Desarrollo de *pipelines* bioinformáticos aplicados a la mejora genética de plantas

Anna Moncusí Moix
Máster en Bioinformática y Bioestadística
Área 33: Microbiología, biotecnología y biología molecular

Nombre Consultor/a: Paloma Pizarro Tobías
Nombre Profesor/a responsable de la asignatura: David Merino Arranz

Fecha Entrega: 06/2018

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Desarrollo de pipelines bioinformáticos aplicados a la mejora genética de plantas*
Nombre del autor:	*Anna Moncusí Moix*
Nombre del consultor/a:	*Paloma Pizarro Tobías*
Nombre del PRA:	*David Merino Arranz*
Fecha de entrega (mm/aaaa):	06/2018
Titulación:	*Máster en Bioinformática y Bioestadística*
Área del Trabajo Final:	*Área 33: Microbiología, biotecnología y biología molecular*
Idioma del trabajo:	*Castellano*
Palabras clave	*bioinformática, biotecnología, genotipo, mejora genética, SNPs*

Resumen del Trabajo (máximo 250 palabras):

El uso de marcadores moleculares ha revolucionado el ritmo y la precisión del análisis genético de las plantas, siendo una disciplina importante para garantizar la seguridad alimentaria mediante el desarrollo de variedades que hagan frente al estrés ambiental y contribuyan al desarrollo sostenible de la agricultura. El objetivo de esta investigación fue desarrollar un pipeline para el análisis bioinformático y bioestadístico de los datos obtenidos en el laboratorio de biotecnología aplicada de la empresa Semillas Fitó. El primer objetivo fue obtener una función en R para realizar estudios de parámetros básicos de los datos de genotipado como la heterocigosidad, PIC, el índice de fijación y el porcentaje de *missing data*. Se ha obtenido una función en R para codificar los datos, lo que permite realizar el segundo objetivo, el análisis de los datos genéticos mediante los softwares R, Flapjack, GGT y TASSEL, seleccionando a partir de una matriz de decisión, R como el mejor software y con más flexibilidad para realizar futuros análisis. Se ha comparado también los softwares MapDisto y MapMaker para analizar las similitudes y diferencias en la construcción de mapas de ligamiento. También se ha representado la distribución de los SNPs en los cromosomas para poder obtener una visión global de la distribución de los datos mediante el software PhenoGram. Por último, el tercer objetivo fue desarrollar una función de R para el análisis de Marker-Assisted Backcrossing.

Abstract (in English, 250 words or less):

The use of molecular markers has revolutionized the rhythm and precision of the genetic analysis of plants, being an important discipline to ensure food security by developing varieties that cope with the environmental stresses and contribute towards sustainable development of agriculture. The objective of this research was to develop a pipeline for bioinformatic and biostatistical analysis of the data obtained in the applied biotechnology laboratory of Semillas Fitó company. The first goal was to obtain a function in R to carry out studies of basic parameters of the genotyping data such as heterozygosity, PIC, inbreeding and the percentage of missing data. A function has been obtained in R to encode the data, which allows to realize the second goal, the analysis of the genetic data comparing R, Flapjack, GGT and TASSEL, selecting from a decision matrix R as the best software and with more flexibility to carry out future analyzes. MapDisto and MapMaker have also been compared to analyze the similarities and differences in the construction of linkage maps. The distribution of the SNPs in the chromosomes has also been represented to obtain a global view of the distribution of the data using the PhenoGram software. Finally, the third goal was to develop an R function for the analysis of Marker-Assisted Backcrossing.

Índice

Lista de figuras

1. Introducción

1.1 Contexto y justificación del Trabajo

El uso de marcadores moleculares ha revolucionado el ritmo y la precisión del análisis genético de las plantas, lo que a su vez ha facilitado la implementación de mejora molecular en cultivos. La mejora molecular en plantas es una industria agrícola vital que necesita ser fomentada y estimulada. Es una disciplina importante para garantizar la seguridad alimentaria mediante el desarrollo de variedades que hagan frente al estrés ambiental y contribuyan al desarrollo sostenible de la agricultura (Tiwari, 2017).

Los *Single Nucleotide Polymorphism* (SNP) son muy populares en genética molecular de plantas debido a su abundancia y susceptibilidad en todo el genoma para las plataformas de detección de estos (Mammadov, 2012). Un SNP es un polimorfismo entre individuos de la misma especie, consiste en una variación en un solo par de bases. Se encuentra con una frecuencia superior al 1%, ya que por debajo de este valor se considera mutación.

La finalidad de este TFM consiste en realizar análisis bioinformáticos y bioestadísticos sobre los datos obtenidos en el laboratorio de marcadores moleculares del departamento de biotecnología aplicada de la empresa Semillas Fitó, generando un pipeline para crear una rutina para el análisis de datos.

Para obtener los datos de genotipado, el primer paso es la llegada de las muestras de hoja deshidratada en placas de 96 pocillos en el laboratorio (imagen 1). Estas placas de recogida de muestra están formadas por 8 filas, nombradas de la A a la H, y 12 columnas, numeradas del 1 al 12. La placa contiene 96 pocillos y cada uno de éstos contiene una muestra de hoja deshidratada de una planta diferente, por lo tanto, en cada placa se puede realizar la extracción de ADN de 96 plantas diferentes.

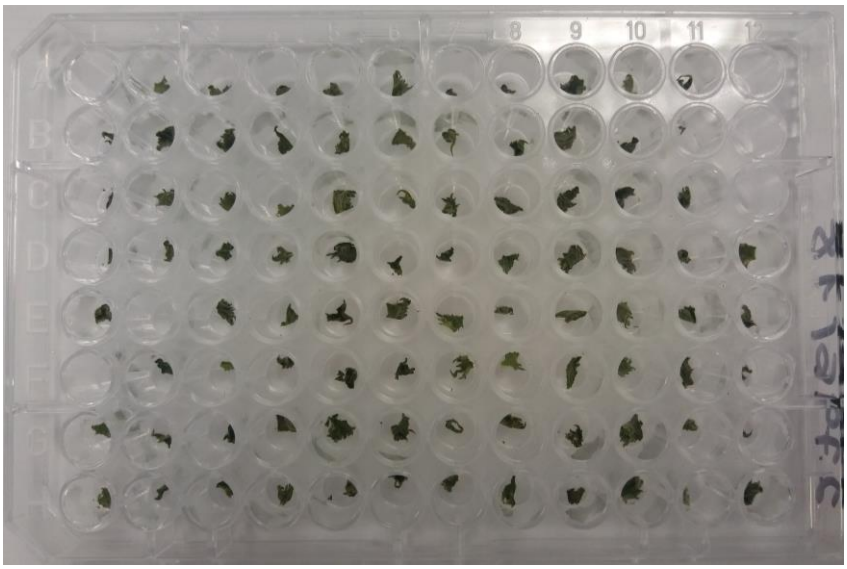


Imagen 1. Placa de recogida de muestras de hoja deshidratada

Los datos de genotipado se obtienen mediante la reacción en cadena de la polimerasa, conocida como PCR. Es una técnica de biología molecular con el objetivo de obtener un gran número de copias de un fragmento de ADN en particular. Cuando se ha finalizado el proceso de extracción de ADN, se preparan las placas de PCR.

Hay dos tipos de placas de PCR, las de 384 pocillos (imagen 2) y las de 1536 pocillos (imagen 3).

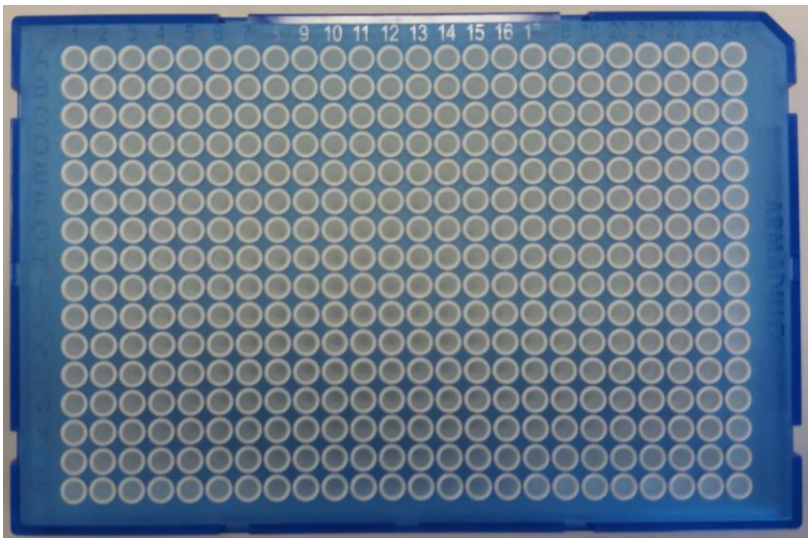


Imagen 2. Placa de PCR de 384 pocillos

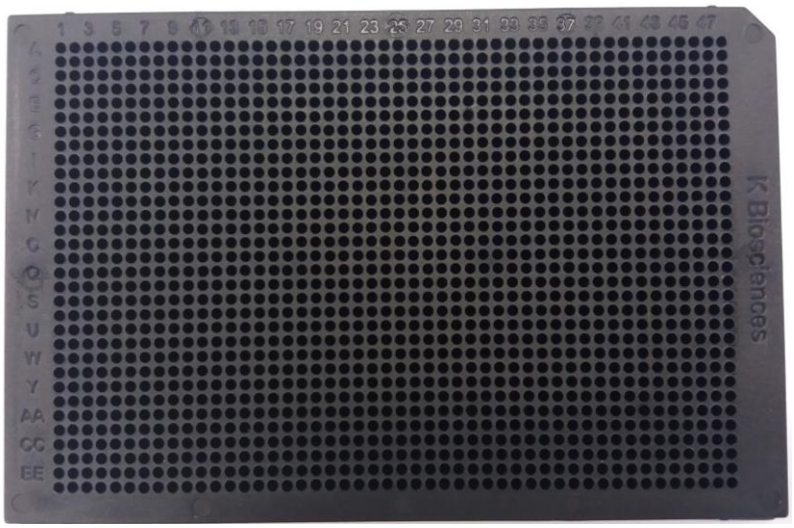


Imagen 3. Placa de PCR de 1536 pocillos

Las placas de PCR de 384 pocillos permiten analizar 4 placas de extracción de 96 plantas simultáneamente, y las placas de 1536 permiten el análisis de 16 placas de extracción.

El sistema de genotipado utilizado en el laboratorio mediante PCR es un sistema de genotipado fluorescente, utiliza dos *primers* iniciadores competitivos, alelo-específicos y un *primer* común inverso. Cada *primer* iniciador lleva una secuencia de cola incorporada que corresponde con uno de los FRET universales (Fluorescent resonance energy transfer).

Una vez completada la PCR, la lectura de los datos permite obtener un gráfico de grupos como el de la imagen 4:

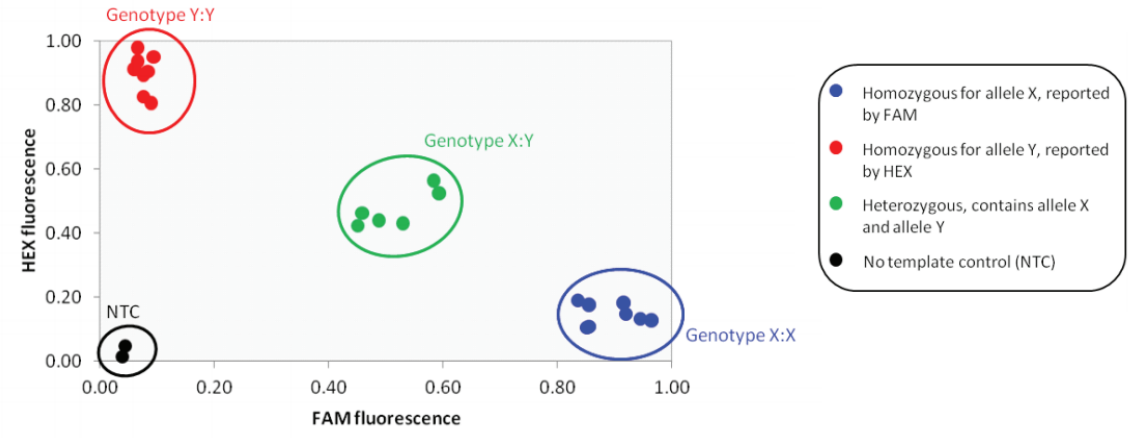


Imagen 4. Gráfico de grupos de genotipado. Imagen obtenida del archivo "A guide to the analysis of KASP genotyping data using cluster plots"

Cada punto representado en el gráfico muestra la señal fluorescente de una única planta o muestra de ADN. Los individuos homocigotos por el alelo Y se sitúan en la parte superior izquierda del gráfico, representados en rojo. Los individuos homocigotos por el alelo X se sitúan en la parte inferior derecha del gráfico, representados en azul. Los individuos heterocigotos se sitúan en la parte central del gráfico, representados en verde. Los puntos negros situados en la parte inferior izquierda indican falta de ADN o *missing data*.

La exportación de este gráfico en formato Excel permite obtener el archivo de genotipado a partir del cual se partirá en este trabajo para realizar los análisis bioinformáticos y bioestadísticos.

1.2 Objetivos del Trabajo

El trabajo de fin de máster presentado en esta memoria tiene como objetivos:

1. **Estudiar los parámetros básicos a partir de datos genotípicos.** Realizar estudios con R de parámetros básicos como:
   - Heterocigosidad
   - PIC *("Polymorphism Information Content")*
   - Inbreeding o índice de fijación
   - Porcentaje de *missing data*.

2. **Analizar los datos genéticos mediante diversos softwares y comparación de los resultados**

   2.1 Analizar distancias genéticas. Realizar análisis de distancias genéticas para una posterior comparativa y evaluación de los resultados, con el fin de encontrar el software más completo para nuestras necesidades. El análisis consta de:
   - Matriz de similitud
   - Matriz de distancias
   - Dendrogramas
   - PCA

   2.2 Construir mapas. Realizar mapas de ligamiento mediante los softwares MapMaker y MapDisto.

   2.3 Visualizar gráficamente genotipos y cromosomas. Representar mediante el software Phenogram la distribución de los SNPs en los cromosomas para obtener una visión global de la distribución de datos.

3. **Desarrollar un pipeline en R para realizar estudios de Marker-Assisted Backcrossing (MABC).**

1.3 Enfoque y método seguido

Inicialmente se valoran todos los recursos para realizar los análisis y, una vez escogidos, se realizan los análisis que se consideren oportunos. Así pues, una buena opción para la realización del trabajo es dedicar las primeras semanas a la búsqueda de paquetes de R y herramientas para el estudio de los datos de genotipado, además de aprender a adaptar los formatos de archivos obtenidos en el laboratorio para los *inputs* de las herramientas.

Una vez encontradas las herramientas adecuadas y toda la información necesaria, la siguiente tarea es avanzar en el estudio con datos reales encriptados, debido a la elevada confidencialidad de éstos.

Finalmente, de forma paralela al análisis de datos reales, se realiza la comparativa entre los diferentes resultados obtenidos en el análisis de datos genómicos y el desarrollo del pipeline en el estudio de MABC, pero dedicándole más tiempo a este último, considerando que presenta una mayor dificultad.

1.4 Planificación del Trabajo

4.1 Tareas

4.1.1. **Tareas objetivo 1.**

- Búsqueda de paquetes de R para realizar los análisis de interés (del 20 al 26 de marzo).

- Análisis de datos reales (del 27 de marzo al 2 de abril).

4.1.2. **Tareas objetivo 2.**

- Búsqueda de herramientas para realizar los estudios de genotipado (del 20 al 26 de marzo).

- Adaptación de los formatos de datos para el input de las diferentes herramientas encontradas (del 27 de marzo al 2 de abril).

- Análisis de datos reales (del 3 al 16 de abril).

- Comparativa de los resultados obtenidos en las diferentes herramientas (del 17 al 30 de abril).

4.1.3. **Tareas objetivo 3.**

- Desarrollar un pipeline para realizar estudios MABC (del 17 de abril al 21 de mayo).

1.5 Breve sumario de productos obtenidos

En el primer objetivo se ha obtenido una función en R la cual permite calcular los parámetros básicos de los archivos de genotipado, para conocer el tipo de datos con los que estamos trabajando, y otra función en R para la codificación de los datos.
El segundo objetivo ha permitido conocer los tipos de softwares para analizar distancias genéticas, realización de mapas de ligamiento y mapas de cromosomas y genotipos. A partir de este estudio se ha podido valorar que, entre los softwares analizados, R presenta mucha versatilidad en los estudios, siendo elegido como el que mejor se adapta a nuestras necesidades.
En el tercer objetivo se ha obtenido una función en R que permite reconocer mediante una matriz de distancia los individuos de mayor interés para los análisis de MABC.

1.6 Breve descripción de los otros capítulos de la memoria

Materiales y métodos, donde se describe detalladamente qué herramientas se usan en el desarrollo del trabajo y cómo se trabaja con ellas.

Resultados, donde se detalla claramente qué resultados se han obtenido mediante el uso de los materiales y métodos establecidos en el punto anterior.

Discusión, una vez se han mostrado los resultados, analizar el porqué, la coherencia de estos, el interés para la población, etc.

2. Materiales y métodos

2.1 Estudiar los parámetros básicos a partir de datos genotípicos

En este estudio, se codifican los datos con doble finalidad: garantizar la confidencialidad de estos y, además, el preprocesado para su introducción en los diferentes softwares que requieren de formatos específicos.

El *output* del software obtenido en el laboratorio contiene columnas con los marcadores moleculares o *assays* y filas con los individuos analizados. Los datos de genotipado están representados mediante nucleótidos. Al tratarse de ADN, estos nucleótidos pueden ser Adenina (A), Citosina (C), Timina (T) o Guanina (G).

El resultado deseado de la codificación es:

2 o A: para el alelo dominante, **0 o B:** para el alelo minoritario, **1 o H:** para los heterocigotos, y **3 o nd:** para los datos faltantes. Esta codificación la puede realizar el programa GGT 2.0. El ejemplo (Imagen 5) contiene datos ficticios con información de genotipado de 10 individuos y 11 marcadores o *assays*, dispuestos en 3 cromosomas:

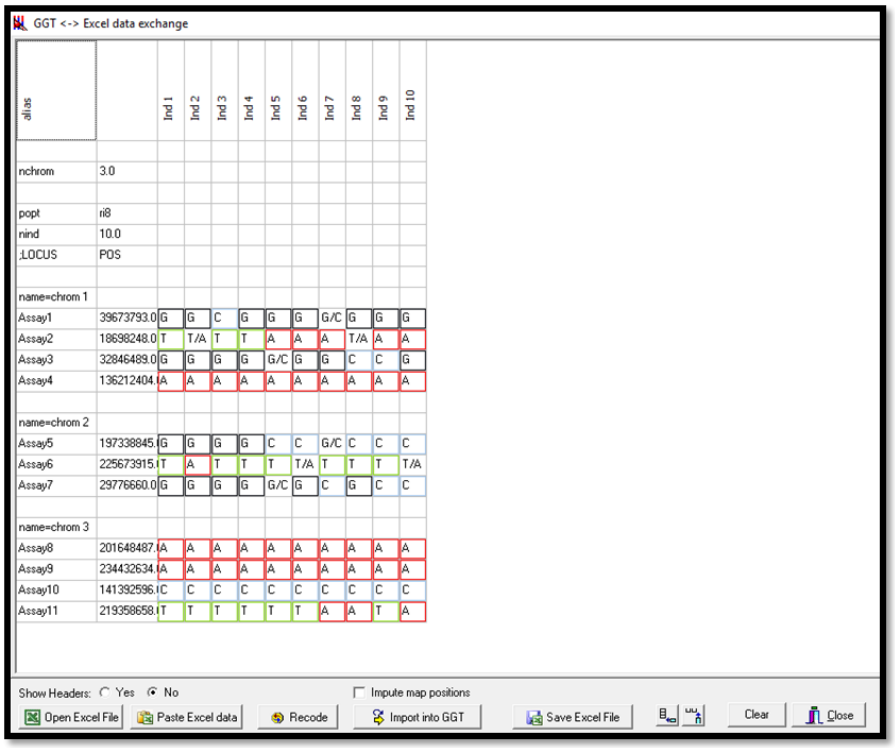


Imagen 5. Software GGT con datos ficticios de genotipado.

Una vez realizada la recodificación, la imagen 6 muestra el resultado obtenido:

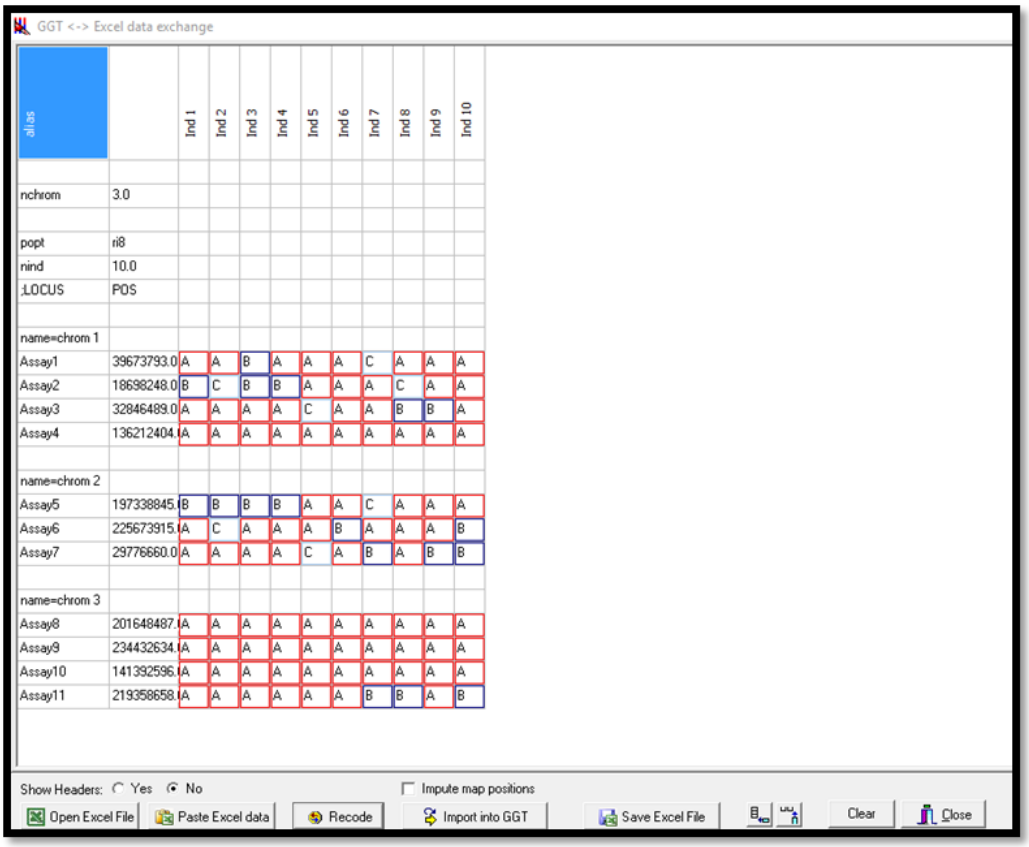


Imagen 6. Codificación de los datos de genotipado mediante el software GGT.

El resultado no es el esperado, ya que se pretende obtener el mismo carácter para los alelos dominantes (2 o A), alelos minoritarios (0 o B), heterocigotos (1 o H o C) y missing data (3 o nd). En cambio, el programa no sigue este criterio, ya que asigna las letras A, B y C según la proporción de alelos, sin discriminar entre homocigotos, heterocigotos y *missing data* (Imagen 1 Anexo).

Con el criterio establecido en este estudio, los resultados obtenidos en la codificación mediante el software GGT 2.0 no son válidos para realizar los análisis de los parámetros básicos con R, ya que el logaritmo de codificación establecido no sigue el patrón esperado. Así pues, se descartó la opción de codificar con dicho software.

La siguiente opción fue realizar una búsqueda de alternativas a la codificación. Una posible solución sería la utilización de una librería de R. Mediante esta búsqueda se encontró la función *data.matrix*(), la cual forma parte del programa base (Imagen 2 Anexo).

Esta función codifica los valores de cada marcador en 1, 2, 3 y 4, donde los números 1 y 3 corresponden a los alelos homocigotos, 2 corresponde a los alelos heterocigotos, y 4 corresponde a los datos faltantes o *missing data*.
Según el interés de la codificación para este trabajo, al analizar los resultados, la función clasifica correctamente los heterocigotos y los datos faltantes, pero no usa una distinción para los alelos, ya que no considera cual es el mayoritario y cuál es el minoritario. Por lo tanto, se descartó esta segunda opción de codificación.

La siguiente opción, y definitiva, fue realizar una función mediante R que realizara la codificación con los requisitos necesarios para el trabajo. La función se detalla a continuación:

```
codigo<-read.csv(file="DADES TFM.csv", sep=";")
DNA<-codigo[,1]
codigo<-codigo[,-1]

Codificacion<-function(x){
  ultim<-data.frame(DNA)
  for(n in 1:length(x)){
    A<-as.numeric(nrow(as.data.frame(which(x[,n]=="A"))))
    T<-as.numeric(nrow(as.data.frame(which(x[,n]=="T"))))
    G<-as.numeric(nrow(as.data.frame(which(x[,n]=="G"))))
    C<-as.numeric(nrow(as.data.frame(which(x[,n]=="C"))))
    nd<-as.numeric(nrow(as.data.frame(which(x[,n]=="nd"))))
    H<-nrow(x)-sum(A,T,G,C,nd)
    library(car)
    Z<-data.frame(if(G==0) {if(A==0) {if(C>T) {recode(x[,n], "'C'=2;
'T'=0; 'nd'='3'; else='1'")}}})
    Z1<-data.frame(if(G==0) {if(A==0) {if(T>C) {recode(x[,n], "'C'=0;
'T'=2; 'nd'='3'; else='1'")}}})
    Z2<-data.frame(if(T==0) {if(G==0) {if(C>A) {recode(x[,n], "'C'=2;
'A'=0; 'nd'='3'; else='1'")}}})
    Z3<-data.frame(if(T==0) {if(G==0) {if(A>C) {recode(x[,n], "'C'=0;
'A'=2; 'nd'='3'; else='1'")}}})
    Z4<-data.frame(if(T==0) {if(A==0) {if(C>G) {recode(x[,n], "'C'=2;
'G'=0; 'nd'='3'; else='1'")}}})
    Z5<-data.frame(if(T==0) {if(A==0) {if(G>C) {recode(x[,n], "'C'=0;
'G'=2; 'nd'='3'; else='1'")}}})
    Z6<-data.frame(if(G==0) {if(C==0) {if(T>A) {recode(x[,n], "'T'=2;
'A'=0; 'nd'='3'; else='1'")}}})
    Z7<-data.frame(if(G==0) {if(C==0) {if(A>T) {recode(x[,n], "'T'=0;
'A'=2; 'nd'='3'; else='1'")}}})
    Z8<-data.frame(if(C==0) {if(A==0) {if(T>G) {recode(x[,n], "'T'=2;
'G'=0; 'nd'='3'; else='1'")}}})
    Z9<-data.frame(if(C==0) {if(A==0) {if(G>T) {recode(x[,n], "'T'=0;
'G'=2; 'nd'='3'; else='1'")}}})
    Z10<-data.frame(if(T==0) {if(C==0) {if(A>G) {recode(x[,n], "'A'=2;
'G'=0; 'nd'='3'; else='1'")}}})
```

```
    Z11<-data.frame(if(T==0) {if(C==0) {if(G>A) {recode(x[,n], "'A'=0;
'G'=2; 'nd'='3'; else='1'")}}})
    Z12<-data.frame(if(G==0) {if(A==0) {if(T==C) {recode(x[,n],
"'T'=2; 'C'=0; 'nd'='3'; else='1'")}}})
    Z13<-data.frame(if(C==0) {if(A==0) {if(T==G) {recode(x[,n],
"'T'=2; 'G'=0; 'nd'='3'; else='1'")}}})
    Z14<-data.frame(if(G==0) {if(C==0) {if(T==A) {recode(x[,n],
"'T'=2; 'A'=0; 'nd'='3'; else='1'")}}})
    Z15<-data.frame(if(T==0) {if(A==0) {if(C==G) {recode(x[,n],
"'C'=2; 'G'=0; 'nd'='3'; else='1'")}}})
    Z16<-data.frame(if(T==0) {if(G==0) {if(C==A) {recode(x[,n],
"'C'=2; 'A'=0; 'nd'='3'; else='1'")}}})
    Z17<-data.frame(if(T==0) {if(C==0) {if(A==G) {recode(x[,n],
"'A'=2; 'G'=0; 'nd'='3'; else='1'")}}})
    intermig<-
c(Z,Z1,Z2,Z3,Z4,Z5,Z6,Z7,Z8,Z9,Z10,Z11,Z12,Z13,Z14,Z15,Z16,Z17)
    intermig<-data.frame(intermig[1])
    ultim<-c(ultim, intermig)
  }
  Assays<-names(codigo)
  Assays<-c("DNA",Assays)
  names(ultim)<-Assays
  return(data.frame(ultim))
}

Datos<-Codificacion(codigo)
```

Inicialmente la función tenia los objetos z hasta z11, los cuales representaban todas las combinaciones de nucleótidos posibles. Al probar la función con un archivo ficticio de 353 columnas, devolvía un archivo con 350 columnas. Buscando el error, se observaron 3 columnas, las cuales, tenían los dos alelos homocigotos con la misma frecuencia (por ejemplo: **45 C**, **45 G**, 13 G/C y 4 nd), por lo que el código de la función no estaba preparado para esta situación, ya que la condición de tener la misma frecuencia no se podía aplicar. Como solución a este problema, se crearon 6 objetos más (z12 hasta z17), donde se establecen criterios aleatorios a la codificación de los alelos con una misma frecuencia.

Una vez obtenidos los datos codificados, podemos analizar los parámetros básicos:

i. *Polymorphism Information Content* (PIC)

El PIC se refiere al valor relativo de cada marcador genético con respecto a la cantidad de polimorfismo (Ngangkham et al. 2018). Esta medida de información depende del número de alelos para ese marcador y sus frecuencias relativas.

Un marcador que genere un nivel bajo de polimorfismo puede ser útil para estudiar especies o géneros dentro de una familia, pero ineficaz para estudiar diferencias entre individuos de una misma población o entre descendientes de un cruzamiento. Si un marcador es poco polimórfico, la mayoría o todos los individuos de una población o especie tendrán el mismo fenotipo. Así pues, un marcador con un elevado valor de PIC presenta mayor interés en el ámbito del *plant breeding*.

ii. Heterocigosidad

La heterocigosidad es una medida de la variación genética: el estado de tener diferentes alelos con respecto a un carácter dado (Jingade et al. 2011). Es una herramienta estadística para explorar los patrones y los efectos fenotípicos de diferentes niveles de variación genética.

iii. Índice de fijación o *inbreeding*

El Índice de fijación representa la reducción de la heterocigosidad debido a cruzamientos realizados de forma no azarosa. Tiende a aumentar el número de homocigotos para un rasgo, incrementando por lo tanto la aparición de rasgos recesivos.

iv. Porcentage de *missing data*

Este valor permite obtener la proporción de datos o información faltante en los archivos obtenidos en el laboratorio. Este valor puede suponer una primera referencia a la calidad de los datos, ya que, una proporción elevada de *missing data* podría indicar que existe algún problema en el proceso de obtención de estos.

Se ha desarrollado una función de R que calcula todos los parámetros, devolviendo un *output* en forma de tabla con todos los datos por *assay*. Además, el parámetro heterocigosidad también se calcula por individuo.

```
tabla<-function(x){
  Data<-c()
  for (i in 1:length(x)) {
    n0<-nrow(as.data.frame(which(x[,i]=="0")))
    n1<-nrow(as.data.frame(which(x[,i]=="1")))
    n2<-nrow(as.data.frame(which(x[,i]=="2")))
    n<-n0+n1+n2
    p<-((2*n2)+n1)/(2*n)
    q<-1-p
    Heterocigosidad<-data.frame(1-((p)^2+(q)^2))
    Inbreeding<-data.frame(1-((n1)/(n*2*p*q)))
    PIC<-data.frame((1-((p)^2+(q)^2))-(2*(((p)^2*(q)^2))))
    MissingData<-
(as.numeric(nrow(as.data.frame(which(x[,i]=="3")))))/nrow(x)
    Q<-data.frame(Heterocigosidad,Inbreeding,PIC,MissingData)
    Data<-rbind(Data,Q)
  }
  Data<-Data[-1,]
  names(Data)<-c("Heterocigosidad","Inbreeding","PIC","Missing data")
  Numeros<-c(1:length(x[,-1]))
  Nombres<-c("Assay")
  Id<-paste(Nombres, Numeros, sep = "")
  row.names(Data)<-Id
  return(Data)
}

Parametros_basicos<-tabla(Dades)
head(Parametros_basicos)
```

```
##        Heterocigosidad Inbreeding       PIC Missing data
## Assay1       0.4601130  0.6681864 0.3542610   0.03676471
## Assay2       0.1198080  0.7329060 0.1126310   0.08088235
## Assay3       0.2761038  0.8366093 0.2379871   0.02205882
## Assay4       0.4027838  0.8631571 0.3216664   0.06617647
## Assay5       0.3731132  0.7984848 0.3035065   0.02205882
## Assay6       0.2661927  0.7513964 0.2307634   0.00000000
```

```
HeterocigosidadIndividuos<-function(x){
  Data<-c()
  for (n in 1:nrow(x)) {
    n0<-nrow(as.data.frame(which(x[n,]=="0")))
    n1<-nrow(as.data.frame(which(x[n,]=="1")))
    n2<-nrow(as.data.frame(which(x[n,]=="2")))
    n<-n0+n1+n2
    p<-((2*n2)+n1)/(2*n)
    q<-1-p
    Heterocigosidad<-data.frame(1-((p)^2+(q)^2))
```

```
    Data<-c(Data,Heterocigosidad)
  }
  Numeros<-c(1:nrow(x[,-1]))
  Nombres<-c("Individuo")
  Id<-paste(Nombres, Numeros, sep = "")
  names(Data)<-Id
  Data<-t(data.frame(Data))
  return(Data)
}

HeterocigosidadIndividual<-HeterocigosidadIndividuos(Dades)
head(HeterocigosidadIndividual)

##                 [,1]
## Individuo1 0.4041795
## Individuo2 0.4095882
## Individuo3 0.3086398
## Individuo4 0.4775567
## Individuo5 0.3694819
## Individuo6 0.3482835
```

2.2 Análisis de datos genéticos mediante diversos softwares y comparación de los resultados

El primer paso para el procesado inicial de los datos ha sido su codificación. El siguiente paso es la adaptación de los datos a los formatos requeridos para cada software.
Es muy importante valorar con qué tipo de softwares se trabaja, qué datos necesitan cada uno como *input*, y las posibles limitaciones que pueden aparecer durante el transcurso del análisis.

Los análisis que se realizan son los siguientes:

i. Análisis de distancias genéticas.

Mediante matrices de distancia (o la conversión de matrices de similitud) se realizan análisis de componentes principales (PCA) y dendrogramas para analizar las distancias genéticas entre individuos a partir de datos de marcadores moleculares o SNP.

ii. Construcción de mapas.

El objetivo en este apartado es ver algunas de las opciones para realizar mapas de ligamiento. La información que nos proporcionan estos mapas es la probabilidad de que dos marcadores sean separados por un evento de recombinación, ya que depende de la distancia que haya entre ellos. La unidad de mapeo es el centimorgan (cM), que corresponde a una unidad de mapa o al 1% de probabilidad de producir recombinación después de la meiosis.

iii. Visualización gráfica de genotipos y cromosomas.

Las herramientas que nos permiten visualizar gráficamente los cromosomas y las posiciones de los marcadores fomentan explorar y compartir resultados complejos, para facilitar una mayor comprensión de los datos.

Los softwares han sido seleccionados por ser gratuitos, relativamente fáciles de usar por su intuitiva interfaz de usuario y por su habitual uso en el ámbito en el que nos encontramos.

A continuación, se detallan qué softwares han sido usados y los correspondientes *inputs*:

RStudio es un entorno de desarrollo integrado (IDE) para R. La dirección para descargar el software R Project es https://www.r-project.org/ y R Studio https://www.rstudio.com/products/rstudio/download/. El objetivo de este trabajo es integrar los conocimientos adquiridos durante el máster con los datos de genotipado, esperando obtener información estadística, gráficos, automatización mediante funciones, etc., para la interpretación de datos y mejora del trabajo diario en el ámbito de la mejora molecular en plantas. Se requiere la incorporación de los datos de genotipado codificados en el formato .csv, mostrado en la imagen 7. Para mostrar los resultados obtenidos en R Studio se usará el paquete R Markdown.

DNA \ Assay	Assay1	Assay2	Assay3	Assay4	Assay5	Assay6	Assay7	Assay8	Assay9	Assay10
Ind1	2	2	2	2	2	0	2	2	2	2
Ind2	2	0	2	2	2	0	2	2	2	2
Ind3	2	3	2	0	2	0	2	2	2	2
Ind4	0	2	2	2	0	2	2	2	2	0
Ind5	2	2	2	2	2	2	2	2	0	2
Ind6	0	2	2	2	0	2	2	2	2	2
Ind7	2	2	2	2	2	2	0	0	2	2
Ind8	2	3	2	2	2	2	2	2	2	2
Ind9	2	2	2	2	2	2	2	2	2	0
Ind10	0	2	2	2	2	2	2	2	2	0

Imagen 7. Input requerido para el software R

Flapjack es una aplicación multiplataforma libre que proporciona visualizaciones interactivas de datos genotípicos de alto rendimiento, lo que permite una navegación rápida y comparaciones entre líneas, marcadores y cromosomas (Milne, 2010). La

dirección para descargar el software es https://ics.hutton.ac.uk/flapjack/download-flapjack/. Este software requiere dos archivos .txt para el *input*: un *Genotype file* (Imagen 8) que contiene la información genética y un *Map file* (Imagen 9), que informa en qué cromosoma se encuentra y qué posición tiene cada *assay* o marcador.

DNA \ Assay	Assay1	Assay2	Assay3	Assay4	Assay5	Assay6	Assay7	Assay8	Assay9	Assay10
Ind1	2	2	2	2	2	0	2	2	2	2
Ind2	2	0	2	2	2	0	2	2	2	2
Ind3	2	3	2	0	2	0	2	2	2	2
Ind4	0	2	2	2	0	2	2	2	2	0
Ind5	2	2	2	2	2	2	2	2	0	2
Ind6	0	2	2	2	0	2	2	2	2	2
Ind7	2	2	2	2	2	2	0	0	2	2
Ind8	2	3	2	2	2	2	2	2	2	2
Ind9	2	2	2	2	2	2	2	2	2	0
Ind10	0	2	2	2	2	2	2	2	2	0

Imagen 8. *Genotype file*, input requerido para el software Flapjack

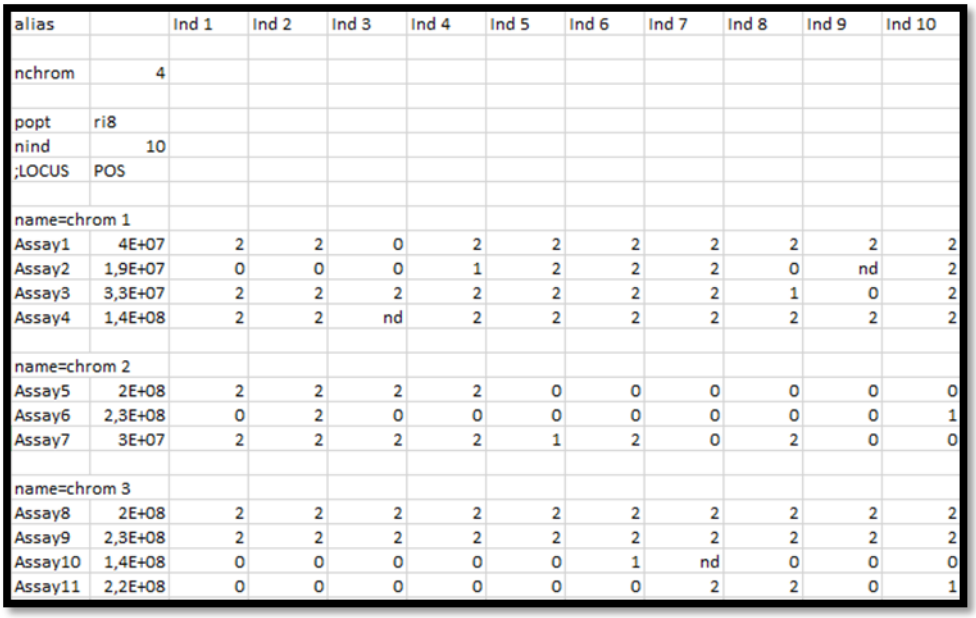
Assay1	Chr11	253224760
Assay2	Chr02	168705693
Assay3	Chr10	4110572
Assay4	Chr07	220036042
Assay5	Chr10	231962789
Assay6	Chr09	90249937
Assay7	Chr09	3957518
Assay8	Chr06	233765528
Assay9	Chr00	73468808
Assay10	Chr12	47495633

Imagen 9. *Map file*, input requerido para el software Flapjack

GGT

GGT es un software libre que permite la visualización de datos de marcadores moleculares (Ralph van Berloo, 2008). La dirección para descargar el software es https://www.wur.nl/en/show/GGT-2.0.htm. Este software presenta dos opciones para cargar los datos:

1. Copiar directamente un archivo Excel con el formato mostrado en la imagen 10, donde se indica los individuos del genotipado, el número de cromosomas, el tipo de población, y la clasificación de cada *assay* en los cromosomas, con las posiciones correspondientes, además de la información genotípica.

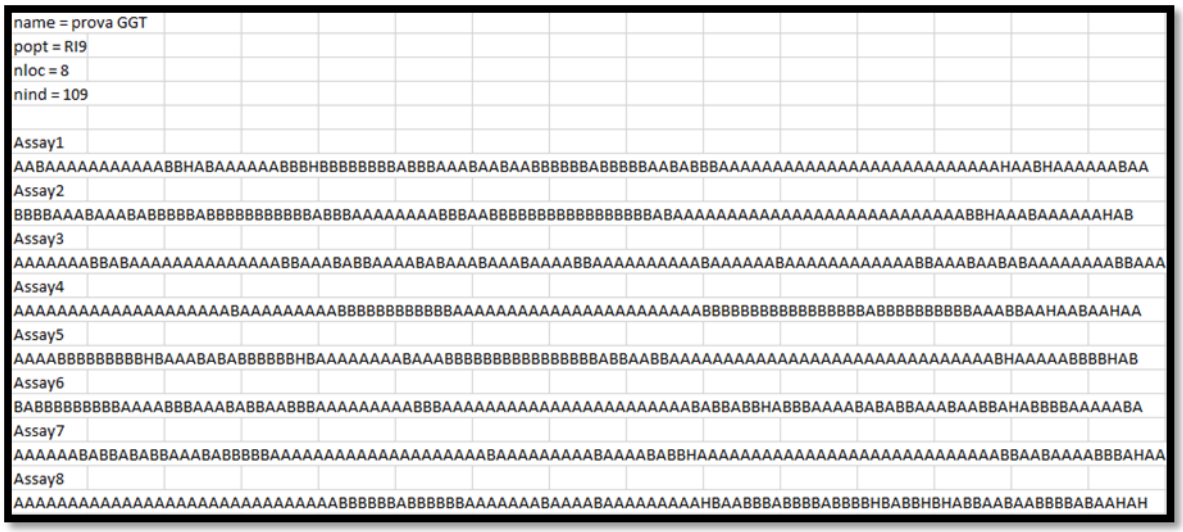


alias		Ind 1	Ind 2	Ind 3	Ind 4	Ind 5	Ind 6	Ind 7	Ind 8	Ind 9	Ind 10
nchrom	4										
popt	ri8										
nind	10										
;LOCUS	POS										
name=chrom 1											
Assay1	4E+07	2	2	0	2	2	2	2	2	2	2
Assay2	1,9E+07	0	0	0	1	2	2	2	0	nd	2
Assay3	3,3E+07	2	2	2	2	2	2	2	1	0	2
Assay4	1,4E+08	2	2	nd	2	2	2	2	2	2	2
name=chrom 2											
Assay5	2E+08	2	2	2	2	0	0	0	0	0	0
Assay6	2,3E+08	0	2	0	0	0	0	0	0	0	1
Assay7	3E+07	2	2	2	2	1	2	0	2	0	0
name=chrom 3											
Assay8	2E+08	2	2	2	2	2	2	2	2	2	2
Assay9	2,3E+08	2	2	2	2	2	2	2	2	2	2
Assay10	1,4E+08	0	0	0	0	0	1	nd	0	0	0
Assay11	2,2E+08	0	0	0	0	0	0	2	2	0	1

Imagen 10. Input requerido para el software GGT

2. Importar dos archivos .txt a través del software (Locus data y Map file):

Locus data (Imagen 11), contiene la información del tipo de población, el número de *assays* e individuos, y la información genotípica.



```
name = prova GGT
popt = RI9
nloc = 8
nind = 109

Assay1
AABAAAAAAAAAAABBHABAAAAAABBBHBBBBBBBBABBBAAABAABAABBBBBBBABBBBBAABABBBAAAAAAAAAAAAAAAAAAAAAAAAAAAHAABHAAAAAABAA
Assay2
BBBBAAABAAABABBBBBABBBBBBBBBBBABBBAAAAAAAABBBAABBBBBBBBBBBBBBBBABAAAAAAAAAAAAAAAAAAAAAAAAAAABBHAAABAAAAAAHAB
Assay3
AAAAAAABBABAAAAAAAAAAAAAABBAAABABBAAAABABAAABAAABAAAABBAAAAAAAAAABAAAAABAAAAAAAAAAABBAAABAABABAAAAAAAABBAAA
Assay4
AAAAAAAAAAAAAAAAAAAABAAAAAAAAABBBBBBBBBBBBAAAAAAAAAAAAAAAAAAAAAABBBBBBBBBBBBBBBBBABBBBBBBBBBAAABBAAHAABAAHAA
Assay5
AAAABBBBBBBBBHBAAABABABBBBBBHBAAAAAAAABAAABBBBBBBBBBBBBBBABBAABBAAAAAAAAAAAAAAAAAAAAAAAAAAAABHAAAAABBBBHAB
Assay6
BABBBBBBBBBAAAABBBAAABABBAABBBAAAAAAAAABBBAAAAAAAAAAAAAAAAAAAAABABBABBHABBBAAAABABABBAAABAABBAHABBBBAAAAABA
Assay7
AAAAAABABBABABBAAABABBBBBAAAAAAAAAAAAAAAAAAAABAAAAAAAAABAAAABABBHAAAAAAAAAAAAAAAAAAAAAAAAAAABBAABAAAABBBAHAA
Assay8
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAABBBBBBABBBBBBAAAAAAABAAAABAAAAAAAAHBAABBBABBBBABBBBHBABBHBHABBAABAABBBBABAAHAH
```

Imagen 11. Locus data, input requerido para el software GGT

Map file (imagen 12), informa en qué cromosoma y qué posición tiene cada *assay* o marcador:

chrom 1	
Assay1	39673793
Assay2	18698248
Assay3	32846489
chrom 2	
Assay4	19733884
Assay5	22567391
Assay6	29776660
chrom 3	
Assay7	20164848
Assay8	23443263

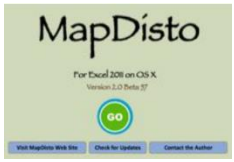
Imagen 12. Map file, input requerido para el software GGT

TASSEL es un software bioinformático libre diseñado para el análisis de la diversidad genómica de los cultivos (Bradbury, 2007). La dirección para descargar el software es http://www.maizegenetics.net/tassel. El input está en formato .txt, y mediante TASSEL se puede convertir en archivo .hmp para poder obtener una *Genotype table*, para realizar los estudios. El input utilizado en el trabajo no está codificado, ya que el programa no admite codificación externa, tiene su propio código mostrado en la imagen 3 del anexo.

Por lo tanto, se ha realizado el estudio con los datos reales, aunque encriptando los nombres de los marcadores (o *assays*) y los individuos (imagen 13).

rs#	alleles	chrom	pos	strand	assembly#	center	protLSID	assayLSID	panel	QCcode	Ind1	Ind2	Ind3	Ind4	Ind5
Assay29	C/A	Chr04	191703	+	TFM	NA	NA	NA	pepper	NA	CC	NN	CC	CC	CC
Assay217	T/G	Chr00	256281	+	TFM	NA	NA	NA	pepper	NA	GG	TG	GG	TG	TG
Assay87	C/G	Chr00	515298	+	TFM	NA	NA	NA	pepper	NA	GG	CG	CG	GG	NN
Assay155	C/T	Chr10	574391	+	TFM	NA	NA	NA	pepper	NA	CC	TT	CC	CC	CC
Assay225	A/T	Chr00	598891	+	TFM	NA	NA	NA	pepper	NA	AA	TT	TT	AA	TA
Assay219	G/A	Chr00	658948	+	TFM	NA	NA	NA	pepper	NA	GG	GA	AA	GA	AA
Assay96	A/C	Chr10	776145	+	TFM	NA	NA	NA	pepper	NA	CC	CC	CC	CC	CC
Assay70	A/C	Chr07	973990	+	TFM	NA	NA	NA	pepper	NA	NN	NN	CC	NN	CC
Assay88	A/G	Chr08	1282222	+	TFM	NA	NA	NA	pepper	NA	AA	GG	AA	AA	AA
Assay122	T/C	Chr00	1425275	+	TFM	NA	NA	NA	pepper	NA	TT	TT	TT	TT	TC
Assay56	C/T	Chr03	1595564	+	TFM	NA	NA	NA	pepper	NA	CC	TT	TT	TT	TT
Assay186	G/A	Chr10	1679555	+	TFM	NA	NA	NA	pepper	NA	GG	NN	GG	AA	NN
Assay43	T/A	Chr10	1795181	+	TFM	NA	NA	NA	pepper	NA	TT	TT	AA	TT	AA
Assay102	A/G	Chr10	1883936	+	TFM	NA	NA	NA	pepper	NA	AA	AA	AA	GG	AA
Assay116	C/T	Chr11	1904733	+	TFM	NA	NA	NA	pepper	NA	CC	CC	CC	CC	CC

Imagen 13. Input requerido para el software TASSEL



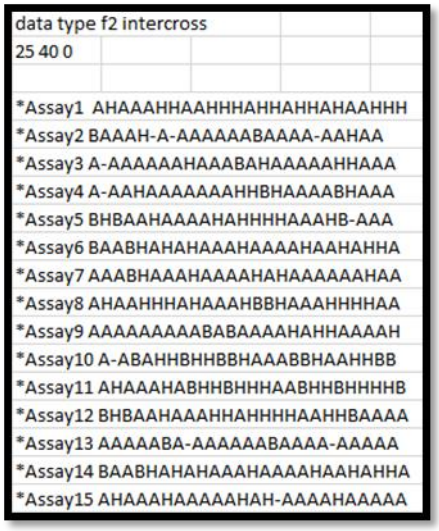
MapDisto es un software libre para realizar mapas de marcadores genéticos, encontrar *linkage groups*, exportar datos y mapas de otros softwares… (Heffelfinger, 2017). La dirección para descargar el software es http://mapdisto.free.fr/Download_Soft/. El *input* (imagen 14) contiene la información genotípica y se inserta al software copiando los datos del archivo Excel. Una vez se han insertado los datos, se configuran las opciones (número de individuos, de marcadores, tipo de población, codificación del genotipado…) y se realiza el análisis.

DNA	Assay1	Assay2	Assay3	Assay4	Assay5	Assay6	Assay7	Assay8	Assay9	Assay10
Ind 1	2	0	2	2	2	0	2	2	2	0
Ind 2	2	0	2	2	2	1	2	2	2	0
Ind 3	0	0	2	2	2	nd	2	2	2	0
Ind 4	2	0	2	2	2	1	2	2	2	0
Ind 5	2	2	1	2	0	0	2	2	2	0
Ind 6	2	2	2	2	0	0	1	2	2	0
Ind 7	2	2	2	nd	0	0	0	2	2	0
Ind 8	2	0	0	1	0	0	1	2	2	1
Ind 9	2	2	0	2	0	0	0	2	2	0
Ind 10	2	2	nd	1	0	0	0	2	1	0

Imagen 14. Input requerido para el software MapDisto

MapMaker

MapMaker es un software libre de mapeo genético (Lander, 1987). La dirección para descargar el software es http://www.softpedia.com/get/Science-CAD/MapMaker.shtml. Se basa en cálculos matemáticos para generar mapas genéticos, ordenando mediante cálculos de frecuencias alélicas los marcadores para formar los distintos grupos de ligamiento. El tipo de archivo requerido es .RAW (imagen 15). La primera columna debe contener antes del nombre del marcador o assay un asterisco (*), lo que permite al programa reconocerlo como marcador. Es muy importante guardar el archivo en la misma carpeta del programa. Éste contiene la información del tipo de población (F2), el número de individuos (29), el número de marcadores (10) y los datos fenotípicos en el archivo (en este caso, 0).



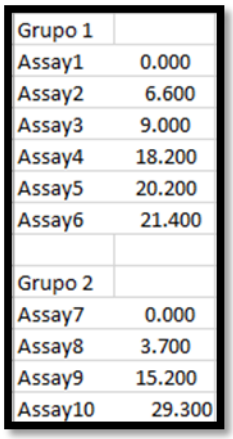
data type f2 intercross
25 40 0
*Assay1 AHAAAHHAAHHHAHHAHHAHAAHHH
*Assay2 BAAAH-A-AAAAAABAAAA-AAHAA
*Assay3 A-AAAAAAHAAABAHAAAAAHHAAA
*Assay4 A-AAHAAAAAAAHHBHAAAABHAAA
*Assay5 BHBAAHAAAAHAHHHHAAAHB-AAA
*Assay6 BAABHAHAHAAAHAAAAHAAHAHHA
*Assay7 AAABHAAAHAAAAHAHAAAAAAHAA
*Assay8 AHAAHHHAHAAAHBBHAAAHHHHAA
*Assay9 AAAAAAAAABABAAAAHAHHAAAAH
*Assay10 A-ABAHHBHHBBHAAABBHAAHHBB
*Assay11 AHAAAHABHHBHHHAABHHBHHHHB
*Assay12 BHBAAHAAAHHAHHHHAAHHBAAAA
*Assay13 AAAAABA-AAAAAABAAAA-AAAAA
*Assay14 BAABHAHAHAAAHAAAAHAAHAHHA
*Assay15 AHAAAHAAAAAHAH-AAAAHAAAAA

Imagen 15. Input requerido para el software MapMaker

MapChart

MapChart es un software para la representación gráfica de mapas de ligamiento genéticos y QTLs (*Quantitative Trait Loci*). La representación puede ser de grupos de ligamiento o de cromosomas y, están representados en una secuencia de barras verticales (Voorrips, 2002). La dirección para descargar el software es https://www.wur.nl/en/show/Mapchart.htm.

El *input* (imagen 16) contiene la información de cada grupo de ligamiento con la información de distancia de los marcadores en centimorgans (cM), y se inserta al software copiando los datos del archivo Excel o txt.



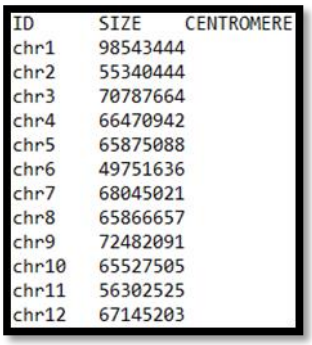
Grupo 1	
Assay1	0.000
Assay2	6.600
Assay3	9.000
Assay4	18.200
Assay5	20.200
Assay6	21.400
Grupo 2	
Assay7	0.000
Assay8	3.700
Assay9	15.200
Assay10	29.300

Imagen 16. Input requerido para el software MapChart

PhenoGram

PhenoGram es una herramienta de software versátil y fácil de usar que fomenta la exploración y el intercambio de información genómica. Mediante la visualización de datos, los investigadores pueden explorar y compartir resultados complejos, lo que facilita una mayor comprensión de estos datos (Wolfe et al. 2013). La dirección web para realizar los análisis online es http://visualization.ritchielab.org/phenograms/plot.

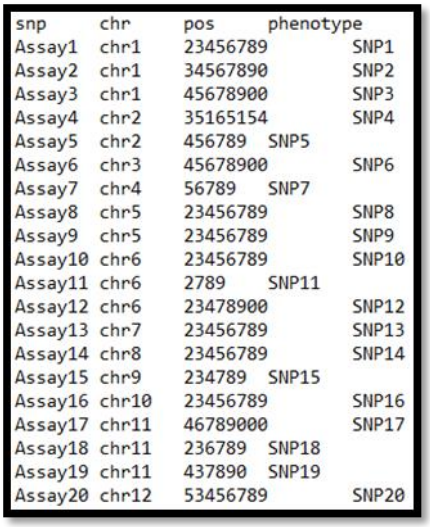
Este software requiere dos inputs, un *Genome file* y un *Input file*, ambos en .txt. El archivo *Genome file* (imagen 17), informa sobre la longitud total de cada cromosoma, además de poder añadir la posición del centrómero:



ID	SIZE	CENTROMERE
chr1	98543444	
chr2	55340444	
chr3	70787664	
chr4	66470942	
chr5	65875088	
chr6	49751636	
chr7	68045021	
chr8	65866657	
chr9	72482091	
chr10	65527505	
chr11	56302525	
chr12	67145203	

Imagen 17. Genome file, input requerido para el software PhenoGram

El *Input file* (imagen 18), indica en qué posición y cromosoma se encuentra cada marcador:



snp	chr	pos	phenotype
Assay1	chr1	23456789	SNP1
Assay2	chr1	34567890	SNP2
Assay3	chr1	45678900	SNP3
Assay4	chr2	35165154	SNP4
Assay5	chr2	456789	SNP5
Assay6	chr3	45678900	SNP6
Assay7	chr4	56789	SNP7
Assay8	chr5	23456789	SNP8
Assay9	chr5	23456789	SNP9
Assay10	chr6	23456789	SNP10
Assay11	chr6	2789	SNP11
Assay12	chr6	23478900	SNP12
Assay13	chr7	23456789	SNP13
Assay14	chr8	23456789	SNP14
Assay15	chr9	234789	SNP15
Assay16	chr10	23456789	SNP16
Assay17	chr11	46789000	SNP17
Assay18	chr11	236789	SNP18
Assay19	chr11	437890	SNP19
Assay20	chr12	53456789	SNP20

Imagen 18. Input file, input requerido para el software PhenoGram

2.3 Marker-Assisted Backcrossing (MABC)

La población mundial está aumentando muy rápidamente, reduciendo el tamaño de tierra cultivable, disminuyendo el agua, surgiendo nuevas enfermedades y plagas, el cambio climático… Debido a estos factores se tienen que desarrollar variedades de cultivo sostenibles con resistencia al estrés biótico y abiótico. Los avances en genética, genómica y fisiología han abierto nuevas oportunidades para reducir el impacto de los factores negativos (Hasan et al. 2015).

El Marker assisted backcrossing (MABC) es uno de los métodos más prometedores en el ámbito de los marcadores moleculares. MABC puede contribuir a desarrollar variedades de elevado interés, incorporando un gen determinado de un individuo donante, que controle alguna resistencia o característica deseada, a un individuo recurrente, el cual ya presenta muchas características de interés para el mejorador. Así pues, el objetivo final es mejorar aún más el individuo recurrente con nuevas características.

Los archivos obtenidos en el laboratorio de MABC son muy voluminosos, ya que un mismo archivo contiene más de un MABC. Para poder entender bien el objetivo de este apartado, se realizará con un archivo ficticio con un solo gen de interés para incorporar al individuo recurrente.

Archivo de ejemplo:

	Assay1	Assay2	Assay3	Assay4	Assay5	Assay6	Assay7	Assay8	Assay9	Assay10
Recurrente	A	C	T	A	G	C	A	T	G	T
Donante	C	T	T	C	A	T	A	C	C	T
Ind1	A	C	T	A	G/A	C	A	T	G	T
Ind2	A	C	T	A	G/A	C	A	T	G	T
Ind3	A	C	T	A	G/A	T	A	T/C	G	T
Ind4	A	C	G	C	G/A	C	A	T/C	G/C	T
Ind5	C	C	T	A	G	C	T	C	G	A
Ind6	C	C	T	A	G	C	A	C	C	A
Ind7	A	C	T	A	A	C	A	C	C	T
Ind8	A	C	T	A	A	T	A	T	G/C	T/A
Ind9	A	T	T	C	A	T	T	T	C	T
Ind10	C	T	G	C	G	C	T	T	G	T/A
Ind11	A	T	T	C	G/A	T	T	T/C	G/C	T
Ind12	A	T	T	C	G	C	T	T/C	G	T
Ind13	A	T	T	C	A	C	A	C	G	T
Ind14	C	C	G	C	G	C	A	C	G	T
Ind15	C	T/C	G	C	G	C	T	T	G/C	A
Ind16	C	T/C	G	A	G	C	T	T	C	A
Ind17	A	C	T	A	G/A	C	A	T	C	T
Ind18	A	C	T	A	G/A	C	T	C	G	T
Ind19	C	C	T/G	A	A	T	A/T	C	G	T
Ind20	C	C	G	A	A	C	A	C	G	T/A

Imagen 19. Input requerido en R para realitzar los anàlisis de MABC

El archivo presenta 20 individuos (Ind1-Ind20), los cuales son la primera generación (F1) entre el cruzamiento del individuo recurrente y el donante. La codificación usada en este tipo de análisis es siempre la B para los alelos del individuo recurrente, ya sea homocigoto o heterocigoto. Por lo tanto, si el individuo donante y los individuos de la F1 presentan el mismo alelo que el recurrente, tendrán una B, y si presentan otro alelo tendrán una A o una H, en función de ser homocigotos o heterocigotos, respectivamente.

Una vez tenemos seleccionados los individuos de la F1 heterocigotos por el marcador de interés, mediante una matriz de distancia seleccionaremos qué individuos se parecen más al recurrente.

La función ha sido desarrollada en R y es la siguiente:

```
Datos<-read.csv("mabc.csv", sep=";")

MABC<-function(x,y){
  library(car)
  x<-x[-2,]
  Nombre_Datos<-x[,1]
  x<-x[,-1]
  Prova<-c()
  for (i in 2:nrow(x)) {
    for (n in 1:length(x)) {
      B<-matrix(if(x[1,n]==x[i,n]) {recode(as.matrix(x[i,n]),
"'C'='B'; 'T'='B'; 'G'='B'; 'A'='B'; 'A/G'='B'; 'G/A'='B'; 'A/T'='B';
'T/A'='B'; 'A/C'='B'; 'C/A'='B'; 'T/C'='B'; 'C/T'='B'; 'T/G'='B';
'G/T'='B'; 'G/C'='B'; 'C/G'='B'")}
                else{recode(as.matrix(x[i,n]), "'C'='A'; 'T'='A';
'G'='A'; 'A'='A'")})
      Prova<-c(Prova,B)
    }
  }
  Prova<-matrix(Prova, ncol=nrow(x)-1)
  Prova<-t(Prova)
  Prova<-data.frame(Prova)
  Recurrent<-rep("B", ncol(x))
  Prova<-rbind(Recurrent,Prova)
  names(Prova)<-names(x)
  PM<-data.frame(recode(as.matrix(Prova), "'A/G'='H'; 'G/A'='H';
'A/T'='H'; 'T/A'='H'; 'A/C'='H'; 'C/A'='H'; 'T/C'='H'; 'C/T'='H';
'T/G'='H'; 'G/T'='H'; 'G/C'='H'; 'C/G'='H'"))
  PM<-`row.names<-`(PM, Nombre_Datos)

  In<-subset(PM, PM[,y]=="H")
  Z<-subset(In, In[,c(y-1)]=="B" & In[,c(y+1)]=="B")
  Q<-subset(In, In[,c(y-1)]=="H" & In[,c(y+1)]=="B")
  S<-subset(In, In[,c(y-1)]=="B" & In[,c(y+1)]=="H")
```

```
  SM<-rbind(Z,Q,S)

  Recurrent<-rep("B", ncol(SM))
  SMr<-rbind(Recurrent,SM)
  SMc<-recode(as.matrix(SMr), "'B'=2; 'H'=1; 'A'=0")
  Nombres_SMc<-rownames(SMr)
  SMc<-matrix(SMc, ncol=ncol(x))
  SMc<-cbind(Nombres_SMc, SMc)
  SMc<-`row.names<-`(SMc, Nombres_SMc)

  TM<-dist(SMc)

  return(TM)
}
```

3. Resultados

3.1 Estudiar los parámetros básicos a partir de datos genotípicos

Los datos de genotipado obtenidos en el laboratorio tienen que ser preprocesados para su posterior análisis, ya sea por compatibilidad de formato, por codificación de los datos o ambos. El primer objetivo de este trabajo ha sido realizar una codificación para convertir los datos nucleotídicos en numéricos, mediante R, para su posterior análisis.
El proceso de codificación realizado con R nos permite obtener los datos con la apariencia de la imagen 20, en la que: 2 representa el alelo mayoritario, 1 los heterocigotos, 0 el alelo minoritario y 3 los datos faltantes o *missing data,* por cada *assay*:

DNA \ Assay	Assay1	Assay2	Assay3	Assay4	Assay5	Assay6	Assay7	Assay8	Assay9	Assay10
Ind1	2	2	2	2	2	0	2	2	2	2
Ind2	2	0	2	2	2	0	2	2	2	2
Ind3	2	3	2	0	2	0	2	2	2	2
Ind4	0	2	2	2	0	2	2	2	2	0
Ind5	2	2	2	2	2	2	2	2	0	2
Ind6	0	2	2	2	0	2	2	2	2	2
Ind7	2	2	2	2	2	2	0	0	2	2
Ind8	2	3	2	2	2	2	2	2	2	2
Ind9	2	2	2	2	2	2	2	2	2	0
Ind10	0	2	2	2	2	2	2	2	2	0
Ind11	2	2	2	2	2	0	2	0	2	2
Ind12	0	2	2	2	0	2	2	2	2	2
Ind13	0	2	2	2	2	2	2	2	0	0
Ind14	0	2	2	2	2	2	2	2	0	2
Ind15	2	2	2	2	2	2	2	2	0	2
Ind16	1	3	2	0	2	2	2	0	2	2
Ind17	2	2	2	2	2	2	2	0	0	2
Ind18	2	2	2	2	2	2	2	2	0	2
Ind19	2	2	2	2	0	2	2	2	2	2
Ind20	2	2	2	3	0	2	0	2	2	2

Imagen 20. Resultado de la codificación de los datos genotípicos mediante R

Una vez obtenida la codificación, los datos ya están listos para su análisis. Los parámetros básicos de interés son heterocigosidad, *inbreeding*, PIC y *missing data*. Aplicando la función de cálculo de los parámetros, se obtiene el siguiente *output* en Excel por cada *assay* (imagen 21) o por cada individuo (imagen 22):

	Heterocigosidad	Inbreeding	PIC	Missing data
Assay1	0,461656805	0,666752115	0,355093302	0,037037037
Assay2	0,120707596	0,732758621	0,113422434	0,081481481
Assay3	0,277777778	0,836363636	0,239197531	0,022222222
Assay4	0,404730411	0,862734418	0,322827058	0,066666667
Assay5	0,367309458	0,79375	0,299851339	0,022222222
Assay6	0,267791495	0,751050097	0,231935353	0
Assay7	0,215944531	0,758091554	0,192628511	0,007407407
Assay8	0,332409972	0,728571429	0,277161777	0,014814815
Assay9	0,36170096	0,918082524	0,296287168	0
Assay10	0,375	0,820895522	0,3046875	0,007407407
Assay11	0,484587145	0,747955748	0,367174795	0,02962963
Assay12	0,407709191	0,763811318	0,324595799	0
Assay13	0,031991539	1	0,03147981	0,088888889
Assay14	0,276103793	0,782145782	0,237987141	0,014814815
Assay15	0,404912658	0,832879581	0,322935527	0,014814815
Assay16	0,330585877	0,683962264	0,275942366	0,007407407
Assay17	0,388001865	0,74423669	0,312729141	0,02962963
Assay18	0,49899755	0,850446429	0,374498272	0,007407407
Assay19	0,349766804	0,767040552	0,288598395	0
Assay20	0,473240143	0,763459841	0,361262026	0,007407407

Imagen 21. Resultado del análisis de los parámetros básicos

Heterocigosidad por individuo:

Individuo1	0,404179
Individuo2	0,417362
Individuo3	0,30864
Individuo4	0,481004
Individuo5	0,369482
Individuo6	0,357476
Individuo7	0,405562
Individuo8	0,355647
Individuo9	0,372854
Individuo10	0,381328
Individuo11	0,418289
Individuo12	0,370763
Individuo13	0,336389
Individuo14	0,324509
Individuo15	0,314411
Individuo16	0,364478
Individuo17	0,324045
Individuo18	0,298995
Individuo19	0,339127
Individuo20	0,429077

Imagen 22. Resultado del análisis de la heterocigosidad por individuo

3.2 Analizar los datos genéticos mediante diversos softwares y comparación de los resultados

i. Analizar distancias genéticas.

Se han analizado las distancias genéticas importando en los cuatro softwares los mismos datos, aunque variando el formato de presentación de estos en función de las necesidades de los softwares. No todos los softwares proporcionan la misma información: pues en R podemos obtener tanto la matriz de similitud como la de distancia, en GGT y TASSEL se obtiene la matriz de distancia, y en Flapjack se obtiene la matriz de similitud. El dendrograma sí se puede obtener en todos los softwares, en cambio, el PCA se puede obtener en R, Flapjack y TASSEL, pero no en GGT. La siguiente tabla detalla los resultados que podemos obtener en cada uno de los programas (tabla 1):

Tabla 1. Análisis que podemos obtener a partir de los 4 softwares

	Matriz de similidud	Matriz de distancia	Dendrograma	PCA
R	Si	Si	Si	Si
Flapjack	Si	No	Si	Si
GGT	No	Si	Si	No
TASSEL	No	Si	Si	Si

A continuación, se muestran los resultados obtenidos en los diferentes análisis para todos los softwares:

MATRIZ DE SIMILITUD

R

	Ind1	Ind2	Ind3	Ind4	Ind5	Ind6	Ind7	Ind8	Ind9
Ind1	1,00	0,76	0,81	0,64	0,69	0,65	0,74	0,78	0,85
Ind2	0,76	1,00	0,77	0,58	0,70	0,65	0,71	0,71	0,70
Ind3	0,81	0,77	1,00	0,64	0,74	0,73	0,74	0,76	0,79
Ind4	0,64	0,58	0,64	1,00	0,69	0,78	0,65	0,66	0,71
Ind5	0,69	0,70	0,74	0,69	1,00	0,82	0,71	0,70	0,73
Ind6	0,65	0,65	0,73	0,78	0,82	1,00	0,68	0,68	0,71
Ind7	0,74	0,71	0,74	0,65	0,71	0,68	1,00	0,71	0,72
Ind8	0,78	0,71	0,76	0,66	0,70	0,68	0,71	1,00	0,82
Ind9	0,85	0,70	0,79	0,71	0,73	0,71	0,72	0,82	1,00
Ind10	0,84	0,75	0,78	0,71	0,68	0,72	0,73	0,75	0,81

Imagen 23. Fragmento inicial de la matriz de similitud obtenida mediante R

FLAPJACK

	Ind1	Ind2	Ind3	Ind4	Ind5	Ind6	Ind7	Ind8	Ind9	Ind10
Ind1	1,00	0,62	0,73	0,51	0,57	0,54	0,64	0,66	0,79	0,79
Ind2	0,62	1,00	0,64	0,41	0,59	0,49	0,57	0,58	0,54	0,59
Ind3	0,73	0,64	1,00	0,51	0,63	0,64	0,65	0,60	0,70	0,69
Ind4	0,51	0,41	0,51	1,00	0,58	0,70	0,56	0,50	0,62	0,60
Ind5	0,57	0,59	0,63	0,58	1,00	0,74	0,61	0,54	0,62	0,55
Ind6	0,54	0,49	0,64	0,70	0,74	1,00	0,58	0,51	0,61	0,63
Ind7	0,64	0,57	0,65	0,56	0,61	0,58	1,00	0,53	0,60	0,63
Ind8	0,66	0,58	0,60	0,50	0,54	0,51	0,53	1,00	0,70	0,63
Ind9	0,79	0,54	0,70	0,62	0,62	0,61	0,60	0,70	1,00	0,74
Ind10	0,79	0,59	0,69	0,60	0,55	0,63	0,63	0,63	0,74	1,00

Imagen 24. Fragmento inicial de la matriz de similitud obtenida mediante Flapjack

MATRIZ DE DISTANCIA

R

	Ind1	Ind2	Ind3	Ind4	Ind5	Ind6	Ind7	Ind8	Ind9
Ind1	0,00	0,24	0,19	0,36	0,31	0,35	0,26	0,22	0,15
Ind2	0,24	0,00	0,23	0,42	0,30	0,35	0,29	0,29	0,30
Ind3	0,19	0,23	0,00	0,36	0,26	0,27	0,26	0,24	0,21
Ind4	0,36	0,42	0,36	0,00	0,31	0,22	0,35	0,34	0,29
Ind5	0,31	0,30	0,26	0,31	0,00	0,18	0,29	0,30	0,27
Ind6	0,35	0,35	0,27	0,22	0,18	0,00	0,32	0,33	0,29
Ind7	0,26	0,29	0,26	0,35	0,29	0,32	0,00	0,29	0,28
Ind8	0,22	0,29	0,24	0,34	0,30	0,33	0,29	0,00	0,18
Ind9	0,15	0,30	0,21	0,29	0,27	0,29	0,28	0,18	0,00
Ind10	0,16	0,25	0,22	0,29	0,32	0,28	0,27	0,25	0,19

Imagen 25. Fragmento inicial de la matriz de distancia obtenida mediante R

GGT

	Ind1	Ind2	Ind3	Ind4	Ind5	Ind6	Ind7	Ind8	Ind9	Ind10
Ind1	0									
Ind2	0,38	0								
Ind3	0,27	0,36	0							
Ind4	0,49	0,59	0,49	0						
Ind5	0,43	0,41	0,37	0,42	0					
Ind6	0,46	0,51	0,36	0,3	0,26	0				
Ind7	0,36	0,43	0,35	0,44	0,39	0,42	0			
Ind8	0,34	0,42	0,4	0,5	0,46	0,49	0,47	0		
Ind9	0,21	0,46	0,3	0,38	0,38	0,39	0,4	0,3	0	
Ind10	0,21	0,41	0,31	0,4	0,45	0,37	0,37	0,37	0,26	0

Imagen 26. Fragmento inicial de la matriz de distancia obtenida mediante GGT

TASSEL

	Ind1	Ind2	Ind3	Ind4	Ind5	Ind6	Ind7	Ind8	Ind9	Ind10
Ind1	0,00	0,29	0,26	0,42	0,35	0,40	0,32	0,22	0,21	0,24
Ind2	0,29	0,00	0,26	0,50	0,38	0,42	0,37	0,32	0,37	0,31
Ind3	0,26	0,26	0,00	0,45	0,33	0,36	0,32	0,28	0,28	0,27
Ind4	0,42	0,50	0,45	0,00	0,36	0,28	0,41	0,39	0,37	0,37
Ind5	0,35	0,38	0,33	0,36	0,00	0,23	0,35	0,33	0,32	0,36
Ind6	0,40	0,42	0,36	0,28	0,23	0,00	0,40	0,37	0,37	0,35
Ind7	0,32	0,37	0,32	0,41	0,35	0,40	0,00	0,32	0,34	0,30
Ind8	0,22	0,32	0,28	0,39	0,33	0,37	0,32	0,00	0,20	0,26
Ind9	0,21	0,37	0,28	0,37	0,32	0,37	0,34	0,20	0,00	0,23
Ind10	0,24	0,31	0,27	0,37	0,36	0,35	0,30	0,26	0,23	0,00

Imagen 27. Fragmento inicial de la matriz de distancia obtenida mediante TASSEL

DENDROGRAMAS

R

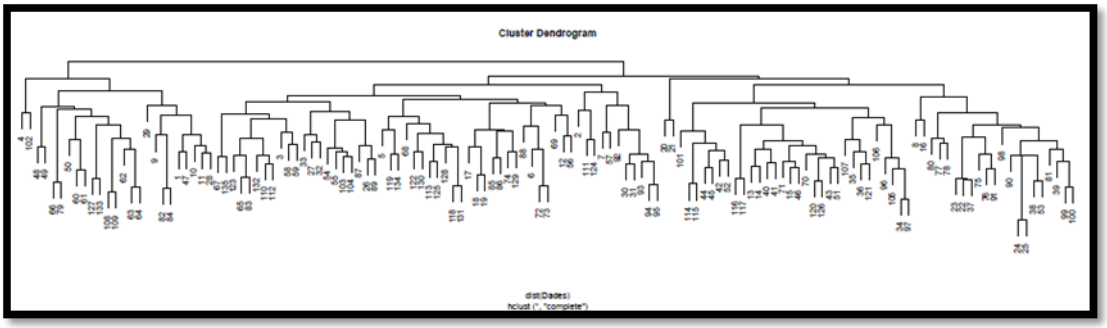


Imagen 28. Dendrograma obtenido mediante R

FLAPJACK

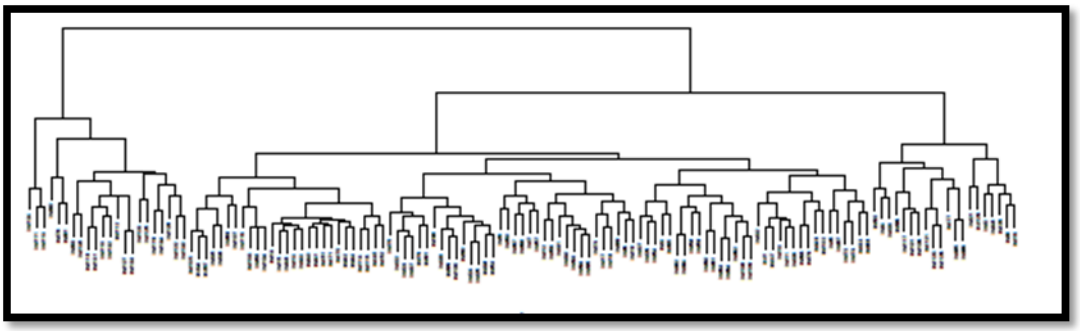


Imagen 29. Dendrograma obtenido mediante Flapjack

GGT

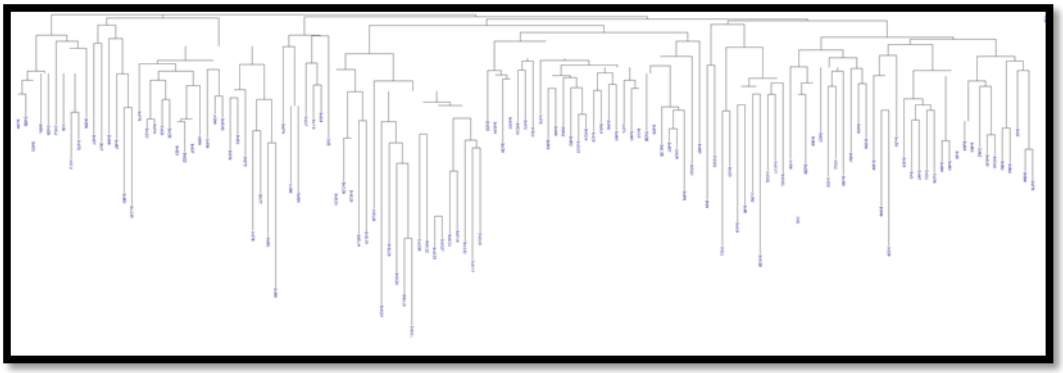


Imagen 30. Dendrograma obtenido mediante GGT

TASSEL

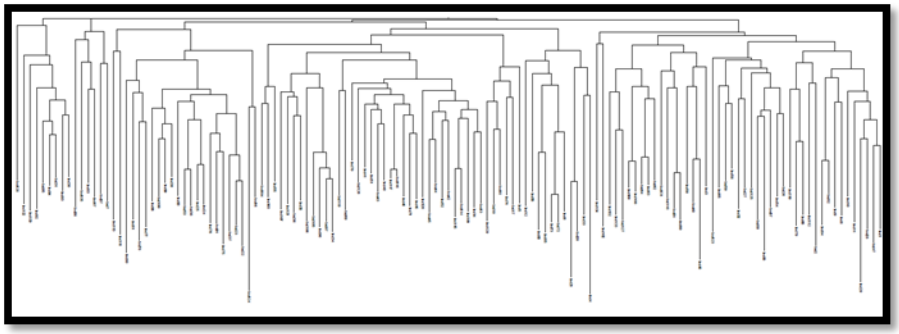


Imagen 31. Dendrograma obtenido mediante TASSEL

PCA

R

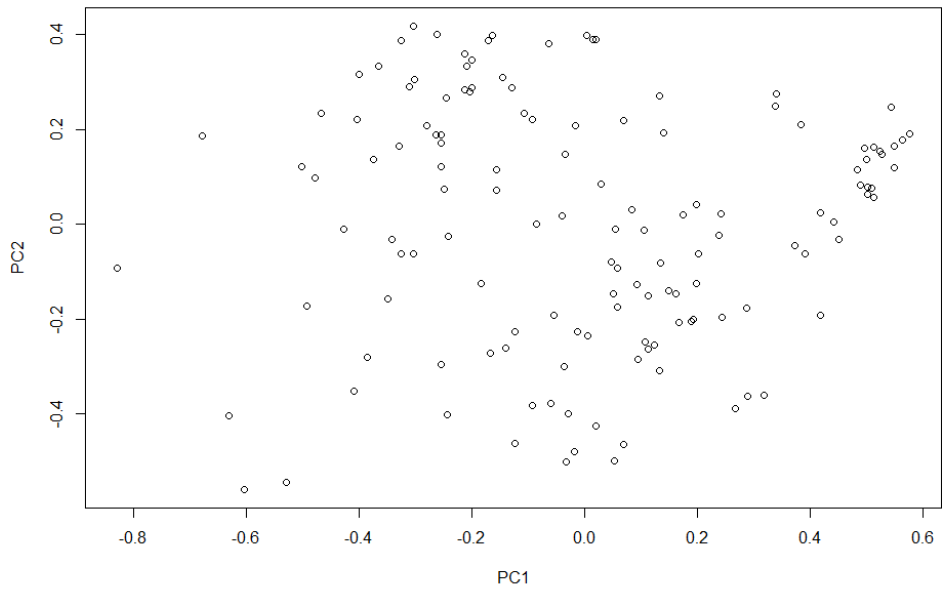


Imagen 32. PCA obtenido mediante R

FLAPJACK

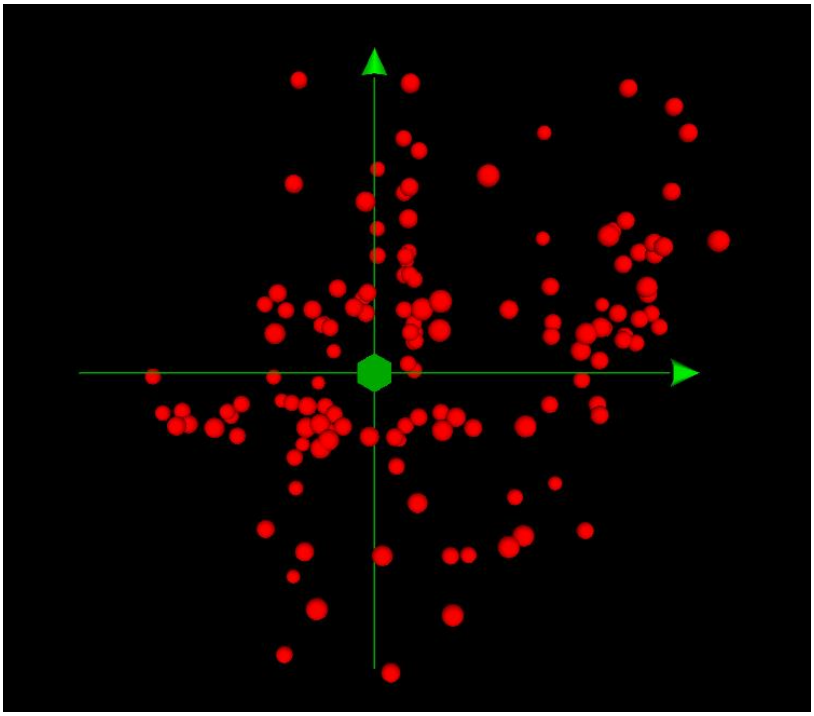


Imagen 33. PCA obtenido mediante Flapjack

TASSEL

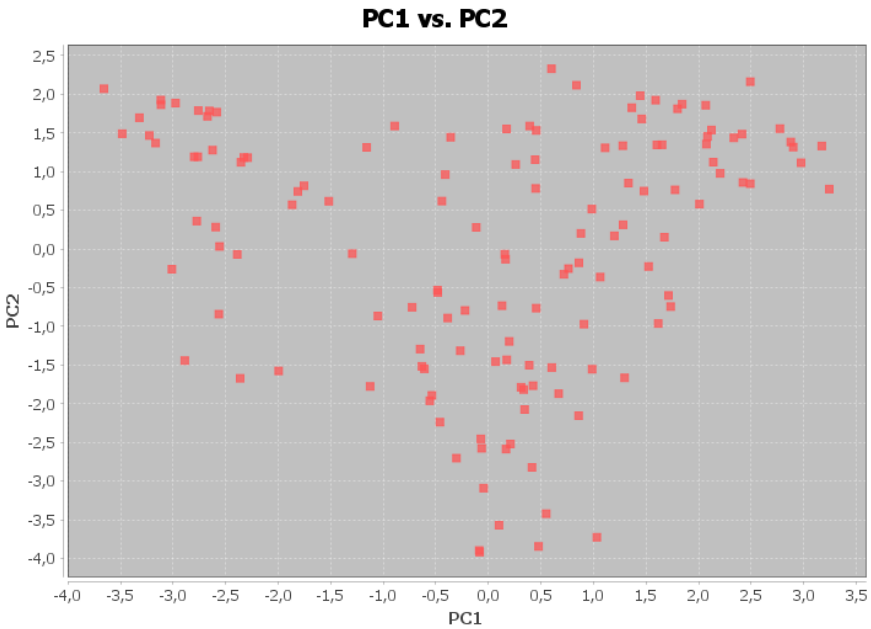


Imagen 34. PCA obtenido mediante TASSEL

La comparación de todos estos resultados de forma manual es una tarea muy larga y laboriosa. Para poder realizar un análisis exhaustivo y relativamente rápido, se han insertado las matrices de distancia (en el caso de no tenerla, se ha obtenido a partir de la matriz de similitud) en R para su comparación:

```
R<-read.table("Distance matrix R.txt")
R<-as.matrix(R)
R[1:5,1:5]

##           Ind1      Ind2      Ind3      Ind4      Ind5
## Ind1 0.0000000 0.2368056 0.1861111 0.3583333 0.3097222
## Ind2 0.2368056 0.0000000 0.2270833 0.4229167 0.2965278
## Ind3 0.1861111 0.2270833 0.0000000 0.3555556 0.2569444
## Ind4 0.3583333 0.4229167 0.3555556 0.0000000 0.3055556
## Ind5 0.3097222 0.2965278 0.2569444 0.3055556 0.0000000

FLAPJACK<-read.table("Distance matrix Flapjack.txt")
FLAPJACK<-as.matrix(FLAPJACK)
FLAPJACK[1:5,1:5]

##           Ind1      Ind2      Ind3      Ind4      Ind5
## Ind1 0.0000000 0.3765690 0.2677825 0.4853557 0.4309623
## Ind2 0.3765690 0.0000000 0.3556485 0.5941423 0.4142259
## Ind3 0.2677825 0.3556485 0.0000000 0.4853557 0.3682008
## Ind4 0.4853557 0.5941423 0.4853557 0.0000000 0.4184101
## Ind5 0.4309623 0.4142259 0.3682008 0.4184101 0.0000000

TASSEL<-read.table("Distance Matrix Tassel.txt", header = TRUE)
TASSEL<-as.matrix(TASSEL)
TASSEL[1:5,1:5]

##           Ind1      Ind2      Ind3      Ind4      Ind5
## Ind1 0.0000000 0.2910798 0.2564655 0.4191489 0.3490991
## Ind2 0.2910798 0.0000000 0.2605634 0.5046729 0.3827751
## Ind3 0.2564655 0.2605634 0.0000000 0.4484979 0.3280543
## Ind4 0.4191489 0.5046729 0.4484979 0.0000000 0.3609866
## Ind5 0.3490991 0.3827751 0.3280543 0.3609866 0.0000000

GGT<-read.table("Distance matrix GGT.txt")
GGT<-as.matrix(GGT)
GGT[65:69,65:69]

##      Ind1 Ind2 Ind3 Ind4 Ind5
## Ind1 0.00 0.38 0.27 0.49 0.43
## Ind2 0.38 0.00 0.36 0.59 0.41
## Ind3 0.27 0.36 0.00 0.49 0.37
## Ind4 0.49 0.59 0.49 0.00 0.42
## Ind5 0.43 0.41 0.37 0.42 0.00
```

Una vez incorporadas todas las matrices, se han realizado los PCAs y dendrogramas, mediante prcomp() y hclust(), respectivamente.

```
PCA_R<-prcomp(R)
plot(PCA_R$x[,1],PCA_R$x[,2],xlab="PC1",ylab="PC2")
```

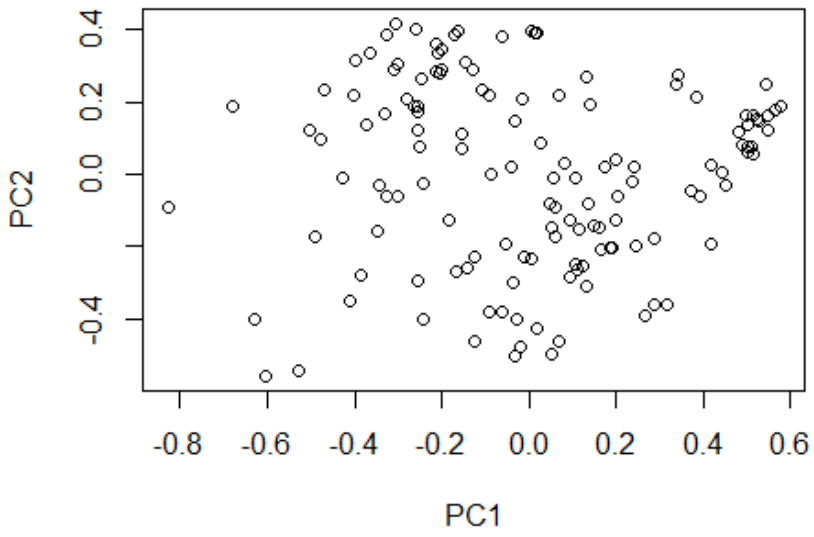


Imagen 35. PCA creado a partir de los datos de R

```
R_clust<-hclust(as.dist(R))
plot(R_clust)
```

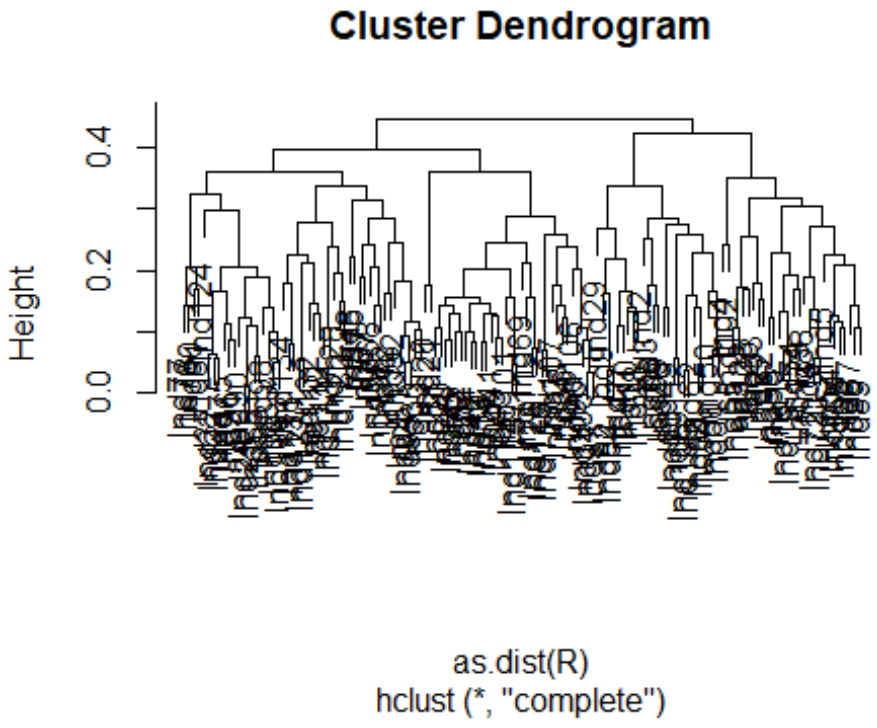


Imagen 36. Dendrograma creado a partir de los datos de R

```
PCA_FLAPJACK<-prcomp(FLAPJACK)
plot(PCA_FLAPJACK$x[,1],PCA_FLAPJACK$x[,2],xlab="PC1",ylab="PC2")
```

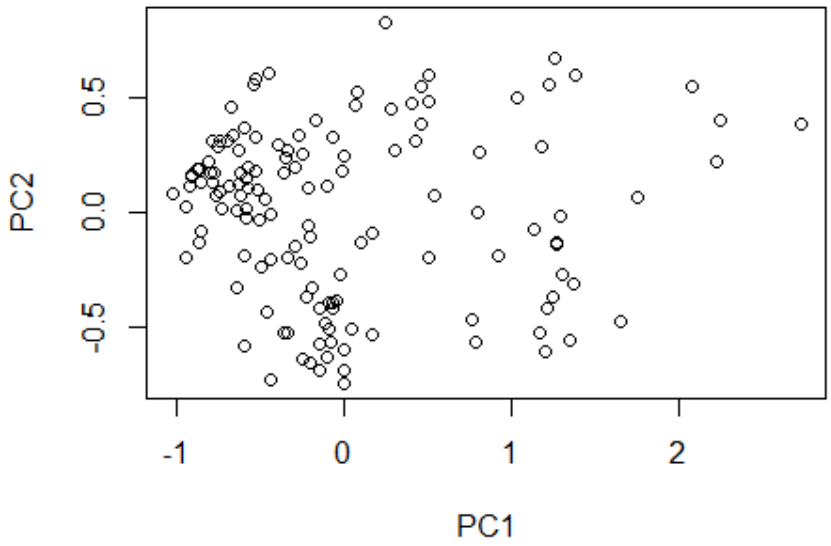


Imagen 37. PCA creado en R a partir de los datos de Flapjack

```
FLAPJACK_clust<-hclust(as.dist(FLAPJACK))
plot(FLAPJACK_clust)
```

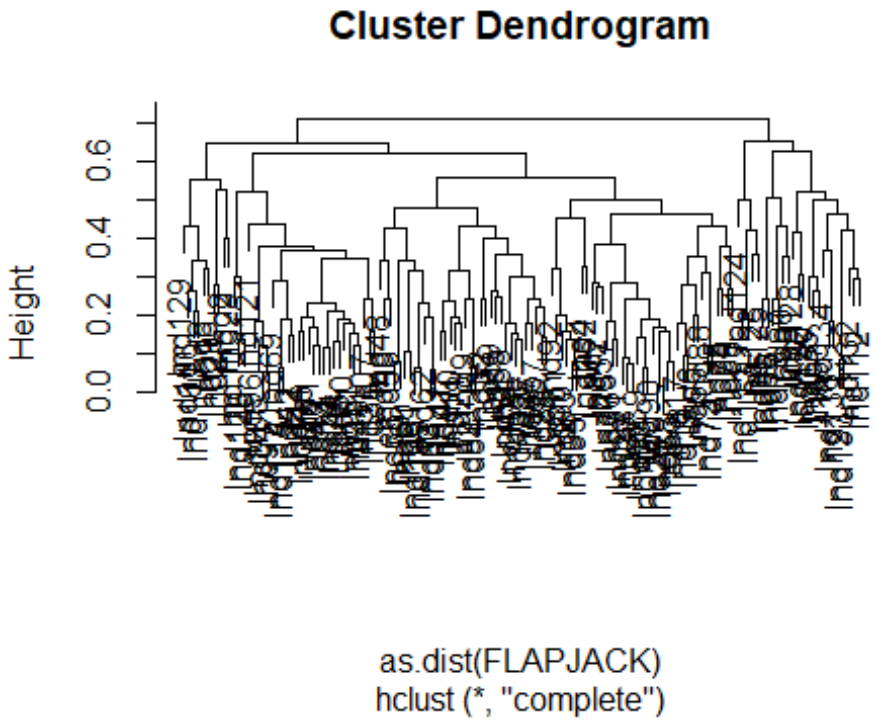


Imagen 38. Dendrograma creado en R a partir de los datos de Flapjack

```
PCA_TASSEL<-prcomp(TASSEL)
plot(PCA_TASSEL$x[,1],PCA_TASSEL$x[,2],xlab="PC1",ylab="PC2")
```

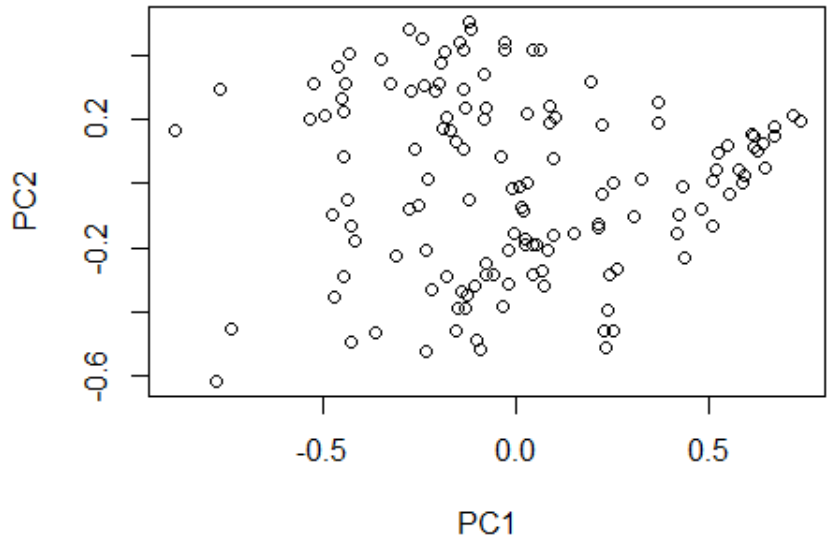


Imagen 39. PCA creado en R a partir de los datos de TASSEL

```
TASSEL_clust<-hclust(as.dist(TASSEL))
plot(TASSEL_clust)
```

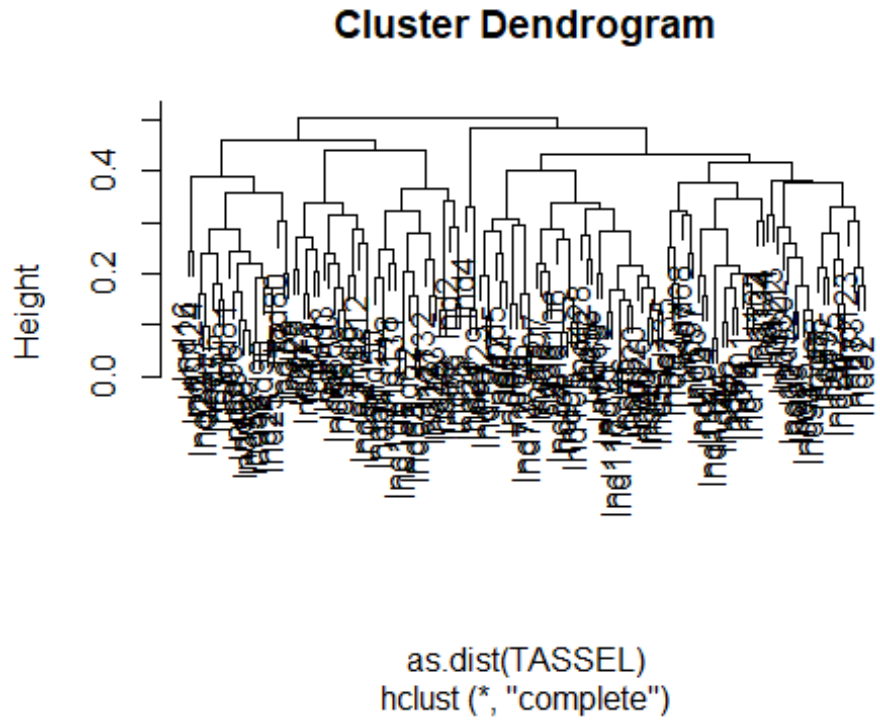


Imagen 40. Dendrograma creado en R a partir de los datos de TASSEL

```
PCA_GGT<-prcomp(GGT)
plot(PCA_GGT$x[,1],PCA_GGT$x[,2],xlab="PC1",ylab="PC2")
```

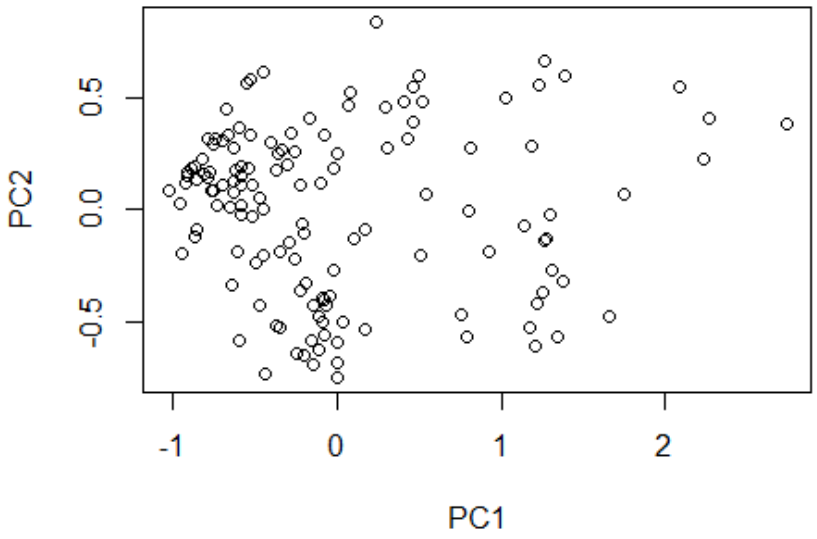


Imagen 41. PCA creado en R a partir de los datos de GGT

```
GGT_clust<-hclust(as.dist(GGT))
plot(GGT_clust)
```

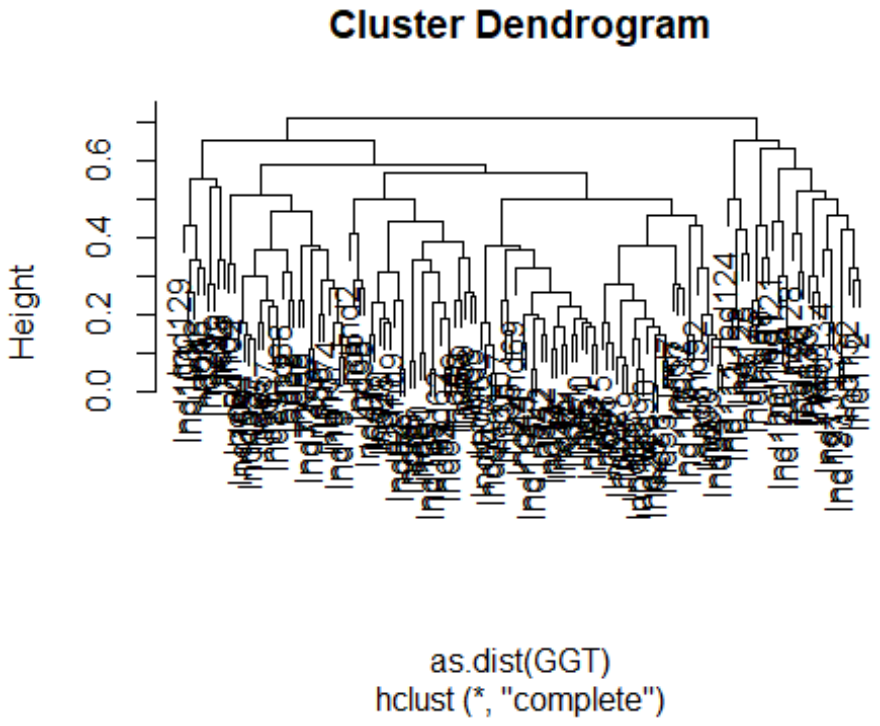


Imagen 42. Dendrograma creado en R a partir de los datos de GGT

Para la comparación de los dendrogramas se ha usado la libreria dendextend():

```
library(dendextend)
TASSEL_cluster<-TASSEL %>% dist %>% hclust %>% as.dendrogram
FLAPJACK_cluster<-FLAPJACK %>% dist %>% hclust %>% as.dendrogram
GGT_cluster<-GGT %>% dist %>% hclust %>% as.dendrogram
R_cluster<-R %>% dist %>% hclust %>% as.dendrogram


R_GGT<-tanglegram(R_cluster, GGT_cluster,
common_subtrees_color_branches=TRUE, main_left = "R", main_right =
"GGT")
```

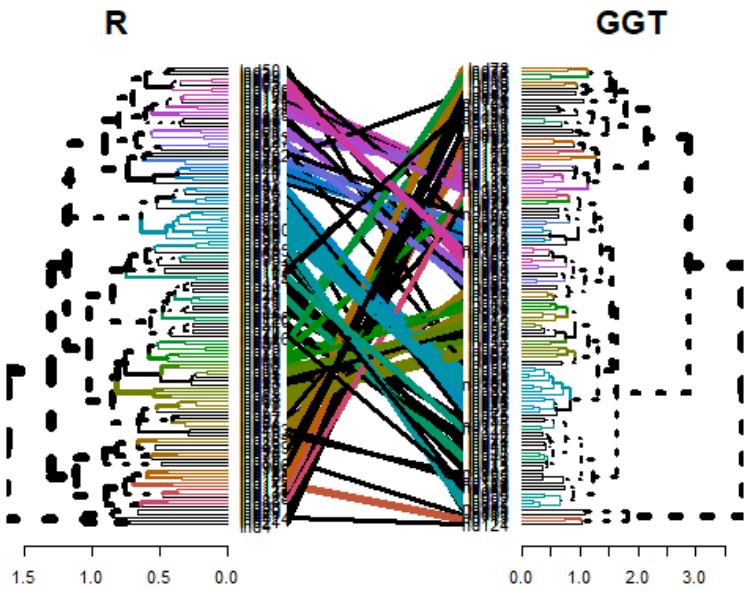


Imagen 43. Comparación de dendrogramas de R y GGT mediante la libreria dendextend()

```
R_FLAPJACK<-tanglegram(R_cluster, FLAPJACK_cluster,
common_subtrees_color_branches=TRUE, main_left = "R", main_right =
"FLAPJACK")
```

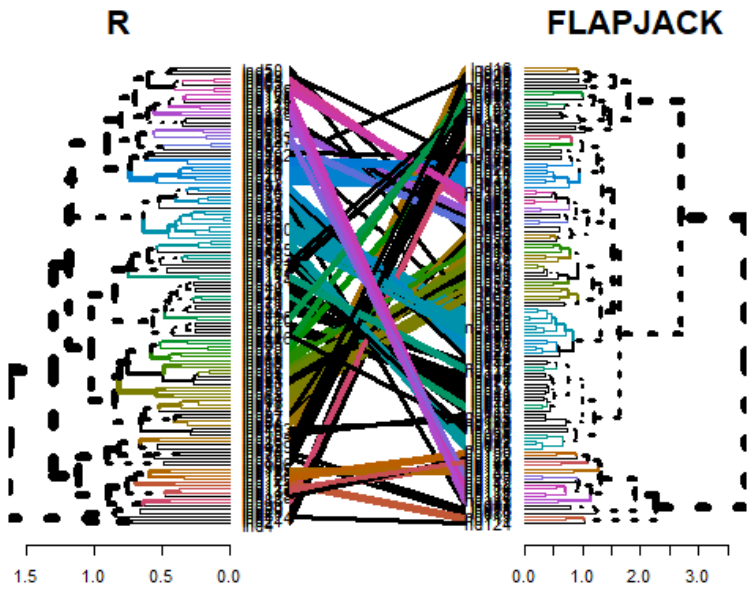


Imagen 44. Comparación de dendrogramas de R y Flapjack mediante la libreria dendextend()

```
R_TASSEL<-tanglegram(R_cluster, TASSEL_cluster,
common_subtrees_color_branches=TRUE, main_left = "R", main_right =
"TASSEL")
```

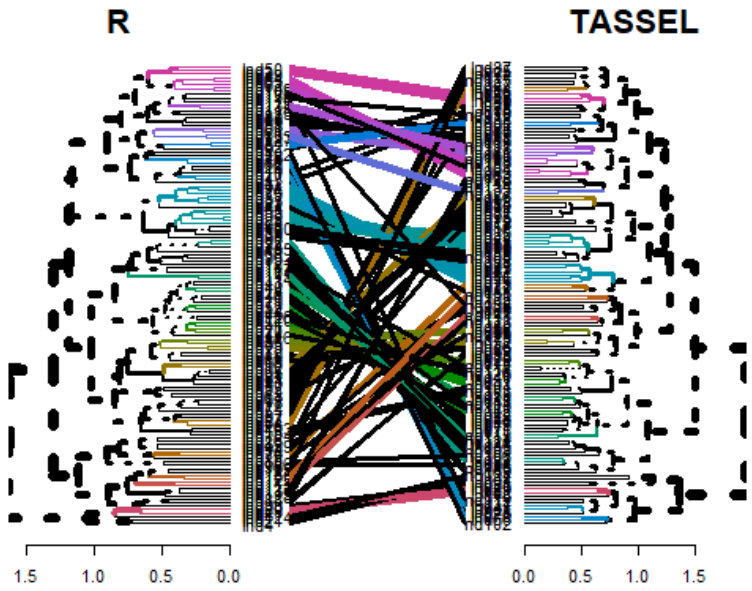


Imagen 45. Comparación de dendrogramas de R y TASSEL mediante la libreria dendextend()

```
GGT_FLAPJACK<-tanglegram(GGT_cluster, FLAPJACK_cluster,
common_subtrees_color_branches=TRUE, main_left = "GGT", main_right =
"FLAPJACK")
```

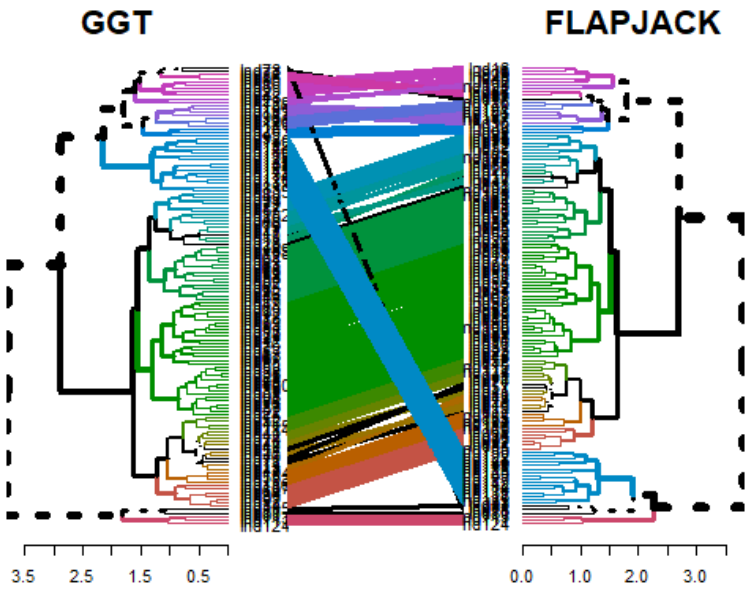


Imagen 46. Comparación de dendrogramas de GGT y Flapjack mediante la libreria dendextend()

```
GGT_TASSEL<-tanglegram(GGT_cluster, TASSEL_cluster,
common_subtrees_color_branches=TRUE, main_left = "GGT", main_right =
"TASSEL")
```

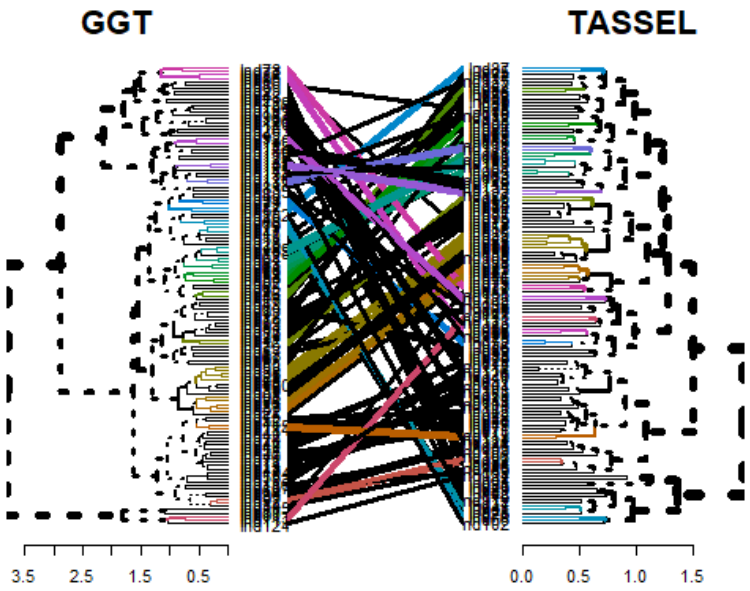


Imagen 47. Comparación de dendrogramas de GGT y TASSEL mediante la libreria dendextend()

```
TASSEL_FLAPJACK<-tanglegram(TASSEL_cluster, FLAPJACK_cluster,
common_subtrees_color_branches=TRUE, main_left = "TASSEL", main_right
= "FLAPJACK")
```

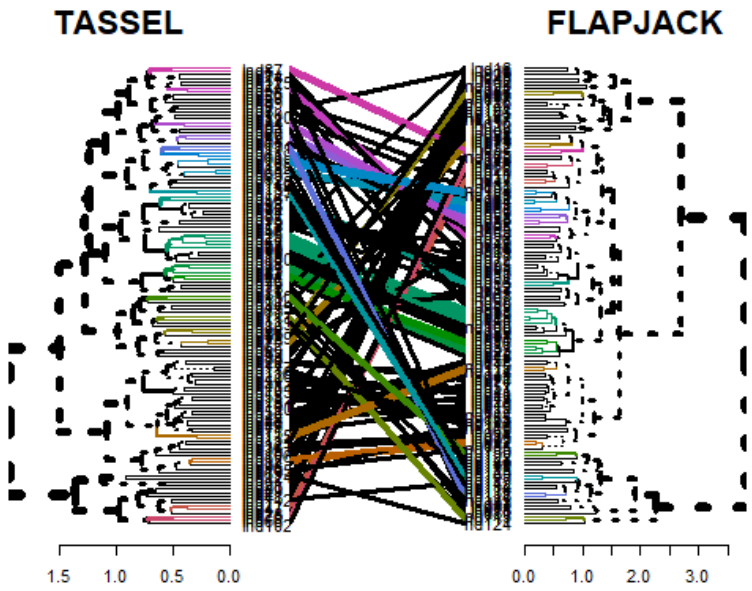


Imagen 48. Comparación de dendrogramas de TASSEL y Flapjack mediante la libreria dendextend()

A continuación, se aplica la función all.equal() que nos proporciona una comparación global de los dendogramas, dos a dos.

```
all.equal(dendlist(GGT_cluster,FLAPJACK_cluster,R_cluster,TASSEL_clust
er))

##
1==2
## "Difference in branch heights -  Mean relative difference:
0.01123828"
##
1==3
##  "Difference in branch heights -  Mean relative difference:
0.3641677"
##
1==4
##  "Difference in branch heights -  Mean relative difference:
0.2283405"
##
2==3
##  "Difference in branch heights -  Mean relative difference:
0.3610871"
##
2==4
##  "Difference in branch heights -  Mean relative difference:
0.2266709"
##
3==4
##  "Difference in branch heights -  Mean relative difference:
0.2236139"
```

ii. Construir mapas de ligamiento

MapMaker

El software utiliza un algoritmo que permite análisis simultáneos de cualquier número de loci o SNPs. También incluye un lenguaje de comandos interactivo que facilita la exploración de datos de ligamiento.

Se ha realizado el análisis de 25 individuos pertenecientes a la segunda generación de cruzamiento, F2, con un total de 40 SNPs. En la imagen 49 se muestra el código para introducir el archivo .RAW al programa mediante "*prepare data*" y el nombre del archivo. Este comando permite al programa a preparar los datos para el análisis. Mediante "sequence all" el programa analiza todos los marcadores.

El comando "group 3 30" crea los grupos de ligamiento basados en el criterio de tener un LOD mayor a 3 y una distancia menor a 30 cM (estos valores pueden ser modificados). El LOD hace referencia al logaritmo de las probabilidades de que dos marcadores se encuentren ligados y por lo tanto se heredan unidos con más frecuencia.

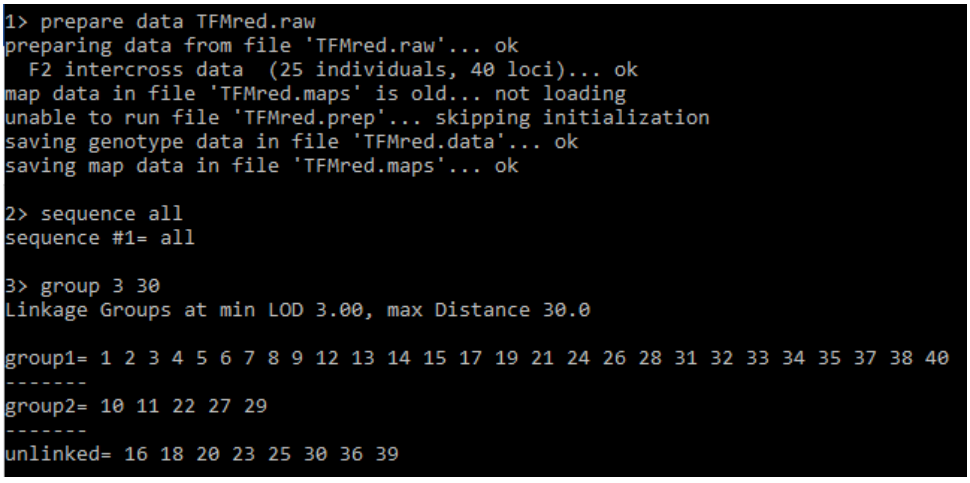


Imagen 49. Código para introducir el archivo .RAW en el software MapMaker y grupos de ligamiento obtenidos

Una vez obtenido los grupos de ligamiento, se pueden obtener las distancias entre ellos mediante el comando "*map*" (imagen 50).

```
4> map
================================================================================
Map:
  Markers          Distance
    1  Assay1       51.6 cM
    2  Assay2       32.1 cM
    3  Assay3       17.5 cM
    4  Assay4       26.5 cM
    5  Assay5       75.0 cM
    6  Assay6       23.3 cM
    7  Assay7       36.1 cM
    8  Assay8      101.9 cM
    9  Assay9      100.3 cM
   10  Assay10      31.4 cM
   11  Assay11     345.4 cM
   12  Assay12      51.5 cM
   13  Assay13      57.7 cM
   14  Assay14      64.0 cM
   15  Assay15      56.7 cM
   16  Assay16      31.4 cM
   17  Assay17     345.4 cM
   18  Assay18     345.4 cM
   19  Assay19      67.9 cM
   20  Assay20     345.4 cM
   21  Assay21     126.3 cM
   22  Assay22     117.6 cM
   23  Assay23      63.6 cM
   24  Assay24     345.4 cM
   25  Assay25     126.3 cM
   26  Assay26     345.4 cM
   27  Assay27     345.4 cM
   28  Assay28      44.7 cM
   29  Assay29     345.4 cM
   30  Assay30      73.4 cM
   31  Assay31      11.2 cM
   32  Assay32       8.7 cM
   33  Assay33       4.2 cM
   34  Assay34      41.0 cM
   35  Assay35     133.4 cM
   36  Assay36     140.0 cM
   37  Assay37      16.4 cM
   38  Assay38      66.6 cM
   39  Assay39      45.4 cM
   40  Assay40    ----------
                  4606  cM   40 markers   log-likelihood= -444.62
================================================================================
```

Imagen 50. Obtención de las distancias entre los marcadores mediante el comando *map*

Estos resultados se pueden representar gráficamente mediante el software MapChart. En el anexo se muestra cómo introducir los datos obtenidos en MapMaker, en la pestaña Data (imagen 4 anexo). En la pestaña Chart obtenemos la representación del mapa, mostrado a continuación (imagen 51).

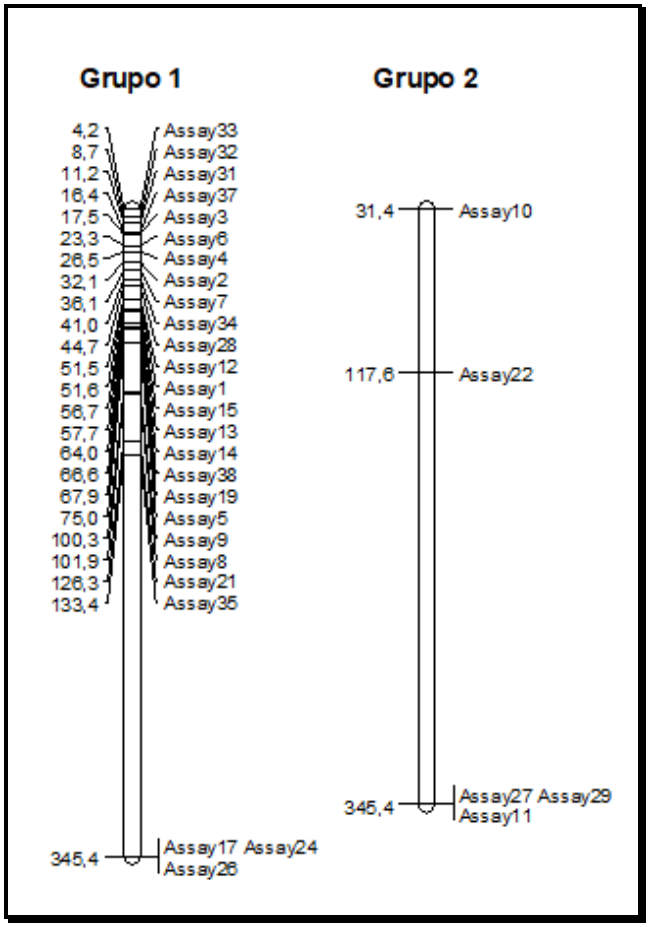


Imagen 51. Representación del mapa de ligamiento mediante el software MapChart a partir de los datos obtenidos en MapMaker

MapDisto

El análisis realizado en MapMaker se repite ahora con el software MapDisto. De esta forma se puede comparar los resultados obtenidos, para valorar los pros y contras de los dos softwares.

Una vez insertados los datos, mediante la opción "*Find linkage groups*" se especifica el valor mínimo de LOD y el valor máximo de frecuencia de recombinación, en este caso, 3 y 0,30 respectivamente. Se observa en la imagen 52 los grupos de ligamiento creados. En este caso, el valor de distancia en centimorgans lo especifica el software en función de los otros dos valores, siendo 34,7.

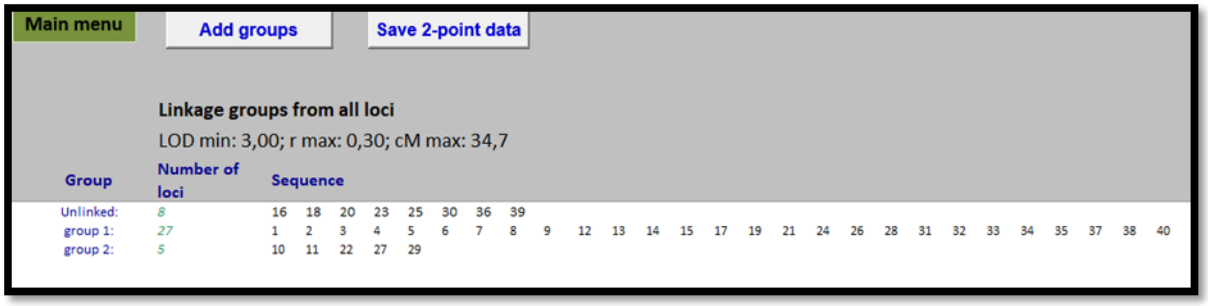


Imagen 52. Grupos de ligamiento obtenidos mediante el software MapDisto

La representación se realiza en el mismo software, mediante “Draw a linkage group” si se quiere especificar un grupo de ligamiento, o “Draw all linkage groups”, obteniendo tantos mapas como grupos han sido creados. En la imagen 53 se puede observar el *output* obtenido (En el anexo se adjunta la imagen 5, la cual contiene la imagen completa)

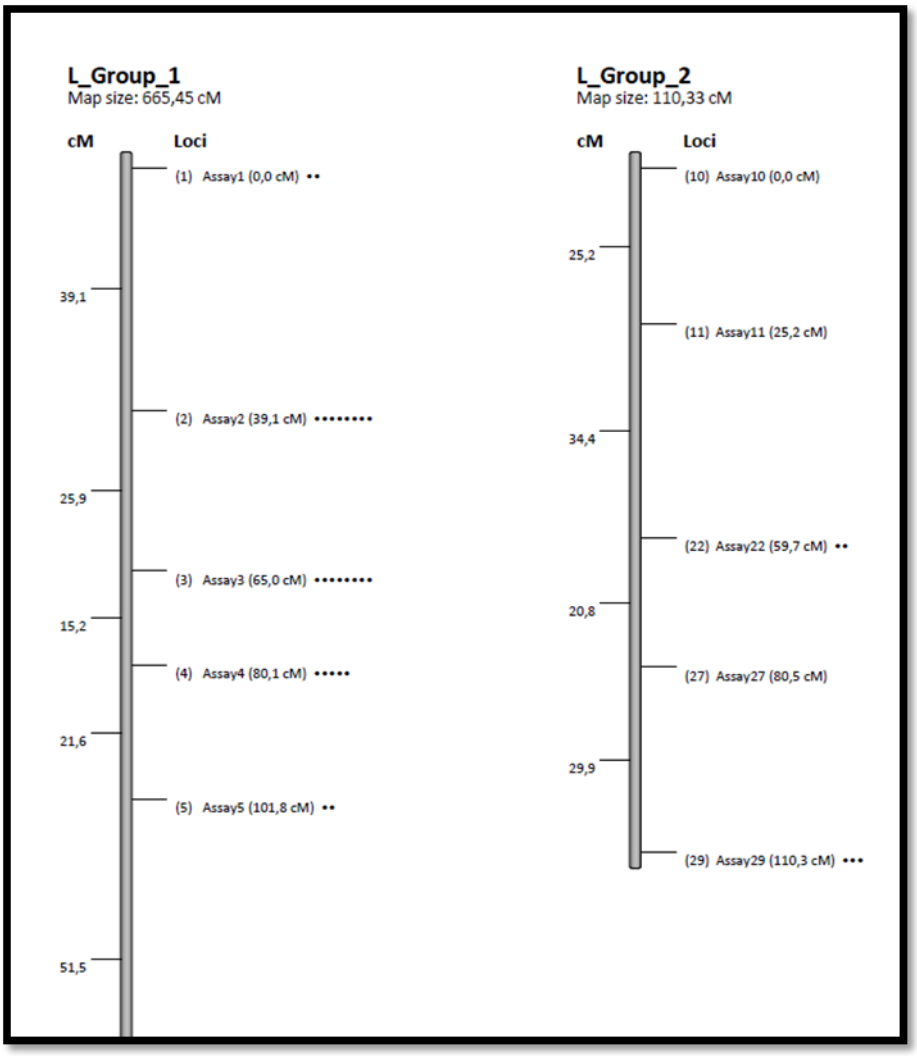


Imagen 53. Representación del mapa de ligamiento obtenido mediante el software MapDisto

iii. Visualizar gráficamente genotipos y cromosomas.

PhenoGram

El ejemplo utilizado en la representación gráfica de cromosomas y genotipos mediante PhenoGram son los 12 cromosomas de pimiento con los respectivos assays. Se puede observar un cromosoma extra, Chr00, el cual representa marcadores de los cuales se desconoce su actual posición, siendo esta ficticia. El resultado de la representación de los marcadores en los cromosomas del 0 al 6 se muestran en la imagen 54, y del cromosoma 7 al 12 en la imagen 55. En el anexo se adjunta un fragmento de la imagen del nombre del marcador que corresponde a cada color (imagen 6 anexo). El modo de representar el color de los marcadores se puede modificar, en función del interés del estudio a analizar.

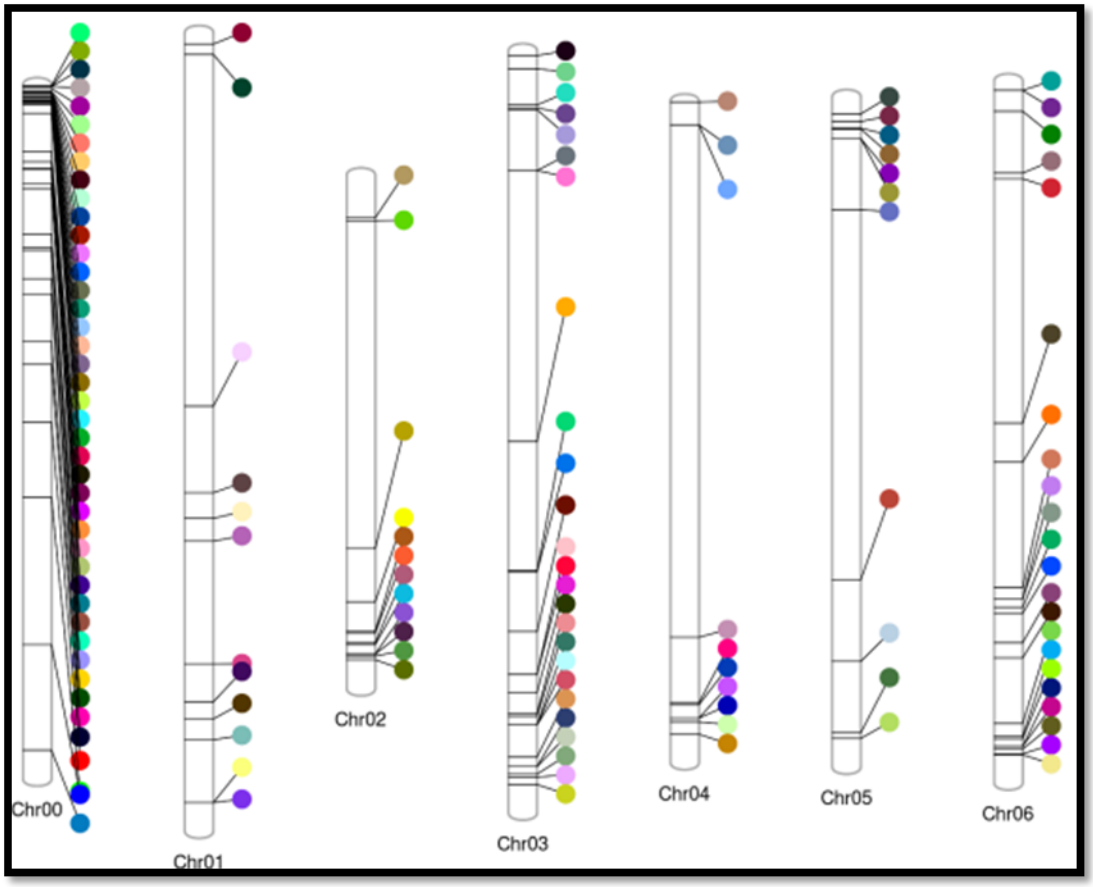


Imagen 54. Representación de la localización de los marcadores ubicados entre los cromosomas 0 y 6

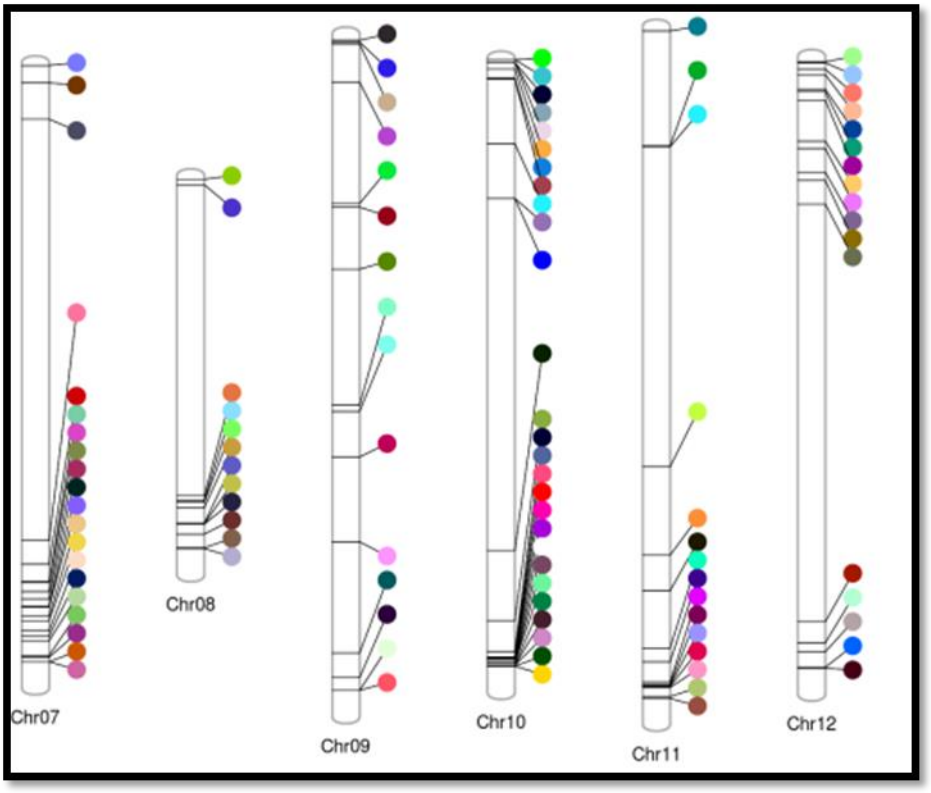


Imagen 55. Representación de la localización de los marcadores ubicados entre los cromosomas 7 y 12

3.3 Marker-Assisted Backcrossing (MABC)

La función desarrollada en R nos devuelve la matriz de distancia entre los individuos que presentan los alelos heterocigotos en el marcador de interés, y su distancia con el individuo recurrente. Los números elevados presentan mayor distancia y los números bajos, menor distancia con el individuo recurrente.

```
A<-MABC(Datos,5);A

##               1     Ind1     Ind2    Ind17
## Ind1  1.048809
## Ind2  1.048809 0.000000
## Ind17 2.345208 2.097618 2.097618
## Ind18 3.146427 2.966479 2.966479 3.633180
```

Para poder entender mejor la función desarrollada, se representan los diferentes pasos por los que va siendo procesado el archivo inicial para ver su evolución. En la imagen 56 se muestra el archivo de ejemplo para el análisis, donde el marcador de interés es el número 5:

X	Assay1	Assay2	Assay3	Assay4	Assay5	Assay6	Assay7	Assay8	Assay9	Assay10
Recurrente	A	C	T	A	G	C	A	T	G	T
Donante	C	T	T	C	A	T	A	C	C	T
Ind1	A	C	T	A	G/A	C	A	T	G	T
Ind2	A	C	T	A	G/A	C	A	T	G	T
Ind3	A	C	T	A	G/A	T	A	T/C	G	T
Ind4	A	C	G	C	G/A	C	A	T/C	G/C	T
Ind5	C	C	T	A	G	C	T	C	G	A
Ind6	C	C	T	A	G	C	A	C	C	A
Ind7	A	C	T	A	A	C	A	C	C	T
Ind8	A	C	T	A	A	T	A	T	G/C	T/A
Ind9	A	T	T	C	A	T	T	T	C	T
Ind10	C	T	G	C	G	C	T	T	G	T/A
Ind11	A	T	T	C	G/A	T	T	T/C	G/C	T
Ind12	A	T	T	C	G	C	T	T/C	G	T
Ind13	A	T	T	C	A	C	A	C	G	T
Ind14	C	C	G	C	G	C	A	C	G	T
Ind15	C	T/C	G	C	G	C	T	T	G/C	A
Ind16	C	T/C	G	A	G	C	T	T	C	A
Ind17	A	C	T	A	G/A	C	A	T	C	T
Ind18	A	C	T	A	G/A	C	T	C	G	T
Ind19	C	C	T/G	A	A	T	A/T	C	G	T
Ind20	C	C	G	A	A	C	A	C	G	T/A

Imagen 56. Archivo de ejemplo para el transcurso del análisis del MABC

El primer paso es la codificación de los datos. Ésta está condicionada a los alelos del individuo denominado como recurrente: si los individuos de la F1 presentan el mismo alelo que el recurrente, tendrán una B, y si presentan otro alelo tendrán una A o una H, en función de ser homocigotos o heterocigotos, respectivamente (imagen 57).

	Assay1	Assay2	Assay3	Assay4	Assay5	Assay6	Assay7	Assay8	Assay9	Assay10
Recurrente	B	B	B	B	B	B	B	B	B	B
Ind1	B	B	B	B	H	B	B	B	B	B
Ind2	B	B	B	B	H	B	B	B	B	B
Ind3	B	B	B	B	H	A	B	H	B	B
Ind4	B	B	A	A	H	B	B	H	H	B
Ind5	A	B	B	B	B	B	A	A	B	A
Ind6	A	B	B	B	B	B	B	A	A	A
Ind7	B	B	B	B	A	B	B	A	A	B
Ind8	B	B	B	B	A	A	B	B	H	H
Ind9	B	A	B	A	A	A	A	B	A	B
Ind10	A	A	A	A	B	B	A	B	B	H
Ind11	B	A	B	A	H	A	A	H	H	B
Ind12	B	A	B	A	B	B	A	H	B	B
Ind13	B	A	B	A	A	B	B	A	B	B
Ind14	A	B	A	A	B	B	B	A	B	B
Ind15	A	H	A	A	B	B	A	B	H	A
Ind16	A	H	A	B	B	B	A	B	A	A
Ind17	B	B	B	B	H	B	B	B	A	B
Ind18	B	B	B	B	H	B	A	A	B	B
Ind19	A	B	H	B	A	A	H	A	B	B
Ind20	A	B	A	B	A	B	B	A	B	H

Imagen 57. Codificación de los datos para el MABC

Una vez codificados los datos, se seleccionan los individuos que contengan una H en el marcador de interés, ya que presentarán un alelo del individuo recurrente y el otro del donante. Además, los *assays* flanqueantes, en este caso *Assay4* y *Assay6*, deben tener alguna de las siguientes combinaciones: B y B; B y H; H y B, debido a la existencia de una mayor probabilidad de producirse recombinación junto con los marcadores flanqueantes por proximidad, teniendo más similitudes con el individuo recurrente.

	Assay1	Assay2	Assay3	Assay4	Assay5	Assay6	Assay7	Assay8	Assay9	Assay10
Ind1	B	B	B	B	H	B	B	B	B	B
Ind2	B	B	B	B	H	B	B	B	B	B
Ind17	B	B	B	B	H	B	B	B	A	B
Ind18	B	B	B	B	H	B	A	A	B	B

Imagen 58. Individuos que cumplen con las condiciones especificades en el MABC

Son 4 los individuos que cumplen las condiciones anteriores (imagen 58). Por lo tanto, se realizará la matriz de distancia de estos, y se verá cuáles de ellos presentan una mayor similitud con el recurrente.

	1	Ind1	Ind2	Ind17
Ind1	1.048809			
Ind2	1.048809	0.000000		
Ind17	2.345208	2.097618	2.097618	
Ind18	3.146427	2.966479	2.966479	3.633180

Imagen 59. Matriz de distancia de los individuos seleccionados por MABC

La matriz de distancia (imagen 59) muestra que los individuos 1 y 2 son los que presentan menos distancia con el recurrente, por lo tanto, son los individuos que presentan un mayor interés de los 20 analizados. Los individuos 17 y 18 presentan todas las condiciones requeridas en el MABC, pero presentan diferencias con el recurrente en algún alelo.

4. Discusión

4.1 Estudiar los parámetros básicos a partir de datos genotípicos

El proceso de codificación nos permite convertir los datos nucleotídicos en numéricos para su posterior análisis. En este trabajo se ha usado la codificación: **2 o A,** para el alelo dominante, **0 o B,** para el alelo minoritario, **1 o H,** para los heterocigotos, y **3 o nd,** para los datos faltantes.

Las conversiones realizadas no siguen siempre un mismo patrón, ya que los softwares no tratan igual los datos según la codificación empleada. Como hemos visto anteriormente, el software GGT asigna las letras A, B, C o D según la proporción de alelos en el archivo (pudiendo ser homocigoto mayoritario, homocigoto minoritario, heterocigoto o datos faltantes), siendo A una mayor proporción, y D la menor proporción. Esto implica que, si un archivo tiene muchos datos faltantes, estos serán asignados automáticamente al valor A.
Otro ejemplo sería el caso de la función de R data.matrix(). Esta función codifica los valores de cada marcador en 1, 2, 3 y 4, donde los números 1 y 3 corresponden a los alelos homocigotos, 2 corresponde a los alelos heterocigotos, y 4 corresponde a los datos faltantes o missing data.
Otro punto importante para tener en cuenta es la codificación empleada sobre los datos faltantes o *missing data*, ya que estos pueden alterar de forma drástica los análisis de datos realizados.

Así pues, se deberían analizar detenidamente qué cambios implican estos modelos de codificación sobre nuestros datos, para poder tener la certeza de no estar cometiendo errores a gran escala. Este razonamiento nos abre puertas para averiguar qué tipo de codificación es la más acertada para este tipo de datos, analizando cuál de ellos se adapta mejor a nuestra necesidad.

4.2 Analizar los datos genéticos mediante diversos softwares y comparación de los resultados

i. Analizar distancias genéticas.

Para realizar el análisis de distancias genéticas, con los datos codificados en R, se han obtenido las matrices de distancia o similitud, a partir de los diferentes softwares R, Flapjack, GGT y TASSEL.
El análisis de las matrices muestra similitudes entre los softwares Flapjack y GGT, en cambio, las matrices de R y TASSEL presentan diferencias. Estas similitudes y diferencias se deben a los métodos usados para calcular las matrices en cada software. Las propiedades genéticas y matemáticas en medidas de distancias y similitudes tienen una importancia crucial para escoger el coeficiente de distancia genética para analizar datos de marcadores moleculares (Reif, 2005).

La comparativa realizada mediante el paquete dendextend() nos permite ver las diferencias y similitudes entre los dendrogramas realizados en R a partir de las matrices de distancia o similidud de los diferentes softwares. Como era de esperar debido a la similitud entre matrices, los dendrogramas de GGT y Flapjack (imagen 46) presentan muchos colores de similitud en su comparación. Se confirma la evidencia mediante la función all.equal(), siendo el valor 0.011 (inferior al nivel de significancia 0.05), en cambio, las demás combinaciones presentan valores superiores a 0.22 (superior al valor de significancia 0.05). Esta es una herramienta potente para el futuro en el que también podremos comparar dendrogramas realizados por un mismo software, aunque mediante diferentes métodos de análisis, perfeccionando así las técnicas de análisis.

El análisis de PCA de los softwares Flapjack y TASSEL no permite una visualización fácil de los resultados. R es una alternativa muy útil para este análisis, un ejemplo de análisis de PCA es el paquete FactoMineR, ya que se puede obtener las etiquetas de los individuos en la imagen, se pueden clasificar los grupos por colores, exportar una lista con los nombres y en qué grupo se encuentran, etc. Todos estos datos nos proporcionan información para poder realizar comparaciones con otros PCAs.

Una matriz de decisión es una herramienta que permite tomar buenas decisiones cuando se trata con factores difíciles de comparar, ayudando a eliminar la subjetividad con el fin de llegar a una conclusión sólida. Es muy importante tener en cuenta que falta la opinión del *breeder*, aunque no se puede aportar en este trabajo. A continuación, se presenta una matriz de decisión (tabla 2) para valorar el mejor software.

Tabla 2. Matriz de decisión para valorar el mejor software

SOFTWARE	Adaptación input	Cargar los datos	Análisis de datos	Exportación de resultados	Flexibilidad del análisis	PUNTUACIÓN
R	10	10	10	10	10	50
Flapjack	10	10	10	8	1	39
GGT	7	10	10	8	5	40
TASSEL	7	3	10	8	1	29

En la matriz de decisión se valoran 5 factores:

- Facilidad de adaptación al input: en R y Flapjack prácticamente se requiere el mismo formato que el obtenido en el laboratorio. GGT y TASSEL requieren más elaboración.

- Cómo incorporar los datos al software: R, Flapjack y GGT presentan formatos muy cómodos para cargar archivos. En cambio, TASSEL requiere crear un *Genotype file* para posteriormente cargar los datos.

- Análisis de datos: en este caso, todos los programas presentan una interfaz de usuario intuitiva que permite realizar los análisis sin ningún tipo de dificultad.

- Exportación de resultados: Flapjack, GGT y TASSEL presentan el mismo formato para exportar los datos, obteniendo las matrices en formato txt y los gráficos en formato JPG o PDF. En cambio, R, además de poder exportar en formato txt los datos o JPG o PDF los gráficos, permite obtener un output mucho más elegante gracias a R Markdown, en el que podemos obtener un *report* con todos los datos de interés y los gráficos para ser presentado.

- Flexibilidad de análisis: no cabe duda de que el mejor software en este aspecto es R. Se puede calcular la matriz de distancia o de similitud a partir de muchos índices mediante la función dist() del paquete proxy, por ejemplo. Hay diversas funciones que permiten calcular dendrogramas mediante diferentes métodos, y también PCAs, además de poder modificar el gráfico en tamaño, color, título, añadir *bootstraps*… GGT tiene un 5 en este apartado ya que permite seleccionar el coeficiente de similaridad entre *Allele sharing*/SM, Jaccard y Euclidean, proporcionando más flexibilidad, y seleccionar el método para calcular el dendrograma, pudiendo ser Neighbor Joining (NJ) o UPGMA.

El resultado de la matriz de decisión permite llegar a la conclusión que R es el software seleccionado más completo para realizar futuros análisis.

Es muy importante tener en cuenta que escoger el método de análisis más adecuado para el análisis de los datos no depende sólo del análisis bioinformático. La experiencia y opinión del mejorador o *breeder* aporta los conocimientos del día a día para poder valorar qué resultados se ajustan mejor a la realidad. Las conclusiones de este objetivo se tendrán que validar en un futuro juntamente con el *breeder*, para llegar a un consenso del análisis adecuado.

ii. Construir mapas.

La construcción de mapas de ligamiento se realiza por medio del análisis de la frecuencia de recombinación de marcadores moleculares entre una población segregante. Los mapas de ligamiento son un paso previo para realizar estudios genéticos cuantitativos, como lo es el mapeo de QTLs.

Se han creado grupos de ligamiento en MapMaker y MapDisto. Ambos han clasificado de la misma forma los grupos, teniendo el primero 27 *assays*, el segundo 5, y el resto de *assays* no han sido clasificados, como se observa en las imágenes 49 y 52 en el apartado de resultados. En cuanto a la representación gráfica de los grupos de ligamiento, MapMaker requiere el software MapChart, en cambio, MapDisto tiene su propia representación.
Otro factor a tener en cuenta es la facilidad de trabajo en cada software. MapMaker requiere un archivo .RAW, con un formato específico de los datos, en cambio, MapDisto presenta más facilidades ya que es copiado un archivo Excel, igual que el obtenido en el laboratorio. Una vez cargados los datos, en MapMaker se trabaja a partir del símbolo del sistema o *command prompt*, y MapDisto contiene una interfaz de usuario intuitiva.

El algoritmo usado para calcular las distancias físicas en MapMaker permite realizar simultáneamente *multipoint analysis* con cualquier número de marcadores. En MapDisto utiliza dos tipos de algoritmos, el primero está basado en el análisis *two-point*, y el segundo en el *multipoint estimates*.

Encontrar el orden correcto de los marcadores que forman los grupos de ligamiento es el paso más difícil en la construcción del mapa. Como se observa en los resultados, el orden del mapa de ligamiento es diferente en los dos softwares, debido a la posibilidad de modificar el orden de estos mediante algoritmos. Hubiera sido muy interesante indagar en este ámbito, pero no se ha podido dedicar todo el tiempo establecido ya que el objetivo de análisis de distancias genéticas ha requerido más tiempo del planificado en el inicio del trabajo. Se seguirá trabajando en estos análisis para poder aplicar los conocimientos en el futuro en el laboratorio, pues tiene mucho interés para la detección de la posición cromosómica de los marcadores.

iii. Visualizar gráficamente genotipos y cromosomas.

Esta herramienta es muy útil para la abundante información que se recopila de los marcadores moleculares. Nos permite la visualización de datos y facilita la difusión de estos, gracias a la flexibilidad de modificación que presenta. Un ejemplo de modificación seria añadir colores al análisis, clasificando los marcadores monomórficos (que no presentan polimorfismo) de un color, y los marcadores polimórficos en otro color, permitiendo una rápida decisión sobre el uso de estos en el día a día del laboratorio. Todas estas modificaciones que se pueden aplicar al software permiten agilizar las decisiones del uso de los marcadores moleculares.

4.3 Marker-Assisted Backcrossing (MABC)

La función creada en R para el análisis del MABC es una herramienta muy útil para la obtención del resultado de una forma rápida. Partiendo de un archivo Excel complejo, se obtiene una matriz de distancia de los individuos que han cumplido todas las condiciones especificadas, seleccionando los individuos que se parecen más al individuo recurrente, aunque con el marcador de interés incorporado a partir del individuo donante.

5. Conclusiones

- La función creada en R para la codificación de datos nos permite garantizar la confidencialidad de estos y, además, el preprocesado para su introducción en los diferentes softwares que requieren de formatos específicos.

- La función creada en R para analizar los parámetros básicos permite obtener información estadística para conocer el tipo de datos con los que estamos trabajando.

- El análisis de distancias genéticas mediante los diferentes softwares ha permitido valorar que R presenta mucha versatilidad en los estudios, siendo elegido como el que mejor se adapta a nuestras necesidades.

- La realización de mapas de ligamiento no ha podido ser explorada como se había planeado en el inicio del trabajo. Queda mucho por investigar esta herramienta, pues tiene mucho interés para la detección de la posición cromosómica de los marcadores.

- PhenoGram es una herramienta versátil que permite la visualización gráfica de cromosomas y genotipos, facilitando la interpretación de estos y agilizando las decisiones del uso de los marcadores moleculares en el laboratorio.

- La función creada en R para el estudio de MABC permite seleccionar, mediante una matriz de distancia, de forma rápida los individuos de la progenie de mayor interés para el análisis.

6. Glosario

DNA	Acido Desoxirribonucleico
FRET	Fluorescence Resonance Energy Transfer
GGT	Graphical GenoTypes
MABC	Marker Assisted BackCrossing
PCA	Principal Component Analysis
PCR	Polymerase Chain Reaction
PIC	Polymorphic Information Content
QTL	Quantitative Trait Loci
SNP	Single Nucleotide Polymorphism
TASSEL	Trait Analysis by aSSociation, Evolution, and Linkage
UPGMA	Unweighted Pair Group Method with Arithmetic mean

7. Bibliografía

Botstein D, White RL, Skolnick M and Davis RW (1980). *Construction of a genetic linkage map in man using restriction fragment length polymorphisms.* Am J Hum Genet 32:314–331.

Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES. (2007) *TASSEL: Software for association mapping of complex traits in diverse samples.* Bioinformatics **23**:2633-2635.

Galili, T. (2015). *dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering*. Bioinformatics, 31(22), 3718–3720. http://doi.org/10.1093/bioinformatics/btv428

Hasan, M. M., Rafii, M. Y., Ismail, M. R., Mahmood, M., Rahim, H. A., Alam, M. A., Latif, M. A. (2015). *Marker-assisted backcrossing: a useful method for rice improvement*. Biotechnology, Biotechnological Equipment, 29(2), 237–254. http://doi.org/10.1080/13102818.2014.995920

Heffelfinger, Fragoso and Lorieux (2017) *Constructing linkage maps in the genomics era with MapDisto 2.0.* DOI: https://doi.org/10.1093/bioinformatics/btx177

Jingade, A. H., Vijayan, K., Somasundaram, P., Srivasababu, G. K., & Kamble, C. K. (2011*). A Review of the Implications of Heterozygosity and Inbreeding on Germplasm Biodiversity and Its Conservation in the Silkworm, Bombyx mori.* Journal of Insect Science, 11, 8. http://doi.org/10.1673/031.011.0108

Lander, E. S., P. Green, J. Abrahamson, A. Barlow, M. J. Daly, S. E. Lincoln, and L. Newburg. 1987. *MAPMAKER: an interactive computer package for constructing primary genetic linkage maps of experimental populations.* Genomics 1: 174–181.

Lorieux M, 2012. *MapDisto: fast and efficient computation of genetic linkage maps.* Molecular Breeding 30: 1231–1235.

Mammadov, J., et al. (2012) *SNP Markers and Their Impact on Plant Breeding*. International Journal of Plant Genomics. Article ID 728398.

Milne I, Shaw P, Stephen G, Bayer M, Cardle L, Thomas WTB, Flavell AJ and Marshall D. (2010). *Flapjack – graphical genotype visualization.* Bioinformatics 26(24), 3133-3134.

Mitton JB, Schuster WSF, Cothran EG, De Fries JC. (1993) *Correlation between the individual heterozygosity of parents and their offspring.* Heredity. 1993;71:59–63.

Ngangkham U, Samantaray S, Yadav MK, Kumar A, Chidambaranathan P, Katara JL (2018) *Effect of multiple allelic combinations of genes on regulating grain size in rice.* PLoS ONE 13(1): e0190684. https://doi.org/10.1371/journal.pone.0190684

Ralph van Berloo; *GGT 2.0: Versatile Software for Visualization and Analysis of Genetic Data*, Journal of Heredity, Volume 99, Issue 2, 1 April 2008, Pages 232–236.

Reif, J.C., Melchinger, A.E., and Frisch, M. (2005). *Genetical and mathematical properties of similarity and dissimilarity coefficients applied in plant breeding and seed bank management*. Crop science, vol. 45.

Tiwari, A. (2017) *Plant Breeding: A Prospect in Developing World. EC Microbiology*. EC Microbiology 8.5: 272-278.

Varshney, R.K. et al (2016) *Analytical and decision support tools for genomics-assisted breeding*. Trends in Plant Science. Volume 21, Pages 354-363.

Voorrips, R.E., 2002. MapChart: Software for the graphical presentation of linkage maps and QTLs. The Journal of Heredity 93 (1): 77-78.

Wolfe, D., Dudek, S., Ritchie, M. D., & Pendergrass, S. A. (2013). *Visualizing genomic information across chromosomes with PhenoGram.* BioData Mining, 6, 18. http://doi.org/10.1186/1756-0381-6-18

8. Anexos

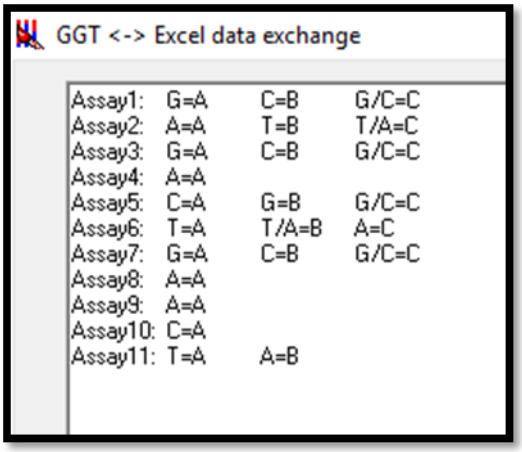


Imagen Anexo 1. Logaritmo de codificación usado en el software GGT.

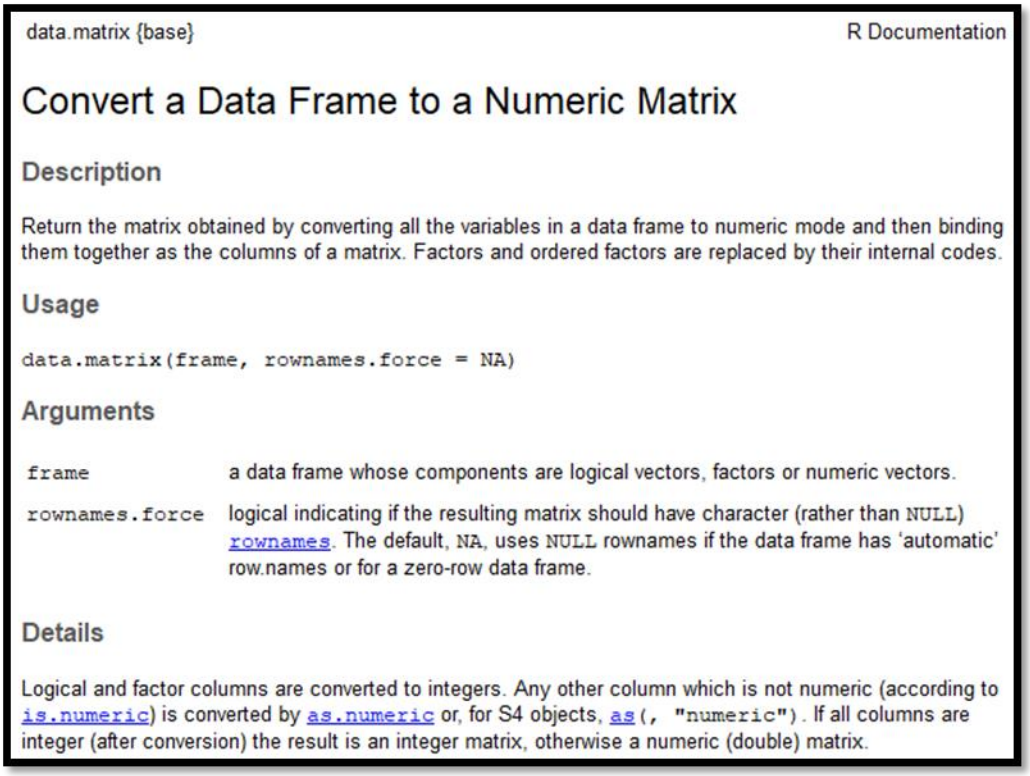


Imagen Anexo 2. Librería data.matrix() para realizar la codificación de datos genotípicos.

Nucleotide Codes (Derived from IUPAC)

Code	Meaning
A	A:A
C	C:C
G	G:G
T	T:T
R	A:G
Y	C:T
S	C:G
W	A:T
K	G:T
M	A:C
+	+:+ (insertion homozygous)
0	+:-
-	-:- (deletion homozygous)
N	Unknown

Imagen Anexo 3. Codificación interna del software TASSEL

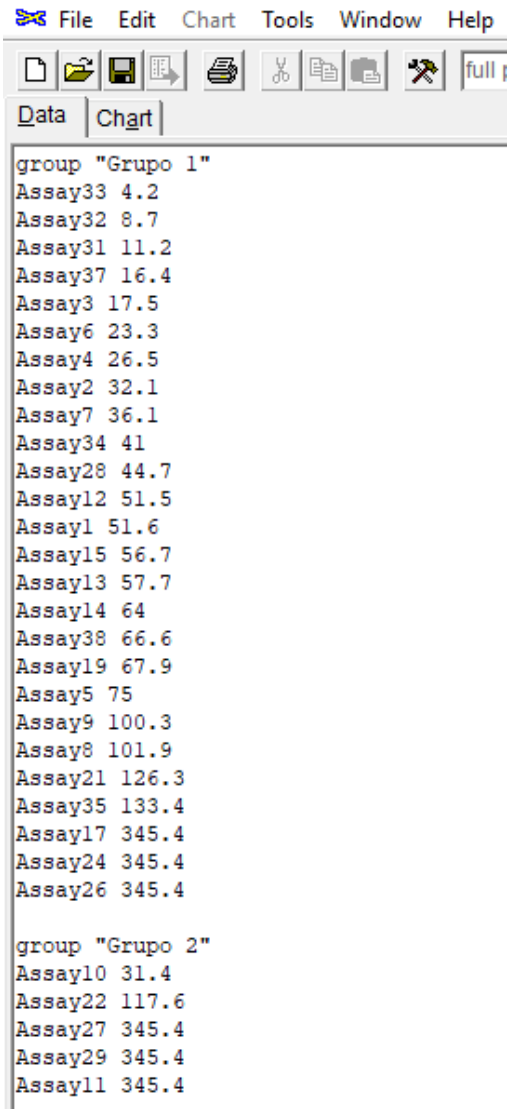
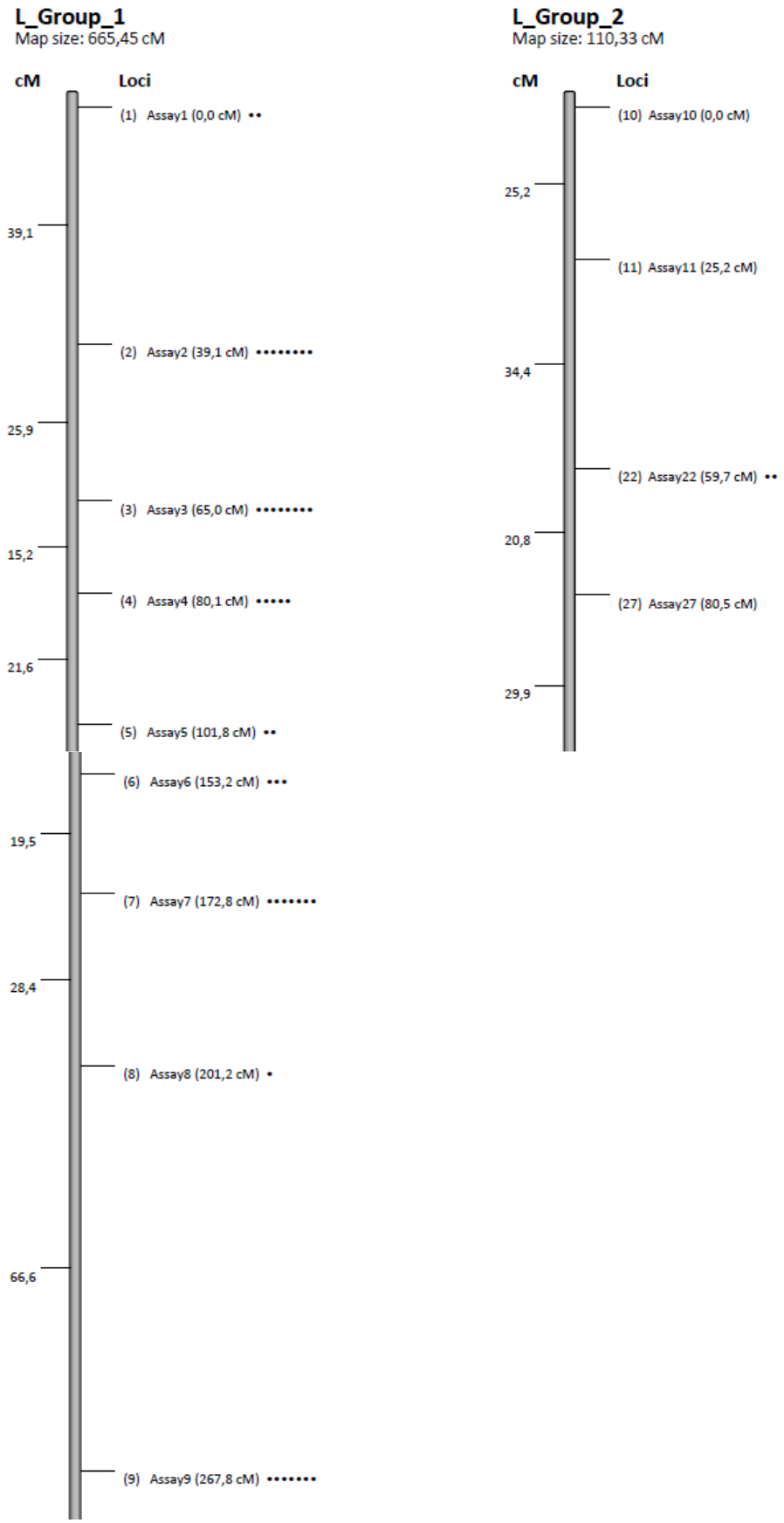
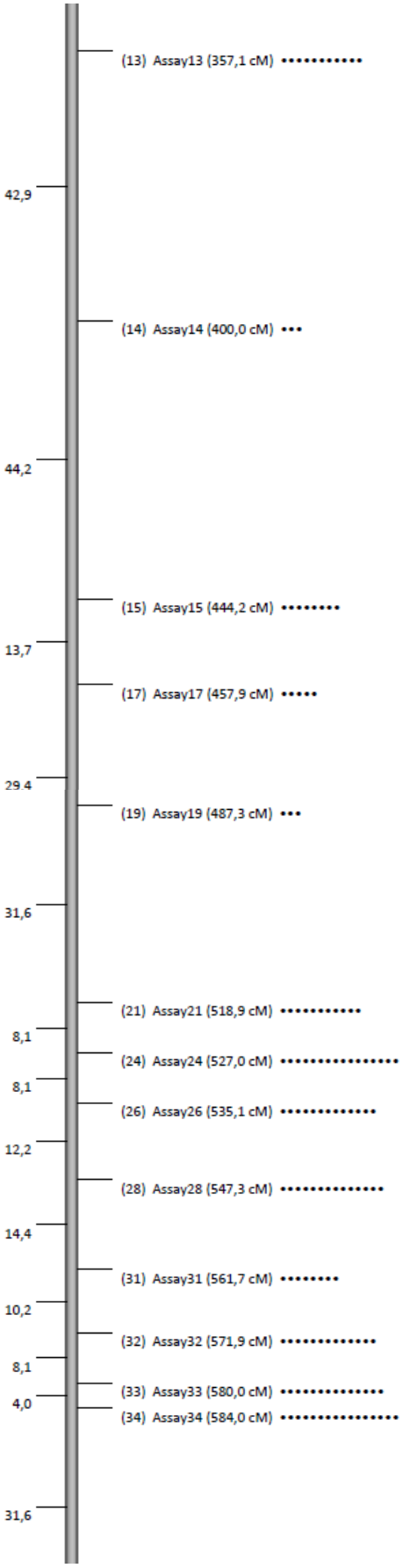


Imagen Anexo 4. . Input para crear el mapa de ligamiento mediante el software MapChart a partir de los datos obtenidos en MapMaker


L_Group_1
Map size: 665,45 cM
cM
Loci
(1) Assay1 (0,0 cM) ••
39,1
(2) Assay2 (39,1 cM) ••••••••
25,9
(3) Assay3 (65,0 cM) ••••••••
15,2
(4) Assay4 (80,1 cM) •••••
21,6
(5) Assay5 (101,8 cM) ••
(6) Assay6 (153,2 cM) •••
19,5
(7) Assay7 (172,8 cM) •••••••
28,4
(8) Assay8 (201,2 cM) •
66,6
(9) Assay9 (267,8 cM) •••••••
L_Group_2
Map size: 110,33 cM
cM
Loci
(10) Assay10 (0,0 cM)
25,2
(11) Assay11 (25,2 cM)
34,4
(22) Assay22 (59,7 cM) ••
20,8
(27) Assay27 (80,5 cM)
29,9


(13) Assay13 (357,1 cM) ••••••••••
42,9
(14) Assay14 (400,0 cM) •••
44,2
(15) Assay15 (444,2 cM) ••••••••
13,7
(17) Assay17 (457,9 cM) •••••
29.4
(19) Assay19 (487,3 cM) •••
31,6
(21) Assay21 (518,9 cM) ••••••••••
8,1
(24) Assay24 (527,0 cM) ••••••••••••••••
8,1
(26) Assay26 (535,1 cM) ••••••••••••
12,2
(28) Assay28 (547,3 cM) ••••••••••••••
14,4
(31) Assay31 (561,7 cM) ••••••••
10,2
(32) Assay32 (571,9 cM) ••••••••••••
8,1
(33) Assay33 (580,0 cM) ••••••••••••••
4,0
(34) Assay34 (584,0 cM) ••••••••••••••••
31,6

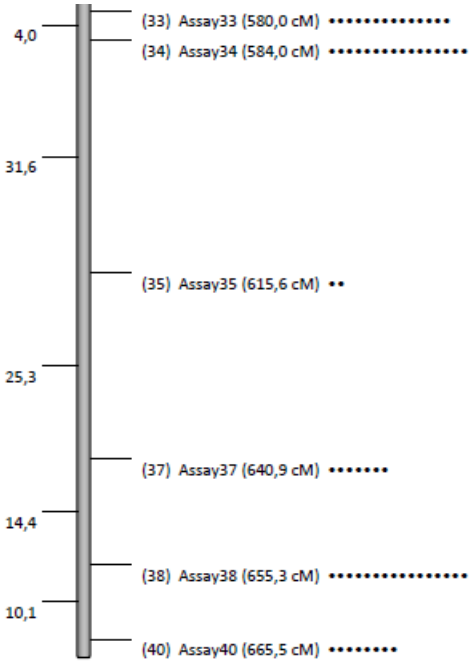


Imagen Anexo 5. Mapa de ligamiento completo realizado mediante el software MapDisto


SNP219
SNP227
SNP104
SNP209
SNP228
SNP217
SNP60
SNP187
SNP214
SNP222
SNP126
SNP233
SNP153
SNP216
SNP21
SNP225
SNP87
SNP122
SNP231
SNP223
SNP226
SNP204
SNP63
SNP9
SNP101
SNP181
SNP171
SNP218
SNP210
SNP229
SNP134
SNP133
SNP211
SNP129
SNP130
SNP131
SNP168
SNP174
SNP176
SNP182
SNP189
SNP192
SNP200
SNP202
SNP206
SNP11
SNP29
SNP39
SNP75
SNP95
SNP105
SNP132
SNP172
SNP178
SNP205
SNP57
SNP66
SNP78
SNP92
SNP107
SNP135
SNP136
SNP138
SNP139
SNP148
SNP162
SNP54
SNP77
SNP80
SNP88
SNP91
SNP110
SNP197
SNP6
SNP7
SNP14
SNP30
SNP58
SNP72
SNP79
SNP100
SNP149
SNP152
SNP154
SNP166
SNP191
SNP195
SNP198
SNP3
SNP5
SNP12
SNP13
SNP43
SNP44
SNP49
SNP53
SNP55
SNP61
SNP68

Imagen Anexo 6. Leyenda de los marcadores moleculares representados mediante PhenoGram