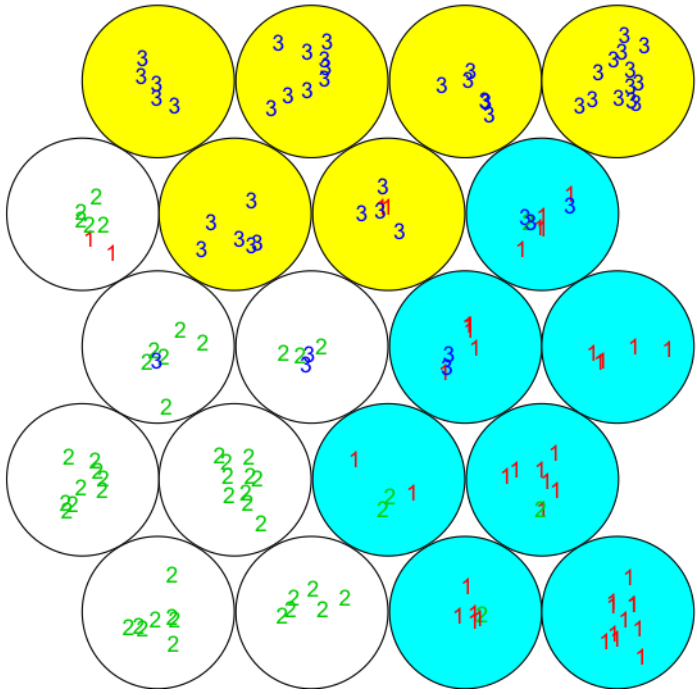


Figura 1.18: Gráfico de la red de nodos obtenidos por SOM. El número de nodo por fila, comenzando desde abajo, de izquierda a derecha.

En el objeto que denominamos *som.datos*, se encuentra un vector que contiene el nodo al que fue asignado cada individuo.

```
clasif.indv<-som.datos$unit.classif
head(clasif.indv)
#> [1] 13 19 10 13 19 16
```

Así, el individuo 1 fue asignados al nodo 13, mientras que los individuos 2 y 5 al nodo 19. En el nodo 16 se encuentra el individuo 6 y así sucesivamente.

Este vector lo denominamos *clasif.indv* y aparecerá como un nuevo objeto en la ventana **Object**.

La función permite predecir los objetos de un mapa para una red de Kohonen entrenada, devolviendo para cada objeto la propiedad asociada con el nodo al que fue asignado. Para *som*, las propiedades de la unidad son calculadas con los argumentos "trainX" y "trainY".

```
som.prediction <- predict(som.datos,
                          newdata = Xtest,
                          trainX = Xtraining1,
                          trainY =
                            factor(clases
                                   [training1]))
table(clases[-training1],
      som.prediction$prediction)
$unit.predictions
```

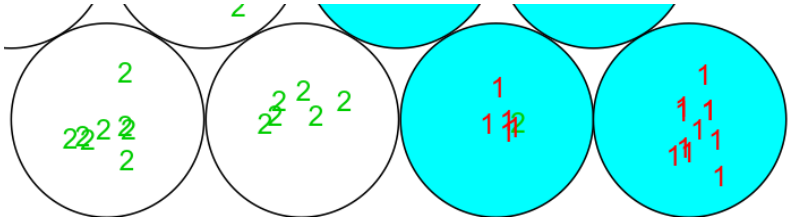


Figura 1.19: Gráfico de los valores predichos asociados a la red de nodos obtenidos por SOM.

Interpretación

A partir del método SOM se agrupó la población de 147 individuos en tres grupos (conjunto de nodos de un mismo color ubicados topológicamente cercanos en la red). Usando la red se predijo el grupo de pertenencia de los individuos de cada nodo. En este ejemplo, dado que el conjunto de entrenamiento es el mismo que el de validación, se observa un alto consenso entre las clasificaciones.

1.2 Validación de agrupamientos

Dado que los algoritmos de cluster definen grupos que no son conocidos *a priori* la partición final de los datos requiere alguna clase de evaluación. El procedimiento que evalúa el resultado del análisis de cluster es conocido como validación del agrupamiento y tiene como finalidad confirmar si la partición de las observaciones o el agrupamiento final obtenido es el que mejor representa la estructura subyacente de los datos (Charrad *et al.*, 2014). Dado que diferentes algoritmos de agrupamiento producen distintos ordenamientos o clasificación de los datos, la evaluación de la efectividad en la clasificación es crítica para tener confianza en los resultados de los agrupamientos. Numerosos índices han sido propuestos combinando información acerca de la compactación intra-grupo y el aislamiento inter-grupos, así como otros factores que se encuentran relacionados a la geometría y propiedades estadísticas de los datos, el número de observaciones y la medida de similaridad/disimilaridad usada. Los índices de validación interna Conectividad, Ancho de Silueta y Dunn han sido propuestos para evaluar las clasificaciones de algoritmos y se encuentran implementados en el paquete *clValid* de *R*. La Conectividad (Handl y Knowles, 2005) está relacionada a la distancia entre observaciones vecinas en un mismo conglomerado, mientras menor es el valor de conectividad mejor, debido a que menor valor indica que los objetos dentro de un mismo conglomerado son más parecidos. El Ancho de Silueta (Kaufman y Rousseeux, 1990) mide la confianza con la que cada observación es asignada a un grupo. Si ha sido bien asignada tendrá valores cercanos a 1, mayor valor indica mayor confianza de la asignación de un objeto a un conglomerado. El máximo valor del índi-

ce determina el número de grupos óptimos. El índice de Dunn (Dunn, 1974) es el cociente entre la mínima distancia entre dos observaciones que no pertenecen a un mismo conglomerado y la máxima distancia entre dos observaciones de un mismo conglomerado. Combina la compactación (homogeneidad dentro del conglomerado) con el grado de separación entre conglomerados (Guy Brock *et al.*, 2008). Mayor valor implica mayor varianza entre conglomerados y menor varianza dentro del conglomerado.

Ejemplo 1.6. Validación de agrupamientos Archivo Datos Trigo

Para ejemplificar se utilizará un conjunto de 599 líneas genotipadas con 1279 marcadores del tipo DArT (McLaren *et al.*, 2000; McLaren *et al.*, 2005) codificados según la presencia/ausencia del marcador molecular como 1 y 0, respectivamente. Esta base de datos es púbica y se encuentra disponible en el paquete BGLR de R (R Core Team, 2019) se compararon diferentes algoritmos de agrupamiento y se validaron los mismos mediante los índices Ancho de Silueta, Dunn y Conectividad. El número sugerido de agrupamientos estará indicado por aquel que maximice dichos índices.

A continuación se presenta el código para realizar la validación de los métodos de agrupamiento jerárquico UPGMA, no jerarquico k-means y basado en redes neuronales SOM. Cada algoritmo fue evaluado para k=2 a 10 agrupamientos.

```
library(clValid)
library(cluster)
library(kohonen)
```

```
set.seed(1234)

#indices de validacion interna para K-means
intern_k<-clValid(
  Datos.Infogen.TRIGO[,2:1280],2:10,
  clMethods="kmeans", metric="euclidean",
  validation="internal")
summary(intern_k)

#indices de validacion interna para UPGMA
intern_U<-clValid(
  Datos.Infogen.TRIGO[,2:1280],2:10,
  clMethods="hierarchical",metric="euclidean",
  validation="internal")
summary(intern_U)

#indices de validacion interna para SOM
intern_SOM<-clValid(
  Datos.Infogen.TRIGO[,2:1280],2:10,
  clMethods="som", metric="euclidean",
  validation="internal")
summary(intern_SOM)
```

Cada uno de los índices de validación interna fue evaluado para cada uno de los k grupos. En función de los criterios de cada índice de validación (mayor es mejor o menor es mejor según el índice) el objeto del tipo *clvalid* devuelve una tabla con el número óptimo de grupos para el conjunto de datos. Dado que puede no haber coincidencia entre los índices en el número óptimo de grupos, se rige por la regla de la mayoría. Por ejemplo, para el método de agrupamiento *K-means*, conectividad y ancho de silueta recomiendan dos grupos mientras que Dunn 8 (Tabla 1.1).

Tabla 1.1: Número óptimo de grupo según los índices de validación interna conectividad, Dunn y ancho de silueta para tres métodos de agrupamiento: k-means, UPGMA y SOM

Índices de validación	K-means	UPGMA	SOM
Conectividad	2	2	2
Dunn	8	2	2
Silueta	2	3	2

2

Modelos de mapeo asociativo

2.1 Mapeo Asociativo

Info-Gen es un Software estadístico para análisis de datos genéticos (Balzarini y Di Rienzo, 2004), en su interfaz con el software R (R CoreTeam, 2019) permite estimar diferentes modelos propuestos en estudios de mapeo asociativo (MA). El mapeo asociativo es una técnica introducida en estudios de mejoramiento genético para identificar genes responsables de la variación de caracteres cuantitativos (Breseghello y Sorrells, 2006; D'hoop *et al.*, 2008; Kraakman *et al.*, 2006; Remington *et al.*, 2001; Stich *et al.*, 2008). El MA se usa en el mejoramiento de especies vegetales para identificar regiones genómicas asociadas a caracteres fenotípicos de herencia compleja como rendimientos y componentes de rendimiento (Aranzana *et al.*, 2005; Thornsberry *et al.*, 2001; Zhu *et al.*, 2008). Desde una perspectiva analítica, el objetivo del MA es evaluar la significancia estadística de las asociaciones entre la información que proveen los marcadores moleculares (variantes genéticas) y la característica fenotípica de interés. Los MA realizados a nivel poblacional y a partir de frecuencias alélicas, tasas de polimorfismos y desequilibrio de ligamiento (LD), han permitido identificar asociaciones reproducibles en numerosas especies vegetales (Flint-García *et al.*, 2005). Una de las principales ventajas del MA es que no requiere del desarrollo de una población específica de mapeo.

Es posible realizarlo a través de un conjunto de genotipos como pueden ser un subconjunto de líneas de un banco de germoplasma (población genotípicamente diversa). El MA explora asociaciones estadísticas entre variaciones del genotipo y fenotipo, algunas de ellas pueden ser reales y otras simplemente pueden darse por azar o por la presencia de estructuras en los datos. El MA puede ser usado para analizar la herencia compartida de una colección de individuos sin ancestría conocida pero donde ha habido recombinación (Yu y Buckler, 2006); la existencia de múltiples generaciones de recombinación conlleva a una mayor resolución de mapeo con mayor oportunidad de disipación del desequilibrio de ligamiento (LD). Los principales mecanismos que provocan el LD son la mutación y la deriva, mientras que la recombinación reduce el LD (Jannink *et al.*, 2009). Por esto, se explicita que para obtener una exitosa detección de asociaciones la densidad de marcadores debe coincidir con el rango de decaimiento del LD (Jannink *et al.*, 2009). Por ejemplo, si el LD decae de forma rápida, se requiere una mayor densidad de marcadores para capturar aquellos marcadores que se encuentran suficientemente cerca de los sitios funcionales (Yu y Buckler, 2006). Dado que el MA depende de la extensión y la distribución del LD entre los marcadores involucrados, es importante calcular empíricamente la magnitud de la correlación entre marcadores. Técnicas gráficas como el **heatmap** (mapas de calor) permiten visualizar patrones de la magnitud de cientos y miles de correlaciones como son las que ocurren en mapeos con alta densidad de marcadores.

2.1.1 Modelos

El modelado de asociaciones marcador-fenotipo incluyendo covariables que reflejen la estructura o la presencia de relaciones de parentesco entre individuos, puede realizarse con distintos modelos estadísticos. Se han propuesto varios tipos de modelos de asociación, principalmente modelos de regresión caracter *vs.* marcador, que pueden ser ajustados en el marco teórico de los modelos lineales mixtos (MLM) (Demidenko, 2004) y la estimación REML (Patterson y Thompson, 1971). El modelo base o modelo de referencia en la etapa de modelación estadística de asociaciones es un modelo de regresión lineal con efectos de marcadores en el vector de coeficientes de regresión, $y = X + e$, donde y es el vector de caracteres fenotípicos de dimensión $n \times 1$, X es la matriz de marcadores moleculares de dimensión $n \times p$, es el vector de parámetros (o efectos fijos) de dimensión $p \times 1$ y e es el vector de errores de dimensión $n \times 1$. Este modelo donde se asume que no hay estructua entre los datos, se denominará en adelante, modelo sin corrección por estructura (Modelo SCE). Pritchard *et al.* (2000) propusieron un modelo de mapeo asociativo, que incorpora las componentes principales (CPs) asociadas a la estructura genética caracterizada por los marcadores moleculares. Usando los datos moleculares, se extraen las CPs significativas para describir la estructura y éstas son usadas para confirmar la matriz de CPs que será usada como covariable en el modelo de asociación. Es usual que con 2 a 5 CPs se pueda explicar o representar la estructura en la población de mapeo, sin embargo, en otras ocasiones la cantidad de CPs es mayor. Otra alternativa es usar el estadístico de Tracy-Widom (1994) para identificar la cantidad de CPs que son necesarias para modelar la estructura. Las componentes principales son obtenidas a partir de un Análisis de Componentes Principales (ACP) sobre los datos mole-

culares. El ACP permite obtener un conjunto de nuevas variables, que se generan como combinación lineal de los datos moleculares originales. Tales variables tienen la característica de no estar correlacionadas y resultan óptimas para señalar estructuras entre los genotipos de la población de mapeo. Así, las CPs conformadas a través de la combinación de marcadores moleculares de los genotipos de la población de mapeo permiten señalar diferencias entre genotipos causadas por la existencia de estructuración genética. Este modelo será denominado en adelante Modelo P. Yu *et al.* (2006) propusieron modelar la estructura genética a partir de la matriz de coancestría, matriz "kinship" (K), entre los efectos aleatorios de genotipos. Este modelo será denominado Modelo K. Luego, Yu *et al.* (2006), propusieron un modelo unificado donde, además de los efectos considerados en el Modelo K, toma los efectos fijos relacionados a la estructura genética de la población obtenidos desde la matriz P proveniente del análisis de componentes principales. Este modelo es denominado PK. Otra forma de incorporar la estructura genética poblacional (EGP) es a partir de la matriz Q que presenta las probabilidades de pertenencia de los indivduos a las p poblaciones en estudio. Esta matriz puede ser obtenida dese el programa *Structure* (Pritchard *et al.*, 2000). El modelo que incorpora la EPG a través de la matriz Q será nombrado en adelante como Modelo Q. Así mismo, el modelo que incluye la matriz de parentezco (K) y la matriz de relaciones Q, se denominará Modelo QK.

2.1.2 Corrección por multiplicidad

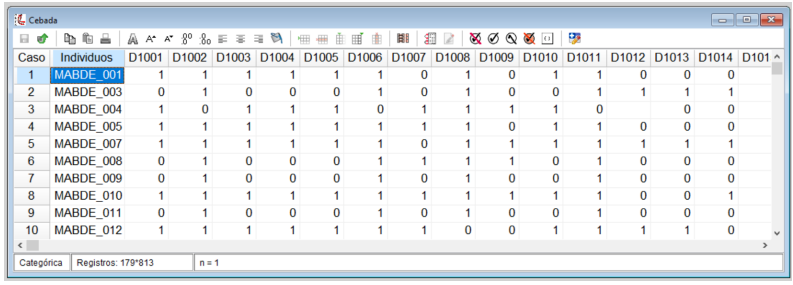
La asociacion entre marcadores y fenotipo pueden no ser consecuencia de un ligamiento real entre marcadores y loci de interés. La alta tasa de falsos positivos (error tipo I) es un problema en el MA y por ello existe la necesidad de corregir las significancias estadísticas por multiplicidad. Los falsos positivos o asociaciones espurias son esperables debido a la gran cantidad de prueba de hipótesis que se realizan sobre los mismos datos, por ejemplo, trabajando con un nivel de significacion del 5 %, 5 de cada 100 marcadores pueden mostrarse asociados al carcater solo por azar. Otra causa de falsos positivos es la falta de independencia entre los individuos de la población de mapeo. Por ejemplo, se supone que si una mutación incrementa la observación de una característica en una población de individuos, entonces podemos esperar que el alelo asociado a tal característica sea más frecuente entre los individuos emparentados que entre el resto de los individuos. Cuando se lleva a cabo un análisis de asociación sin considerar los efectos de la estructura poblacional, se aumenta el riesgo de detectar asociaciones espurias entre marcadores y el fenotipo de interés. Un método de control del error tipo I es la correccion de valores p por la aproximación de Bonferroni (1935). Este puede resultar excesivamente conservador cuando la cantidad de pruebas de hipótesis involucradas es alta. Un criterio alternativo de corrección de valores p es el propuesto por Benjamini y Hochberg (1995) el cual estima la proporción esperada de hipótesis falsas rechazadas respecto de todas aquellas rechazadas, proporción denominada tasa de falsos descubrimientos o FDR (del inglés, *False Discovery Rate*). La tasa FDR se calcula en base a las proporciones de falsos positivos (FP) y verdaderos positivos (VP). Los FP son todos aquellos valores p significativos vinculados a marcadores que no están asociados al fenoti-

po y los VP son todos aquellos marcadores positivos que efectivamente están asociados al fenotipo. El método ha sido desarrollado para mantener la tasa de error tipo I en el valor nominal o pre-especificado por el experimentador. Cheverud (2001) propuso una idea de ajuste de valores p para pruebas correlacionadas como son las que podría emerger en el caso de LD. La idea es probar las hipótesis como si ellas fueran independientes estimando previamente un número efectivo (Meff) de pruebas independientes. Li y Ji (2005) propusieron una estimación del Meff basada en la correlación entre marcadores moleculares y diseñaron un procedimiento basado en el Meff para controlar el error tipo I en la familia de hipótesis que se prueban en análisis de QTL clásico. Esta prueba, no considera la presencia de estructura genetica poblacional. Peña *et al.* (2013) propusieron una prueba similiar conceptualmente a la de Li y Ji pero ajustando por la presencia de estructura genética.

2.2 Ajuste de Modelos

2.2.1 Descripción del conjunto de datos de prueba usados para ilustración

Los datos utilizados como ejemplo se denominan *Cebada.igdb* y corresponden a un programa de mejoramiento de cebada (Comadran *et al.*, 2009). Esta base de datos está conformado por 179 individuos (filas del archivo) y 811 marcadores moleculares del tipo Diversity Array Technology (DArT®) ubicados como columnas de la base de datos. La información genotípica está expresada como presencia/ausencia del marcador molecular representadas por 1 y 0, respectivamente. El archivo de datos tiene una columna (variable) con información sobre el carácter fenotípico que en este ejemplo se denomina Rendimiento (y). El archivo cuenta con otra columna denominada Individuos, que identifica cada observación (fila del archivo) y es del tipo variable de clasificación. El formato de la base de datos se presenta a continuación (Figura 2.1). En los Modelos que incluyan la información provista por la matriz Q, se usará el archivo *CebadaQ.igdb* que proporciona, además, las columnas que conforman la matriz de relaciones genéticas Q.



Cebada

Caso	Individuos	D1001	D1002	D1003	D1004	D1005	D1006	D1007	D1008	D1009	D1010	D1011	D1012	D1013	D1014	D101
1	MABDE_001	1	1	1	1	1	1	0	1	0	1	1	0	0	0	
2	MABDE_003	0	1	0	0	0	1	0	1	0	0	1	1	1	1	
3	MABDE_004	1	0	1	1	1	0	1	1	1	1	0		0	0	
4	MABDE_005	1	1	1	1	1	1	1	1	0	1	1	0	0	0	
5	MABDE_007	1	1	1	1	1	1	0	1	1	1	1	1	1	1	
6	MABDE_008	0	1	0	0	0	1	1	1	1	0	1	0	0	0	
7	MABDE_009	0	1	0	0	0	1	0	1	0	0	1	0	0	0	
8	MABDE_010	1	1	1	1	1	1	1	1	1	1	1	0	0	1	
9	MABDE_011	0	1	0	0	0	1	0	1	0	0	1	0	0	0	
10	MABDE_012	1	1	1	1	1	1	1	0	0	1	1	1	1	0	

Categórica | Registros: 179*813 | n = 1

Figura 2.1: Formato del archivo de datos para mapeo asociativo en Info-Gen. Las filas reperesentan las observaciones (individuos), las columnas representan un marcador codificado como binario. Archivo Cebada.igdb.

2.2.2 Modelo SCE

Modelo sin corrección por estructura genética poblacional

La ecuación del modelo sin corrección por estructura (SCE) corresponde a la del modelo de regresión lineal $y = X\beta + e$, donde y es el carácter fenotípico (variable respuesta) que se asocia a cada marcador molecular (variable regresora) a través de un coeficiente de regresión β que debe ser estimado desde los datos para cada marcador y e es un término de error aleatorio. Este modelo no contempla ninguna corrección por estructura genética.

Para ajustar este modelo en *Info-Gen*, ir al Menú **Mejoramiento**, sumbenú **Mapeo Asociativo**, opción **1. Modelo SCE**. Se desplegarán tres opciones, una de **Datos de prueba** que permitirá abrir el archivo de ejemplo denominado *Cebada.igdb*. Antes de ajustar un modelo, es necesario tener una base de datos sobre la cual *Info-Gen* realizará los ajustes. Luego, seleccionar **Run...**, para indicar la acción de ajustar un modelo (Figura 2.2).

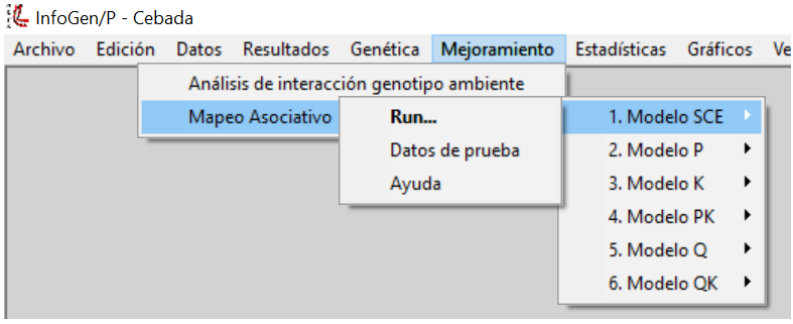


Figura 2.2: Opciones de Modelos de Mapeo Asociativo que pueden ajustarse en Info-Gen. El comando Run permite ajustar el modelo, el comando Datos de prueba abre un conjunto de datos de ejemplo, el comando Ayuda abre un tutorial.

Al hacer click en el comando **Run...** se abrirá la ventana selector de variables, en la cual se listan las columnas

de la base de datos activa. En la solapa Variables se visualizan tres espacios; en el primero llamado **Individuos** deberá seleccionar desde el lado izquierdo la columna que contiene la identificación de los individuos, en este ejemplo se denomina "Individuos" y con el botón con forma de flecha transportarlo al espacio de **Individuos**. En el segundo espacio denominado **Marcadores Moleculares** se deberan traspasar, desde el sector izquierdo de la ventana de selector de variables, las columnas del archivo que contienen información de marcadores moleculares. En el último espacio llamado **Carácter Fenotípico** deberá ingresarse la variable que contiene la información fenotípica (variable respuesta, y). En este archivo de ejemplo la columnas que contiene el carácter fenotípico se denomina **Rendimiento**. Luego, seleccionar el botón **Aceptar** en la parte inferior izquierda del selector de variables (Figura 2.3).

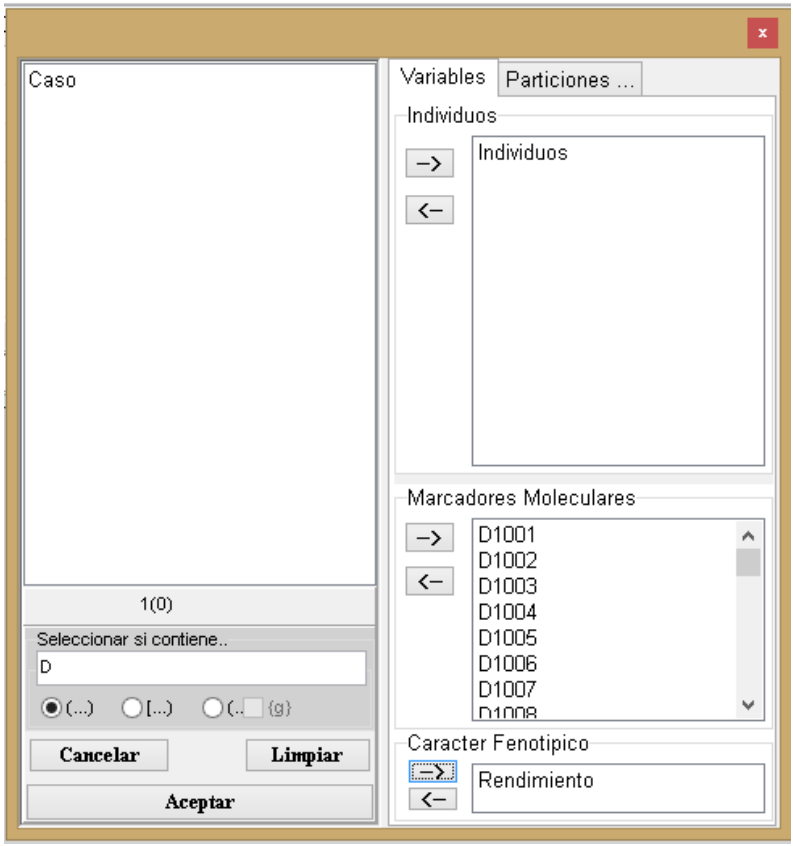


Figura 2.3: Ventana de selector de variable de Info-Gen que se despliega al seleccionar el menú Mejoramiento, sumbenú Mapeo Asociativo, opción 1. Modelo SCE. Archivo de ejemplo Cebada.igdb.

La ventana siguiente permite fijar el valor de α y el método de corrección por multiplicidad que permite calcular el umbral para el rechazo de la hipótesis de no asociacion (Figura 2.4). Las opciones disponibles en *Info-Gen* para calcular el umbral de significancia para disminuir la tasa de falsos positivos son: Li-Ji que estima el número efectivo de pruebas independientes por método propuesto por Li y Ji (Li y Ji, 2005) y FDR la tasa de falsos positivos. Si se selecciona la opción $-log10(valor \quad p)$ no se realiza

corrección por multiplicidad, con lo cual el umbral para la aceptación de la hipótesis nula se estima como el logaritmo del valor p de cada prueba de hipótesis. Además del método de corrección por multiplicidad, el usuario puede seleccionar el nivel de significancia (valor de alfa) que por defecto es 0.05 pero puede ser modificado.

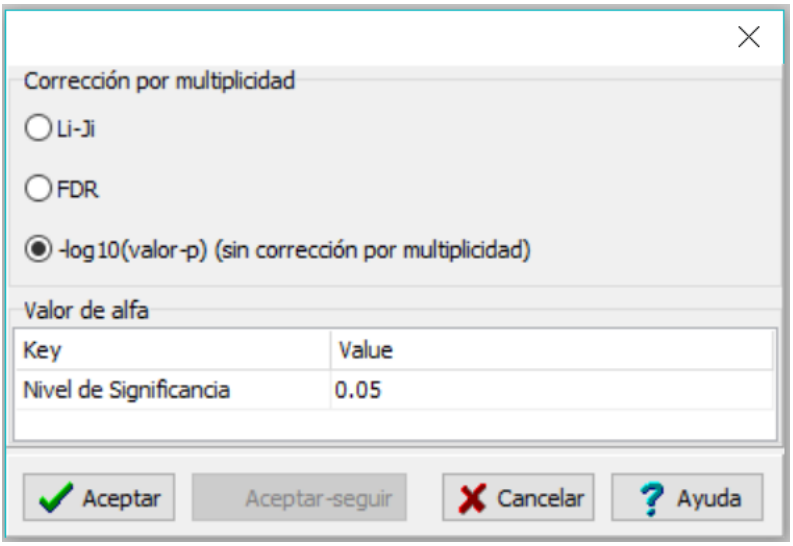


Figura 2.4: Ventana de la opción modelo SCE (Menú Mejoramiento, submenú mapeo asociativo) en Info-Gen.

Resultados

En la primer tabla de la ventana Resultados (Figura 2.5) se muestra una breve descripción de la **Población de Mapeo** indicando la cantidad de individuos y de marcadores moleculares usados en el análisis. La segunda tabla contiene las **Opciones seleccionadas**, *i.e.*, nivel de significancia y corrección por multiplicidad. La tercer tabla denominada **Valores p. Modelo SCE**, presenta los valores p, asociados a la prueba de hipótesis contrastada para cada marcador, indicando el Marcador, el Orden (o ubicación del marcador) del marcador y el valor p obtenido luego

de corrección seleccionada. La tabla denominada **Marcadores seleccionados (Valor p<Umbral)**, contiene un listado de los marcadores seleccionado según el valor de α y el umbral seleccionados en las opciones de corrección por multiplicidad (Figura 2.5).

Población de Mapeo

Resumen	Cantidad
Número de Individuos	179
Número de Marcadores	811

Opciones seleccionadas

Opciones	Valores usados
Nivel de Sign.	0.05
Correc. por multip.	Sin corrección

Valores-p. Modelo SCE

Marcador	Orden	valor-p
D1001	1	0.03
D1002	2	0.09
D1003	3	0.01
D1004	4	0.01
D1005	5	0.01
.	.	.
.	.	.
.	.	.
D7131	810.00	0.02
D7132	811.00	1.5E-03

Marcadores seleccionados (Valor-p< Umbral)

Marcador	Orden	valor-p
D1001	1	0.03
D1003	3	0.01
D1004	4	0.01
D1005	5	0.01
D1010	10	0.01
.	.	.
.	.	.
.	.	.
D7131	810.00	0.02
D7132	811.00	1.5E-03

Figura 2.5: Ajuste del modelo SCE (Menú Mejoramiento, submenú mapeo asociativo) en Info-Gen con un nivel de significación (alfa) de 0.05 y sin aplicar correción por multiplicidad (-log10(valor p)). Archivo Cebada.igdb.

En el gráfico puede observarse una línea de corte en el umbral seleccionado. Los marcadores que superan dicha línea de corte son los que fueron estadísticamente significativos indicando una asociación entre el genotipo y el fenotipo (Figura 2.6).

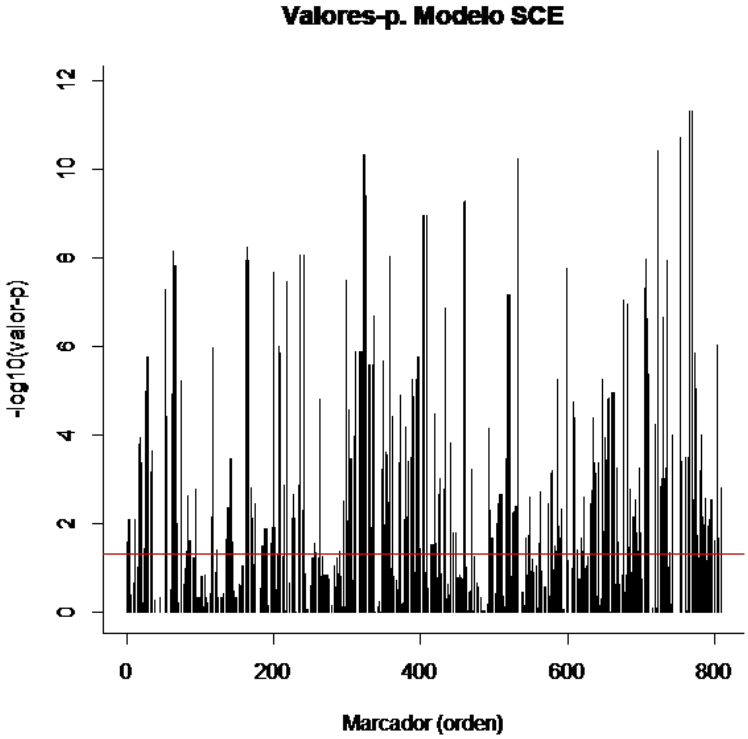


Figura 2.6: Gráfico de valores p obtenidos al ajustar un modelo SCE (Menú Mejoramiento, submenú mapeo asociativo) en Info-Gen con un nivel de significación (alfa) de 0.05 y ninguna correción por multiplicidad para cada marcador. Archivo Cebada.igdb. La línea roja indica el umbral para determinar la significancia de la asociación, valores por encima del umbral indica que el marcador está asociado al carácter bajo estudio.

2.2.3 Modelo P

Modelo con corrección por estructura estimada a través de un Análisis Multivariado de Componentes Principales

La ecuación del Modelo P corresponde a la de un modelo de regresión lineal $y = X\beta + Sv + e$, donde y es el carácter fenotípico (variable respuesta) que se asocia a cada marcador molecular (variable regresora) a través de un coeficiente de regresión β que debe ser estimado desde los datos, S es la matriz de estructura genética (matriz P construida con las componentes principales del ACP realizado sobre los datos moleculares, usualmente se seleccionan aquellas CPs estadísticamente significativas según Tracy-Widom), v es el vector de efectos fijos de la estructura poblacional, e representa el término de error aleatorio.

Pasos en Info-Gen para ajustar un Modelo P

Ir al Menú **Mejoramiento**, sumbenú **Mapeo Asociativo**, opción **2. Modelo P**. Al seleccionar la opción Datos de prueba se abrirá el archivo de ejemplo *Cebada.igdb*. Al apretar la opción **Run...**, se indica la acción de ajustar el modelo seleccionado (Figura 2.7).

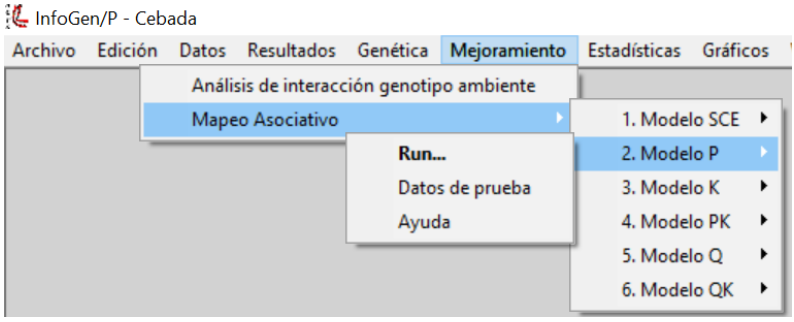


Figura 2.7: Opciones de Modelos de Mapeo Asociativo que pueden ajustarse en Info-Gen. El comando Run permite ajustar el modelo, el comando Datos de prueba abre un conjunto de datos de ejemplo, el comando Ayuda abre un tutorial.

Al hacer click en el comando **Run...** se abrirá la ventana selector de variables, en la cual se listan las columnas de la base de datos activa. En la solapa Variables se visualizan tres espacios; en el primero llamado **Individuos** deberá seleccionar desde el lado izquierdo la columna que contiene la identificación de los individuos, en este ejemplo se denomina "Individuos" y con el botón (->) transportarlo al espacio de Individuos del lado derecho de la ventana. En el segundo espacio denominado **Marcadores Moleculares** se deberan traspasar las columnas del archivo que contienen información de marcadores molecularеas. En el último espacio llamado **Carácter Fenotípico** deberá ingresarse la variable que contiene la información fenotípica (variable respuesta), en este archivo de ejemplo la columna que contiene el carácter fenotípico se denomina "Rendimiento". Luego seleccionar el botón **Aceptar** en la parte inferior izquierda del selector de variables (Figura 2.8).

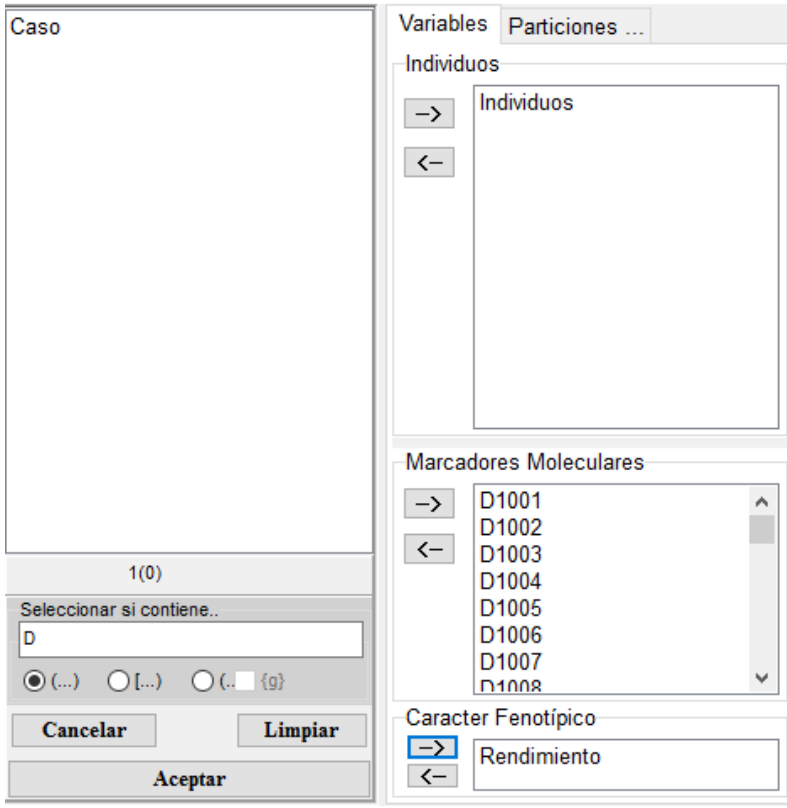


Figura 2.8: Ventana de selector de variable de Info-Gen que se despliega al seleccionar el menú Mejoramiento, submenú Mapeo Asociativo, opción 2. Modelo P. Archivo Cebada.igdb.

La siguiente ventana permite fijar el valor de α y el método de corrección por multiplicidad que permite calcular el umbral para el rechazo de la hipótesis de no asociación (Figura 2.9). Las opciones disponibles en *Info-Gen* para calcular el umbral de significancia para disminuir la tasa de falsos positivos son: Li-Ji que estima el número efectivo de pruebas independientes por método propuesto por Li y Ji (Li y Ji, 2005) y FDR la tasa de falsos positivos. Si se selecciona $-log10(valor \quad p)$ no se realizará corrección por multiplicidad, con lo cual el umbral para la aceptación de la hipótesis nula se estima como el logaritmo del valor p

de cada prueba de hipotesis. Además del método de corrección por multiplicidad, el usuario puede seleccionar el nivel de significancia (valor de alfa), por defecto es 0.05 pero puede ser modificado y la Cantidad de CP (componentes principales) a usar en el ajuste del Modelo P. TW es la opción por defecto, seleccionando las componentes principales significativas por Tracy-Widom.

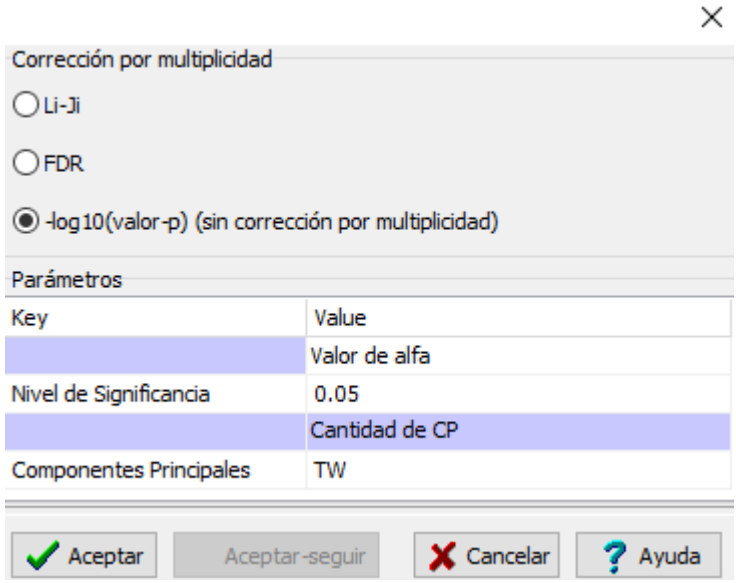


Figura 2.9: Ventana de la opción Modelo P (Menú Mejoramiento, submenú Mapeo Asociativo) en Info-Gen.

Resultados

En la primer tabla de la ventana Resultados (Figura 2.10) se muestra una breve descripción de la **Población de Mapeo** indicando la cantidad de individuos, de marcadores moleculares y las CP significativas según TW usados en el análisis. La segunda tabla contiene los coeficientes de las **Componentes Principales significativas (Tracy-Widom)**. La tercera tabla contiene las **Opciones seleccionadas**, *i.e.*, nivel de significancia, corrección por mul-

tiplicidad y cantidad de CPs. La cuarta tabla presenta los **Valores p. Modelo P** asociados a la prueba de hipótesis contrastada para cada marcador, indicando el Marcador, el Orden del marcador (o ubicación del marcador) y el valor p obtenido luego de la corrección seleccionada. La tabla denominada **Marcadores seleccionados (Valor p<Umbral)**, contiene un listado solo con los marcadores seleccionado según el valor de α y el umbral seleccionados en las opciones de corrección por multiplicidad (Figura 2.10).

Población de Mapeo

Resumen	Cantidad
Número de Individuos	179
Número de Marcadores	811
CP significativas (T-W)	17

Componentes Principales significativos (Tracy-Widom)

PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13	PC14	PC15	PC16	PC17
-4.60	6.56	-2.20	-5.00	0.01	-7.30	3.40	-2.52	4.00	-1.05	0.51	-2.72	6.52	-1.89	4.28	-2.7E-03	0.76
-6.62	13.01	-3.55	-6.53	-3.39	5.01	-2.51	4.32	-0.17	3.37	-2.88	3.80	1.75	-1.26	2.83	-0.13	-2.33
9.66	-4.61	2.65	1.72	0.26	2.00	0.58	-2.39	1.51	3.13	-3.50	-5.17	1.89	-0.51	-0.53	-0.61	-4.73
-7.23	2.53	-2.12	-0.93	8.73	-7.82	-0.12	4.15	-3.74	-2.52	-0.87	2.76	3.86	5.81	1.24	2.83	-5.43
-9.23	10.17	-6.27	-2.76	4.92	-8.07	6.17	-6.76	0.85	4.85	-2.16	2.36	-0.61	-1.92	-1.02	-0.77	-2.21
-6.91	9.00	-2.25	-0.85	1.81	-5.20	-0.85	-0.74	-0.72	2.81	3.67	-1.20	4.30	-4.65	-7.38	0.01	2.12

Opciones seleccionadas

Opciones	Valores usados
Nivel de Sign.	0.05
Correc. por multip.	Sin corrección
N° PCs usadas	17

Valores-p del modelo P

Marcador	Orden	valor-p
D1001	1.00	4.3E-03
D1002	2.00	0.03
D1003	3.00	6.8E-04
D1004	4.00	6.8E-04
.	.	.
.	.	.
.	.	.
D7130	809.00	0.43
D7131	810.00	2.9E-03
D7132	811.00	7.0E-05

Marcadores seleccionados (Valor-p < alfa)

Marcador	Orden	valor-p
D1001	1.00	4.3E-03
D1002	2.00	0.03
D1003	3.00	6.8E-04
D1004	4.00	6.8E-04
.	.	.
.	.	.
.	.	.
D7125	804.00	3.4E-03
D7126	805.00	4.1E-09
D7131	810.00	2.9E-03

Figura 2.10: Ajuste del modelo P (Menú Mejoramiento, submenú mapeo asociativo) en Info-Gen con un nivel de significación de 0.05 y sin aplicar correción por multiplicidad. Archivo Cebada.igdb.

En el gráfico puede observarse una línea de corte en el umbral seleccionado. Los marcadores que superan dicha línea de corte son los que fueron estadísticamente significativos

indicando una asociación entre el genotipo y el fenotipo (Figura 2.11).

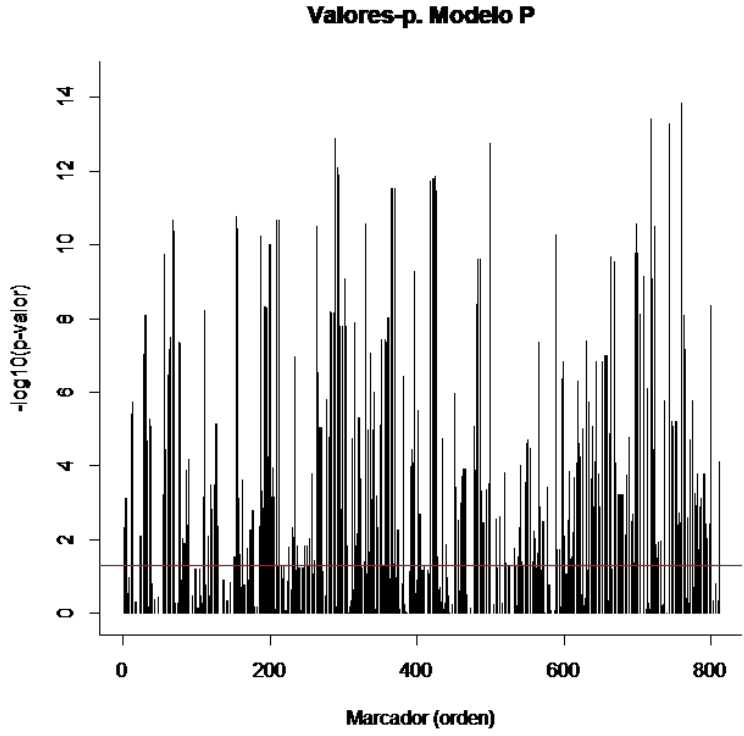


Figura 2.11: Gráfico de los valores p obtenidos al ajustar un modelo P (Menú Mejoramiento, submenú Mapeo Asociativo) en Info-Gen con un nivel de significación de 0.05 y ninguna corrección por multiplicidad para cada marcador. La línea roja indica el umbral para determinar la significancia de la asociación, valores por encima del umbral indica que el marcador está asociado al caracter bajo estudio. Archivo Cebada.igdb.

Entre las salidas del Modelo P se muestran dos gráficos relacionados al Análisis de Componentes Principales, realizado sobre los marcadores moleculares para capturar la estructura genética poblacional en las componentes principales significativas (CPs) según Tracy-Widom. Uno de

los gráficos muestra la proporción de la varianza explicada por cada componente principal que resultó estadísticamente significativo según la propuesta de Tracy Widom. Sobre el eje de las ordenadas se muestra la proporción de varianza explicada en función de la cantidad de componentes principales (CPs). En línea azul (curva que asciende) se representa la varianza acumulada asociada a cada componente significativa según Tracy-Widom y en línea de color rojo (curva que desciende) la proporción de la varianza total explicada. Debido a que la primer componente principal explica más variabilidad que la segunda y la segunda más variabilidad que la tercera y así sucesivamente, la proporción de la varianza acumulada va disminuyendo (Figura 2.12). El otro gráfico es un diagrama de dispersión de las observaciones (Individuos) en el espacio de las dos primeras CP (Figura 2.13).

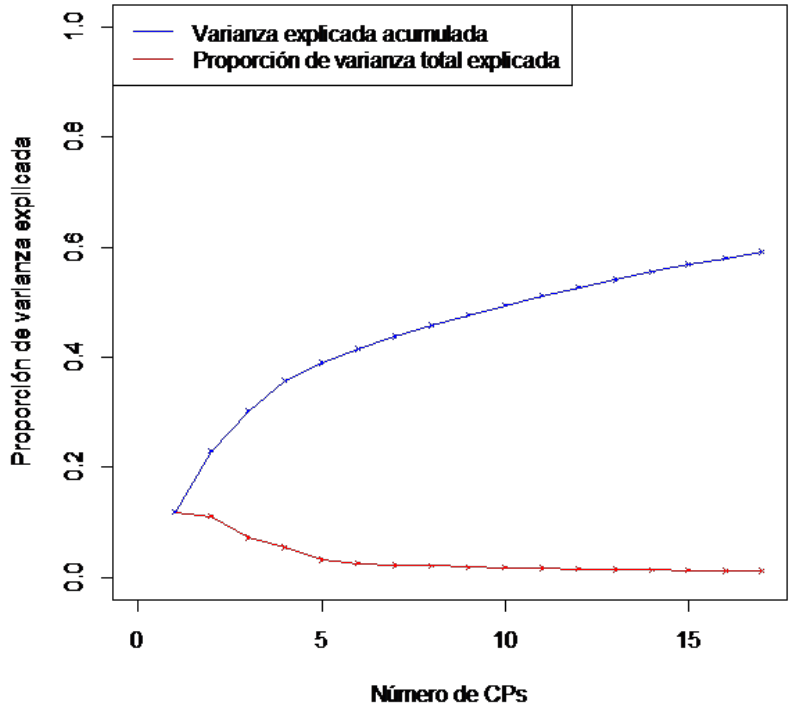


Figura 2.12: Proporción de la varianza total explicada por cada componente principal significativa según Tracy-Widon. En línea azul se representa la varianza explicada por cada eje y en rojo la proporción de la varianza explicada por cada componente significativa.

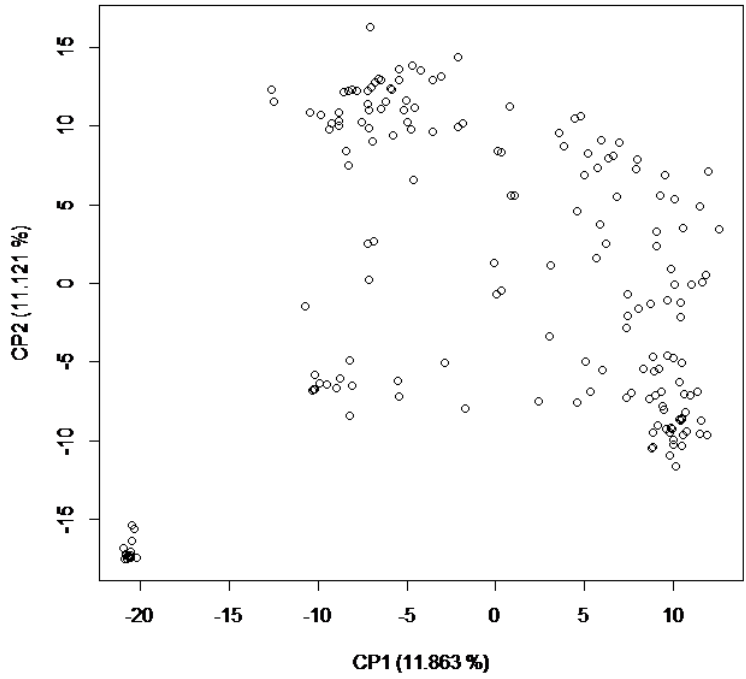


Figura 2.13: Gráfico de dispersión de las dos primeras componentes principales significativas obtenidas a partir de una Análisis de Componentes Principales sobre la información genética.

2.2.4 Modelo K

Modelo con corrección por estructura usando la matriz de parentesco

La ecuación del modelo K corresponde a la de un modelo de regresión lineal $y = X\beta + Zu + e$, donde y es el carácter fenotípico (variable respuesta) que se asocia a cada marcador molecular (variable regresora) a través de un coeficiente de regresión β que debe ser estimado desde los datos, Z es una matriz de incidencia asociada al vector u de efectos poligénicos y e representa el término de error aleatorio. Se supone que el vector u se distribuye independientemente del vector e con matriz de varianzas y covarianzas dada por $Var(u) = \sigma_g^2$ donde K es la matriz de parentesco (Kinship) obtenido mediante el paquete EMMA (Kang *et al.*, 2008) y la $Var(e) = \sigma_e^2 I$. La matriz de varianzas y covarianzas de los fenotipos es $V = \sigma_g^2 ZKZ' + {}_e^2 I$.

Pasos en *Info-Gen* para ajustar un Modelo K

Ir al Menú **Mejoramiento**, sumbenú **Mapeo Asociativo**, opción **3. Modelo K**. Se desplegarán tres opciones, una de **Datos de prueba** que permitirá abrir el archivo de ejemplo denominado *Cebada.igdb*. Antes de ajustar un modelo, es necesario tener una base de datos sobre la cual *Info-Gen* realizará los ajustes. Luego, seleccionar **Run...**, para indicar la acción de ajustar un modelo (Figura 2.14).

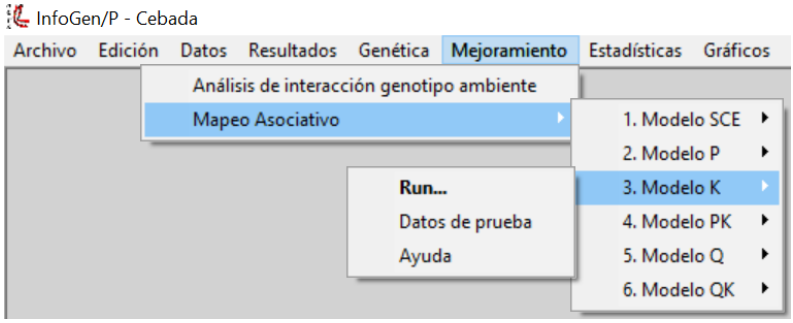


Figura 2.14: Gráfico de dispersión de las dos primeras componentes principales significativas obtenidas a partir de una Análisis de Componentes Principales sobre la información genética.

Al hacer click en el comando **Run...** se abrirá la ventana selector de variables, en la cual se listan las columnas de la base de datos activa. En la solapa Variables se visualizan tres espacios; en el primero llamado **Individuos** deberá seleccionar desde el lado izquierdo la columna que contiene la identificación de los individuos, en este ejemplo se denomina "Individuos" y con el botón (->) transportarlo al espacio de Individuos. En el segundo espacio denominado **Marcadores Moleculares** se deberan indicar las columnas del archivo que contienen información de marcadores moleculareas con los que se obtendrá la matriz de parentesco K. En el último espacio llamado **Carácter Fenotípico** deberá ingresarse la variable que contiene la información fenotípica (variable respuesta), en este archivo de ejemplo la columnas que contiene el carácter fenotípico se denomina Rendimiento. Luego seleccionar el botón **Aceptar** en la parte inferior izquierda del selector de variables (Figura 2.15).

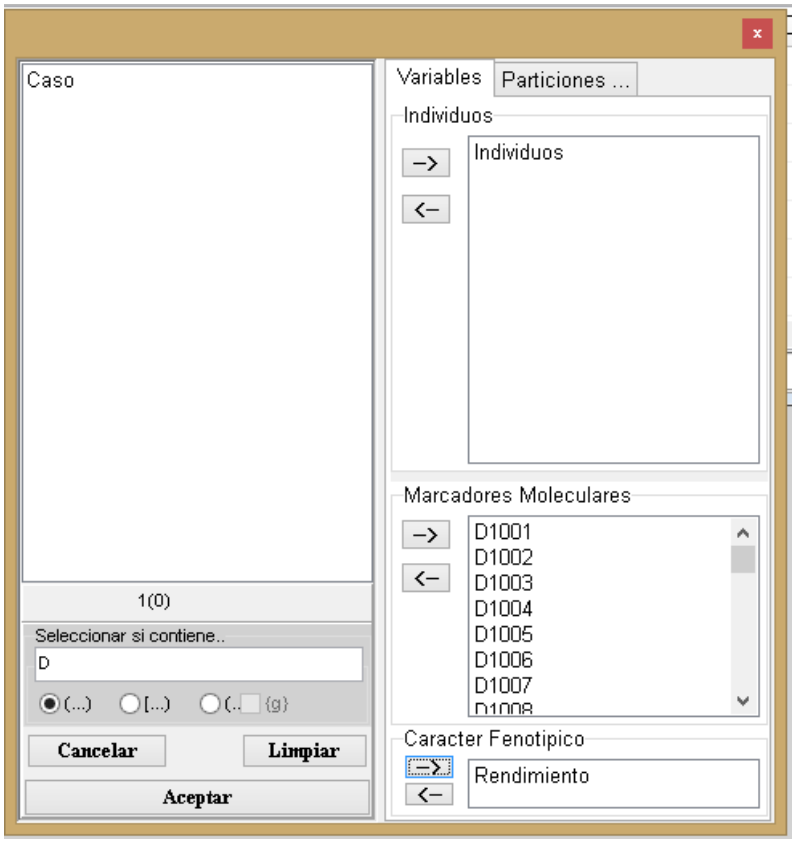


Figura 2.15: Ventana de selector de variable de Info-Gen que se despliega al seleccionar el menú Mejoramiento, submenú Mapeo Asociativo, opción 3. Modelo K. Archivo Cebada.igdb.

La ventana siguiente permite fijar el valor de α y el método de corrección por multiplicidad que permite calcular el umbral para el rechazo de la hipótesis de no asociacion (Figura 2.16). Las opciones disponibles en *Info-Gen* para calcular el umbral de significancia para disminuir la tasa de falsos positivos son: Li-Ji que estima el número efectivo de pruebas independientes por método propuesto por Li y Ji (Li y Ji, 2005) y FDR la tasa de falsos positivos. Si se selecciona $-log10(valor \quad p)$ no se realiza corrección por multiplicidad, con lo cual el umbral para la aceptación de

la hipótesis nula se estima como el logaritmo del valor p de cada prueba de hipótesis. Además del método de corrección por multiplicidad, el usuario puede seleccionar el nivel de significancia (valor de alfa) que por defecto es 0.05 pero puede ser modificado.

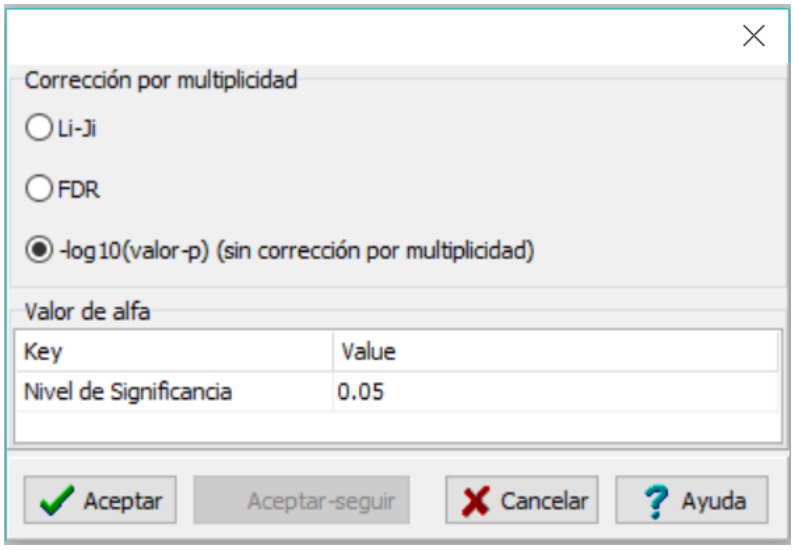


Figura 2.16: Ventana de la opción Modelo K (Menú Mejoramiento, sumbenú Mapeo asociativo) en Info-Gen.

Resultados

En la primer tabla de la ventana Resultados (Figura 2.17) se muestra una breve descripción de la **Población de Mapeo** indicando la cantidad de individuos y de marcadores moleculares usados en el análisis. La segunda tabla contiene las **Opciones seleccionadas**, *i.e.*, nivel de significancia y corrección por multiplicidad (Figura 2.16). La tercer tabla denominada **Valores p. Modelo K**, presenta los valores p, asociados a la prueba de hipótesis contrastada para cada marcador, indicando el Marcador, el Orden (o ubicación) del marcador y el valor p obtenido luego de corrección seleccionada. La tabla denominada **Marcadores**

seleccionados (Valor p< Umbral), contiene un listado de los marcadores seleccionado según el valor de α y el umbral seleccionados en las opciones de corrección por multiplicidad. Los valores de parentesco estimados pueden visualizarse en una nueva tabla de datos denominada *Matriz K.igdb* (Figura 2.17).

Población de Mapeo

Resumen	Cantidad
Número de Individuos	179
Número de Marcadores	811

Opciones seleccionadas

Opciones	Valores usados
Nivel de Sign.	0.05
Correc. por multip.	Sin corrección

Valores-p. Modelo K

Marcador	Orden	valor-p
D1001	1	0.74
D1002	2	0.18
D1003	3	0.55
D1004	4	0.55
D1005	5	0.55
.	.	.
.	.	.
.	.	.
D7131	810	0.10
D7132	811	0.48

Marcadores seleccionados (Valor-p< Umbral)

Marcador	Orden	valor-p
D1014	14	0.05
D1058	58	0.04
D1061	61	0.01
D1062	62	0.01
D1063	63	0.01
.	.	.
.	.	.
.	.	.
D7084	763	0.02
D7090	769	0.04

Buscar [Matriz Kinship] en la tabla [Matriz Kinship]

Figura 2.17: Ajuste del modelo K (Menú Mejoramiento, submenú mapeo asociativo) en Info-Gen con un nivel de significación de 0.05 y sin aplicar correción por multiplicidad (-log10(valor p)). Archivo Cebada.igdb.

En el gráfico puede observarse una línea de corte en el umbral seleccionado. Los marcadores que superan dicha línea de corte son los que fueron estadísticamente significativos indicando una asociación entre el genotipo y el fenotipo (Figura 2.18).

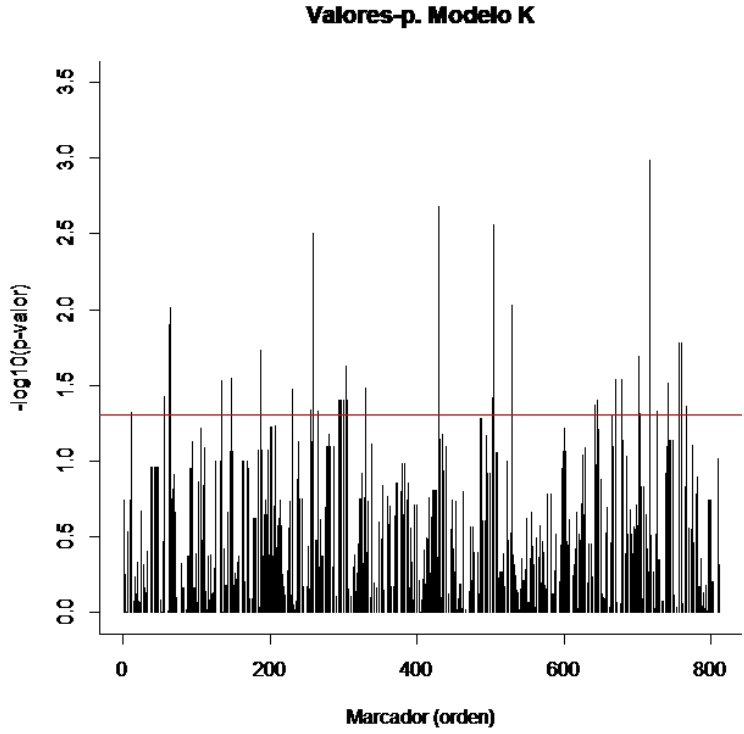


Figura 2.18: Ajuste del modelo K (Menú Mejoramiento, submenú mapeo asociativo) en Info-Gen con un nivel de significación de 0.05 y sin aplicar correción por multiplicidad (-log10(valor p)). Archivo Cebada.igdb.

2.2.5 Modelo Q

Modelo con corrección por estructura estimada a través de la matriz de relaciones genéticas

La ecuación del modelo Q corresponde a la de un modelo de regresión lineal $y = X\beta + Sv + e$, donde y es el carácter fenotípico (variable respuesta) que se asocia a cada marcador molecular (variable regresora) a través de un coeficiente de regresión β que debe ser estimado desde los datos, S es la matriz de estructura genética (matriz Q construida a partir de la probabilidad de pertenencia de cada individuo a cada subpoblación), v es el vector de efectos fijos de la estructura poblacional y e representa el término de error aleatorio.

Pasos en *Info-Gen* para ajustar un Modelo K

Ir al Menú **Mejoramiento**, sumbenú **Mapeo Asociativo**, opción **5. Modelo Q**. Se desplegarán tres opciones, una de **Datos de prueba** que permitirá abrir el archivo de ejemplo denominado *CebadaQ.igdb*. Antes de ajustar un modelo, es necesario tener una base de datos sobre la cual *Info-Gen* realizará los ajustes. El archivo *CebadaQ.igdb* contiene además de la información de los marcadores moleculares, la clasificación de los individuos y el caracter fenotípico, tres columnas que representan la matriz Q (Q1, Q2 y Q3). Luego, seleccionar **Run...**, para indicar la acción de ajustar un modelo (Figura 2.19).

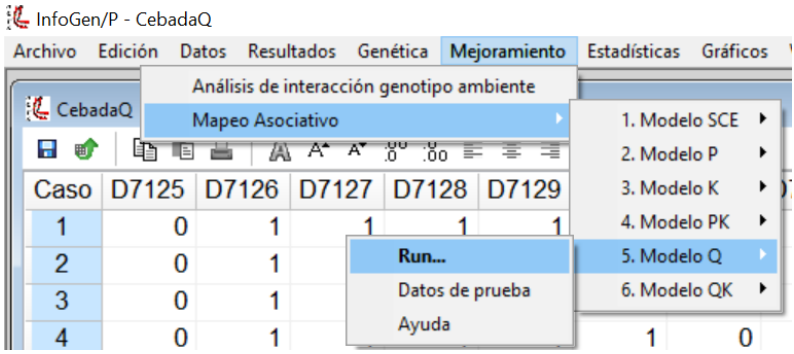


Figura 2.19: Opciones de Modelos de Mapeo Asociativo que pueden ajustarse en Info-Gen. El comando Run permite ajustar el modelo, el comando Datos de prueba abre un conjunto de datos de ejemplo, el comando Ayuda abre un tutorial.

Al hacer click en el comando **Run...** se abrirá la ventana selector de variables, en la cual se listan las columnas de la base de datos activa. En la solapa Variables se visualizan cuatro espacios; en el primero llamado **Individuos** deberá seleccionar desde el lado izquierdo la columna que contiene la identificación de los individuos, en este ejemplo se denomina "Individuos" y con el botón (->) transportarlo al espacio de Individuos, en el segundo espacio denominado **Marcadores Moleculares** se deberán indicar las columnas del archivo que contienen información de marcadores molecularеas. En el tercer espacio llamado **Matriz Q**, se deberán indicar las columnas que contienen la información de relaciones genéticas. En el último espacio llamado **Carácter Fenotípico** deberá ingresarse la variable que contiene la información fenotípica (variable respuesta), en este archivo de ejemplo la columnas que contiene el carácter fenoítpico se denomina Rendimiento. Luego seleccionar el botón **Aceptar** en la parte inferior izquierda del selector de variables (Figura 2.20).

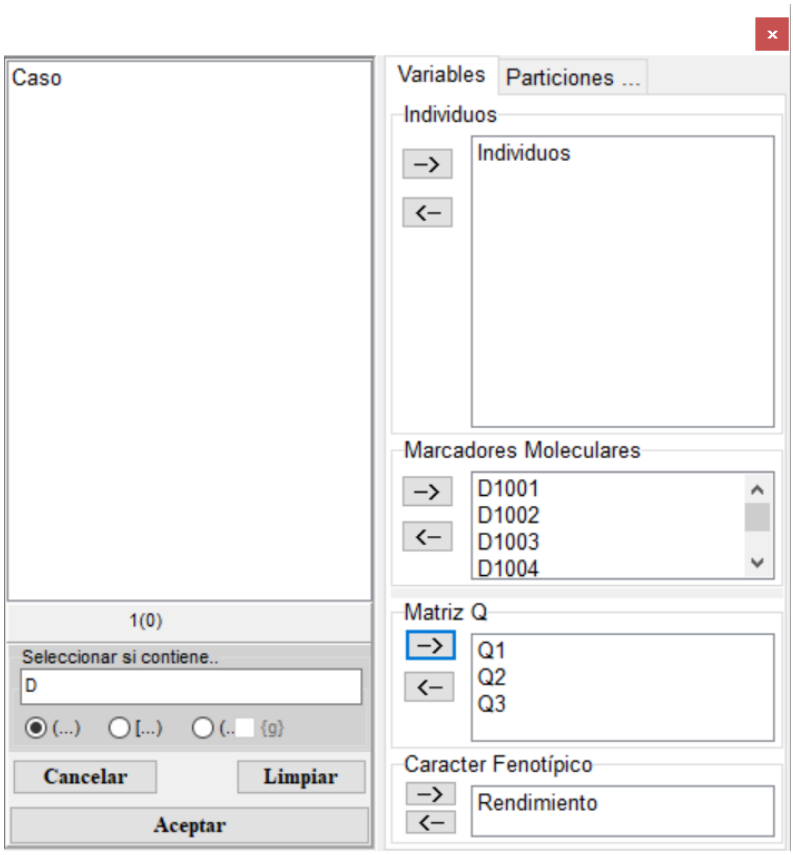


Figura 2.20: Ventana de selector de variable de Info-Gen que se despliega al seleccionar el menú Mejoramiento, submenú Mapeo Asociativo, opción 5. Modelo Q. Archivo de ejemplo CebadaQ.igdb.

La ventana siguiente permite fijar el valor de α y el método de corrección por multiplicidad que permite calcular el umbral para el rechazo de la hipótesis de no asociación (Figura 2.21). Las opciones disponibles en *Info-Gen* para calcular el umbral de significancia para disminuir la tasa de falsos positivos son: Li-Ji que estima el número efectivo de pruebas independientes por método propuesto por Li y Ji (Li y Ji, 2005) y FDR la tasa de falsos positivos. Si se selecciona $-log10(valor \quad p)$ no se realizará corrección por

multiplicidad, con lo cual el umbral para la aceptación de la hipótesis nula se estima como el logaritmo del valor p de cada prueba de hipótesis. Además del método de corrección por multiplicidad, el usuario puede seleccionar el nivel de significancia (valor de alfa) que por defecto es 0.05 pero puede ser modificado.

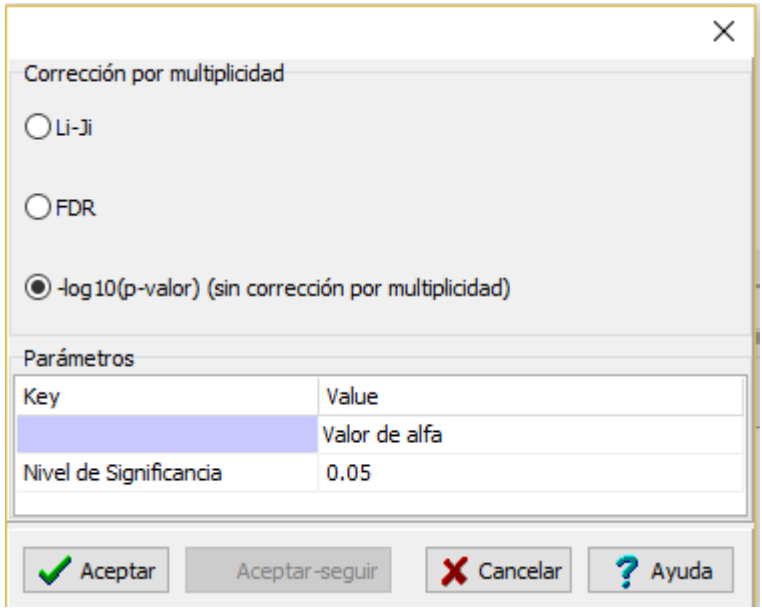


Figura 2.21: Ventana de la opción modelo Q (Menú Mejoramiento, submenú mapeo asociativo) en Info-Gen.

Resultados

En la primer tabla de la ventana Resultados (Figura 2.22) se muestra una breve descripción de la **Población de Mapeo** indicando la cantidad de individuos y de marcadores moleculares usados en el análisis. La segunda tabla contiene las **Opciones seleccionadas**, *i.e.*, nivel de significancia y corrección por multiplicidad (Figura 2.22). La tercer tabla denominada Valores p. Modelo Q, presenta los valores p, asociados a la prueba de hipótesis contrastada para cada marcador, indicando el Marcador, el Orden (o

ubicación) del marcador y el valor p obtenido luego de corrección seleccionada. La tabla denominada **Marcadores seleccionados (Valor p< Umbral)**, contiene un listado de los marcadores seleccionado según el valor de α y el umbral seleccionados en las opciones de corrección por multiplicidad (Figura 2.22).

Población de Mapeo

Resumen	Cantidad
Número de Individuos	179
Número de Marcadores	811

Opciones seleccionadas

Opciones	Valores usados
Nivel de Sign.	0.05
Correc. por multip.	Sin corrección

Valores-p. Modelo Q

Marcador	Orden	valor-p
D1001	1	0.016
D1002	2	0.068
D1003	3	0.004
D1004	4	0.004
D1005	5	0.004
.	.	.
.	.	.
.	.	.
D7131	810	0.010
D7132	811	0.001

Marcadores seleccionados (Valor-p< alfa)

Marcador	Orden	valor-p
D1001	1	0.016
D1003	3	0.004
D1004	4	0.004
D1005	5	0.004
D1010	10	0.004
.	.	.
.	.	.
.	.	.
D7131	810.00	0.010
D7132	811.00	0.001

Figura 2.22: Ajuste del modelo Q (Menú Mejoramiento submenú mapeo asociativo) en Info-Gen con un nivel de significación de 0.05 y sin aplicar correción por multiplicidad (-log10(valor p)). Archivo CebadaQ.igdb.

En el gráfico puede observarse una línea de corte en el umbral seleccionado. Los marcadores que superan dicha línea de corte son los que fueron estadísticamente significativos indicando una asociación entre el genotipo y el fenotipo (Figura 2.23).

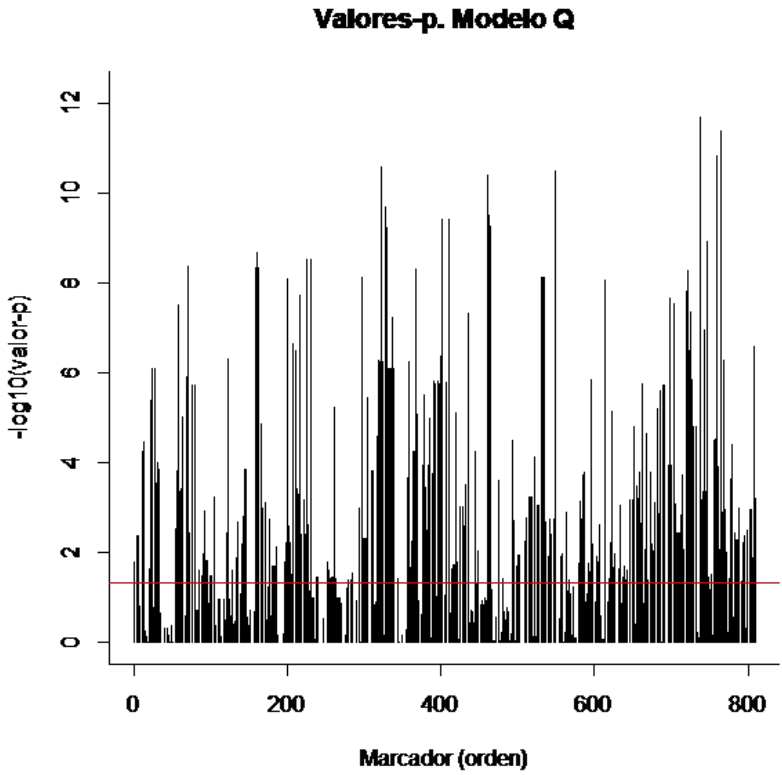


Figura 2.23: Gráfico de valores p obtenidos al ajustar un modelo Q (Menú Mejoramiento, submenú mapeo asociativo) en InfoGen con un nivel de significaciónde 0.05 y ninguna correción por multiplicidad para cada marcador. Archivo CebadaQ.igdb. La línea roja indican el umbral para determinar la significancia de la asociación, valores por encima del umbral indica que el marcador está asociado al carácter bajo estudio.

2.2.6 Modelo PK

Modelo con corrección por estructura estimada a través de un Análisis Multivariado de Componentes Principales y usando la matriz de parentesco

La ecuación del modelo PK corresponde a la de un modelo de regresión lineal $y = X\beta + Sv + Zu + e$, donde y es el carácter fenotípico (variable respuesta) que se asocia a cada marcador molecular (variable regresora) a través de un coeficiente de regresión β que debe ser estimado desde los datos, S es la matriz de estructura genética (matriz P construida con los componentes principales del Análisis de Componentes Principales (ACP) realizado sobre los datos moleculares, usualmente se seleccionan aquellas Componentes Principales estadísticamente significativas según Tracy-Widom), v es el vector de efectos fijos de la estructura poblacional, Z es una matriz de incidencia asociada al vector u de efectos poligénicos. Se supone que el vector u se distribuye independientemente del vector e que representa el término de error aleatorio. La matriz de varianzas y covarianzas de los efectos genéticos puede expresarse como $Var(u) = \sigma_g^2 K$ donde K es la matriz de parentesco (*Kinship*) obtenido mediante el paquete EMMA (Kang *et al.*, 2008) de R y $Var(e) = \sigma_e^2 I$. La matriz de varianzas y covarianzas de los fenotipos se expresa como $V = \sigma_g^2 ZKZ' + \sigma_e^2 I$.

Pasos en *Info-Gen* para ajustar un Modelo PK

Ir al Menú **Mejoramiento**, submenú **Mapeo Asociativo**, opción **4. Modelo PK**. Al seleccionar la opción **Datos de prueba** se abrirá el archivo de ejemplo Cebada.igdb. Al apretar la opción **Run...**, se indica la acción de ajustar el modelo seleccionado (Figura 2.24).

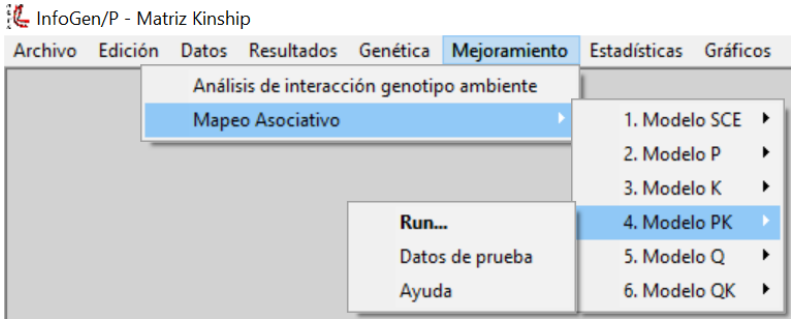


Figura 2.24: Opciones de Modelos de Mapeo Asociativo que pueden ajustarse en Info-Gen. El comando Run permite ajustar el modelo, el comando Datos de prueba abre un conjunto de datos de ejemplo, el comando Ayuda abre un tutorial.

Al hacer click en el comando **Run...** se abrirá la ventana selector de variables, en la cual se listan las columnas de la base de datos activa. En la solapa Variables se visualizan tres espacios; en el primero llamado **Individuos** deberá seleccionar desde el lado izquierdo la columna que contiene la identificación de los individuos, en este ejemplo se denomina "Individuos" y con el botón (->) transportarlo al espacio de Individuos, en el segundo espacio denominado **Marcadores Moleculares** se deberán indicar las columnas del archivo que contienen información de marcadores moleculares. En el último espacio llamado **Carácter Fenotípico** deberá ingresarse la variable que contienen la información fenotípica (variable respuesta), en este archivo de ejemplo la columnas que contiene el caracter fenotípico se denomina Rendimiento. Luego, seleccionar el botón **Aceptar** en la parte inferior izquierda del selector de variables (Figura 2.25).

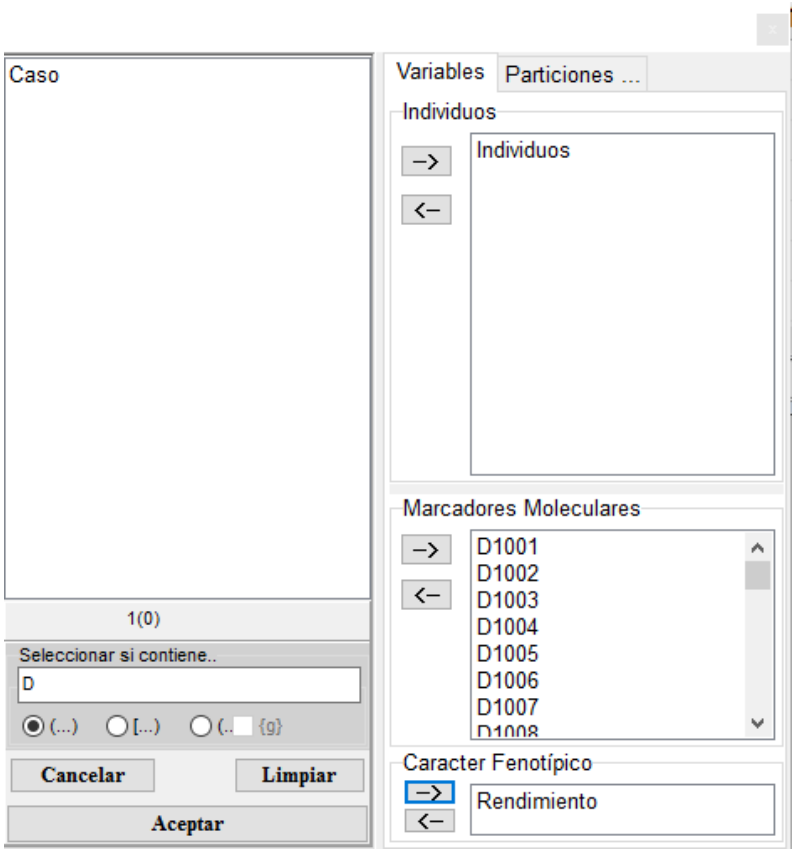


Figura 2.25: Ventana de selector de variable de Info-Gen que se despliega al seleccionar el menú Mejoramiento, submenú Mapeo Asociativo, opción 4. Modelo PK. Archivo Cebada.igdb.

La ventana siguiente permite fijar el valor de α y el método de corrección por multiplicidad que permite calcular el umbral para el rechazo de la hipótesis de no asociación (Figura 2.26). Las opciones disponibles en *Info-Gen* para calcular el umbral de significancia para disminuir la tasa de falsos positivos son: Li-Ji que estima el número efectivo de pruebas independientes por método propuesto por Li y Ji (Li y Ji, 2005) y FDR la tasa de falsos positivos. Si se selecciona $-log10(valor \quad p)$ no se realizará corrección por multiplicidad, con lo cual el umbral para la aceptación de

la hipótesis nula se estima como el logaritmo del valor p de cada prueba de hipótesis. Además del método de corrección por multiplicidad, el usuario puede seleccionar el nivel de significancia (valor de alfa) que por defecto es 0.05 pero puede ser modificado y la Cantidad de CP (componentes principales) a usar en el ajuste del Modelo P, TW es la opción por defecto, seleccionando las componentes principales significativas por Tracy-Widom.

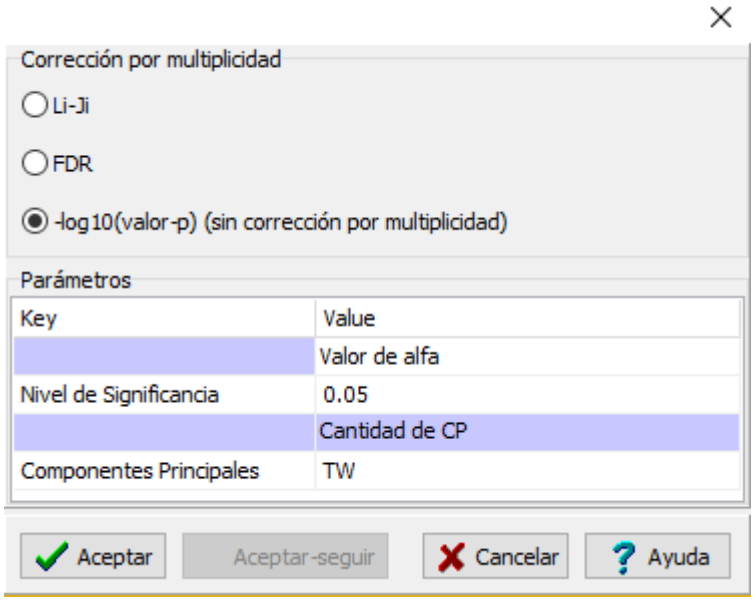


Figura 2.26: Ventana de la opción Modelo PK (Menú Mejoramiento, submenú Mapeo Asociativo) en Info-Gen.

Resultados

En la primer tabla de la ventana Resultasos (Figura 2.27) se muestra una breve descripción de la **Población de Mapeo** indicando la cantidad de individuos, de marcadores moleculares y las CP significativas según TW usados en el análisis. La segunda tabla contiene los coeficientes de las **Componentes Principales significativas (Tracy-Widom)**. La tercera tabla contiene las **Opciones selec-**

cionadas, *i.e.*, nivel de significancia, corrección por multiplicidad y cantidad de CPs. La cuarta tabla presenta los **Valores p. Modelo PK** asociados a la prueba de hipótesis contrastada para cada marcador, indicando el Marcador, el Orden del marcador (o ubicación del marcador) y el valor p obtenido luego de la corrección seleccionada. La tabla denominada **Marcadores seleccionados (Valor p< Umbral)**, contiene un listado sólo con los marcadores seleccionado según el valor de α y el umbral seleccionados en las opciones de corrección por multiplicidad (Figura 2.27).

Población de Mapeo

Resumen	Cantidad
Número de Individuos	179
Número de Marcadores	811
CP significativas (T-W)	17

Componentes Principales significativos (Tracy-Widom)

PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13	PC14	PC15	PC16	PC17
-4.60	6.56	-2.20	-5.00	0.01	-7.30	3.40	-2.52	4.00	-1.05	0.51	-2.72	6.52	-1.89	4.28	-2.7E-03	0
-6.62	13.01	-3.55	-6.53	-3.39	5.01	-2.51	4.32	-0.17	3.37	-2.88	3.80	1.75	-1.26	2.83	-0.13	-2
9.66	-4.61	2.65	1.72	0.26	2.00	0.58	-2.39	1.51	3.13	-3.50	-5.17	1.89	-0.51	-0.53	-0.61	-4
-7.23	2.53	-2.12	-0.93	8.73	-7.82	-0.12	4.15	-3.74	-2.52	-0.87	2.76	3.86	5.81	1.24	2.83	-5
-9.23	10.17	-6.27	-2.76	4.92	-8.07	6.17	-6.76	0.85	4.85	-2.16	2.36	-0.61	-1.92	-1.02	-0.77	-2
-6.91	9.00	-2.25	-0.85	1.81	-5.20	-0.85	-0.74	-0.72	2.81	3.67	-1.20	4.30	-4.65	-7.38	0.01	2

Opciones seleccionadas

Opciones	Valores usados
Nivel de Sign.	0.05
Correc. por multip.	Sin corrección
N° PCs usadas	17

Valores-p del modelo PK

Marcador	Orden	valor-p
D1001	1	0.14
D1002	2	0.03
D1003	3	0.18
D1004	4	0.18
.	.	.
.	.	.
.	.	.
D7130	809	0.09
D7131	810	0.23
D7132	811	0.38

Marcadores seleccionados (Valor-p < 0.05)

Marcador	Orden	valor-p
D1002	2	0.030
D1011	11	0.030
D1014	14	0.030
D1018	18	0.040
.	.	.
.	.	.
.	.	.
D7076	755	0.015
D7079	758	0.004
D7092	771	0.012

Buscar [Matriz Kinship] en la tabla [Matriz Kinship]

Figura 2.27: Ajuste del modelo PK (Menú Mejoramiento submenú Mapeo Asociativo) en Info-Gen con un nivel de significación de 0.05 y sin aplicar correción por multiplicidad (-log10(valor p)). Archivo Cebada.igdb.

En el gráfico puede observarse una línea de corte en el umbral seleccionado. Los marcadores que superan dicha línea de corte son los que fueron estadísticamente significativos

indicando una asociación entre el genotipo y el fenotipo (Figura 2.28).

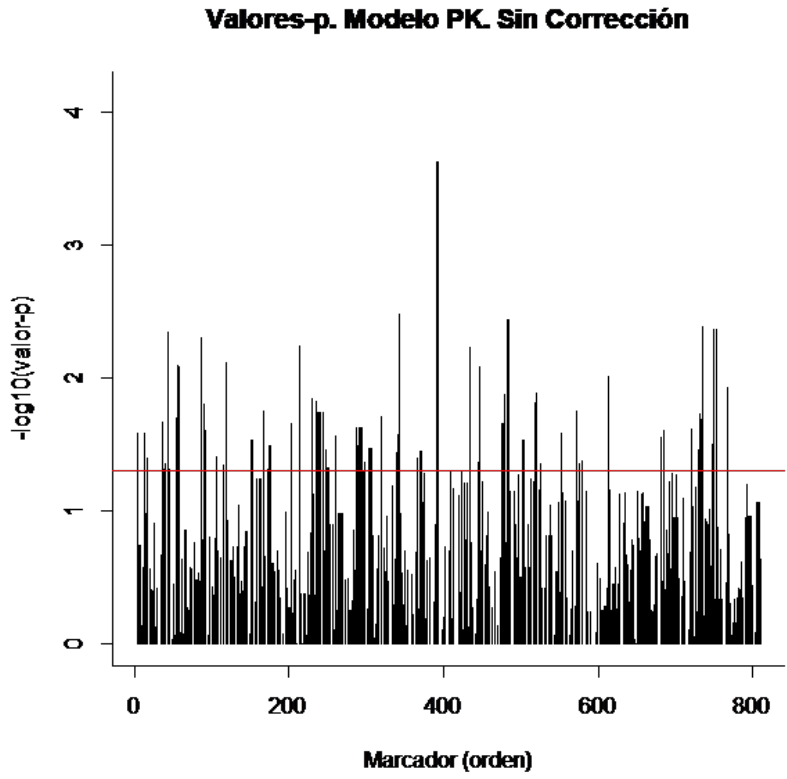


Figura 2.28: Ajuste del modelo PK (Menú Mejoramiento submenú Mapeo Asociativo) en Info-Gen con un nivel de significación de 0.05 y sin aplicar correción por multiplicidad (-log10(valor p)). Archivo Cebada.igdb.

Entre las salidas del Modelo PK se muestran dos gráficos relacionados al Análisis de Componentes Principales, realizado sobre los marcadores moleculares para capturar la estructura genética poblacional en las componentes principales significativas (CPs) según Tracy-Widom. Uno de los gráficos muestra la proporción de la varianza explicada por cada componente principal que resultó estadísticamente significativo según la propuesta de Tracy Widom. Sobre

el eje de las ordenadas se muestra la proporción de varianza explicada en función de la cantidad de componentes principales (CPs). En línea azul (descendente) se representa la varianza acumulada asociada a cada componente significativa según Tracy-Widom y en línea de color rojo la proporción de la varianza total explicada (curva ascendente). Debido a que la primer componente principal explica más variabilidad que la segunda, y la segunda más variabilidad que la tercera y así sucesivamente, la proporción de la varianza acumulada asociada a cada componente va disminuyendo (Figura 2.29). El otro gráfico es un diagrama de dispersión de las observaciones (Individuos) en el espacio de las dos primeras CP donde podrían visualizarse patrones de agrupamiento a la variabilidad entre individuos (Figura 2.30).

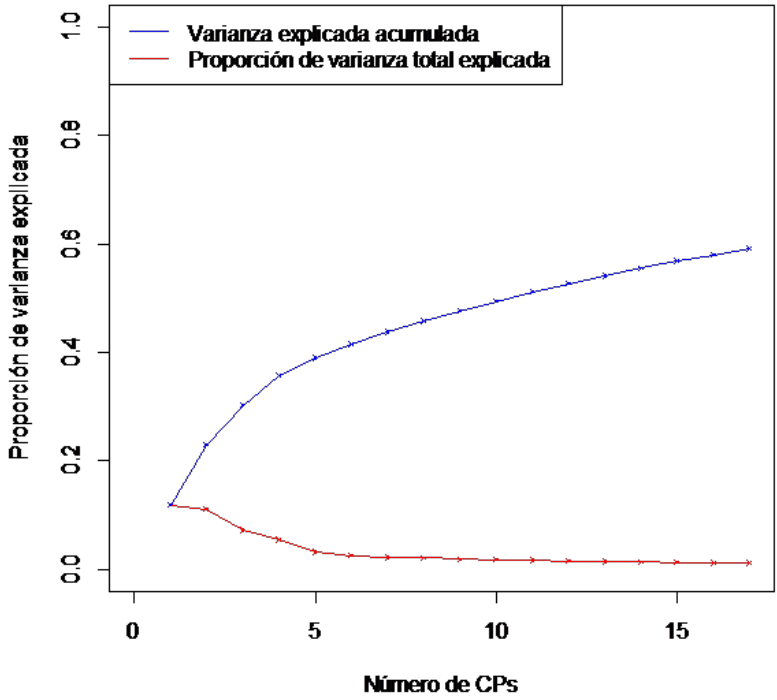


Figura 2.29: Proporción de la varianza total explicada por cada componente principal significativa según Tracy-Widon. En línea azul se representa la varianza explicada por cada eje y en rojo la proporción de la varianza explicada por cada componente significativa.

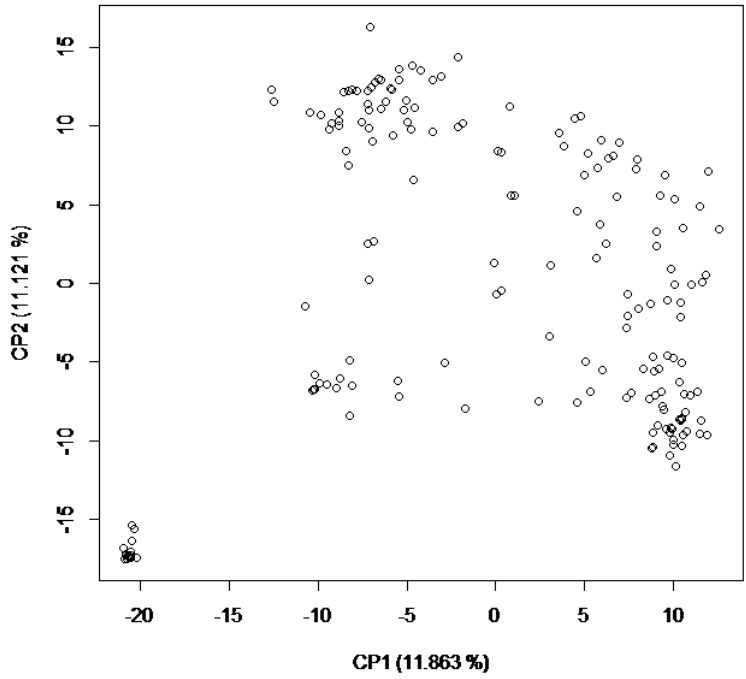


Figura 2.30: Gráfico de dispersión de las dos primeras componentes principales significativas obtenidas a partir de una Análisis de Componentes Principales sobre la información genética.

2.2.7 Modelo QK

Modelo con corrección por estructura usando la matriz de parentesco K y la matriz de relaciones genéticas Q

La ecuación del modelo QK corresponde a la de un modelo de regresión lineal, $y = X\beta + Sv + Zu + \epsilon$, donde y es el carácter fenotípico (variable respuesta) que se asocia a cada marcador molecular (variable regresora) a través de un coeficiente de regresión β que debe ser estimado desde los datos, S es la matriz de estructura genética (matriz Q que contiene la probabilidad de pertenencia de cada individuo a una subpoblación), v es el vector de efectos fijos de la estructura poblacional, Z es una matriz de incidencia asociada al vector u de efectos poligénicos. Se supone que el vector u se distribuye independientemente del vector ϵ que representa el término de error aleatorio. La matriz de varianzas y covarianzas de los efectos genéticos puede expresarse como $Var(u) = \sigma_g^2 K$ donde K es la matriz de parentesco (*Kinship*) obtenido mediante el paquete EMMA (Kang *et al.*, 2008) de R y $Var(e) = \sigma_e^2 I$. La matriz de varianzas y covarianzas de los fenotipos se expresa como $V = \sigma_g^2 ZKZ' + \sigma_e^2 I$.

Pasos en *Info-Gen* para ajustar un Modelo PK

Ir al Menú **Mejoramiento**, submenú **Mapeo Asociativo**, opción **6. Modelo QK**. Se desplegarán tres opciones, una de **Datos de prueba** que permitirá abrir el archivo de ejemplo denominado *CebadaQ.igdb*. Antes de ajustar un modelo, es necesario tener una base de datos sobre la cual *Info-Gen* realizará los ajustes. El archivo *CebadaQ.igdb* contiene además de la información de los marcadores moleculares, la clasificación de los individuos y el carácter fenotípico, tres columnas que representan la matriz Q (Q1, Q2 y Q3). Esta matriz puede obtenerse siguiendo los pasos detallados en la sección *conglomerados Bayesianos* del

Capítulo 1 de este libro, ya sea a través de Structure o con el código de R provisto. Una vez obtenida la matriz Q pueden adicionarse las columnas a la matriz que contiene la información genotípica y base de datos fenotípica usando los comandos del menú Datos de *Info-Gen* para insertar nuevas columnas y el comando Edición para pegar con nombres de columnas. Luego, seleccionar **Run...**, para indicar la acción de ajustar un modelo (Figura 2.31).

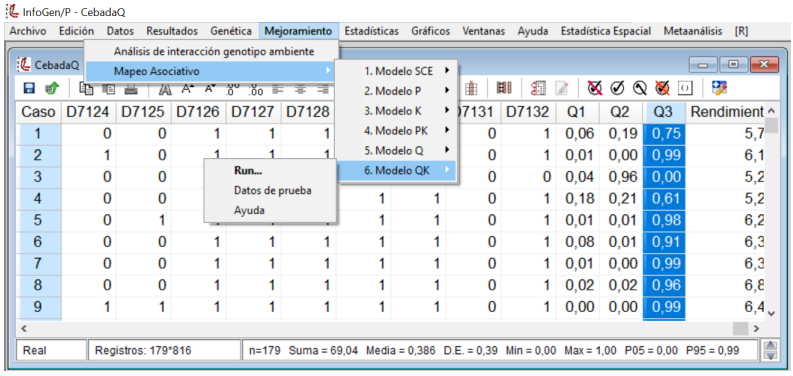


Figura 2.31: Opciones de Modelos de Mapeo Asociativo que pueden ajustarse en Info-Gen. El comando Run permite ajustar el modelo, el comando Datos de prueba abre un conjunto de datos de ejemplo, el comando Ayuda abre un tutorial.

Al hacer click en el comando **Run...** se abrirá la ventana selector de variables, en la cual se listan las columnas de la base de datos activa. En la solapa Variables se visualizan cuatro espacios; en el primero llamado **Individuos** deberá seleccionar desde el lado izquierdo la columna que contiene la identificación de los individuos, en este ejemplo se denomina "Individuos" y con el botón (->) transportarlo al espacio de Individuos, en el segundo espacio denominado **Marcadores Moleculares** se deberan indicar las columnas del archivo que contienen información de marcadores moleculareas. En el tercer espacio llamado **Matriz Q**, se deberán indicar las columnas que contienen la información

de relaciones genéticas obtenidas por Structure o R según sección 1.1.3 del Capítulo 1. En el último espacio llamado **Carácter Fenotípico** deberá ingresarse la variable que contiene la información fenotípica (variable respuesta), en este archivo de ejemplo la columnas que contiene el carácter fenoítpico se denomina Rendimiento. Luego seleccionar el botón **Aceptar** en la parte inferior izquierda del selector de variables (Figura 2.32).

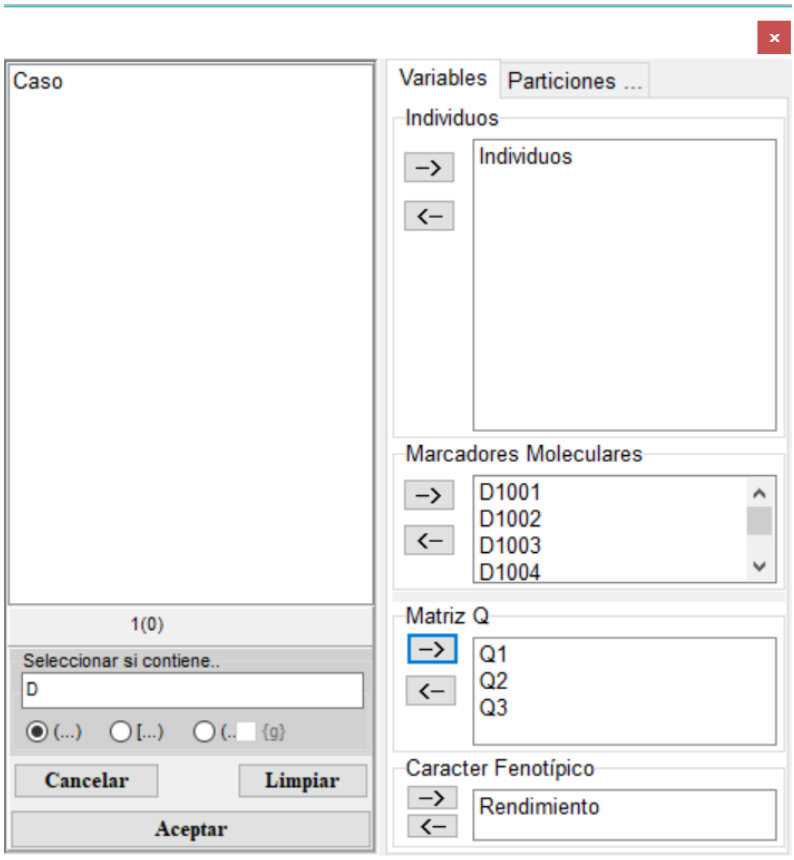


Figura 2.32: Ventana de selector de variable de Info-Gen que se despliega al seleccionar el menú Mejoramiento, submenú Mapeo Asociativo, opción 6. Modelo QK. Archivo de ejemplo CebadaQ.igdb.

La ventana siguiente permite fijar el valor de α y el método de corrección por multiplicidad que permite calcular el umbral para el rechazo de la hipótesis de no asociación (Figura 2.33). Las opciones disponibles en *Info-Gen* para calcular el umbral de significancia para disminuir la tasa de falsos-positivos son: Li-Ji que estima el número efectivo de pruebas independientes por método propuesto por Li y Ji (Li y Ji, 2005) y FDR la tasa de falsos positivos. Si se selecciona $log10(valor \quad p)$ no se realizará corrección por multiplicidad, con lo cual el umbral para la aceptación de la hipótesis nula se estima como el logaritmo del valor p de cada prueba de hipótesis. Además del método de corrección por multiplicidad, el usuario puede seleccionar el nivel de significancia (valor de α) que por defecto es 0.05 pero puede ser modificado.

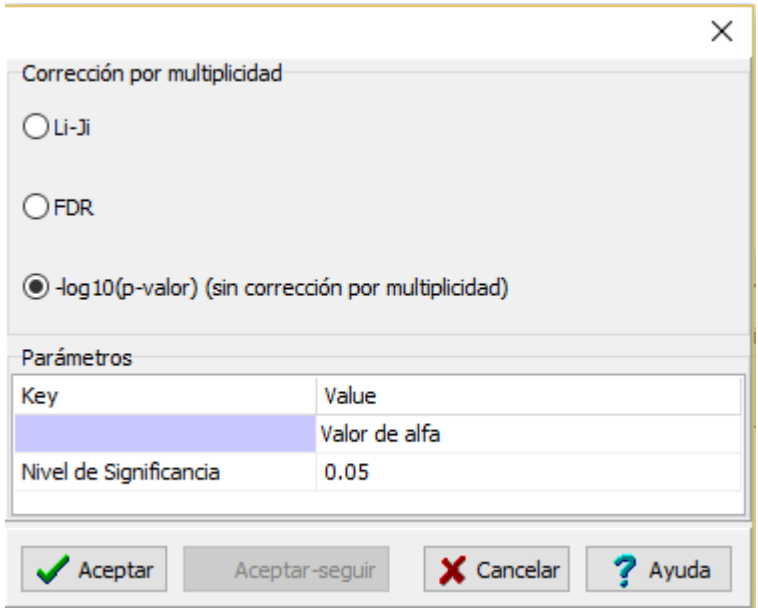


Figura 2.33: Ventana de la opción modelo QK (Menú Mejoramiento, submenú mapeo asociativo) en Info-Gen.

Resultados

En la primer tabla de la ventana Resultados (Figura 2.33) se muesstra una breve descripción de la **Población de Mapeo** indicando la cantidad de individuos y de marcadores moleculares usados en el análisis. La segunda tabla contiene las **Opciones seleccionadas**, *i.e.*, nivel de significancia y corrección por multiplicidad (Figura 2.34). La tercer tabla denominada **Valores p. Modelo QK**, presenta los valores p, asociados a la prueba de hipótesis contrastada para cada marcador, indicando el Marcador, el Orden (o ubicación) del marcador y el valor p obtenido luego de corrección seleccionada. La tabla denominada **Marcadores seleccionados** (Valor p<Umbral), contiene un listado de los marcadores seleccionado según el valor de α y el umbral seleccionados en las opciones de corrección por multiplicidad.

Población de Mapeo

Resumen	Cantidad
Número de Individuos	179
Número de Marcadores	811

Opciones seleccionadas

Opciones	Valores usados
Nivel de Sign.	0.05
Correc. por multip.	Sin corrección

Valores-p. Modelo QK

Marcador	Orden	valor-p
D1001	1	0.82
D1002	2	0.21
D1003	3	0.63
D1004	4	0.63
D1005	5	0.63
.	.	.
.	.	.
.	.	.
D7131	810	0.04
D7132	811	0.44

Marcadores seleccionados (Valor-p< alfa)

Marcador	Orden	valor-p
D1014	14	0.04
D1058	58	0.03
D1061	61	0.01
D1062	63	0.01
D1063	69	0.01
.	.	.
.	.	.
.	.	.
D7090	769	0.04
D7131	810	0.04

Buscar [Matriz Kinship] en la tabla [Matriz Kinship]

Figura 2.34: Ajuste del modelo Q (Menú Mejoramiento, submenú mapeo asociativo) en Info-Gen con un nivel de significación de 0.05 y sin aplicar correción por multiplicidad (-log10(valor p)). Archivo CebadaQ.igdb.

En el gráfico puede observarse una línea de corte en el umbral seleccionado. Los marcadores que superan dicha línea de corte son los que fueron estadísticamente significativos indicando una asociación entre el genotipo y el fenotipo (Figura 2.35).

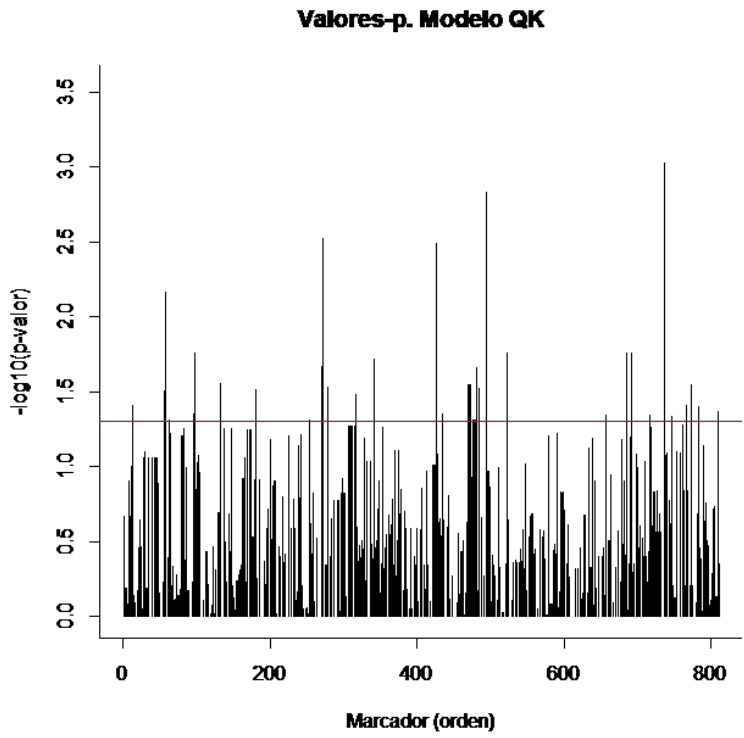


Figura 2.35: Gráfico de valores p obtenidos al ajustar un modelo QK (Menú Mejoramiento, submenú mapeo asociativo) en Info-Gen con un nivel de significación de 0.05 y ninguna correción por multiplicidad para cada marcador. Archivo CebadaQ.igdb. La línea roja indica el umbral para determinar la significancia de la asociación, valores por encima del umbral indican que el marcador está asociado al caracter bajo estudio.

Interpretación

A continuación se presenta cada uno de los gráficos de los valores p obtenidos al ajustar los distintos modelos de asociativos. La línea de corte indica el umbral para determinar la significancia de la asocaición, así, vallores p por encima de dicho umbral indican que el marcador está asociado al caracter bajo estudio. Cada modelo ajustado identifica un número diferente de marcadores moleculares estadísticamente singiciativos asociados al fenotipo. Estos valores pueden cambiar según el nivel de significanción usado y la corrección por multiplicidad seleccionado (Figura 2.36).

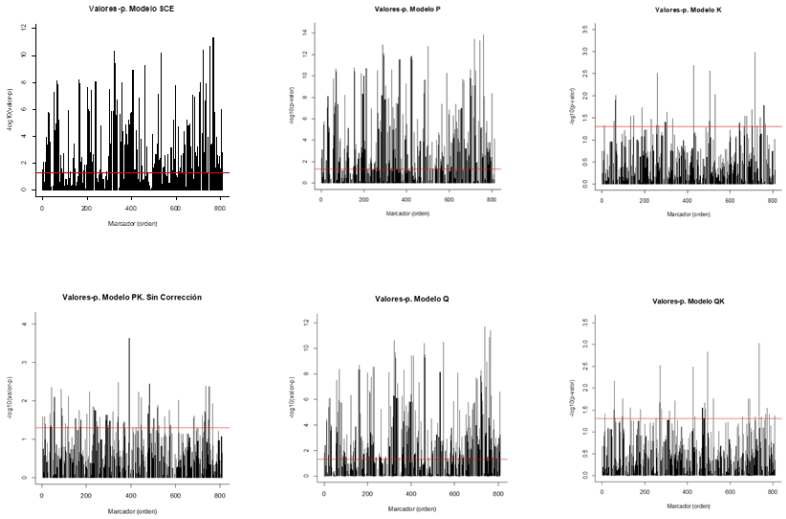


Figura 2.36: Gráfico de valores p obtenidos al ajustar cada uno de los modelos a partir del Menú Mejoramiento, submenú mapeo asociativo en Info-Gen con un nivel de significación de 0.05 y ninguna correción por multiplicidad para cada marcador. Archivo CebadaQ.igdb. La línea roja indica el umbral para determinar la significancia de la asociación, valores por encima del umbral indican que el marcador está asociado al carácter bajo estudio.

2.2.8 Implementación en la interfaz de R en Info-Gen

Preparación de los archivos de datos

Las funciones que presentamos para ajustar modelos de mapeo asociativo en R (Gutierrez *et al.*, 2016) requieren tener tres archivos de datos en formato .txt, uno que contenga la información genotípica, otro con la información fenotípica y un tercer archivo con la ubicación de cada marcador en el cromosoma. En este ejemplo los datos que trabajaremos para los modelos de mapeo de asociación son:

- QAssociation_geno.txt: posee información de 179 individuos y 810 marcadores. La información genotípica está expresada como presencia/ausencia (Tabla 2.1).

Tabla 2.1: Formato del archivo de extensión .txt con la información genotípica codificada con presencia=1 y ausencia=0. La primer fila contiene los nombres de los marcadores, la primer columna contiene información de los genotipos o individuos (notar que no lleva nombre dicha columna).

	M1	M2	M3	M4	M5	M6	M7	M8	M9
GENO_001	1	0	0	0	1	1	0	1	0
GENO_003	1	1	0	1	1	0	1	0	1
GENO_004	0	1	1	1	0	1	0	1	1
GENO_005	1	0	0	1	0	0	0	0	1

- QAssociation_pheno.txt: posee información de los individuos (genotipos), del grupo y del carácter fenotípico (yield) (Tabla 2.2).

Tabla 2.2: Formato del archivo de extensión .txt con la información fenotípica (caracteres morfológicos) para cada individuo. La primer columna indica el nombre de los genotipos, es un factor de clasificación, la segunda columna contiene información del grupo al que fueron asignados los individuos. Se puede adicionar una o más columnas indicando el grupo al que pertenecen o la probabilidad de pertenencia a un grupo. Las columnas siguientes contienen la información fenotípica, puede colocarse una columna por cada carácter fenotípico evaluado.

Genotipo	Grupo	Caracter 1	Caracter 2	Caracter 3
GENO_001	3	5,77	3,44	4,34
GENO_003	3	6,12	3,97	4,17
GENO_004	4	5,25	2,84	3,36

- Qassociation_map.txt: posee información respecto a la ubicación de cada marcador en cada cromosoma, grupo de ligamiento y posición dentro del grupo de ligamiento (Tabla 2.3).

Tabla 2.3: Formato del archivo de extensión .txt con la información de la ubicación de los marcadores (mapa). No lleva encabezado, la primer columna contiene los nombres de los marcadores moleculares, la segunda columna el cromosoma (o grupo de ligamiento) en el cuál ha sido mapeado el marcador y la tercer columna la posición del marcador dentro del cromosoma o grupo de ligamiento.

M1	3	4,07
M2	3	10,70
GM3	4	10,70

Códigos en R

Las funciones que se usaran para el análisis de mapeo deben cargarse previo al ajuste de los modelos, para ello puede seguir los siguientes pasos.

1. Descargar desde http://www.agro.unc.edu.ar/~estadisticaaplicada la carpeta denominada Mapeo al cual podrá acceder desde la sección Mapeo Asociativo. La carpeta Mapeo contiene los códigos de las funciones, los archivos de datos de ejemplo y las funciones para ajustar los modelos del análisis de mapeo asociativo (Gutierrez et al., 2016)
2. Abrir *Info-Gen*, luego el intérprete de R.
3. Abrir el script denominado source.R, que se encuentra en la carpeta Mapeo descargada desde http://www.agro.unc.edu.ar/~estadisticaaplicada. Al correr el script completo, se cargaran las funciones.
4. La primer fila del script con la función setwd() cambia el directorio de trabajo. Sugerimos colocar el directorio donde han guardaro en su máquina la carpeta que contiene los script descargados desde http://www.agro.unc.edu.ar/~estadisticaaplicada. Para indicar el directorio de trabajo debe usarse la barra / y no \.
5. Es posible que sea necesario instalar una serie de paquetes, para ello puede correr el script que se llama instala.paquetes.R que se encuentra en la carpeta Mapeo http://www.agro.unc.edu.ar/~estadisticaaplicada
6. Para realizar el análisis de Mapeo de Asociación, luego de cargar las fuentes requeridas, ir al menú del intérprete de R y abrir el Script denominado

Mapeo. Se abrirá una nueva solapa para cargar los compandos del script Mapeo.R que contiene las funciones para leer los archivos de datos, describir los datos, realizar gráficos, ajustar los modelos y opciones para realizar correcciones por multiplicidad. A continuación se presenta este código:

```
# Lee los datos desde los archivos de
# extensión .txt y los compila juntos creando
#  un nuevo objeto
qtl.data <-
  load.data(
    P.file = "QAssociation_pheno.txt",
    G.file = "QAssociation_geno.txt",
    map.file = "QAssociation_map.txt",
    cross = "am",  heterozygotes = "FALSE")

#Descripción de los datos
summary(qtl.data)

# Análisis de diagnóstico del fenotipo
qtl.diagnostics(
  file = qtl.data, boxplot = "TRUE",
  qqplot = "TRUE", scatterplot = "TRUE",
  plotorder = "TRUE")

#Muestra el mapa genético
marker.dist(file = qtl.data, chr = "ALL",
  markernames = FALSE,comparison = "TRUE")

#Permite ver si tenemos datos faltantes
# y la composición genotípica
geno.plots(qtl.data)
```

```
#Lista datos faltantes de marcadores y genotipos
#y marcadores con potencial de distorsión de
#segregación
summary.markers(qtl.data, p.val = 0.01)
list.problems(qtl.data,threshold = "FALSE",
              quant = 0.001)

#Identifica problemas potenciales
LD.plots(file = qtl.data, structure = "FALSE",
         heterocigotes = "TRUE")

#Modelo sin corrección por estructura
(naive.am <- am(file = qtl.data,method = "naive",
      provide.K = FALSE, covariates = FALSE ,
      trait = "yield", threshold = 2, p = 0.05,
      out.file ="AM naïvemodel"))$selected

#Gráfico de los resultados
p.file <- naive.am$p.val
xyplot(-log10(p.file[, 3]) ~ p.file[, 2] |
    factor(p.file[, 1]),
  type = "h",layout = c(length(unique(
    p.file[, 1])), 1), col = "red",
  xlab = "Posición del Cromosoma",
  ylab = " -log10(P)",
  main = "Modelo sin corrección por
  estructura", scales = list(x = "free"),
  ylim = c(0, (max(-log10( p.file[, 3]
  )) + 0.5)))

write.table( p.file, file = paste("Modelo sin
               correccion", ".txt",sep = ""),
  append = TRUE, row.names = FALSE,
```

```
  col.names = FALSE,quote = FALSE)

# Realiza el PCA y selecciona las componentess
# significativas por Tracy-Widom
pca <- pca.analysis(file = qtl.data,
                    p.val = 0.05)

#Modelo PCA con efecto fijo
(pca.am <-am(file = qtl.data, method = "fixed",
      provide.K = FALSE,
      covariates = pca$scores,trait = "yield",
      threshold = 2,p = 0.05,
      out.file ="AM fixed PCA modef"))$selected

#Gráfico de los resultados
p.file <- pca.am$p.val
xyplot(-log10(p.file[, 3]) ~p.file[, 2] |
    factor(p.file[, 1]), type = "h",
  layout = c(length(unique(p.file[, 1])), 1),
  col = "red",xlab ="Posición del Cromosoma",
  ylab = "-log10(P)",main =
"Modelo con corrección PCA fijo",
  scales = list(x = "free"),
ylim = c(0, (max(-log10(p.file[, 3])) + 0.5)))

write.table(p.file, file = paste(
  "Modelo PCA fijo", ".txt", sep = ""),
  append = TRUE, row.names = FALSE,
  col.names = FALSE, quote = FALSE)

#Modelo PCA con efectos
#aleatorios
(pcaR.am <-am(file = qtl.data,method =
```

```
                "mixed.random",
      provide.K = FALSE,covariates = pca$scores,
      trait = "yield", threshold = 2,p = 0.05,
      out.file = "Modelo con corrección "
    ))$selected

#Gráfico de los resultados
p.file <- pcaR.am$p.val
xyplot(-log10(p.file[, 3]) ~ p.file[, 2] |
    factor(p.file[, 1]),
  type = "h",layout = c(length(unique(
    p.file[, 1])), 1),
  col = "red",xlab ="Posición del Cromosoma",
  ylab = "-log10(P)",
  main = "Modelo con corrección PCA aleatorio",
  scales = list(x = "free"),
  ylim = c(0, (max(-log10(p.file[, 3]))
               + 0.5)))

write.table(p.file,file = paste("Modelo PCA
                                aleatorio",
    ".txt", sep = ""), append = TRUE,
  row.names = FALSE,col.names = FALSE,
  quote = FALSE)

#Mapeo asociativo con
#corrección por parentesco
(k.am <- am(file = qtl.data,method =
      "mixed.nostructure",
    provide.K = FALSE,covariates = FALSE,
    trait = "yield", threshold = 2,p = 0.05,
    out.file ="AM mixed model nocovariate"
  ))$selected
```

```
# Gráfico de los resultados
p.file <- k.am$p.val
xyplot(-log10(p.file[, 3]) ~ p.file[, 2] |
    factor(p.file[, 1]), type = "h",
  layout = c(length(unique(p.file[, 1])), 1),
  col = "red", xlab =
  " Posición del Cromosoma ",
  ylab = "-log10(P)",
  main ="Modelo con corrección por parentesco",
  scales = list(x = "free"),
  ylim = c(0,(max(-log10(p.file[, 3])) + 0.5)))

write.table(p.file, file = paste("Modelocon
correccion por parentesco", ".txt",
  sep = ""), append = TRUE,row.names = FALSE,
  col.names = FALSE,quote = FALSE)


#Identificación de la
#covariable de estructura
covariate <-(read.table(
  file ="QAssociation_pheno.txt ",
  header = TRUE))[, 2]

#Mapeo asociativo con
#corrección por estructura
(g.am <-am(file = qtl.data,method = "fixed",
      provide.K = FALSE,covariates =
        covariate,trait = "yield",
      threshold = 2,p = 0.05,out.file = "AM
fixed Groups model"))$selected

# Gráfico de los resultados
p.file <- g.am$p.val
```

```
xyplot(-log10(p.file[, 3]) ~p.file[, 2] |
    factor(p.file[, 1]),type = "h",
  layout = c(length(unique(p.file[, 1])), 1),
  col = "red", xlab = "Chromosome position",
  ylab = "-log10(P)",main = "Association
mapping groups fixed",
scales = list(x = "free"),
ylim = c(0,(max(-log10(p.file[,3])) + 0.5)))

write.table(p.file,file = paste("Modelo
                covariable grupo",
                ".txt", sep = ""),
  append = TRUE,row.names = FALSE,
  col.names = FALSE,quote = FALSE)

# Correccion por multiplicidad


# Correccion por Li&Ji
# (Li y Ji, 2008)
(naive.am <-am(file = qtl.data,
               method = "naive",
      provide.K = FALSE,
      covariates = FALSE,
      trait = "yield",
      threshold = "Li&Ji", p = 0.05,
      out.file = "AM naive model"))
$selected


# Correccion por "FDR" -
# false discovery rate
# (Benjamini y Hochberg, 1995)
```

```
(naive.am <-am(file = qtl.data, method = "naive",
      provide.K = FALSE, covariates = FALSE,
      trait = "yield", threshold = "FDR",
      p = 0.05,out.file = "AM naive model")
)$selected

# otras correcciones por
# multiplicidad

# Cargar datos
p_valor <-read.table(file = paste("Modelo
sin correcion", ".txt",sep = ""), sep = "",
    header = FALSE)
adjp = NULL

pvBH <- p.adjust(p_valor[, 3],method = "BH",
  n = row(p_valor))
pvBo <- p.adjust(p_valor[, 3],
                 method = "bonferroni",
  n = row(p_valor))
pvBY <- p.adjust( p_valor[, 3],method = "BY",
  n = row(p_valor))

adjp <- cbind(p_valor,pvBH, pvBo,pvBY)

write.table(adjp,file = paste("correcciones",
    ".txt", sep = ""), quote = FALSE,
  row.names = FALSE)
```

3

Instructivo para instalar Info-Gen y R

Info-Gen es un software para análisis de datos genéticos desarrollado por docentes investigadores de Estadística y Biometría de la Facultad de Ciencias Agropecuarias de la Universidad Nacional de Córdoba (Balzarini y Di Rienzo, 2004). Se encuentra disponible en www.info-gen.com.ar.

Info-Gen puede conectarse con el programa *R* a través de una interface. Para ello es necesario que el usuario tenga instalado *Info-Gen* (`http://www.info-gen.com.ar/`) y *R* (`http://www.r-project.org/`). Para usar *Info-Gen* son necesarias dos acciones:

- **instalación** del software ejecutando el instalador infogeninstaller.exe y
- **activación** de *Info-Gen* con una Clave de Activación

Si *Info-Gen* no es activado a través de su Clave de Activación (para versión Estudiantil o Profesional) funciona como una versión libre.

¿Cómo instalar *Info-Gen*?

Desde la página de *Info-Gen*, http://www.info-gen.com.ar/, el usuario puede descargar el instalador del software. El instalador es un archivo ejecutable de la versión actualizada a la fecha denominado *infogeninstaller.exe*. Recomendamos guardar el instalador en cualquier lugar de su disco y luego ejecutarlo. Si Ud. lo desea puede ejecutarlo directamente sin guardar previamente el instalador.

Para instalar *Info-Gen* debe hacer doble click en el ejecutable (*infogeninstaller.exe*) aparecerán ventanas de dialogo que lo guiarán hacia la instalación del mismo para ello por favor realizar click en el botón denominado **Next** y luego en **Finish** para terminar la instalación. *Info-Gen* se instala en el disco *C:\Archivos de Programas\InfoGen*. Automáticamente se creará un ícono de acceso directo en el escritorio de su computadora esto indica que Ud. ha instalado correctamente el software.

¿Cómo activar *Info-Gen*?

Al abrir *Info-Gen* por primera vez el mismo le pedirá una Clave de Activación. Actualmente *Info-Gen* cuenta con dos tipos de licencias: Profesional y Estudiantil. La Licencia Estudiantil es libre y cuenta con todas las herramientas estadísticas de la Licencia Profesional. Para solicitar una Clave de Activación Estudiantil deberá ingresar a la página y seleccionar la solapa Versión Estudiantil sobre el margen superior derecho de la misma. Aparecerá un formulario que le solicita los datos personales. La Clave de activación será enviada automáticamente al correo electrónico registrado.

¿Cómo hacer para usar el software *R* desde la interfaz de *Info-Gen*?

R Project, más conocido como *R*, es un lenguaje de programación que ha sido desarrollado principalmente para análisis estadístico. Es un lenguaje libre que permite generar algoritmos (conjunto de instrucciones) para procesar datos. *Info-Gen* tiene incorporada una interfaz que permite ejecutar éstos algoritmos de manera más sencilla, ya que facilita el manejo de los objetos involucrados en el procesamiento de los mismos.

R es un **lenguaje orientado a objetos**. Esto significa que R lee, genera y trabaja sobre **objetos** que el mismo software crea o que lee desde otros ambientes. El ambiente o entorno de trabajo es aquel en el que se incluyen todos los objetos relacionados con un trabajo específico. Hemos implementado una versión de *Info-Gen* que utiliza tecnología propia para vincularse con R. A continuación se describen los pasos necesarios para que R funcione bajo la aplicación de *Info-Gen*.

1. Si tiene una actualización pendiente de Windows, actualice, reinicie su computadora y luego continúe con este procedimiento.
2. Instalar la última versión de *R* desde este link `https://cran.r-project.org/bin/windows/base/R-3.6.1-win.exe` con TODAS las opciones por defecto.

- Si su máquina es de 32 bits instale *R.3.4.4* desde este link `https://cran.r-project.org/bin/windows/base/old/3.4.4/R-3.4.4-win.exe`.
- Si el sistema operativo de su computadora es Windows 7 debe instalar previamente *Microsoft .NET Frameworke*

4.7.2. Lo puede descargar desde este link `http://go.microsoft.com/fwlink/?linkid=863265`.

3. Descargar el instalador de *Info-Gen* desde el sitio web de la aplicación (www.info-gen.com.ar)
4. Instalar la aplicación desde el instalador descargado en 3.
5. Cuando la instalación termina. La aplicación se abre en modo administrador. Si todo funcionó bien, la barra de menú debería mostrar una [R] (Figura 3.1).

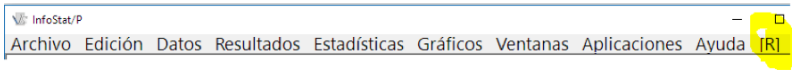


Figura 3.1: Barra de Menú de Info-Gen.

6. Sin salir de la Aplicación ir al menú **Ayuda** y seleccionar **Instalar librerías especializadas**. Este menú tiene dos etapas. La primera descomprime los paquetes de *R* zipeadas y las instala. La segunda etapa verifica que todos los paquetes se hayan instalado correctamente (para ello necesita Internet). Si algún componente falta, lo va a descargar e instalar. Darle tiempo. Si Ud. ha cambia la versión de *R* (por ejemplo de *R-3.6.0* a *R-3.6.1*) debe actualizar las librerías instaladas en sus sistema (ver menú *R* para esa opción). Para que los paquetes se instales en *C:/Archivos de Programas/R/library* debe correr *Info-Gen* en modo administrador, este modo habilita la escritura en dicho directorio. De lo contrario se instalarán en en una carpeta pública, usualmente *C:/user/documents/winR/R/library*

Introducción al manejo de *Info-Gen*

Posee un ambiente amigable de trabajo, tanto para la lectura de datos, edición de los mismos, análisis y obtención de gráficos. La visualización y manejo de los datos es muy similar a un archivo de excel (Figura 3.2).

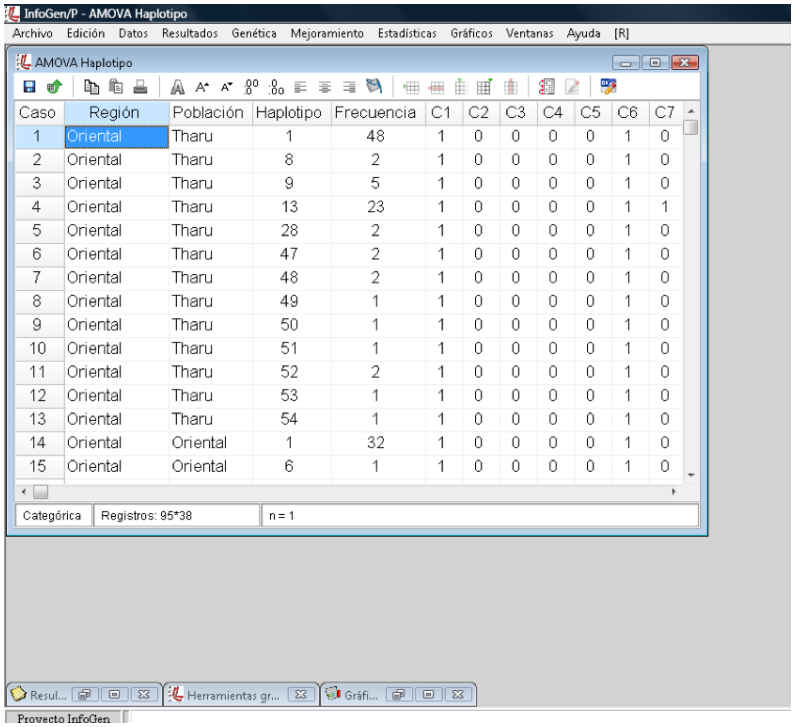


Figura 3.2: Ambiente de trabajo en Info-Gen.

Posee facilidades para la lectura de datos provenientes de marcadores moleculares dominantes y codominantes. Las bases de datos son guardadas con la extensión *.igdb*. Dispone de bases de datos de prueba que ejemplifica cada uno de los análisis que realiza el programa (Figura 3.3). La carpeta que contiene las bases de datos de prueba se encuentra en el C:\Archivos de Programas(x86)\InfoGen\Datos.

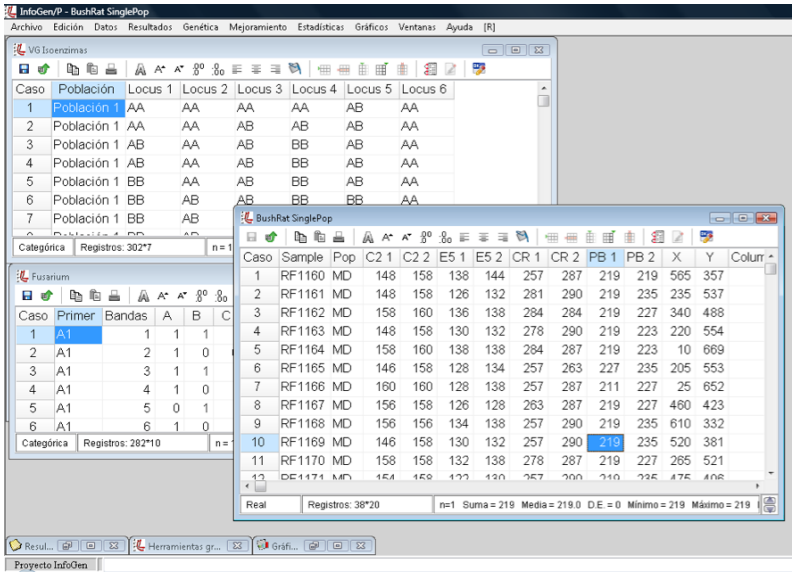


Figura 3.3: Bases de datos de ejemplos disponibles desde el Menú Archivo, opción Abrir Datos de Prueba.

Dispone de un menú específico para análisis de datos genéticos (Figura 3.4).

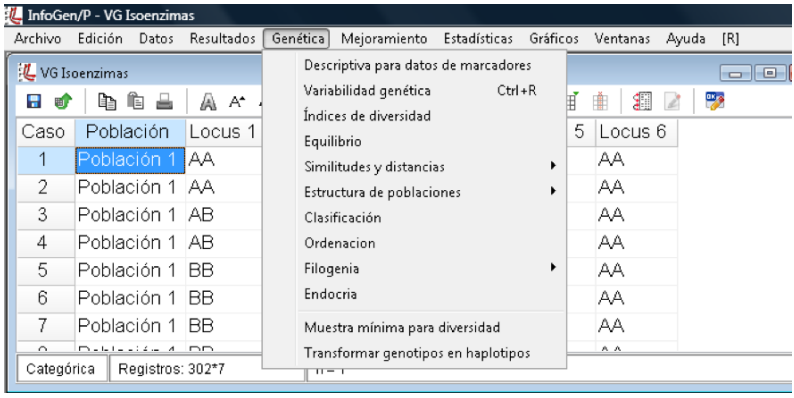


Figura 3.4: Menú Genética de Info-Gen., opciones de los análisis para datos genéticos.

Los resultados son visibles en el mismo entorno de trabajo y pueden guardase con la extensión *.itres* (Figura 3.5).

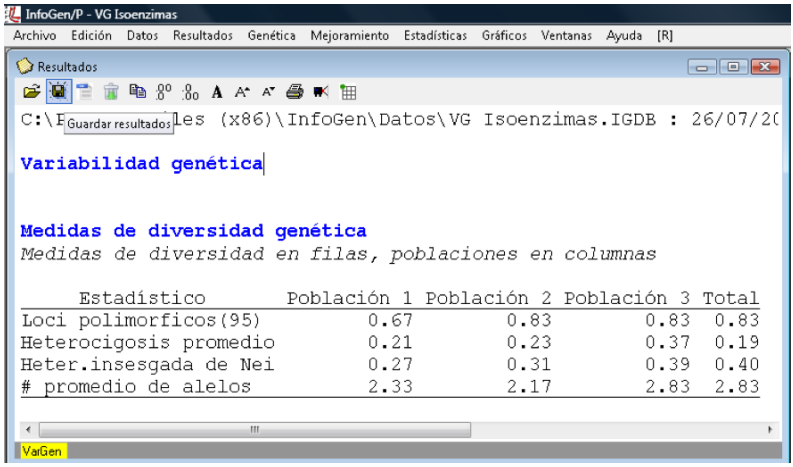


Figura 3.5: Ventana Resultados de Info-Gen.

Posee un menú de gráficos, los mismos pueden ser editados a través de herramientas gráficas de manera sencilla, así como guardar los gráficos para futuras ediciones en formato *.igb*(Figura 3.6).

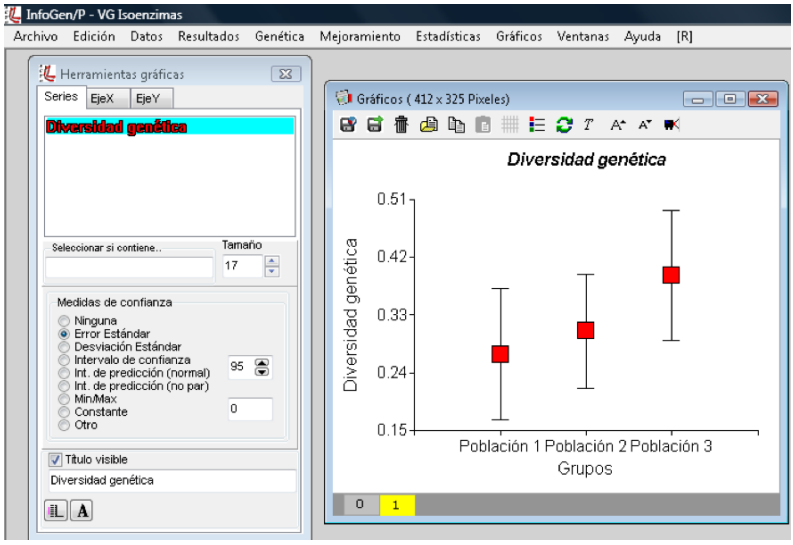


Figura 3.6: Ventanas Gráficos y Herramientas Gráficas de Info-Gen.

Con el mouse sobre el gráfico, apretando el botón derecho se despliega un menú de opciones gráficas. La opción guardar como permite seleccionar diferentes formatos, como por ejemplo *.jpg* o *.png* (Figura 3.7).

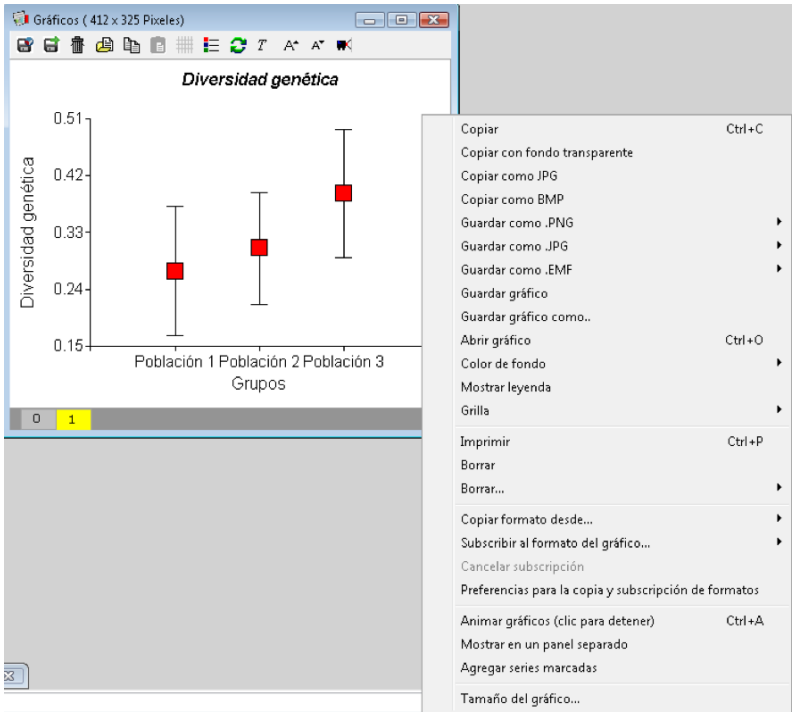


Figura 3.7: Opciones que se despliegan al apretar el botón derecho del mouse sobre el gráfico.

El Menú Genética, opción "Clasificación" de *Info-Gen* ofrece la oportunidad de implementar análisis de conglomerados jerárquicos y no jerarquicos para agrupar objetos descriptos por un conjunto de valores de varias variables. Los objetos generalmente representan las filas de la tabla de datos. Ocasionalmente, estos procedimientos son usados para agrupar variables en lugar de observaciones (es decir conglomerar columnas en lugar de filas). La ventana "selector de variables" permite seleccionar las variables

del archivo que se usarán en el análisis e indicar una o más variables como criterio de clasificación con el objetivo de resumir varios registros en un único caso. La ventana llamada *Análisis de conglomerados* la cual tiene tres solapas: *Jerárquicos*, *No jerárquicos* y *Medidas resumen*. Esta última es útil en caso que se haya indicado un criterio de clasificación de registros ya que se podrá escoger la medida de resumen con la que se sintetizaran todos los registros de cada nivel del criterio de clasificación, por ejemplo resumir registros del mismo individuo a través de ambientes. En la solapa *Jerárquicos* y *No jerárquicos*, se puede elegir el método (por defecto se selecciona automáticamente el agrupamiento promedio entre los jerárquicos o *K-means* como algoritmo no jerárquico) y el tipo de distancia (por defecto Euclídea promedio) a utilizar en la conformación de conglomerados. Tanto para los conglomerados no jerárquicos como para los jerárquicos, cuando se está agrupando casos (conglomerar filas) o variables (conglomerar columnas), mediante la activación del casillero "Guardar clasificación", *Info-Gen* genera una nueva columna en la tabla de datos activa que contiene la designación del número de grupo al que fue asignada cada observación. El número de grupos debe ser especificado de antemano en el casillero "Número de conglomerados".

3.1 Introducción a la interfaz de *Info-Gen* con *R*

Info-Gen también proporciona un intérprete de *R* que permite realizar análisis de conglomerados a través de script programados en *lenguaje R*. Para realizar los análisis de este modo, el primer paso es abrir el software *Info-Gen*, ir a menú Archivo opción "Abrir tabla". Se abrirá un selector para buscar la ubicación del archivo sobre el que se quiere realizar el análisis. Luego, ir al menú [R] y se abrirá el intérprete de *R* en *Info-Gen* (Figura 3.8). Para cargar la tabla activa como un data frame, operable en *R*, ir al primer ícono de acceso directo de la barra de herramientas de la ventana *Objects*. Para cargar un *Script* ir al segundo ícono de acceso directo que se encuentra en la barra de herramientas del intérprete de *R* y seleccionar el script deseado (Figura 3.9). Teniendo cargado el archivo de datos como un objeto de *R* y el script o código de instrucciones para *R*, será posible ejecutar los comandos usado el botón con ícono triangular. Mayores indicaciones sobre la instalación de *Info-Gen* y de *R*, así como del uso del intérprete de *R* en *Info-Gen* pueden accederse desde www.info-gen.com.ar.

Para acceder al ambiente de trabajo de *R* en *Info-Gen* hacer click sobre el menú [R].

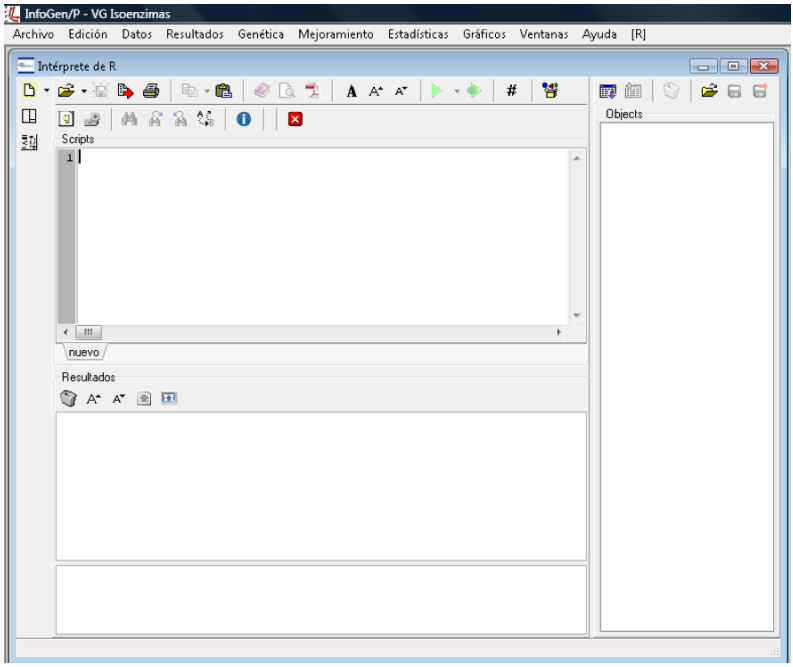


Figura 3.8: Intérprete de R en Info-Gen.

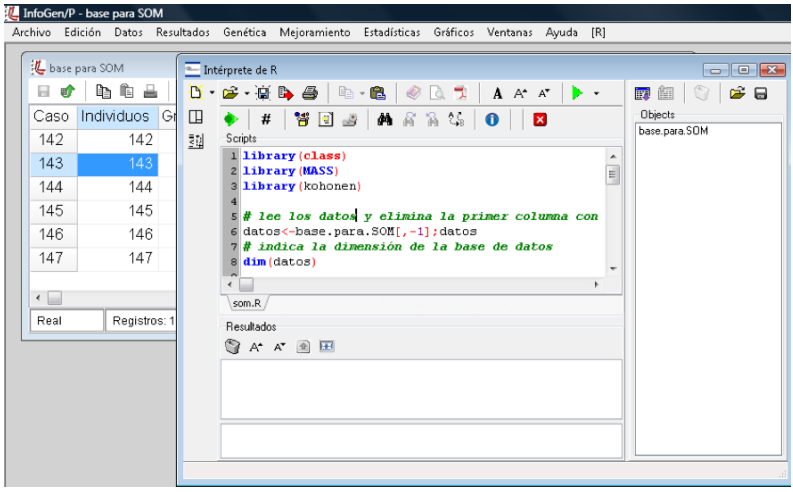


Figura 3.9: Intérprete de R en Info-Gen.

Para crear, leer o trabajar sobre los objetos, *R* necesita ins-

trucciones o códigos. Las funciones que hacen posible que las instrucciones se lleven a cabo están contenidas en paquetes (*packages*). *R* en la instalación incoprora una serie de paquetes. Sin embargo, cuando se quieren usar funciones específicas, suele ser necesario instalar los paquetes que las contienen. Una vez instalados dichos paquetes, todas las funciones contenidas en los mismos deben ser cargadas usando la función "library()" cada vez que se usan en el script o códigos de comandos. Las funciones que se cargan en "library()" equivalen a los menús del programa *Info-Gen* y los argumentos de las funciones a las opciones. Por ejemplo, para ajustar una regresión lineal en *Info-Gen*, debemos ir al menú Estadísticas, submenú "Regresión Lineal", para que se abra la ventana *selector de variables*. Una vez elegidas las variables, se hace click en *Aceptar* y se abren las opciones para el cálculo de regresión en una o más solapas o pestañas (Figura 3.10).

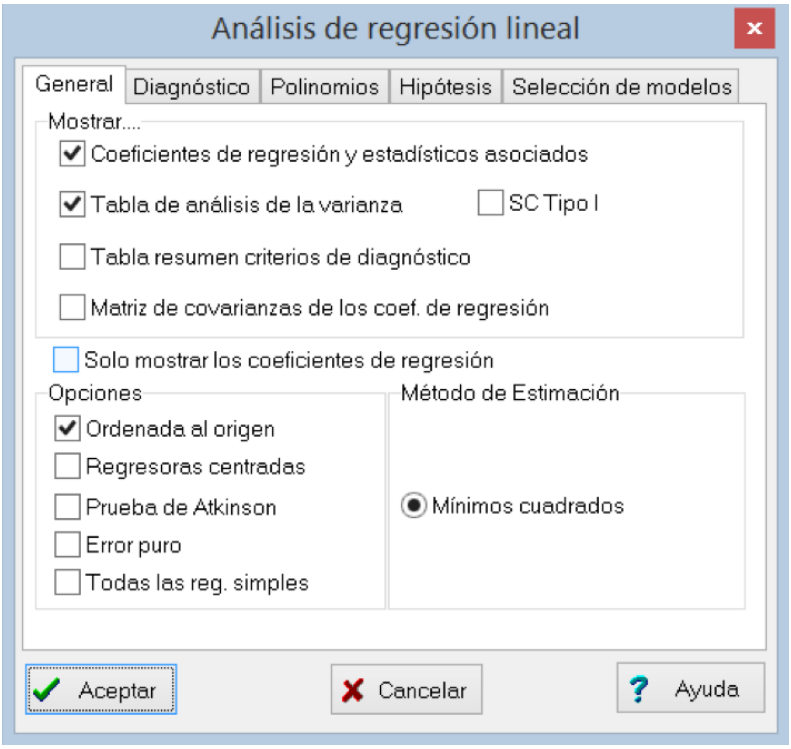


Figura 3.10: Intérprete de R en Info-Gen.

La función "lm" de la librería Estadísticas permite ejecutar una regresión lineal y tiene la siguiente sintaxis:

```
lm(formula, data, subset, weights, na.action,
   method = "qr", model=TRUE, x = FALSE,
   y = FALSE, qr = TRUE, singular.ok = TRUE,
   contrasts = NULL, offset, ...)
```

Cada término entre paréntesis y delimitado por comas es un argumento. Por ejemplo, el argumento *fórmula* es el que indica cuál es la variable dependiente y cuál la independiente en la regresión, así cómo cuál es el modelo

lineal que las une, mientras que *data* indica cuál es el objeto (*data.frame*) sobre el que se ajustará el modelo. Cuando no sabemos qué hace una función, usamos el comando "help(función)".

La ventana del inérprete de *R* se divide en cuatro paneles: *Editor de Scripts*, *Resultados o Salidas*, *Mensajes* y *Objetos* (Figura 3.11).

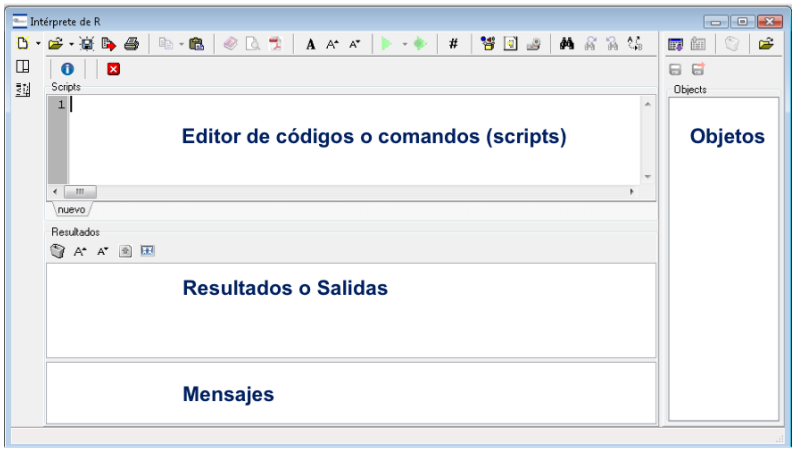


Figura 3.11: Paneles del intérprete de R en Info-Gen.

El intérprete del *lenguaje R* en *Info-Gen* permite trabajar con *R* desde *Info-Gen* escribiendo y ejecutando Scripts. El editor también permite al usuario cargar scripts previamente escritos. El editor destaca con distintos colores palabras clave de *R*, números, símbolos, palabras reservadas y comentarios (Figura 3.12).

```
Scripts
1 lm(formula, data, subset, weights, na.action,
2 method = "qr", model=TRUE, x = FALSE, y = FALSE,
3 qr = TRUE, singular.ok = TRUE, contrasts = NULL, offset, ...)
```

Figura 3.12: Líneas de comandos en el panel de script del intérprete de R en Info-Gen.

Una ayuda sobre un tema puede ser solicitada seleccionando el tema en el editor y pulsando el botón de ayuda disponibles en la barra de herramientas que encabeza el marco de trabajo. La ventana de ayuda se mostrará como una ventana independiente que tiene extensión *.html*.

R puede cargar archivos desde *Info-Gen* sin salir de la aplicación. Por ejemplo, se puede abrir en *Info-Gen*, desde el menú ARCHIVOS, opción *abrir tabla*, el archivo *Datos Cebada.igdb*. Luego, *R* necesita que sea convertido en un objeto. Uno de los objetos más frecuentes son los archivos de datos que contiene variables de clasificación y variables numéricas, al cual *R* denomina "Data Frame". Para ello, en el panel de "Objets" en *R*, hay una opción que permite convertir un archivo activo en un objeto (Figura 3.13).

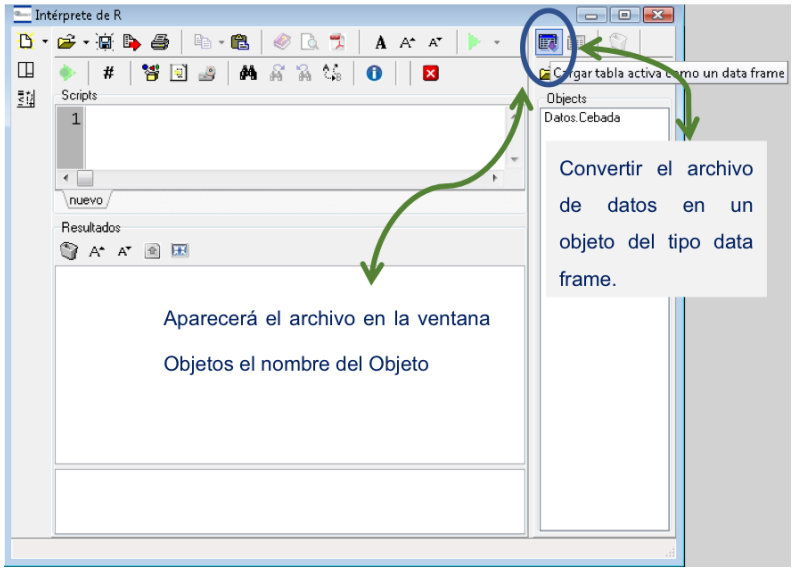


Figura 3.13: Convertir el archivo de datos de Info-Gen en un objeto de R de clase data grame. El nuevo objeto figurará listado en el panel de objetos.

Una vez escritas las líneas de comando, las mismas deben ser "corridas". Pueden ser corridas de a una línea por

vez o todas juntas. Para ello se puede usar el botón de la barra de accesos directos del intérprete de *R* que tiene forma de triángulo de color verde. También combinando las teclas *Ctrl+R* se corren las líneas seleccionadas. Si se pretende correr una sola línea es suficiente que el cursor se encuentre en dicha fila y apretar *Ctrl+R*; en cambio si se desea correr más de una línea simultáneamente, deben ser seleccionadas todas juntas (quedan sombreadas) y luego accionar *Ctrl+R* o el íncono destinado a tal fin. Las funciones se aplican sobre los datos que están listados en el panel de "Objetos" y los resultados se visualizan en el panel "Resultados".

Ejemplificaremos el uso de la interface de *Info-Gen* con *R* con la implementación de una regresión lineal usando la función "lm". Para ello es necesario escribir la función que se desea ejecutar en la ventana "Script". Ajustaremos el rendimiento, variable denominada yield en el archivo de datos, en función del genotipo, que está denominada en el archivo de datos como marcador D1001. Para escribir esto en el argumento fórmula escribiremos *yield~D1001*. La información o datos serán leídos desde el objeto "Datos.Cebada", que se indica con el argumento *data=Datos.Cebada*. La función *lm(yield~D1001, data=Datos.Cebada)* debe ser corrida para visualizar los resultados usando para ello las teclas *Ctrl+R* (Figura 3.14).

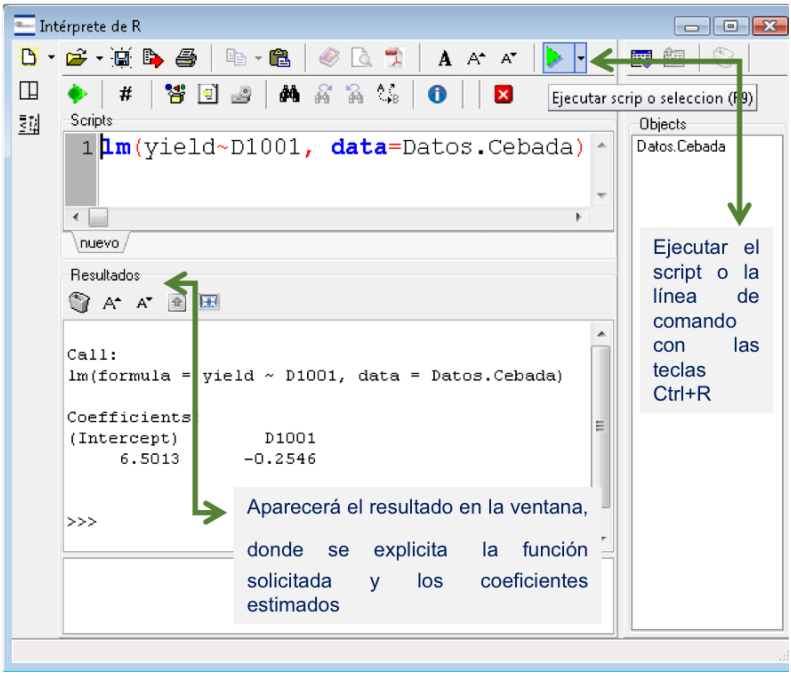


Figura 3.14: Al correr los comandos escritos en el script, para el cual lee la información desde los objetos data frame, los resultados se muestran en el panel de Resultados.

Una opción es colocar un nombre antes de la función de manera que al correr el script se generará un nuevo objeto. Para ello se coloca un nombre, por ejemplo, *lm.101* y debe colocarse entre este nombre y la función *lm()* un símbolo de asignación, que puede ser el símbolo matemático igual o una flecha de asignación conformado por el signo menos y el símbolo mayor (->). Los resultados de la función o el comando que se corre quedaran asignados al nuevo objeto denominado *lm.101* en este ejemplo. Para visualizar los resultados guardados en el nuevo objeto, puede correrse en una nueva de comando colocando línea el nombre del objeto o usar otras funciones como *summary(lm.101)* para visualizar los resultados (Figura 3.15).

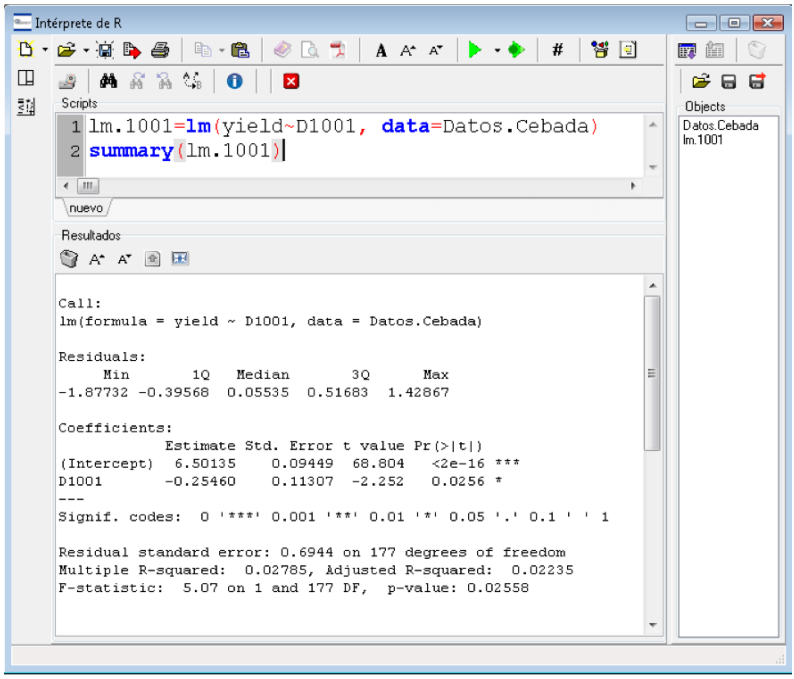


Figura 3.15: Los resultados de las funciones pueden ser guardados en nuevos objetos y visualizar sus resultados con la función summary() o corriendo el nombre del objeto.

Referencias

Aranzana, M. J., Kim, S., Zhao, K., Bakker, E., Horton, M., Jakob, K., Lister, C., Molitor, J., Shindo, C., Tang, C., Toomajian, C., Traw, B., Zheng, H., Bergelson, J., Dean, C., Marjoram, P. y Nordborg, M. (2005). *Genome-Wide Association Mapping in Arabidopsis Identifies Previously Known Flowering Time and Pathogen Resistance Genes*. PLoS Genet 1(5).

Balzarini M.G. y Di Rienzo J.A. *InfoGen versión 2018*. FCA, Universidad Nacional de Córdoba, Argentina. URL http://www.info-gen.com.ar

Balzarini, M., y Di Rienzo, J. (2004). *Info-Gen: Software para análisis estadístico de datos genéticos*. Universidad Nacional de Córdoba. Córdoba. Argentina.

Benjamini, Y. y Hochberg, Y. (1995). *Controlling the false discovery rate: A practical and powerful approach to multiple testing*. J R Stat Soc Ser B 57: 289 - 300.

Bonferroni, C. E. (1935). *Il calcolo delle assicurazioni su gruppi di teste*. Studi in Onore del Professore Salvatore Ortu Carboni.: 13-60.

Breseghello, F. y Sorrells, M. E. (2006). *Association Mapping of Kernel Size and Milling Quality in Wheat (Triticum aestivum L.) Cultivars*. Genetics 172(2): 1165-1177.

Brock, G., Pihur, V., Datta, S. y Datta, S. (2011). *clValid, an R package for cluster validation*. Journal of Statistical Software (Brock et al., March 2008).

Bruno, C. y Balzarini, M. (2010). *Distancias genéticas entre perfiles moleculares obtenidos desde marcadores multilocus multialélicos*. Revista de la Facultad de Ciencias Agrarias UNCuyo 41(3), 11.

Charrad, M., Ghazzali, N., Boiteau, V., Niknafs, A., y Charrad, M. M. (2014). *Package 'nbclust'*. Journal of statistical software, 61, 1-36.

Cheverud, J. M. (2001). *A simple correction for multiple comparisons in interval mapping genome scans*. Heredity 87(Pt 1): 52-58.

Comadran, J., Thomas, W. T. B., Eeuwijk, F. Á., Ceccarelli, S., Grando, S., Stanca, A. M., Pecchioni, N., Akar, T., Al-Yassin, A., Benbelkacem, A., Ouabbou, H., Bort, J., Romagosa, I., Hackett, C. A. y Russell, J. R. (2009). *Patterns of genetic diversity and linkage disequilibrium in a highly structured Hordeum vulgare association-mapping population for the Mediterranean basin*. Theoretical and Applied Genetics 119(1): 175-187.

D'hoop, B., Paulo, M., Mank, R., Eck, H. y Eeuwijk, F. (2008). *Association mapping of quality traits in potato (Solanum tuberosum L.)*. Euphytica 161(1-2): 47-60.

Demidenko, E. (2004). *Mixed Models: Theory and Application*. New Jersey: John Wiley & Sons.

Dunn, J. C. (1974). *Well-separated clusters and optimal fuzzy partitions*. Journal of cybernetics, 4(1), 95-104.

Fernández, E. A. y Balzarini, M. (2007). *Improving cluster visualization in self-organizing maps: Application in gene expression data analysis*. Comput. Biol. Med. 37(12): 1677-1689.

Flint-Garcia, S., Thuillet A.C., Yu J., Pressoir G., Romero S.M., Mitchell S.E., Doebley J., Kresovich S., Goodman M.M. y Buckler E.S. (2005). *Maize association population:*

a high-resolution platform for quantitative trait locus dissection. The Plant Journal 44: 10.

Frichot, E. y François, O. (2015). *LEA: an R package for landscape and ecological association studies*. Methods in Ecology and Evolution, 6(8), 925-929.

Gutiérrez, L., Quero, G. , Fernández, S., Brandariz, S. y Simondi, S. 2016. "Lmem.gwaser: Linear Mixed Effects Models for Genome-Wide Association Studies". Disponible en https://rdrr.io/cran/lmem.gwaser/man/gwas.analysis.html

Handl, J. y Knowles, J. (2005) *Exploiting the trade-off—the benefits of multiple objectives in data clustering*. In Coello,L.A. et al. (eds), Proceedings of the Third International Conference on Evolutionary Multicriterion Optimization. Springer- Verlag, Berlin, pp. 547–560.

Jaccard, P. (1901). *Étude comparative de la distribution florale dans une portion des Alpes et des Jura*. Bull Soc Vaudoise Sci Nat, 37, 547-579.

Jannink, J.-L., Iwata, H., Bhat, P. R., Chao, S., Wenzl, P. y Muehlbauer, G. J. (2009). *Marker imputation in barley association studies*. The Plant Genome 2(1): 11-22.

Kang, H. M., Zaitlen, N. A., Wade, C. M., Kirby, A., Heckerman, D., Daly, M. J. y Eskin, E. (2008). *Efficient Control of Population Structure in Model Organism Association Mapping*. Genetics 178(3): 1709-1723.

Kaufman, L. y Rousseeuw, P.J. (1990). *Finding groups in data: an introduction to cluster analysis*. John Wiley and Sons, New York, pp. 342.

Kohonen, T. (1997). *Self-organizing maps*. 117.

Kraakman, A. T. W., Martínez, F., Mussiraliev, B., Eeuwijk, F. A. y Niks, R. E. (2006). *Linkage Disequilibrium*

Mapping of Morphological, Resistance, and Other Agronomically Relevant Traits in Modern Spring Barley Cultivars. Molecular Breeding 17(1): 41-58.

Lee, C. Abdool, A. y Huang, C. (2009).*PCA-based population structure inference with generic clustering algorithms*. BMC Bioinformatics 10(Suppl 1), S73.

Li, J. y Ji, L. (2005). *Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix*. Heredity 95(3): 221-227.

Malavera, P., Bruno, C. y Balzarini, M. (2018). *Control de falsos descubrimientos en mapeo asociativo con poblaciones estructuradas false discovery rate control in association mapping with genetically structured populations*. Journal of Basic and Applied Genetics, 29(1), 37-49.

McLaren, C. G., Bruskiewich, R. M., Portugal, A. M. y Cosico, A. B. (2005). *The international rice information system. A platform for meta-analysis of rice crop data*. Plant Physiology, 139(2), 637-642.

McLaren, C. G., Ramos, L., López, C. y Eusebio, W. (2000). *Applications of the genealogy management system. ICIS International Crop Information System: Technical Development Manual-version 6*. Mexico DF (Mexico): Centro Internacional de Mejoramiento de Maiz y Trigo (CIMMYT), Mexico DF.

Nikolic, N. Park, Y. S. Sancristobal, M. Lek, S. y Chevalet, C. (2009). *What do artificial neural networks tell us about the genetic structure of populations? The example of European pig populations*. Genet Res (Camb) 91(02), 121-132.

Odong, T. van Heerwaarden, J. Jansen, J. van Hintum, T. y van Eeuwijk, F. (2011). *Determination of genetic structure of germplasm collections: are traditional hierarchical*

clustering methods appropriate for molecular marker data? TAG Theoretical and Applied. Genetics 123(2), 195-205.

Patterson, H. D. y Thompson, R. (1971). *Recovery of inter-block information when block sizes are unequal*. Biometrika 58(3): 545-554.

Peña Malavera, A., Gutierrez, L. y Balzarini, M. (2014). *Componentes principales en mapeo asociativo*. BAG. Journal of basic and applied genetics 25: 32-40.

Pritchard, J., M. Stephens y P. Donnelly (2000). *Inference of population structure using multilocus genotype data*. Genetics, 155, 945 – 959.

R Core Team (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Remington, J. S. y Klein, J. O. (2001). *Infectious diseases of the fetus and newborn infant*. WB Saunders,.

Rezaee, R., Lelieveldt, B.P.F. y Reiber, J.H.C. (1998). *A New Cluster Validity Index for the Fuzzy c-Mean*. Pattern Recognition Letters, 19, 237–246.

Roux, O. Gevrey, M. Arvanitakis, L. Gers, C. Bordat, D. y Legal, L. (2007). *ISSR-PCR: Tool for discrimination and genetic structure analysis of Plutella xylostella populations native to different geographical areas*. Molecular Phylogenetics and Evolution 43(1), 240-250.

Stich, B., Möhring, J., Piepho, H.-P., Heckenberger, M., Buckler, E. S. y Melchinger, A. E. (2008). *Comparison of Mixed-Model Approaches for Association Mapping*. Genetics 178(3): 1745-1754.

Thornsberry, J., Goodman, M., Doebley, J., Kresovich, S., Nielsen, D. y Buckler, E. (2001). *Dwarf8 polymorphisms*

associate with variation in flowering time. Nat Genet 28: 286 - 289.

Tracy, C. A. y Widom, H. (1994). *Level-spacing distributions and the Airy kernel*. Comm. Math. Phys. 159(1): 23.

Ultsch, A. (Ed) (1999). *Data mining and knowledge discovery with emergent selforganizing feature maps for multivariate time series*. E. Oja, S. Kaski.

West, B. T., Welch, K. B. y Galecki, A. T. (2014). *Linear mixed models: a practical guide using statistical software*. Chapman and Hall/CRC.

Yu, J. y Buckler, E. (2006). *Genetic association mapping and genome organization of maize*. Curr Opin Biotech 17: 155 - 160.

Yu, J. y Buckler, E. (2006). *Genetic association mapping and genome organization of maize*. Curr Opin Biotech 17: 155 - 160.

Zhu, C., Gore, M., Buckler, E. y Yu, J. (2008). *Status and Prospects of Association Mapping in Plants*. The Plant Genome 1(1): 16.

Impreso por Editorial Brujas • diciembre de 2019 • Córdoba–Argentina