VNIVERSITAT DE VALÈNCIA

Facultat de Matemáticas
Departament d'Estadística i Investigació Operativa

PROPUESTA BAYESIANA PARA LA ELABORACIÓN DE MAPAS GENÉTICOS

TESIS DOCTORAL

Mónica Alacreu García

2016

Dirigida por:

Emilio A. Carbonell Guevara

M. Asunción Martínez Mayoral

UNIVERSITAT DE VALÈNCIA
Facultat de Matemàtiques
Departament d'Estadística i Investigació Operativa

PROPUESTA BAYESIANA PARA LA ELABORACIÓN DE MAPAS GENÉTICOS

Programa de doctorado 130 E Estadística y Optimización.

Tesis doctoral.
Realizada por: Mónica Alacreu García
Dirigida por: Emilio A. Carbonell Guevara y
M. Asunción Martínez Mayoral
Tutelada por: M. Teresa Rabena Pérez

Don Emilio A. Carbonell Guevara, Profesor de Investigación del Instituto Valenciano de Investigaciones Agrarias (I.V.I.A.) y M. Asunción Martínez Mayoral, Profesor Doctor del Departamento de Estadística, Matemáticas e Informática (Centro de Investigación Operativa) de la Universidad Miguel Hernández

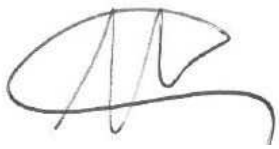
CERTIFICAN que la presente memoria de investigación:

"Propuesta bayesiana para la elaboración de mapas genéticos"

ha sido realizado bajo su dirección por Mónica Alacreu García y constituye su tesis para optar al grado de Doctor en Ciencias Matemáticas.

Y para que así conste, en cumplimiento con la normativa vigente, autorizan su presentación ante la Facultad de Matemáticas de la Universitat de València para que pueda ser tramitada su lectura y defensa pública.

Valencia, 14 de Enero de 2016.



Emilio A. Carbonell Guevara M. Asunción Martínez Mayoral

A todas las personas que me quieren.

Agradecimientos

Tengo una sensación extraña a la par que feliz al redactar estas palabras que, aunque dan comienzo a mi tesis doctoral, se escriben al final de todo, al final de un periodo infinito de mi vida. El camino ha sido largo y complicado pero también estimulante y gratificante. Ahora puedo decir: "ja!, lo conseguí".

Quiero expresar mi gratitud al Instituto Valenciano de Investigaciones Agrarias, cuya financiación ha hecho posible el siguiente trabajo de investigación a cargo de la beca doctoral: "Desarrollo y aplicación de metodologías estadísticas para la construcción de mapas genéticos en especies leñosas". Allí, durante varios años, la Unidad de Biometría fue mi casa y el Laboratorio de Genética la "nevera" que me proporcionó datos para investigar. Aprovecho para agradecer toda la colaboración a su responsable María José Asíns y a todo su equipo.

El trabajo ha transcurrido en paralelo a distintas etapas de mi vida, con muchísimas alegrías y con alguna que otra pena. Por ello, hay mucha gente involucrada a la que quiero agradecer su relación directa o indirecta con todo este logro personal.

Empiezo por mis padres académicos, Emilio y Asun. Sin duda, soy una privilegiada por los directores de tesis que tengo. Os agradezco vuestro tiempo, vuestra infinita paciencia, vuestra atención, vuestra disponibilidad, vuestra permisividad, vuestro apoyo, vuestra comprensión, vuestra crítica y un gran etc. Soy consciente de lo afortunada que he sido por beneficiarme de todo lo que sabéis. Simplemente, os admiro. Emilio, he de reconocer que me supera que siempre tengas razón. Sabes que envidio esa cabecita que Dios te ha dado. Te voy a echar de menos y sé que, en el fondo, tú a mi también. Te dedico la

siguiente imagen. Los dos sabemos el significado que tiene. Siempre serás mi jefe.

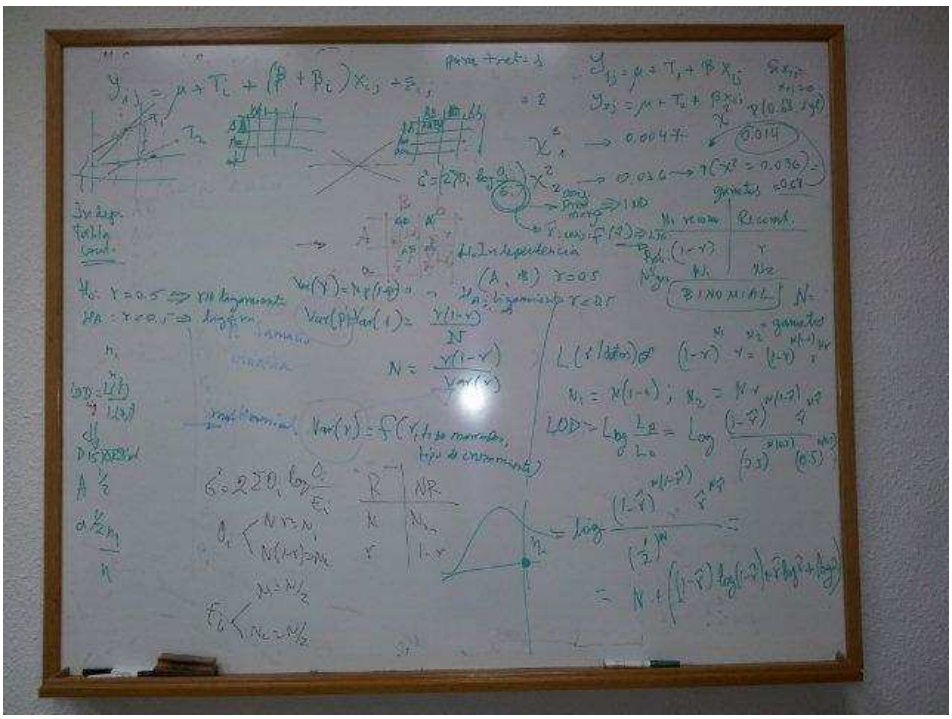


Figura 1: Pizarra.

Querida Carmen Armero, te agradezco tu tiempo y dedicación en mis comienzos, cuando empezaba a desplegar mis alas. Sé que el trabajo lo leerás con atención y bajo una perspectiva crítica, como no podía ser de otra manera, pero también sé que lo leerás con cariño y satisfacción. Por eso y más cosas, siempre serás una persona especial para mi.

"Amigas de la facu", sé que os vais a sentir orgullosas de mi y de este trabajo. Por eso, quiero daros las gracias. Por todo lo que compartimos durante muchos años, por seguir estando ahí, os dedico el resultado de mi esfuerzo. Todo lo que nos proponemos, lo conseguimos: licenciarnos, ser madres, ser trabajadoras y en definitiva, seguir aprendiendo. Espero que tengamos muchos

más éxitos que compartir.

Gracias a mis compañeros del IVIA. Jordi, Pilar, Marisol, Víctor y Guillermo. Qué suerte haberos conocido. Todos vosotros, poco a poco, os hicisteis importantes en vida. Siempre guardaré el mejor de los recuerdos de esa etapa de mi vida y sé que vosotros también. Menudas risas, eh? Por eso, un poco del alma de esta tesis es vuestro también.

Gracias a mis compañeros de la Universidad CEU Cardenal Herrera. Otra etapa importante en mi vida con nueva gente importante también. Estoy encantada de formar parte de esta casa porque tengo compañeros estupendos con los que comparto el día a día. Especialmente, quiero darte las gracias a ti, Paloma, por lo fácil que es trabajar y compartir asignaturas contigo. Me siento muy afortunada de tu compañía y ojalá sea así durante muchos años.

Para finalizar, quiero agradecer el apoyo de todas esas personas que, durante años, me han sufrido y me han arropado sin entender muy bien por qué significaba tanto para mi realizar una tesis doctoral. Me refiero a toda mi familia. Gracias a tod@s, a los que estáis y a los que estáis pero un poco más lejos, por vuestro ánimo y por vuestro refuerzo siempre positivo. Especialmente a mi marido y a mi hijo. Sin vosotros no soy nadie. En realidad, sois lo más importante de mi vida y por eso os dedico todo mi trabajo.

¡¡¡Gracias!!!

Prólogo

En los últimos años, los avances en la investigación biotecnológica, permiten manipular genes o grupos de genes de forma específica en el genoma de los organismos vegetales para producir cultivos con mejores características. Su objetivo es la obtención de especies mejoradas con peculiaridades deseables como un crecimiento más rápido, capaces de una adaptación al medio en el que se desarrollan, más resistentes a plagas o a enfermedades o simplemente proporcionando frutos más grandes, uniformes y de mejor calidad.

La investigación y el desarrollo sobre los marcadores moleculares está contribuyendo a la comprensión de los resultados derivados de la herencia de caracteres cuantitativos y a una mayor eficacia de las técnicas de mejora genética en la agricultura. Una estrategia de gran importancia en los programas de mejora consiste en la selección asistida por marcadores, que se basa en la detección de los factores responsables de la variación de los caracteres cuantitativos a mejorar. Es decir, los QTLs (del inglés, Quantitative Trait Loci). La localización de QTLs y los efectos que producen, se pueden inferir combinando la información de los genotipos y los fenotipos de individuos que provienen de poblaciones en desequilibrio, tales como la de los diseños experimentales controlados, Retrocruce y F_2. Es decir, cruzamientos que se llevan a cabo con el objetivo de conseguir características deseables en el cultivo que se quiere mejorar. Un análisis de localización de QTLs se fundamenta en la correcta estimación del mapa genético que define a la población de la que proceden. Sobre esta necesidad primaria se ha desarrollado la siguiente tesis doctoral.

El objetivo motivador del siguiente trabajo de investigación es el desarrollo de una metodología bayesiana, precisa, capaz de estimar mapas genéticos en tres escenarios experimentales: poblaciones con un diseño Retrocruce, poblaciones con un diseño F_2 con marcadores exclusivamente codominantes y por último, poblaciones con un diseño F_2 en el que intervienen conjuntamente

marcadores dominantes y codominantes. La estimación de un mapa genético implica elaborar una estrategia para establecer el reparto de marcadores en los distintos grupos de ligamiento y una vez designada esa pertenencia, definir la ordenación de los marcadores dentro de cada grupo de ligamiento. Estas dos fases han sido las investigadas a lo largo de los siguientes capítulos.

Existen metodologías muy bien definidas y contrastadas en el ámbito frecuentista (Capítulo 1, sección 1.4.). Sin embargo, ninguna de ellas, por construcción, se acompaña de un nivel adecuado de fiabilidad del mapa basado en probabilidades. Generalmente, el mapa se valida basándose en la localización de los mismos marcadores en otra población similar o en la existencia de mapas físicos con marcadores genéticos anclados en él. Por otro lado, se ha dado muy poca atención a los métodos bayesianos en este ámbito (Capítulo 1, sección 1.6.), por lo que el siguiente estudio puede suponer un avance en este sentido.

Índice general

Capítulo 1

Introducción.

1.1. Un poco de biología.

La **genética** como ciencia nació en 1900, cuando varios investigadores descubrieron el trabajo del monje austríaco Gregor Mendel que, aunque fue publicado en 1866, había sido ignorado en la práctica. Poco después del redescubrimiento de los trabajos de Mendel, los científicos sugirieron que las unidades mendelianas de la herencia, los **genes**, se localizaban en los cromosomas. Ello condujo a un estudio profundo de la división celular.

Cada célula de un organismo superior está formada por un material de aspecto gelatinoso, el **citoplasma**. Este material citoplasmático rodea un cuerpo denominado **núcleo**, que contiene, habitualmente, el mismo número par de

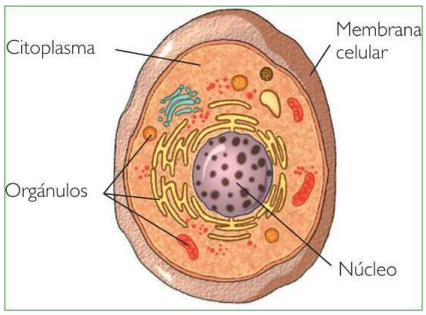


Figura 1.1: *Célula. http://www.escuelapedia.com/.*

cromosomas característico de la especie involucrada. Por ello, cada célula se dice que es **diploide**. Sin embargo, las células germinales de la especie o **gametos**, responsables de la reproducción, contienen en su núcleo la mitad de cromosomas. Son, por tanto, **haploides**.

La composición química de los cromosomas es ADN (ácido desoxirribonucleico) y proteínas y físicamente tienen un aspecto de hebras filamentosas.

Los animales, que se reproducen sexualmente, se forman a partir de la unión o fecundación de dos gametos de progenitores diferentes. En plantas, la autofecundación (gametos femenino y masculino del mismo progenitor) es un sistema de reproducción habitual. La célula resultante, **cigoto**, contiene toda la dotación doble de cromosomas de la especie. La mitad de estos cromosomas proceden de un progenitor y la otra mitad del otro. Estas parejas de **cromosomas homólogos**, tienen un estrecho parecido entre sí, en forma, tamaño y en la secuencia de ADN que representan. En los sucesivos gráficos, una pareja de cromosomas homólogos se representarán en color rojo y azul para distinguir el parental del que proceden.

Cualquier especie pluricelular está formada por miles de millones de células individuales organizadas en tejidos y órganos, que cumplen funciones específicas. Estas células han surgido a partir del cigoto por un proceso de división celular.

1.1.1. Mitosis

La Mitosis es un proceso de división celular asociada a las células somáticas (aquellas que no van a convertirse en células germinales sexuales o gametos). Mediante este proceso, los organismos pluricelulares, reemplazan células muertas o desgastadas, permitiendo la cicatrización, el crecimiento y la formación de nuevos tejidos. Los organismos unicelulares se reproducen por Mitosis también. En la Figura 1.2 se representa un esquema gráfico, reducido, de su proceso y a continuación, se describen brevemente sus etapas:

Interfase:

En el núcleo de la célula, cada hebra de ADN forma una copia idéntica a la inicial. Las hebras de ADN duplicadas, **cromátidas** hermanas, se mantienen unidas por el centrómero. La finalidad de esta duplicación es entregar, a cada célula hija, la misma cantidad de material genético que posee la célula original. Además, también se duplican otros orgánulos celulares como, por ejemplo, los

centríolos que participan directamente en la Mitosis.

Profase:

Las hebras de ADN se condensan y van adquiriendo la forma de cromosoma. Desaparecen la membrana nuclear y el nucléolo. Los centríolos se ubican en puntos opuestos de la célula y comienzan a formarse unos finos filamentos que, en conjunto, se llaman huso mitótico. En ese momento, el núcleo (ya sin membrana) y todos los componentes celulares están dispersos dentro del citoplasma.

Metafase:

Las fibras del huso mitótico se unen a cada centrómero de los cromosomas. Estos se ordenan en el plano ecuatorial de la célula, cada uno unido a su duplicado.

Anafase:

Los centrómeros se duplican. Cada duplicado del cromosoma se separa y es atraído a su correspondiente polo, a través de las fibras del huso. La Anafase constituye la fase crucial de la Mitosis, porque en ella se realiza la distribución de las dos copias de la información genética original.

Telofase:

En ella se desintegra el huso mitótico. La membrana nuclear y el nucléolo reaparecen, los nuevos cromosomas pierden su forma definida y se transforman en hebras filamentosas de ADN. Se forman dos núcleos idénticos encerrando cada una de las dos copias cromosómicas. El citoplasma comienza a separarse en la región de la línea ecuatorial en dos porciones iguales (citoquinesis) hasta que forma dos células diploides idénticas entre sí.

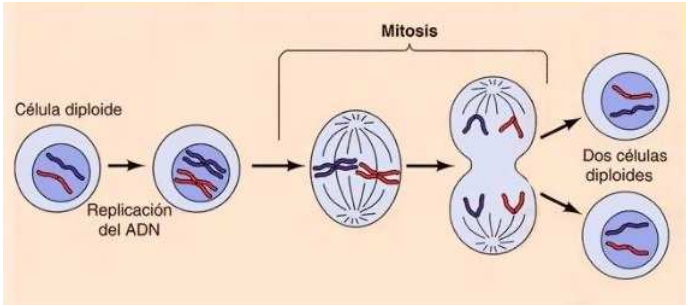


Figura 1.2: *Esquema gráfico del proceso de Mitosis. Imagen obtenida de "La Célula", Cooper 2002, ISBN 84-7101-356-8.*

1.1.2. Meiosis

La Meiosis es un proceso de división celular exclusivo para la formación de los gametos (células germinales o sexuales). Su característica principal radica en que a partir de una célula diploide se crean cuatro células haploides diferentes entre sí. Es decir, cuatro células hijas con la mitad de cromosomas que la célula madre y cuyos cromosomas transferidos son un mosaico de los cromosomas originales. La Meiosis consta de dos divisiones sucesivas de la célula (Meiosis I y Meiosis II) que ocurren tras una única duplicación previa del ADN, similar a la Intrafase de la Mitosis. Cada división meiótica se divide formalmente en los estados de: Profase, Metafase, Anafase y Telofase. De estas, la más compleja y de más larga duración es la Profase I. En la Figura 1.3 aparece representado este proceso de división.

Meiosis I

Profase I

Esta fase da comienzo con el material cromosómico (ADN) ya duplicado. Los hebras filamentosas se condensan, acortándose y engrosándose, dando lugar a los cromosomas. Se hace visible que cada cromosoma está constituido por dos cromátidas hermanas unidas por un centrómero. Muy pronto, los cromosomas homólogos (cada uno proveniente de un parental) se atraen entre sí y contactan íntimamente en toda su extensión. Este proceso de apareamiento se conoce como sinapsis y da lugar a una estructura de cuatro cromátidas entrelazadas, la tétrada. Mientras integran una tétrada, las cromátidas no hermanas intercambian porciones homólogas, fenómeno conocido como **sobrecruzamiento**. Esta recombinación de material hereditario contribuye a la variación de la descendencia.

Durante la Profase I, la célula sufre cambios similares a los explicados en la Mitosis. Los centríolos (si existen) se separan y aparecen el huso mitótico. La membrana nuclear y el nucléolo terminan desintegrándose.

Básicamente, la principal diferencia entre la Profase I en la Meiosis y la Profase de la Mitosis radica en la sinapsis.

Metafase I

En esta etapa la membrana nuclear y los nucléolos han desaparecido. En esta fase los centrómeros no se dividen; esta ausencia de división presenta una diferencia importante con respecto a la Mitosis. Cada cromosoma de origen paterno queda enfrentado a su homólogo de procedencia materna pero el hemisferio celular que ocupa cualquiera de ellos depende sólo de la casualidad.

Los dos centrómeros de una pareja de cromosomas homólogos se unen a las fibras del huso mitótico de polos opuestos.

Anafase I

Como en la Mitosis, esta etapa comienza con los cromosomas moviéndose hacia los polos. Cada grupo incluye una mezcla casual de cromosomas maternos y paternos, lo que se traduce finalmente en una amplia variedad de combinaciones cromosómicas. Tras la separación de los cromosomas homólogos, ocurre realmente la haploidía.

Telofase I

Esta etapa tiene aspectos variables dependiendo del tipo de organismo. En muchos organismos, esta etapa ni siquiera se producen; no se forma de nuevo la membrana nuclear y las células pasan directamente a la Meiosis II. En otros organismos dura poco; los cromosomas se alargan y se hacen difusos, y se forma una nueva membrana nuclear. En todo caso, nunca se produce una nueva duplicación de ADN y no cambia el estado genético de los cromosomas.

Meiosis II

Profase II

Esta fase se caracteriza por la presencia de un número haploide de cromosomas compactos recombinados y por el rompimiento de la membrana nuclear, mientras aparecen nuevamente las fibras del huso. Los centriolos se desplazan hacia los polos opuestos de las células.

Metafase II

En esta fase, los cromosomas se disponen en el plano ecuatorial. En este caso, las cromátidas aparecen parcialmente separadas una de la otra en lugar de permanecer perfectamente adosadas, como en la Mitosis.

Anafase II

Los centrómeros se separan y las cromátidas son arrastradas por las fibras del huso hacia los polos opuestos.

Telofase II

Esta fase concluye de forma similar a la Telofase de la Mitosis, dando lugar a cuatro células haploides que contienen diferente material genético que la célula diploide original.

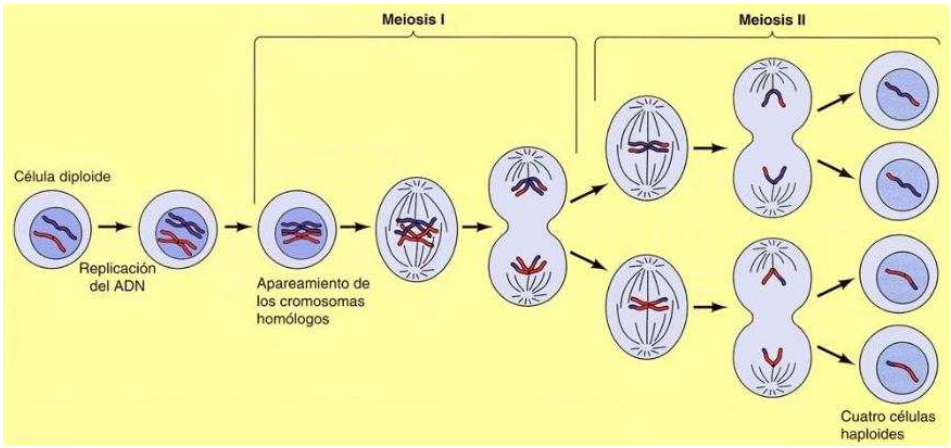


Figura 1.3: *Esquema gráfico del proceso de Meiosis. Imagen obtenida de "La Célula", Cooper 2002, ISBN 84-7101-356-8.*

	Comparativa	
	MITOSIS	MEIOSIS
Células implicadas	Se produce en células somáticas.	Se produce en las células madre de los gametos.
Número de divisiones	UNA sola división celular.	DOS divisiones celulares.
En la ANAFASE ...	... se separan cromátidas hermanas.	... de la primera división celular, se separan pares de cromosomas homólogos y en la de la segunda división celular, se separan cromátidas.
Sobrecruzamiento	No se produce.	Se produce entre cromosomas homólogos.
Duración	Corta.	Larga.
Resultado	Dos células hijas diploides con idéntica información genética	Cuatro células hijas haploides genéticamente distintas (gametos).
Finalidad	Crecimiento y renovación de células y tejidos. Mantenimiento de la vida del organismo.	Continuidad de la especie y dotación de la variabilidad genética.

En la fecundación, la unión de los gametos combina dos conjuntos de genes (o **marcadores**, como los denominaremos posteriormente), uno de cada parental. Por lo tanto, cada **gen** (o secuencia de ADN que puede codificar para una proteína u otro elemento (ARN, por ejemplo, con función celular conocida)) que ocupa una posición específica (**locus**, o en plural **loci**) en un cromosoma y que afecta a un carácter particular, está representado por dos copias o **alelos**, que son secuencias de ADN muy parecidas, una procedente de la madre y otra del padre. Cada copia se localiza en la misma posición sobre cada uno de los cromosomas homólogos. Cuando las dos copias son idénticas se dice que el individuo es **homocigoto** para ese marcador en particular. Cuando no son idénticas, es decir, cuando cada parental ha aportado formas distintas (alelos) del mismo marcador o alelo, se dice que el individuo es **heterocigoto** para dicho marcador. En adelante, se llamará al marcador homocigótico o heterocigótico cuando en realidad se quiere decir que es el individuo el que manifiesta esta condición sobre el marcador. Ambos alelos están contenidos en los cromosomas homólogos del individuo, pero si uno es **dominante**, sólo se detecta éste, quedando el heterocigoto confundido con el homocigoto para el alelo dominante. Si se manifiestan ambos alelos, el marcador es **codominante**. Generalmente, los alelos de un marcador se designan por una única letra; un alelo se representa con la letra mayúscula (principalmente en el caso de que sea dominante) y el **recesivo** con la minúscula, o alternativamente para el caso de codominancia, mediante una letra con subíndice. En la Figura 1.4 se representan algunos de los elementos de una pareja de cromosomas homólogos.

Por ejemplo, la capacidad de una persona para pigmentar la piel, el cabello o los ojos, depende de la presencia de un alelo particular (A), mientras que la ausencia de esta capacidad, denominada albinismo, es consecuencia de otro alelo (a) del mismo marcador. Los efectos de A son dominantes; los de a, recesivos. Por lo tanto, los individuos heterocigotos (Aa), así como los homocigotos (AA), tienen una pigmentación normal. Las personas homocigotos para el alelo que da lugar a una ausencia de pigmentación (aa) son albinas. Cada hijo de una pareja en la que ambos son heterocigotos (Aa) tiene un 25 % de probabilidades de ser homocigoto (AA), un 50 % de ser heterocigoto (Aa), y un 25 % de ser homocigoto (aa). Sólo los individuos que son (aa) serán albinos. Observamos que cada hijo tiene una posibilidad entre cuatro de ser albino, pero no es exacto decir que en una familia, una cuarta parte de los

hijos estarán afectados. Ambos alelos estarán presentes en el material genético del descendiente heterocigoto, quien originará gametos que contendrán uno u otro alelo.

Una pareja de marcadores heterocigotos situados en un mismo cromosoma, pueden encontrarse en **fase de acoplamiento**, si los alelos que proceden del mismo parental están sobre el mismo cromosoma o en **fase de repulsión**, si sobre cada cromosoma aparecen alelos de distinto parental. Recordemos que según la nomenclatura utilizada, distinguimos los alelos que proceden de distintos parentales con mayúsculas y minúsculas En la Figura 1.4 los marcadores B, E y F son heterocigotos. La pareja B - E están en fase de repulsión y la B - F en fase de acoplamiento.

Se distingue entre la apariencia o característica manifestada de un organismo, y los marcadores y alelos que posee. Los caracteres observables representan lo que se denomina el **fenotipo** del organismo, y su composición genética se conoce como **genotipo**.

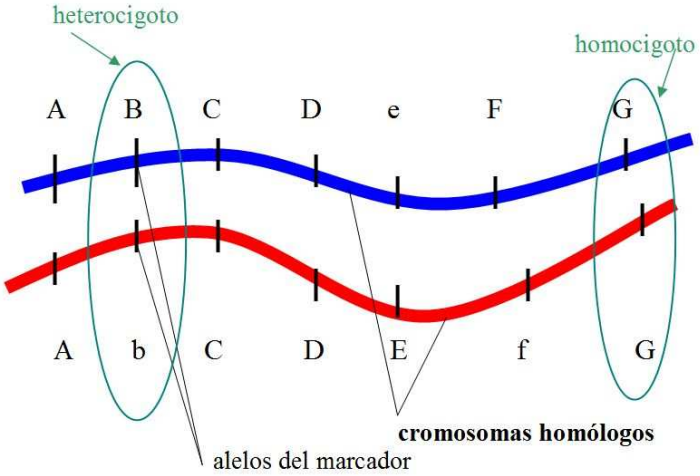


Figura 1.4: *Elementos en un cromosoma*

1.2. Fracción de recombinación.

El principio de Mendel según el cual "los marcadores que controlan diferentes caracteres son heredados de forma independiente uno de otro", es cierto sólo cuando los marcadores existen en cromosomas diferentes. El genetista estadounidense Thomas Hunt Morgan y sus colaboradores demostraron en una amplia serie de experimentos con la mosca del vinagre, que los marcadores se disponen de forma lineal en los cromosomas y que, cuando éstos se encuentran en el mismo cromosoma, se heredan como un bloque mientras el propio cromosoma permanezca intacto. Los marcadores que se heredan de esta forma se dice que están **ligados**. Sin embargo, Morgan y su grupo observaron también que este ligamiento rara vez es completo. Las combinaciones de características alélicas de cada parental pueden reorganizarse entre algunos de sus descendientes. Como se ha explicado en el apartado anterior, durante la Meiosis, una pareja de cromosomas homólogos puede intercambiar material genético cuando se producen los sobrecruzamientos.

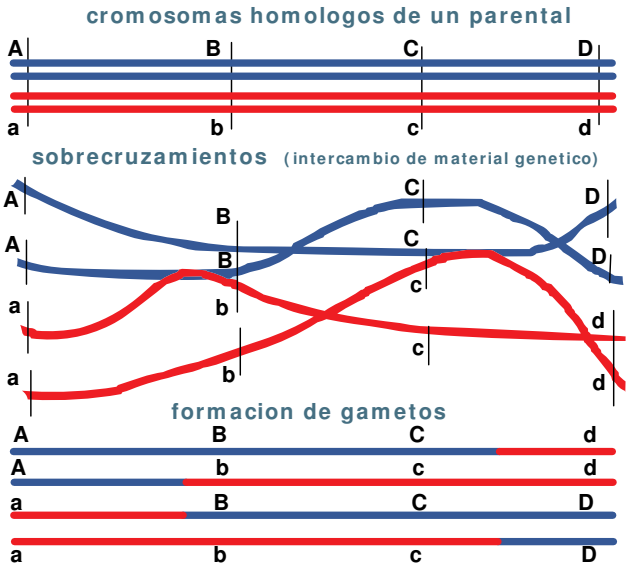


Figura 1.5: *Formación de gametos.*

Es importante hacer notar que, si el número de sobrecruzamientos entre

dos marcadores heterocigotos contiguos es impar, se produce una **recombinación**, que da lugar a gametos que no aparecen en los padres y que llamamos **gametos recombinantes**. Si el número de sobrecruzamientos es cero o par, el sobrecruzamiento no comporta recombinación, como se puede ver en la Figura 1.6.

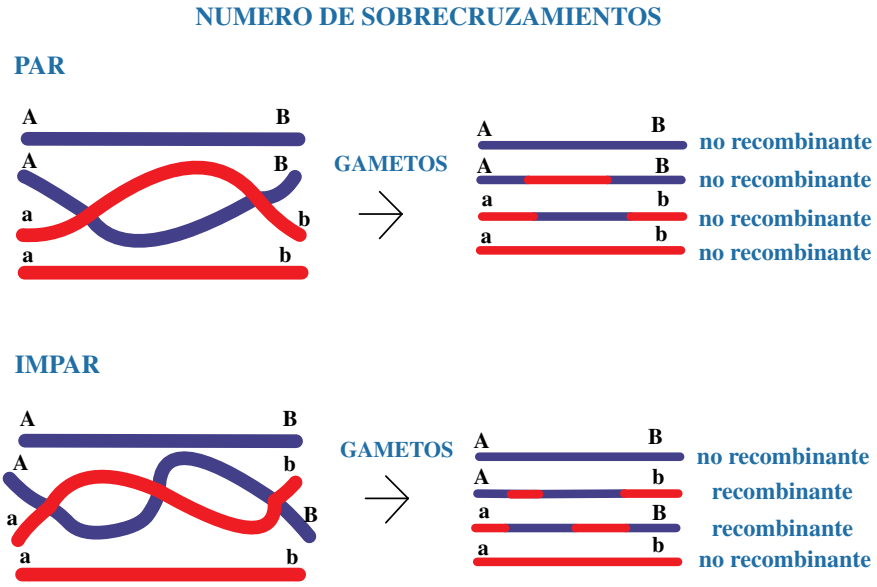


Figura 1.6: *Formación de gametos.*

Una hipótesis es que los sobrecruzamientos se producen más o menos al azar a lo largo de los cromosomas. Si los marcadores están relativamente alejados, los gametos recombinantes serán habituales; si están más o menos próximos, los gametos recombinantes serán poco frecuentes. Luego, la **frecuencia de recombinación** entre dos marcadores depende de la distancia que los separe en el cromosoma.

Si definimos como $2r$ la probabilidad de un sobrecruzamiento, la Figura 1.7 muestra el paso previo al cálculo de las probabilidades gaméticas para un grupo de ligamiento compuesto de dos marcadores heterocigotos:

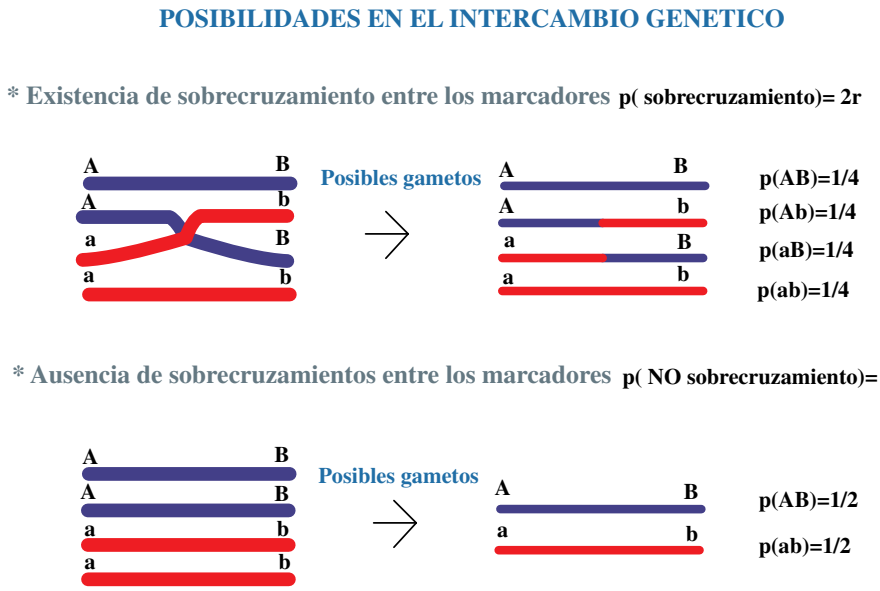


Figura 1.7: *Formación de gametos.*

Según observamos en la Figura 1.7, si llamamos "s" a la existencia de sobrecruzamiento y "ns" al no sobrecruzamiento, las probabilidades gaméticas vienen dadas por la expresión (1.1):

$$\begin{aligned} P(AB) = &\quad P(AB|s) * P(s) + P(AB|ns) * P(ns) = \tfrac{1}{4}(2r) + \tfrac{1}{2}(1-2r) = \tfrac{1}{2}(1-r) \\ P(Ab) = &\quad P(Ab|s) * P(s) + P(Ab|ns) * P(ns) = \tfrac{1}{4}(2r) + 0(1-2r) = \tfrac{1}{2}r \\ P(aB) = &\quad P(aB|s) * P(s) + P(aB|ns) * P(ns) = \tfrac{1}{4}(2r) + 0(1-2r) = \tfrac{1}{2}r \\ P(ab) = &\quad P(ab|s) * P(s) + P(ab|ns) * P(ns) = \tfrac{1}{4}(2r) + \tfrac{1}{2}(1-2r) = \tfrac{1}{2}(1-r) \end{aligned} \tag{1.1}$$

Entonces la probabilidad de gametos recombinantes resulta:

$$P(gametos\ recombinantes) = P(Ab) + P(aB) = r \tag{1.2}$$

De ahí que la **fracción de recombinación** "r" entre dos marcadores mida

la proporción de descendientes diferentes a los padres. Al definirse $2r$ como una probabilidad, $0 \leq 2r \leq 1$. $2r = 0$ ($r = 0$) indica que los loci de un par de marcadores ligados están tan cercanos que no tienen posibilidad de sufrir sobrecruzamientos. El otro caso extremo sería que los loci estén tan distanciados que los sobrecruzamientos entre esa pareja ocurran con toda seguridad. En ese caso, $2r = 1$ y $r = 0.5$ Luego la fracción de recombinación, r, varía entre 0 y 0.5 cM. En realidad, una fracción de recombinación igual a 0.5 representa a pares de marcadores no ligados o que pertenecen a distinto **grupo de ligamiento** (término más correcto y que necesita menos suposición que el de "cromosoma").

1.3. Mapa genético vs Mapa físico.

Un mapa representa la distribución de marcadores (y/o genes) en grupos de ligamiento (cromosomas) y su ordenación dentro de cada grupo y está basado en la definición de una distancia. Para un mapa genético la distancia es una función de la fracción de recombinación mientras que para un mapa físico la distancia está definida por el numero de pares de bases entre los marcadores. Cuanto más próximos estén dos marcadores con mayor probabilidad se heredaran conjuntamente. El mapa genético se obtiene mediante el análisis de la segregación en una descendencia o progenie dada. Si se repitiera la descendencia y se volviera a realizar el análisis de segregación (incluso utilizando los mismos marcadores), el mapa obtenido presentaría ciertas variaciones con respecto al anterior.

La elaboración de un mapa genético requiere un análisis de ligamiento que consta de cinco pasos:

0. Estudio de la distorsión en la segregación dentro de cada marcador. La distorsión en la segregación puede ser de origen gamético (uno de los alelos del heterocigoto se hereda con mayor probabilidad que el otro) o zigótico (la probabilidad de los genotipos de la descendencia no coincide con el producto de las probabilidades gaméticas; es decir, falta de independencia en la unión de los gametos). En la práctica, este paso es muy importante porque la distorsión puede venir de una errónea codificación de genotipos a partir de la lectura de los marcadores. Aunque la detección de éste y otros errores en la matriz de genotipos es una etapa fundamental (y limitante en la calidad del mapa) no se va tratar en esta Tesis. Actualmente existen metodologías y programas

muy establecidos en la literatura tales como Joinmap (Van Ooijen 2006 [85]) o R-Qtl (Broman et al. 2003 [13]) que incluyen herramientas para detectar estos problemas realizando un análisis exploratorio de datos En lo sucesivo, se considerará que en los análisis de ligamiento este estudio estará previamente hecho.

1. Un análisis por parejas para todas las posibles combinaciones de los diferentes loci para estimar la frecuencia de recombinación. El análisis está basado en la comparación de las frecuencias observadas y esperadas de los posibles genotipos. El número de posibles genotipos depende del número de alelos de los dos loci que se consideran en cada momento y del diseño experimental de cruzamientos (Retrocruce, F_2, líneas puras recombinantes, haploides duplicados, etc.). La propuesta más común para estimar las fracciones de recombinación a partir de las frecuencias genotípicas observadas es la máxima verosimilitud.

2. Agrupar los marcadores en distintos grupos de ligamiento. Los elementos clave para construir grupos de ligamiento son los valores de las estimas de las fracciones de recombinación, su nivel de significación y la información que tengamos del genoma (como el número de cromosomas, etc.)

3. Determinar la posición relativa de los marcadores (orden) dentro de los grupos de ligamiento. Este paso es clave para la obtención de un mapa genético de calidad. Aquí, el problema es cuantificar la precisión de la estimación del orden de marcadores.

4. Por último, estimar las fracciones de recombinación multipunto entre loci adyacentes, dada la ordenación obtenida en el paso anterior.

Un mapa físico muestra la localización física de los genes y otras secuencias de ADN de interés. La elaboración de un mapa físico es similar al ensamblaje de un rompecabezas ya que la secuencia de una molécula larga de ADN ha de construirse a partir de una serie de secuencias más cortas. Esto se realiza rompiendo la molécula en fragmentos, determinando la secuencia en cada uno de ellos y mediante el uso de técnicas de computación, buscar solapamientos para construir la secuencia final. El uso de los polimorfismos de un solo nucleótido o "single nucleotide polymorphisms" (conocidos por sus siglas SNPs) como marcadores de ADN para el genotipado de plantas, ha incrementado la posibilidad de cuantificar la variación en dianas específicas de ADN. Más importante aun, la información sobre millones de SNPs potenciales o pequeñas delecciones-inserciones y sus secuencias circundantes, establece las bases del genotipado de alto rendimiento (He et al. 2014 [30]). La gran demanda

de datos de secuencias a bajo costo ha producido como consecuencia el desarrollo de tecnologías de secuenciación de alto rendimiento (o secuenciación de siguiente generación) que pueden producir miles o millones de secuencias concurrentemente. Los métodos de secuenciación de siguiente generación (o NGS) se basan en secuenciaciones paralelas masivas y técnicas de imagen para producir varios cientos de millones de bases de ADN por cada carrera (Shendure and Ji 2008 [70]). Diversas plataformas como la Roche 454 FLX Titanium (Thudi et al. 2012 [83]), la Illumina MiSeq y HiSeq2500 (Bentley et al. 2008 [6]), Ion Torrent PGM (Rothberg et al. 2011 [67]) se han desarrollado recientemente que han permitido bajar los costos de la secuenciación de ADN (Deschamps et al. 2012 [18]; Quail et al. 2012 [61]) por lo que la relación entre mapa genético y mapa físico es más estrecha. De hecho, los mapas genéticos y físicos deben reflejar la misma estructura del genoma basada en el orden de los genes o marcadores y las distancias entre ellos. Sin embargo, la relación entre unidades de distancia de recombinación genética y longitud física del ADN, en pares de bases (pb), no es constante. La unidad de distancia genética es el Morgan o, más frecuentemente, el centimorgan (cM): distancia entre dos loci a la que se produce un promedio de 0.01 sobrecruzamientos. En realidad, la probabilidad de sobrecruzamiento varía enormemente de un punto a otro. Incluso hay zonas en el genoma (cerca del centrómero, por ejemplo) o sexo de una especie (machos de la especie Drosophila) o especies completas en donde no se producen sobrecruzamientos. Además, si alguno de los parentales es heterocigoto para una inversión o una traslocación, dentro de este cambio estructural no se detectará recombinación. Así pues, la resolución de un mapa genético está limitada por la posibilidad de detección de recombinaciones. Si en una determinada región del genoma no pueden ser detectadas recombinaciones entre marcadores, tampoco se podrá averiguar la posición relativa entre ellos y en consecuencia su mapa genético. Sin embargo, esta limitación no es un impedimento para desarrollar un mapa físico. Aunque las distancias entre marcadores pueden variar, el orden relativo debe ser el mismo en los dos mapas. La no correspondencia entre el mapa genético y el físico puede implicar errores en la construcción de alguno de los dos mapas (principalmente en el genético). En realidad, ambos mapas son complementarios. El cruce y comparación entre el mapa genético y físico de un genoma permite validar posibles errores, así como identificar y anclar marcadores de interés en el mapa físico que pueden ser utilizados posteriormente en la elaboración del mapa genético.

Nuestro objetivo final es la obtención de un mapa genético. A continuación presentamos los procedimientos clásicos para resolver los cuatro puntos que conforman un análisis de ligamiento, expuestos anteriormente.

1.4. Elaboración de mapas genéticos desde un enfoque frecuentista

Como ya hemos visto, la elaboración de un mapa genético se fundamenta básicamente en estimar las fracciones de recombinación entre parejas de loci, repartir los loci en grupos de ligamiento, dentro de cada grupo deducir el orden relativo de los loci o su localización en el cromosoma, y por último estimar las distancia entre los loci. En este apartado vamos a presentar los criterios tradicionales utilizados para obtener grupos de ligamiento y ordenar loci, una vez estimadas las fracciones de recombinación entre pares de loci.

1.4.1. Análisis por parejas de marcadores

La estimación de la fracción de recombinación entre dos marcadores depende del diseño experimental del cruzamiento del que provenga el genotipo. Los cruces que consideraremos en este trabajo son el Retrocruce y el F_2. En primer lugar consideramos un Retrocruce, que consiste en fecundar un individuo F_1 (resultado del cruce entre dos parentales totalmente homocigotos alternativos) con uno de sus padres. Cuando se selecciona el parental homocigoto recesivo, el Retrocruce se denomina cruzamiento de prueba (Figura 1.8).

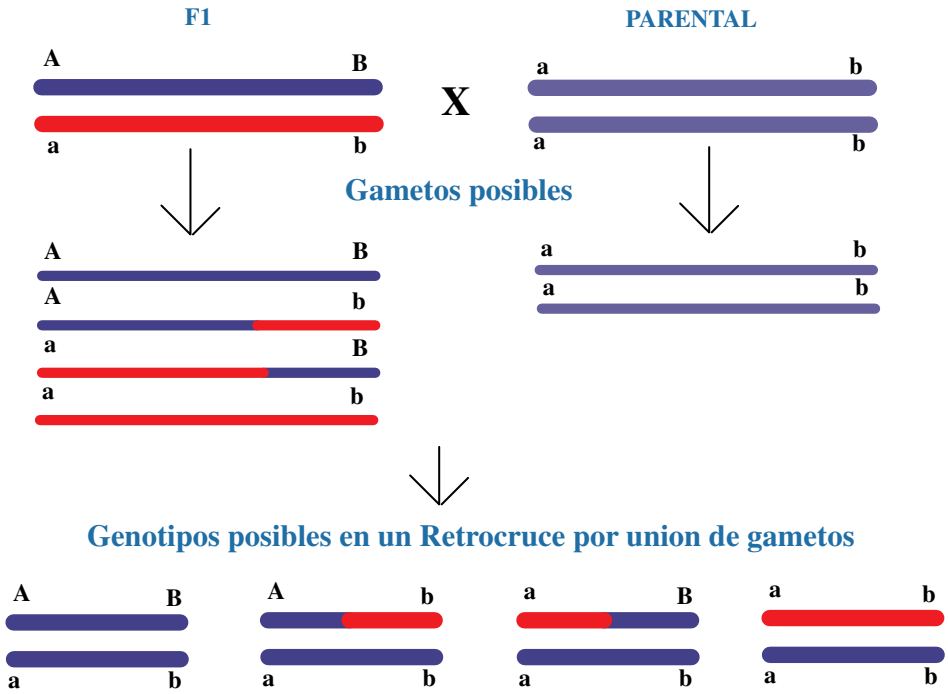


Figura 1.8: *Retrocruce (denominado cruzamiento de prueba, ya que el parental es homocigoto recesivo).*

Para un modelo con dos marcadores donde los abuelos tienen genotipos $AABB$ y $aabb$, uno de los padres del Retrocruce será $AaBb$. Considerando el otro padre, por ejemplo $aabb$, se producen los individuos viables por unión de gametos al azar, que se obtienen conforme a las probabilidades que aparecen en el Cuadro 1.1, siendo r la fracción de recombinación. En el cuadro se muestran, por filas cada uno de los gametos que genera uno de los parentales, junto con las frecuencias gaméticas, y en columnas los gametos generado por el otro parental, también con su frecuencia gamética. La configuración posible

de descendientes (genotipos), junto con sus frecuencias genotípicas aparece en las celdas de cruce. Estas frecuencias genotípicas se obtienen como producto de las frecuencias gaméticas de los gametos que los producen, asumiendo que no hay distorsión posible.

Frecuencias gaméticas y genotípicas de un Retrocruce	ab 1
AB $\frac{1}{2}(1-r)$	$AaBb$ $\frac{1}{2}(1-r)$
Ab $\frac{1}{2}r$	$Aabb$ $\frac{1}{2}r$
aB $\frac{1}{2}r$	$aaBb$ $\frac{1}{2}r$
ab $\frac{1}{2}(1-r)$	$aabb$ $\frac{1}{2}(1-r)$

Cuadro 1.1: Frecuencias gaméticas y genotípicas de un Retrocruce.

En definitiva, disponemos del Cuadro 1.2 con las frecuencias genotípicas esperadas y observadas:

Genotipos	Frecuencias esperadas (p_i)	Frecuencias observadas (f_i)
$AaBb$	$\frac{1}{2}(1-r)$	f_1
$Aabb$	$\frac{1}{2}r$	f_2
$aaBb$	$\frac{1}{2}r$	f_3
$aabb$	$\frac{1}{2}(1-r)$	f_4

Cuadro 1.2: Frecuencias genotípicas esperadas y observadas de un Retrocruce.

Considerando una distribución *Multinomial*, para las frecuencias genotípicas observadas, f_i, dadas las esperadas, p_i, el logaritmo de la función de verosimilitud de la fracción de recombinación r es:

$$l(r) = \sum_{i=1}^{4} f_i log(p_i) = (f_1 + f_4)log(1-r) + (f_2 + f_3)log(r) \tag{1.3}$$

Igualando la primera derivada a cero,

$$l'(r) = \frac{f_2 + f_3}{r} - \frac{f_1 + f_4}{1-r} = 0 \tag{1.4}$$

se obtiene la estimación máximo verosímil de la fracción de recombinación:

$$\hat{r} = \frac{f_2 + f_3}{n} \tag{1.5}$$

donde $n = f_1 + f_2 + f_3 + f_4$ es el número de individuos de la muestra. En definitiva se obtiene mediante la frecuencia observada de recombinación, tal como es lógico.

La generación F_2, como se puede ver en la Figura 1.9, se obtiene por autofecundación de la generación F_1, descendientes directos de los parentales homocigotos.

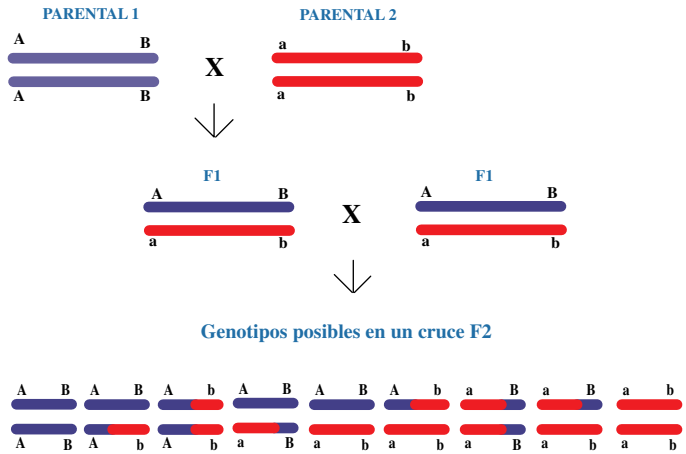


Figura 1.9: *Generación F_2.*

Al igual que hemos visto en el Retrocruce, los individuos se producen por unión de gametos al azar. En este caso, cada uno de los padres genera cuatro gametos posibles. Las frecuencias genotípicas (conjuntas) se obtienen como producto de las gaméticas marginales, conforme a las probabilidades que aparecen en el Cuadro 1.3, por considerar ausencia de distorsión.

Frecuencias gaméticas y genotípicas de una F_2	AB $\frac{1}{2}(1-r)$	Ab $\frac{1}{2}r$	aB $\frac{1}{2}r$	ab $\frac{1}{2}(1-r)$
AB $\frac{1}{2}(1-r)$	$AABB$ $\frac{1}{4}(1-r)^2$	$AABb$ $\frac{1}{4}r(1-r)$	$AaBB$ $\frac{1}{4}r(1-r)$	$AaBb$ $\frac{1}{4}(1-r)^2$
Ab $\frac{1}{2}r$	$AABb$ $\frac{1}{4}r(1-r)$	$AAbb$ $\frac{1}{4}r^2$	$AaBb$ $\frac{1}{4}r^2$	$Aabb$ $\frac{1}{4}r(1-r)$
aB $\frac{1}{2}r$	$AaBB$ $\frac{1}{4}r(1-r)$	$AaBb$ $\frac{1}{4}r^2$	$aaBB$ $\frac{1}{4}r^2$	$aaBb$ $\frac{1}{4}r(1-r)$
ab $\frac{1}{2}(1-r)$	$AaBb$ $\frac{1}{4}(1-r)^2$	$Aabb$ $\frac{1}{4}r(1-r)$	$aaBb$ $\frac{1}{4}r(1-r)$	$aabb$ $\frac{1}{4}(1-r)^2$

Cuadro 1.3: Frecuencias gaméticas y genotípicas de la generación F_2.

En realidad cuando se unen, por ejemplo, el gameto Ab del padre y AB de la madre, se obtiene el mismo genotipo que cuando se une el gameto AB del padre y Ab de la madre. Igualmente, cuando se une el gameto AB de un parental con el ab de otro, da lugar al mismo genotipo que cuando se unen los gametos Ab y aB, aunque tengan una composición gamética distinta. Con lo cual se obtienen 9 genotipos diferentes, que se resumen en el Cuadro 1.4.

Como antes, asumiendo una distribución *Multinomial* para las frecuencias genotípicas observadas f_i, el logaritmo de la función de verosimilitud es:

$$\begin{aligned} l(r) = \; & \textstyle\sum_{i=1}^{9} f_i log(p_i) = 2(f_1+f_9)log(1-r) + (f_2+f_4+f_6+f_8) \\ & log[r(1-r)] + f_5 log(1-2r+2r^2) + 2(f_3+f_7)log(r) \end{aligned} \tag{1.6}$$

Genotipos	Frec. esperadas (p_i)	Frec. observadas (f_i)
$AABB$	$\frac{1}{4}(1-r)^2$	f_1
$AABb$	$\frac{1}{2}r(1-r)$	f_2
$AAbb$	$\frac{1}{4}r^2$	f_3
$AaBB$	$\frac{1}{2}r(1-r)$	f_4
$AaBb$	$\frac{1}{2}(1-2r+2r^2)$	f_5
$Aabb$	$\frac{1}{2}r(1-r)$	f_6
$aaBB$	$\frac{1}{4}r^2$	f_7
$aaBb$	$\frac{1}{2}r(1-r)$	f_8
$aabb$	$\frac{1}{4}(1-r)^2$	f_9

Cuadro 1.4: Frecuencias genotípicas esperadas y observadas de la generación F_2.

Sin embargo, en este caso el método de la máxima verosimilitud no proporciona una solución directa para la estimación de la fracción de recombinación, ya que la resolución de la ecuación $l'(r) = 0$ no da lugar a una expresión explícita para r. Por lo tanto, se debe de recurrir a métodos de aproximación como el de Newton-Raphson o el algoritmo EM. Además del motivo matemático, también influye el motivo genético. Como se razonaba tras el Cuadro 1.3, de la observación de los genotipos, no se puede identificar si los gametos que lo forman son recombinantes o no.

1.4.2. Criterios para formar grupos de ligamiento.

Un **grupo de ligamiento** se define biológicamente como un grupo de genes con su localización en el mismo cromosoma, y estadísticamente como un grupo de loci heredados conjuntamente según ciertos criterios estadísticos. En realidad, el resultado de ambas definiciones debería ser el mismo, pero en ocasiones los genes de un mismo cromosoma pueden agruparse estadísticamente en distintos grupos de ligamiento ya que podemos no observar un largo segmento del cromosoma, por no disponer de marcadores en esa zona y concluir que hay dos grupos de ligamiento. Veamos algunos criterios clásicos para formar grupos de loci.

Para poder formar los grupos de ligamiento es preciso estimar las fracciones de recombinación entre todas las parejas de loci, así como la significación del ligamiento entre ellos. Para ello se establece el contraste (1.7):

$$\begin{cases} H_0 : loci\ i\ y\ j\ no\ ligados\ (r \geq 0.5) \\ H_A : loci\ i\ y\ j\ ligados\ (r < 0.5) \end{cases} \tag{1.7}$$

Denotando como *L(.)* a la función de verosimilitud, y en particular $L(r_k)$ a la verosimilitud bajo la hipótesis nula (k=0) y alternativa (k=A), la resolución del contraste se puede abordar de formas diferentes:

1. Planteando la razón de verosimilitudes:

$$LR = \frac{L(r_A)}{L(r_0)} = \frac{L(\hat{r}_A)}{L(0.5)} \tag{1.8}$$

que se distribuye asintóticamente según una χ^2 con 1 grado de libertad.

2. Sin embargo, en el entorno genético se suele emplear el cálculo del "LOD score" en vez de la razón de verosimilitudes:

$$\text{LOD} = log_{10}[\frac{L(\hat{r}_A)}{L(0.5)}] = log_{10}(L(\hat{r}_A)) - log_{10}(L(0.5)) \tag{1.9}$$

Consideramos $r_{i,j}, z_{i,j}$ y $p_{i,j}$ que denotan respectivamente la fracción de recombinación, "LOD score" y p - valor obtenido para el contraste de ligamiento entre dos loci i y j. Se pueden utilizar distintas combinaciones de estos tres elementos como criterio para deducir si los loci i y j pertenecen al mismo grupo de ligamiento:

$$\{[r_{i,j} \leq c]\ y\ [p_{i,j} \leq b]\}\ o\ bien\ \{[r_{i,j} \leq c]\ o\ [p_{i,j} \leq b]\} \tag{1.10}$$

y también

$$\{[r_{i,j} \leq c]\ y\ [z_{i,j} \geq a]\}\ o\ bien\ \{[r_{i,j} \leq c]\ o\ [z_{i,j} \geq a]\}, \tag{1.11}$$

donde c es el valor máximo de la fracción de recombinación para ser declarado un grupo de ligamiento (en general, se suele considerar c=0.5 pero se puede

hacer variar hacia un valor menor), b es la significación máxima del p-valor para considerar un ligamiento y a es el valor mínimo del LOD score para considerar un ligamiento.

3. Alternativamente, para resolver el contraste, se puede llevar a cabo una prueba de independencia clásica definida por la tabla de contingencia que generan las frecuencias genotípicas observadas. De ese modo se evita el problema de la distorsión en la segregación (JoinMap 1995 [77]).

Procedimientos

En la práctica, para la construcción de un mapa genético, los grupos de ligamiento se elaboran combinando distintos criterios de agrupación, después elegido un criterio, se ensayan distintos valores de los parámetros para comprobar que los grupos de ligamiento son fiables.

En términos prácticos, se suele razonar de la siguiente forma con respecto al conjunto de grupos de ligamiento obtenido:

1. Se considera que el grupo de ligamiento obtenido es fiable si éste no cambia sobre una amplia gama de criterios.

2. Debe ser biológicamente coherente; es decir, estar de acuerdo con estudios previos y/o conocimientos que se tengan de la situación y orden de marcadores anclados en mapas físicos.

3. Es conveniente revisar los criterios que proporcionan grupos de ligamiento relativamente grandes con respecto al número total de marcadores, pues pueden ser signo de agrupaciones incorrectas.

4. Que aparezcan un gran número de marcadores genéticos sin agrupar puede ser indicador de una baja calidad en los datos (tamaño de población pequeña, pequeño número de marcadores o errores en la codificación de los datos).

1.4.3. Ordenación de los loci dentro de un grupo de ligamiento.

El orden de los loci se define como la disposición de los genes o marcadores genéticos en un grupo de ligamiento. Para m genes hay $m!$ posibles órdenes o $\frac{m!}{2}$, si se ignora su orientación. Es decir, por ejemplo, $ABC = CBA$. En esta sección vamos a tratar el caso de la ordenación de 3 loci y luego la generali-

zación al caso de muchos loci (multiloci).

1.4.3.1. La ordenación de 3 loci.

Consideramos la situación en la que hay 3 loci ligados, A, B y C. Existen pues tres posibles órdenes de los loci: ABC, ACB y BAC. Las frecuencias genotípicas esperadas no se pueden definir si no está determinado el orden de los loci en el genoma. Sin embargo, las 3 fracciones de recombinación entre cada par de loci r_{AB}, r_{AC}, r_{BC}, se pueden obtener sin conocer las posiciones de los loci en el genoma. Considerando que el orden de los tres loci es ABC, los tres marcadores generan dos intervalos: uno entre A y B que se denotará como el intervalo 1 (*int1*) y otro entre B y C que será el intervalo 2 (*int2*).

$$A \overbrace{\qquad}^{int1} B \overbrace{\qquad}^{int2} C$$

Según ese orden se considera, por simplificar, que $r_1 = r_{AB}$ y $r_2 = r_{BC}$. Lógicamente, los intervalos dependerán del orden dado. Si el orden de los loci fuera ACB, entonces $r_1 = r_{AC}$ y $r_2 = r_{BC}$; por lo tanto los valores de r_1 y r_2 cambian según el orden de los marcadores. Para un Retrocruce del tipo cruzamiento de prueba, las frecuencias esperadas para los 8 genotipos se muestran en el Cuadro 1.5. Esta tabla es una ampliación del Cuadro 1.2 que ya se mostró en la Sección 1.4.1.

La letra ζ de las celdas de frecuencias denota el **coeficiente de coincidencia** y 1 - ζ se define como la **interferencia**. Si no hay interferencia, entonces 1 - $\zeta = 0$ y las recombinaciones ocurren independientemente en los dos intervalos, es decir los intervalos no se interfieren; en ese caso, la frecuencia esperada de los dobles recombinantes (es decir, recombinación en los dos intervalos) será $r_1 r_2$. Si las recombinaciones en los intervalos AB y BC no son independientes, la frecuencia observada de los dobles recombinantes puede no ser significativamente igual a la esperada, bajo la hipótesis de independencia. Si usamos $recomb_{12}$ para denotar la frecuencia observada de los dobles recombinantes, el coeficiente de coincidencia es $\zeta = \frac{recomb_{12}}{r_1 r_2}$.

La interferencia es negativa, 1 - ζ <0, si ζ >1, es decir si significativamente hay más dobles recombinaciones observadas que esperadas bajo la hipótesis de independencia. La interferencia es positiva, 1 - ζ >0, si ζ <1, es decir si significativamente hay menos dobles recombinaciones observadas que esperadas bajo la hipótesis de independencia. Habrá ausencia de interferencia, 1 - ζ

Genotipo	Frecuencias observadas	Frec. esperadas sin interferencia (p_i)	Frec. esperadas con interferencia (p_{ζ_i})
AaBbCc	f_1	$0.5(1-r_1)(1-r_2)$	$0.5(1-r_1-r_2+\zeta r_1 r_2)$
AaBbcc	f_2	$0.5(1-r_1)r_2$	$0.5(r_2-\zeta r_1 r_2)$
AabbCc	f_3	$0.5r_1 r_2$	$0.5\zeta r_1 r_2$
Aabbcc	f_4	$0.5(1-r_2)r_1$	$0.5(r_1-\zeta r_1 r_2)$
aaBbCc	f_5	$0.5(1-r_2)r_1$	$0.5(r_1-\zeta r_1 r_2)$
aaBbcc	f_6	$0.5r_1 r_2$	$0.5\zeta r_1 r_2$
aabbCc	f_7	$0.5(1-r_1)r_2$	$0.5(r_2-\zeta r_1 r_2)$
aabbcc	f_8	$0.5(1-r_1)(1-r_2)$	$0.5(1-r_1-r_2+\zeta r_1 r_2)$

Cuadro 1.5: Frecuencias genotípicas esperadas sin y con interferencia.

= 0, si $\zeta = 1$, es decir si las dobles recombinaciones observadas y esperadas coinciden significativamente. Por último, la interferencia es completa, 1 - ζ = 1, si ζ=0, es decir, las recombinaciones en los intervalos son mutuamente excluyentes. Si hay recombinación en un intervalo, no la habrá en el adyacente, luego no se observarán dobles recombinaciones.

El logaritmo de la función de verosimilitud para un modelo Retrocruce de tres loci bajo el orden ABC y asumiendo que las frecuencias observadas de recombinación f_i se distribuyen según una distribución *Multinomial* dadas las frecuencias esperadas y la interferencia, es:

$$\begin{aligned} l_{ABC}(r_1, r_2, \zeta) = \ & \textstyle\sum_{i=1}^{8} f_i \, log p_{\zeta_i} = (f_1+f_8)log(1-r_1-r_2+\zeta r_1 r_2)+ \\ & +(f_3+f_6)log(\zeta r_1 r_2) + (f_2+f_7)log(r_2-\zeta r_1 r_2)+ \\ & +(f_4+f_5)log(r_1-\zeta r_1 r_2) \end{aligned} \tag{1.12}$$

Las estimaciones máximo verosímiles de las fracciones de recombinación para los pares de loci r_{AB}, r_{AC} y r_{BC} se obtienen mediante una extensión de lo demostrado en (1.5):

$$\begin{aligned} \hat{r}_{AB} = \ & (f_3+f_4+f_5+f_6)/n \\ \hat{r}_{AC} = \ & (f_2+f_4+f_5+f_7)/n \\ \hat{r}_{BC} = \ & (f_2+f_3+f_6+f_7)/n \end{aligned} \tag{1.13}$$

donde $n = \sum_{i=1}^{8} f_i$ es el tamaño de la muestra. Sus estimaciones no dependen del orden especificado. Sin embargo, la relación entre las estimaciones depende de este orden. Para el orden ABC, la relación es:

$$\begin{aligned} r_{AB} &= r_1 \\ r_{AC} &= r_1 + r_2 - 2\zeta r_1 r_2 \\ r_{BC} &= r_2, \end{aligned} \tag{1.14}$$

de donde podemos deducir el valor estimado para r_1 y r_2.

La estimación máximo verosímil del coeficiente de coincidencia es:

$$\hat{\zeta} = \frac{2n(f_3 + f_6)}{(f_3 + f_4 + f_5 + f_6)(f_2 + f_3 + f_6 + f_7)} = \frac{r_{\hat{AB}} + r_{\hat{BC}} - r_{\hat{AC}}}{2r_{\hat{AB}} r_{\hat{BC}}} \tag{1.15}$$

que es función del orden de los loci.

Al igual que ocurría para un modelo F_2 de dos loci, son necesarios métodos de aproximación para obtener las estimas de las fracciones de recombinación e interferencia para un modelo F_2 de 3 loci.

Funciones de mapeo

Como ya hemos visto en un modelo 3 loci, las fracciones de recombinación basadas en modelos dos loci no son aditivas. Cuando el número de loci aumenta, la complejidad de las relaciones entre las posiciones y las fracciones de recombinación aumentan. Para solucionar este problema es necesario el uso de funciones de mapeo que permiten la conversión de las fracciones de recombinación a distancias, que sí guardan una relación aditiva, por construcción. Existen un gran número de funciones de mapeo y su utilización depende de las especificaciones biológicas propias de cada problema. A continuación mostramos algunas de ellas:

	Func. mapeo D=*F(r)*	**Inversa $r=F^{-1}(D)$**	ζ	**Interferencia**
Morgan (1928)	r	D	0	Completa
Haldane (1919)	-0.5log(1-2r)	$0.5(1-e^{-2\|D\|})$	1	Ausencia
Kosambi (1944)	$\frac{1}{2}tanh^{-1}2r = \frac{1}{4}log\frac{1+2r}{1-2r}$	$\frac{1}{2}tanh(2D) = \frac{1}{2}\frac{e^{4D}-1}{e^{4D}+1}$	$2r$	*f(r)*
Carter & Falconer (1951)	$0.5(tan^{-1}2r + tanh^{-1}2r)$		$8r^3$	Fuerte
Felsenstein (1979)	$\frac{1}{2(k-2)}log\frac{1-2r}{1-2(k-1)r}$	$\frac{1-e^{2(k-2)D}}{2[1-(k-1)e^{2(k-2)D}]}$	k-(k-1)2r	k=1: Ausencia k<1: positiva k>1: negativa
Karlin (binomial) (1984)	$0.5n(1-(1-2r)^{\frac{1}{n}})$	$0.5(1-(1-\frac{2D}{n})^n)$		

Cuadro 1.6: Algunas funciones de mapeo.

Como en nuestro trabajo asumimos ausencia de interferencia, hemos utilizado como función de mapeo la función Haldane, aunque se podría emplear cualquier otra sin pérdida de generalidad.

A continuación revisamos algunos de los métodos que se suelen utilizar para obtener el orden de tres loci:

Método del doble cruzamiento

Consiste en localizar los genotipos más raros y asignarlos como dobles recombinantes. Si somos capaces de identificar las clases recombinantes, podemos determinar el orden de los loci.

Método de la fracción de recombinación entre dos loci

Consiste en determinar el orden de los loci mediante las magnitudes de las estimaciones de las tres fracciones de recombinación. La estimación más grande estará asociada a los loci de los extremos. El otro locus se colocará en el intermedio de los extremos.

Método de la verosimilitud

También es común comparar las verosimilitudes de los tres posibles órdenes. El orden con mayor verosimilitud es considerado el orden más probable. Hay que tener en cuenta que este método no será adecuado en modelos en los que asumimos presencia de interferencia y la incluimos en el modelo para ser estimada, ya que al ser un modelo completamente parametrizado, la verosimilitud de los órdenes posibles sería la misma y no permitiría discriminar.

1.4.3.2. La ordenación de multiloci.

En general, la ordenación conjunta de muchos loci se basa en minimizar el número total de sobrecruzamientos (Sturtevant 1913 [79]) de modo que la longitud del mapa sea mínima. El orden estadístico para muchos loci se puede considerar como una extensión de la propuesta para 3 loci. Para un grupo de ligamiento con un gran número de loci es imposible elaborar un modelo completamente parametrizado, ya que el número de genotipos observados también es grande, por lo que es necesario recurrir a técnicas de optimización. Para la ordenación de loci con un número de loci mayor que 6 es razonable asumir un modelo con ausencia de interferencia, ya que bajo esta suposición biológica la computación se simplifica.

Veamos algunos de los métodos más conocidos en la ordenación multiloci:

Método de los 3 loci

Para un grupo de ligamiento con más de 3 loci, el problema de ordenación se puede dividir en subproblemas de 3 loci. Por ejemplo, 4 loci ligados se pueden organizar como 4 posibles combinaciones de 3 loci. Para cada una de las tripletas hay 3 posibles órdenes y el mejor orden se puede determinar utilizando las propuestas descritas anteriormente. En general, para m loci, hay $\binom{m}{3}$ posibles combinaciones de 3 loci. Sin embargo, alguna de las tripletas puede no tener significado al contradecir el hecho de que los 3 loci pertenezcan al mismo grupo de ligamiento. Para las tripletas con los tres loci ligados se puede determinar el mejor orden. En la práctica, se puede escoger como primera la tripleta con los marcadores más juntos. Posteriormente, basándonos en esta tripleta, es posible añadir el siguiente locus más próximo y hacer un análisis de 3 loci para las combinaciones del nuevo locus y las posibles combinaciones 2 loci. Si hay una contradicción del nuevo locus respecto al orden, el locus no

se puede colocar en el mapa. Este procedimiento se repite hasta que se han integrado todos los loci. Las desventajas de este método son:

1. Algunos de los loci no se pueden localizar en el mapa debido a resultados contradictorios entre las distintas tripletas.
2. Se puede obtener un orden local óptimo en lugar de un orden global óptimo.
3. La información del orden de los loci se puede perder por la subdivisión de los loci en múltiples tripletas.

Método de la máxima verosimilitud

Lander y Green (1987) [43] desarrollaron un método de ordenación de loci basado en máxima verosimilitud. Considerando una secuencia lineal de m loci. El logaritmo de la función de verosimilitud para el orden 1,2,3,...., m-1, m, proviene de asumir un modelo *Binomial* para las frecuencias de recombinación observadas, $R_{i,i+1}$, entre loci adyacentes, dada la fracción de recombinación entre ellos, $r_{i,i+1}$. Esto es:

$$l(r_{i,j}) = \sum_{i=1}^{m-1} R_{i,i+1}[r_{i,i+1} log r_{i,i+1} + (1 - r_{i,i+1}) log(1 - r_{i,i+1})] \qquad (1.16)$$

$R_{i,i+1}$ puede ser diferente para diferentes combinaciones de loci, debido a los datos faltantes o a la falta de información de ligamiento para 2 loci adyacentes. Posteriormente se comparan las verosimilitudes de los diferentes órdenes posibles para encontrar el "mejor", identificado como el que tiene mayor verosimilitud.

Método de la mínima suma de fracciones de recombinación adyacentes (*SARF*)

Este método está basado en los índices *SARF*. Se define el índice *SARF* (Falk 1989 [20]) como la suma de fracciones de recombinación adyacentes para un orden dado 1,2,3,...m-1, m. Este índice se puede expresar pues, según:

$$SARF = \sum_{i=1}^{m-1} r_{i,i+1} \qquad (1.17)$$

Éste y alguno de los siguientes métodos, se basan en intentar que el mapa sea lo más corto posible ya que resulta lógico que la ordenación se haga de modo que las distancias resultantes entre los loci sean las mínimas posibles.

Si no hay datos faltantes y no hay distorsión en la segregación, hay una gran correlación entre las soluciones mínimo *SARF* y máxima verosimilitud de los órdenes.

Método del mínimo producto de fracciones de recombinación adyacentes (*PARF*)

Similarmente, Wilson (1988) [90] definió el *PARF* como el producto de las fracciones de recombinación adyacentes, dando como solución aquella ordenación con mínimo *PARF*. Para evitar que *PARF* sea cero, se puede reemplazar una fracción de recombinación nula por una tolerancia pequeña.

Método de la máxima suma de LOD score adyacentes (*SALOD*)

El *SALOD* (Weeks y Lange 1987 [87]), para un orden específico, se define como la suma de los LOD score entre marcadores adyacentes y la ordenación elegida es la de menor *SALOD*.

Olson y Boehnke (1990) [57] apuntaron que este criterio es sensible al contenido de información de los loci y a la tendencia a localizar juntos los loci altamente informativos. Por esta razón, propusieron modificaciones del *SALOD* basándose en las definiciones de Boststein et al. (1980) [12]. Realizaron un estudio de simulación para comparar el poder de ordenación de los métodos. Usaron los métodos anteriores descritos, excepto el método de máxima verosimilitud. Concluyeron que *SARF* y una modificación, *SALOD*, son los mejores métodos de ordenación de loci. Sin embargo, mencionaron que el funcionamiento de los métodos para ordenar loci puede variar de acuerdo a los diferentes ajustes genéticos.

Otros métodos

Dado un orden de los loci, el mapa de las distancias entre los loci puede ser estimado utilizando **mínimos cuadrados** (Buetow y Chakravarti 1987 [15]; Weeks y Lange 1987 [87]). Este método es una vía para obtener el mapa de distancias multi-punto mínimo cuadráticas.

También se pueden utilizar métodos como la mínima suma de probabilidad

de dobles recombinantes (***PDR***) y el método para encontrar la ordenación de genes con un mínimo de cruces obligatorios (***MOC***), propuesto por Thompson (1987) [82].

Todos estos métodos se pueden abordar mediante algoritmos de búsqueda exhaustiva o de aproximación. Los algoritmos de aproximación normalmente no garantizan soluciones óptimas; sin embargo, los resultados son bastante buenos para aplicaciones prácticas. En la ordenación de loci, el algoritmo **seriation** (Buetow y Chakravarti 1987 [15]; Buetow 1987 [14]) se utiliza para obtener mínimo *SARF*, y el **simulated annealing** (SA) (Jasen et al. 2001 [34]) se usa como algoritmo de aproximación para obtener mínimo *SALOD*. La búsqueda exhaustiva proporciona soluciones óptimas en el caso que se aplique sobre un problema específico con espacio de búsqueda restringido. Como ejemplo, el algoritmo **branching and bound** (BB) se propuso para la ordenación de loci (Thompson 1987 [82]).

Para el método *SARF*, la matriz de fracciones de recombinación entre dos marcadores es equivalente a la matriz de distancias para el problema del agente viajero (TSP). Los algoritmos como, por ejemplo, simulated annealing (SA) (Kirkpatrick 1983 [38]; Aarts y Korst 1989 [1]; Jansen et al 2001 [34]) y algoritmo genético (Holland 1975 [31]; Goldberg 1989 [25]) se han usado para la búsqueda de soluciones aproximadas para el TSP. También se han obtenido soluciones óptimas para este problema, utilizando algoritmos de búsqueda exhaustiva, como por ejemplo BB (Miller y Pekín 1991 [54]).

1.4.3.3 Probabilidad y confiabilidad de los órdenes estimados

Las fuentes de error en la determinación del orden correcto de los loci pueden ser el muestreo de los datos y la ineficiencia de los criterios y algoritmos utilizados en el análisis de ligamiento. Algunos investigadores piensan que el orden de los loci es único y que no está sujeto a error por muestreo o por la metodología empleada en su obtención, pudiéndose determinar mediante biología molecular, como por ejemplo en los mapas físicos. Sin embargo, el orden de los loci "siempre" se determina utilizando técnicas estadísticas sobre los datos, luego el orden de los loci está sujeto, cuanto menos, al error de muestreo. La pregunta de interés es cómo cuantificar los errores o cómo proporcionar una medida de confianza de la ordenación. Este hecho no es exclusivo en genética. En cualquier modelización relacional estadística, el objetivo de encontrar un modelo que estime de la forma más precisa posible la relación entre las

variables involucradas, no consigue nunca certeza absoluta de que el modelo propuesto y finalmente elegido sea el único y verdadero que describe la relación buscada. Veamos varias propuestas:

Método de la verosimilitud

Intuitivamente, el cociente de verosimilitudes entre dos órdenes alternativos podría usarse para cuantificar la confianza de un orden estimado. Sin embargo, como observaron Lathrop et al. (1987) [47], la interpretación de los cocientes de verosimilitudes es difícil, debido a la carencia de una distribución teórica que pueda conducir al cálculo de un nivel de significación, ya que la comparación se suele hacer entre órdenes alternativos emplazados en diferentes regiones del espacio paramétrico si no son modelos jerarquizados. La carencia de grados de libertad para la diferencia entre los diversos órdenes de loci se puede compensar utilizando la teoría standard para el test del cociente de verosimilitudes. Sin embargo, y a pesar de las dificultades mencionadas, se han utilizado propuestas de verosimilitud para cuantificar la confianza de los órdenes (Edwards 1987 [19]; Lathrop et al. 1987 [47]; Smith 1990 [73]; Keats et al. 1991 [37]).

El cociente de verosimilitudes, al ser relativo, sólo permite afirmar que un orden es mejor que el alternativo, pero no cuantificar la confianza del orden. En todo método estadístico, es necesario adjuntar la varianza o la probabilidad de error de los procedimientos de estimación utilizados, para mostrar la confianza de la estimación, aun cuando el estimador máximo verosímil es el mejor e insesgado. Además, el cociente de verosimilitudes no puede cuantificar la superioridad de un orden sobre otro. Asimismo, la estimación del orden puede o no, ser insesgada, debido a los criterios y a los algoritmos utilizados.

Propuesta bootstrap

Presupone poder repetir un experimento varias veces y en cada uno de ellos, estimar el orden de los m marcadores. La metodología considera una matriz en la que las columnas representan las distintas posiciones del genoma o loci (*Loc1, Loc2,...,Locm*) y las filas los distintos marcadores (*M1, M2,...,Mm*). Finalizados los experimentos, un elemento de la matriz, $P_{i,j}$, almacena la frecuencia relativa con la que el marcador *Mi* ha ocupado la posición *Locj*. Si todos los marcadores han ocupado su posición correcta (*Mk* ha ocupado siempre *Lock*) para cada uno de los experimentos, la matriz de frecuencias

coincide con la matriz identidad, que proporciona una confianza del 100 % sobre el orden. Confianza 0 % sobre el orden se obtiene si cada marcador se ubica con la misma frecuencia en cada posición. Es decir, $P_{i,j} = \frac{1}{m}$. En estos términos se define la probabilidad del orden correcto (POC) como

$$PCO = \frac{1}{m} \sum P_{ii} \tag{1.18}$$

donde P_{ii} son las probabilidades de la diagonal en la matriz de frecuencias. En la práctica no se suele trabajar con la repetición de experimentos. Lo habitual es disponer de una única muestra. La matriz de frecuencias se puede estimar utilizando un muestreo técnico, como por ejemplo remuestreo (bootstrap). La repetición de muestras bootstrap se pueden llevar a cabo con los datos experimentales de un grupo de ligamiento.

Apoyo global y apoyo de intervalo (Global and Interval support)
Algunos métodos para dar credibilidad a un grupo de ligamiento se basan en cálculos asociados al "LOD score", $log_{10}(L_1/L_2)$, con L_1 y L_2 definidos bajo una medida de credibilidad específica (Keats et al., 1991 [37]).

El **apoyo global o "global support"** indica la evidencia de que un locus pertenezca a un mapa. Se calcula tomando L_1 como la verosimilitud cuando el locus se inserta en el mapa y L_2 como la verosimilitud cuando el locus no está en el mapa. Un ligamiento de un locus a un mapa se declara como "indudable" cuando el valor del apoyo global es de 3 o más.

El **apoyo de intervalo o "interval support"** mide la evidencia de que un locus esté en un intervalo particular de un mapa, siendo L_1 la verosimilitud cuando el locus está en el intervalo y L_2 la verosimilitud localizando el marcador en cualquier otro intervalo. Por último, para expresar el apoyo para un orden dado, L_1 se toma como la verosimilitud del mejor orden y L_2 como la verosimilitud del orden dado (Ott, 1999 [59]).

Los loci con un apoyo de intervalo de al menos 3 se llaman "**framework loci**", y un mapa compuesto íntegramente por estos loci se llama "**framework map**" (Keats et al., 1991 [37]).

Además de los tres tipos de apoyo multipunto anteriores, se utiliza una cuarta para dar una medida de la evidencia de que un conjunto de loci formen

un grupo de ligamiento. Esta medida se denomina **"LOD score" generalizado** y se obtiene como el cociente entre la verosimilitud L_1 bajo el mejor orden y la obtenida cuando se asume que ningún loci está ligado, L_0.

$$log(L_1/L_0) = log(\frac{L(r_{1,2}, ..., r_{m-1,m})}{L(0.5, 0.5, ..., 0.5)}) \tag{1.19}$$

donde $r_{i,j}$ es la fracción de recombinación entre los loci i y j.

Combinación de Jackknife y Bootstrap

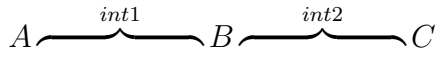
El procedimiento jackknife combinado con remuestreo se utiliza para evaluar el impacto de un locus sobre la estimación total del orden genético. La estrategia consiste en estimar el PCO para cada uno de los loci usando bootstrap cuando se descarta el locus en los datos (jackknifing). Si la estimación del PCO de la ordenación genética se incrementa significativamente cuando se descarta el locus, se entiende que el locus tiene una mala influencia sobre la estimación del orden. Por otra parte, si el PCO del orden estimado decrece significativamente cuando el locus se descarta, entonces el locus es esencial para la obtención de un orden con alta confianza.

1.4.4. Estimación de la fracciones de recombinación conjuntas dentro de un mapa

Es conocido que en un análisis de ligamiento, un modelo 'multipunto' proporciona más información que el modelo 'dos loci'. Sin embargo, existen 3 dificultades en relación al análisis de ligamiento para la implementación de modelos multiloci y la construcción del mapa genético:

1. La función de verosimilitud multiloci tiene un gran número de parámetros desconocidos si incluimos los parámetros de interferencia. Aun en ausencia de interferencia, la función de verosimilitud es compleja.

2. En la construcción de un modelo multipunto es habitual disponer de menos datos de los necesarios. El modelo multipunto está basado en la posible combinación de recombinaciones entre los loci; así pues, sería necesario disponer de grandes bancos de datos para observar al menos una recombinación en cada uno de los intervalos. Por ejemplo, para 3 loci ABC, hay 4 posibles combinaciones de recombinación en los dos intervalos AB y BC:



int1= 1 y *int2*= 1, *int1*= 1 y *int2*= 0, *int1*= 0 y *int2*= 1 y *int1*= 0 y *int2*= 0. Es decir, 11, 10, 01 y 00, donde "1" denota la recombinación en el intervalo y "0" la no recombinación en el intervalo. En general, hay 2^{m-1} combinaciones para m loci. Luego cuando m es grande, es impracticable obtener todas las posibles combinaciones de recombinación, debido al tamaño de la población, los marcadores no informativos y las fases de ligamiento desconocidas.

3. Construir un modelo multiloci es computacionalmente complejo, ya que es difícil obtener el orden más probable entre las $m!/2$ posibilidades de un modelo con m loci. Aun cuando el orden es simple, sigue siendo difícil estimar la verosimilitud multipunto con un gran número de parámetros desconocidos.

No se confunda las 2^{m-1} posibles combinaciones de recombinación que determinan los (m-1) intervalos que determinan m marcadores adyacentes en una determinada ordenación (y que caracteriza la dificultad 2), con los $m!/2$ posibles órdenes que existen para los m marcadores que determinan un grupo de ligamiento, aunque sólo uno de los órdenes es el correcto (que caracterizaría la dificultad 3)

Las funciones de mapeo son importantes en la construcción de un modelo multipunto porque disminuyen el número de parámetros. De hecho, la estimación de las interferencias es directa, al estar basada en la función de mapeo empleada. Sin embargo, la conveniencia de las funciones de mapeo depende de si las suposiciones biológicas bajo las que se utilizan son verdaderas o falsas.

Durante las últimas décadas, se han desarrollado muchas propuestas para construir modelos multipunto y estimar las correspondientes fracciones de recombinación conjuntas. A continuación veremos algunos métodos para obtener el estimador máximo verosímil de la fracción de recombinación entre dos loci, teniendo en cuenta el resto de loci.

Mínimos cuadrados

Para estimar un mapa de distancias multipunto, Stam (1993) [76] propuso un método de mínimos cuadrados, que es una versión modificada del algoritmo original introducido por Jensen y Jorgensen (1975a y b) [35] y [36] para estimar el mapa de distancias de la cebada. El método de mínimos cuadrados fue implementado en un paquete informático para mapeo genético, JoinMap

[86] (Stam 1993) [76]. Métodos similares fueron también descritos por Lalouel (1977) [41] y Weeks y Lange (1987) [87].

Se considera la siguiente notación para desarrollar el algoritmo de mínimos cuadrados, (Liu 1998) [49]:

$d_{i,j}$ = distancia multipunto entre los loci i y j.
$r_{i,j}$ = fracción de recombinación entre los loci i y j.
$\hat{r}_{i,j}$ = estimación de la fracción de recombinación $r_{i,j}$.
$F()$ = una función de mapeo (ver Cuadro 1.6).
$D_{i,j} = F(\hat{r}_{i,j})$ = distancia entre los loci i y j convertida directamente de la estimación de fracción de recombinación, $\hat{r}_{i,j}$, entre los loci i y j según la función de mapeo $F()$.
$s_{\hat{r}_{i,j}}$ = desviación típica de $\hat{r}_{i,j}$.
$s_{D_{i,j}}$ = desviación típica de $D_{i,j}$.
$E(D_{i,j})$ = esperanza de $D_{i,j}$.

Con esta notación se considera que si el locus i está situado a la izquierda del locus j, es decir $i<j$, se cumple que $E(D_{i,j}) = \sum_{k=i}^{j-1} d_{k,k+1}$. En definitiva, si se considera un mapa con m marcadores y un determinado orden, habrá m-1 distancias multipunto entre pares de loci adyacentes a estimar. Jensen y Jorgensen (1995) [35] asumen que la cantidad

$$\frac{D_{i,j} - \sum_{k=i}^{j-1} d_{k,k+1}}{s_{D_{i,j}}}, \tag{1.20}$$

se distribuye según una normal de media cero y varianza 1. $s_{D_{i,j}}$ se puede aproximar en función de la desviación típica del estimador de la frecuencia de recombinación entre pares de marcadores $s_{r_{i,j}}$, de tal forma que

$$s_{D_{i,j}} = s_{r_{i,j}} \Big| \frac{dF}{dr_{i,j}}(\hat{r}_{i,j}) \Big|, \tag{1.21}$$

donde $F()$ es la función de mapeo que convierte la fracción de recombinación a distancia entre dos loci (ver Cuadro 1.6). Como ya se mencionó en apartados anteriores la función de mapeo que hemos empleado es la función Haldane. En ese caso la expresión anterior sería:

$$s_{D_{i,j}} = s_{\hat{r}_{i,j}} \Big| \frac{d}{dr_{i,j}}[-0.5log(1-2\hat{r}_{i,j})] \Big| = \frac{s_{\hat{r}_{i,j}}}{1-2\hat{r}_{i,j}}, \tag{1.22}$$

Jesen y Jorgensen (1975) [35] definen la función de verosimilitud como

$$L = Constante \times exp\Big[-\frac{1}{2} \sum_{j=2}^{k+1} \sum_{i=1}^{j-1} \big(\frac{D_{i,j} - \sum_{k=i}^{j-1} d_{k,k+1}}{s_{D_{i,j}}} \big)^2 \Big] \qquad (1.23)$$

Derivando con respecto a $d_{k,k+1}$ el logaritmo de la expresión (1.23), se obtienen n-1 ecuaciones lineales:

$$\frac{dLog(L)}{dd_{k,k+1}} = \sum_{j=2}^{k+1} \sum_{i=1}^{j-1} \big(\frac{D_{i,j} - \sum_{k=i}^{j-1} d_{k,k+1}}{s^2_{D_{i,j}}} \big) = 0 \qquad (1.24)$$

Resolviendo estas ecuaciones se obtienen las distancias multipunto entre los loci adyacentes $d_{k,k+1}$.

Una modificación de este método se utilizará en los próximos Capítulos de esta Tesis, por lo que se muestra su funcionamiento mediante un ejemplo simple.

Ejemplo, (Liu 1998) [49]:

Se considera la siguiente información necesaria para estimar las distancias multipunto entre tres loci:

Marcadores	$\hat{r}_{i,j}$ $(s_{\hat{r}_{i,j}})$	$d_{i,j}$	$D_{i,j}$ $(s_{D_{i,j}})$ Haldane
1, 2	0.1 (0.03)	$d_{1,2}$	0.11 (0.038)
2, 3	0.15 (0.04)	$d_{2,3}$	0.18 (0.057)
1, 3	0.3 (0.13)	$d_{1,2} + d_{2,3}$	0.46 (0.325)

El logaritmo de la función de verosimilitud viene dada por la expresión:

$Log(L) = (\frac{0.11 - d_{1,2}}{0.038})^2 + (\frac{0.18 - d_{2,3}}{0.057})^2 + (\frac{0.46 - (d_{1,2} + d_{2,3})}{0.325})^2$

Derivando respecto $d_{1,2}$ y $d_{2,3}$ se obtiene el siguiente sistema de ecuaciones lineales:

$$\frac{0.11 - d_{1,2}}{0.038^2} + \frac{0.46 - d_{1,2} - d_{2,3}}{0.325^2} = 0$$

$$\frac{0.18 - d_{2,3}}{0.057^2} + \frac{0.46 - d_{1,2} - d_{2,3}}{0.325^2} = 0,$$

cuya resolución proporciona la siguiente estimación de las distancias multipunto:

$$\widehat{d}_{1,2} = 0.112 \; y \; \widehat{d}_{2,3} = 0.188$$

Algoritmo EM

El algoritmo EM es un grupo de procedimientos para obtener una estimación de la máxima verosimilitud cuando los datos experimentales son incompletos (Dempster et al. 1977 [17]). En nuestro contexto, el algoritmo EM se ha usado para estimar fracciones de recombinación y frecuencias alélicas (Ott 1977 [58]; Weir 1996 [89]; Liu 1998 [49]). Las estimaciones de las fracciones de recombinación multipunto (Lander y Green 1987 [43] y 1991 [44]) incluyen 4 pasos:

1. Partir de un punto inicial $r^{(old)}$.

2. Paso E: Usar $r^{(old)}$ como si fuera la verdadera fracción de recombinación y calcular el número esperado de recombinantes para cada intervalo.

3. Paso M: Usar el valor esperado como si fuera el verdadero valor y estimar la máxima verosimilitud $r^{(new)}$ para la fracción de recombinación.

4. Iterar los pasos E y M hasta que la verosimilitud converja a un máximo. $(r^{(old)}{=}r^{(new)})$

Este es el enfoque que se implementa en el programa Mapmaker [48] (Lander et al. 1987 [45]), que será base de comparación con el método descrito en esta Tesis

Ejemplo, (Liu 1998) [49]:

Para desarrollar el algoritmo EM, se considera la siguiente información sobre un mapa genético definido por tres marcadores, ordenados según ABC, en el que los intervalos contiguos son AB y BC:

	AB	BC	AC
verdaderas fracciones de recombinación	r_1	r_2	
nº recombinantes	Rc_1	Rc_2	
totales	$N_{1,2}$	$N_{2,3}$	$N_{1,3}$
nº recombinantes observados	$Rco_{1,2}$	$Rco_{2,3}$	$Rco_{1,3}$

Asumiendo ausencia de interferencia:
1: $r^{(old)} = (r_1, r_2)$
2: Paso E:
El número de recombinantes estimado en AB y BC, respectivamente, es:

$$Rc_1 = Rco_{1,2} + a_1 Rco_{1,3} + a_2(N_{1,3} - Rco_{1,3})$$
$$Rc_2 = Rco_{2,3} + (1 - a_1) Rco_{1,3} + a_2(N_{1,3} - Rco_{1,3})$$

donde, a_1 representa la probabilidad de que suceda una recombinación en el intervalo AB, dado que ha habido recombinación en el intervalo AC y a_2 representa la probabilidad de que suceda una recombinación en el intervalo AB, dado que no ha habido recombinación en el intervalo AC. Es decir,

$$a_1 = \frac{r_1(1-r_2)}{r_1(1-r_2)+r_2(1-r_1)}$$

$$a_2 = \frac{r_1 r_2}{r_1 r_2+(1-r_1)(1-r_2)}$$

3: Paso maximización, M:
$r_1^{(new)} = \frac{Rc_1}{N_{1,2}+N_{1,3}}$
$r_2^{(new)} = \frac{Rc_2}{N_{2,3}+N_{1,3}}$
4: Iterar los pasos E y M hasta que $r^{(new)} \simeq r^{(old)}$ con la precisión deseada.

Simulación

Otra propuesta para calcular el mapa de distancias multipunto es encontrar una función de mapeo que ajuste bien los datos. Esta propuesta implica comparar la verosimilitud multipunto de los datos utilizando distintas funciones de mapeo (Weeks et al. 1993 [88]). Como la distribución de la diferencia de verosimilitudes es desconocida, es necesaria la simulación para determinar la significación de la diferencia de verosimilitudes.

1.5. Ventajas del enfoque bayesiano

El análisis estadístico en el ámbito de la investigación genética se aborda, generalmente, desde un enfoque clásico, basado en contrastes de hipótesis, estimación e intervalos de confianza. Sin embargo, el enfoque bayesiano, aunque menos popular, ofrece la posibilidad de incorporar información a priori y además obtener conclusiones e interpretaciones mucho más intuitivas y próximas a los verdaderos intereses del investigador. Existen muchos trabajos en los que se explican las limitaciones de la estadística clásica frente a las ventajas del enfoque bayesiano: Shoemaker et al (1999) [71], Blasco (2001) [9], Beaumont et al (2004) [4], Blasco (2005) [10] o Blasco (2011) [11].

Enfoque clásico/ frecuentista

Los métodos clásicos definen la probabilidad en términos frecuentistas. La probabilidad tiene sentido si el experimento se puede repetir hipotéticamente muchas veces, bajo idénticas circunstancias y entonces la variabilidad existente entre los resultados observables es modelizable en términos de distribuciones de probabilidad, que dirigen todo análisis posterior. Además, los contrastes de hipótesis se resuelven calculando un p-valor cuya interpretación sería:

"Asumiendo la hipótesis nula como cierta, si el mismo experimento se repitiera muchas veces, la proporción de veces que el estadístico de contraste toma un valor tan o más extremo que el observado es p".

El p-valor está definido en términos frecuentistas e involucra exclusivamente a los datos y a la distribución de los estadísticos utilizados, derivada o aproximada a partir de la distribución asumida para los datos observados.

Esencialmente, proporcionan evidencia en contra de una hipótesis. A menudo, asumimos una hipótesis nula de no diferencia entre dos cantidades. Entonces se lleva a cabo el experimento y se calcula, para los datos, el valor muestral del estadístico de contraste apropiado. Para evaluar la evidencia en contra de la hipótesis nula, se compara el valor muestral del estadístico con su distribución bajo la hipótesis nula. Los valores extremos de los resultados observados se toman como evidencia en contra de la hipótesis nula. Un p-valor en un test de significación suele ser reinterpretado después como una medida de certidumbre o probabilidad sobre la falsedad de la hipótesis nula, hecho que no es cierto, pues en ningún momento la aproximación frecuentista tolera

asumir probabilidades sobre los parámetros desconocidos, y por consiguiente sobre hipótesis alguna sobre éstos. Además, es importante señalar que la significación estadística no siempre implica significación biológica.

Enfoque bayesiano

Desde la perspectiva bayesiana, la probabilidad se utiliza como una medida directa de la incertidumbre sobre los parámetros. En estadística bayesiana, se asume que todo aquello sujeto a variabilidad o incertidumbre, es modelizable a través de probabilidad. Así pues, la incertidumbre inicial existente sobre los parámetros desconocidos implícitos en un problema cualquiera es modelizable en términos de una distribución de probabilidad a priori, cuya variabilidad refleja el conocimiento disponible y la certidumbre que existe sobre su valor. La inferencia se basa en la distribución posterior $\pi(\theta|X)$, que se obtiene tras actualizar la distribución a priori con la distribución o modelo asumido para los datos y que aporta información sobre la variabilidad de éstos en torno a los parámetros desconocidos, según el Teorema de Bayes. Las conclusiones se proporcionan en términos de la probabilidad (posterior) de cualquier hipótesis de interés formulada sobre el parámetro objetivo.

Ventajas de los métodos bayesianos

El tratamiento de las preguntas de interés directo

En muchos casos, con los métodos bayesianos se pueden tratar las preguntas de interés de forma más directa que con los clásicos.

Un típico objeto de investigación sobre el análisis de ligamiento es determinar el grado de ligamiento entre dos loci. En un ajuste clásico, se investiga la evidencia en contra de la hipótesis nula de no ligamiento; en un enfoque bayesiano, se puede calcular la probabilidad de ligamiento, dados los datos, para un experimento en particular evaluable a través de la distribución posterior de los parámetros. El enfoque bayesiano contesta a la pregunta: "Dados los resultados observados, ¿cuál es la probabilidad de que dos loci estén separados por hasta m centimorgan?". Los análisis más tradicionales contestan a la pregunta: "Si dos loci estuvieran separados por m centimorgan, ¿cómo de verosímiles serían los resultados observados?" Parece que es de mayor interés para un investigador y más próximo a la realidad dar respuesta a la primera

pregunta.

Incorporación de la información a priori

Una distribución a priori ha de reflejar el conocimiento e incertidumbre existente sobre todos los parámetros implicados en un modelo, tengan más o menos interés por sí mismos. Existen alternativas informativas, como las distribuciones a priori conjugadas, capaces de capturar información y a la vez combinar paramétricamente con la verosimilitud para dar lugar a distribuciones posteriores en la misma familia. También existen alternativas mínimo o no informativas, usualmente impropias, que sirven para representar situaciones de ausencia de información sobre los parámetros del modelo.

Facilitar la comparación de muchos modelos/ hipótesis alternativos

En una aproximación clásica, trabajar con varias hipótesis puede ser un problema ya que sólo se pueden comparar dos hipótesis a la vez y esto supone considerar tantos pares de contrastes como sean necesarios; además, las conclusiones se formulan de forma indirecta. Sin embargo, en la aproximación bayesiana, se puede calcular la probabilidad posterior de cada hipótesis, pudiéndose interpretar de forma directa los resultados obtenidos.

Desventajas de los métodos bayesianos

Una de las críticas más comunes de los métodos bayesianos es que la elección de la distribución a priori es subjetiva. Esta crítica está relacionada con el hecho que la distribución posterior, en algunos casos, es muy sensible a la elección de la a priori. En estos casos, si dos investigadores utilizan los mismos datos y diferentes a prioris, pueden llegar a expresar diferentes conclusiones. Esta situación está relacionada con considerar la probabilidad más bien como un grado de creencia o incertidumbre que como una frecuencia a largo plazo. Con todo, esta crítica se diluye con la utilización de distribuciones a priori mínimo-informativas.

Otra dificultad a la que se enfrentan los métodos bayesianos es que su implementación, en general, puede ser muy compleja. Se deben especificar las distribuciones a priori de los parámetros y aunque se elijan de forma conveniente, la integración sobre los parámetros de interés se puede complicar en la práctica, especialmente en problemas de dimensión alta (con muchos pa-

rámetros). Sin embargo, el creciente desarrollo de técnicas como los métodos MCMC han hecho posible que los modelos bayesianos sean más accesibles.

Adicionalmente, los métodos bayesianos requieren de gran capacidad computacional, ya que, con frecuencia el uso de distribuciones mínimo o nada informativas deriva en distribuciones posteriores no analíticas y complejas, aunque no imposibles, de simular. Por ello, su aplicación puede ser complicada en el caso de emplear los marcadores actuales (SNPs) que pueden alcanzar un número de varios millares o incluso millones.

1.6. Estudios previos de mapas genéticos mediante métodos bayesianos

Dentro del mundo de la Genética y la Mejora Vegetal y Animal, se ha acogido con gran auge el aporte de los métodos bayesianos (Sorensen y Gianola, 2002 [74]), principalmente en la detección de factores responsables de la variación de un carácter cuantitativo (QTLs). Por ejemplo, Sillanpaa et al. (1998) [72], Van den Berg et al. (2013) [84] y Bink et al. (2014) [8]. Por otro lado, para llevar a cabo un análisis de QTLs es muy importante la disponibilidad de mapas genéticos que sean precisos y fiables, ya que "los errores en la construcción de un mapa genético pueden tener un impacto significativo respecto la localización de QTL en las etapas posteriores del análisis" (Anfock et al. 2014 [2]). Sin embargo, aunque el enfoque bayesiano puede ser una alternativa interesante al clásico frecuentista, son escasos los estudios para la obtención de mapas genéticos basados en esta metodología. Tal como se ha mencionado anteriormente, la enorme dificultad del tratamiento de datos procedentes de marcadores dominantes en fase de repulsión empleando tamaños de muestra reducidos y la rapidísima evolución, en los diez últimos años, de los tipos de marcadores moleculares y sistemas de detección hacia una mayor cobertura genómica y, a veces, directamente anclados en el mapa físico, podría ser una explicación de la falta de estudios recientes de este tema.

Stephens y Smith (1993) [78] basándose en Smith (1990) [73] formularon el problema de ordenar genes dentro del marco bayesiano. Con objeto de simplificar la estructura de la verosimilitud, supusieron que las recombinaciones podían ser observadas sin ambigüedad a partir de los datos, una suposición que no es correcta en algunos tipos de marcadores y en muchos diseños experi-

mentales de cruzamientos como F_2, parejas de medios hermanos, etc. Además, consideraron que el problema de la ordenación de marcadores era puramente un tema de estimación de parámetros y no de selección entre modelos alternativos (las diversas ordenaciones posibles).

Rogatko y Zacks (1993) [65] presentaron una alternativa para ordenar genes de una forma secuencial o paso a paso, basada en Teoría Bayesiana de la Decisión. En su trabajo tratan los diferentes órdenes de los marcadores como distintos modelos estadísticos candidatos, pero dedican muy poca atención a la estimación de las frecuencias de recombinación y a las probabilidades posteriores; por el contrario, se centran en la formulación de una regla bayesiana de decisión para identificar el "mejor" orden.

George et al (1999) [23] presentaron una alternativa bayesiana para el análisis del orden de marcadores en un tipo de familias en que la información es incompleta, como por ejemplo en los medios hermanos. Utilizaron MCMC para llevar a cabo los cálculos.

Las familias de medios hermanos están formadas de modo que un único parental (generalmente el padre en animales) se cruza con un gran número de individuos del otro sexo y produce un descendiente de cada cruce. Se dispone del genotipo del parental y de toda su descendencia, pero se desconoce el genotipo del otro parental (generalmente la madre en animales o el padre en plantas alógamas). Propusieron un modelo para el caso en que las recombinaciones no pudieran ser observadas sin ambigüedad, tratando los diferentes órdenes como modelos estadísticos competidores, identificando el "mejor" orden por medio de la distribución a posteriori. Dentro de la estrategia MCMC para seleccionar modelos, discuten propuestas para cambiar entre modelos alternativos y para determinar la probabilidad de aceptación de un determinado salto entre modelos basada en salto reversible. Dado un orden, para cambiar y moverse entre modelos con órdenes diferentes hacen "adelantar" o "retroceder" los marcadores de un mapa. En ambos casos se parte de un marcador seleccionado al azar que actúa como pivote. Para un cambio hacia delante, un marcador seleccionado al azar a la izquierda del pivote se pasa a la derecha del mismo. Reversiblemente, para un salto hacia atrás, un marcador al azar situado a la derecha del pivote salta a colocarse a su izquierda. Lógicamente, la distancia entre el pivote y el marcador que se mueve se mantiene, pero varían las distancias con respecto al resto de marcadores en el mapa, que deben ser calculadas cada vez, generando un coste computacionalmente pesado.

Rosa et al. (2002) [66] discutieron el problema de los sesgos introducidos en las estimaciones de los mapas cuando se ignoran los errores en la codificación de los datos y sugirieron un método robusto más realista para realizar inferencias, relacionando con las posiciones de los marcadores y las distancias entre ellos. Los datos procedían de diseños basados en retrocruces y duplo-haploides, empleando marcadores codominantes de modo que todos los genotipos son observables. Mediante una actualización simultánea del orden y las frecuencias de recombinación por Metropolis-Hastings (MH), proponen un nuevo orden, de acuerdo a una densidad generadora de candidatos. La elección de esa densidad es muy importante para la implementación eficiente del MCMC, especialmente si el número de marcadores es grande. Un proceso basado en densidades equiprobables podría generar un gran número de órdenes improbables o inconsistentes, lo que aumentaría la tasa de rechazo en el paso M-H. Para una mejor implementación, sugieren diversas formas para la generación de órdenes alternativos: intercambiar marcadores adyacentes, intercambiar dos marcadores elegidos al azar y rotar segmentos de longitud aleatoria (inversiones).

Siguiendo esta misma idea, York at al (2005) [91] presentaron una extensión al trabajo de Rosa et al (2002) [66] de modo que funciona eficientemente para el caso de un gran numero de datos (tanto individuos como marcadores) y para diseños tipo F_2. En su enfoque, modelizan directamente la presencia de cromosomas, de manera que para un conjunto denso de marcadores eliminan la necesidad de identificar previamente los grupos de ligamiento. Al igual que Rosa et al (2002) [66], su interés principal es el efecto de la codificación errónea y la existencia de datos faltantes. Similarmente, describen formas de proponer órdenes alternativos, añadiendo la posibilidad de traslocaciones o cambios de marcadores entre cromosomas, además de las inversiones. Para realizar inferencias sobre el orden se basan en la probabilidad posterior máxima (MAP), de modo que el orden de los marcadores viene dado por aquel que aparece más frecuentemente en la cadena, y para inferencias sobre la fracción de recombinación emplean MAP o la esperanza posterior (estimada como la media de las fracciones de recombinación a lo largo de la cadena). Para contrastar su método, emplearon datos simulados, comparando los resultados con los frecuentistas proporcionados por el programa Mapmaker [48] (Lander et al. 1987 [45]). Los resultaron fueron similares, pero no mejores que el enfoque frecuentista.

Simultáneamente a la investigación anterior, George (2005) [24], realizó

una propuesta MCMC para ordenar de forma conjunta marcadores a partir de datos observados en pedigríes, considerando la posibilidad de individuos no observados. El estudio se realizó utilizando datos simulados y reales. Concretamente, utilizó 4 muestras de datos simulados provenientes de dos mapas genéticos de 78 cM y 8 cM, ambos definidos por 8 marcadores. De cada mapa genético extrajo aleatoriamente dos muestras, una sin datos faltantes y otra con un 50 % de datos faltantes aproximadamente. Además trabajó con datos reales respecto a un cromosoma de 123 cM definido por 12 marcadores, con un 50 % de datos faltantes. Los marcadores eran codominantes, multialélicos y con frecuencias alélicas conocidas. Las fracciones de recombinación sin interferencia. George reconoció que su propuesta presentaba problemas como en el caso de trabajar con marcadores estrechamente ligados, especialmente cuando existen datos faltantes. Los resultados de los distintos órdenes obtenidos fueron valorados respecto a sus probabilidades posteriores, obteniendo mejores resultados en comparación con los obtenidos por el programa CRI-MAP (Lander y Green, 1987 [46]).

Ninguno de los trabajos anteriores estudiaron la bondad de su método en casos de marcadores dominantes estrechamente ligados, que es el caso que más problemas presenta, principalmente cuando el número de individuos analizados no es muy grande. El trabajo de York et al (2005) [91] podría considerarse una primera aproximación al problema a través de incluir el genotipo dominante como un caso de dato faltante, pero sigue sin tener en cuenta la posibilidad de que los marcadores estén en fase de repulsión que es, precisamente, el caso que mayores problemas presenta y será discutido en profundidad en la presente tesis.

1.7. Motivación e introducción a los capítulos de la tesis

En los últimos años, los avances en la investigación biotecnológica, permiten manipular genes o grupos de genes de forma específica en el genoma de los organismos vegetales para producir cultivos con mejores características. Esta manipulación genética, tiene como objetivo la obtención de especies mejoradas con peculiaridades deseables como un crecimiento más rápido, capaces de una adaptación al medio en el que se desarrollan, más resistentes a plagas o a

enfermedades o simplemente proporcionando frutos más grandes y de mejor calidad.

La investigación y el desarrollo sobre los marcadores moleculares está contribuyendo a la comprensión de los resultados derivados de la herencia cuantitativa y a una mayor eficacia de las técnicas de mejora genética en la agricultura. Una estrategia de gran importancia en los programas de mejora consiste en la selección asistida por marcadores, que se basa en la detección de QTLs. Múltiples investigaciones se han desarrollado en este sentido: Lande y Thompson (1990) [42], Hospital (2009) [32] Gupta et al (2010) [29], Kumar et al (2011) [40], Foolad et al (2012) [21] o Yue (2014) [92]. La localización de QTLs y los efectos que producen, se pueden inferir combinando la información de los genotipos y los fenotipos de individuos que provienen de poblaciones en desequilibrio, tales como la de los diseños experimentales controlados, Retrocruce y F_2. Estos son cruzamientos artificiales, que se llevan a cabo con el objetivo de conseguir características deseables en el cultivo que se quiere mejorar. Sin embargo, como se apuntó en el apartado anterior, debe tenerse en cuenta que un análisis de localización de QTLs se fundamenta en la correcta estimación del mapa genético que define a la población de la que proceden. Sobre esta necesidad primaria se ha desarrollado la siguiente tesis doctoral.

El objetivo motivador del siguiente trabajo de investigación es el desarrollo de una metodología bayesiana, precisa, capaz de estimar mapas genéticos en tres escenarios experimentales: poblaciones con un diseño Retrocruce, poblaciones con un diseño F_2 con marcadores exclusivamente codominantes y por último, poblaciones con un diseño F_2 en el que intervienen conjuntamente marcadores dominantes y codominantes tanto en fase de acoplamiento como en fase de repulsión. La estimación de un mapa genético implica elaborar una estrategia para establecer el reparto de marcadores en los distintos grupos de ligamiento y una vez designada esa pertenencia, definir la ordenación de los marcadores dentro de cada grupo de ligamiento. Estas dos fases han sido las investigadas a lo largo de los siguientes capítulos.

En los Capítulos 2 y 3, se modelizan cada una de las poblaciones de estudio según la aproximación tradicional (con distribuciones *Binomiales* y *Multinomiales*), y se añade la modelización previa de los parámetros con familias próximas a las familias conjugadas. Las primeras conclusiones se extraen a partir de las distribuciones posteriores para las fracciones de recombinación entre cada pareja de marcadores. Dado que dichas distribuciones no resultan

analíticas, se requiere el uso de simulación con algoritmos como Metropolis-Hastings. Con las simulaciones se estiman y aproximan los cálculos de probabilidad requeridos para proseguir con la ordenación de marcadores. En cada iteración de la cadena, se desarrolla un algoritmo de ordenación de los marcadores basado en las distancias mínimas entre ellos, *SARF*. Una vez obtenido el orden de los marcadores, de nuevo en cada iteración, se calculan las distancias multipunto entre los marcadores contiguos, según el método de mínimos cuadrados. Tras este proceso, se obtienen distintas estimaciones alternativas del mapa genético (modelos), cada una de ellas valorada por una probabilidad que permite diferenciar modelos más o menos probables en función de los datos observados. Las distancias entre marcadores contiguos se calculan finalmente a partir del valor esperado de la distribución final de las distancias multipunto, y su error con la desviación típica de dicha distribución. Para llevar a cabo la comprobación de la metodología, se prediseñan dos mapas genéticos definidos por 20 marcadores (Figura 1.10).

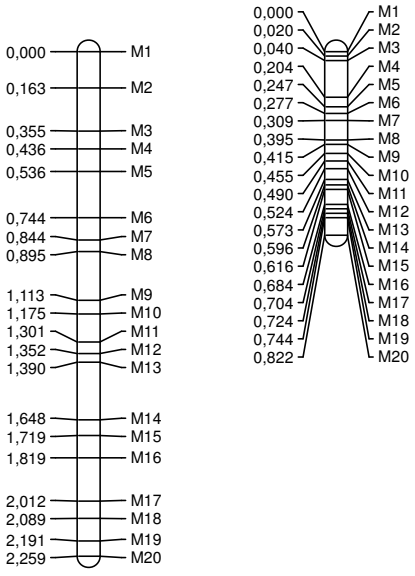


Figura 1.10: *Mapa genético menos denso (izquierda) y mapa genético más denso (derecha), prediseñados para recrear las distintas poblaciones de los Capítulos 2 y3.*

Aunque ambos mapas son densos, para considerar una situación "comple-

ja", uno es menos denso que el otro. Es decir, en uno de ellos los 20 marcadores se concentran en un mapa de tamaño global superior a 2 Morgan y en el otro los marcadores se concentran en menos de 1 Morgan. Para distinguirlos, en general, los llamaremos a partir de ahora "mapa menos denso" y "mapa más denso", respectivamente. En base a estos mapas, se extraen muestras de individuos de distintos tamaños, considerando que los mapas representan a cada una de las 3 poblaciones de interés. La metodología se ensaya con cada una de las muestras y, en algunos casos, se detectan problemas en la estimación de los mapas genéticos reales.

Con el objetivo de investigar exhaustivamente el problema de estimación detectado en los capítulos anteriores, y con el fin de investigar modelizaciones alternativas que puedan resultar más eficaces para la ordenación, en los Capítulos 4, 5 y 6 se reduce el mapa genético a la mínima expresión, con sólo tres o cuatro marcadores a ordenar, como se representan a continuación:

$$M_1 \overbrace{\overbrace{}^{r_1=0.02} M_2 \overbrace{}^{r_3=0.02} M_3}^{r_2=0.0392} \tag{1.25}$$

$$\overbrace{}^{r_5}$$
$$M_1 \underbrace{}_{r_1=0.017158} M_2 \underbrace{}_{r_4=0.030935} M_3 \underbrace{}_{r_6=0.018935} M_4, \tag{1.26}$$
$$\underbrace{}_{r_2}$$
$$\underbrace{}_{r_3}$$

$$\overbrace{}^{r_5}$$
$$M_1 \underbrace{}_{r_1=0.030935} M_2 \underbrace{}_{r_4=0.018935} M_3 \underbrace{}_{r_6=0.143953} M_4, \tag{1.27}$$
$$\underbrace{}_{r_2}$$
$$\underbrace{}_{r_3}$$

Mapas genéticos utilizados en los Capítulos 4, 5 y/o 6.

Se plantea una modelización previa para las fracciones de recombinación, así como una distribución previa mínimo-informativa para los posibles órdenes entre los tres marcadores. Implicando parámetros y órdenes en la modelización, se pretende actualizar la información sobre éstos y derivar probabilidades más precisas para fracciones de recombinación y mayor certidumbre para el modelo real. En los tres capítulos se proponen distintas variantes sobre la elección de la distribución propuesta (proposal) que utiliza el algoritmo. Las muestras consideradas en estos capítulos para ensayar la metodología, se obtienen de tres mapas prediseñados que representan distintas poblaciones F_2, en función de los tipos de marcadores designados.

Aunque los resultados no han sido totalmente satisfactorios, tras el estudio realizado en estos tres capítulos, se intuye desde el punto de vista práctico, que quizás se obtendrían mejores resultados si se estimara un mapa genético preliminar, eliminando alguno de los dos marcadores de las parejas estrechamente ligadas del tipo dominante en repulsión y se diseñara una estrategia para su posterior incorporación. Sobre esta idea se trabaja en el Capítulo 7.

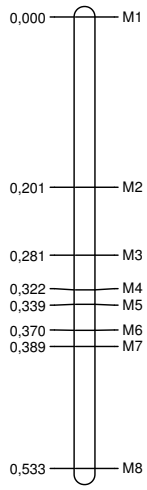


Figura 1.11: *Mapa genético, con 8 marcadores, para recrear las distintas poblaciones de los Capítulos 7, 8 y 9.*

Se elabora una metodología bayesiana equivalente a la introducida en el Capítulo 3 pero esta vez, se simula de la distribución posterior de las fracciones de recombinación para cada pareja de marcadores a través del programa

WinBugs [51], optimizando la computación y convergencia. Se diseña un algoritmo de ordenación entre marcadores acorde con la experiencia adquirida en los capítulos anteriores. Los resultados son tímidamente mejores que los obtenidos hasta ahora pero no tan satisfactorios como cabía esperar. En esta ocasión, el mapa genético teórico utilizado aparece representado en la Figura 1.11

En los Capítulos 8 y 9 se continúa simulando de la distribución posterior de las fracciones de recombinación para cada pareja de marcadores a través del programa OpenBugs [75] en los mismos términos que en el capítulo anterior y se ensaya un algoritmo de ordenación mucho más exhaustivo y laborioso que los ideados hasta el momento. En él, la estimación del mapa genético se inicia con una pareja de marcadores y prosigue incluyendo el resto de marcadores uno a uno, valorando su ubicación en el mapa entre todas las posibles. Además, tras la obtención del mapa completo, se proponen todas las posibles permutaciones entre cuatripletas de marcadores contiguos hasta obtener la estimación definitiva del mapa genético. Bajo esta metodología se obtienen resultados satisfactorios incluso en las situaciones que resultaban problemáticas anteriormente. Con el fin de comparar el comportamiento del método propuesto respecto de los métodos tradicionales implementados con los programas de referencia JoinMap [86] y Mapmaker [48], se aborda la simulación de múltiples muestras a partir del mapa teórico, y se realiza un análisis basado en la distribución en el muestreo aplicando los diferentes métodos de ordenación. Se concluye que la metodología bayesiana prácticamente iguala o mejora los resultados de la mejor de las metodologías frecuentistas en las tres poblaciones estudiadas. Sobretodo en el caso de tamaños muestrales pequeños, proporcionando además medidas de incertidumbre para la estimación del mapa obtenida. Los resultados en estos capítulos provienen de distintas muestras que representan las tres poblaciones de interés cuyo mapa genético aparece en la Figura 1.11.

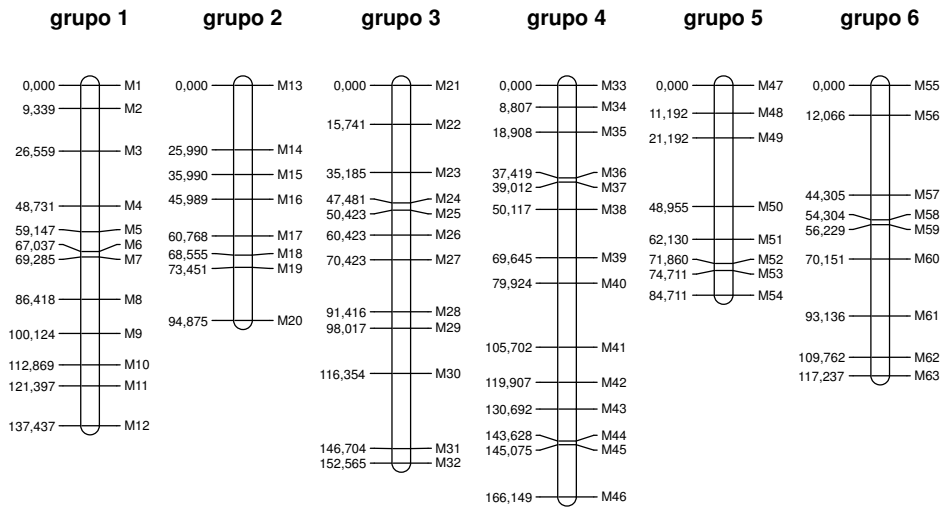
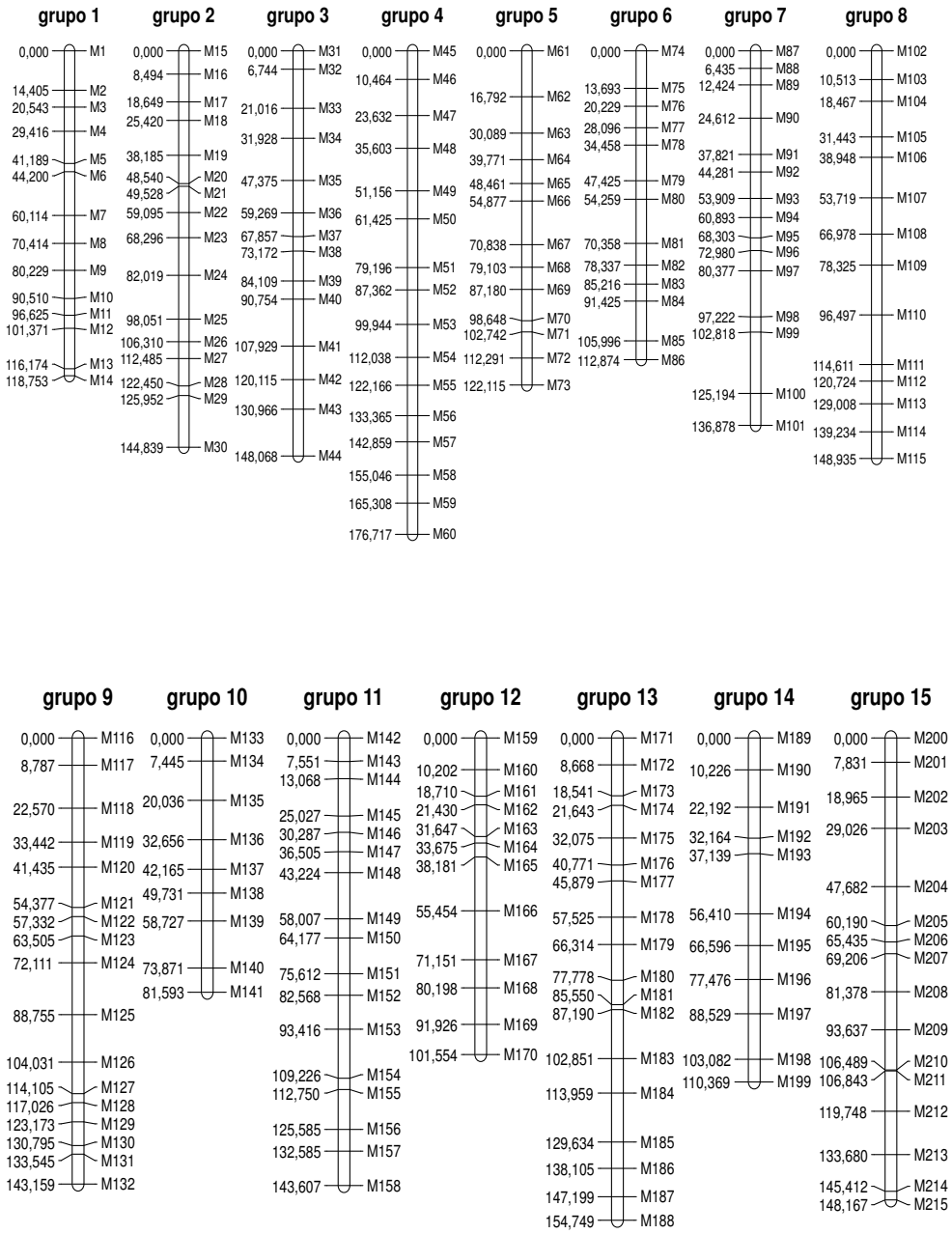


Figura 1.12: *Mapa genético de una población Retrocruce y F_2 con 63 marcadores repartidos en 6 grupos de ligamiento. Capítulo 10.*

Considerada razonable la metodología bayesiana diseñada para estimar fracciones de recombinación entre parejas de marcadores y el algoritmo de ordenación de marcadores para la obtención del mapa genético de la población, en el Capítulo 10 se amplía la metodología para cubrir un paso previo al ya investigado. Se trata de trabajar en un entorno más realista, en el que existen distintos grupos de ligamiento o cromosomas y es necesario determinar el reparto de marcadores en cada uno de los grupos de ligamiento para, posteriormente, estimar el mapa genético de la población. El algoritmo implementado valora, para cada pareja de marcadores, la condición de asociación y la de proximidad, que supone un ejercicio bayesiano de comparaciones múltiples, cuya resolución involucra al nivel de significatividad global empleado. De nuevo se ensaya la metodología sobre nuevas poblaciones Retrocruce y F_2 prediseñadas, en función de los mapas genéticos representados en las Figuras 1.12 y 1.13. Para considerar casi todos los tipos de dificultades más frecuentes que podrían presentar en datos reales.


grupo 1
0,000 M1
14,405 M2
20,543 M3
29,416 M4
41,189 M5
44,200 M6
60,114 M7
70,414 M8
80,229 M9
90,510 M10
96,625 M11
101,371 M12
116,174 M13
118,753 M14
grupo 2
0,000 M15
8,494 M16
18,649 M17
25,420 M18
38,185 M19
48,540 M20
49,528 M21
59,095 M22
68,296 M23
82,019 M24
98,051 M25
106,310 M26
112,485 M27
122,450 M28
125,952 M29
144,839 M30
grupo 3
0,000 M31
6,744 M32
21,016 M33
31,928 M34
47,375 M35
59,269 M36
67,857 M37
73,172 M38
84,109 M39
90,754 M40
107,929 M41
120,115 M42
130,966 M43
148,068 M44
grupo 4
0,000 M45
10,464 M46
23,632 M47
35,603 M48
51,156 M49
61,425 M50
79,196 M51
87,362 M52
99,944 M53
112,038 M54
122,166 M55
133,365 M56
142,859 M57
155,046 M58
165,308 M59
176,717 M60
grupo 5
0,000 M61
16,792 M62
30,089 M63
39,771 M64
48,461 M65
54,877 M66
70,838 M67
79,103 M68
87,180 M69
98,648 M70
102,742 M71
112,291 M72
122,115 M73
grupo 6
0,000 M74
13,693 M75
20,229 M76
28,096 M77
34,458 M78
47,425 M79
54,259 M80
70,358 M81
78,337 M82
85,216 M83
91,425 M84
105,996 M85
112,874 M86
grupo 7
0,000 M87
6,435 M88
12,424 M89
24,612 M90
37,821 M91
44,281 M92
53,909 M93
60,893 M94
68,303 M95
72,980 M96
80,377 M97
97,222 M98
102,818 M99
125,194 M100
136,878 M101
grupo 8
0,000 M102
10,513 M103
18,467 M104
31,443 M105
38,948 M106
53,719 M107
66,978 M108
78,325 M109
96,497 M110
114,611 M111
120,724 M112
129,008 M113
139,234 M114
148,935 M115
grupo 9
0,000 M116
8,787 M117
22,570 M118
33,442 M119
41,435 M120
54,377 M121
57,332 M122
63,505 M123
72,111 M124
88,755 M125
104,031 M126
114,105 M127
117,026 M128
123,173 M129
130,795 M130
133,545 M131
143,159 M132
grupo 10
0,000 M133
7,445 M134
20,036 M135
32,656 M136
42,165 M137
49,731 M138
58,727 M139
73,871 M140
81,593 M141
grupo 11
0,000 M142
7,551 M143
13,068 M144
25,027 M145
30,287 M146
36,505 M147
43,224 M148
58,007 M149
64,177 M150
75,612 M151
82,568 M152
93,416 M153
109,226 M154
112,750 M155
125,585 M156
132,585 M157
143,607 M158
grupo 12
0,000 M159
10,202 M160
18,710 M161
21,430 M162
31,647 M163
33,675 M164
38,181 M165
55,454 M166
71,151 M167
80,198 M168
91,926 M169
101,554 M170
grupo 13
0,000 M171
8,668 M172
18,541 M173
21,643 M174
32,075 M175
40,771 M176
45,879 M177
57,525 M178
66,314 M179
77,778 M180
85,550 M181
87,190 M182
102,851 M183
113,959 M184
129,634 M185
138,105 M186
147,199 M187
154,749 M188
grupo 14
0,000 M189
10,226 M190
22,192 M191
32,164 M192
37,139 M193
56,410 M194
66,596 M195
77,476 M196
88,529 M197
103,082 M198
110,369 M199
grupo 15
0,000 M200
7,831 M201
18,965 M202
29,026 M203
47,682 M204
60,190 M205
65,435 M206
69,206 M207
81,378 M208
93,637 M209
106,489 M210
106,843 M211
119,748 M212
133,680 M213
145,412 M214
148,167 M215

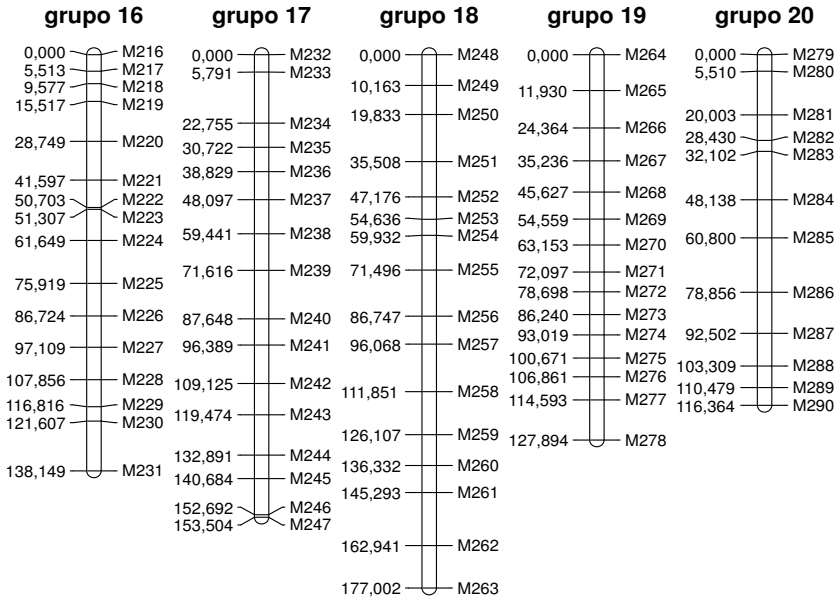


Figura 1.13: *Mapa genético de una población F_2 con 290 marcadores repartidos en 20 grupos de ligamiento. Capítulos 10 y 11*

En el Capítulo 11, se trabaja en el entorno más desfavorable de todos los planteados con anterioridad, es decir una población F_2 con marcadores codominantes y dominantes definida por 20 grupos de ligamiento, cuyo mapa genético aparece en la Figura 1.13. En este escenario, se prueba la metodología completa sobre muestras con distintos porcentajes de datos faltantes y se compara con los programas de referencia JoinMap [86] y Mapmaker [48].

En el Capítulo 12, se prueba la metodología bayesiana completa sobre datos reales de cítricos. Concretamente, una población definida por 52 marcadores en la que los datos han sido codificados como un Retrocruce y una población, tipo F_2, definida por 53 marcadores codominantes y dominantes conjuntamente. En este caso, el desconocimiento absoluto de los mapas genéticos reales de las poblaciones impide la valoración exacta de los resultados. Sin embargo, permite hacer un estudio comparativo con los resultados obtenidos por los programas JoinMap [86] y Mapmaker [48] y con los resultados previamente publicados por Raga et al. (2012) [63] y Bernet et al. (2010) [7].

Para finalizar, en el Capítulo 13, se elabora un discusión sobre los resultados obtenidos en los distintos capítulos y se establecen algunas líneas futuras de investigación recomendables para complementar el trabajo descrito en esta tesis doctoral.

Capítulo 2

Diseño Retrocruce (cruzamiento de prueba)

Recuérdese que un una población con un diseño Retrocruce provenía del cruce de un individuo F_1 (resultado del cruce entre dos parentales totalmente homocigotos) con uno de sus padres (Figura 1.8) y que por unión al azar de los gametos de cada parental, a cada marcador sólo le cabía la posibilidad de ser heterocigoto (Aa) u homocigoto (aa), para cada individuo.

2.1. Metodología

Para llevar a cabo el estudio, se ha diseñado un mapa genético, que representa a un diseño del tipo de cruzamiento Retrocruce, con una estructura de m marcadores. De él se extrae una muestra de n individuos independientes. El objetivo es estimar el mapa genético a partir de la información de la muestra de los n individuos. Puesto que se conoce el mapa de partida, podremos, en este caso, estudiar la bondad del método propuesto.

En primer lugar, definimos una variable observable (vectorial) $y_{.j}$, cuyo k-ésimo elemento, y_{kj}, es 0 si el individuo k es homocigoto para el marcador j y 1 si es el individuo k es heterocigoto para el marcador j. Esto es, para k= 1,...,n y j= 1,..., m:

$$y_{kj} = \begin{cases} 0, \ si \ aa \\ 1, \ si \ Aa \end{cases} \tag{2.1}$$

Por ejemplo, como se observa en el Cuadro 2.1.

	Marcadores						
Individuos	1	...	i	...	j	...	m
1	0	...	0	...	1	...	1
.	.	...	.	...	.	...	.
.	.	...	.	...	.	...	.
k	1	...	1	...	1	...	1
.	.	...	.	...	.	...	.
.	.	...	.	...	.	...	.
n	0	...	0	...	1	...	0

Cuadro 2.1: *Codificación de los marcadores en un diseño Retrocruce.*

Con esta codificación se define una variable binaria para cada par de marcadores i, j y para cada individuo k=1,...., n, que identifica si entre dos marcadores se da o no una recombinación:

$$z_{k,i,j} = \begin{cases} 1, si\ y_{ki} \neq y_{kj} \text{ (hay recombinación)} \\ 0, si\ y_{ki} = y_{kj} \text{ (no hay recombinación)} \end{cases} \tag{2.2}$$

o lo que es lo mismo

$$z_{k,i,j} = \begin{cases} 1, si\ i\ y\ j\ tienen\ diferente\ genotipo \\ 0, si\ i\ y\ j\ tienen\ el\ mismo\ genotipo \end{cases} \tag{2.3}$$

es decir, $z_{k,i,j} \sim Bernoulli(r_{i,j})$, donde $r_{i,j}$ es la proporción esperada de descendientes recombinantes para los marcadores i y j o fracción de recombinación entre i y j. Recordamos que es habitual en genética la nomenclatura r para referirse a estas probabilidades de recombinación.

De este modo, podemos definir la frecuencia absoluta observada de recombinantes o de recombinación, $R_{i,j} = \sum_{k=1}^{n} z_{k,i,j} \sim Binomial(n, r_{i,j})$

Extendiéndolo a todas las posibles combinaciones de m marcadores tomados de 2 en 2, se obtienen $m(m\text{-}1)/2$ frecuencias absolutas:

$$\mathbf{R}^* = \{R_{i,j} :\ i = 1, ..., m-1; j = i+1, ..., m\}$$

Asumiendo independencia entre todas las fracciones de recombinación entre pares de marcadores,

$$\boldsymbol{r}^* = \{r_{i,j} : \ i = 1, ..., m-1; j = i+1, ..., m\},$$

la función de probabilidad para los datos disponibles, $f(\mathbf{R}^*|\mathbf{r}^*)$, se obtiene como el producto de las correspondientes funciones de probabilidad de cada una de las variables binomiales independientes $R_{i,j}$ con parámetros n y $r_{i,j}$:

$$f(\mathbf{R}^*|\mathbf{r}^*) = \prod_{i=1}^{m-1} \prod_{j=i+1}^{m} f(R_{i,j}|r_{i,j}) \propto \prod_{i=1}^{m-1} \prod_{j=i+1}^{m} r_{i,j}^{R_{i,j}} (1 - r_{i,j})^{n-R_{i,j}} \tag{2.4}$$

Dado que la fracción de recombinación $r_{i,j} \in [0,\ 0.5]$, $\forall i, j$, la distribución a priori que asumimos para estas probabilidades de recombinación, es una *Beta* truncada en dicho intervalo, *BetaT*. Recordemos que la distribución *Beta* es la distribución conjugada para un modelo *Binomial*, de modo que esta modelización ofrecerá un análisis lo más próximo posible a un análisis conjugado. Así, dada la independencia entre todas las fracciones de recombinación, la distribución a priori para el vector de fracciones de recombinación, $\mathbf{r}^*$, será igual al producto de las distribuciones *BetaT* independientes:

$$\pi(\boldsymbol{r}^*) = \prod_{i=1}^{m-1} \prod_{j=i+1}^{m} BetaT(r_{i,j}|\alpha, \beta; \varepsilon) \tag{2.5}$$

siendo ε= 0.5 el parámetro de truncamiento según la definición de la *Beta* truncada en el Apéndice A, y (α, β)= (1, 1), los parámetros que proporcionan una modelización uniforme (distribución plana) en el intervalo [0,0.5].

Combinando la función de probabilidad de los datos con la distribución a priori de los parámetros, se obtiene la expresión de la distribución **posterior**:

$$\pi(\mathbf{r}^*|\mathbf{R}^*) \propto f(\mathbf{R}^*|\mathbf{r}^*)\pi(\mathbf{r}^*) \tag{2.6}$$

A la vista de la modelización, y dada la independencia asumida entre las fracciones de recombinación, $r_{i,j}$, en la distribución a priori, y las frecuencias

de recombinación, $R_{i,j}$, en la modelización de los datos, la distribución posterior, $\pi(\boldsymbol{r}^*|\mathbf{R}^*)$, se obtiene como el producto de distribuciones posteriores marginales:

$$\begin{aligned}\pi(\mathbf{r}^*|\mathbf{R}^*) &= \prod_{i=1}^{m-1}\prod_{j=i+1}^{m}\pi(r_{i,j}|R_{i,j}) \\ &\propto \prod_{i=1}^{m-1}\prod_{j=i+1}^{m} f(R_{i,j}|r_{i,j})\pi(r_{i,j}) \\ &\propto r_{i,j}^{R_{i,j}+\alpha-1}(1-r_{i,j})^{n-R_{i,j}}(\varepsilon-r_{i,j})^{\beta-1}I_{[0,\varepsilon]}(r_{i,j}) \\ &\propto BetaT(r_{i,j}|R_{i,j}+\alpha,\beta;\varepsilon)(1-r_{i,j})^{n-R_{i,j}}\end{aligned} \tag{2.7}$$

siendo $I_{[0,\varepsilon]}(\cdot)$ la función indicatriz en el intervalo [0, ε].

Así, para simular de la distribución posterior conjunta habremos de simular de las marginales, $\pi(r_{i,j}|R_{i,j})$. Para simular de estas marginales usaremos un algoritmo Metropolis-Hastings (MH), utilizando como distribución propuesta (proposal):

$$p(r_{i,j}) = \quad BetaT(r_{i,j}|R_{i,j}+\alpha,\beta;\varepsilon) \tag{2.8}$$

y como probabilidad de salto para aceptar una simulación $r_{i,j}$ en la iteración t+1 habiendo aceptado previamente $r_{i,j}^{(t)}$ en la iteración t:

$$prob.salto(r_{i,j}^{(t)}, r_{i,j}) = \quad min\{1, (\tfrac{\varepsilon-r_{i,j}}{\varepsilon-r_{i,j}^{(t)}})^{\beta-1}I_{[0,\varepsilon]}(r_{i,j})\} \tag{2.9}$$

Este procedimiento proporciona un porcentaje de saltos aceptable (superior al 70 %) para la mayoría de las distribuciones posteriores de las $m(m\text{-}1)/2$ fracciones de recombinación distintas (en concreto más del 75 % de ellas superan dicha cota de salto).

Una vez finalizada la simulación, se dispone de una cadena de fracciones de recombinación $\{r^{(t)}\}_{t=1}^{nsim}$, con $r^{(t)} = \{r_{i,j}^{(t)}\}_{i=1,\ldots,m-1;\ j=i+1,\ldots,m}$, que provienen de las distribuciones posteriores marginales de cada una de las fracciones de recombinación, $\pi(r_{i,j}|R_{i,j})$. Por lo tanto, cada $r^{(t)}$ almacena las m(m-1)/2 simulaciones de las fracciones de recombinación correspondientes a todas las parejas de marcadores posibles.

En cada iteración, t, con la información que proporciona $r^{(t)}$, se lleva a cabo un algoritmo de ordenación de marcadores para estimar el mapa genético, basado en las distancias mínimas entre marcadores, *SARF* (Buetow y Chakravarti 1987 [15]; Buetow 1987 [14]). Dicha ordenación de marcadores, o mapa, será una de las m! posibles permutaciones, $o(1, ..., m)$, de los m marcadores. Denotaremos a continuación a la ordenación de marcadores obtenida en la iteración t, como $O^{(t)} = \{o_1^{(t)}, ..., o_m^{(t)}\}$, donde $o_i^{(t)}$ representa al marcador que ocupa la i-ésima posición en la ordenación $O^{(t)}$ de los m marcadores.

El algoritmo de ordenación de marcadores está basado en los siguientes pasos:

1. Se seleccionan los dos marcadores más próximos en términos de su fracción de recombinación, que dan comienzo al "*mapa*":

$$mapa = \{(i^*, j^*) \in \{1, ..., m\} / \; r_{i^*,j^*} = \min_{i,j}\{r_{i,j}^{(t)}\}\}$$

Llamaremos *m.mapa* al número de marcadores que contiene el *mapa* en cada momento. En este paso *m.mapa*= 2.

2. Para cada uno de los marcadores candidatos a entrar en el *mapa*, se consideran los diferentes órdenes que se obtienen al ubicarlo en todas las posibles posiciones respecto a los marcadores que ya están fijados en el *mapa*. Es decir, se consideran las (*m.mapa*+1)! permutaciones:

$$o(l, mapa), \; \forall l \in \{1, ..., m\} - \{mapa\}.$$

3. Dada la k-ésima de estas permutaciones de marcadores, $\{o_{k1}^{(t)}, ...,$ $o_{k(m.mapa+1)}^{(t)}\}$, en la que el marcador l ocupa la posición i-ésima (es decir, $o_{ki}^{(t)} = l$), se calcula la suma de las fracciones de recombinación entre los marcadores adyacentes,

$$SARF.l_k = \sum_{j=1}^{m.mapa} r_{o_{kj}^{(t)}, o_{k(j+1)}^{(t)}}, \; \forall k \in \{1, ..., (m.mapa + 1)!\}.$$

4. La permutación asociada al mínimo de los *SARF* anteriores se considera el nuevo *mapa*.

Los pasos del 2 al 4 se repiten hasta la total incorporación de todos los marcadores en el *mapa* (en ese caso, *m.mapa*=*m*).

Este algoritmo se aplica a cada una de las simulaciones $r^{(t)}$. Nótese que en cada iteración, t, los marcadores que inician la ordenación pueden ser diferentes, ya que en cada iteración, las estimaciones de las fracciones de recombinación entre marcadores varían y en consecuencia, el orden de entrada de los marcadores en el mapa también puede variar.

La finalización del algoritmo de ordenación da lugar a una cadena de mapas, $\{O^{(t)}\}_{t=1}^{nsim}$. Para estimar la verdadera ordenación de los marcadores en el mapa genético de la población, basta con identificar los diferentes mapas resultantes de la cadena y contar cuántas veces se repite cada uno de ellos. El mapa que aparece más veces en la cadena, se considera la ordenación más probable del mapa genético de la población o modelo más probable y su probabilidad asociada se calcula como el número de veces que aparece en la cadena dividido por el número total de simulaciones, *nsim*. Los siguientes modelos obtenidos tendrán asociadas probabilidades menores. Cabe señalar que, durante la identificación de los diferentes mapas, por ejemplo, los órdenes {1, 2, 3,..., m-1, m} y {m, m-1,..., 3, 2, 1} se consideran iguales porque, en ambos resultados, los marcadores contiguos son los mismos.

Por otra parte, también en base a la cadena de mapas $\{O^{(t)}\}_{t=1}^{nsim}$ y de forma análoga a como se estima la probabilidad de los distintos modelos, se puede estimar la probabilidad con la que cada marcador se ha ubicado en su posición correcta, contando el número de iteraciones en la que ha ocurrido este hecho y dividiendo por el número total de iteraciones, *nsim*.

Para finalizar, y haciendo extensible la nomenclatura utilizada en este apartado, se convierte la distribución posterior de las fracciones de recombinación, $\{r^{(t)}\}_{t=1}^{nsim}$, en una distribución posterior de distancias, $\{D^{(t)}\}_{t=1}^{nsim}$, con $D^{(t)} = \{D_{i,j}^{(t)}\}_{i=1,\ldots,m-1;\ j=i+1,\ldots,m}$, obtenidas con la función de mapeo Haldane. Es decir, para cada fracción de recombinación $r_{i,j}^{(t)}$, $D_{i,j}^{(t)} = F(r_{i,j}^{(t)})$. Véase el Cuadro 1.6. Este procedimiento es fácilmente extensible a otras funciones, como Kosambi. A partir de estas distancias, se obtiene la distribución posterior de las distancias multipunto, $\{d^{(t)}\}_{t=1}^{nsim}$, con $d^{(t)} = \{d_{i,j}^{(t)}\}_{i=1,\ldots,m-1;\ j=i+1,\ldots,m}$, utilizando el método de mínimos cuadrados, cuyas ecuaciones se definen según la expresión (1.24). Nótese que, en cada iteración t, la conversión de fracciones de recombinación, $r^{(t)}$, a distancias multipunto, $d^{(t)}$, se lleva a cabo, dado un determinado orden de los marcadores, $O^{(t)}$. La media de la cadena de distancias multipunto, obtenidas a partir de las simulaciones de la distribución posterior para las fracciones de recombinación, sirve como estimación de las

distancias reales entre cada pareja de marcadores y el error asociado, como la desviación típica de dichas distancias multipunto.

2.2. Resultados

Para ensayar la metodología se ha diseñado un mapa con una estructura de m= 20 marcadores (mapa genético menos denso en la Figura 1.10) cuya ordenación y distancias contiguas se incluyen en el Cuadro 2.2.

Para ilustrar la aplicación del método de estimación del mapa genético propuesto anteriormente, y el efecto del tamaño de muestra sobre dicha estimación, consideramos tres situaciones:

1. Tamaño muestral n= 200 individuos simulados de la población.
2. Tamaño muestral n= 100 individuos simulados de la población.
3. Tamaño muestral n= 50 individuos simulados de la población.

Se ha simulado utilizando el entorno de programación para análisis estadístico R [62]. La longitud de las cadenas de simulación utilizadas para implementar el método propuesto es de $nsim$= 3000. La convergencia de las cadenas de simulación se ha probado utilizando el paquete coda de R [60].

A continuación, se representan 3 figuras (Figura 2.1, Figura 2.2, y Figura 2.3), una por cada tamaño de muestra. En cada gráfica se puede ver el verdadero mapa genético de la población junto con los 5 modelos más probables estimados. En cada modelo, se señalan en color rojo los marcadores que no han sido localizados en su posición correcta así como la probabilidad que cuantifica la proporción de veces que ese modelo ha aparecido en la cadena de simulación y que representa su probabilidad estimada. Por ejemplo, en la Figura 2.1, en el modelo 2 se observa una permutación entre los marcadores M9 y M10. Este modelo se ha observado en una proporción de 0.1540.

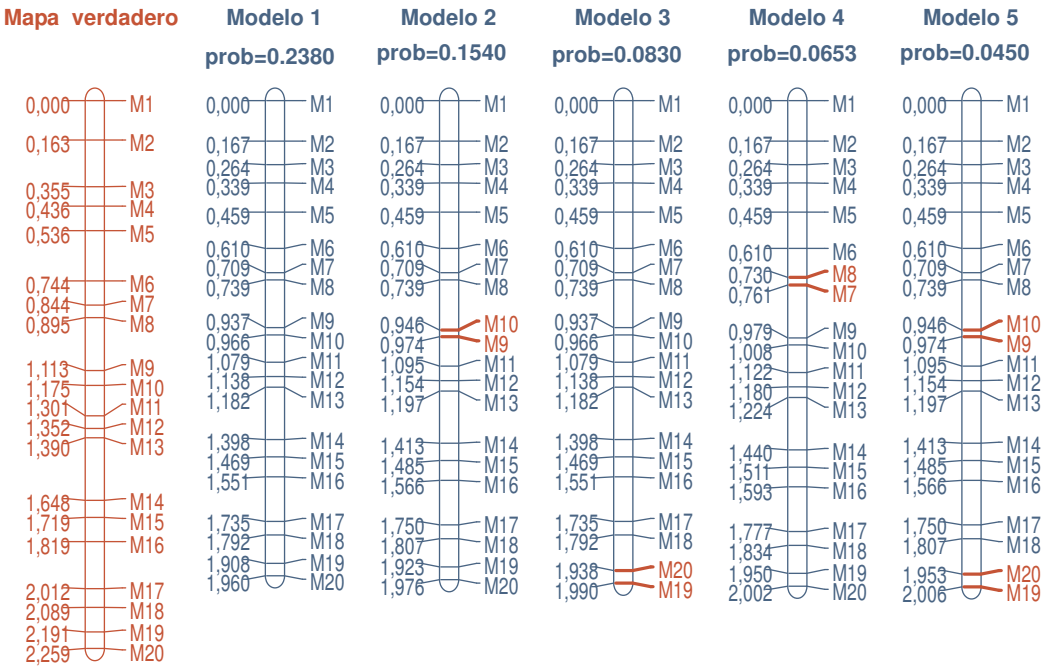


Figura 2.1: *Mapa verdadero de una población Retrocruce junto con la estima de los 5 modelos multipunto más probables de una muestra con 200 individuos.*

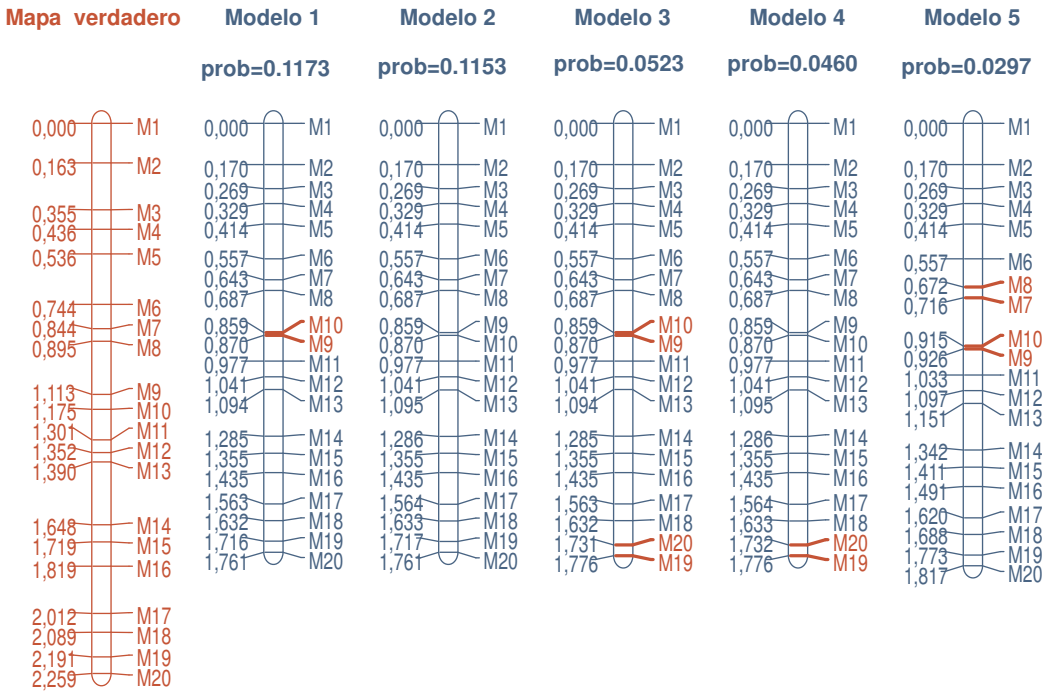


Figura 2.2: *Mapa verdadero de una población Retrocruce junto con la estima de los 5 modelos multipunto más probables de una muestra con 100 individuos.*

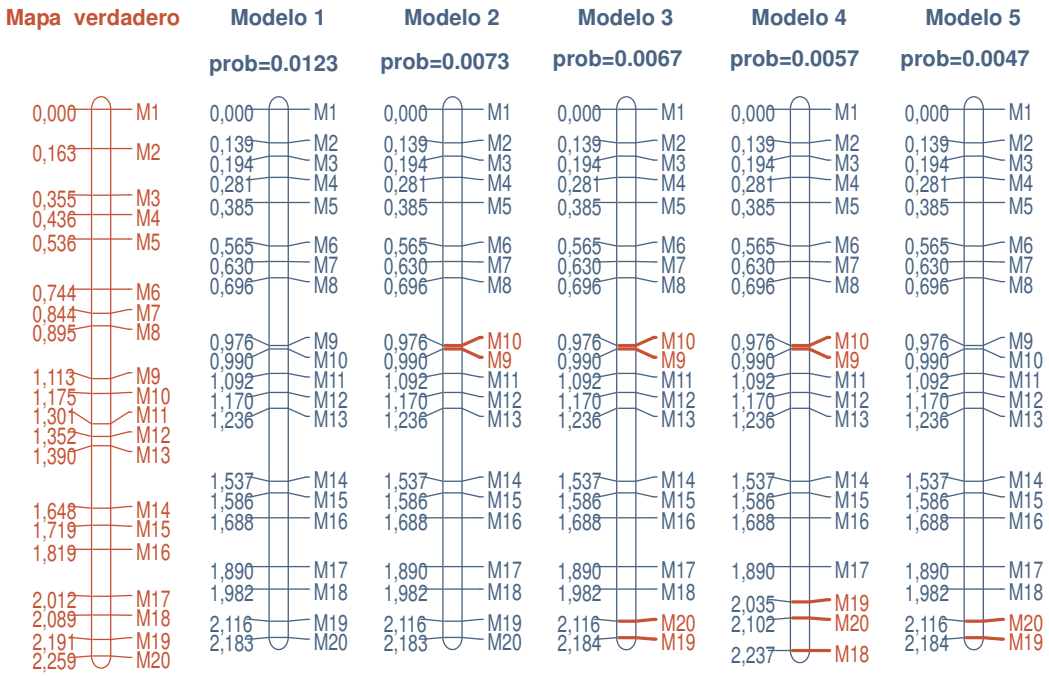


Figura 2.3: *Mapa verdadero de una población Retrocruce junto con la estima de los 5 modelos multipunto más probables de una muestra con 50 individuos.*

En el Cuadro 2.2 se resumen las distancias verdaderas entre los marcadores contiguos del mapa genético, junto con la media posterior de la distribución multipunto para el modelo bayesiano más probable representado en la Figura 2.1.

	distancias reales	modelo bayesiano
		media ± desv. típ.
$d_{1,2}$	0.162756	0.1665 ± 0.0320
$d_{2,3}$	0.191711	0.0978 ± 0.0184
$d_{3,4}$	0.081566	0.0749 ± 0.0156
$d_{4,5}$	0.100000	0.1202 ± 0.0216
$d_{5,6}$	0.207871	0.1509 ± 0.0275
$d_{6,7}$	0.100000	0.0984 ± 0.0250
$d_{7,8}$	0.050557	0.0306 ± 0.0221
$d_{8,9}$	0.218596	0.1979 ± 0.0743
$d_{9,10}$	0.061823	0.0285 ± 0.0073
$d_{10,11}$	0.125967	0.1137 ± 0.0514
$d_{11,12}$	0.051199	0.0587 ± 0.0370
$d_{12,13}$	0.037476	0.0438 ± 0.0268
$d_{13,14}$	0.257857	0.2159 ± 0.0657
$d_{14,15}$	0.071767	0.0713 ± 0.0262
$d_{15,16}$	0.100000	0.0815 ± 0.0190
$d_{16,17}$	0.192264	0.1839 ± 0.0393
$d_{17,18}$	0.076754	0.0574 ± 0.0167
$d_{18,19}$	0.102130	0.1156 ± 0.0283
$d_{19,20}$	0.068696	0.0528 ± 0.0133

Cuadro 2.2: *Distancias reales entre marcadores contiguos en Morgans y distancias multipunto del modelo bayesiano más probable con n= 200 individuos.*

En el Apéndice B, aparecen con más detalle las medias posteriores y las desviaciones típicas posteriores para cada pareja de marcadores, obtenidas según las simulaciones para cada tamaño muestral (Cuadros del B.1 al B.6)

Otro resultado de interés es el cálculo de la probabilidad con la que cada marcador ocupa todas las posibles posiciones en el mapa. Este resultado se muestra de forma gráfica en las Figuras 2.4, 2.5 y 2.6, que se corresponden con los tamaños muestrales n=200, 100 y 50, respectivamente.

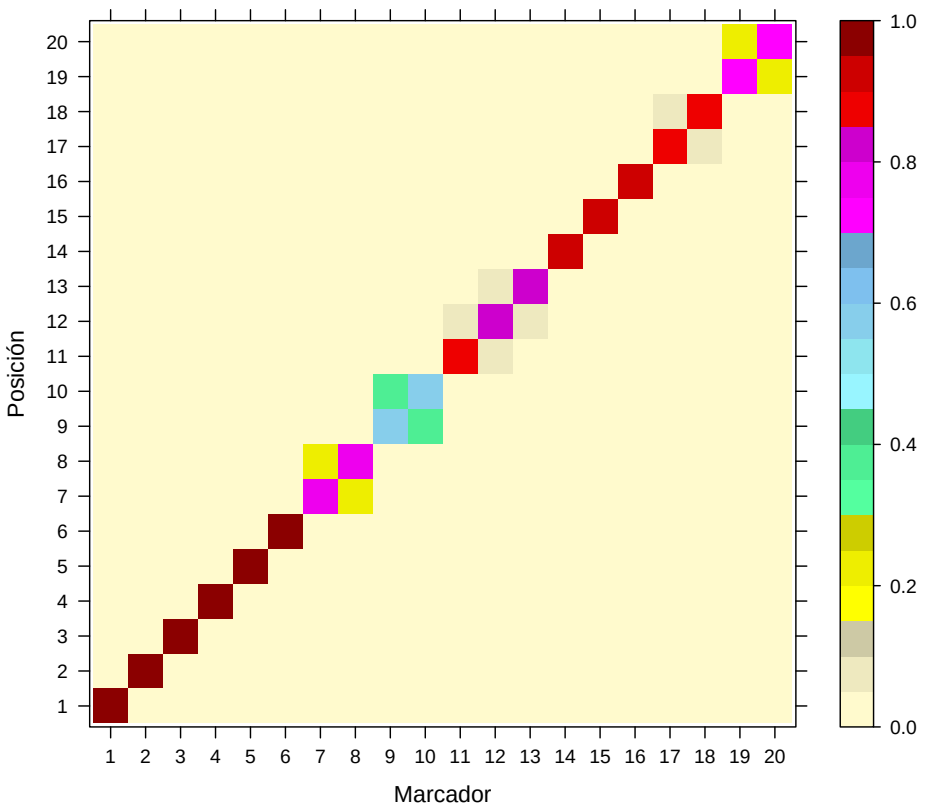


Figura 2.4: *Probabilidad con la que cada marcador aparece en cada una de las posibles posiciones del mapa genético según la simulación con 200 individuos.*

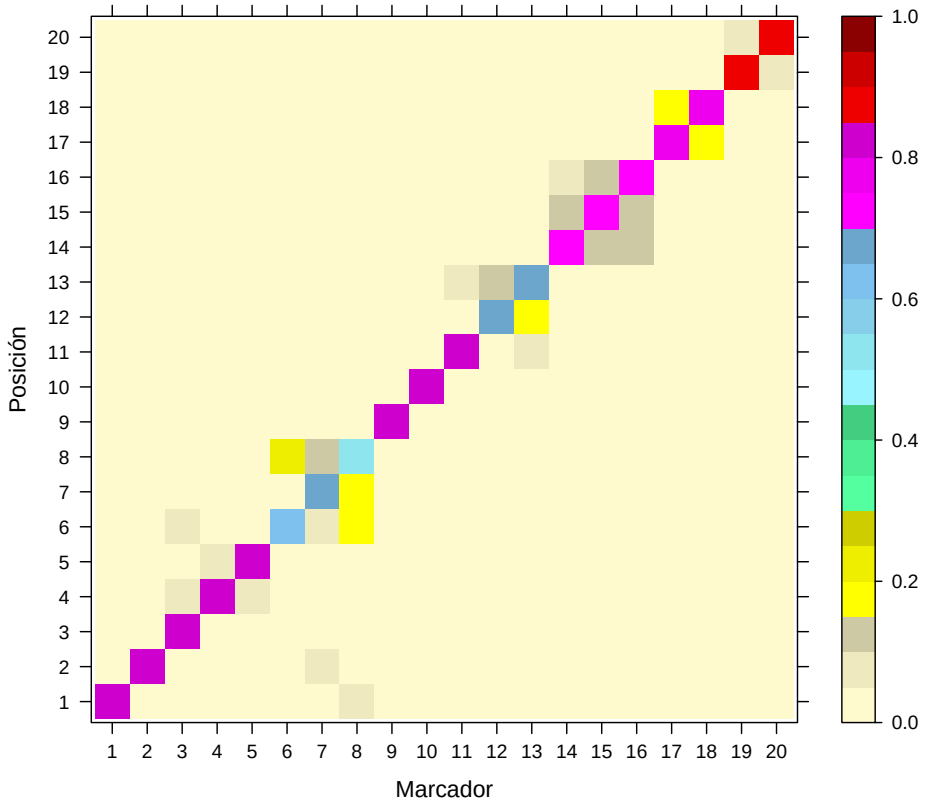


Figura 2.5: *Probabilidad con la que cada marcador aparece en cada una de las posibles posiciones del mapa genético según la simulación con 100 individuos.*

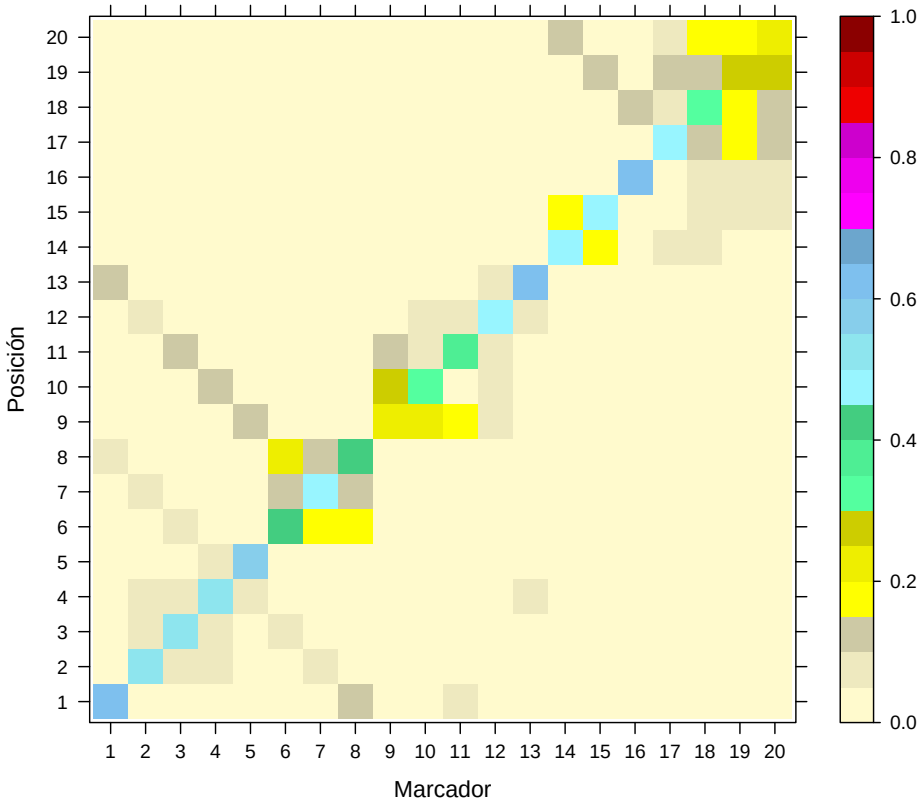


Figura 2.6: *Probabilidad con la que cada marcador aparece en cada una de las posibles posiciones del mapa genético según la simulación con 50 individuos.*

2.3. Discusión

A la vista de los resultados, se concluye que la metodología empleada parece estimar correctamente el orden de los marcadores del mapa genético de la población, incluso cuando el número de datos disponibles es relativamente pequeño (n=50). Cabe señalar que en la muestra de 100 individuos el modelo correcto aparece en segunda posición, pero con una probabilidad muy cercana a la del primer modelo (0.1173 vs 0.1153). Ambos se diferencian en una permutación de una pareja de marcadores estrechamente ligados, por lo que es comprensible que tengan probabilidades parecidas.

De la evaluación de los resultados, se deduce que la pérdida de información que supone la reducción del tamaño muestral tiene distintas consecuencias:

1. Por una parte, implica el aumento de modelos alternativos que repercute en la reducción de la probabilidad del modelo más probable. En los ejemplos particulares que se muestran, la probabilidad del modelo más probable estimada con 200 individuos es de 0.2380. Esta probabilidad se reduce a 0.1173 con 100 individuos y a 0.0123 con 50 individuos. Este resultado es lógico: un mayor tamaño de muestra produce conclusiones con mayor certidumbre, que en términos de mapas genéticos se corresponde con mayor probabilidad para el mapa "ganador.º más factible.

2. Por otra parte, en las Figuras 2.4, 2.5 y 2.6, se observa que la metodología aplicada sobre 200 individuos es capaz de ubicar cada marcador, de forma casi inequívoca, en su posición correcta obteniendo probabilidades próximas a cero para posiciones alternativas. La reducción del tamaño muestral, hace que los marcadores más cercanos en la realidad, aumenten las probabilidades de ser ubicados en posiciones alternativas relativamente próximas. De nuevo, mayor tamaño muestral provoca mayor eficiencia al colocar a cada marcador en el lugar del mapa que le corresponde.

3.También se detecta un efecto en la estimación de la longitud total del mapa genético. Con 200 individuos, se obtienen modelos más cortos que el verdadero mapa genético. Esta condensación se asevera para 100 individuos. Sin embargo, con 50 individuos los modelos se vuelven a alargar, manteniéndose, aun así, más cortos que el verdadero mapa genético. Quizás, este comportamiento se puede atribuir tanto al algoritmo de ordenación basado en distancias mínimas ($SARF$) como al sesgo que se comete al estimar las fracciones de recombinación entre parejas de marcadores en una población Retrocruce. Hühn

y Piepho (2008) [26], en su estudio sobre el sesgo de las de las fracciones de recombinación bajo la función de mapeo Karlin, concluyeron que en poblaciones Retrocruce, es siempre positivo, aunque disminuye según aumenta el tamaño muestral. También demostraron, a través de una comparativa entre las funciones de mapeo Karlin, Haldane y Kosambi, que el comportamiento de los sesgos medianos son similares, por lo que cabe esperar que el resultado con la función de mapeo Haldane no difiera mucho del resultado con la función de mapeo Karlin.

Finalmente, como se puede ver en el Cuadro 2.2, los marcadores contiguos más próximos en la realidad son: {M7, M8}, {M9, M10}, {M11, M12, M13} y {M19, M20}. Por otra parte, en general, los marcadores cuyas distancias multipunto se han estimado con una mayor variabilidad, en términos de la desviación típica posterior son: {M1, M2}, {M5, M6}, {M8, M9} y {M13, M14}. Precisamente, se puede comprobar que casi todos los modelos alternativos al correcto, independientemente del tamaño muestral, involucran a una o varias de estas parejas de marcadores.

Capítulo 3

Diseño F_2

Este capítulo se dedica al estudio del cruce controlado F_2 que consiste en la autofecundación de individuos F_1 (Figura 1.9). La extensión a generaciones avanzadas de autofecundación, hasta llegar a RILs (Líneas Recombinantes Puras), se basa en el cálculo generalizado de las frecuencias genotípicas esperadas (Martin y Hospital 2006 [52]), similares a las presentadas en los Cuadros 1.2 y 1.4 de la Introducción. En las siguientes secciones se estudiará principalmente el caso de marcadores codominantes y se pondrán de manifiesto los posibles problemas de la aplicación del método a marcadores dominantes.

3.1. Diseño F_2 con marcadores codominantes.

3.1.1. Metodología

Al igual que en el capítulo anterior, se parte de un mapa genético concreto correspondiente a una población F_2 con una estructura de m marcadores y de él se extrae una muestra de n individuos independientes. Para cada marcador se asume observable la condición de ser homocigoto (AA o aa) o heterocigoto (Aa). Por lo tanto, en este diseño se codifica con un 0 si el individuo k en el marcador j es homocigoto (aa), con un 1 si es heterocigoto (Aa) y con un 2 si es homocigoto (AA). Es decir,

$$y_{kj} = \begin{cases} 0, \ si \ aa \\ 1, \ si \ Aa \\ 2, \ si \ AA \end{cases} \tag{3.1}$$

De este modo, el banco de datos tiene una estructura como la que se muestra en el Cuadro 3.1.

	Marcadores						
Individuos	1	...	i	...	j	...	m
1	0	...	2	...	1	...	1
.	.	...	.	...	.	...	.
.	.	...	.	...	.	...	.
k	1	...	1	...	2	...	1
.	.	...	.	...	.	...	.
.	.	...	.	...	.	...	.
n	2	...	0	...	1	...	0

Cuadro 3.1: *Codificación de los marcadores en un modelo F_2.*

La metodología es similar a la expuesta en el diseño Retrocruce. Sin embargo, para permitir una generalización a otro tipo de diseño de cruzamiento a partir de una F_2 (F_n), la forma de obtener las frecuencias observadas varía con respecto al diseño Retrocruce. En este caso, la combinación de dos marcadores cualesquiera da lugar al reparto de los n individuos en los 9 genotipos posibles, como se muestra en el Cuadro 3.2, y que coincide con lo ya justificado en el Capítulo 1, mediante el Cuadro 1.4, aunque en este caso utilizando una notación más acorde y específica a la metodología que se va a desarrollar. Comparando el Cuadro 1.4 con el Cuadro 3.2, $p_s = \gamma_s$ y $f_s = R_{s,i,j}, \ s \in \{1, ..., 9\}$

marcadores i / j	BB	Bb	bb
AA	$AABB$ $R_{1,i,j}$ $\gamma_1 = 0.25(1 - r_{i,j})^2$	$AABb$ $R_{2,i,j}$ $\gamma_2 = 0.5r_{i,j}(1 - r_{i,j})$	$AAbb$ $R_{3,i,j}$ $\gamma_3 = 0.25r_{i,j}^2$
Aa	$AaBB$ $R_{4,i,j}$ $\gamma_4 = 0.5r_{i,j}(1 - r_{i,j})$	$AaBb$ $R_{5,i,j}$ $\gamma_5 = 0.5(1 - 2r_{i,j} + 2r_{i,j}^2)$	$Aabb$ $R_{6,i,j}$ $\gamma_6 = 0.5r_{i,j}(1 - r_{i,j})$
aa	$aaBB$ $R_{7,i,j}$ $\gamma_7 = 0.25r_{i,j}^2$	$aaBb$ $R_{8,i,j}$ $\gamma_8 = 0.5r_{i,j}(1 - r_{i,j})$	$aabb$ $R_{9,i,j}$ $\gamma_9 = 0.25(1 - r_{i,j})^2$

Cuadro 3.2: *Genotipos, frecuencias observadas y esperadas para dos marcadores de un cruce F_2, sin distorsión en la segregación.*

En cada celda se muestra tanto el genotipo como la frecuencia observada ($R_{s,i,j}$) y frecuencia relativa esperada (γ_s) de cada genotipo. De nuevo, $r_{i,j}$ es la fracción de recombinación entre los marcadores i y j. Nótese que $\sum_{s=1}^{9} R_{s,i,j} = n$, tamaño muestral.

Para la generalización a diseños F_n, la configuración genotípica de los marcadores dos a dos sería la misma que en F_2 (9 genotipos posibles). Sin embargo, las frecuencias esperadas serían diferentes.

Como el vector $\mathbf{R_{i,j}} = (R_{1,i,j}, ..., R_{9,i,j})$ sigue una distribución *Multinomial*:

$$\mathbf{R_{i,j}} \sim Multinomial(n; \gamma), \ con \ \gamma = (\gamma_1, ..., \gamma_9)$$

La función de probabilidad de los datos, dados los parámetros n y $r_{i,j}$ se expresa como:

$$
\begin{aligned}
f(\mathbf{R_{i,j}}|r_{i,j}) &= \tfrac{n!}{\prod_{s=1}^{9} R_{s,i,j}!} \prod_{s=1}^{9} (\gamma_s)^{R_{s,i,j}} \\
&\propto [(1-r_{i,j})^2]^{R_{1,i,j}} [r_{i,j}(1-r_{i,j})]^{R_{2,i,j}} \cdot \\
&\cdot [r_{i,j}^2]^{R_{3,i,j}} [r_{i,j}(1-r_{i,j})]^{R_{4,i,j}} \cdot \\
&\cdot [(1-2r_{i,j}+2r_{i,j}^2)]^{R_{5,i,j}} [r_{i,j}(1-r_{i,j})]^{R_{6,i,j}} \cdot \\
&\cdot [r_{i,j}^2]^{R_{7,i,j}} [r_{i,j}(1-r_{i,j})]^{R_{8,i,j}} [(1-r_{i,j})^2]^{R_{9,i,j}}
\end{aligned}
\tag{3.2}
$$

Dadas las restricciones sobre $r_{i,j}$ que ya destacamos en el capítulo anterior, $r_{i,j} \in [0, 0.5]$, asumimos a priori una distribución previa plana (poco informativa) en dicho intervalo, en forma de *Beta* truncada. Dicha distribución no conjuga perfectamente con el modelo de los datos, pero nos permitirá llevar a cabo un análisis posterior basado en simulación. Es decir,

$$
\pi(r_{i,j}) = BetaT(r_{i,j}|\alpha, \beta; \varepsilon), \ con\ \alpha = \beta = 1 \tag{3.3}
$$

Asumiendo independencia entre todos los vectores de frecuencias observadas $\mathbf{R_{1,2}}$, $\mathbf{R_{1,3}}$, ..., $\mathbf{R_{m\text{-}1,m}}$, e independencia a priori entre todas las fracciones de recombinación de pares de marcadores distintos $\boldsymbol{r}^* = \{r_{i,j} : \ i = 1, ..., m-1; j = i+1, ..., m\}$, se verifica que las distribuciones posteriores resultan independientes entre sí. En definitiva, la distribución posterior para $\mathbf{r}^*$ se obtiene a partir del producto de la distribución posterior de cada $r_{i,j}$ dado $\mathbf{R_{i,j}}$. Es decir:

$$
\pi(\mathbf{r}^*|\mathbf{R}^*) = \prod_{i=1}^{m-1} \prod_{j=i+1}^{m} \pi(r_{i,j}|\mathbf{R_{i,j}}) \ \ \text{con}
$$

$$\begin{aligned}\pi(r_{i,j}|\mathbf{R_{i,j}}) &\propto f(\mathbf{R_{i,j}}|r_{i,j})\pi(r_{i,j}) \\ &\propto r_{i,j}^{R_{2,i,j}+R_{4,i,j}+R_{6,i,j}+R_{8,i,j}+2(R_{3,i,j}+R_{7,i,j})+\alpha-1}. \\ &\cdot(1-r_{i,j})^{2(R_{1,i,j}+R_{9,i,j})+R_{2,i,j}+R_{4,i,j}+R_{6,i,j}+R_{8,i,j}} \ . \\ &\cdot\ (1-2r_{i,j}+2r_{i,j}^2)^{R_{5,i,j}}(\varepsilon-r_{i,j})^{\beta-1}I_{[0,\varepsilon]}(r_{i,j})\end{aligned} \tag{3.4}$$

donde $I_{[0,\varepsilon]}(\cdot)$ es la función indicatriz en el intervalo [0, ε].

Llamando a $= R_{2,i,j} + R_{4,i,j} + R_{6,i,j} + R_{8,i,j} + 2(R_{3,i,j} + R_{7,i,j}) + \alpha - 1$ y b $= 2(R_{1,i,j} + R_{9,i,j}) + R_{2,i,j} + R_{4,i,j} + R_{6,i,j} + R_{8,i,j}$ se obtiene la **distribución posterior**:

$$\pi(r_{i,j}|\mathbf{R_{i,j}}) \propto r_{i,j}^{\mathrm{a}}(1-r_{i,j})^{\mathrm{b}}(1-2r_{i,j}+2r_{i,j}^2)^{R_{i,j}^5}(\varepsilon-r_{i,j})^{\beta-1}I_{[0,\varepsilon]}(r_{i,j}) \tag{3.5}$$

Para simular de $\pi(r_{i,j}|\mathbf{R_{i,j}})$ a través del algoritmo Metropolis-Hastings, utilizamos como distribución propuesta (proposal) para obtener candidatos, $r_{i,j}$, una distribución *Normal* truncada en (0, 0.5):

$$p(r_{i,j}) = NormalT(\hat{\mu}, \hat{\sigma}^2; 0, 0.5) \tag{3.6}$$

Esta elección está justificada porque, si se considera la densidad posterior evitando el factor de truncamiento, $(\varepsilon - r_{i,j})^{\beta-1}$:

$$g(r_{i,j}) = r_{i,j}^{\mathrm{a}}(1-r_{i,j})^{\mathrm{b}}(1-2r_{i,j}+2r_{i,j}^2)^{R_{5,i,j}} \tag{3.7}$$

esta densidad se puede aproximar asintóticamente por una distribución $Normal(\hat{\mu}, \hat{\sigma}^2)$, con media, $\hat{\mu}$, la moda de dicha densidad y con varianza, $\hat{\sigma}^2$, la inversa cambiada de signo del Hessiano (segunda derivada) de la logposterior evaluada en la moda, según Tanner (1996) [81]. Estos parámetros son los que se utilizan en (3.6). Para mejorar la eficiencia computacional, para el cálculo de la moda, se utiliza el logaritmo de $g(r_{i,j})$:

$$log(g(r_{i,j})) = \text{a } log(r_{i,j}) + \text{b } log(1 - r_{i,j}) + R_{5,i,j}\ log(1 - 2r_{i,j} + 2r_{i,j}^2) \quad (3.8)$$

El algoritmo de búsqueda del máximo es el propuesto por Byrd et. al. 1995 [16], e implementado en la función *optim* de R.

La probabilidad de salto para aceptar un candidato, $r_{i,j}$, en la cadena, en el paso t+1, habiendo aceptado previamente $r_{i,j}^{(t)}$ en la iteración t, vendrá dada por:

$$\begin{aligned} prob.salto(r_{i,j}^{(t)}, r_{i,j}) &= min\Big\{1, \tfrac{\pi(r_{i,j}|\mathbf{R_{i,j}})}{p(r_{i,j})} \tfrac{p(r_{i,j}^{(t)})}{\pi(r_{i,j}^{(t)}|\mathbf{R_{i,j}})}\Big\} \\ &= min\Big\{1, q(r_{i,j})/q(r_{i,j}^{(t)})\Big\}, \end{aligned} \quad (3.9)$$

con

$$q(r) = r^{\text{a}}(1 - r)^{\text{b}}(\varepsilon - r)^{\beta-1}(1 - 2r + 2r^2)^{R_{5,i,j}} exp((r - \hat{\mu})^2/2\hat{\sigma}^2) \quad (3.10)$$

Para hacer más eficaz la computación de la probabilidad de salto, se utilizaran logaritmos.

Una vez finalizada la simulación, se dispone de una cadena de fracciones de recombinación $\{r^{(t)}\}_{t=1}^{nsim}$ con $r^{(t)} = \{r_{i,j}^{(t)}\}_{i=1,\ldots,m-1;j=i+1,\ldots,m}$, que provienen de la distribución posterior de cada una de las fracciones de recombinación, $\pi(r_{i,j}|\mathbf{R_{i,j}})$. Aplicando el mismo algoritmo de ordenación que en el capítulo anterior, se obtiene una cadena de mapas, $\{O^{(t)}\}_{t=1}^{nsim}$. A continuación, siguiendo idéntico procedimiento, se obtiene la distribución posterior de las distancias multipunto entre todas las parejas de marcadores, $\{d^{(t)}\}_{t=1}^{nsim}$ con $d^{(t)} = \{d_{i,j}^{(t)}\}_{i=1,\ldots,m-1;j=i+1,\ldots,m}$. Las distintas estimaciones del mapa genético de la población (modelos), sus probabilidades asociadas y las probabilidades de ubicación de cada uno de los marcadores, en las distintas posiciones del mapa genético, se obtienen con el mismo razonamiento que en el capítulo anterior.

3.1.2. Resultados

Dado que cada vez es más habitual que los mapas genéticos estén definidos por un mayor número de marcadores estrechamente ligados, para ensayar la metodología se han diseñado dos mapas con una estructura de m= 20 marcadores, uno menos denso que otro (Figura 1.10). Es decir, en uno de los mapas todos los marcadores se concentran en algo más de 2 Morgan; sin embargo, el otro mapa tiene una longitud de menos de 1 Morgan. El diseño del primer mapa genético es idéntico al creado para ensayar la metodología en la población Retrocruce que se investigó en el capítulo anterior. La intención es comprobar y comparar el funcionamiento de la metodología en las dos situaciones. En adelante, para simplificar, se hablará de población con mapa menos denso, o no tan denso, para denominar aquella representada por el mapa genético más largo y población con mapa más denso, para referirnos a la población representada por el mapa genético más corto, dado que ambos mapas tienen el mismo número de marcadores y difieren en la distancia entre marcadores contiguos

De cada una de las poblaciones, se ha extraído una muestra de tres tamaños diferentes (n= 200, n= 100 y n= 50 individuos), para estudiar el efecto del tamaño de la muestra sobre la eficiencia de la metodología propuesta. Se ha aplicado la metodología simulando, en el entorno de programación para análisis estadístico R [62], cadenas de longitud *nsim*= 3000, tras la comprobación de su convergencia con funciones pertenecientes al paquete coda de R [60].

Los resultados se han organizado en dos secciones (una para cada población). Dentro de cada una de ellas, aparecen los resultados siguiendo la misma presentación que en el capítulo anterior. Es decir, en primer lugar las gráficas del mapa genético de la población junto con los 5 modelos bayesianos más probables, después un cuadro resumen que incluye las distancias reales entre marcadores contiguos, las distancias medias multipunto posteriores y las desviaciones típicas posteriores del modelo bayesiano más probable para la muestra con 200 individuos y por último las gráficas que representan la probabilidad con la que cada marcador ocupa cada una de las posibles posiciones en el mapa genético. En el Apéndice C, también organizado en secciones, se disponen de las medias posteriores y de las desviaciones típicas posteriores, para cada pareja de marcadores, de cada muestra.

Población F_2, con mapa menos denso, con todos los marcadores codominantes

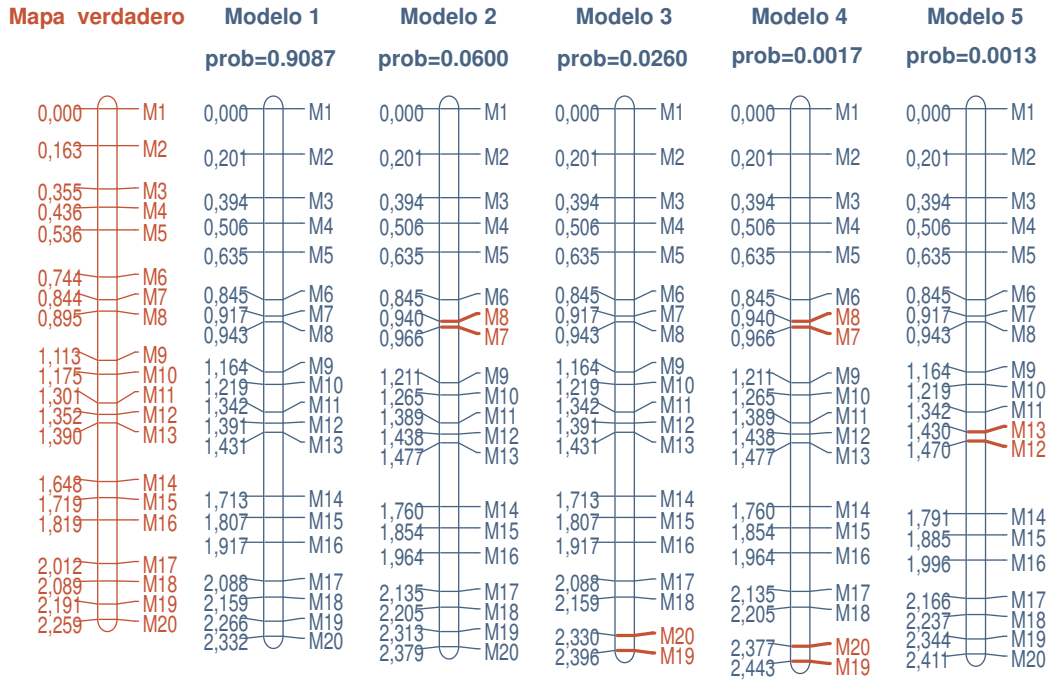


Figura 3.1: *Mapa verdadero de una población F_2, con mapa menos denso, junto con la estima de los 5 modelos multipunto más probables de una muestra con 200 individuos.*

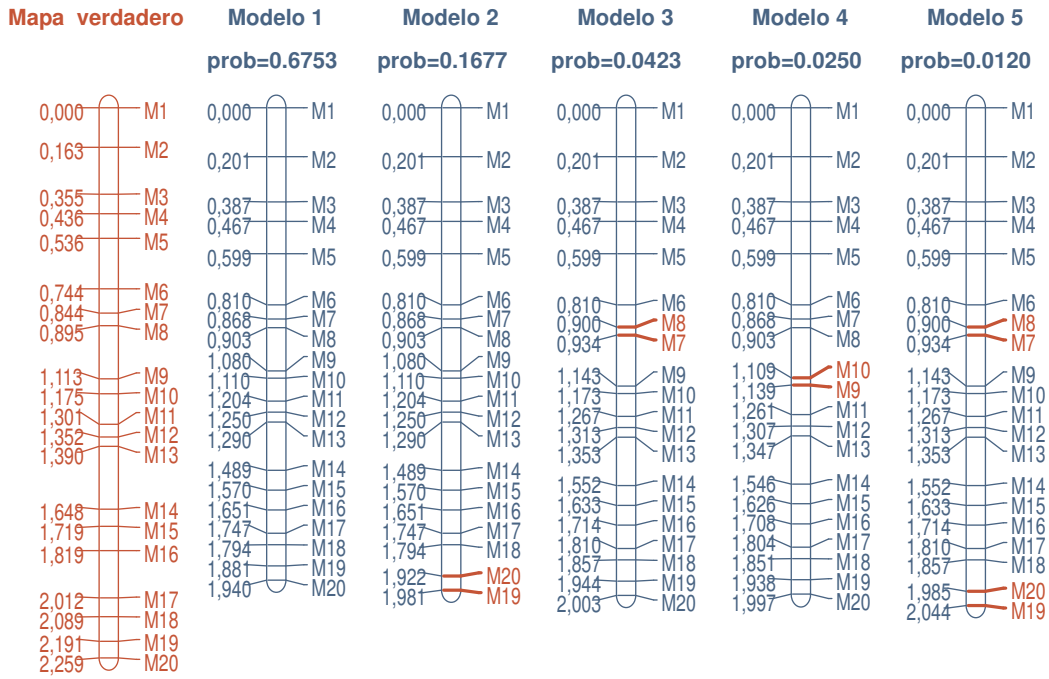


Figura 3.2: *Mapa verdadero de una población F_2, con mapa menos denso, junto con la estima de los 5 modelos multipunto más probables de una muestra con 100 individuos.*

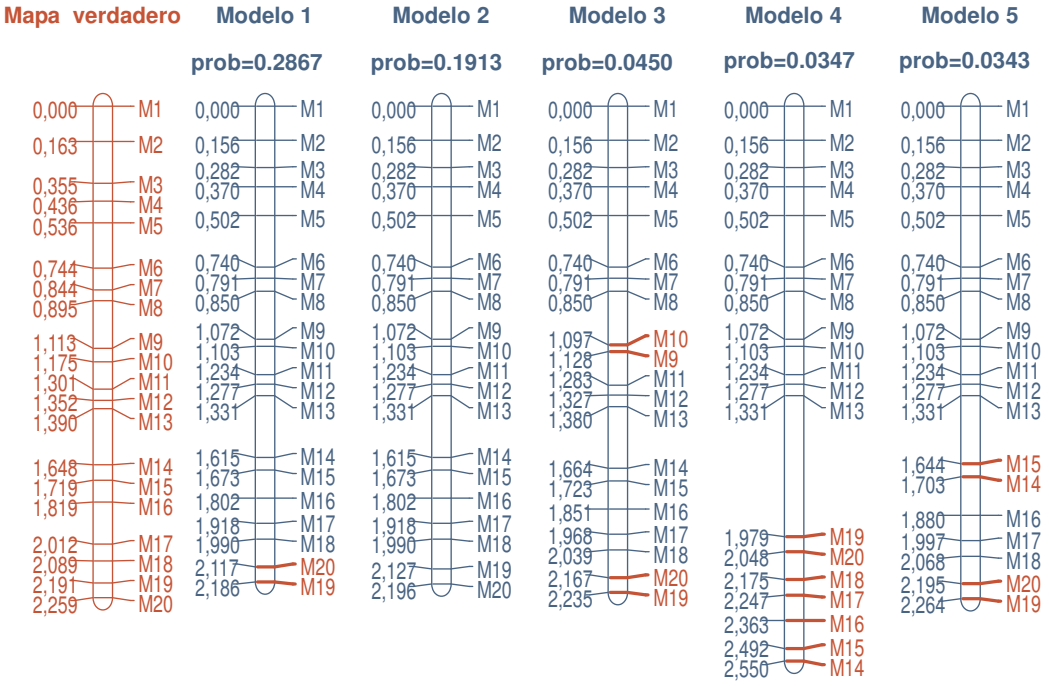


Figura 3.3: *Mapa verdadero de una población F_2, con mapa menos denso, junto con la estima de los 5 modelos multipunto más probables de una muestra con 50 individuos.*

El Cuadro 3.3 está relacionado con el modelo bayesiano multipunto más probable representado en la Figura 3.1.

	distancias reales	modelo bayesiano
		media ± desv. típ.
$d_{1,2}$	0.162756	0.2014 ± 0.0158
$d_{2,3}$	0.191711	0.1930 ± 0.0128
$d_{3,4}$	0.081566	0.1112 ± 0.0096
$d_{4,5}$	0.100000	0.1298 ± 0.0100
$d_{5,6}$	0.207871	0.2095 ± 0.0110
$d_{6,7}$	0.100000	0.0720 ± 0.0082
$d_{7,8}$	0.050557	0.0257 ± 0.0046
$d_{8,9}$	0.218596	0.2216 ± 0.0102
$d_{9,10}$	0.061823	0.0544 ± 0.0054
$d_{10,11}$	0.125967	0.1235 ± 0.0067
$d_{11,12}$	0.051199	0.0493 ± 0.0051
$d_{12,13}$	0.037476	0.0392 ± 0.0050
$d_{13,14}$	0.257857	0.2826 ± 0.0121
$d_{14,15}$	0.071767	0.0939 ± 0.0083
$d_{15,16}$	0.100000	0.1103 ± 0.0086
$d_{16,17}$	0.192264	0.1706 ± 0.0092
$d_{17,18}$	0.076754	0.0706 ± 0.0061
$d_{18,19}$	0.102130	0.1074 ± 0.0111
$d_{19,20}$	0.068696	0.0664 ± 0.0069

Cuadro 3.3: *Distancias reales entre marcadores contiguos en Morgans, distancias medias multipunto posteriores y desviaciones típicas posteriores del modelo bayesiano más probable respecto a la muestra con 200 individuos*

En las Figuras 3.4, 3.5 y 3.6 se representa la probabilidad con la que cada marcador ocupa cada una de las posibles posiciones en el mapa según los resultados obtenidos para las muestras representadas en las Figuras 3.1, 3.2 y 3.3, respectivamente.

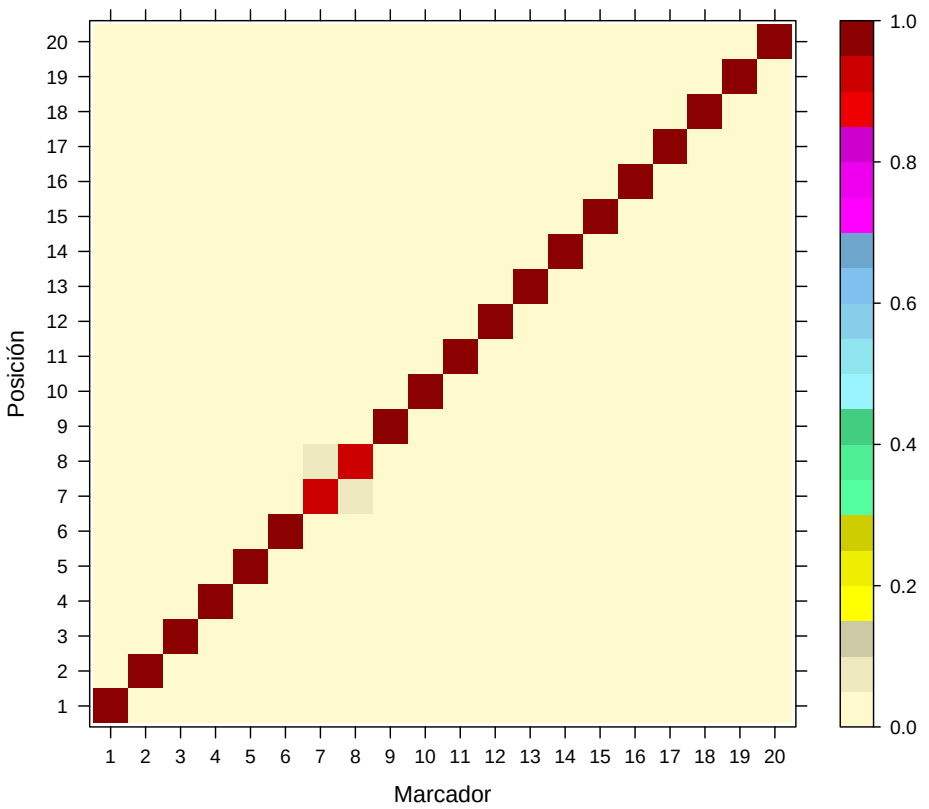


Figura 3.4: *Probabilidad con la que cada marcador aparece en cada una de las posibles posiciones del mapa genético menos denso, según la simulación de la muestra con 200 individuos.*

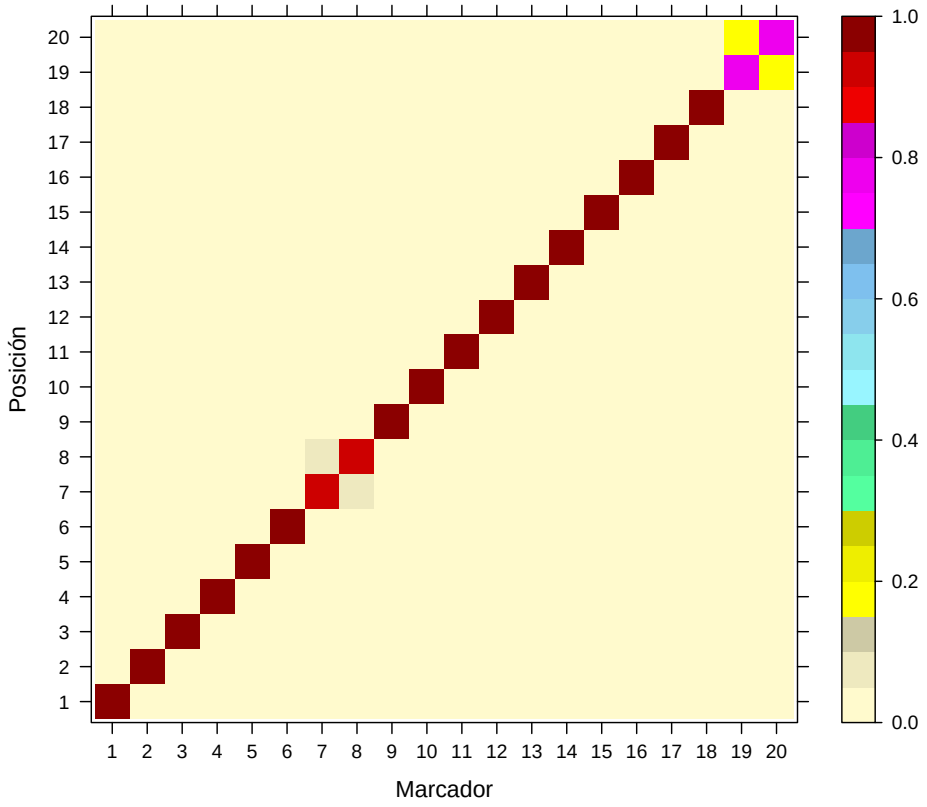


Figura 3.5: *Probabilidad con la que cada marcador aparece en cada una de las posibles posiciones del mapa genético menos denso, según la simulación de la muestra con 100 individuos.*

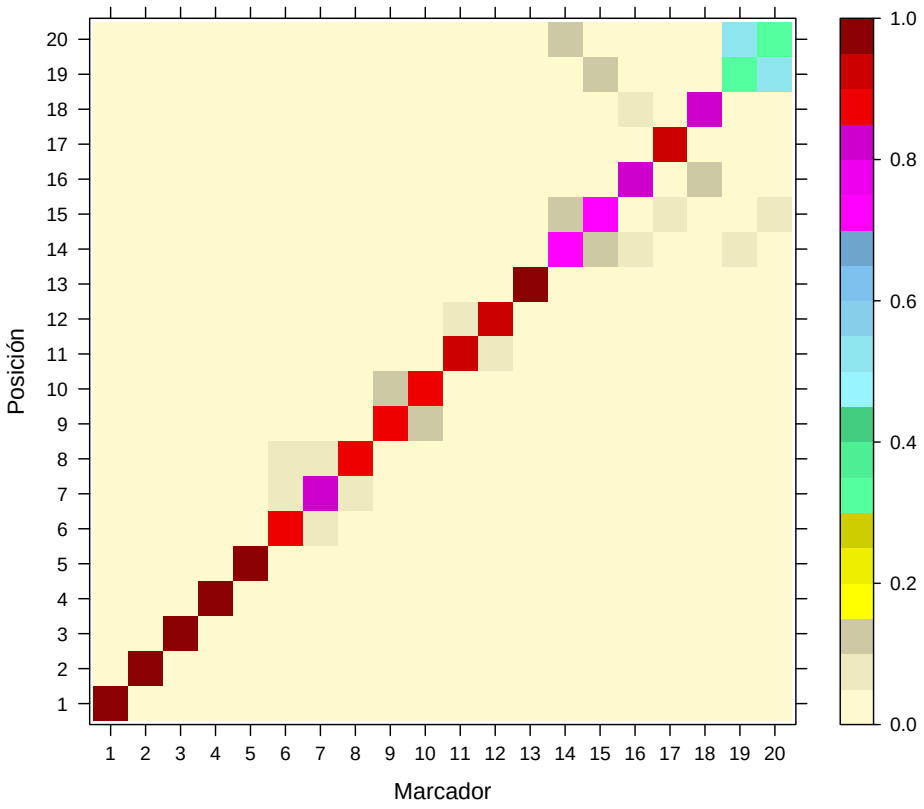


Figura 3.6: *Probabilidad con la que cada marcador aparece en cada una de las posibles posiciones del mapa genético menos denso, según la simulación de la muestra con 50 individuos.*

Población F_2, con mapa más denso, con todos los marcadores codominantes

A continuación, se presentan unos resultados similares a los anteriores pero para muestras que provienen de una población con mapa más denso.

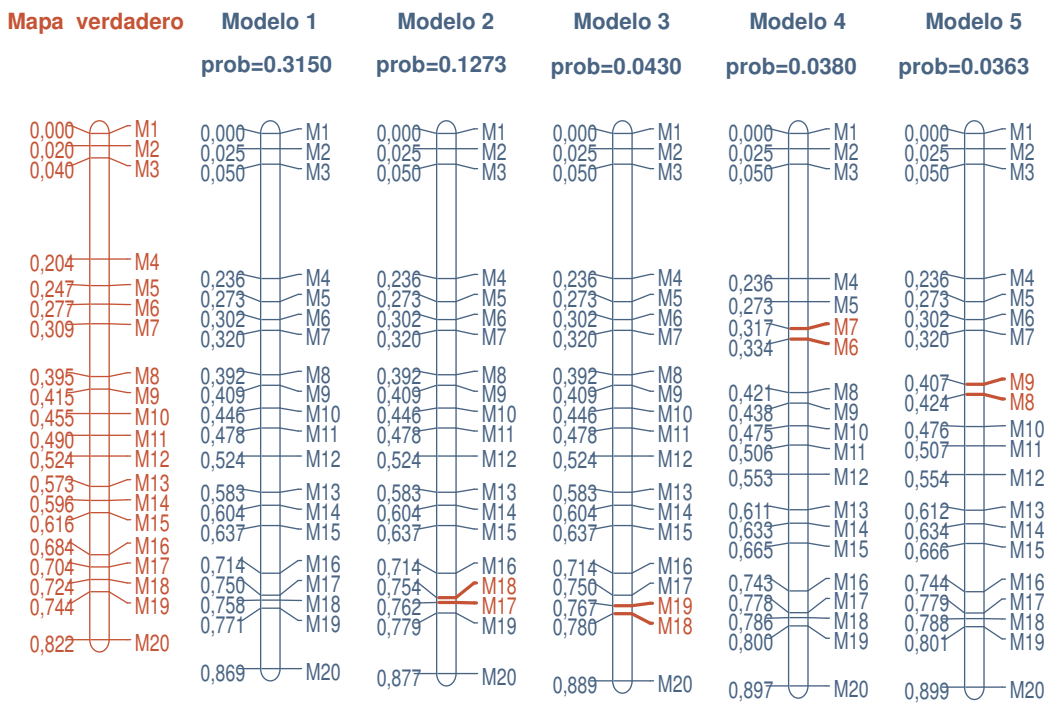


Figura 3.7: *Mapa verdadero de una población F_2, con mapa más denso, junto con la estima de los 5 modelos multipunto más probables de una muestra con 200 individuos.*

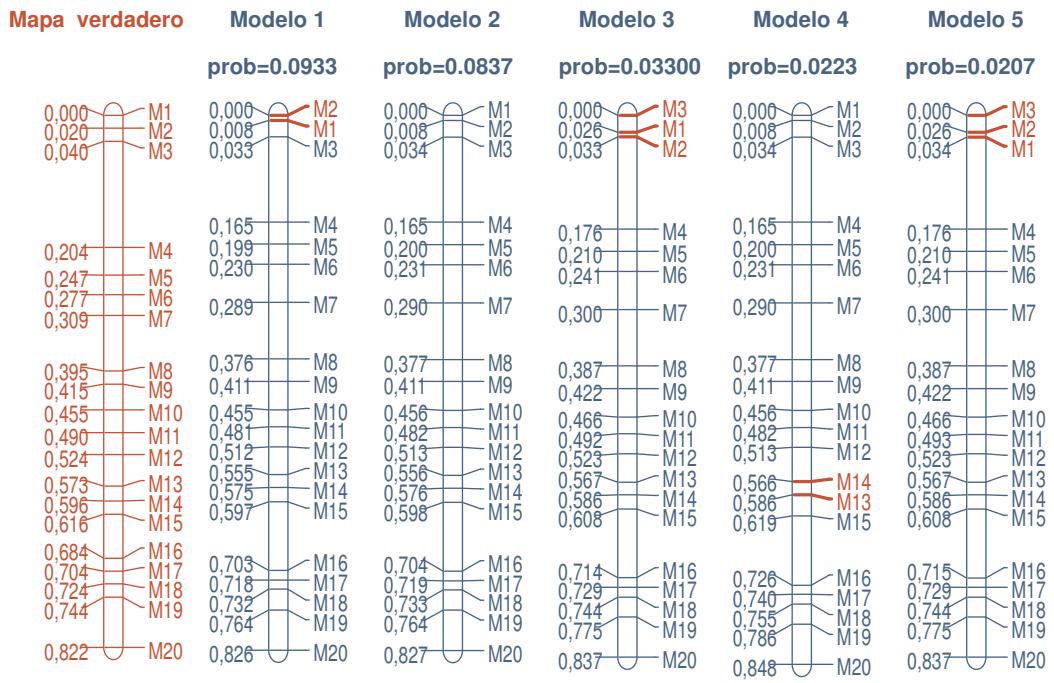


Figura 3.8: *Mapa verdadero de una población F_2, con mapa más denso, junto con la estima de los 5 modelos multipunto más probables de una muestra con 100 individuos.*

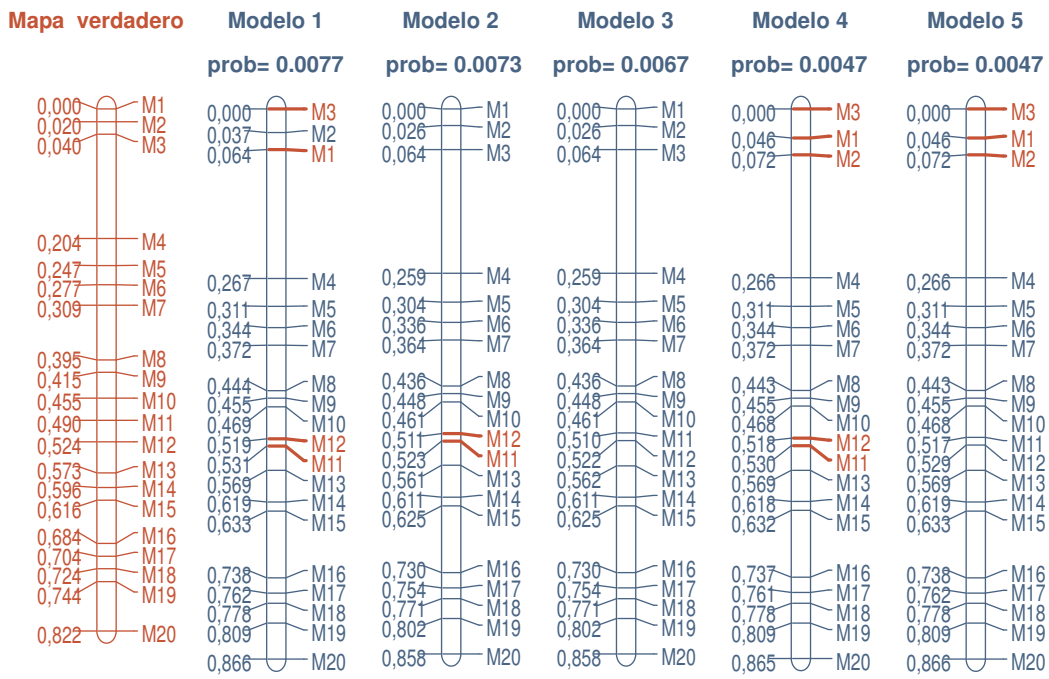


Figura 3.9: *Mapa verdadero de una población F_2, con mapa más denso, junto con la estima de los 5 modelos multipunto más probables de una muestra con 50 individuos.*

El Cuadro 3.4 se relaciona con el modelo bayesiano multipunto más probable representado en la Figura 3.7

	distancias reales	modelo bayesiano
		media ± desv. típ.
$d_{1,2}$	0.020000	0.0248 ± 0.0084
$d_{2,3}$	0.020000	0.0252 ± 0.0089
$d_{3,4}$	0.163737	0.1859 ± 0.0253
$d_{4,5}$	0.043085	0.0367 ± 0.0109
$d_{5,6}$	0.029992	0.0296 ± 0.0101
$d_{6,7}$	0.032026	0.0173 ± 0.0104
$d_{7,8}$	0.086469	0.0725 ± 0.0127
$d_{8,9}$	0.020000	0.0173 ± 0.0093
$d_{9,10}$	0.039634	0.0369 ± 0.0129
$d_{10,11}$	0.034757	0.0315 ± 0.0150
$d_{11,12}$	0.034757	0.0466 ± 0.0175
$d_{12,13}$	0.048186	0.0582 ± 0.0150
$d_{13,14}$	0.022916	0.0217 ± 0.0191
$d_{14,15}$	0.020000	0.0324 ± 0.0176
$d_{15,16}$	0.068384	0.0778 ± 0.0262
$d_{16,17}$	0.020000	0.0351 ± 0.0085
$d_{17,18}$	0.020000	0.0084 ± 0.0048
$d_{18,19}$	0.020000	0.0133 ± 0.0060
$d_{19,20}$	0.078403	0.0977 ± 0.0118

Cuadro 3.4: *Distancias reales entre marcadores contiguos en Morgans, distancias medias multipunto posteriores y desviaciones típicas posteriores del modelo bayesiano más probable respecto a la muestra con 200 individuos.*

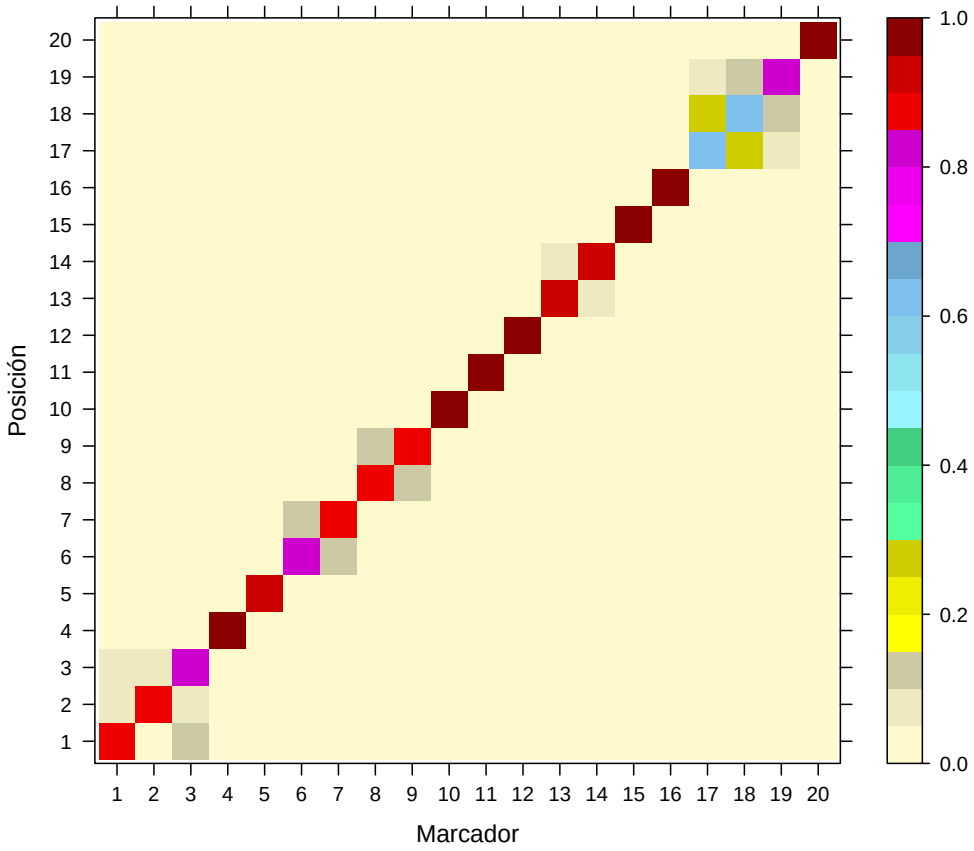


Figura 3.10: *Probabilidad con la que cada marcador aparece en cada una de las posibles posiciones del mapa genético más denso según la simulación de la muestra con 200 individuos.*

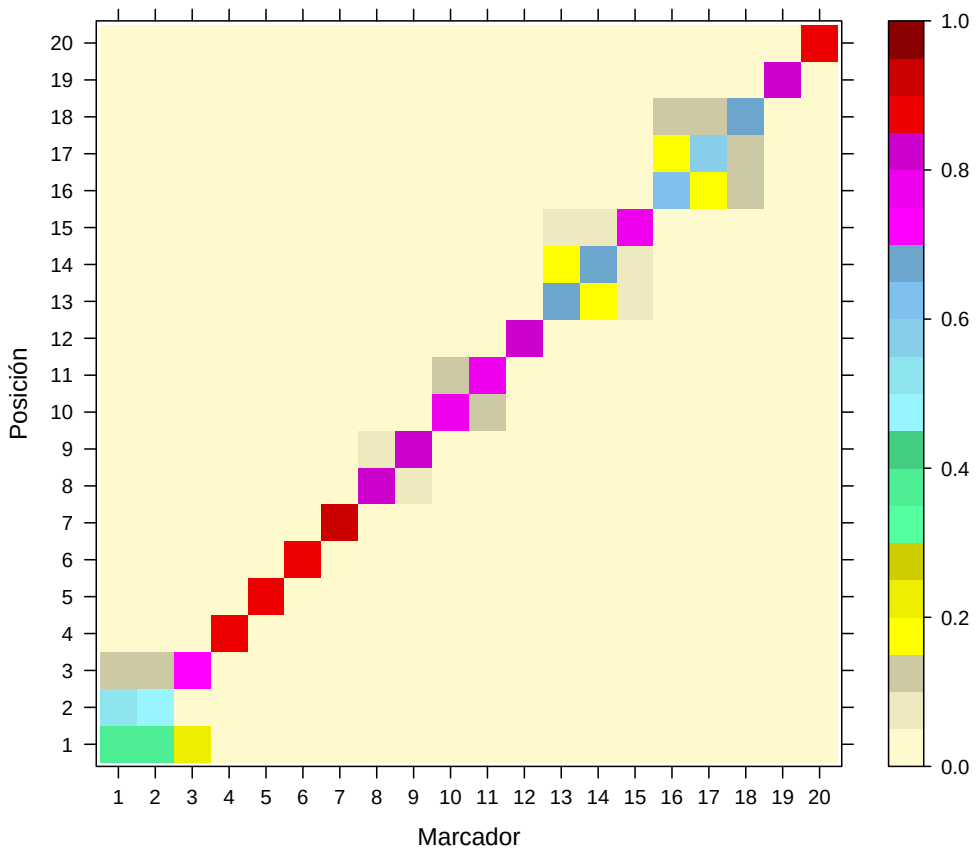


Figura 3.11: *Probabilidad con la que cada marcador aparece en cada una de las posibles posiciones del mapa genético más denso según la simulación de la muestra con 100 individuos.*

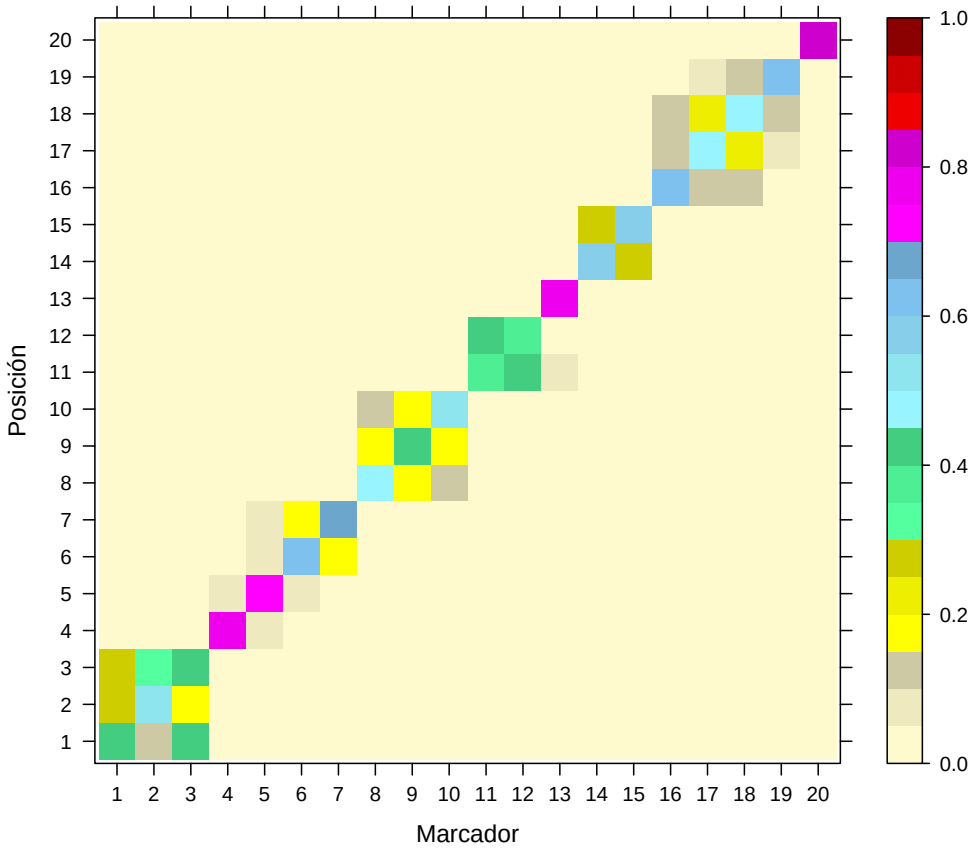


Figura 3.12: *Probabilidad con la que cada marcador aparece en cada una de las posibles posiciones del mapa genético más denso según la simulación de la muestra con 50 individuos.*

3.1.3. Discusión

Efecto de la densidad de marcadores.

De los resultados expuestos en esta sección se deduce que la metodología ha sido satisfactoria para estimar mapas genéticos de poblaciones con todos los marcadores codominantes, independientemente de la densidad del mapa genético de la población. Sólo se detectan problemas de estimación, destacables, en el caso hipotético, de trabajar con una muestra muy pequeña (50 individuos), procedente de una población con marcadores muy próximos. En ese caso, el modelo más probable difiere del mapa genético original en alguna permutación de marcadores adyacentes muy ligados. Trabajar con una muestra tan pequeña es poco real en algunos tipos de cultivos como hortícolas y cereales, aunque es más frecuente en estudios con frutales y especies forestales.

En la población con mapa más denso, la probabilidad del modelo más probable disminuye con respecto a la que se obtiene en la población con mapa menos denso. Una posible explicación sobre esta reducción sería que los marcadores de la población con mapa más denso están mucho más cercanos entre sí, por lo que existe una mayor probabilidad de estimar permutaciones entre marcadores, que dan lugar a una gama más amplia de modelos alternativos. Así, el modelo más probable obtiene una probabilidad menor. Esto puede repercutir en posteriores estudios en los que se vean involucrados marcadores SNPs. Los marcadores tipo SNPs son polimorfismos de un solo nucleótido ("Single Nucleotide Polymorphisms") en los que el simple cambio de un nucleótido en una secuencia genómica da lugar a distintos alelos.

En cuanto a la longitud de los modelos más probables con respecto al mapa genético, se puede considerar que son bastante similares independientemente de la densidad del mapa genético de la población. Parece que los modelos subestiman las distancias reales del mapa genético de la población con mapa menos denso y sobrestiman las distancias reales del mapa genético de la población con mapa más denso. Aunque este hecho puede ser específico de la muestra de datos seleccionada, en el Capítulo 9 se estudiará con mayor profundidad la estabilidad del método definitivo y de las conclusiones ante distintas muestras.

También se observa que, para las muestras procedentes de la población con mapa más denso, se obtienen un mayor número de modelos alternativos que provocan una disminución de la probabilidad tanto de los mejores modelos como la de los sucesivos modelos. En el caso de la población con mapa más

denso, cuando el tamaño muestral es pequeño, las probabilidades de los primeros y segundos modelos difieren poco entre sí. No obstante, este resultado puede estar condicionado a la muestra específica considerada.

Efecto del tamaño muestral en cada población.

En general, se observa que según se reduce el tamaño muestral a 200, 100 y 50 individuos, la probabilidad de los modelos más probables va disminuyendo en ambas poblaciones. En la población con mapa menos denso: 0.9087, 0.6753 y 0.2867 y en la población con mapa más denso: 0.3150, 0.0933 y 0.0077, respectivamente. Sin embargo, según se reduce el tamaño muestral, la probabilidad del siguiente modelo alternativo va aumentando en la población con mapa menos denso: 0.0600, 0.1677 y 0.1913, pero disminuye en la población con mapa más denso: 0.1273, 0.0837 y 0.0073. En definitiva, parece que, en la población con mapa menos denso, la reducción del tamaño muestral repercute en una reducción de la fiabilidad que la metodología otorga al modelo más probable.

En las Figuras 3.4, 3.5 y 3.6 se observa como, aun reduciendo el tamaño muestral de las muestras que provienen de la población con mapa menos denso, los marcadores quedan localizados en su posición correcta de forma casi inequívoca. Tan sólo cuando se trabaja con 50 individuos, para algunos marcadores muy ligados, se obtienen pequeñas probabilidades de ubicación en posiciones contiguas. En las figuras equivalentes para la población con mapa más denso, 3.10, 3.11 y 3.12, se observa un incremento de las probabilidades de ubicación a posiciones más allá de las contiguas, incluso para el tamaño muestral de 200 individuos. Como cabe esperar, la incertidumbre de la ubicación de los marcadores en sus posiciones correctas, es mayor cuanto menor es el tamaño muestral.

Comparativa entre población Retrocruce y F_2 con todos los marcadores codominantes.

En este caso se comparan los resultados de la población Retrocruce y los de la población F_2, con mapa menos denso, con todos los marcadores codominantes, pues recordemos que las muestras empleadas se extrajeron del mismo mapa genético. Se observa que la probabilidad del modelo más probable en la población Retrocruce es inferior a la obtenida para la población F_2, independientemente del tamaño muestral. Concretamente, en las muestras de

200, 100 y 50 individuos, en el caso Retrocruce fue 0.2380, 0.1173 y 0.0123, respectivamente, frente a 0.9087, 0.6753 y 0.2867, en el caso F_2. Parece que en ambos escenarios, la metodología subestima la longitud del mapa genético independientemente del tamaño muestral, aunque más acentuado en el caso Retrocruce. En las dos poblaciones sucede el mismo patrón: de 200 a 100 individuos se encogen los modelos y de 100 a 50 individuos se alargan pero sin llegar a la estimación de 200 individuos. En definitiva, parece que la metodología es capaz de estimar mejor el mapa genético si la muestra procede de una población F_2 que de una población Retrocruce, a pesar de que en ésta última, los recombinantes son directamente observables y por tanto, la estimación de las fracciones de recombinación se realiza de forma explícita. Quizás la explicación esté en que en el caso F_2, es observable la condición de ser homocigoto (AA o aa) y heterocigoto (Aa) lo que da lugar a 9 genotipos posibles por cada pareja de marcadores implicados. Esto supone una mayor información, en el reparto de individuos, que la que se maneja en el caso Retrocruce, que para cada pareja de marcadores sólo se contemplan dos genotipos posibles.

3.2. Diseño F_2 con marcadores dominantes.

3.2.1. Metodología

Una situación más realista respecto a la planteada en las secciones anteriores sería la de un diseño F_2 con no todos los marcadores codominantes, es decir un diseño en el que algunos de los marcadores que aparecen son dominantes, y por lo tanto sólo manifiestan uno de los dos alelos que lo forman, por lo que el nivel de información es menor.

Siguiendo una metodología similar a la desarrollada para un diseño F_2 con marcadores codominantes, se considera una población F_2 con una estructura de m marcadores, no todos codominantes, de la que se extrae una muestra de n individuos independientes. Cada uno de estos marcadores quedará definido según su tipo, dependiendo de los genotipos que sean observables. Se denotará un marcador como D1 si sólo son observables los genotipos $A\_$, aa y como D2 si sólo son observables los genotipos AA, $a\_$. Si el marcador es codominante (C) serán observables todos sus genotipos AA, Aa, aa.

Por lo tanto, en este diseño se codifica con un 0 si el individuo k en el marcador j es homocigoto (aa), con un 1 si es heterocigoto (Aa), con un 2 si es homocigoto (AA), con un 3 si es $(A\_)$ y con un 4 si es $(a\_)$, es decir

$$y_{kj} = \begin{cases} 0, \ si \ aa \\ 1, \ si \ Aa \\ 2, \ si \ AA \\ 3, \ si \ A\_ \\ 4, \ si \ a\_ \end{cases} \tag{3.11}$$

De este modo el banco de datos tiene una estructura como la que se muestra en el Cuadro 3.5.

	Marcadores						
Individuos	1	...	i	...	j	...	m
1	0	...	3	...	1	...	2
.	.	...	.	...	.	...	.
.	.	...	.	...	.	...	.
k	1	...	3	...	2	...	2
.	.	...	.	...	.	...	.
.	.	...	.	...	.	...	.
n	2	...	0	...	0	...	4

Cuadro 3.5: *Codificación de los marcadores en un diseño F_2 con marcadores dominantes.*

En este caso, la combinación de dos marcadores cualesquiera, da lugar al reparto de los n individuos en 9, 6 o 4 genotipos posibles, que dependen del tipo de ambos marcadores, como se muestra en los Cuadros del 3.6 al 3.11.

Análogamente al Capítulo 2, en cada celda de dichos cuadros, se muestra tanto el genotipo como la frecuencia observada ($R_{s,i,j}$)y esperada (γ_s) de cada genotipo. $r_{i,j}$ es la frecuencia esperada de recombinación entre los marcadores i y j o fracción de recombinación. Nótese que $\sum_{s=1}^{ngeno} R_{s,i,j} = n$, donde *ngeno* denota el número de genotipos según la combinación de marcadores y n es el tamaño muestral.

El Cuadro 3.6 es idéntico al Cuadro 3.2 de la sección anterior, que ya se justificó con el Cuadro 1.4 del Capítulo 1. Respecto al Cuadro 3.7, por ejemplo, las frecuencias de recombinación observadas y esperadas del genotipo *A_B_* se obtienen sumando las respectivas de los genotipos *AABB*, *AABb*, *AaBB* y *AaBb* del Cuadro 3.6. Y las del genotipo *aaB_*, sumando las respectivas de los genotipos *aaBB* y *aaBb* del Cuadro 3.6. En definitiva, las frecuencias de recombinación observadas y esperadas en los Cuadros del 3.7 al 3.11 se obtienen de forma directa o sumando las respectivas frecuencias del Cuadro 3.6.

marcadores C-C	BB	Bb	bb
AA	$AABB$ $R_{1,i,j}$ $\gamma_1 = 0.25(1 - r_{i,j})^2$	$AABb$ $R_{2,i,j}$ $\gamma_2 = 0.5r_{i,j}(1 - r_{i,j})$	$AAbb$ $R_{3,i,j}$ $\gamma_3 = 0.25r_{i,j}^2$
Aa	$AaBB$ $R_{4,i,j}$ $\gamma_4 = 0.5r_{i,j}(1 - r_{i,j})$	$AaBb$ $R_{5,i,j}$ $\gamma_5 = 0.5(1 - 2r_{i,j} + 2r_{i,j}^2)$	$Aabb$ $R_{6,i,j}$ $\gamma_6 = 0.5r_{i,j}(1 - r_{i,j})$
aa	$aaBB$ $R_{7,i,j}$ $\gamma_7 = 0.25r_{i,j}^2$	$aaBb$ $R_{8,i,j}$ $\gamma_8 = 0.5r_{i,j}(1 - r_{i,j})$	$aabb$ $R_{9,i,j}$ $\gamma_9 = 0.25(1 - r_{i,j})^2$

Cuadro 3.6: *Para la combinación de marcadores C-C, genotipos observables y frecuencias de recombinación observadas y esperadas.*

marcadores D1-D1	$B\_$	bb
$A\_$	$A\_B\_$ $R_{1,i,j}$ $\gamma_1 = 0.25(3 - 2r_{i,j} + r_{i,j}^2)$	A_bb $R_{2,i,j}$ $\gamma_2 = 0.25r_{i,j}(2 - r_{i,j})$
aa	$aaB\_$ $R_{3,i,j}$ $\gamma_3 = 0.25r_{i,j}(2 - r_{i,j})$	$aabb$ $R_{4,i,j}$ $\gamma_4 = 0.25(1 - r_{i,j})^2$

Cuadro 3.7: *Para la combinación de marcadores D1-D1, genotipos observables y frecuencias de recombinación observadas y esperadas.*

marcadores D1-D2	BB	$b\_$
$A\_$	A_BB $R_{1,i,j}$ $\gamma_1 = 0.25(1 - r_{i,j}^2)$	$A\_b\_$ $R_{2,i,j}$ $\gamma_2 = 0.25 r_{i,j}(2 + r_{i,j}^2)$
aa	$aaBB$ $R_{3,i,j}$ $\gamma_3 = 0.25 r_{i,j}^2$	$aab\_$ $R_{4,i,j}$ $\gamma_4 = 0.25(1 - r_{i,j}^2)$

Cuadro 3.8: *Para la combinación de marcadores D1-D2, genotipos observables y frecuencias de recombinación observadas y esperadas.*

marcadores D2-D2	BB	$b\_$
AA	$AABB$ $R_{1,i,j}$ $\gamma_1 = 0.25(1 - r_{i,j})^2$	$AAb\_$ $R_{2,i,j}$ $\gamma_2 = 0.25 r_{i,j}(2 - r_{i,j})$
$a\_$	a_BB $R_{3,i,j}$ $\gamma_3 = 0.25 r_{i,j}(2 - r_{i,j})$	$a\_b\_$ $R_{4,i,j}$ $\gamma_4 = 0.25(3 - 2r_{i,j} + r_{i,j}^2)$

Cuadro 3.9: *Para la combinación de marcadores D2-D2, genotipos observables y frecuencias de recombinación observadas y esperadas.*

marcadores C-D1	$B\_$	bb
AA	$AAB\_$ $R_{1,i,j}$ $\gamma_1 = 0.25(1 - r_{i,j}^2)$	$AAbb$ $R_{2,i,j}$ $\gamma_2 = 0.25 r_{i,j}^2$
Aa	$AaB\_$ $R_{3,i,j}$ $\gamma_3 = 0.5(1 - r_{i,j} + r_{i,j}^2)$	$Aabb$ $R_{4,i,j}$ $\gamma_4 = 0.5 r_{i,j}(1 - r_{i,j})$
aa	$aaB\_$ $R_{5,i,j}$ $\gamma_5 = 0.25 r_{i,j}(2 - r_{i,j})$	$aabb$ $R_{6,i,j}$ $\gamma_6 = 0.25(1 - r_{i,j})^2$

Cuadro 3.10: *Para la combinación de marcadores C-D1, genotipos observables y frecuencias de recombinación observadas y esperadas.*

marcadores C-D2	BB	$b\_$
AA	$AABB$ $R_{1,i,j}$ $\gamma_1 = 0.25(1 - r_{i,j})^2$	$AAb\_$ $R_{2,i,j}$ $\gamma_2 = 0.25 r_{i,j}(2 - r_{i,j})$
Aa	$AaBB$ $R_{3,i,j}$ $\gamma_3 = 0.5 r_{i,j}(1 - r_{i,j})$	$Aab\_$ $R_{4,i,j}$ $\gamma_4 = 0.5(1 - r_{i,j} + r_{i,j}^2)$
aa	$aaBB$ $R_{5,i,j}$ $\gamma_5 = 0.25 r_{i,j}^2$	$aab\_$ $R_{6,i,j}$ $\gamma_6 = 0.25(1 - r_{i,j}^2)$

Cuadro 3.11: *Para la combinación de marcadores C-D2, genotipos observables y frecuencias de recombinación observadas y esperadas.*

Equivalentemente a la sección anterior, el vector $\mathbf{R_{i,j}} = (R_{1,i,j}, ..., R_{ngeno,i,j})$ sigue un modelo *Multinomial*:

$$\mathbf{R_{i,j}} \sim Multinomial(n; \gamma),\ con\ \gamma = (\gamma_1, ..., \gamma_{ngeno})$$

La función de probabilidad de los datos, dado los parámetros n y $r_{i,j}$ se expresa como:

$$f(\mathbf{R_{i,j}}|r_{i,j}) \ = \frac{n!}{\prod_{s=1}^{ngeno} R_{i,j}^{s}!} \prod_{s=1}^{ngeno} (\gamma_s)^{R_{s,i,j}} \tag{3.12}$$

Como distribución a priori de $r_{i,j}$ utilizamos una distribución *Beta* truncada en $[0, \varepsilon]$, con $\varepsilon = 0.5$, es decir

$$\pi(r_{i,j}) = BetaT(\alpha, \beta; \varepsilon),\ con\ \alpha = \ \beta = \ 1 \tag{3.13}$$

Asumiendo independencia entre todos los vectores de frecuencias observadas $\mathbf{R_{1,2}}$, $\mathbf{R_{1,3}}$, ..., $\mathbf{R_{m\text{-}1,m}}$, e independencia a priori entre todas las fracciones de recombinación de pares de marcadores distintos, $\mathbf{r^*} = (r_{1,2},\ r_{1,3},\ ..., r_{m-1,m})$, tendremos que también las distribuciones posteriores resultan independientes entre sí. En definitiva, al igual que que ocurría en el diseño F_2 con sólo marcadores codominantes, simular de la distribución posterior para $\mathbf{r^*}$ es equivalente a simular independientemente de la distribución posterior de cada $r_{i,j}$ dado $\mathbf{R_{i,j}}$

Combinando la función de probabilidad de los datos y la distribución a priori, se obtiene una posterior, $\pi(r_{i,j}|\mathbf{R_{i,j}})$, para cada posible combinación de marcadores i y j, de la que se simula a través de Metropolis-Hastings, utilizando como distribución propuesta (proposal) una distribución *Normal* truncada, para obtener candidatos. Los parámetros de dicha *Normal* truncada se aproximan en idénticos términos a la sección anterior.

Nótese que la densidad posterior $g(r_{i,j})$ y la probabilidad de salto dependerán del tipo de combinación entre los marcadores i y j y serán equivalentes a las mostradas en el caso en que i y j son codominantes en (3.7) y (3.9), respectivamente.

Como ya vimos, una vez finalizada la simulación, se dispone de una cadena de fracciones de recombinación, $\{r^{(t)}\}_{t=1}^{nsim}$, que da lugar a una cadena de

mapas, $\{O^{(t)}\}_{t=1}^{nsim}$, según el algoritmo basado en las distancias mínimas entre marcadores ($SARF$), cuyos pasos básicos ya se detallaron anteriormente, obteniendo la distribución posterior de las distancias multipunto, $\{d^{(t)}\}_{t=1}^{nsim}$, en base al método de mínimos cuadrados cuyas ecuaciones se definen según la expresión (1.24).

3.2.2. Resultados

Para exponer unos resultados comparables a los obtenidos en el diseño F_2 con todos los marcadores codominantes, se consideran los dos mismos mapas genéticos de 20 marcadores (uno menos denso y el otro más denso), sólo que en este caso no todos los genotipos de los marcadores son observables. Los tipos de cada marcador y las distancias reales entre los marcadores adyacentes quedan resumidas en los Cuadros 3.12 y 3.13.

De cada población se han extraído una muestra de tamaño 200 individuos, ya que, como se comentará en la discusión, no se ha considerado necesario endurecer la condición sobre el tamaño muestral. Se ha aplicado la metodología simulando cadenas de longitud $nsim$= 3000, en los mismos términos que anteriormente. Los resultados se organizan del mismo modo que en las secciones anteriores.

Población F_2, con mapa menos denso, con marcadores codominantes y dominantes

En la Figura 3.13 se representa el mapa genético menos denso junto con los 5 modelos bayesianos (estimaciones) más probables obtenidos con la metodología descrita.

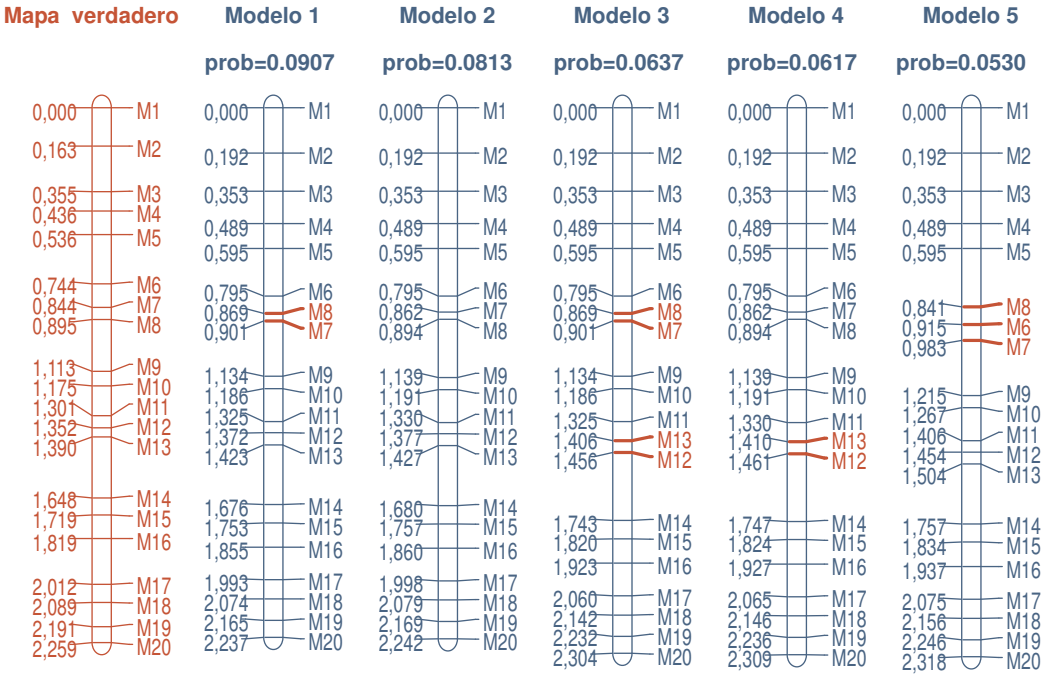


Figura 3.13: *Mapa verdadero de una población F_2, con mapa menos denso, con marcadores dominantes, junto con la estima de los 5 modelos multipunto más probables de una muestra con 200 individuos.*

En el Cuadro 3.12 se muestra el tipo de combinación entre marcadores adyacentes, junto con sus verdaderas distancias, así como la estima de la media posterior y la desviación típica posterior de la distribución de distancias multipunto para la muestra representada en la Figura 3.13, según el Modelo 2, que coincide con la ordenación real de los marcadores.

	tipo comb.	distancias reales	modelo bayesiano
			media ± desv. típ.
$d_{1,2}$	CC	0.162756	0.1924 ± 0.0340
$d_{2,3}$	CD1	0.191711	0.1608 ± 0.0403
$d_{3,4}$	D1D2	0.081566	0.1359 ± 0.0764
$d_{4,5}$	D2C	0.100000	0.1063 ± 0.0340
$d_{5,6}$	CC	0.207871	0.1992 ± 0.0411
$d_{6,7}$	CD1	0.100000	0.0673 ± 0.0249
$d_{7,8}$	D1D2	0.050557	0.0324 ± 0.0507
$d_{8,9}$	D2C	0.218596	0.2444 ± 0.1034
$d_{9,10}$	CC	0.061823	0.0520 ± 0.0095
$d_{10,11}$	CD1	0.125967	0.1388 ± 0.0681
$d_{11,12}$	D1D2	0.051199	0.0474 ± 0.0642
$d_{12,13}$	D2D1	0.037476	0.0505 ± 0.0716
$d_{13,14}$	D1C	0.257857	0.2527 ± 0.1295
$d_{14,15}$	CD2	0.071767	0.0771 ± 0.0326
$d_{15,16}$	D2C	0.100000	0.1028 ± 0.0308
$d_{16,17}$	CD1	0.192264	0.1379 ± 0.0412
$d_{17,18}$	D1C	0.076754	0.0811 ± 0.0184
$d_{18,19}$	CD1	0.102130	0.0903 ± 0.0277
$d_{19,20}$	D1C	0.068696	0.0724 ± 0.0229

Cuadro 3.12: *Tipo de combinación entre marcadores contiguos, distancias reales en Morgans, distancias medias multipunto posteriores y desviaciones típicas posteriores del modelo bayesiano correcto respecto a una muestra con 200 individuos.*

La Figura 3.14 se relaciona con la muestra representada en la Figura 3.13 y en el Cuadro 3.12

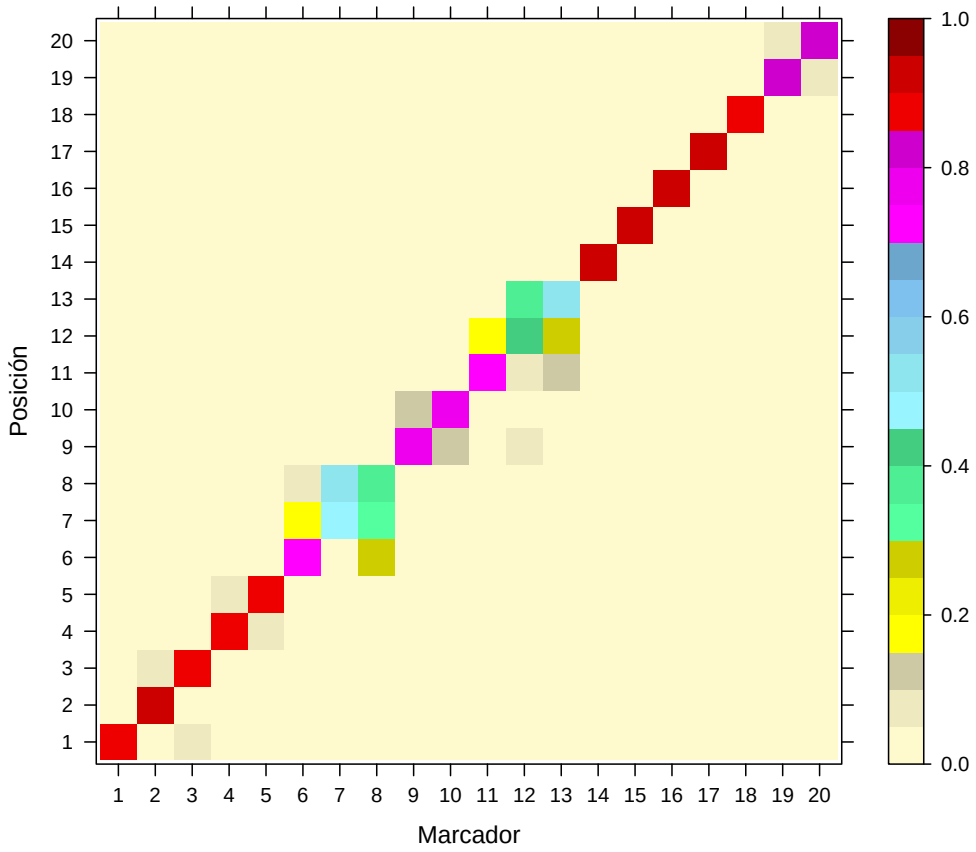


Figura 3.14: *Probabilidad con que cada marcador aparece en cada una de las posibles posiciones del mapa genético menos denso con algunos marcadores dominantes, según la simulación de la muestra con 200 individuos.*

Población F_2, con mapa más denso, con marcadores codominantes y dominantes

A continuación, se detallan resultados equivalentes a los anteriores para la población con mapa más denso con algunos marcadores dominantes.

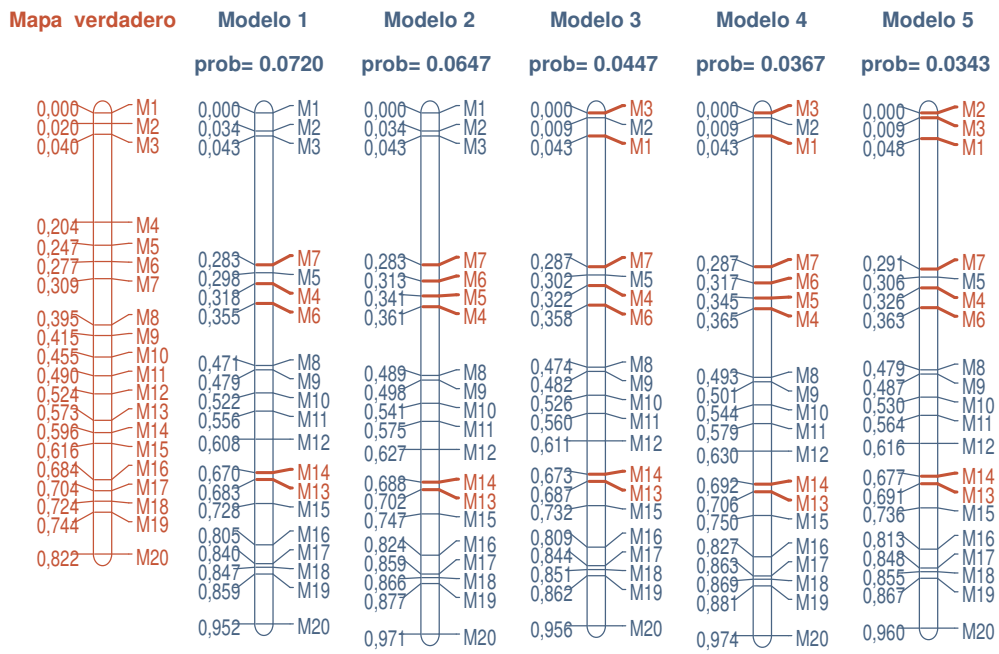


Figura 3.15: *Mapa verdadero de una población F_2, con mapa más denso, con algunos marcadores dominantes, junto con la estima de los 5 modelos bayesianos multipunto más probables de una muestra con 200 individuos.*

	tipo comb.	distancias reales	modelo bayesiano
			media ± desv. típ.
$d_{1,2}$	D1C	0.020000	0.0344 ± 0.0339
$d_{2,3}$	CD1	0.020000	0.0089 ± 0.0041
$d_{3,4}$	D1C	0.163737	0.2548 ± 0.0429
$d_{4,5}$	CD1	0.043085	0.0202 ± 0.0196
$d_{5,6}$	D1C	0.029992	0.0276 ± 0.0253
$d_{6,7}$	CD2	0.032026	0.0304 ± 0.0307
$d_{7,8}$	D2D1	0.086469	0.1406 ± 0.0491
$d_{8,9}$	D1D1	0.020000	0.0082 ± 0.0056
$d_{9,10}$	D1C	0.039634	0.0432 ± 0.0204
$d_{10,11}$	CC	0.034757	0.0343 ± 0.0200
$d_{11,12}$	CD1	0.034757	0.0515 ± 0.0299
$d_{12,13}$	D1C	0.048186	0.0688 ± 0.0289
$d_{13,14}$	CD2	0.022916	0.0139 ± 0.0286
$d_{14,15}$	D2C	0.020000	0.0446 ± 0.0215
$d_{15,16}$	CD1	0.068384	0.0771 ± 0.0281
$d_{16,17}$	D1D2	0.020000	0.0351 ± 0.0160
$d_{17,18}$	D2C	0.020000	0.0066 ± 0.0083
$d_{18,19}$	CC	0.020000	0.0116 ± 0.0098
$d_{19,20}$	CC	0.078403	0.0935 ± 0.0197

Cuadro 3.13: *Tipo de combinación entre marcadores contiguos, distancias reales en Morgans, distancia media posterior multipunto y desviación típica posterior del modelo bayesiano correcto respecto a la muestra con 200 individuos.*

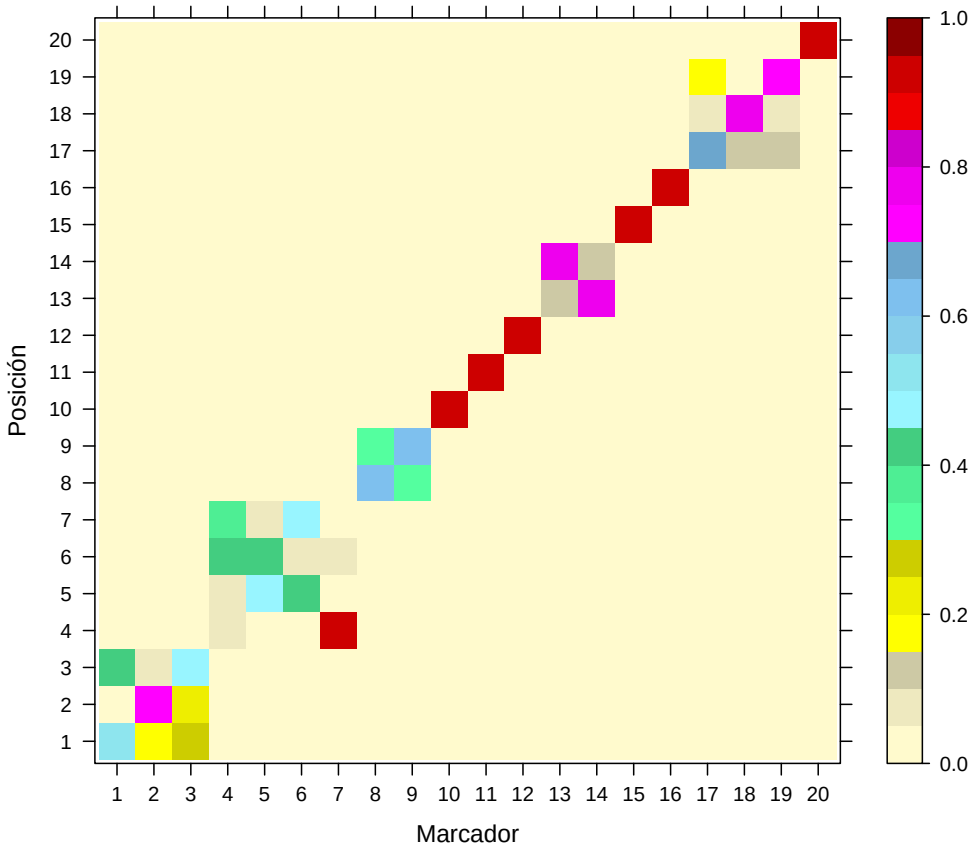


Figura 3.16: *Probabilidad con que cada marcador aparece en cada una de las posibles posiciones del mapa genético más denso, con algunos marcadores dominantes, según la simulación de la muestra con 200 individuos.*

3.2.3. Discusión

Nótese que, en esta sección, se ha trabajado con un único tamaño muestral (200 individuos). No se ha considerado necesario endurecer las condiciones sobre el tamaño de muestra para obtener conclusiones sobre la metodología, puesto que incluso para 200 individuos se manifiestan claramente los problemas asociados a la metodología.

Efecto de la densidad de marcadores.

En la muestra que provienen de la población con mapa genético menos denso, el modelo correcto aparece como segundo modelo alternativo al más probable y difieren entre sí en la permutación de una pareja de marcadores contiguos fuertemente ligados.

En el caso de la muestra de la población con mapa más denso, no se obtiene el modelo correcto entre los 5 más probables.

Las Figuras 3.14 y 3.16 muestran incertidumbre sobre la localización de los marcadores. Menos de la mitad de los marcadores quedan correctamente ubicados, con una alta probabilidad. Este hecho, genera una mayor variedad de modelos alternativos que repercute en una disminución de sus probabilidades asociadas. Además, los modelos alternativos que más se repiten, difieren poco entre si. Las diferencias estriban en una o dos permutaciones de marcadores fuertemente ligados, por ese motivo, sus probabilidades asociadas tampoco difieren mucho entre sí.

Comparativa entre población F_2 con todos los marcadores codominantes y F_2 con no todos los marcadores codominantes.

Parece que en este caso se sigue cumpliendo la tendencia que se detectó en el diseño F_2 con todos los marcadores codominantes. Los modelos obtenidos con la muestra de la población con mapa menos denso subestiman las distancias reales entre los marcadores y los modelos obtenidos con la muestra de la población con mapa más denso sobrestiman las distancias reales entre los marcadores.

La probabilidad del modelo correcto, en esta última sección, se reducen drásticamente frente a la obtenida con todos los marcadores codominantes.

Se concluye que los resultados no sólo se ven afectados por la densidad de marcadores del mapa genético de la población, sino también por el tipo de los marcadores que participan en él. Los dos mapas genéticos han sido diseñados para recoger todo tipo de dificultades, como la de contener marcadores adyacentes fuertemente ligados en repulsión (D1D2 y D2D1). Esta combinación de marcadores, $\{i, j\}$, resulta especialmente problemática porque el único genotipo claramente distinguible ($aaBB$ para D1D2 o $AAbb$ para D2D1) tiene asociada una frecuencia de recombinación esperada de $0.25r_{i,j}^2$ (Cuadro 3.8). Si los marcadores, en ese caso, son muy cercanos, su fracción de recombinación, $r_{i,j}$, será próxima a cero y serán poco probables individuos con ese genotipo, por lo que, la estimación se ve resentida por un problema de falta de información.

Falta de información que ha sido estudiada, por otros autores, bajo distintas perspectivas. Por ejemplo, Säll y Nilsson (1994) [69] estudiaron la robustez de las estimaciones máximo verosímiles de las fracciones de recombinación, entre marcadores dominantes, en base al sesgo debido a distintos factores como son el tamaño muestral y la clasificación errónea de los individuos. Concluyeron que entre marcadores dominantes en fase de acoplamiento, la estimación es robusta y el sesgo insignificante para tamaños muestrales superiores a 20 individuos. La clasificación errónea fue tipificada como la causa más grave de sesgo, tendiendo a aumentar las estimaciones de las fracciones de recombinación, particularmente cuando las fracciones de recombinación eran pequeñas. Sin embargo, para marcadores dominantes en fase de repulsión, concluyeron que el sesgo podía ser muy grave incluso para tamaños muestrales de varios cientos de individuos. En este caso, se observaba sesgo negativo, especialmente para fracciones de recombinación superiores a 0.05.

Knapp et al (1995) [39] también estudiaron el sesgo de las estimaciones de las fracciones de recombinación entre marcadores dominantes llegando a conclusiones equivalentes a otros estudios realizados con anterioridad por otros investigadores. Señalaron que, aunque otros autores atribuyen el problema del sesgo a la varianza del estimador máximo verosímil de las fracciones de recombinación, el verdadero problema viene provocado por el error cuadrático medio, definido como $sesgo^2 + varianza$. También apuntaron que aunque se suele considerar como orden más probable de los loci aquel que minimiza la longitud del mapa genético, esto sólo sería cierto para poblaciones que obtienen estimaciones insesgadas de las fracciones de recombinación.

Mester et al. (2003) [53] en su propuesta para la ordenación de marcadores, advirtieron que cuanto más alta es la proporción de marcadores dominantes en fase de repulsión, menor es la calidad de la ordenación multilocus. Asimismo, señalaron problemas para determinar el orden de marcadores en regiones de alta densidad y tamaño muestral insuficiente.

Tan y Fu (2007) [80] En su propuesta para la estimación de fracciones de recombinación entre marcadores dominantes, advirtieron que el algoritmo EM estima mal las fracciones de recombinación entre marcadores dominantes porque no distingue bien la fase en que se relacionan. Reconocen que su método tiene deficiencias y que necesita de un tamaño muestral alrededor de 300 individuos F_2, pero que puede ser un buen método para complementar el algoritmo EM.

Hühn y Piepho (2008) [26] en su estudio sobre el sesgo de las estimaciones de las fracciones de recombinación entre marcadores, bajo la función de mapeo Karlin, señalaron que disminuye rápidamente con el incremento del tamaño muestral en marcadores dominantes en fase de acomplamiento, como pasaba también en la población Retrocruce. Sin embargo, si los marcadores están en fase de repulsión, el comportamiento del sesgo es inestable y depende fuertemente de la magnitud de la fracción de recombinación y del tamaño muestral disponible. Para tamaños muestrales pequeños, el sesgo es positivo y bastante grande. A medida que el tamaño muestral aumenta, el sesgo se hace negativo, alcanzando un mínimo y luego se aproxima lentamente a cero. Todo ello afectado por fluctuaciones importantes a modo de dientes de sierra.

Añadido a estos problemas, siempre se puede sufrir el que implica el muestreo en sí mismo. Ya se sabe que al extraer una muestra de una población, por el azar de los individuos observados, al aplicar la metodología desarrollada, quizás se esté estimando "correctamente" un hipotético mapa genético que no es exactamente el de la población original.

Así pues, se concluye que la metodología propuesta no es adecuada para estimar mapas genéticos de poblaciones en las que no son observables todos los genotipos de todos los marcadores. En este caso, queda de manifiesto que, aun trabajando con muestras de 200 individuos, se dispone de menos información de la necesaria para hacer estimaciones satisfactorias del mapa genético.

Las probabilidades obtenidas por los distintos modelos (estimaciones) no son convincentes a la hora de decidir cuál es el modelo correcto. De hecho, para la muestra de la población con mapa genético más denso, el modelo correcto

no está entre los 5 más probables. En ese mismo escenario, los problemas en la ubicación de los marcadores pasan a manifestarse por bloques.

Capítulo 4

Ordenación de tripletas de marcadores. Simulación marginal.

4.1. Introducción

En el capítulo anterior se han observado problemas en la estimación de las fracciones de recombinación para muestras que provienen de una población con un diseño F_2 en el que no todos los marcadores son codominantes, especialmente en el caso en que los marcadores están muy ligados. Este problema de estimación afecta posteriormente a la ordenación de los marcadores, basada en distancias mínimas ($SARF$). Para estudiar en profundidad este problema vamos a reducir el número de marcadores que definen el mapa genético de la población a la mínima expresión (3 marcadores) y a diseñar una modelización que estime las fracciones de recombinación entre marcadores y los ordene conjuntamente de una manera más fiable.

Se considera el problema de estimar las fracciones de recombinación, $\mathbf{r} = (r_1; r_2; r_3) \equiv (r_{1,2}; r_{1,3}; r_{2,3})$ entre tres marcadores M_1, M_2 y M_3, no todos ellos codominantes.

$$M_1 \overbrace{\underbrace{\qquad}_{r_1 \equiv r_{1,2}} M_2 \underbrace{\qquad}_{r_3 \equiv r_{2,3}}}^{r_2 \equiv r_{1,3}} M_3$$

Para ello, se dispone del número de recombinantes observados entre cada par de marcadores, $\mathbf{R} = (R_1; R_2; R_3) \equiv (\mathbf{R_{1,2}}; \mathbf{R_{1,3}}; \mathbf{R_{2,3}})$, cuya distribución depende de sus fracciones de recombinación respectivas $(r_1; r_2; r_3)$. Además, cada R_i es un vector cuya dimensión coincide con el número de genotipos observables en la combinación entre el par de marcadores involucrados. Cuando ambos marcadores son codominantes (C), su dimensión es 9, como se puede ver en el Cuadro 3.6; cuando ambos son dominantes (D1D1, D1D2 o D2D2), su dimensión es de 4, como aparece en los Cuadros 3.7, 3.8 y 3.9 y cuando uno es codominante y el otro dominante (CD1 o CD2), la dimensión de R_i es 6, según se representa en los Cuadros 3.10 y 3.11.

4.1.1. Ejemplo

Con el fin de aclarar bien las implicaciones de esta notación asumida, consideramos un mapa genético con tres marcadores, el primero M_1 codominante y los siguientes M_2 y M_3 dominantes. A continuación, se muestra un esquema del mapa genético, con las fracciones de recombinación implicadas y los tipos de marcador:

$$M_1(\mathrm{C}) \overbrace{\overbrace{\qquad}^{r_1} M_2(\mathrm{D1}) \overbrace{\qquad}^{r_3}}^{r_2} M_3(\mathrm{D2})$$

Las coordenadas de los vectores de frecuencias observadas de recombinantes, R_i, y las coordenadas de los vectores de frecuencias genotípicas esperadas, $\rho_i(r_i)$, para cada pareja de marcadores, se pueden representar en las siguientes tablas:

CD1	$B\_$	bb
AA	$R_{1,1,2}$ $0.25(1-r_1^2)$	$R_{2,1,2}$ $0.25r_1^2$
Aa	$R_{3,1,2}$ $0.5(1-r_1+r_1^2)$	$R_{4,1,2}$ $0.5r_1(1-r_1)$
aa	$R_{5,1,2}$ $0.25r_1(2-r_1)$	$R_{6,1,2}$ $0.25(1-r_1)^2$

CD2	CC	$c\_$
AA	$R_{1,1,3}$ $0.25(1-r_2)^2$	$R_{2,1,3}$ $0.25r_2(2-r_2)$
Aa	$R_{3,1,3}$ $0.5r_2(1-r_2)$	$R_{4,1,3}$ $0.5(1-r_2+r_2^2)$
aa	$R_{5,1,3}$ $0.25r_2^2$	$R_{6,1,3}$ $0.25(1-r_2^2)$

D1D2	CC	$c\_$
$B\_$	$R_{1,2,3}$ $0.25(1-r_3^2)$	$R_{2,2,3}$ $0.25(2+r_3^2)$
bb	$R_{3,2,3}$ $0.25r_3^2$	$R_{4,2,3}$ $0.25(1-r_3^2)$

Cuadro 4.1: Frecuencias de recombinación observadas y esperadas para cada pareja de marcadores.

Es decir, en este caso y denotando los marcadores M_1, M_2 y M_3 como loci A, B y C, respectivamente, los vectores implicados serían:

$\mathbf{R} = (R_1; R_2; R_3)$, donde $R_1 = (R_{1,1,2}, ..., R_{6,1,2}) \in N^6$, $R_2 = (R_{1,1,3}, ..., R_{6,1,3}) \in N^6$ y $R_3 = (R_{1,2,3}, ..., R_{4,2,3}) \in N^4$

$\rho_1(r_1) = (\ 0.25(1-r_1^2),\ 0.25r_1^2,\ 0.5(1-r_1+r_1^2), ...,\ 0.25(1-r_1)^2\ )$

$\rho_2(r_2) = (\ 0.25(1-r_2)^2,\ 0.25r_2(2-r_2),\ 0.5r_2(1-r_2), ...,\ 0.25(1-r_2^2)\ )$

$\rho_3(r_3) = (\ 0.25(1-r_3^2),\ 0.25(2+r_3^2),\ 0.25r_3^2,\ 0.25(1-r_3^2)\ )$

4.2. Modelización

La modelización que se propone tiene en cuenta los siguientes aspectos:

4.2.1. Previa para el orden

Para una tripleta de marcadores, existen tres ordenaciones posibles:

$$\begin{aligned} O_1 = O_{123} &= M_1M_2M_3 \\ O_2 = O_{132} &= M_1M_3M_2 \\ O_3 = O_{312} &= M_3M_1M_2 \end{aligned} \tag{4.1}$$

Cuando no tenemos ninguna información, cada ordenación O_i tiene la misma probabilidad a priori de ser cierta:

$$\pi(O_i) = \frac{1}{3}, \quad i = 1,\ 2,\ 3. \tag{4.2}$$

4.2.2. Previa para las fracciones de recombinación

La distribución a priori para las fracciones de recombinación, $\mathbf{r} = (r_1; r_2; r_3)$, ha de responder a las siguientes consideraciones:

• Dado que, conocido el orden O_i entre los tres marcadores, es conocida la posición de los marcadores extremos, la distancia entre ellos se puede calcular en términos de las distancias al marcador central. Además, estas distancias al marcador central son independientes entre sí (hablamos de independencia condicional, conocido el orden). En términos de fracciones de recombinación y asumiendo ausencia de interferencia, esto se resume como sigue:

Si el orden es O_1, hay independencia entre r_1 y r_3, mientras que $r_2 = r_1 + r_3 - 2r_1r_3$.
Si el orden es O_2, hay independencia entre r_2 y r_3, mientras que $r_1 = r_2 + r_3 - 2r_2r_3$.
Si el orden es O_3, hay independencia entre r_1 y r_2, mientras que $r_3 = r_2 + r_1 - 2r_2r_1$.

• En general, para un orden O_k se puede considerar la independencia entre dos componentes del vector $\mathbf{r} = (r_1; r_2; r_3)$, identificadas por r_{k_1} y r_{k_2} y la dependencia de la tercera componente r_{k_3} con respecto a las dos anteriores. r_{k_3} representa la distancia entre los dos marcadores extremos y cumplirá la relación

$$r_{k_3} = g(r_{k_1}, r_{k_2}) = r_{k_1} + r_{k_2} - 2r_{k_1}r_{k_2} \tag{4.3}$$

en función de las distancias entre los marcadores contiguos r_{k_1} y r_{k_2}.
Así pues,

	r_{k_1}	r_{k_2}	r_{k_3}
$O_1 = M_1M_2M_3$	$r_1 \equiv r_{1,2}$	$r_3 \equiv r_{2,3}$	$r_2 \equiv r_{1,3}$
$O_2 = M_1M_3M_2$	$r_2 \equiv r_{1,3}$	$r_3 \equiv r_{2,3}$	$r_1 \equiv r_{1,2}$
$O_3 = M_3M_1M_2$	$r_2 \equiv r_{1,3}$	$r_1 \equiv r_{1,2}$	$r_3 \equiv r_{2,3}$

Cuadro 4.2: Relación entre las distancias de dos marcadores según el orden.

Utilizando las consideraciones detalladas antes, la distribución a priori para $\mathbf{r} = (r_1; r_2; r_3)$, condicionada a que se da el orden O_k, y asumiendo desconocimiento previo sobre las fracciones de recombinación de los marcadores extremos al central (cualquier valor entre 0 y 0.5 es igualmente plausible), se puede expresar como:

$$\pi(\mathbf{r}|O_k) = BetaT(r_{k_1}|1,1;0.5)BetaT(r_{k_2}|1,1;0.5)P_I(r_{k_3}|\ r_{k_1}, r_{k_2}) \qquad (4.4)$$

donde $P_I(r_{k_3}|\ r_{k_1}, r_{k_2})$ es la distribución con masa puntual de probabilidad igual a 1 en el valor $r_{k_3} = g(r_{k_1}, r_{k_2})$.

4.2.3. Modelización de los datos

Como siempre, existe independencia condicional entre R_i y R_j , conocidas sus fracciones de recombinación respectivas r_i y r_j y dado un orden concreto entre los marcadores. Además, cada R_i representa un reparto multinomial de los n individuos observados sobre los diferentes genotipos observables para el par de marcadores y justifica a continuación la dependencia de las probabilidades genotípicas en el orden entre marcadores. Así pues, la función de probabilidad de los datos dados los parámetros se escribirá como:

$$f(\mathbf{R}|\mathbf{r}, O_k) = \prod_{i=1}^{3} f(R_i|r_i, O_k) = \prod_{i=1}^{3} Multinomial(R_i|\ n; \rho_{i,k}(r_i)) \qquad (4.5)$$

donde $\rho_{i,k}(r_i)$ proporciona el vector de frecuencias genotípicas esperadas para r_i bajo la ordenación O_k. Es decir, dependiente del orden existente entre los marcadores indexados por i y de su fracción de recombinación r_i asociada al par de marcadores indexados por i.

En definitiva, el **modelo jerárquico bayesiano** que proponemos queda expresado en tres niveles, según:

$$\begin{array}{lll} I. & \pi(O_k) = & \frac{1}{3}, \quad k = 1,2,3 \\ II. & \pi(\mathbf{r}|O_k) = & BetaT(r_{k_1}|1,1;0.5)BetaT(r_{k_2}|1,1;0.5)P_I(r_{k_3}|\ r_{k_1}, r_{k_2}) \\ III. & f(\mathbf{R}|\mathbf{r},O_k) = & \prod_{i=1}^{3} f(R_i|r_i,O_k) = \prod_{i=1}^{3} Multinomial(R_i|\ n;\rho_{i,k}(r_i)) \end{array} \tag{4.6}$$

Esta modelización, es equivalente a asumir que la distribución marginal a priori de las fracciones de recombinación es una mixtura de distribuciones condicionadas por el orden:

$$\pi(\mathbf{r}) = \sum_{i=1}^{3} \pi(\mathbf{r}|\ O_i)\pi(O_i) \tag{4.7}$$

4.3. Distribución posterior. Simulación

Combinando la función de probabilidad de los datos dados los parámetros y la distribución a priori, la distribución posterior de interés la podemos calcular como:

$$\pi(\mathbf{r}|\mathbf{R}) = \sum_{k=1}^{3} \pi(\mathbf{r}, O_k|\mathbf{R}) = \sum_{k=1}^{3} \pi(\mathbf{r}|\mathbf{R}, O_k)\pi(O_k|\mathbf{R}) \tag{4.8}$$

donde

$$\pi(\mathbf{r}|\mathbf{R}, O_k) \propto f(\mathbf{R}|\mathbf{r}, O_k)\pi(\mathbf{r}|O_k) \tag{4.9}$$

con $f(\mathbf{R}|\mathbf{r}, O_k)$ y $\pi(\mathbf{r}|O_k)$ detalladas en (4.5) y (4.4), respectivamente.

Seguidamente, $\pi(O_k|\mathbf{R})$ se puede calcular según:

$$\pi(O_k|\mathbf{R}) = \frac{f(\mathbf{R}|O_k)\pi(O_k)}{\sum_{j=1}^{3} f(\mathbf{R}|O_j)\pi(O_j)} \propto \frac{f(\mathbf{R}|O_k)}{\sum_{j=1}^{3} f(\mathbf{R}|O_j)} \tag{4.10}$$

donde

$$f(\mathbf{R}|O_k) = \int_r f(\mathbf{R}|\mathbf{r}, O_k)\pi(\mathbf{r}|O_k)dr \tag{4.11}$$

Para simular de la distribución posterior de las fracciones de recombinación, $\pi(\mathbf{r}|\mathbf{R})$, y así poder inferir sobre dichas fracciones (y después sobre las distancias entre los marcadores) nos valdremos de su expresión como mixtura de las distribuciones posteriores condicionadas al orden de los marcadores, y utilizaremos un mecanismo de simulación basado en Reversible Jump (Green (1995) [28]; Lunn et al. (2009) [50]; Robert y Casella (2013) [64]; Gelman (2014) [22]), que consiste básicamente, en un método de simulación Metropolis-Hastings generalizado, que ofrece la posibilidad de saltar de un modelo a otro. La dinámica de dicho algoritmo será la siguiente:

1. Simularemos un orden, O_{k^*}, de la distribución posterior $\pi(O_k|\mathbf{R})$. En realidad, lo haremos de una aproximación a dicha distribución dada su dependencia (4.10) de la integral no analítica (4.11) que proporciona esta probabilidad. Para aproximar dichas probabilidades posteriores se utilizará una aproximación numérica de $f(\mathbf{R}|O_k)$ mediante cuadratura Gausiana (Véase Apéndice D), que denotaremos como $\widehat{f}(\mathbf{R}|O_k)$, esto es:

$$p(k) = \frac{\widehat{f}(\mathbf{R}|O_k)}{\sum_{j=1}^{3} \widehat{f}(\mathbf{R}|O_j)}, k = 1, 2, 3. \tag{4.12}$$

Simulamos a continuación el orden O_{k^*} con estas probabilidades $p(k)$, según el procedimiento convencional. Es decir, tomar el menor k^* tal que $u < \sum_{j=1}^{k^*} p(k_j)$, siendo $u \sim Uniforme(0, 1)$.

2. Con dicho orden O_{k^*}, simularemos un candidato de la distribución $\pi(\mathbf{r}|\mathbf{R}, O_{k*})$. Dada la forma compleja de esta distribución, dependiente del orden de los marcadores, optamos por simular candidatos $r_{k_i^*}$ para i=1 ,2 bajo el orden simulado previamente, O_k^*, mediante distribuciones *Normales* truncadas en el intervalo [0,0.5], con media la moda de $f(R_{k_i^*}|r_{k_i^*}, O_{k^*})$, $\hat{\mu}_{k_i^*}$, y varianza la inversa de la matriz de información de Fisher evaluada en dicha moda, $\hat{\sigma}_{k_i^*}$,

$$q_{k^*}(r_{k_1^*}; r_{k_2^*}) = \prod_{i=1}^{2} NormalT(r_{k_i^*}|\hat{\mu}_{k_i^*}, \hat{\sigma}_{k_i^*}; 0, 0.5) \tag{4.13}$$

y calculamos mediante dichas simulaciones $r_{k_3^*} = g(r_{k_1^*}; r_{k_2^*})$, con *g()* definida en (4.3). Tenemos ya el vector candidato de fracciones de recombinación r_{k^*}.

A continuación, aceptamos en la iteración, *(t)*, de la cadena de simulación al candidato (r_{k^*}, O_{k^*}) con probabilidad de salto dependiente de los elementos ya aceptados en la iteración anterior $(\mathbf{r}^{(t-1)}, O^{(t-1)})$:

$$prob.salto = min\{1, \frac{G(r_{k^*}, O_{k^*}, \mathbf{R})}{G(r^{(t-1)}, O^{(t-1)}, \mathbf{R})}\} \tag{4.14}$$

con $G(\mathbf{r}_k, O_k, \mathbf{R}) = \frac{\pi^*(\mathbf{r}_k|\mathbf{R}, O_k)}{q_k(r_{k_1}; r_{k_2}) \cdot p(k)}$ donde $\pi^*(\mathbf{r}|\mathbf{R}, O_k) = f(\mathbf{R}|\mathbf{r}, O_k)\pi(\mathbf{r}|O_k)$ vienen dados en (4.9), $q_k()$ viene definido en (4.13) y $p()$ en (4.12).

En el caso de no aceptar el candidato, repetimos los pasos 1 y 2 hasta aceptar un candidato para la fracción de recombinación y el orden, $(\mathbf{r}^{(t)}, O^{(t)})$.

4.4. Ilustración del método

Para ilustrar el método de ordenación de tres marcadores, se proponen dos situaciones:

• Una tripleta de la forma CD1C, consistente en un marcador codominante, a continuación otro dominante y en el otro extremo otro codominante, como se puede ver esquemáticamente a continuación:

$$M_1(\mathrm{C}) \underbrace{\overbrace{\qquad}^{r_1=0.02} M_2(\mathrm{D1}) \overbrace{\qquad}^{r_3=0.02}}_{r_2=0.0392} M_3(\mathrm{C})$$

• Una tripleta de la forma CD1D2, en la que dos de los marcadores contiguos son dominantes en fase de repulsión, de forma que el mapa genético es más complicado de estimar:

$$M_1(\mathrm{C}) \underbrace{\overbrace{\qquad}^{r_1=0.02} M_2(\mathrm{D1}) \overbrace{\qquad}^{r_3=0.02}}_{r_2=0.0392} M_3(\mathrm{D2})$$

Con el fin de concluir de modo general sobre la eficiencia del método de ordenación propuesto, en cada una de las situaciones mencionadas, se consideran 500 muestras de tamaño n= 200 (razonablemente grande para aportar información "suficiente" de la inferencia), y sobre ellas se aplica la metodología descrita anteriormente. La longitud de las cadenas de Metropolis-Hastings es de 100.000 (la computación es rápida), con el fin de garantizar convergencia.

Se obtiene, en cada muestra, las estimaciones y errores (posteriores) de las fracciones de recombinación r_1, r_2 y r_3 conseguidas a partir de las distribuciones posteriores simuladas, así como las probabilidades posteriores de los órdenes posibles O_1, O_2 y O_3. Asimismo concluiremos sobre el funcionamiento del algoritmo de simulación propuesto para la distribución posterior de dichas fracciones de recombinación. Los resultados se resumen en los siguientes apartados.

4.4.1. Tripleta C D1 C

• El funcionamiento del algoritmo de simulación es razonable, atendiendo a los porcentajes de salto:

mediana=55.39 %, 3er cuartil= 58.44 % y máximo=66.73 %

• La media y la desviación típica de las medias posteriores de las fracciones de recombinación, aparecen en el Cuadro 4.3 :

$r_1 = 0.02$	$r_2 = 0.0392$	$r_3 = 0.02$
media $\pm$ sd	media $\pm$ sd	media $\pm$ sd
0.0236 $\pm$ 0.0093	0.0416 $\pm$ 0.0107	0.0226 $\pm$ 0.0092

Cuadro 4.3: Medias y desviaciones típicas de las medias posteriores.

• Las probabilidades posteriores de los órdenes posibles O_1= CD1C, O_2= CCD1 y O_3= CCD1 se muestran en la Figura 4.1

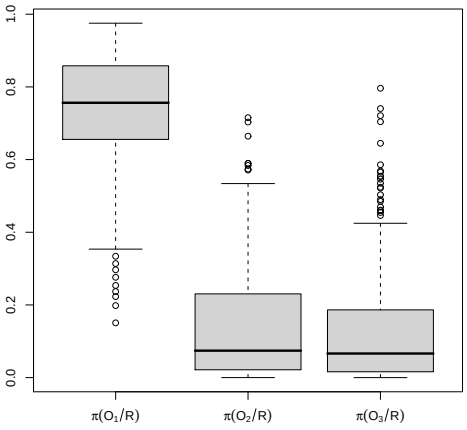


Figura 4.1: *Diagrama de cajas para las estimas posteriores de los órdenes dados los datos.*

4.4.2. Tripleta C D1 D2

• Del mismo modo, en este caso, el funcionamiento del algoritmo de simulación es razonable, atendiendo a los porcentajes de salto:

mediana=61.3 %, 3er cuartil 64.31 % y máximo=69.49 %

• La media y la desviación típica de las medias posteriores de las fracciones de recombinación, aparecen en el Cuadro 4.4:

$r_1 = 0.02$	$r_2 = 0.0392$	$r_3 = 0.02$
media $\pm$ sd	media $\pm$ sd	media $\pm$ sd
0.0239 $\pm$ 0.0110	0.0400 $\pm$ 0.0110	0.0223 $\pm$ 0.0101

Cuadro 4.4: Medias y desviaciones típicas de las medias posteriores.

• Las probabilidades posteriores de los órdenes posibles O_1= CD1D2, O_2= CD2D1 y O_3= D2CD1 se muestran en la Figura 4.2.

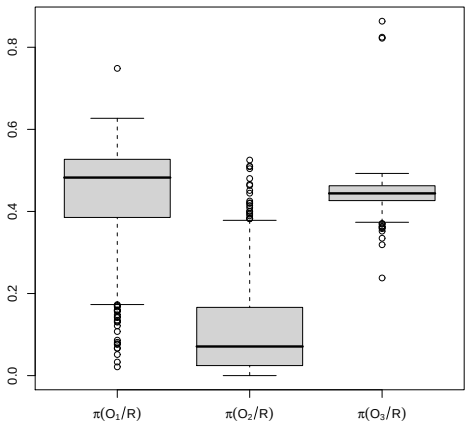


Figura 4.2: *Diagrama de cajas para las estimas posteriores de los órdenes dados los datos.*

4.5. Conclusión

Las conclusiones que derivamos a través de la aplicación del método de ordenación propuesto en las dos situaciones, son las siguientes:

• La probabilidad de salto en las cadenas MH es razonable: en torno al 60 % en ambas situaciones.

• La variabilidad que se aprecia entre las estimaciones posteriores de las fracciones de recombinación es realmente pequeña (en términos de desviaciones típicas). Además, dichas estimaciones resultan considerablemente precisas: en prácticamente todos los casos presentados el valor real de r_i queda a menos de una desviación típica de la estimación media (promedio del valor esperado de la distribución posterior para las 500 muestras). Encontramos algo más problemática la estimación de r_3 para la situación más desfavorable CD1D2 en la que dicha distancia era la existente entre los marcadores dominantes en fase de repulsión D1 y D2; en este caso, el valor real queda a 3 desviaciones típicas del promedio.

• El orden real queda claramente identificado en la situación CD1C, con unas probabilidades posteriores superiores a 0.75 en más de la mitad de las 500 muestras. Esta metodología parece resolver con éxito la estimación de las fracciones de recombinación y la ordenación conjunta de tripletas de marcadores fuertemente ligados, aun en el caso en que no son observables todos los genotipos de uno de los marcadores (D1).

• Siguen sin obtenerse resultados satisfactorios en el caso extremo en que no son observables todos los genotipos de dos de los marcadores de la tripleta, como ocurre en la situación CD1D2, en que a la vista de la Figura 4.2, no llega a identificarse el orden real con una probabilidad posterior diferenciada del resto de posibilidades de ordenación. En esta situación tampoco se estima eficientemente la fracción de recombinación entre los marcadores D1 y D2.

Capítulo 5

Ordenación de tripletas de marcadores. Simulación conjunta.

Con el objetivo de mejorar los problemas detectados en el capítulo anterior, se proponen algunas modificaciones sobre el algoritmo basado en la modelización con mixturas.

5.1. Descripción del algoritmo

Se considera idéntica notación y el mismo mecanismo de simulación descrito en el capítulo anterior.

Recordemos que si en una determinada iteración se obtenía el vector $(\mathbf{r}; O_k)$, en la siguiente iteración se sorteaba un nuevo orden (modelo) O_{k^*}, de entre los órdenes posibles $\{O_1; O_2; O_3\}$, según las probabilidades posteriores de los modelos, vistas en (4.10) y se simulaba un vector candidato $\mathbf{r}^{(k^*)}$, a participar en la cadena Metropolis Hastings, utilizando una distribución propuesta (proposal) basada en $\pi(\mathbf{r}|\mathbf{R}, O_k)$ (véase (4.9)).

En este capítulo se modifican las distribuciones proposal unidimensionales para simular candidatos a entrar en la cadena MH para aproximar la distribución posterior de las fracciones de recombinación r_{k_1} y r_{k_2}, sustituyéndolas por una distribución bivariante, también basada en la verosimilitud. La distribución proposal ahora es una *Normal* bivariada truncada en

(0, 0.5), $NormalT(r_{k_1}, r_{k_2}|\ \hat{\mu}, \hat{\sigma}; 0, 0.5)$, centrada en la moda de la verosimilitud para r_{k_1} y r_{k_2}, *L()*, y con matriz de varianzas-covarianzas la inversa del hessiano evaluada en dicha moda, donde

$$\begin{aligned} L(r_{k_1}, r_{k_2}|R) = \quad & Multinomial(R_{k_1}|n; \rho_{k_1}(r_{k_1}))\cdot \\ & Multinomial(R_{k_2}|n; \rho_{k_2}(r_{k_2}))\cdot \\ & Multinomial(R_{k_3}|n; \rho_{k_3}(r_{k_1} + r_{k_2} - 2r_{k_1}r_{k_2})) \end{aligned} \tag{5.1}$$

para las fracciones de recombinación de los marcadores contiguos bajo el orden O_{k^*}, con el siguiente mapa genético:

$$M_1 \overbrace{\overbrace{\qquad}^{r_{k_1}} M_2 \overbrace{\qquad}^{r_{k_2}}}^{r_{k_3}=r_{k_1}+r_{k_2}-2r_{k_1}r_{k_2}} M_3$$

El cálculo de la moda se realiza, como habitualmente, en términos del logaritmo de la verosimilitud, $l(r_{k_1}, r_{k_2}|R)$.

Obtenido el vector candidato a entrar en la cadena, todo el procedimiento descrito en el capítulo anterior se reproduce para simular las distribuciones posteriores de las fracciones de recombinación y de los órdenes.

5.2. Ilustración del método

Como esta propuesta está pensada para mejorar los resultados de estimación de las fracciones de recombinación y la ordenación de marcadores dominantes fuertemente ligados en fase de repulsión, la metodología se ensaya sobre la tripleta de marcadores más problemática, con mapa genético:

$$M_1(\mathrm{C}) \underbrace{\overbrace{\qquad}^{r_1=0.02} M_2(\mathrm{D1}) \overbrace{\qquad}^{r_3=0.02}}_{r_2=0.0392} M_3(\mathrm{D2})$$

Se consideran las mismas tablas que en el Cuadro 4.1 para construir las expresiones de las logverosimilitudes conjuntas en función del orden obtenido en el sorteo.

• Así, si en la nueva iteración, el orden sorteado es:

$$O_1 \equiv M_1(\mathrm{C}) \underbrace{\overbrace{\qquad}^{r_1} M_2(\mathrm{D1}) \overbrace{\qquad}^{r_3}}_{r_2} M_3(\mathrm{D2})$$

entonces $r_{k_1} = r_1$, $r_{k_2} = r_3$ y el logaritmo de la verosimilitud se expresa según:

$$\begin{aligned} l(r_1, r_3|R) \quad & \propto R_{1,1,2} \cdot log(1 - r_1^2) + R_{2,1,2} \cdot log(r_1^2) + ... \\ & + R_{6,1,2} \cdot log((1 - r_1)^2) \\ & + R_{1,2,3} \cdot log(1 - r_3^2) + R_{2,2,3} \cdot log(2 + r_3^2) \\ & + R_{3,2,3} \cdot log(r_3^2) + R_{4,2,3} \cdot log(1 - r_3^2) \\ & + R_{1,1,3} \cdot log((1 - r_2)^2) + R_{2,1,3} \cdot log(r_2(2 - r_2)) + ... \\ & + R_{6,1,3} \cdot log(1 - r_2^2) = \\ & = R_{1,1,2} \cdot log(1 - r_1^2) + R_{2,1,2} \cdot log(r_1^2) + ... \\ & + R_{6,1,2} \cdot log((1 - r_1)^2) \\ & + R_{1,2,3} \cdot log(1 - r_3^2) + R_{2,2,3} \cdot log(2 + r_3^2) \\ & + R_{3,2,3} \cdot log(r_3^2) + R_{4,2,3} \cdot log(1 - r_3^2) \\ & + R_{1,1,3} \cdot log((1 - r_1 - r_3 + 2r_1r_3)^2) \\ & + R_{2,1,3} \cdot log((r_1 + r_3 - 2r_1r_3)(2 - r_1 - r_3 + 2r_1r_3)) + ... \\ & + R_{6,1,3} \cdot log(1 - (r_1 + r_3 - 2r_1r_3)^2), \end{aligned} \tag{5.2}$$

siendo R=($R_{1,2}$; $R_{2,3}$; $R_{1,3}$) y $r_2 = r_1 + r_3 - 2 \cdot r_1r_3$.

- Del mismo modo, si el orden sorteado es:

$$O_2 \equiv M_1(\text{C}) \underbrace{\overset{r_2}{\frown} M_3(\text{D2}) \overset{r_3}{\frown}}_{r_1} M_2(\text{D1})$$

entonces $r_{k_1} = r_2$, $r_{k_2} = r_3$ y el logaritmo de la verosimilitud se expresa según:

$$\begin{aligned} l(r_2, r_3|R) \quad & \propto R_{1,1,3} \cdot log((1-r_2)^2) + R_{2,1,3} \cdot log(r_2(2-r_2)) + ... \\ & + R_{6,1,3} \cdot log(1-r_2^2) \\ & + R_{1,2,3} \cdot log(1-r_3^2) + R_{2,2,3} \cdot log(2+r_3^2) \\ & + R_{3,2,3} \cdot log(r_3^2) + R_{4,2,3} \cdot log(1-r_3^2) \\ & + R_{1,1,2} \cdot log(1-r_1^2) + R_{2,1,2} \cdot log(r_1^2) + ... \\ & + R_{6,1,2} \cdot log((1-r_1)^2) \\ & = R_{1,1,3} \cdot log((1-r_2)^2) + R_{2,1,3} \cdot log(r_2(2-r_2)) + ... \\ & + R_{6,1,3} \cdot log(1-r_2^2) \\ & + R_{1,2,3} \cdot log(1-r_3^2) + R_{2,2,3} \cdot log(2+r_3^2) \\ & + R_{3,2,3} \cdot log(r_3^2) + R_{4,2,3} \cdot log(1-r_3^2) \\ & + R_{1,1,2} \cdot log(1-(r_2+r_3-2r_2r_3)^2) \\ & + R_{2,1,2} \cdot log((r_2+r_3-2r_2r_3)^2) + ... \\ & + R_{6,1,2} \cdot log((1-(r_2+r_3-2r_2r_3))^2) \end{aligned} \tag{5.3}$$

siendo R=($R_{1,3}$; $R_{2,3}$; $R_{1,2}$) y $r_1 = r_2 + r_3 - 2 \cdot r_2 r_3$.

• Por último, si el orden sorteado es:

$$O_3 \equiv M_3(\text{D2}) \underbrace{\overset{r_2}{\frown} M_1(\text{C}) \overset{r_1}{\frown}}_{r_3} M_2(\text{D1})$$

entonces $r_{k_1} = r_2$, $r_{k_2} = r_1$ y el logaritmo de la verosimilitud se expresa según:

$$\begin{aligned} l(r_2, r_1|R) \quad & \propto R_{1,1,3} \cdot log((1-r_2)^2) + R_{2,1,3} \cdot log(r_2(2-r_2)) + ... \\ & + R_{6,1,3} \cdot log(1-r_2^2) \\ & + R_{1,1,2} \cdot log(1-r_1^2) + R_{2,1,2} \cdot log(r_1^2) + ... \\ & + R_{6,1,2} \cdot log((1-r_1)^2) \\ & + R_{1,2,3} \cdot log(1-r_3^2) + R_{2,2,3} \cdot log(2+r_3^2) \\ & + R_{3,2,3} \cdot log(r_3^2) + R_{4,2,3} \cdot log(1-r_3^2) \\ & = R_{1,1,3} \cdot log((1-r_2)^2) + R_{2,1,3} \cdot log(r_2(2-r_2)) + ... \\ & + R_{6,1,3} \cdot log(1-r_2^2) \\ & + R_{1,1,2} \cdot log(1-r_1^2) + R_{2,1,2} \cdot log(r_1^2) + ... \\ & + R_{6,1,2} \cdot log((1-r_1)^2) \\ & + R_{1,2,3} \cdot log(1-(r_2+r_1-2r_2r_1)^2) \\ & + R_{2,2,3} \cdot log(2+(r_2+r_1-2r_2r_1)^2) \\ & + R_{3,2,3} \cdot log((r_2+r_1-2r_2r_1)^2) \\ & + R_{4,2,3} \cdot log(1-(r_2+r_1-2r_2r_1)^2) \end{aligned} \tag{5.4}$$

siendo R=($R_{1,3}$; $R_{1,2}$; $R_{2,3}$) y $r_3 = r_2 + r_1 - 2 \cdot r_2 r_1$.

A continuación, para cada orden sorteado, presentamos las gráficas de la función de verosimilitud bivariante para r_{k_1} y r_{k_2} seguidas de las gráficas de las distribuciones *Normales* bivariantes que se utilizan para simular los candidatos MH para la distribución posterior de las fracciones de recombinación. Las gráficas se han particularizando para la siguiente muestra, que es bastante representativa de lo que ocurre en general:

$R_{1,2}$	$B\_$	bb
AA	49	0
Aa	94	1
aa	2	54

$R_{1,3}$	CC	$c\_$
AA	46	3
Aa	5	90
aa	0	56

$R_{2,3}$	CC	$c\_$
$B\_$	51	94
bb	0	55

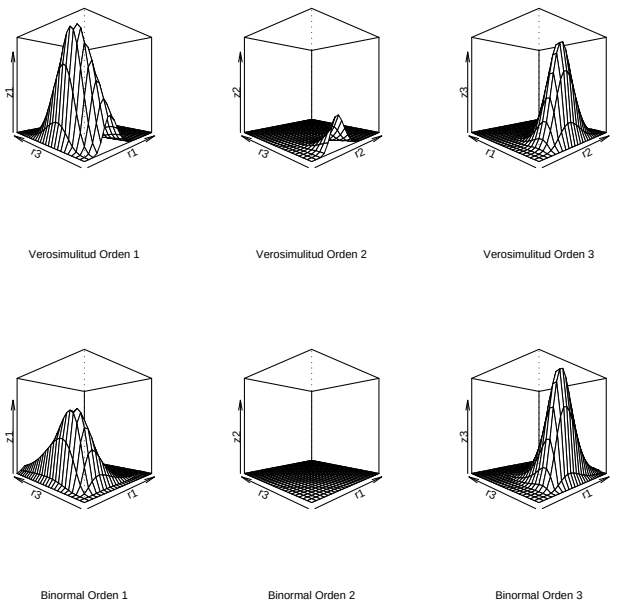


Figura 5.1: *Representación, según un ángulo de -45º, de las verosimilitudes bivariantes y de las* Normales *bivariantes que las aproximan, para cada uno de los 3 órdenes, en el intervalo (0,0.1).*

Para obtener la moda de la verosimilitud bivariante, se ha utilizando la función *mle( )* del entorno R [62].

Nótese en la Figura 5.1, que no se aprecia la forma de la gráfica de la *Normal* bivariante referente al orden 2. Sin embargo, si se ajusta la escala de la representación (Figura 5.2), se observa una gran varianza para la fracción de recombinación r_3 que repercute en la no convergencia del método de simulación. Por este motivo, cuando se simula la distribución posterior de las fracciones de recombinación, la cadena Metropolis Hastings se queda enganchada si se acepta un candidato de orden 2.

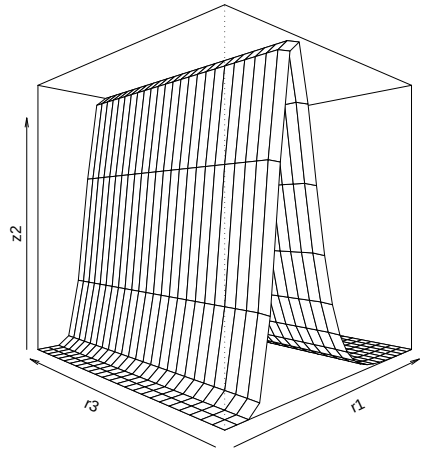


Figura 5.2: *Representación, según un ángulo de -45º, de la* Normal *bivariante que aproxima la verosimilitud bivariante referente al orden 2, en el intervalo (0,0.1).*

Los resultados obtenidos tras simular una cadena de longitud 50000, para la muestra especificada, son:

r_1= 0.02 media± sd	r_2= 0.0392 media± sd	r_3= 0.02 media± sd	prob. salto MH	prob. orden 1	prob. orden 2	prob. orden 3
0.0272± 0.0088	0.0284± 0.0089	0.0016± 0.0080	0.0468	0.0480	0.9385	0.0134

Cuadro 5.1: *Medias y desviaciones típicas posteriores de las simulaciones generadas de la distribución posterior con el algoritmo utilizado.*

5.3. Conclusiones

• Se confiaba en esta modificación en la distribución proposal, para simular de la distribución posterior de las fracciones de recombinación, supondría una mejora por el hecho de incluir en la modelización de la verosimilitud, toda la información disponible sobre las frecuencias de recombinación observadas entre las tres parejas de marcadores que definen una tripleta. Sin embargo, encontramos deficiencias graves con esta nueva modelización:

- La estimación de r_3 sigue siendo deficiente.

- Surgen problemas de la estimación de los parámetros de la distribución proposal que se utiliza para el Orden 2, lo que provoca la no convergencia del método.

- En consecuencia, la probabilidad posterior del orden 2 (que no es el correcto), resulta prácticamente igual a 1, dejando anulados a los órdenes 1 (el real) y 3.

• Como ya se señaló en la discusión del Capítulo 3, en el caso de la combinación D1D2, el único genotipo inequívocamente distinguible es el $bbCC$ que tiene asociada una frecuencia esperada de $0.25r_3^2$ (Cuadro 3.8 o última tabla del Cuadro 4.1). En este caso, cuando la distancia entre los dos marcadores es pequeña, la frecuencia relativa esperada es próxima a cero, ya que la fracción de recombinación se encuentra elevada al cuadrado. Por lo tanto, es necesario un tamaño de muestra elevado para que la frecuencia observada, $R^3_{2,3}$, no sea cero. Por ejemplo, para una distancia próxima a 1cM, la frecuencia esperada del genotipo $bbCC$ es de 0.000025, por lo que para tamaños de muestra razonables (en torno a 100 o 200 individuos) es muy probable que no se observe ningún individuo de este tipo. De hecho, en los datos observados del ejemplo, la frecuencia absoluta de esa combinación es cero. Esto influye negativamente en el cálculo de la verosimilitud y en el desarrollo de la metodología.

• En definitiva, en los diseños F_2 en los que intervienen conjuntamente marcadores dominantes y codominantes, se observan especialmente problemas en el caso de marcadores estrechamente ligados en fase de repulsión (D1D2).

Capítulo 6

Ordenación de tripletas de marcadores. Simulación conjunta y Slice Sampler. Generalización.

Como se acaba de ver en el capítulo anterior, el uso de la distribución *Normal* bivariada como distribución propuesta (proposal) para simular fracciones de recombinación de los marcadores contiguos bajo un cierto orden, no produce resultados satisfactorios. Sin embargo, nos parece acertado seguir trabajando con verosimilitudes conjuntas que incluyen, de algún modo, toda la información recogida por la muestra. Así pues, a continuación se describe un nuevo algoritmo de simulación multipunto para tripletas de marcadores, basado en una combinación de los algoritmos Slice Sampler (Neal 2003 [56]; Robert y Casella 2013 [64]) y Metropolis-Hastings, en la que la distribución propuesta (proposal) es la propia logverosimilitud conjunta. El objetivo sigue siendo estimar las fracciones de recombinación entre marcadores y ordenarlos correctamente.

6.1. Descripción del algoritmo

Imaginemos que en la iteración t-1 se obtiene de la cadena de simulación de la distribución posterior $\pi(\mathbf{r}, O_k|\mathbf{R})$ el vector $(\mathbf{r}^{(t-1)}; O_k^{(t-1)})$, donde $O_k^{(t-1)}$ es un orden de entre los posibles $\{O_1; O_2; O_3\}$ y $\mathbf{r}^{(t-1)}=(r_{k_1}^{(t-1)}, r_{k_2}^{(t-1)}, r_{k_3}^{(t-1)})$ representa las fracciones de recombinación simuladas para las parejas de marcadores que proceden del mapa genético sujeto al orden $O_k^{(t-1)}$:

$$M_1 \overbrace{\overset{r_{k_1}^{(t-1)}}{\frown} M_2 \overset{r_{k_2}^{(t-1)}}{\frown}}^{r_{k_3}^{(t-1)}=r_{k_1}^{(t-1)}+r_{k_2}^{(t-1)}-2r_{k_1}^{(t-1)}r_{k_2}^{(t-1)}} M_3$$

Asumimos, igual que en el capítulo anterior, que r_{k_1} y r_{k_2} corresponden a las fracciones de recombinación de marcadores adyacentes, (tal y como se muestra en el esquema anterior).

En la iteración t, se sortea un orden, $O_k^{(t)}$, que puede ser diferente al $O_k^{(t-1)}$. Se localiza el último $\widetilde{r} = (\widetilde{r}_{k_1}, \widetilde{r}_{k_2}, \widetilde{r}_{k_3})$ que entró en la cadena bajo ese mismo orden $O_k^{(t)}$ y se considera:

$$z = l(\widetilde{r}_{k_1}, \widetilde{r}_{k_2}|\mathbf{R}) - e, \quad e \sim Exponencial(1),$$

con *l()* el logaritmo de la función de verosimilitud, *L()*, de las fracciones de recombinación contiguas.

Para obtener un candidato, $\mathbf{r^*}$, el algoritmo simula $r_{k_1^*}$ y $r_{k_2^*}$ de un recinto rectangular,

$$[\mathrm{L}_{k_1}^*, \mathrm{U}_{k_1}^*] \times [\mathrm{L}_{k_2}^*, \mathrm{U}_{k_2}^*] \subset [0, 0.5]^2,$$

susceptible de ser encogido para aproximarse más eficientemente a la distribución objetivo. Recuérdese que $r_{k_3^*}$ se calculaba a través de la función $g(r_{k_1^*}, r_{k_2^*})$ dependiente del tipo de marcadores (4.3). Es decir, en pseudocódigo:

Para i=1,2:
$u_i \sim Uniforme(0,1)$
$\mathrm{L}^*_{k_i} = \widetilde{r}_{k_i} - w * u_i$
$\mathrm{U}^*_{k_i} = \mathrm{L}^*_{k_i} + w$
donde w es una aproximación global del ancho de los intervalos $[\mathrm{L}^*_{k_i}, \mathrm{U}^*_{k_i}]$, definida al inicio del algoritmo.
Repetir mientras $(z < l(r_{k_1^*}, r_{k_2^*}|\mathbf{R}))$:
$u_i \sim Uniforme(0,1)$ i= 1, 2.
$r_{k_1^*} = \mathrm{L}^*_{k_i} + u_i * (\mathrm{U}^*_{k_i} - \mathrm{L}^*_{k_i})$

Se calcula la probabilidad de salto (*prob.salto*) del nuevo candidato, $\mathbf{r^*}$, a entrar en la cadena Metropolis-Hastings mediante (4.14)
Se simula $u \sim Uniforme(0,1)$ y si $u < prob.salto$ se acepta en la cadena el nuevo vector $(\mathbf{r}_k^{(t)}; O_k)$

En caso de que el candidato $\mathbf{r^*}$ no sea más verosímil que el anterior $\widetilde{\mathbf{r}}$, (incorporado en la cadena de simulación para ese mismo orden) se encoge el recinto y se simula un nuevo candidato. Nótese que si se produce un encogimiento sucesivo, esto puede dar lugar a un recinto con un único punto no satisfactorio, en cuyo caso se iniciaría de nuevo la iteración t sorteando un nuevo orden.

6.2. Ilustración del método y discusión

Para poner en marcha el algoritmo se considera la misma muestra de 200 individuos que se utilizó para ejemplificar la metodología aplicada en el capítulo anterior. Es decir, la muestra extraída de una población con mapa genético:

$$\underbrace{M_1(\mathrm{C})\underbrace{\qquad}_{r_1=0.02}M_2(\mathrm{D1})\underbrace{\qquad}_{r_3=0.02}M_3(\mathrm{D2}),}_{r_2=0.0392} \tag{6.1}$$

cuyas frecuencias de recombinación observadas se recuerdan en las siguientes tablas:

$R_{1,2}$	$B\_$	bb
AA	49	0
Aa	94	1
aa	2	54

$R_{1,3}$	CC	$c\_$
AA	46	3
Aa	5	90
aa	0	56

$R_{2,3}$	CC	$c\_$
$B\_$	51	94
bb	0	55

En este caso, las probabilidades posteriores de los ordenes dados los datos son ($\underbrace{0.52165247}_{orden1=ABC}$, $\underbrace{0.03459091}_{orden2=ACB}$, $\underbrace{0.03475662}_{orden3=CAB}$)

Recuérdese que las gráficas de las tres verosimilitudes bivariantes según los órdenes posibles aparecen en la Figura 5.1.

Tras un estudio sobre la sensibilidad del algoritmo respecto a valores iniciales considerablemente diferentes, se concluye la convergencia al mismo punto. Además, se comprueba que los candidatos son aceptados en la misma proporción que indican las probabilidades posteriores de los órdenes, por lo que la convergencia del algoritmo no dependen de los valores iniciales.

Así pues, utilizando, por ejemplo, como valores iniciales rini=c(0.01480452, 0.02757598, 0.02669249), se obtienen los siguientes resultados como medias y desviaciones típicas posteriores de una cadena de longitud 50.000:

r_1=0.02 media ± sd	r_2=0.0392 media ± sd	r_3=0.02 media ± sd	prob. orden 1	prob. orden 2	prob. orden 3
0.0194 ± 0.0100	0.0456 ± 0.0144	0.0432 ± 0.0238	0.5233	0.0338	0.4429

Nótese que a diferencia de las metodologías utilizadas en los dos capítulos anteriores, no se ofrecen resultados sobre los porcentajes de salto ya que cada iteración produce una nueva simulación.

Tras los resultados obtenidos se observa que el proceso de simulación evita prácticamente la inclusión en la cadena de candidatos que provienen del orden con menor probabilidad. Sin embargo, no es capaz de moderar la interferencia que existe entre los dos órdenes más probables de los cuales procede la muestra. Esto influye de manera negativa en la estima de la media posterior de las fracciones de recombinación y en consecuencia sobre la ordenación de los marcadores del mapa genético. Quizás sería necesario investigar modificaciones sobre la expresión que determina la probabilidad de salto de una iteración a la iteración siguiente.

6.3. Generalización a más de tres marcadores

Una vez vistos los resultados para tripletas, se propone aplicar la metodología descrita anteriormente, para la ordenación de cuatro marcadores.

Para ejemplificar este problema, se van a considerar 4 muestras de 3 poblaciones definidas por 4 marcadores y de cada una de ellas se va a calcular la probabilidad posterior de los órdenes de los que procede cada muestra utilizando integración numérica mediante cuadratura gaussiana. Nótese que en los tres últimos capítulos se ha trabajado con tripletas de marcadores que determinaban tan sólo 3 órdenes posibles. Sin embargo, cuando el mapa genético se compone de 4 marcadores, aumenta a 12 órdenes posibles. Estos son:

Orden1	$M_1\ M_2\ M_3\ M_4$
Orden2	$M_1\ M_2\ M_4\ M_3$
Orden3	$M_1\ M_3\ M_2\ M_4$
Orden4	$M_1\ M_3\ M_4\ M_2$
Orden5	$M_1\ M_4\ M_2\ M_3$
Orden6	$M_1\ M_4\ M_3\ M_2$
Orden7	$M_2\ M_1\ M_3\ M_4$
Orden8	$M_2\ M_1\ M_4\ M_3$
Orden9	$M_2\ M_3\ M_1\ M_4$
Orden10	$M_2\ M_4\ M_1\ M_3$
Orden11	$M_3\ M_1\ M_2\ M_4$
Orden12	$M_3\ M_2\ M_1\ M_4$

Recordamos que, en un mapa genético con m marcadores se definen $\frac{m!}{2}$ órdenes posibles.

Con el fin de no extraer conclusiones basadas en una única muestra, se consideran 4 muestras de 200 individuos extraídas independientemente de una población cuyo mapa genético tiene todos los marcadores codominantes:

$$\overbrace{\underbrace{M_1(\mathrm{C}) \underbrace{\quad}_{r_1=0.017158} M_2(\mathrm{C}) \underbrace{\quad}_{r_4=0.030935} M_3(\mathrm{C})}_{r_2} \underbrace{\quad}_{r_6=0.018935} M_4(\mathrm{C}),}^{r_3} \tag{6.2}$$
$$\overbrace{\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad}^{r_5}$$

Para cada una de las 4 muestras se calculan, mediante cuadratura Gausina, las probabilidades posteriores de los órdenes posibles, que se muestran en el Cuadro 6.1:

$\pi(O_i\|R)$	muestra1	muestra2	muestra3	muestra4
Orden1	0.76705	0.58094	0.69473	0.77566
Orden2	0.02688	0.11288	0.02949	0.04054
Orden3	0.00000	0.00293	0.00000	0.00000
Orden4	0.01757	0.09141	0.00273	0.01839
Orden5	0.00000	0.00068	0.00000	0.00000
Orden6	0.03528	0.19299	0.00331	0.02900
Orden7	0.07803	0.00642	0.22557	0.04601
Orden8	0.05473	0.00800	0.01446	0.03252
Orden9	0.00000	0.00000	0.00000	0.00000
Orden10	0.00000	0.00000	0.00000	0.00000
Orden11	0.00726	0.00137	0.01346	0.02169
Orden12	0.01320	0.00236	0.01626	0.03619

Cuadro 6.1: Estimaciones posteriores de los órdenes obtenidas con integración numérica mediante cuadratura gaussiana.

Aunque hay diferencias entre las probabilidades estimadas, como se puede ver, no existe lugar a dudas del orden asumido para simularlas.

Consideremos ahora las mismas 4 muestras que en el caso anterior, sólo que codificadas de manera que los marcadores del mapa genético ya no son todos codominantes. En particular, se codifican las 4 muestras como si se hubiesen extraído de una población con el siguiente mapa genético:

$$\overbrace{\underbrace{\underbrace{M_1(\mathrm{D1})}_{r_1=0.017158}\underbrace{M_2(\mathrm{C})}_{r_4=0.030935}\underbrace{M_3(\mathrm{D1})}_{r_6=0.018935}}_{r_2}\underbrace{M_4(\mathrm{D2})}}^{r_5}, \qquad (6.3)$$

r_3

En este caso, las probabilidades posteriores de los órdenes se estiman en el siguiente cuadro:

$\pi(O_i\|R)$	muestra1	muestra2	muestra3	muestra4
Orden1	0.00277	0.01579	0.00760	0.02196
Orden2	0.00005	0.00077	0.00565	0.00084
Orden3	0.00014	0.01174	0.00001	0.00431
Orden4	0.08284	0.02804	0.05339	0.01956
Orden5	0.00001	0.00531	0.00023	0.00158
Orden6	0.29529	0.30938	0.14864	0.19842
Orden7	0.36091	0.01554	0.38028	0.12704
Orden8	0.19630	0.08761	0.38845	0.20218
Orden9	0.04247	0.45251	0.00157	0.36390
Orden10	0.01479	0.06656	0.00643	0.03641
Orden11	0.00135	0.00020	0.00254	0.00152
Orden12	0.00307	0.00654	0.00522	0.02229

Cuadro 6.2: Estimaciones posteriores de los órdenes obtenidas con integración numérica mediante cuadratura gaussiana.

Nótese que el orden que obtiene mayor probabilidad posterior no coincide en ningún caso con el verdadero orden de los marcadores en la población, que es el Orden1. Si en este caso aplicáramos los algoritmos de simulación propuestos en los últimos tres capítulos, es de esperar que las cadenas de simulación incluirían esencialmente fracciones de recombinación procedentes de marcadores ordenados de forma incorrecta.

Por último, modificamos el mapa genético anterior eliminando su primer marcador, que es dominante, y añadimos en la cola un marcador codominante como aparece a continuación:

$$\overbrace{\overbrace{\underbrace{M_1(\mathrm{C})}_{r_1=0.030935}\underbrace{M_2(\mathrm{D1})}_{r_4=0.018935}\underbrace{M_3(\mathrm{D2})}_{r_6=0.143953}}^{r_2}M_4(\mathrm{C})}^{r_3},\qquad\overbrace{\qquad\qquad\qquad\qquad}^{r_5} \tag{6.4}$$

En este caso, las probabilidades posteriores de los órdenes calculadas con cuadratura gaussiana son:

$\pi(O_i\|R)$	muestra1	muestra2	muestra3	muestra4
Orden1	0.82319	0.91542	0.71422	0.94890
Orden2	0.00000	0.00004	0.00001	0.00009
Orden3	0.17658	0.07674	0.28432	0.03983
Orden4	0.00000	0.00002	0.00001	0.00002
Orden5	0.00000	0.00000	0.00000	0.00000
Orden6	0.00000	0.00000	0.00000	0.00000
Orden7	0.00007	0.00253	0.00070	0.00460
Orden8	0.00000	0.00000	0.00000	0.00000
Orden9	0.00001	0.00015	0.00001	0.00007
Orden10	0.00000	0.00000	0.00000	0.00000
Orden11	0.00008	0.00275	0.00067	0.00290
Orden12	0.00007	0.00235	0.00006	0.00360

Cuadro 6.3: Estimaciones posteriores de los órdenes obtenidas con integración numérica mediante cuadratura gaussiana.

El orden correcto obtiene la probabilidad más alta en las 4 muestras.

Como hemos podido comprobar con estas tres poblaciones, y como ya advirtieron Mester et al. (2003) [53] la proporción de marcadores dominantes en fase de repulsión en el mapa genético influye de forma negativa en la ordenación definitiva de los marcadores. Quizás por ese mismo motivo se ve afectada la estima de las probabilidades posteriores de los órdenes, haciéndose extensible a todo el proceso de simulación.

Por la experiencia adquirida sobre el comportamiento de las estimas basadas en sólo tres marcadores, se concluye que, desde el punto de vista práctico, se obtendrían mejores resultados si se estimara un mapa genético preliminar, eliminando alguno de los dos marcadores de la pareja estrechamente ligada D1D2 o D2D1 y diseñar una estrategia para su final incorporación al mapa.

Capítulo 7

Ordenación basada en los marcadores codominantes

Como se ha visto en los últimos capítulos, basar la ordenación de marcadores en las probabilidades posteriores de sus órdenes y en la estimación de sus distancias entre sí, conduce a serios problemas cuando en el mapa genético de un diseño F_2 aparecen marcadores contiguos estrechamente ligados en fase de repulsión.

Por otra parte, Knapp et al (1995) [39], en su estudio sobre la estimación de las fracciones de recombinación entre marcadores dominantes, recomienda crear dos mapas por separado de manera que en cada uno de ellos los marcadores sólo se relacionen en fase acoplamiento. Mester et al. (2003) [53] analizaron que cuando todos los marcadores dominantes están en fase de acoplamiento, la proporción de marcadores dominantes y codominantes no tenía un grave efecto en la calidad de la ordenación. Sin embargo, con marcadores dominantes en fase de repulsión el resultado era bien diferente. Cuanto más alta era la proporción de marcadores en fase de repulsión, menor era la calidad de la ordenación. De ese modo, justificaron un algoritmo ideado en varias fases en el que proponían la división y ordenación de los marcadores en dos grupos, ambos en fase de acoplamiento, según recomendaban Knapp et al (1995) [39], para después integrarlos en un único mapa global. Reconocieron que el resultado podía tener dificultades motivadas por perturbaciones que afectaran a los mapas locales y al global y/o si la densidad de marcadores codominantes compartidos entre los mapas locales era relativamente baja. Posteriormente, Jasen

(2009) [33] apuntó que los enfoques en los que los marcadores codominantes se comparten en dos mapas locales, asumen que los marcadores codominantes carecen de errores. Supuesto que en la práctica no es así.

Según los resultados experimentales vistos hasta el momento, parece adecuado que, durante el proceso de ordenación, se ofrezca mayor peso a las estimas de las fracciones de recombinación que se obtienen entre marcadores codominantes, ya que se ha comprobado que son las más precisas y fiables. Así pues, siguiendo la idea de separar los marcadores en dos grupos, en este capítulo, se diseña una estrategia basada en la construcción de un mapa genético preliminar de marcadores codominantes y la posterior incorporación de los marcadores dominantes uno a uno. Esta nueva estrategia de ordenación se va a acompañar de una forma más eficaz de reproducir las distribuciones posteriores de las fracciones de recombinación entre marcadores.

7.1. Descripción del algoritmo

Como ya se ha comentado en capítulos anteriores, el uso de la distribución normal como distribución pivote para simular de las distribuciones posteriores de las fracciones de recombinación mediante Metropolis-Hastings, no produce estimas precisas en el caso de parejas de marcadores dominantes cercanos, especialmente en las del tipo D1D2 o D2D1.

Con el objetivo de mejorar las estimas posteriores de las fracciones de recombinación, vamos a emplear para simular el programa Winbugs [51] bajo el entorno R [62]. Una de las ventajas más destacables es que la simulación se lleva a cabo de forma simplificada mediante algoritmos optimizados, por lo que las cadenas de Markov finalizan con mayor rapidez. Además, tras el proceso de simulación, se obtiene información detallada que permite valorar cómodamente cuestiones como la convergencia de la cadena. La desventaja es que cualquier modificación sobre el funcionamiento del algoritmo de simulación es limitada.

La modelización propuesta en este capítulo es similar a la empleada en el Capítulo 3. La función de probabilidad de los datos, dada la fracción de recombinación entre los marcadores i y j, $r_{i,j}$, se detalla en (3.12).

Como distribución a priori de $r_{i,j}$ se sigue utilizando una distribución *Beta* truncada en $[0, \varepsilon]$, con $\varepsilon = 0.5$. Es decir,

$$\pi(r_{i,j}) \sim BetaT(\alpha, \beta; \varepsilon),$$

con $\alpha = \beta = 1$

Una vez introducida la modelización en Winbugs [51], el programa produce simulaciones de la distribución posterior para las fracciones de recombinación, utilizando los algoritmos optimizados de Gibbs Sampling y Slice Sampler de WinBugs [51]. Gibbs Sampling se ocupa de saltar de un parámetro a otro, mientras que Slice Sampler proporciona las simulaciones de la distribución posterior de cada uno de los parámetros. Nótese que en este caso, la cadena de simulación de fracciones de recombinación se compone únicamente de candidatos, todos diferentes, que acepta el algoritmo.

En definitiva, tras el proceso de simulación se obtiene una cadena $\{r^{(t)}\}_{t=1}^{nsim}$, de longitud *nsim*, que representa la distribución posterior de las fracciones de recombinación de cada pareja de marcadores independientes. Es decir, $r^{(t)} = \{r_{i,j}^{(t)}\}_{i=1,\ldots,\ m-1;\ j=i+1,\ldots,\ m;\ t=1,\ldots,\ nsim}$, con $r_{i,j}$ la estimación de la fracción de recombinación entre los marcadores i y j.

Imaginemos que se desea estimar un mapa genético con una estructura de m marcadores, de los cuales m_c son codominantes y m_d dominantes (m= m_c+m_d).

Para cada marcador dominante, M_k, se puede calcular:

$$s_k = \sum_l sd_{k,l}, \forall M_l \text{ codominantes}$$

donde $sd_{k,l}$ representa la desviación típica de la distribución posterior de la fracción de recombinación entre el marcador dominante M_k y el marcador codominante M_l.

Para dar mayor fiabilidad a la construcción del mapa y entendiendo que una mayor precisión va vinculada a una mayor fiabilidad, de menor a mayor s_k, se confecciona un "ranking" de los marcadores dominantes. Es decir, el marcador dominante con menor s_k es el primero en el "ranking" y el marcador dominante con mayor s_k es el último en el "ranking".

Tras haber obtenido $\{r^{(t)}\}_{t=1}^{nsim}$ y el "ranking" de los marcadores dominantes, comienza la fase de ordenación de los m marcadores en cada iteración. Nótese que el "ranking" de los marcadores dominantes es independiente de la iteración, t, ya que está basado en la desviación típica de la distribución posterior de cada fracción de recombinación,

En cada iteración, t, se construye un mapa preliminar únicamente con los m_c marcadores codominantes, basado en sus distancias mínimas ($SARF$), mediante el algoritmo explicado en el Capítulo 3. Por lo tanto, en cada iteración, el mapa preliminar resultante, puede ser diferente, aunque están involucrados los mismos m_c marcadores codominantes.

Seguidamente, se incorporan al mapa preliminar, uno a uno, los m_d marcadores dominantes según el "ranking" confeccionado.

La ubicación de cada marcador dominante se establece según los siguientes pasos:

1. Se selecciona el marcador codominante del mapa preliminar, M_l^*, más próximo al candidato dominante, M_k^*, (según el "ranking"), en términos de fracciones de recombinación, que servirá de referente para ubicar al dominante en el mapa. Es decir,

$$M_l^* = \{M_j \text{ codominante } / \ r_{k,l} = min\{\{r_{k,j}^{(t)}\} \forall \ M_j \text{ codominante}\}\}$$

2. Se considera el marcador codominante del mapa preliminar, M_l^{**}, más próximo al referente M_l^*, en términos de fracciones de recombinación.

Las posibilidades a tener en cuenta son:

a) M_l^{**} es anterior a M_l^*, esquemáticamente M_l^{**}_____ M_l^*

b) M_l^* es anterior a M_l^{**}, esquemáticamente M_l^*_____ M_l^{**}

En ambos casos, el candidato, M_k^*, se ubica a la izquierda o bien a la derecha del marcador referente M_l^*. Es decir:

a) M_l^{**}_____ M_l^*_

. $\uparrow_{M_k^*}$ $\uparrow_{M_k^*}$

b) _ M_l^*_____M_l^{**}

. $\uparrow_{M_k^*}$ $\uparrow_{M_k^*}$

Esta localización se establece en función de si la fracción de recombinación estimada entre los dos marcadores codominantes M_l^* y M_l^{**} es mayor o menor que la fracción de recombinación estimada entre M_k^* y M_l^{**}. La localización del marcador dominante candidato a entrar en el mapa, no tiene conflicto con los dominantes ya introducidos (respecto a ese referente) porque, de todos ellos, se conoce a qué lado fueron colocados y además que habían sido más próximos al referente que el propio candidato.

Los pasos del 1 y 2 se repiten hasta que se introducen todos los marcadores en el mapa. A continuación, para comprobar posibles mejoras sobre

el orden obtenido, se proponen variaciones sobre la localización de marcadores contiguos como son permutaciones de parejas de marcadores, tripletas y cuatripletas en busca de mapas con mínimo *SARF*.

Una vez finalizado el proceso ordenación, se obtiene la cadena de ordenaciones, $\{O^{(t)}\}_{t=1}^{nsim}$. Del mismo modo que sucedía en el Capítulo 3, aquella ordenación de marcadores que se repite más veces en la cadena de ordenaciones, es considerada la estimación más probable del mapa y su probabilidad asociada coincide con la proporción de veces que aparece en la cadena. Los siguientes modelos tendrán asociadas probabilidades menores. A continuación se lleva a cabo la conversión de las fracciones de recombinación a distancias multipunto dado el orden obtenido por los marcadores, en cada iteración, obteniendo así, la distribución posterior de distancias multipunto entre cada pareja de marcadores, i y j, $\{d^{(t)}\}_{t=1}^{nsim}$.

7.2. Resultados y conclusiones

A continuación, se ofrece una comparativa de los resultados obtenidos con la estrategia utilizada en el Capítulo 3, basada en la ordenación conjunta de los marcadores dominantes y codominantes con mínimo *SARF*, con la nueva propuesta, en la que se ordenan primero los marcadores codominantes y se introducen posteriormente los marcadores dominantes.

Para comparar los resultados de ambos algoritmos se ha diseñado una nueva población F_2 definida por 8 marcadores, cuyo mapa genético aparece en la Figura 1.11. El tipo considerado para cada marcador se especifica en el siguiente esquema:

$$M_1(\mathrm{C})\_M_2(\mathrm{D2})\_M_3(\mathrm{D1})\_M_4(\mathrm{D1})\_M_5(\mathrm{C})\_M_6(\mathrm{D1})\_M_7(\mathrm{D2})\_M_8(\mathrm{C})$$

El diseño de este nuevo mapa genético, para ensayar la nueva metodología, quedará justificada al inicio del Capítulo 9. Con el objetivo de no basar nuestras conclusiones en el resultado respecto una única muestra, de esta población se extraen 5 muestras con un tamaño de 200 individuos. Para cada una de ellas, se simula la distribución posterior de las fracciones de recombinación entre cada pareja de marcadores utilizando Winbugs [51]. La longitud de las cadenas se ha fijado en nsim=3000 desechando las 1000 primeras (burning=1000). Tras comprobar la convergencia y estabilidad de la cadena, se

aplican las dos estrategias de ordenación, que llamamos de forma simplificada "Orden (Cap.3)" y "Orden (Cap.7)". En los siguientes cuadros se resumen las 10 estimaciones más probables del mapa genético (modelos) junto con la probabilidad obtenida con cada estrategia y se señala en negrita el modelo que coincide con el verdadero mapa genético.

Muestra 1	Orden (Cap.3)	Prob	Orden (Cap.7)	Prob
Modelo 1	1 2 6 5 4 3 7 8	0.0790	**1 2 3 4 5 6 7 8**	**0.1820**
Modelo 2	**1 2 3 4 5 6 7 8**	**0.0730**	1 2 3 6 4 5 7 8	0.1120
Modelo 3	1 2 3 4 6 5 7 8	0.0395	1 2 3 5 4 6 7 8	0.1110
Modelo 4	1 2 6 5 3 4 7 8	0.0285	1 2 3 6 5 4 7 8	0.0530
Modelo 5	1 3 2 6 5 4 7 8	0.0230	1 2 3 4 5 7 6 8	0.0430
Modelo 6	1 2 6 4 5 3 7 8	0.0230	1 2 3 4 6 5 7 8	0.0385
Modelo 7	1 3 2 4 5 6 7 8	0.0225	1 2 3 5 6 4 7 8	0.0315
Modelo 8	1 2 4 6 5 3 7 8	0.0210	1 2 3 7 6 4 5 8	0.0310
Modelo 9	1 2 4 3 5 6 7 8	0.0210	1 2 3 5 4 7 6 8	0.0300
Modelo 10	1 2 6 4 3 5 7 8	0.0205	1 2 7 3 4 5 6 8	0.0295

Muestra 2	Orden (Cap.3)	Prob	Orden (Cap.7)	Prob
Modelo 1	1 2 3 4 6 5 7 8	0.1045	1 2 3 5 6 4 7 8	0.1115
Modelo 2	1 2 3 7 5 6 4 8	0.0690	1 2 3 5 6 7 4 8	0.0820
Modelo 3	**1 2 3 4 5 6 7 8**	**0.0470**	1 2 3 7 5 6 4 8	0.0705
Modelo 4	1 2 6 5 4 3 7 8	0.0465	**1 2 3 4 5 6 7 8**	**0.0600**
Modelo 5	1 2 3 7 6 5 4 8	0.0395	1 2 3 7 6 5 4 8	0.0540
Modelo 6	1 2 6 5 7 3 4 8	0.0355	1 2 5 6 4 7 3 8	0.0510
Modelo 7	1 2 4 3 7 5 6 8	0.0330	1 2 3 6 5 4 7 8	0.0495
Modelo 8	1 2 3 4 7 5 6 8	0.0320	1 2 3 4 6 5 7 8	0.0410
Modelo 9	1 2 4 3 6 5 7 8	0.0290	1 2 5 6 7 4 3 8	0.0375
Modelo 10	1 2 3 7 4 5 6 8	0.0280	1 2 3 6 5 7 4 8	0.0335

Muestra 3	Orden (Cap.3)	Prob	Orden (Cap.7)	Prob
Modelo 1	**1 2 3 4 5 6 7 8**	**0.2400**	**1 2 3 4 5 6 7 8**	**0.2650**
Modelo 2	1 2 3 4 5 7 6 8	0.0980	1 2 3 5 4 6 7 8	0.1785
Modelo 3	1 2 3 4 6 5 7 8	0.0610	1 2 3 4 5 7 6 8	0.0885
Modelo 4	1 2 3 5 4 7 6 8	0.0590	1 2 3 5 4 7 6 8	0.0625
Modelo 5	1 2 3 5 4 6 7 8	0.0510	1 8 7 6 5 4 3 2	0.0345
Modelo 6	1 2 3 4 5 6 8 7	0.0470	1 8 7 6 4 5 3 2	0.0315
Modelo 7	1 2 3 6 5 4 7 8	0.0460	1 2 5 4 6 7 3 8	0.0240
Modelo 8	1 2 3 4 7 5 6 8	0.0450	1 2 3 6 4 5 7 8	0.0205
Modelo 9	1 2 3 6 4 5 7 8	0.0250	1 2 3 5 6 4 7 8	0.0160
Modelo 10	1 2 3 5 6 4 7 8	0.0240	1 2 5 4 6 3 7 8	0.0155

Muestra 4	Orden (Cap.3)	Prob	Orden (Cap.7)	Prob
Modelo 1	**1 2 3 4 5 6 7 8**	**0.1675**	**1 2 3 4 5 6 7 8**	**0.1905**
Modelo 2	1 2 3 4 6 5 7 8	0.1130	1 2 3 4 6 5 7 8	0.1650
Modelo 3	1 2 6 5 3 4 7 8	0.0845	1 2 3 6 4 5 7 8	0.1020
Modelo 4	1 2 4 3 5 6 7 8	0.0635	1 2 3 5 6 4 7 8	0.0420
Modelo 5	1 2 4 3 6 5 7 8	0.0525	1 8 7 6 5 4 3 2	0.0360
Modelo 6	1 3 4 2 6 5 7 8	0.0390	1 2 3 6 5 4 7 8	0.0340
Modelo 7	1 2 3 5 6 4 7 8	0.0295	1 2 4 3 6 5 7 8	0.0255
Modelo 8	1 2 3 6 5 4 7 8	0.0260	1 2 3 5 4 6 7 8	0.0220
Modelo 9	1 2 6 3 4 5 7 8	0.0210	1 2 3 7 4 6 5 8	0.0215
Modelo 10	1 2 6 8 7 5 4 3	0.0180	1 8 7 5 6 4 3 2	0.0180

Muestra 5	Orden (Cap.3)	Prob	Orden (Cap.7)	Prob
Modelo 1	1 2 6 5 4 3 7 8	0.0750	1 2 3 6 4 5 7 8	0.1875
Modelo 2	**1 2 3 4 5 6 7 8**	**0.0645**	**1 2 3 4 5 6 7 8**	**0.1810**
Modelo 3	1 2 3 4 6 5 7 8	0.0540	1 2 3 5 4 6 7 8	0.0575
Modelo 4	1 2 3 6 4 5 7 8	0.0415	1 2 3 4 6 5 7 8	0.0480
Modelo 5	1 2 4 5 6 3 7 8	0.0390	1 2 3 6 5 4 7 8	0.0455
Modelo 6	1 2 6 4 5 3 7 8	0.0385	1 2 3 4 5 7 6 8	0.0410
Modelo 7	1 2 3 6 5 4 7 8	0.0345	1 2 4 5 6 7 3 8	0.0320
Modelo 8	1 2 3 5 4 6 7 8	0.0330	1 2 3 7 6 4 5 8	0.0265
Modelo 9	1 2 3 5 6 4 7 8	0.0215	1 2 5 4 6 7 3 8	0.0225
Modelo 10	1 2 6 3 4 5 7 8	0.0215	1 2 6 3 4 5 7 8	0.0200

En función de los resultados, parece que se obtiene una pequeña mejora con el nuevo método de ordenación. En todas las muestras, es más alta la proporción de veces que se estima el modelo correcto utilizando la nueva estrategia que utilizando la estrategia del Capítulo 3. Por otra parte, para cada muestra, los dos algoritmos clasifican el modelo correcto de forma similar. Por ejemplo, se puede ver que si en la primera muestra se obtiene mejor resultado con el nuevo método, ya que el modelo correcto aparece en primera posición, en la segunda muestra se obtiene mejor clasificación con el viejo método, aunque en este caso los modelos más probables no coinciden con el correcto. En definitiva, aunque los resultados anteriores muestran una tímida mejora, la probabilidad sigue siendo baja y por lo tanto, la metodología diseñada no cumple con las expectativas.

Capítulo 8

Ordenación basada en la información de todos los marcadores

Según los resultados obtenidos en el trascurso de los capítulos anteriores, se hace evidente la necesidad de desarrollar un procedimiento de estimación y ordenación que involucre la información que proporcionan todos los marcadores conjuntamente, tanto codominantes como dominantes. Así pues, a continuación se desarrolla un nuevo algoritmo de ordenación más exhaustivo que los ensayados hasta el momento.

8.1. Descripción del algoritmo

Se diseña un proceso de estimación de mapas genéticos que consta de 2 pasos. En primer lugar, simulación de la distribución posterior de las fracciones de recombinación y en segundo lugar, ordenación de los marcadores conjuntamente con la corrección de las distancias multipunto.

La modelización se lleva a cabo en idénticos términos que en el Capítulo 3 y 7. De nuevo, simular de la distribución posterior de $\mathbf{r^*} = (r_{1,2}, r_{1,3}, ..., r_{m-1,m},)$ es equivalente a simular independientemente de la distribución posterior de cada fracción de recombinación, $r_{i,j}$, dado el vector de frecuencias de recombinación $\mathbf{R_{i,j}}$ que determina cada pareja distinta de marcadores M_i y M_j. Se simula a través de la combinación entre los algoritmos Gibbs Sampling y

Slice Sampler implementada en el programa OpenBugs [75] y se obtiene la cadena de simulación $\{r^{(t)}\}_{t=1}^{nsim}$, que representa la distribución posterior de las fracciones de recombinación para cada pareja de marcadores. Nótese que la cadena consta de *nsim* filas o iteraciones y $m(m\text{-}1)/2$ columnas o parejas de marcadores, siendo m el número de marcadores del mapa genético.

A continuación se lleva a cabo la ordenación de los marcadores en cada iteración, t, de la cadena, a través del siguiente **algoritmo de ordenación**:

Se seleccionan los dos marcadores cuya fracción de recombinación se determina como más precisa, en función de la desviación típica de la distribución posterior de fracciones de recombinación,

$$mapa = \{(M_k, M_l)/\ sd_{k,l} = min\{\{sd_{i,j}\}_{i=1,\ldots,m-1;j=i+1,\ldots,m}\}\}$$

donde $sd_{i,j}$ recordemos que es la desviación típica de la distribución posterior entre el marcador dominante M_i y el codominante M_j, y con ellos se inicia el *mapa*, insertando el resto de marcadores según los pasos que se indica a continuación:

1. De todos los marcadores restantes, se consideran *candidatos* a entrar en el *mapa* sólo aquellos más cercanos a los que ya están incluidos en el mapa. Es decir, que tienen una fracción de recombinación estimada con los que ya están incluidos en el *mapa*, menor o igual a 0.3 cM.

$$candidatos = \{M_i\ \notin\ mapa/\ r_{i,j} < 0.3,\ \forall\ M_j\ \in mapa\}$$

Y de ellos, se selecciona aquel cuyas fracciones de recombinación, con el resto de marcadores incluidos en el mapa, son las más precisas. Esta medida de precisión se calcula como la suma de las inversas de las desviaciones típicas de la distribución posterior de las fracciones de recombinación, del candidato con cada marcador incluido en el *mapa* y se elige aquel que obtiene mayor suma. Es decir, para cada $M_i\ \in\ candidatos$ se calcula

$$s_i = \sum_j \frac{1}{sd_{i,j}},\ \forall M_j\ \in\ mapa$$

y se selecciona $M_c = \{M_i\ \in\ candidatos | s_c = max_i\{s_i\}\}$

Este criterio de selección viene motivado por un estudio observacional, en el que se aprecia que la estimación más precisa, en términos de desviaciones típicas posteriores, se produce para fracciones recombinación muy pequeñas o muy grandes. En las Figuras 8.1 y 8.2 se muestran dos ejemplos de este patrón. Uno para una muestra que procede de una población Retrocruce y otro para una muestra que procede de una población F_2 con algunos marcadores dominantes. En la leyenda aparece la distinción del tipo de marcadores a la que corresponde cada estimación. Ambas poblaciones tienen el mismo mapa genético, representado en la Figura 1.12 y que se estudiará más a fondo en el Capítulo 10.

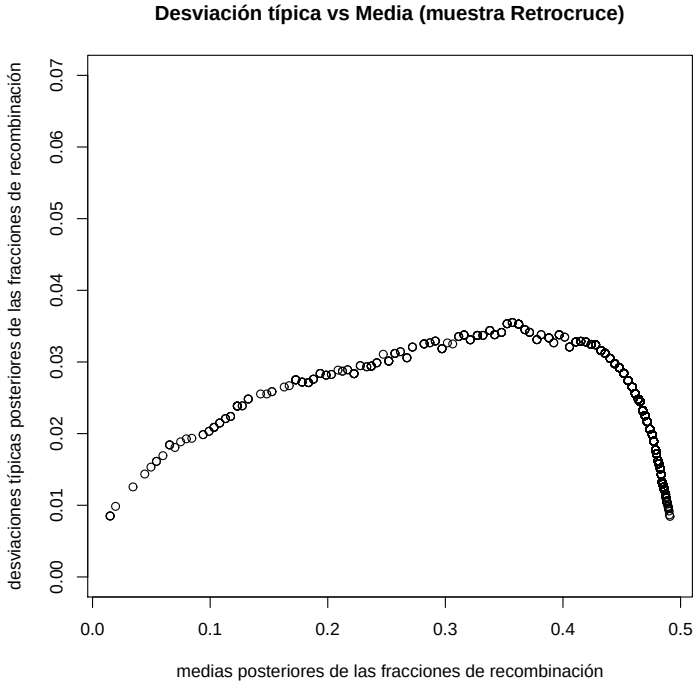


Figura 8.1: *Relación entre las medias y las desviaciones típicas posteriores de las fracciones de recombinación para una muestra que proviene de una población Retrocruce.*

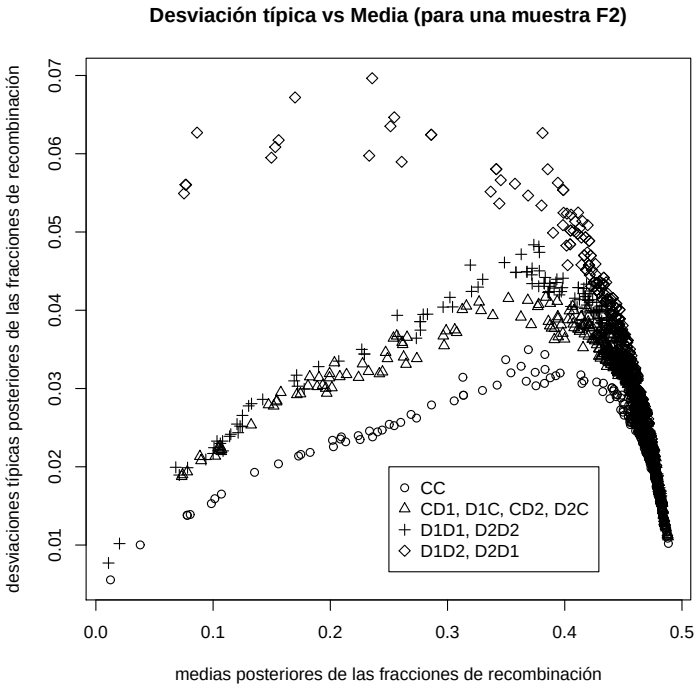


Figura 8.2: *Relación entre las medias y las desviaciones típicas posteriores de las fracciones de recombinación para una muestra que proviene de una población F_2 con marcadores algunos dominantes.*

2. Para determinar la posición definitiva en la que M_c se ubica en el *mapa*, se consideran todas las posibles inserciones del candidato, incluyéndolo tanto al principio, como al final y entre los marcadores ya colocados. Para cada una de estas posibles ordenaciones se dispone de dos estimas de las fracciones de recombinación entre parejas de marcadores: Por un lado, como siempre, las fracciones de recombinación estimadas inicialmente, $\{O^{(t)}\}_{t=1}^{nsim}$, por simulación de la distribución posterior, mediante OpenBugs [75], que, como ya sabemos, son independientes del orden de los marcadores y que podríamos llamar fracciones de recombinación "iniciales". Por otro lado, las obtenidas como inversa de la función de mapeo Haldane sobre las distancias multipunto entre marcadores, calculadas en función de cada una de las ordenaciones propuestas, que podríamos llamar fracciones de recombinación "multipunto". Es decir, ca-

da fracción de recombinación "multipunto", se calcula como $F^{-1}(d_{i,j})$, para la distancia multipunto, $d_{i,j}$ entre M_i y M_j bajo un determinado orden. Se escoge la ordenación en la que más se parecen las fracciones de recombinación "iniciales" y "multipunto". Esta selección se realiza en términos del criterio de información bayesiano BIC. Esa ordenación pasará a considerarse el *mapa*. Si tras este proceso, algún marcador candidato no consigue ser ubicado en ninguna de las posibles posiciones, por ejemplo por obtener distancias multipunto negativas, se reserva su entrada y se selecciona un nuevo candidato, de entre todos los que faltan por ubicar, como se explica en el paso anterior.

Los pasos 1 y 2 se repiten hasta obtener el "mapa completo" que contiene a todos los marcadores ordenados de manera preliminar.

3. Para finalizar, se considera una "ventana" que abarca a cuatro marcadores, que recorre el "mapa completo" de principio a fin. Sobre los marcadores contenidos en esa "ventana" se proponen todas sus posibles permutaciones, dando lugar a variantes del "mapa completo" y se examina si la variante mejora el mapa en términos del criterio de información bayesiano BIC, basado en la función de verosimilitud evaluada en su máximo. Si es así, la variante se guarda como "mapa completo". Una vez terminado el proceso se obtiene el "mapa definitivo".

Nótese que en cada iteración, los dos marcadores que se utilizan como "corazón" del mapa y respecto a los que se van insertando todos los demás, son los mismos en cada iteración, puesto que, como se ha explicado anteriormente, se seleccionan en función de la desviación típica de la distribución posterior de las fracciones de recombinación, que es la misma en cada iteración. Sin embargo, los candidatos a entrar en el mapa pueden ser propuestos en un orden de entrada diferente y el mapa definitivo resultante puede ser distinto, porque en ambos pasos la elección se establece en función de las fracciones de recombinación estimadas para cada pareja de marcadores, que son distintas en cada iteración.

Este procedimiento es más exhaustivo y laborioso que los ensayados en capítulos anteriores ya que, en cada iteración, cada modelo se obtiene tras la evaluación de múltiples ordenaciones distintas.

Finalizado el proceso de ordenación se obtiene conjuntamente una cadena de mapas, $\{O^{(t)}\}_{t=1}^{nsim}$ y una cadena de distancias multipunto, $\{d^{(t)}\}_{t=1}^{nsim}$, que representa la distribución posterior de las distancias entre marcadores. Como siempre, aquel modelo que se repite más veces en la cadena se considera la es-

timación más probable del mapa genético y su probabilidad asociada coincide con la proporción de veces que aparece en la cadena.

8.2. Ilustración del método

A continuación, se evalúa la bondad del método utilizando dos poblaciones diferentes. En primer lugar, se ensaya el método sobre las muestras del capítulo anterior para señalar posibles mejoras. En segundo lugar, se ofrece una comparación de los resultados obtenidos con la estrategia seguida en el Capítulo 3 y la actual metodología, sobre 5 muestras procedentes de la población definida por m= 20 marcadores, no todos codominantes, con mapa genético más denso. En ambos casos, se ha utilizado un tamaño muestral de n= 200 individuos.

Como se puede observar en las siguientes tablas, el procedimiento descrito en este capítulo obtiene mejores resultados que los vistos hasta ahora. Para todas las muestras el modelo correcto aparece una proporción de veces superior al de los otros métodos. Tan sólo la muestra 2 se manifiesta especialmente problemática ya que el modelo más probable no coincide con el correcto con ninguno de los 3 métodos, aunque con el último empleado, el modelo más probable tan sólo difiere del correcto en la ubicación de un marcador. Es posible que en este caso, la desviación provenga de la propia aleatoriedad del muestreo. Además, el primer modelo muestra una probabilidad muy superior en relación a los otros métodos y dentro del método, también al resto de modelos alternativos. Cabe señalar que el motivo de esta alta probabilidad no se debe a atribuir a que el orden de entrada de los marcadores, en cada iteración, pueda ser poco variado, pues el algoritmo utilizado en el Capítulo 7, proponía siempre los marcadores en el mismo orden de entrada (los dominantes) y sin embargo, se obtenían más modelos alternativos, hecho que provocaba una menor probabilidad asignada al modelo más probable.

M.1	Orden (Cap.3)	Prob	Orden (Cap.7)	Prob	Orden (Cap.8)	Prob
Mod1	1 2 6 5 4 3 7 8	0.079	**1 2 3 4 5 6 7 8**	**0.182**	**1 2 3 4 5 6 7 8**	**0.814**
Mod2	**1 2 3 4 5 6 7 8**	**0.073**	1 2 3 6 4 5 7 8	0.112	1 2 3 4 5 7 6 8	0.0605
Mod3	1 2 3 4 6 5 7 8	0.040	1 2 3 5 4 6 7 8	0.111	1 2 3 5 4 6 7 8	0.025
Mod4	1 2 6 5 3 4 7 8	0.029	1 2 3 6 5 4 7 8	0.053	1 3 7 4 5 6 2 8	0.012
Mod5	1 3 2 6 5 4 7 8	0.023	1 2 3 4 5 7 6 8	0.043	1 8 7 6 5 4 3 2	0.012
Mod6	1 2 6 4 5 3 7 8	0.023	1 2 3 4 6 5 7 8	0.039	1 8 7 3 4 5 6 2	0.008
Mod7	1 3 2 4 5 6 7 8	0.023	1 2 3 5 6 4 7 8	0.032	1 2 3 5 4 7 6 8	0.007
Mod8	1 2 4 6 5 3 7 8	0.021	1 2 3 7 6 4 5 8	0.031	1 8 7 3 5 4 6 2	0.006
Mod9	1 2 4 3 5 6 7 8	0.021	1 2 3 5 4 7 6 8	0.030	1 2 6 5 4 7 3 8	0.004
Mod10	1 2 6 4 3 5 7 8	0.021	1 2 7 3 4 5 6 8	0.030	1 3 7 5 4 6 2 8	0.004

M.2	Orden (Cap.3)	Prob	Orden (Cap.7)	Prob	Orden (Cap.8)	Prob
Mod1	1 2 3 4 6 5 7 8	0.105	1 2 3 5 6 4 7 8	0.112	1 2 3 7 4 5 6 8	0.403
Mod2	1 2 3 7 5 6 4 8	0.069	1 2 3 5 6 7 4 8	0.082	**1 2 3 4 5 6 7 8**	**0.137**
Mod3	**1 2 3 4 5 6 7 8**	**0.047**	1 2 3 7 5 6 4 8	0.070	1 2 3 4 7 5 6 8	0.128
Mod4	1 2 6 5 4 3 7 8	0.047	**1 2 3 4 5 6 7 8**	**0.060**	1 2 8 7 5 6 4 3	0.034
Mod5	1 2 3 7 6 5 4 8	0.040	1 2 3 7 6 5 4 8	0.054	1 2 3 5 6 7 4 8	0.032
Mod6	1 2 6 5 7 3 4 8	0.036	1 2 5 6 4 7 3 8	0.051	1 2 3 7 4 6 5 8	0.029
Mod7	1 2 4 3 7 5 6 8	0.033	1 2 3 6 5 4 7 8	0.050	1 2 3 4 7 6 5 8	0.025
Mod8	1 2 3 4 7 5 6 8	0.032	1 2 3 4 6 5 7 8	0.041	1 3 7 5 6 4 2 8	0.024
Mod9	1 2 4 3 6 5 7 8	0.029	1 2 5 6 7 4 3 8	0.038	1 2 3 7 5 6 4 8	0.023
Mod10	1 2 3 7 4 5 6 8	0.028	1 2 3 6 5 7 4 8	0.034	1 2 3 5 6 4 7 8	0.021

M.3	Orden (Cap.3)	Prob	Orden (Cap.7)	Prob	Orden (Cap.8)	Prob
Mod1	**1 2 3 4 5 6 7 8**	**0.240**	**1 2 3 4 5 6 7 8**	**0.265**	**1 2 3 4 5 6 7 8**	**0.707**
Mod2	1 2 3 4 5 7 6 8	0.098	1 2 3 5 4 6 7 8	0.179	1 2 3 4 5 7 6 8	0.151
Mod3	1 2 3 4 6 5 7 8	0.061	1 2 3 4 5 7 6 8	0.089	1 2 3 5 4 6 7 8	0.022
Mod4	1 2 3 5 4 7 6 8	0.059	1 2 3 5 4 7 6 8	0.063	1 2 8 7 6 5 4 3	0.016
Mod5	1 2 3 5 4 6 7 8	0.051	1 8 7 6 5 4 3 2	0.035	1 3 7 4 5 6 2 8	0.011
Mod6	1 2 3 4 5 6 8 7	0.047	1 8 7 6 4 5 3 2	0.032	1 2 3 5 4 7 6 8	0.009
Mod7	1 2 3 6 5 4 7 8	0.046	1 2 5 4 6 7 3 8	0.024	1 2 8 6 7 5 4 3	0.009
Mod8	1 2 3 4 7 5 6 8	0.045	1 2 3 6 4 5 7 8	0.021	1 2 3 4 7 5 6 8	0.008
Mod9	1 2 3 6 4 5 7 8	0.025	1 2 3 5 6 4 7 8	0.016	1 3 5 4 7 6 2 8	0.008
Mod10	1 2 3 5 6 4 7 8	0.024	1 2 5 4 6 3 7 8	0.016	1 8 7 6 5 4 3 2	0.007

M.4	Orden (Cap.3)	Prob	Orden (Cap.7)	Prob	Orden (Cap.8)	Prob
Mod1	**1 2 3 4 5 6 7 8**	**0.169**	**1 2 3 4 5 6 7 8**	**0.191**	**1 2 3 4 5 6 7 8**	**0.913**
Mod2	1 2 3 4 6 5 7 8	0.113	1 2 3 4 6 5 7 8	0.165	1 2 3 4 6 5 7 8	0.063
Mod3	1 2 6 5 3 4 7 8	0.085	1 2 3 6 4 5 7 8	0.102	1 2 3 4 5 7 6 8	0.007
Mod4	1 2 4 3 5 6 7 8	0.064	1 2 3 5 6 4 7 8	0.042	1 3 2 4 5 6 7 8	0.003
Mod5	1 2 4 3 6 5 7 8	0.053	1 8 7 6 5 4 3 2	0.036	1 2 3 7 4 5 6 8	0.003
Mod6	1 3 4 2 6 5 7 8	0.039	1 2 3 6 5 4 7 8	0.034	1 3 2 4 6 5 7 8	0.002
Mod7	1 2 3 5 6 4 7 8	0.030	1 2 4 3 6 5 7 8	0.026	1 3 7 4 5 6 2 8	0.002
Mod8	1 2 3 6 5 4 7 8	0.026	1 2 3 5 4 6 7 8	0.022	1 2 8 7 6 5 4 3	0.001
Mod9	1 2 6 3 4 5 7 8	0.021	1 2 3 7 4 6 5 8	0.023	1 2 3 4 7 5 6 8	0.001
Mod10	1 2 6 8 7 5 4 3	0.018	1 8 7 5 6 4 3 2	0.018	1 2 4 3 5 6 7 8	0.001

M.5	Orden (Cap.3)	Prob	Orden (Cap.7)	Prob	Orden (Cap.8)	Prob
Mod1	1 2 6 5 4 3 7 8	0.075	1 2 3 6 4 5 7 8	0.188	**1 2 3 4 5 6 7 8**	**0.788**
Mod2	**1 2 3 4 5 6 7 8**	**0.065**	**1 2 3 4 5 6 7 8**	**0.1810**	1 2 3 5 4 6 7 8	0.067
Mod3	1 2 3 4 6 5 7 8	0.054	1 2 3 5 4 6 7 8	0.058	1 2 3 4 5 7 6 8	0.043
Mod4	1 2 3 6 4 5 7 8	0.042	1 2 3 4 6 5 7 8	0.048	1 2 3 6 4 5 7 8	0.012
Mod5	1 2 4 5 6 3 7 8	0.039	1 2 3 6 5 4 7 8	0.046	1 2 6 4 5 7 3 8	0.007
Mod6	1 2 6 4 5 3 7 8	0.039	1 2 3 4 5 7 6 8	0.041	1 2 6 5 4 7 3 8	0.006
Mod7	1 2 3 6 5 4 7 8	0.035	1 2 4 5 6 7 3 8	0.032	1 3 7 4 5 6 2 8	0.006
Mod8	1 2 3 5 4 6 7 8	0.033	1 2 3 7 6 4 5 8	0.027	1 2 3 6 5 4 7 8	0.005
Mod9	1 2 3 5 6 4 7 8	0.022	1 2 5 4 6 7 3 8	0.023	1 2 3 5 4 7 6 8	0.005
Mod10	1 2 6 3 4 5 7 8	0.022	1 2 6 3 4 5 7 8	0.020	1 8 7 6 5 4 3 2	0.004

En las siguientes figuras se resumen los cinco modelos más probables estimados con la metodología empleada en el Capítulo 3 y los estimados con la actual metodología, para cinco muestras procedentes de la población F_2 con mapa genético más denso, definida por m= 20 marcadores, no todos codominantes.

Como se puede observar, con el procedimiento del Capítulo 3, ninguna de las muestras obtiene el modelo correcto entre los cinco más probables. Es más, de la baja probabilidad obtenida para los modelos más probables, concluye sobre la baja fiabilidad del método en el caso de la participación de marcadores estrechamente ligados en fase de repulsión.

Como se puede apreciar para dos de las muestras, la metodología consigue reproducir el modelo correcto como el más probable. Para otras dos muestras, el modelo correcto aparece como segundo modelo más probable, difiriendo

del que tiene mejor probabilidad en una permutación de dos marcadores muy próximos. Tan sólo en una de las muestras el modelo correcto aparece en cuarta posición, por detrás de otras alternativas con permutaciones de marcadores contiguos.

Por otra parte, parece que la metodología consigue estimar de una forma más apropiada las distancias entre marcadores contiguos, de manera que la longitud del mapa resultante es muy próxima a la real.

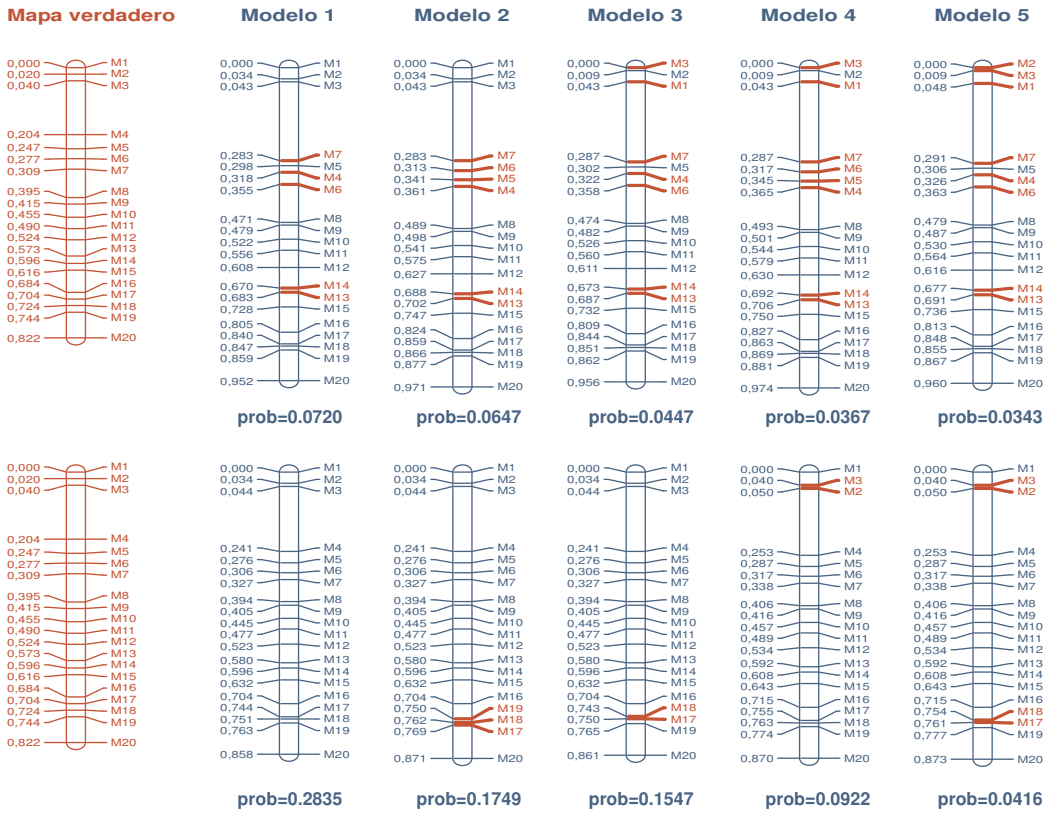


Figura 8.3: *Mapa verdadero de una población F_2, con mapa más denso, con 20 marcadores, no todos codominantes, junto con la estima de los 5 modelos bayesianos más probables respecto una muestra con 200 individuos (muestra 1). Parte superior, según la metodología descrita en el Capítulo 3. Parte inferior, según la metodología descrita en el Capítulo 8.*

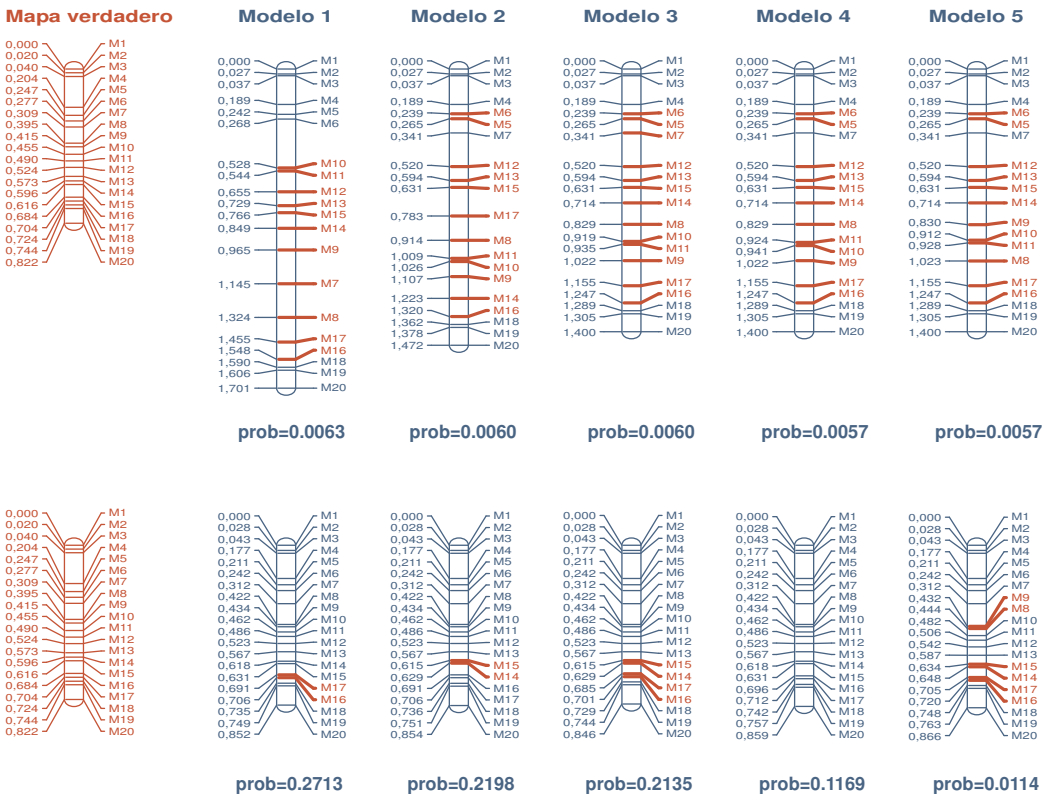


Figura 8.4: *Mapa verdadero de una población F_2, con mapa más denso, con 20 marcadores, no todos codominantes, junto con la estima de los 5 modelos bayesianos más probables respecto una muestra con 200 individuos (muestra 2).Parte superior, según la metodología descrita en el Capítulo 3. Parte inferior, según la metodología descrita en el Capítulo 8.*

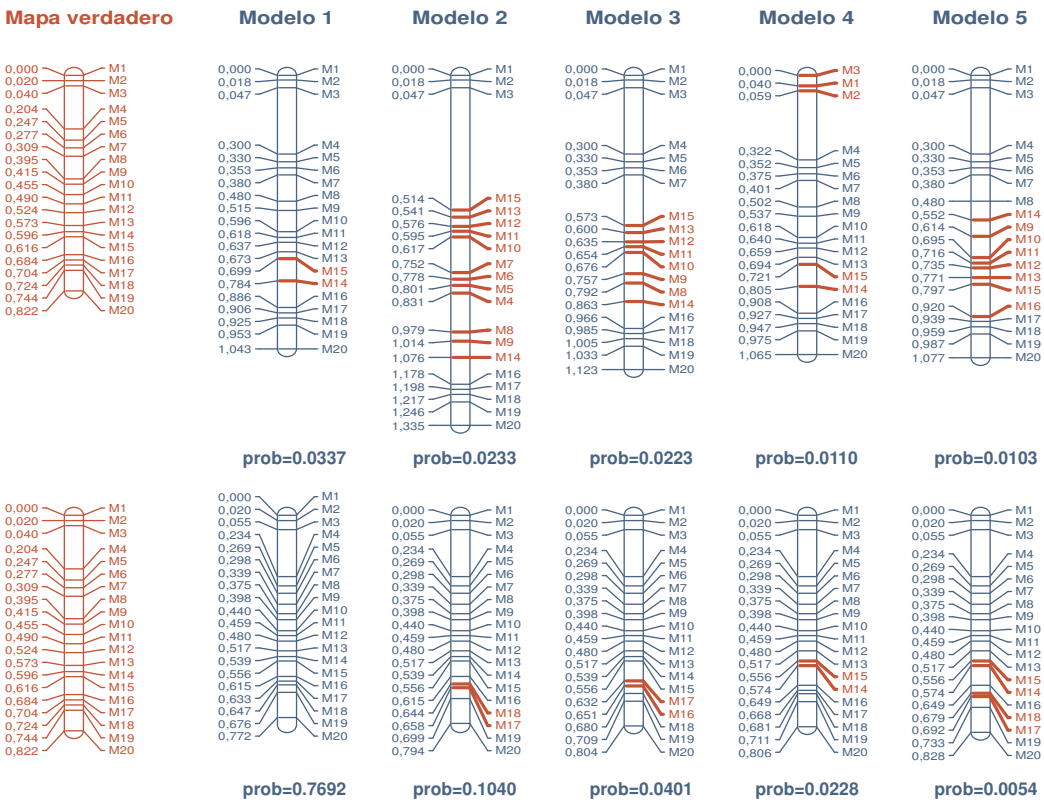


Figura 8.5: *Mapa verdadero de una población F_2, con mapa más denso, con 20 marcadores, no todos codominantes, junto con la estima de los 5 modelos bayesianos más probables respecto una muestra con 200 individuos (muestra 3).Parte superior, según la metodología descrita en el Capítulo 3. Parte inferior, según la metodología descrita en el Capítulo 8.*

Figura 8.6: *Mapa verdadero de una población F_2, con mapa más denso, con 20 marcadores, no todos codominantes, junto con la estima de los 5 modelos bayesianos más probables respecto una muestra con 200 individuos (muestra 4).Parte superior, según la metodología descrita en el Capítulo 3. Parte inferior, según la metodología descrita en el Capítulo 8.*

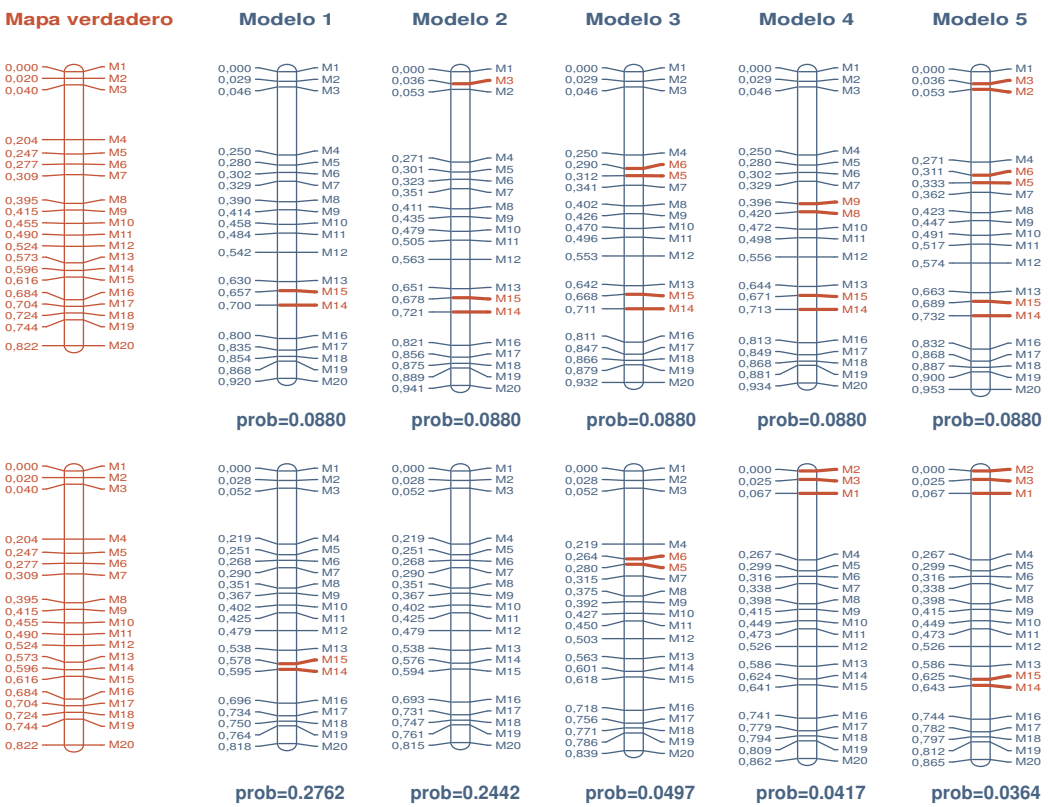


Figura 8.7: *Mapa verdadero de una población F_2, con mapa más denso, con 20 marcadores, no todos codominantes, junto con la estima de los 5 modelos bayesianos más probables respecto una muestra con 200 individuos (muestra 5).Parte superior, según la metodología descrita en el Capítulo 3. Parte inferior, según la metodología descrita en el Capítulo 8.*

8.3. Discusión

La metodología de estimación de las fracciones de recombinación entre las distintas parejas de marcadores se realiza mediante el programa OpenBugs [75] bajo el entorno R, de manera que la simulación se lleva a cabo de forma simplificada mediante algoritmos optimizados, por lo que las cadenas de Markov finalizan con mayor rapidez. Además, tras el proceso de simulación se obtiene información detallada que permite valorar cómodamente cuestiones como la convergencia de la cadena. Parece que la optimización sobre la estimación de las fracciones de recombinación entre parejas de marcadores también repercute de forma positiva en la justa obtención de las distancias multipunto, dando lugar a longitudes de mapas muy próximas a la realidad.

El algoritmo de ordenación ideado es más exhaustivo y laborioso que los ensayados en capítulos anteriores, ofreciendo resultados muy satisfactorios en situaciones problemáticas que recogen todo tipo de dificultades, como la de contener marcadores adyacentes fuertemente ligados en fase de repulsión.

Es probable, que el éxito del algoritmo de ordenación radique en los pasos de selección de los marcadores en función del criterio de precisión utilizado y en el barrido de permutaciones de cuatripletas de marcadores propuestas en el último paso del algoritmo. Nótese que con este criterio, "la tendencia" del algoritmo es seleccionar aquellos marcadores candidatos, de entre los más cercanos a los que ya están ubicados en el mapa, que son codominantes antes que los que son dominantes, ya que, como se muestra en la Figura 8.2, las combinaciones CC proporcionan menor variabilidad, pero esta tendencia no es estricta como se obligaba en el Capítulo 7, al insertar los marcadores dominantes en un mapa preliminar de marcadores codominantes. Las permutaciones de cuatripletas de marcadores planteadas en el último paso del algoritmo de ordenación buscan el equilibrio entre el gasto computacional y la mejora de la reubicación de marcadores cercanos aprovechando la información conjunta del resto de marcadores

Nótese que el método se ha llevado a cabo en un entorno poco explorado hasta ahora por otros autores. La metodología desarrollada, tiene en cuenta que las recombinaciones no son inequívocamente observables en el caso de dominancia en poblaciones F_2. Los datos simulados proceden de un mapa genético relativamente corto (menos de un 1 cM). Se obtienen órdenes alternativos como consecuencia de un problema de selección entre múltiples ordenacio-

nes distintas. La metodología filtra aquellos órdenes que son improbables o inconsistentes. Se obtiene una probabilidad que procede de una distribución posterior y que cuantifica la fiabilidad de cada ordenación. Como consecuencia, el método descrito en este capítulo, se considera el óptimo tanto para la obtención de las frecuencias de recombinación y distancias multipunto, como para la obtención de la ordenación de los marcadores. Por tanto, este es el método que se empleará en los capítulos posteriores.

Capítulo 9

Estudio de la estabilidad del método y comparación con métodos frecuentistas.

El objetivo de este capítulo es la comparación de los resultados obtenidos por la metodología bayesiana, formalizada en el capítulo anterior para estimar un mapa genético, con respecto a los resultados obtenidos por dos programas de uso extendido y que operan bajo una metodología frecuentista, como son Mapmaker [48] de acceso libre y JoinMap [77], con licencia comercial.

La limitación que tienen estos programas para automatizar la simulación de la distribución en el muestreo ha condicionado que las versiones de los programas utilizadas, en este Capítulo, sean las citadas en el párrafo anterior y que el número de marcadores que define el mapa genético utilizado en el estudio no supere los 8 marcadores. Por ese último motivo, en el Capítulo 7 ya se diseñó un mapa genético con esta condición (Figura 1.11), y se va a utilizar para recrear los tres escenarios investigados durante los capítulos anteriores. Es decir, en primer lugar se considera que el mapa proviene de una población Retrocruce, posteriormente se considera que el mapa proviene de una población F_2 con todos los marcadores codominantes y por último, se considera que el mapa proviene de una población F_2 con marcadores codominantes y dominantes, con idéntico diseño que en el Capítulo 7.

En cada uno de estos tres escenarios también se investiga la influencia del tamaño muestral, de manera que, para cada uno de los tres tipos de población

se han simulado 500 muestras, de 200, 100 y 50 individuos.

9.1. Retrocruce

A continuación, se presentan los resultados obtenidos de la simulación de los tres tamaños muestrales en la población Retrocruce.

9.1.1. Resultados obtenidos de 500 muestras de 200 individuos

El Cuadro 9.1 ofrece un resumen de las distancias reales entre marcadores contiguos, junto con las distancias medias obtenidas según el modelo bayesiano y las estimaciones obtenidas por los dos programas de referencia.

	distancias reales	modelo bayesiano con 200 indv.	modelo Mapmaker con 200 indv.	modelo JoinMap con 200 indv.
		media±desv. típ.	media±desv. típ.	media±desv. típ.
$d_{1,2}$	0.201217	0.2130 ± 0.0336	0.1768 ± 0.0199	0.2167 ± 0.0338
$d_{2,3}$	0.079390	0.0771 ± 0.0200	0.0725 ± 0.0169	0.0780 ± 0.0207
$d_{3,4}$	0.040954	0.0394 ± 0.0133	0.0387 ± 0.0129	0.0401 ± 0.0139
$d_{4,5}$	0.017158	0.0164 ± 0.0080	0.0159 ± 0.0087	0.0158 ± 0.0093
$d_{5,6}$	0.030935	0.0294 ± 0.0123	0.0297 ± 0.0123	0.0304 ± 0.0129
$d_{6,7}$	0.018935	0.0189 ± 0.0088	0.0179 ± 0.0094	0.0179 ± 0.0100
$d_{7,8}$	0.143953	0.1419 ± 0.0276	0.1270 ± 0.0195	0.1446 ± 0.0276

Cuadro 9.1: *Distancias reales entre marcadores contiguos junto con las distancias medias y desviaciones típicas obtenidas con el modelo bayesiano y los dos modelos frecuentistas, según 500 muestras con 200 individuos.*

En la Figura 9.1 se resume gráficamente la información del Cuadro 9.1. Nótese que bajo las representaciones gráficas de los mapas genéticos aparecen los porcentajes de acierto obtenidos por cada metodología al estimar el mapa genético real. Por ejemplo, con el método bayesiano 474 muestras de 500 han obtenido como modelo más probable el modelo correcto. Es decir, un porcentaje de acierto del $\frac{474}{500}100\,\% = 94.8\,\%$

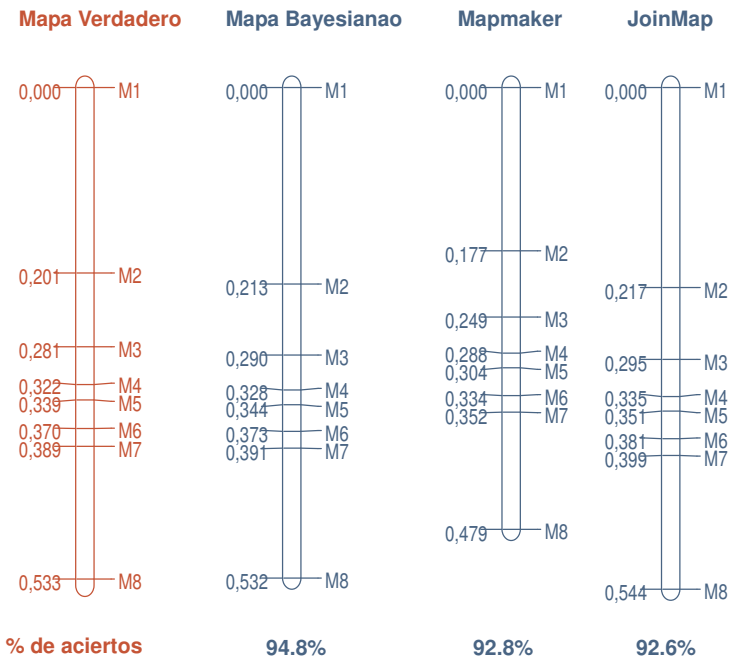


Figura 9.1: *Mapa de una población Retrocruce con 8 marcadores, junto con los modelos bayesiano, Mapmaker y JoinMap, para 500 muestras con 200 individuos.*

Otro resultado interesante es el que se muestra en la Figura 9.2. Como ya es sabido en el enfoque bayesiano, tras la simulación, para cada una de las muestras, se obtiene una colección de modelos, con probabilidades asociadas, que estiman el mapa genético de la población. A continuación se representan gráficamente las probabilidades de los dos mejores modelos de las 474 muestras, cuyo modelo más probable ha resultado ser el correcto.

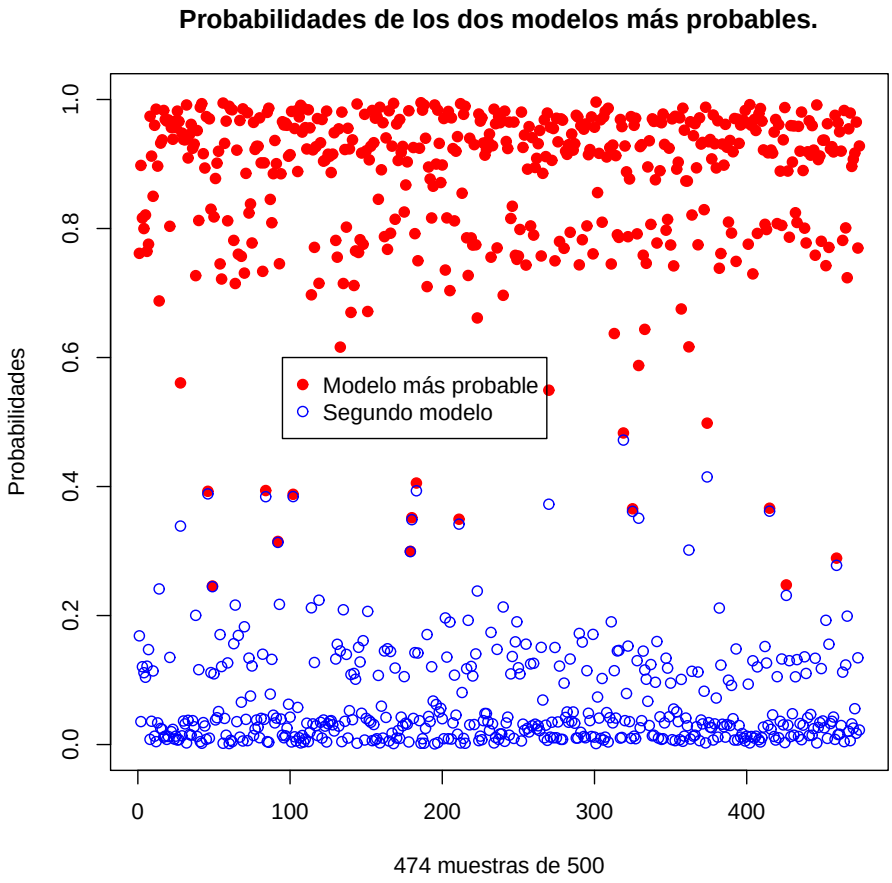


Figura 9.2: *Probabilidades de los dos mejores modelos de las muestras cuyo modelo más probable ha resultado ser el correcto.*

Como se puede observar, las probabilidades de los modelos más probables son ampliamente superiores a la de los segundos modelos, de manera que no existe lugar a dudas sobre cuál es el modelo definitivo.

En el Cuadro 9.2 se muestra el promedio de las probabilidades asociadas a los modelos que ocupan la posición j-ésima, dado que el modelo correcto ha salido en la posición i-ésima.

	veces en 500 muestras	j=1	j=2	j=3
i=1	474	**0.8747643**	0.06771416	0.01589643
i=2	23	0.4573499	**0.3357906**	0.04119497
i=3	2	0.1459266	0.1387536	**0.1329023**

Cuadro 9.2: *Para la distribución con muestras de 200 individuos, promedio de las probabilidades asociadas a los modelos que ocupan la posición j-ésima, dado que el modelo correcto ha salido en la posición i-ésima.*

Se observa que cuando el modelo correcto se registra como modelo más probable (en 474 muestras), la probabilidad media difiere mucho de la probabilidad media del resto de modelos (0.8747643 vs $\leq$ 0.06771416). Sin embargo, cuando el modelo correcto aparece en segunda posición o más allá, las probabilidades medias de los modelos de las posiciones superiores no difieren mucho respecto a las probabilidades medias de las posiciones inmediatamente siguientes ("0.4573499 vs 0.3357906" y "0.1459266 vs 0.1329023"). En estas situaciones, el investigador tiene un elemento de juicio que le permite tomar decisiones en base a afirmaciones probabilísticas.

9.1.2. Resultados obtenidos de 500 muestras de 100 individuos

Rebajando el tamaño muestral a 100 individuos, se presentan los resultados equivalentes a la sección anterior.

En el Cuadro 9.3 se resumen de las distancias reales entre los marcadores contiguos del mapa genético junto con las distancias medias estimadas mediante el modelo bayesiano y los modelos frecuentistas.

	distancias reales	modelo bayesiano con 100 indv.	modelo Mapmaker con 100 indv.	modelo JoinMap con 100 indv.
		media±desv. típ.	media±desv. típ.	media±desv. típ.
$d_{1,2}$	0.201217	0.2165 ± 0.0490	0.1789 ± 0.0307	0.2184 ± 0.0507
$d_{2,3}$	0.079390	0.0794 ± 0.0283	0.0741 ± 0.0241	0.0794 ± 0.0291
$d_{3,4}$	0.040954	0.0379 ± 0.0180	0.0383 ± 0.0190	0.0398 ± 0.0211
$d_{4,5}$	0.017158	0.0191 ± 0.0090	0.0167 ± 0.0123	0.0169 ± 0.0129
$d_{5,6}$	0.030935	0.0300 ± 0.0153	0.0306 ± 0.0174	0.0315 ± 0.0185
$d_{6,7}$	0.018935	0.0220 ± 0.0107	0.0187 ± 0.0132	0.0188 ± 0.0141
$d_{7,8}$	0.143953	0.1437 ± 0.0379	0.1289 ± 0.0289	0.1466 ± 0.0388

Cuadro 9.3: *Distancias reales entre marcadores contiguos junto con las distancias medias y desviaciones típicas obtenidas por el modelo bayesiano y los dos modelos frecuentistas, según 500 muestras con 100 individuos.*

Del mismo modo que en el apartado anterior, en la Figura 9.3 se resume gráficamente la información del Cuadro 9.3, especificando bajo cada representación gráfica el porcentaje de aciertos obtenidos por cada metodología al estimar el mapa genético real.

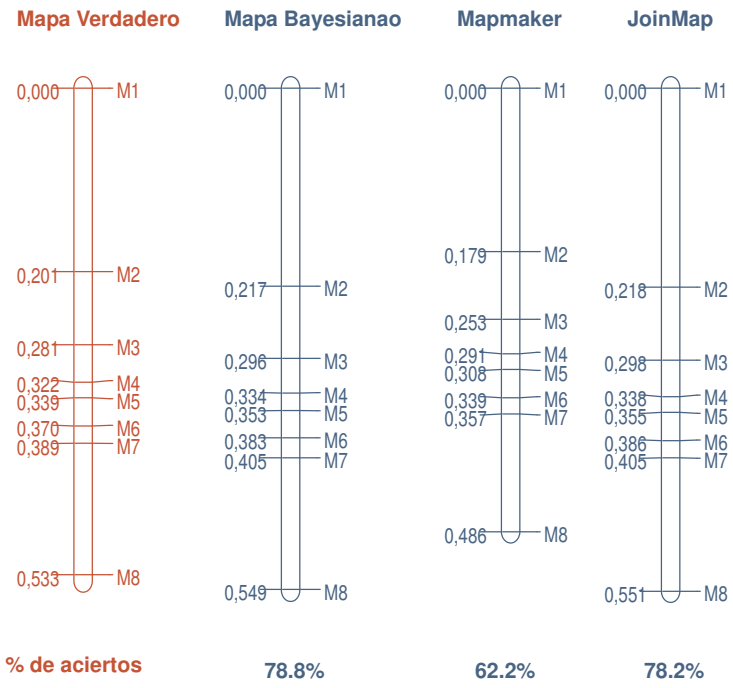


Figura 9.3: *Mapa de una población Retrocruce con 8 marcadores, junto con los modelos bayesiano, Mapmaker y JoinMap, para muestras de 100 individuos.*

En la Figura 9.4 se representan las 394 muestras de 500 cuyo modelo más probable ha resultado ser el correcto junto con las probabilidades de los segundos modelos más probables.

Como se puede observar, aunque las nubes de puntos que representan las probabilidades de los modelos más probables y las de los segundos modelos están claramente separas, se aprecia un acercamiento entre ellas en comparación al resultado equivalente, obtenido para muestras de 200 individuos.

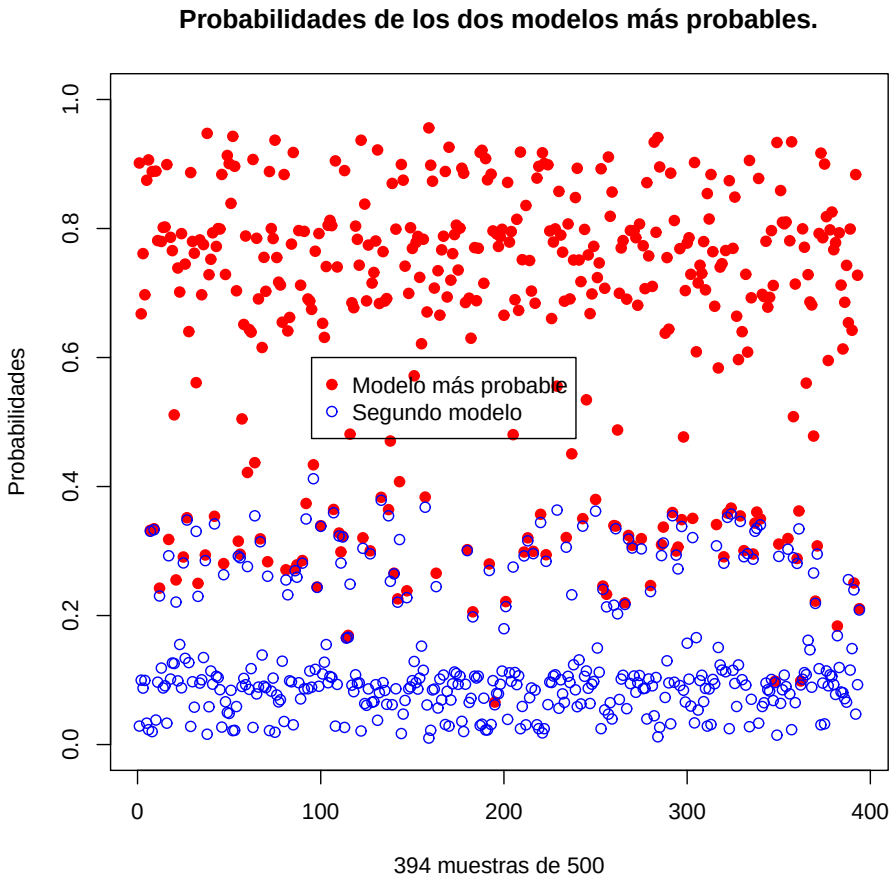


Figura 9.4: *Probabilidades de los dos mejores modelos de las muestras cuyo modelo más probable ha resultado ser el correcto.*

En el Cuadro 9.4 se muestra el promedio de las probabilidades asociadas a los modelos que ocupan la posición j-ésima, dado que el modelo correcto ha salido en la posición i-ésima.

	veces en 500 muestras	j=1	j=2	j=3
i=1	394	**0.6556092**	0.1337729	0.04142361
i=2	91	0.3307905	**0.2861699**	0.05458876
i=3	5	0.2509997	0.1427002	**0.119368**

Cuadro 9.4: *Para la distribución con muestras de 100 individuos, promedio de las probabilidades asociadas a los modelos que ocupan la posición j-ésima, dado que el modelo correcto ha salido en la posición i-ésima.*

De nuevo, para las 394 muestras en que el modelo más probable ha resultado ser el correcto, las probabilidades medias entre los modelos primero y segundo están bien diferenciadas (0.6556092 vs 0.1337729). Sin embargo, esta diferenciación tan clara se atenúa según el modelo correcto aparece en segunda posición o inferior.

9.1.3. Resultados obtenidos de 500 muestras de 50 individuos

Por último, se presentan los resultados obtenidos de 500 muestras con 50 individuos.

En el Cuadro 9.5 se resumen las distancias reales entre marcadores contiguos del mapa genético, junto con la estimación de las distancias medias obtenidas según la metodología bayesiana y la de los programas frecuentistas de referencia.

	distancias reales	modelo bayesiano con 50 indv.	modelo Mapmaker con 50 indv.	modelo JoinMap con 50 indv.
		media±desv. típ.	media±desv. típ.	media±desv. típ.
$d_{1,2}$	0.201217	0.2834 ± 0.0823	0.2316 ± 0.0552	0.2914 ± 0.0956
$d_{2,3}$	0.079390	0.1040 ± 0.0436	0.0946 ± 0.0381	0.1029 ± 0.0498
$d_{3,4}$	0.040954	0.0520 ± 0.0269	0.0496 ± 0.0316	0.0509 ± 0.0343
$d_{4,5}$	0.017158	0.0302 ± 0.0142	0.0222 ± 0.0210	0.0223 ± 0.0218
$d_{5,6}$	0.030935	0.0426 ± 0.0213	0.0404 ± 0.0278	0.0408 ± 0.0288
$d_{6,7}$	0.018935	0.0346 ± 0.0153	0.0236 ± 0.0213	0.0240 ± 0.0224
$d_{7,8}$	0.143953	0.1869 ± 0.0602	0.1653 ± 0.0464	0.1968 ± 0.0681

Cuadro 9.5: *Distancias reales entre marcadores contiguos junto las distancias medias y desviaciones típicas obtenidas por el modelo bayesiano y los dos modelos frecuentistas, según 500 muestras con 50 individuos.*

En la Figura 9.5 se resume gráficamente la información del Cuadro 9.5 y el porcentaje de aciertos obtenidos por cada metodología. Al disminuir el tamaño de la muestra a valores de 50 individuos, el enfoque bayesiano y JoinMap proporcionan estimas de las distancias superiores a las reales. Como de forma sistemática Mapmaker subestima las distancias, el resultado final se aproxima más alas distancias reales. La mejora de Mapmaker se podría considerar, por tanto , un artificio.

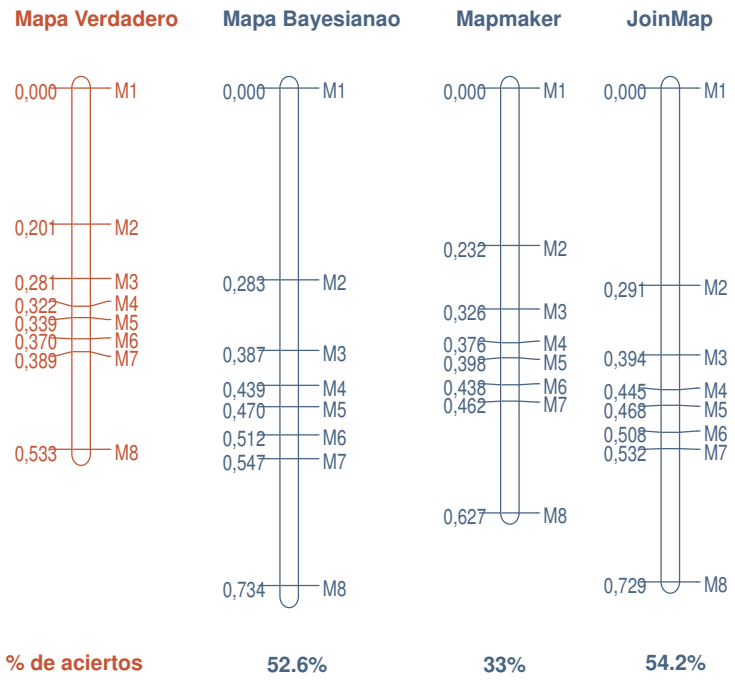


Figura 9.5: *Mapa de una población Retrocruce con 8 marcadores, junto con los modelos bayesiano, Mapmaker y JoinMap, para muestras de 50 individuos.*

En la Figura 9.6 se representan las 263 muestras de 500 cuyo modelo más probable ha resultado ser el correcto junto con las probabilidades de los segundos modelos más probables.

Es evidente que un gran número de muestras obtienen probabilidades muy parecidas en los dos primeros modelos estimados. La reducción del tamaño muestral influye sobre el parecido de las probabilidades de los dos modelos más probables.

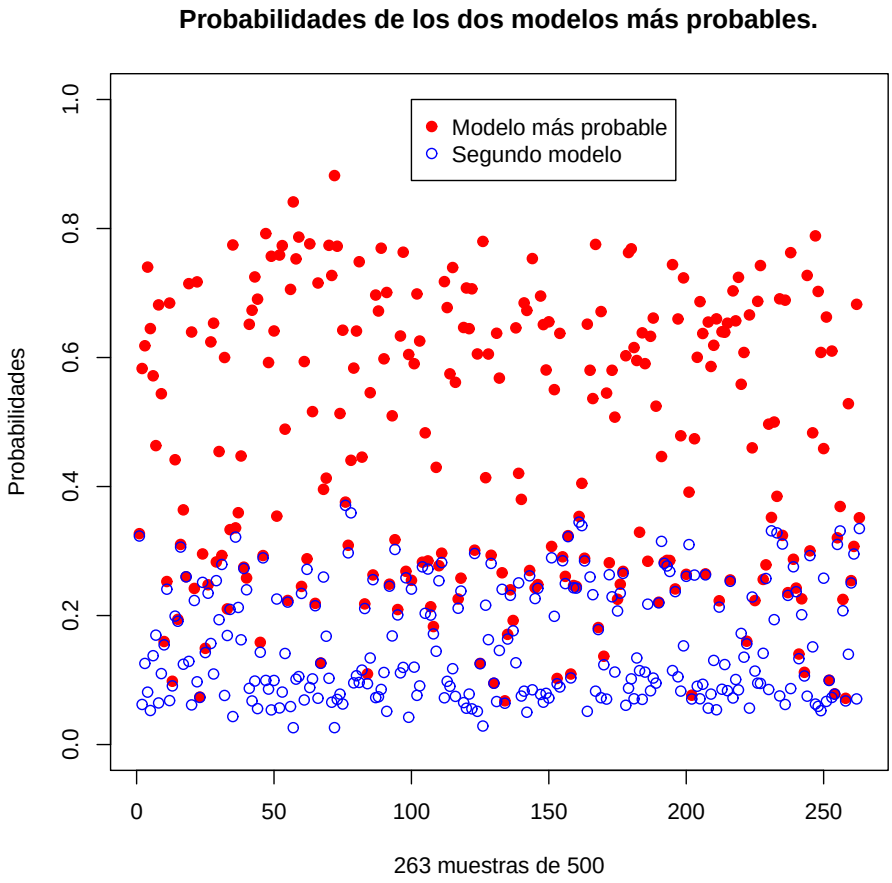


Figura 9.6: *Probabilidades de los dos mejores modelos de las muestras cuyo modelo más probable ha resultado ser el correcto.*

En el Cuadro 9.6 se muestra el promedio de las probabilidades asociadas a los modelos que ocupan la posición j-ésima, dado que el modelo correcto ha salido en la posición i-ésima.

	veces en 500 muestras	j=1	j=2	j=3
i=1	263	**0.4676707**	0.156692	0.05708518
i=2	140	0.2548059	**0.2234483**	0.06182296
i=3	28	0.123872	0.1037373	**0.08153999**

Cuadro 9.6: *Para la distribución con muestras de 50 individuos, promedio de las probabilidades asociadas a los modelos que ocupan la posición j-ésima, dado que el modelo correcto ha salido en la posición i-ésima.*

La probabilidad media obtenida de las muestras cuyo modelo más probable ha resultado ser el correcto, se reduce considerablemente con respecto al resultado equivalente obtenido para muestras de tamaño 200 individuos. Aún así, las probabilidades medias de la primera fila del cuadro difieren más, entre sí, que las probabilidades medias de la segunda y tercera fila del cuadro. Lo que demuestra que cuando el modelo correcto se obtiene como más probable, es indiscutible su fiabilidad. Sin embargo, cuando el modelo correcto se obtiene en posiciones siguientes a la primera, los modelos que lo superan son prácticamente equiprobables al correcto.

9.2. F_2 con todos los marcadores codominantes

En esta sección se trabaja con idéntico mapa genético que en la sección anterior pero las muestras de individuos se obtienen considerando que el mapa genético representa a una población F_2 con los 8 marcadores codominantes.

Recordemos que los resultados se obtienen en base a 500 muestras pero con distinto tamaño muestral: 200, 100 y 50 individuos.

9.2.1. Resultados obtenidos de 500 muestras de 200 individuos

En el Cuadro 9.7 se resume, igual que en la sección anterior, las distancias reales entre marcadores contiguos, junto con las estimaciones de los modelos posteriores bayesiano y frecuentistas.

	distancias reales	modelo bayesiano con 200 indv.	modelo Mapmaker con 200 indv.	modelo JoinMap con 200 indv.
		media±desv. típ.	media±desv. típ.	media±desv. típ.
$d_{1,2}$	0.201217	0.1955 ± 0.0267	0.1643 ± 0.0178	0.1994 ± 0.0273
$d_{2,3}$	0.079390	0.0729 ± 0.0148	0.0682 ± 0.0126	0.0740 ± 0.0152
$d_{3,4}$	0.040954	0.0369 ± 0.0096	0.0361 ± 0.0089	0.0380 ± 0.0100
$d_{4,5}$	0.017158	0.0160 ± 0.0064	0.0158 ± 0.0064	0.0162 ± 0.0069
$d_{5,6}$	0.030935	0.0273 ± 0.0083	0.0273 ± 0.0081	0.0285 ± 0.0088
$d_{6,7}$	0.018935	0.0173 ± 0.0067	0.0168 ± 0.0065	0.0173 ± 0.0071
$d_{7,8}$	0.143953	0.1400 ± 0.0211	0.1231 ± 0.0154	0.1425 ± 0.0213

Cuadro 9.7: *Distancias reales entre marcadores junto las distancias medias y desviaciones típicas obtenidas por el modelo bayesiano y los dos modelos frecuentistas, según 500 muestras con 200 individuos.*

En la Figura 9.7 se resume gráficamente la información del Cuadro 9.7, junto con el porcentaje de aciertos obtenidos por cada metodología al estimar el mapa genético real.

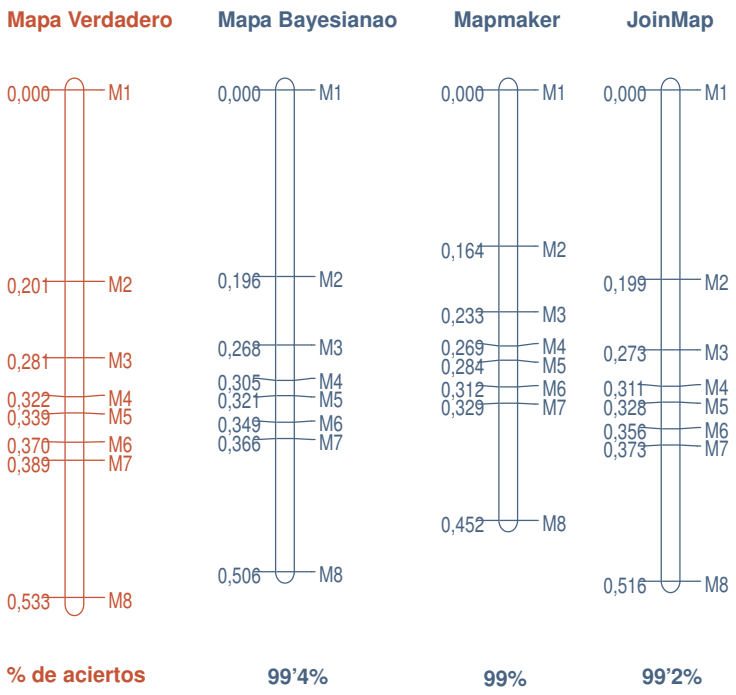


Figura 9.7: *Mapa de una población F_2 con 8 marcadores todos codominantes, junto con los modelos bayesiano, Mapmaker y JoinMap, para muestras de 200 individuos.*

En este caso, 497 muestras de las 500 obtienen como modelo más probable el modelo correcto, lo que supone un porcentaje de acierto del $\frac{497}{500}100\,\% = 99.4\,\%$

En la Figura 9.8 se representan las 497 muestras cuyo modelo más probable ha resultado ser el correcto junto con las probabilidades de los segundos modelos más probables.

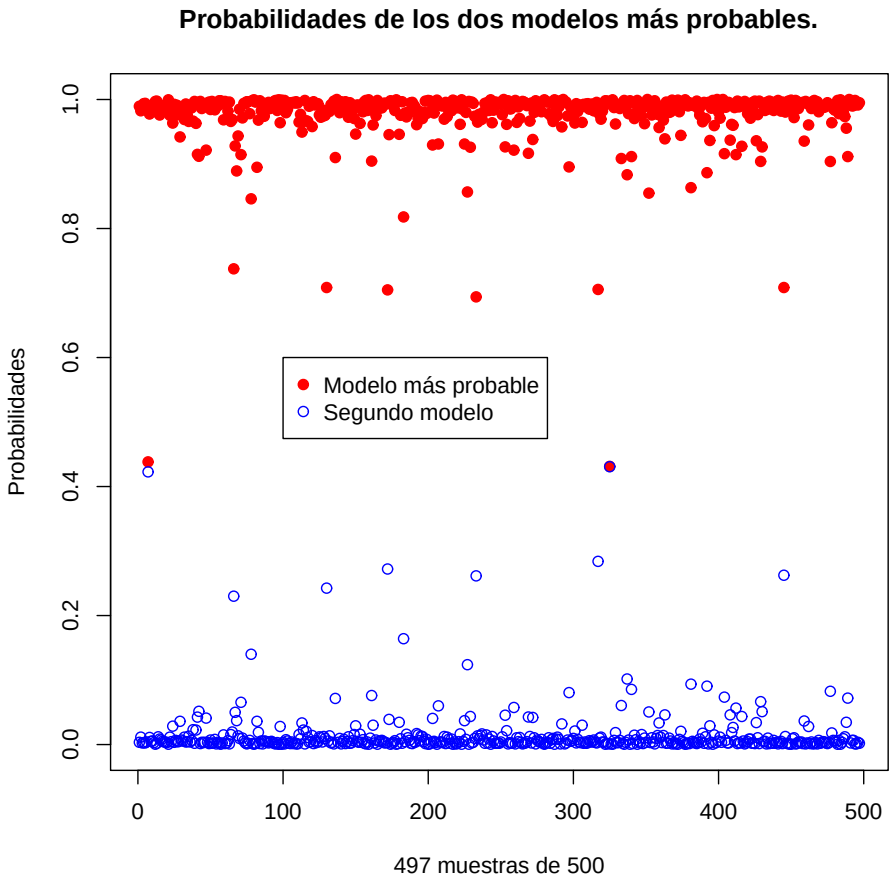


Figura 9.8: *Probabilidades de los dos mejores modelos de las muestras cuyo modelo más probable ha resultado ser el correcto.*

Se puede observar que las probabilidades asociadas a los modelos más probables son muy próximas a 1, por lo que el modelo correcto obtiene una alta fiabilidad.

En el Cuadro 9.8 se muestra el promedio de las probabilidades asociadas a los modelos que ocupan la posición j-ésima, dado que el modelo correcto ha salido en la posición i-ésima.

	veces en 500 muestras	j=1	j=2	j=3
i=1	497	**0.9747385**	0.01524294	0.003501541
i=2	3	0.4537385	**0.4423693**	0.020899084

Cuadro 9.8: *Para la distribución con muestras de 200 individuos, promedio de las probabilidades asociadas a los modelos que ocupan la posición j-ésima, dado que el modelo correcto ha salido en la posición i-ésima.*

9.2.2. Resultados obtenidos de 500 muestras de 100 individuos

En el Cuadro 9.9 se ofrece un resumen con las distancias reales entre marcadores contiguos, junto con las estimaciones de los modelos posteriores bayesiano y frecuentistas.

	distancias reales	modelo bayesiano con 100 indv.	modelo Mapmaker con 100 indv.	modelo JoinMap con 100 indv.
		media±desv. típ.	media±desv. típ.	media±desv. típ.
$d_{1,2}$	0.201217	0.1890 ± 0.0355	0.1601 ± 0.0238	0.1929 ± 0.0362
$d_{2,3}$	0.079390	0.0715 ± 0.0205	0.0671 ± 0.0174	0.0726 ± 0.0212
$d_{3,4}$	0.040954	0.0361 ± 0.0137	0.0361 ± 0.0128	0.0379 ± 0.0147
$d_{4,5}$	0.017158	0.0158 ± 0.0080	0.0151 ± 0.0089	0.0154 ± 0.0097
$d_{5,6}$	0.030935	0.0271 ± 0.0115	0.0278 ± 0.0118	0.0288 ± 0.0127
$d_{6,7}$	0.018935	0.0171 ± 0.0085	0.0161 ± 0.0093	0.0163 ± 0.0100
$d_{7,8}$	0.143953	0.1337 ± 0.0288	0.1196 ± 0.0213	0.1368 ± 0.0292

Cuadro 9.9: *Distancias reales entre marcadores junto las distancias medias y desviaciones típicas obtenidas por el modelo bayesiano y los dos modelos frecuentistas, según 500 muestras con 100 individuos.*

En la Figura 9.9 se resume gráficamente la información del Cuadro 9.9, junto al porcentaje de aciertos obtenidos por las distintas metodologías al estimar de forma correcta el mapa genético de la población.

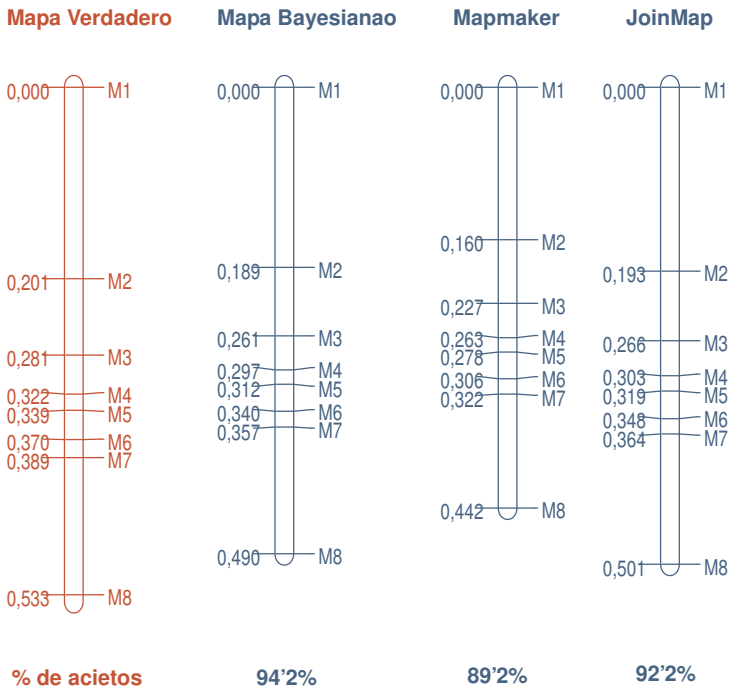


Figura 9.9: *Mapa de una población F_2 con 8 marcadores todos codominantes, junto con los modelos bayesiano, Mapmaker y JoinMap, para muestras de 100 individuos.*

Como se puede ver en la Figura 9.9, bajo cada mapa, aparece el porcentaje de aciertos obtenido. En el caso del modelo bayesiano, 471 muestras, de 500, obtienen el modelo correcto como modelo más probable, lo que supone un porcentaje de éxito del $\frac{471}{500}100\,\% = 94.2\,\%$

En la Figura 9.10 se representan las 471 muestras cuyo modelo más probable ha resultado ser el correcto junto con las probabilidades de los segundos modelos más probables.

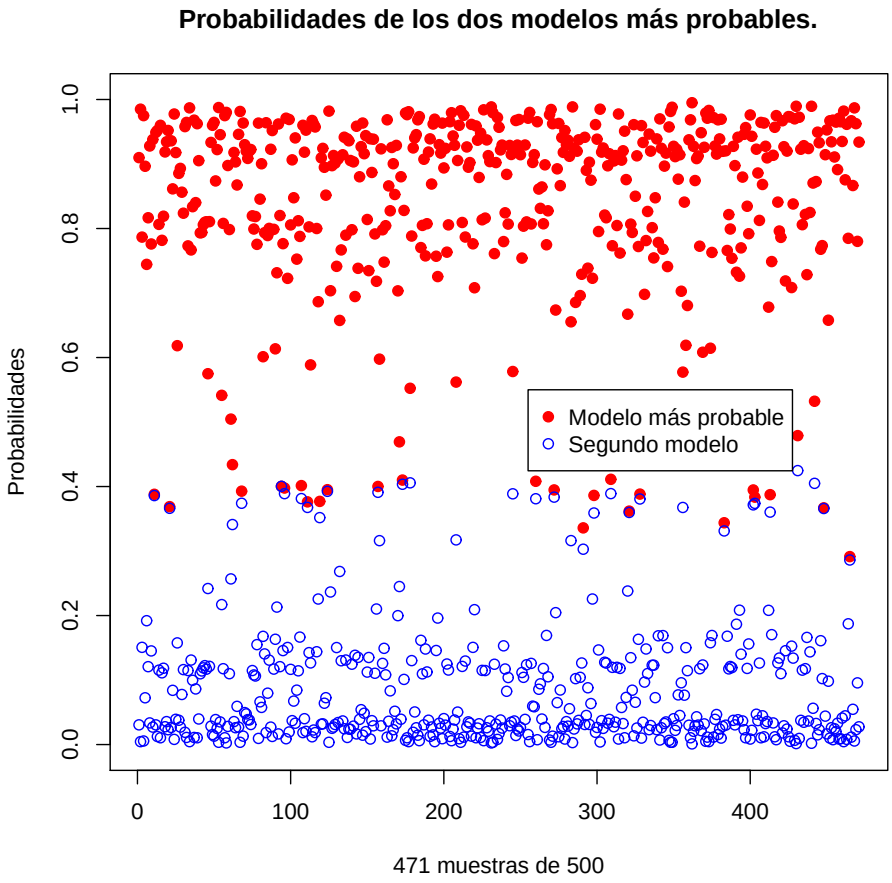


Figura 9.10: *Probabilidades de los dos mejores modelos de las muestras cuyo modelo más probable ha resultado ser el correcto.*

Se observa gráficamente que, aun en el caso de trabajar con un tamaño muestral de 100 individuos, las probabilidades de los dos modelos más probables están bastante alejadas, por lo que no queda lugar a dudas sobre la preferencia del primer modelo.

En el Cuadro 9.10 se muestra el promedio de las probabilidades asociadas a los modelos que ocupan la posición j-ésima, dado que el modelo correcto ha salido en la posición i-ésima.

	veces en 500 muestras	j=1	j=2	j=3
i=1	471	**0.842926**	0.08650804	0.01917817
i=2	28	0.4352395	**0.337802**	0.0405518
i=3	1	0.1712846	0.1687657	**0.163728**

Cuadro 9.10: *Para la distribución con muestras de 100 individuos, promedio de las probabilidades asociadas a los modelos que ocupan la posición j-ésima, dado que el modelo correcto ha salido en la posición i-ésima.*

Se sigue cumpliendo el patrón en que, para muestras que obtienen el modelo correcto como más probable, el promedio de las probabilidades de los siguientes modelos están muy alejadas del primer modelo. Sin embargo, para muestras en que el modelo correcto no aparece en primera posición, el promedio de las probabilidades de los modelos que están por delante, no difieren excesivamente del del modelo correcto.

9.2.3. Resultados obtenidos de 500 muestras de 50 individuos

En el Cuadro 9.11 se ofrece un resumen con las distancias reales entre marcadores contiguos, junto con los modelos bayesiano y frecuentitas y en la Figura 9.11 se resume gráficamente la información del Cuadro 9.11:

	distancias reales	modelo bayesiano con 50 indv.	modelo Mapmaker con 50 indv.	modelo JoinMap con 50 indv.
		media±desv. típ.	media±desv. típ.	media±desv. típ.
$d_{1,2}$	0.201217	0.1963 ± 0.0553	0.1802 ± 0.0372	0.2053 ± 0.0592
$d_{2,3}$	0.079390	0.0832 ± 0.0288	0.0796 ± 0.0252	0.0837 ± 0.0304
$d_{3,4}$	0.040954	0.0420 ± 0.0193	0.0425 ± 0.0201	0.0437 ± 0.0218
$d_{4,5}$	0.017158	0.0202 ± 0.0094	0.0178 ± 0.0128	0.0177 ± 0.0130
$d_{5,6}$	0.030935	0.0329 ± 0.0163	0.0336 ± 0.0182	0.0343 ± 0.0190
$d_{6,7}$	0.018935	0.0239 ± 0.0112	0.0207 ± 0.0138	0.0210 ± 0.0145
$d_{7,8}$	0.143953	0.1612 ± 0.0465	0.1447 ± 0.0331	0.1669 ± 0.0485

Cuadro 9.11: *Distancias reales entre marcadores junto las distancias medias y desviaciones típicas obtenidas por el modelo bayesiano y los dos modelos frecuentistas, según 500 muestras con 50 individuos.*

En este caso, aun trabajando con un tamaño muestral tan desfavorable, se puede observar en la Figura 9.11, que el porcentaje de aciertos del modelo bayesiano es satisfactorio y bastante superior al de los programas de referencia, sobre todo a Mapmaker. Del orden del $\frac{402}{500}100\,\% = 80.4\,\%$

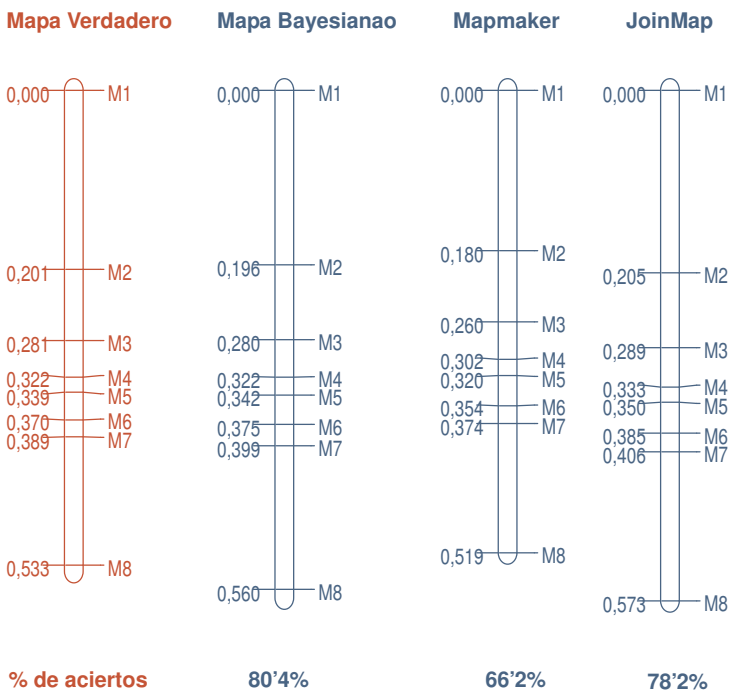


Figura 9.11: *Mapa de una población F_2 con 8 marcadores todos codominantes, junto con los modelos bayesiano, Mapmaker y JoinMap, para muestras de 50 individuos.*

En la Figura 9.12 se representan las 402 muestras de 500 cuyo modelo más probable ha resultado ser el correcto junto con las probabilidades de los segundos modelos más probables.

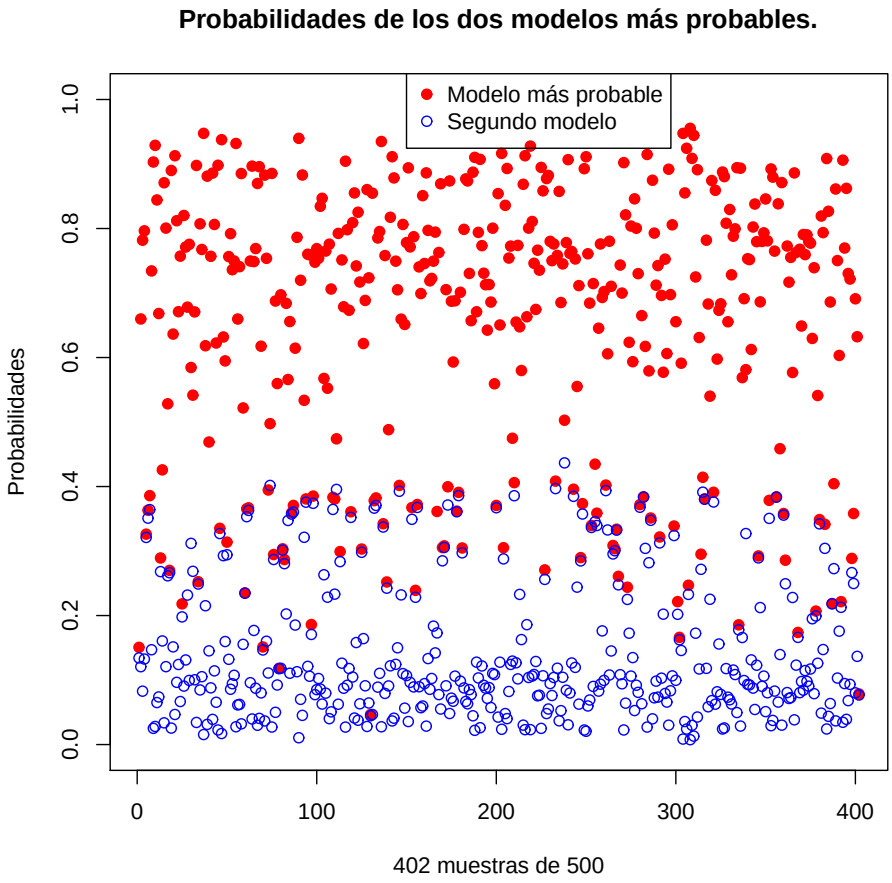


Figura 9.12: *Probabilidades de los dos mejores modelos de las muestras cuyo modelo más probable ha resultado ser el correcto.*

De nuevo, como en las secciones anteriores, las probabilidades de los dos mejores modelos, cuando el primero ha sido el correcto, están muy alejadas entre si, por lo que no hay lugar a dudas del modelo preferente.

En el Cuadro 9.12 se muestra el promedio de las probabilidades asociadas a los modelos que ocupan la posición j-ésima, dado que el modelo correcto ha salido en la posición i-ésima.

	veces en 500 muestras	j=1	j=2	j=3
i=1	402	**0.6561862**	0.1424293	0.04185293
i=2	84	0.3571302	**0.28105**	0.054566
i=3	9	0.2869317	0.24363	**0.1097659**

Cuadro 9.12: *Para la distribución con muestras de 50 individuos, promedio de las probabilidades asociadas a los modelos que ocupan la posición j-ésima, dado que el modelo correcto ha salido en la posición i-ésima.*

9.3. F_2 con marcadores codominantes y dominantes conjuntamente

Para finalizar el capítulo, se considera una población F_2 definida por 8 marcadores no todos codominantes, con el mismo mapa genético que en las secciones anteriores, y se simula, con cada una de las tres metodologías (bayesiana, Mapmaker y JoinMap) para 500 muestras de 200, 100 y 50 individuos.

9.3.1. Resultados obtenidos de 500 muestras de 200 individuos

Empezamos con los resultados obtenidos para las muestras con 200 individuos. En el Cuadro 9.13 se ofrece un resumen con las distancias reales entre los marcadores contiguos, los tipos de los mismos y las estimaciones de los modelos posteriores bayesiano y frecuentista.

	tipo de cruce	distancias reales	modelo bayesiano con 200 indv.	modelo Mapmaker con 200 indv.	modelo JoinMap con 200 indv.
			media±sd	media±sd	media±sd
$d_{1,2}$	CD2	0.201217	0.1908 ± 0.0312	0.1664 ± 0.0204	0.1944 ± 0.0318
$d_{2,3}$	D2D1	0.079390	0.0841 ± 0.0270	0.0741 ± 0.0237	0.0761 ± 0.0317
$d_{3,4}$	D1D1	0.040954	0.0340 ± 0.0137	0.0340 ± 0.0133	0.0343 ± 0.0149
$d_{4,5}$	D1C	0.017158	0.0167 ± 0.0076	0.0149 ± 0.0091	0.0157 ± 0.0102
$d_{5,6}$	CD1	0.030935	0.0265 ± 0.0115	0.0264 ± 0.0124	0.0255 ± 0.0129
$d_{6,7}$	D1D2	0.018935	0.0309 ± 0.0113	0.0225 ± 0.0128	0.0253 ± 0.0209
$d_{7,8}$	D2C	0.143953	0.1378 ± 0.0242	0.1237 ± 0.0173	0.1447 ± 0.0261

Cuadro 9.13: *Distancias reales entre marcadores junto las distancias medias y desviaciones típicas obtenidas por el modelo bayesiano y los dos modelos frecuentistas, según 500 muestras con 200 individuos.*

En la Figura 9.13 se resume gráficamente la información del Cuadro 9.13, así como el porcentaje de aciertos obtenido por cada una de las metodologías. De nuevo, la metodología bayesiana presenta un porcentaje de aciertos muy superior esta vez a JoinMap y similar a Mapmaker.

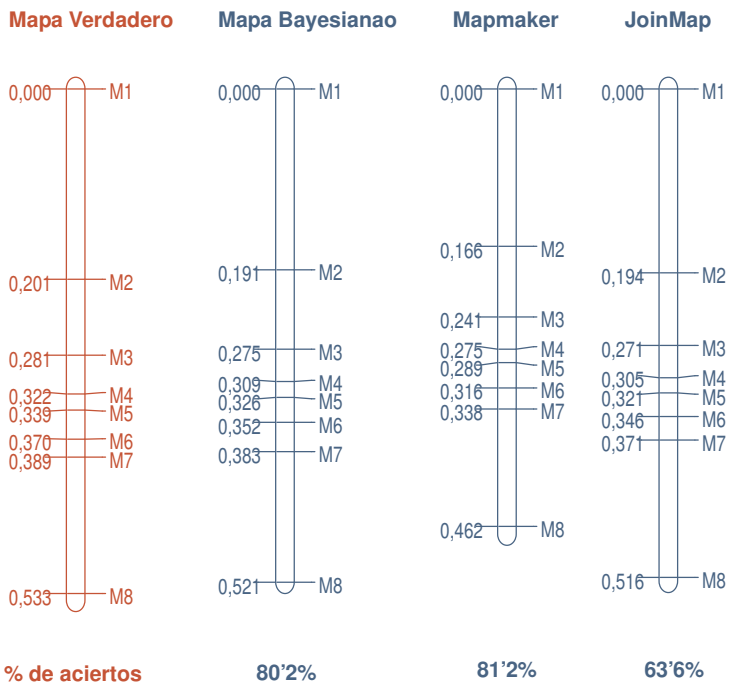


Figura 9.13: *Mapa de una población F_2 con 8 marcadores, no todos codominantes, junto con los modelos bayesiano, Mapmaker y JoinMap, para muestras de 200 individuos.*

En la Figura 9.14 se representan las 401 muestras de 500 cuyo modelo más probable ha resultado ser el correcto junto con las probabilidades de los segundos modelos más probables.

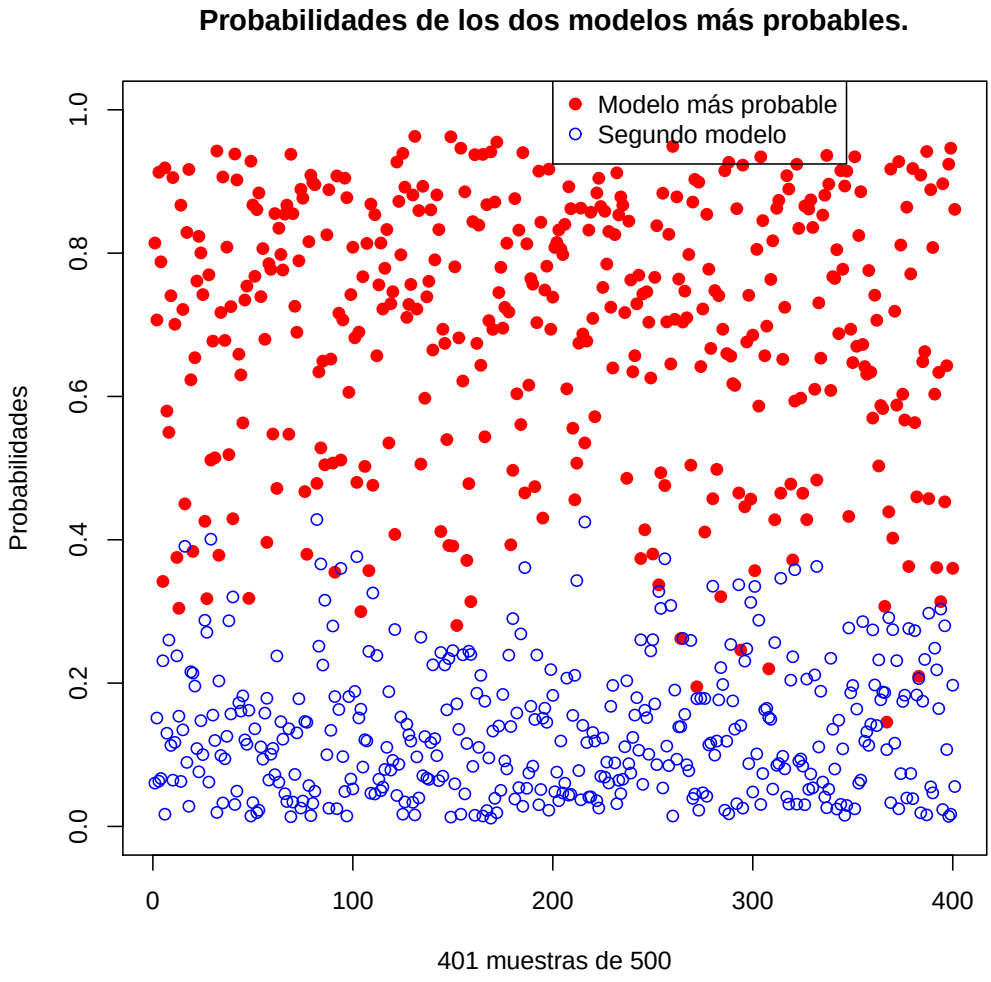


Figura 9.14: *Probabilidades de los dos mejores modelos de las muestras cuyo modelo más probable ha resultado ser el correcto.*

En el Cuadro 9.14 se muestra el promedio de las probabilidades asociadas a los modelos que ocupan la posición j-ésima, dado que el modelo correcto ha salido en la posición i-ésima.

	veces en 500 muestras	j=1	j=2	j=3
i=1	401	**0.6970636**	0.1333879	0.04108846
i=2	69	0.4713913	**0.2536051**	0.06072527
i=3	16	0.3912939	0.1860908	**0.1090338**

Cuadro 9.14: *Para la distribución con muestras de 200 individuos, promedio de las probabilidades asociadas a los modelos que ocupan la posición j-ésima, dado que el modelo correcto ha salido en la posición i-ésima.*

9.3.2. Resultados obtenidos de 500 muestras de 100 individuos

A continuación se presentan resultados similares a los anteriores pero para muestras con 100 individuos.

En el Cuadro 9.15 se ofrece un resumen con las distancias reales entre marcadores contiguos, los tipos de los mismos y las estimaciones obtenidas por las propuestas bayesina y frecuentistas.

	tipo de cruce	distancias reales	modelo bayesiano con 100 indv.	modelo Mapmaker con 100 indv.	modelo JoinMap con 100 indv.
			media±sd	media±sd	media±sd
$d_{1,2}$	CD2	0.201217	0.1930 ± 0.0421	0.1625 ± 0.0282	0.1981 ± 0.0485
$d_{2,3}$	D2D1	0.079390	0.0850 ± 0.0318	0.0698 ± 0.0313	0.0664 ± 0.0427
$d_{3,4}$	D1D1	0.040954	0.0383 ± 0.0179	0.0360 ± 0.0180	0.0369 ± 0.0199
$d_{4,5}$	D1C	0.017158	0.0210 ± 0.0095	0.0150 ± 0.0122	0.0168 ± 0.0134
$d_{5,6}$	CD1	0.030935	0.0305 ± 0.0140	0.0278 ± 0.0167	0.0270 ± 0.0170
$d_{6,7}$	D1D2	0.018935	0.0399 ± 0.0131	0.0262 ± 0.0163	0.0271 ± 0.0226
$d_{7,8}$	D2C	0.143953	0.1337 ± 0.0329	0.1206 ± 0.0254	0.1424 ± 0.0358

Cuadro 9.15: *Distancias reales entre marcadores junto las distancias medias y desviaciones típicas obtenidas por el modelo bayesiano y los dos modelos frecuentistas, según 500 muestras con 100 individuos.*

En la Figura 9.15 se resume gráficamente la información del Cuadro 9.15. Se mantiene el patrón señalado en el apartado anterior. Es decir, la metodología bayesiana proporciona resultados similares a Mapmaker y ampliamente superiores a JoinMap

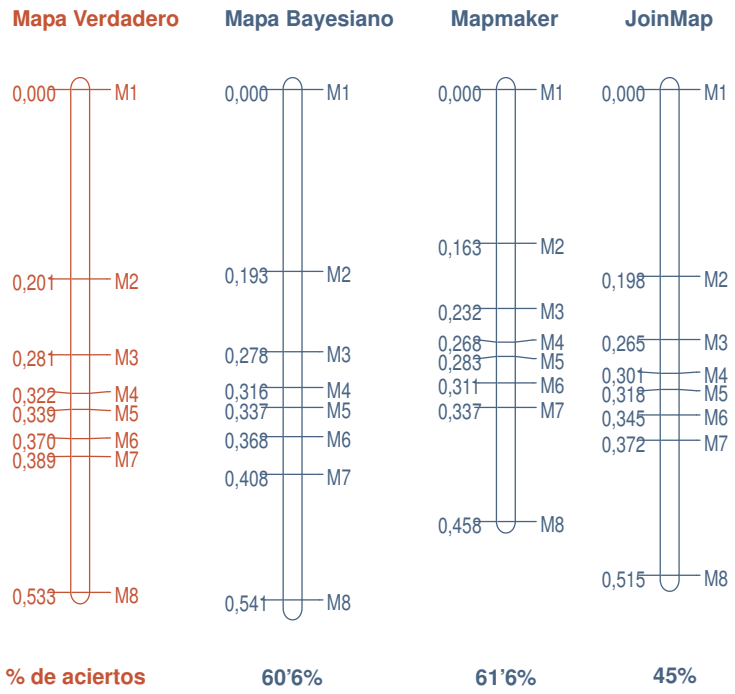


Figura 9.15: *Mapa de una población F_2 con 8 marcadores, no todos codominantes, junto con los modelos bayesiano, Mapmaker y JoinMap, para muestras de 100 individuos.*

En la Figura 9.16 se representan las 303 muestras de 500 cuyo modelo más probable es el modelo correcto junto con las probabilidades de los segundos modelos más probables.

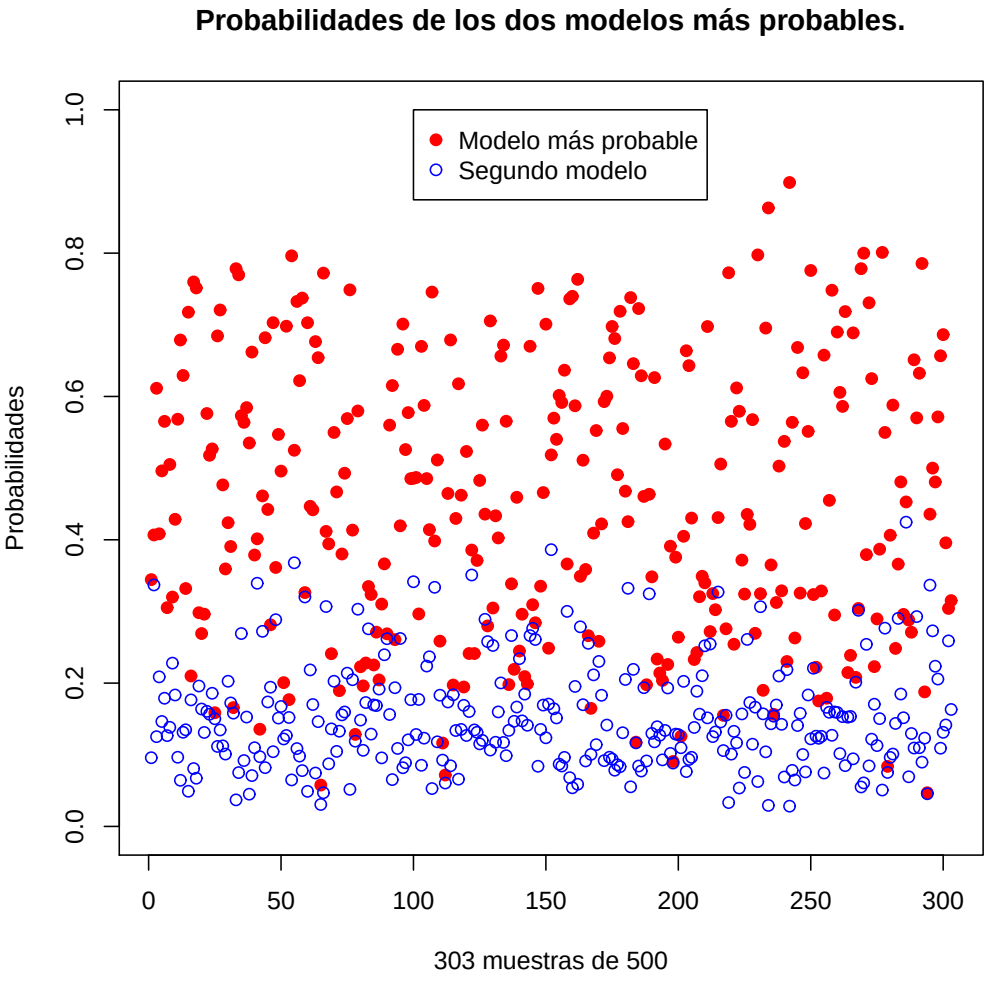


Figura 9.16: *Probabilidades de los dos mejores modelos de las muestras cuyo modelo más probable ha resultado ser el correcto.*

En el Cuadro 9.16 se muestra una comparación de los modelos más probables frente al resto de modelos, utilizando los resultados de nuestra propuesta. En él se ofrece el promedio de las probabilidades asociadas a los modelos que ocupan la posición j-ésima, dado que el modelo correcto ha salido en la

posición i-ésima.

	veces en 500 muestras	j=1	j=2	j=3
i=1	303	**0.4535319**	0.1525681	0.07183277
i=2	102	0.3449574	**0.1906287**	0.07029722
i=3	28	0.2372939	0.1490195	**0.09539315**

Cuadro 9.16: *Para la distribución con muestras de 100 individuos, promedio de las probabilidades asociadas a los modelos que ocupan la posición j-ésima, dado que el modelo correcto ha salido en la posición i-ésima.*

9.3.3. Resultados obtenidos de 500 muestras de 50 individuos

Por último finalizamos con resultados similares a los anteriores pero esta vez obtenidos para las muestras con 50 individuos.

En el Cuadro 9.17 se ofrece un resumen con las distancias reales entre marcadores contiguos, los tipos de los mismos y las estimaciones obtenidas tanto por la metodología bayesiana como por los programas Mapmaker [48] y JoinMap [77].

	tipo de cruce	distancias reales	modelo bayesiano con 50 indv.	modelo Mapmaker con 50 indv.	modelo JoinMap con 50 indv.
			media±sd	media±sd	media±sd
$d_{1,2}$	CD2	0.201217	0.2099 ± 0.0641	0.1878 ± 0.0602	0.2113 ± 0.0911
$d_{2,3}$	D2D1	0.079390	0.1478 ± 0.0577	0.0964 ± 0.0681	0.1062 ± 0.0780
$d_{3,4}$	D1D1	0.040954	0.0500 ± 0.0226	0.0420 ± 0.0303	0.0401 ± 0.0293
$d_{4,5}$	D1C	0.017158	0.0353 ± 0.0149	0.0163 ± 0.0197	0.0167 ± 0.0188
$d_{5,6}$	CD1	0.030935	0.0452 ± 0.0186	0.0346 ± 0.0309	0.0324 ± 0.0264
$d_{6,7}$	D1D2	0.018935	0.0744 ± 0.0263	0.0446 ± 0.0314	0.0499 ± 0.0379
$d_{7,8}$	D2C	0.143953	0.1634 ± 0.0475	0.1494 ± 0.0418	0.1769 ± 0.0581

Cuadro 9.17: *Distancias reales entre marcadores junto las distancias medias y desviaciones típicas obtenidas por el modelo bayesiano y los dos modelos frecuentistas, según 500 muestras con 50 individuos.*

En la Figura 9.17 se resume gráficamente la información del Cuadro 9.17 y los porcentajes de acierto obtenidos por las tres metodologías. En este caso extremo, la metodología bayesiana presenta los mejores resultados, siendo de nuevo JoinMap el método que proporciona un porcentaje de aciertos muy bajo (17 %).

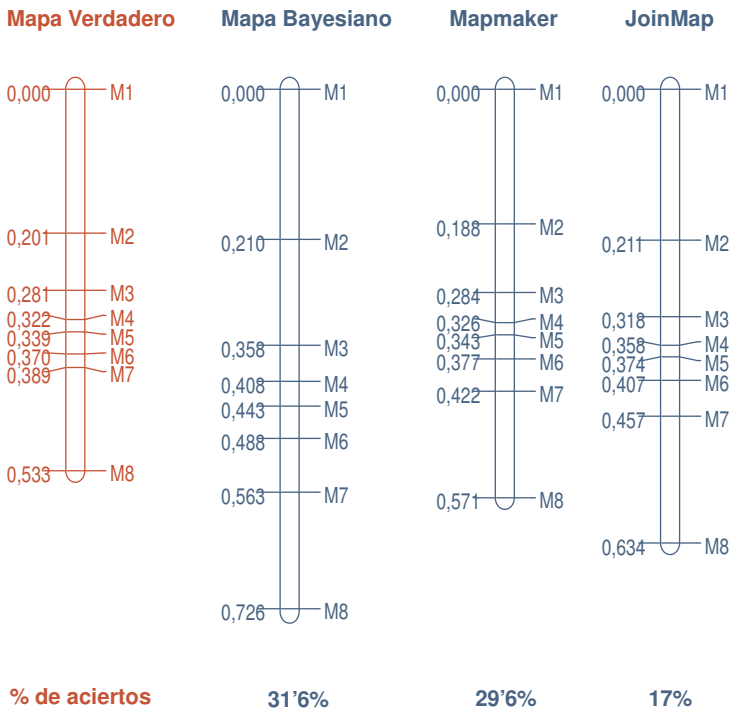


Figura 9.17: *Mapa de una población F_2 con 8 marcadores, no todos codominantes, junto con los modelos bayesiano, Mapmaker y JoinMap, para muestras de 50 individuos.*

En la Figura 9.18 se representan las 158 muestras de 500 cuyo modelo más probable ha resultado ser el correcto junto con las probabilidades de los segundos modelos más probables.

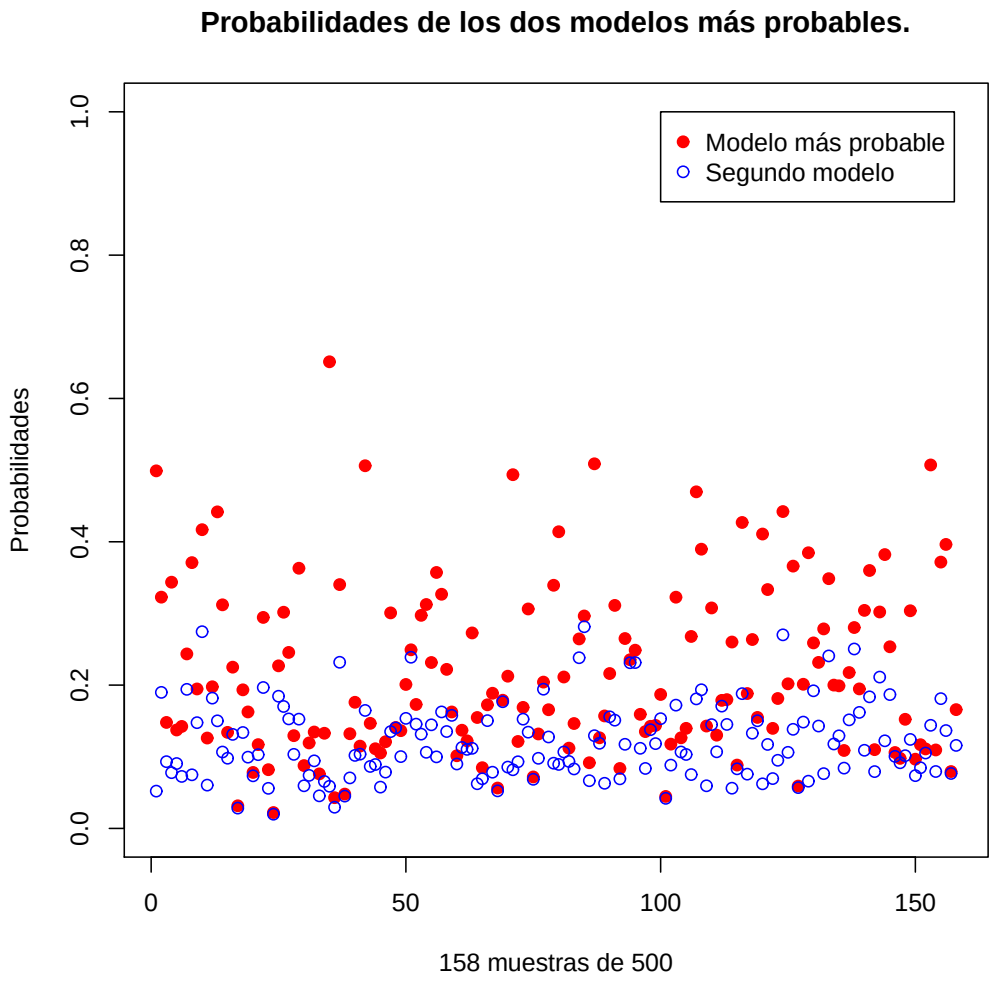


Figura 9.18: *Probabilidades de los dos mejores modelos de las muestras cuyo modelo más probable ha resultado ser el correcto.*

Un resultado obtenido con nuestro método, que compara los mejores modelos frente al resto es el que se ofrece en el Cuadro 9.18. En él se muestra el promedio de las probabilidades asociadas a los modelos que ocupan la posición j-ésima, dado que el modelo correcto ha salido en la posición i-ésima.

	veces en 500 muestras	j=1	j=2	j=3
i=1	158	**0.2174828**	0.1202916	0.06544908
i=2	64	0.1573313	**0.1209118**	0.06289023
i=3	52	0.1472714	0.09549471	**0.06904495**

Cuadro 9.18: *Para la distribución con muestras de 50 individuos, promedio de las probabilidades asociadas a los modelos que ocupan la posición j-ésima, dado que el modelo correcto ha salido en la posición i-ésima.*

9.4. Discusión

Los escenarios investigados han sido tres poblaciones con el mismo mapa genético, Figura 1.11, (Retrocruce, F_2 con todos los marcadores codominantes y F_2 con marcadores dominantes y codominantes conjuntamente) por tres tamaños muestrales (50, 100 y 200 individuos), dando lugar a un total de 9 estudios realizados con 500 muestras cada uno. En 7 de ellos, la metodología bayesiana obtiene porcentajes de acierto más elevados, respecto a la ordenación de los marcadores, que los obtenidos por los programas Mapmaker [48] y JoinMap [77]. Tan sólo en los casos en que las muestras de 50 individuos provienen de Retrocruce o las muestras de 100 individuos provienen de F_2 con no todos los marcadores codominantes, la metodología bayesiana queda en segunda posición, cerca del mejor resultado obtenido por Joinmap y Mapmaker, respectivamente.

Efecto del tamaño muestral en cada población.

En las tres poblaciones e independientemente de la metodología, la pérdida de individuos en el tamaño muestral, se traduce en una reducción del porcentaje de acierto al estimar la ordenación de los marcadores del mapa genético, que implica un aumento de la variabilidad posterior de las distancias medias multipunto, que finalmente repercute en una mayor longitud de la estimación del mapa genético de la población. Concretamente, se observa que con la metodología bayesiana, una reducción de 200 individuos a 50 individuos en el tamaño muestral, ha reducido el porcentaje de acierto de la ordenación de marcadores del 94.8 % al 52.6 % para una población Retrocrece, del 99.4 % al 80.4 % para una población F_2 con todos los marcadores codominates y del

80.2 % al 31.6 % para una población F_2 con marcadores codominantes y dominantes conjuntamente, respectivamente. Tanto en Retrocruce como en F_2 con marcadores codominantes y dominantes, la longitud media estimada, por la metodología bayesiana, con muestras de 200 individuos es prácticamente la real, por lo que al reducir el tamaño muestral la longitud media es sobreestimada. Sin embargo, en la población F_2 con sólo marcadores codominantes, la longitud media estimada por la metodología bayesiana con muestras de 200 individuos es menor que la real. Experimenta un encogimiento al reducir el tamaño muestral a 100 individuos y luego se produce una sobreestimación con muestras de 50 individuos. Este patrón se observa con las tres metodologías. Puede deberse a que en la población F_2 con todos los marcadores codominantes, la reducción del tamaño muestral reduce el porcentaje de aciertos en la ordenación de marcadores de una forma mucho más suave que en las otras dos poblaciones; de 200 a 100 individuos tan sólo se reduce un 5 % el porcentaje de aciertos. Sin embargo, de 100 a 50 individuos este porcentaje se reduce casi un 14 %. En las otras dos poblaciones la reducción del tamaño muestral provoca una reducción del porcentaje de aciertos mucho más drástica: entre el 16 % y el 29 %

Efecto del tipo de población.

Independientemente del la población y del tamaño muestral, el modelo bayesiano medio más probable coincide con el mapa genético real y su probabilidad media asociada, con muestras de 200 individuos, es de 0.8748, 0.9747 y 0.8429 para Retrocruce, para F_2 con todos los marcadores codominantes y para F_2 con marcadores codominantes y dominantes conjuntamente, respectivamente. Como se puede ver, altas probabilidades que determinan inequívocamente la ordenación de los marcadores. Como cabe esperar estas probabilidades medias se reducen en las tres poblaciones, según se reduce el tamaño muestral. La falta de información hace que aparezcan más modelos alternativos, que reducen la frecuencia con que aparece el modelo más repetido y por tanto, su probabilidad de ser observado. Se deduce que las muestras que proceden de la población F_2 con todos los marcadores codominantes estiman con mayor fiabilidad el mapa genético, seguidas de las muestras de la población Retrocruce y de la población F_2 con marcadores codominantes y dominantes.

Independientemente de la población y del tamaño muestral, el programa Mapmaker, siempre ofrece modelos más cortos que los correspondientes a las

otras dos metodologías, tal como también observaron experimentalmente, Ruiz y Asíns (2003) [68]. La longitud obtenida por la metodología bayesiana y la de JoinMap, son similares entre si. Independientemente del porcentaje de aciertos observados respecto a la ordenación de los marcadores, la metodología bayesiana obtiene mapas un poco más largos en Retrocruce, con muestras de 50 individuos, y en F_2 con marcadores codominantes y dominantes. Por el contrario, JoinMap estima una la longitud mayor en Retrocruce, con muestras de 200 y 100 individuos y en F_2 con todos los marcadores codominantes.

Definitivamente, entre los resultados frecuentistas, el programa JoinMap [77] se muestra más eficaz, en cuanto a la correcta ordenación de los marcadores, que Mapmaker [48] en las poblaciones Retrocruce y F_2 con todos los marcadores codominantes. Sin embargo, obtiene resultados bastante peores en la población F_2 con no todos los marcadores codominantes. Quizás, si se hubiera podido automatizar en este Capítulo una versión de JoinMap superior a la 2.0, sus resultados hubieran mejorado. Es posible que las versiones superiores hayan mejorado el problema de estimación de los marcadores dominantes en repulsión. La metodología bayesiana, mejora (en 7 de los escenarios) o prácticamente iguala (en 2 de los escenarios) los resultados de la mejor de las metodologías frecuentistas. Nótese que, aunque 50 individuos es un tamaño muestral irreal en plantas autógamas como cereales, tomates, etc; puede ser normal en el ámbito de los frutales provocado por problemas de espacio. Incluso en esta situación tan desfavorable, la metodología bayesiana se muestra robusta, especialmente en situaciones de falta de información como es el caso de la población F_2 con marcadores codominantes y dominantes en fase de repulsión.

El algoritmo de ordenación bayesiano y el que aplica JoinMap [77], en general, se basan en que las fracciones de recombinación "multipunto" se parezcan a las iniciales, según el método de mínimos cuadrados. Por lo que, si en algún momento se produce cierto sesgo al estimar las fracciones de recombinación iniciales, en parte, es de esperar que dicho sesgo se herede al ajustar las distancias multipunto, bajo cierto orden, y que éste repercuta en la longitud de la estimación del mapa genético. Sin embargo, la metodología multipunto utilizada por Mapmaker (algoritmo EM), sistemáticamente parece que acorte el mapa multipunto. Para investigar este hecho, en el caso más problemático de existencia de incertidumbre en los datos (F_2 con marcadores dominantes), se ha llevado a cabo un pequeño estudio (Apéndice E), sobre muestras de

tamaño 200 individuos, respecto a la influencia de la tolerancia utilizada en el algoritmo EM. El algoritmo EM, según se ha comentado en el Capítulo 1, necesita de una "tolerancia" como condición de parada o convergencia en la estimación de las fracciones de recombinación. Como se puede observar en el Apéndice E, si no se ajusta debidamente este término de tolerancia, es posible que las fracciones de recombinación entre marcadores no se estimen correctamente. Por ejemplo, en el caso en que dos marcadores distan entre 5 cM y 10 cM, el algoritmo EM tiende a subestimar las fracciones de recombinación entre las parejas de marcadores investigadas (excepto la pareja D1-D2), para tolerancias de 10^{-6} a 10^{-3}. Para la pareja D1-D2, el algoritmo puede sobreestimar o subestimar la fracción de recombinación gravemente dependiendo de la distancia que exista entre los marcadores y de la tolerancia utilizada. Éste, puede ser uno de los motivos por el que el programa Mapmaker, casi siempre obtenga longitudes del mapa, más cortas al real.

Capítulo 10

Determinación de grupos de ligamiento.

En los capítulos anteriores hemos trabajado con muestras de datos que provenían de poblaciones definidas por un único cromosoma. Como ya hemos visto, en el Capítulo 8, se ha definido satisfactoriamente una metodología que permite asignar probabilidades a ordenaciones alternativas de los marcadores pertenecientes a un mismo cromosoma o grupo de ligamiento.

Una vez resuelto de forma razonablemente satisfactoria el problema de la estimación de fracciones de recombinación y la ordenación de los marcadores en un único grupo de ligamiento, en este capítulo se va a trabajar en un entorno más realista, sobre poblaciones definidas por varios cromosomas. Para ello, se amplía la metodología para determinar qué marcadores forman parte de un mismo grupo de ligamiento, asignando probabilidades a los alternativos grupos de ligamiento. Una vez determinada la estructura de grupos de ligamiento más probable, se podrían ordenar los marcadores dentro de cada grupo de ligamiento aplicando el algoritmo de ordenación visto en el Capítulo 8.

10.1. Descripción del algoritmo

El algoritmo de agrupación de marcadores en los distintos grupos de ligamiento, se ha elaborado en base al cumplimiento de dos condiciones, que de forma resumida llamaremos asociación y proximidad. A continuación se detalla en qué consisten los dos:

En primer lugar, la condición de asociación se fundamenta en determinar para cada uno de los m marcadores que definen el mapa genético de la población, su asociación con el resto de marcadores. Para ello, se plantean $m(m\text{-}1)/2$ test de independencia:

$$\begin{cases} H_{0_{i,j}}: \text{ Los marcadores } M_i \text{ y } M_j \text{ son independientes.} \\ H_{1_{i,j}}: \text{ Los marcadores } M_i \text{ y } M_j \text{ son dependientes.} \end{cases}$$

con M_i y $M_j \in \{M_1, ..., M_m\},\ M_i \neq M_j$

Asociados a estos $m(m-1)/2$ tests de independencia, podemos evaluar la probabilidad posterior de $H_{1_{i,j}}$ en términos del Factor Bayes, $FB_{i,j}$, que asumiendo $P(H_{0_{i,j}})= 0.5$, se puede expresar como:

$$P(\text{Asociación entre } M_i \text{ y } M_j|Datos) = \frac{FB_{i,j}}{FB_{i,j}+1} \tag{10.1}$$

El factor Bayes es fácilmente evaluable con la función ctable() en el paquete LearnBayes [3] de R [62], a partir de la tabla de contingencia que genera el vector de frecuencias de recombinación observadas $R_{ij} = (R_{1,i,j}, ..., R_{ngeno,i,j})$, con *ngeno* el número de genotipos distinguibles para la pareja de marcadores M_i y M_j. Así, las probabilidades (10.1) se evalúan también rápidamente.

A continuación, utilizamos la propuesta de resolución de un problema de comparaciones múltiples desde la perspectiva Bayesiana dada por Müller et al. (2006) [55], adaptando el método para comparaciones múltiples de Benjamini y Hochberg (1995) [5], que tiene en cuenta el número de comparaciones múltiples planteadas y el nivel de significatividad global α pretendido. Dado el "ranking" de las probabilidades (10.1), dicho procedimiento propone calcular una probabilidad umbral w_{j*} entre todas ellas, y rechazar las hipótesis de independencia a favor de la asociación para todas aquellas parejas de marcadores M_i y M_j que la superen.

Como resultado, para cada marcador M_i se obtiene una lista de marcadores (listado1) con los que aceptamos asociación:

$$list1.M_i = \{M_j \in \{M_1, ...M_m\}|\ P(\text{Asociación entre } M_i \text{ y } M_j|Datos) > w_j*\}$$

Alternativamente, en este paso, se podría haber optado por resolver $H_0 : r_{i,j} = 0.5$ (Véase (1.8)). Se ha preferido el enfoque anterior porque no necesita asumir ausencia de distorsión en la segregación, según Van Ooijen 2006 [85] (apartado Grouping test statistics)

En segundo lugar, para aplicar la condición de proximidad se obtiene la distribución posterior conjunta de las fracciones de recombinación, simulando independientemente de cada fracción de recombinación, $r_{i,j}$, dado el vector de frecuencias de recombinación $\mathbf{R_{i,j}}$ que determina cada pareja de marcadores distinta, M_i y M_j. Esta cadena de simulación, $\{r^{(t)}\}_{t=1}^{nsim}$, se obtiene en idénticos términos y bajo la misma modelización que se utilizó en el Capítulo 8. Recordamos que la cadena de simulación consta de *nsim* iteraciones o filas y $m(m\text{-}1)/2$ columnas o fracciones de recombinación entre todas las parejas de marcadores. En base a esta cadena de simulación, en cada iteración, se elabora para cada marcador una lista de marcadores próximos a él (listado 2), entendiendo como próximo al marcador de referencia todo aquel que mantenga una fracción de recombinación estimada menor o igual a un valor establecido por el investigador genético, que en la presente tesis se ha fijado en 0.35 M. Por ejemplo, el listado 2 para M_i es:

$$list2.M_i = \{M_j \in \{M_1, ..., M_m\},\ M_j \neq M_i / r_{i,j} < 0.35\}$$

Nótese que el listado 1 es el mismo en todo el proceso y es independiente de las fracciones de recombinación estimadas para cada pareja de marcadores. Sin embargo, el listado 2 puede variar en cada iteración, ya que en una de las iteraciones un marcador puede considerarse próximo a otro y en la siguiente puede considerarse alejado.

Por último, combinando estos dos listados en cada iteración, se obtiene para cada marcador un listado preliminar de marcadores que cumplen las dos condiciones deseadas: ser próximos y además asociados al marcador de referencia. Por ejemplo, el listado preliminar para M_i es:

$$list.pre.M_i = \{list1.M_i \cap list2.M_i\}$$

Una vez conocida esta clasificación, se elabora la agrupación definitiva de marcadores que forman los grupos de ligamiento. Este proceso se lleva a cabo en cada iteración, aplicando la propiedad transitiva. Es decir, si un marcador "M_i" está agrupado a un marcador "M_j" según la lista preliminar y,

a su vez, el marcador "M_j" está agrupado a un marcador "M_k", se considera que los marcadores "M_i", "M_j" y "M_k" deben pertenecen al mismo grupo de ligamiento, $grupo_s$. Es decir, un grupo de ligamiento se forma como:

$$grupo\ s = \{\bigcup_i list.pre.M_i / \bigcap_i list.pre.M_i \neq \{\emptyset\}\}$$

con s= 1,...,$n.gr$, donde $n.gr$ representa el número de grupos de ligamiento diferentes.

Finalizado el proceso, se obtiene una cadena, $\{gr^{(t)}\}_{t=1}^{nsim}$. En cada iteración, t, se guarda la estructura de grupos de ligamiento correspondiente,

$$gr^{(t)} = \{\bigcup_{s=1}^{n.gr} \{grupo\ s\} / \bigcap_{s=1}^{n.gr} \{grupo\ s\} = \{\emptyset\}\}.$$

La estructura que se repite más veces en la cadena se considera la estructura de grupos de ligamiento más probable y su probabilidad asociada coincide con la proporción de veces que aparece en la cadena.

10.2. Resultados

Para valorar la bondad de la metodología se han diseñado dos mapas genéticos. Uno de ellos, definido por 63 marcadores, repartidos en 6 grupos de ligamiento (Figura 1.12) y otro definido por 290 marcadores, repartidos en 20 grupos de ligamiento (Figura 1.13). Es decir, vamos a considerar un problema con "pocos" grupos de ligamiento y otro con "muchos" grupos de ligamiento. El objetivo es observar el efecto del número de cromosomas en la utilización del nivel global de significatividad. Del primer mapa genético se han extraído 2 muestras de 200 individuos, cada una de ellas. Una de las muestras se ha extraído considerando que el mapa representa a una población Retrocruce, a la que hemos llamado 6Gr63MarcRetro, y la otra muestra considerando que proviene de una población F_2, en la que algunos marcadores son codominantes y otros dominantes (6Gr63MarcF2). El segundo mapa genético sólo representa una población F_2, en la que no todos los marcadores son codominantes. De esta población se ha extraído una única muestra de 200 individuos (20Gr290Marc).

El diseño esquemático de los dos mapas genéticos se detalla a continuación:

El reparto real de los 63 marcadores en los grupos de ligamiento, según el primer mapa genético prediseñado es:

grupo 1= {1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12}
grupo 2= {13, 14, 15, 16, 17, 18, 19, 20}
grupo 3= {21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32}
grupo 4= {33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46}
grupo 5= {47, 48, 49, 50, 51, 52, 53, 54}
grupo 6= {55, 56, 57, 58, 59, 60, 61, 62, 63}

El tipo de cada marcador prefijado para el caso en el que el mapa representa una población F_2 es:

grupo 1= {C, C, D1, D2, C, D2, D1, C, D1, D1, C, C}
grupo 2= {C, C, C, C, C, C, C, C}
grupo 3= {D1, C, C, D2, D1, D1, C, C, D1, C, D1, D1}
grupo 4= {D1, C, C, D1, D2, D1, C, D1, D1, D2, D2, C, C, C}
grupo 5= {D2, D2, D2, D2, D2, D2, D2, D2}
grupo 6= {D1, D1, D1, D1, D1, D1, D1, D1, D1}

Equivalentemente, el reparto real de los 290 marcadores del segundo mapa genético diseñado según una población F_2, junto al tipo de marcador considerado es:

grupo 1 = {1(D1), 2(C), 3(C), 4(C), 5(C), 6(D1), 7(D2), 8(C), 9(D1), 10(D2), 11(C), 12(D2), 13(C), 14(D2)}

grupo 2 = {15(D1), 16(D1), 17(D2), 18(D2), 19(D2), 20(C), 21(D2), 22(C), 23(C), 24(D2), 25(D1), 26(D2), 27(D2), 28(D2), 29(C), 30(D1)}

grupo 3 = {31(D2), 32(D1), 33(D1), 34(D2), 35(D2), 36(C), 37(D2), 38(D1), 39(D1), 40(D1), 41(D2), 42(D2), 43(C), 44(D2)}

grupo 4 = {45(C), 46(D2), 47(C), 48(C), 49(D1), 50(D1), 51(D2), 52(D1), 53(D2), 54(D2), 55(D2), 56(D2), 57(D2), 58(D2), 59(D1), 60(D2)}

grupo 5 = {61(D1), 62(D1), 63(D1), 64(D2), 65(C), 66(C), 67(D1), 68(D1), 69(D1), 70(C), 71(C), 72(C), 73(D2)}

grupo 6 = {74(C), 75(D2), 76(C), 77(D1), 78(D2), 79(D2), 80(C), 81(D2), 82(C), 83(D1), 84(C), 85(D1), 86(D1)}

grupo 7 = {87(C), 88(C), 89(C), 90(D2), 91(D1), 92(D1), 93(D1), 94(D2), 95(C), 96(C), 97(D2), 98(D1), 99(D2), 100(C), 101(D1)}

grupo 8 = {102(D1), 103(D2), 104(D1), 105(C), 106(D1), 107(C), 108(D1), 109(D2), 110(C), 111(D1), 112(D2), 113(D2), 114(D1), 115(C)}

grupo 9 = {116(D1), 117(C), 118(D2), 119(D2), 120(C), 121(C), 122(D2), 123(C), 124(C), 125(D2), 126(D1), 127(C), 128(C), 129(D2), 130(D2), 131(C), 132(C)}

grupo 10 = {133(D1), 134(C), 135(D2), 136(D2), 137(D2), 138(C), 139(D2), 140(D2), 141(C)}

grupo 11 = {142(D1), 143(C), 144(C), 145(D1), 146(D1), 147(D2), 148(C), 149(C), 150(C), 151(C), 152(C), 153(C), 154(D1), 155(C), 156(D1), 157(C), 158(C)}

grupo 12 = {159(C), 160(D1), 161(D1), 162(C), 163(D1), 164(D2), 165(D1), 166(C), 167(C), 168(D2), 169(D2), 170(C)}

grupo 13 = {171(D2), 172(D2), 173(D2), 174(D2), 175(D2), 176(D1), 177(C), 178(D2), 179(C), 180(D2), 181(D1), 182(D2), 183(D2), 184(D2), 185(C), 186(D1), 187(D1), 188(D2)}

grupo 14 = {189(D1), 190(D1), 191(D2), 192(C), 193(D2), 194(C), 195(C), 196(C), 197(D1), 198(C), 199(D2)}

grupo 15 = {200(D1), 201(D2), 202(D1), 203(D2), 204(C), 205(D1), 206(C), 207(C), 208(D1), 209(D1), 210(D2), 211(D2), 212(C), 213(D2), 214(D1), 215(D2)}

grupo 16 = {216(C), 217(D1), 218(C), 219(D2), 220(D2), 221(C), 222(D2), 223(D2), 224(C), 225(C), 226(C), 227(C), 228(C), 229(C), 230(D2), 231(D2)}

grupo 17 = {232(D2), 233(D1), 234(C), 235(D1), 236(D2), 237(D1), 238(D1), 239(D2), 240(C), 241(C), 242(C), 243(C), 244(C), 245(C), 246(D1), 247(C)}

grupo 18 = {248(D2), 249(D1), 250(D1), 251(D1), 252(D1), 253(C), 254(D1), 255(D2), 256(D2), 257(C), 258(C), 259(D1), 260(D1), 261(D1), 262(C), 263(C)}

grupo 19 = {264(D1), 265(C), 266(D1), 267(C), 268(D1), 269(C), 270(D1), 271(C), 272(C), 273(C), 274(D1), 275(D2), 276(D2), 277(D2), 278(D1)}

grupo 20 = {279(D1), 280(D1), 281(C), 282(D2), 283(D1), 284(D2), 285(C), 286(C), 287(C), 288(C), 289(D2), 290(C)}

A continuación, se ofrecen los resultados obtenidos por la metodología bayesiana para las tres muestras y una comparativa con los resultados obtenidos mediante la metodología frecuentista según los programas de referencia: Mapmaker [48] y JoinMap [85]. La versión de JoinMap utilizada en adelante ha sido la 4.0, en lugar de la 2.0, que se utilizó en el capítulo anterior, por las limitaciones que ya se comentaron.

Con el objetivo de valorar su posible efecto, se ha ensayado la metodología bajo tres niveles de significatividad globales (α= 0.05, α= 0.01, α= 0.005). Del mismo modo, en la comparativa de resultados de los programas de referencia, se han ofrecido los resultados bajo los tres LOD más habituales (LOD= 3, LOD= 4, LOD= 5).

En las tablas, el modelo bayesiano ofrece distintas agrupaciones alternativas (filas), cada una de ellas valorada según su probabilidad. Sin embargo, los programas de referencia, una única agrupación, sin ninguna medida que cuantifique su bondad o fiabilidad. Nótese que, sólo se ha especificado el reparto de marcadores en la primera agrupación. En este capítulo no se ha estudiado el orden de los marcadores, sólo la agrupación de los mismos en los distintos grupos de ligamiento. Cuando el nombre de un grupo se expresa con varios índices, separados por puntos, significa que los marcadores que lo componen son unión de los grupos originales llamados con esos índices. Por ejemplo, el "grupo 2.20" contiene los marcadores del "grupo 2" y del "grupo 20".

6Gr63Marc-Retro	modelo bayesiano		
	α=0.05	α=0.01	α=0.005
Agrupación 1	6 grupos de ligamiento con prob= 0.9995 grupo 1 grupo 2 grupo 3 grupo 4 grupo 5 grupo 6	6 grupos de ligamiento con prob= 0.9995 grupo 1 grupo 2 grupo 3 grupo 4 grupo 5 grupo 6	6 grupos de ligamiento con prob= 0.9995 grupo 1 grupo 2 grupo 3 grupo 4 grupo 5 grupo 6
Agrupación 2	7 grupos prob= 0.0005	7 grupos prob= 0.0005	7 grupos prob= 0.0005

(a) Agrupaciones alternativas obtenidas por la metodología bayesiana, considerando que la muestra proviene de la población Retrocruce. Resultados obtenidos bajo tres niveles de significatividad globales

6Gr63Marc-Retro	JoinMap		
	LOD 3	LOD 4	LOD 5
Agrupación única	6 grupos de ligamiento grupo 1 grupo 2 grupo 3 grupo 4 grupo 5 grupo 6	6 grupos de ligamiento grupo 1 grupo 2 grupo 3 grupo 4 grupo 5 grupo 6	6 grupos de ligamiento grupo 1 grupo 2 grupo 3 grupo 4 grupo 5 grupo 6

(b) Agrupación obtenida por JoinMap, considerando que la muestra proviene de la población Retrocruce. Resultados obtenidos según tres LODs.

6Gr63Marc-Retro	MapMaker		
	LOD 3	LOD 4	LOD 5
Agrupación única	6 grupos de ligamiento grupo 1 grupo 2 grupo 3 grupo 4 grupo 5 grupo 6	6 grupos de ligamiento grupo 1 grupo 2 grupo 3 grupo 4 grupo 5 grupo 6	6 grupos de ligamiento grupo 1 grupo 2 grupo 3 grupo 4 grupo 5 grupo 6

(c) Agrupación obtenida por MapMaker, considerando que la muestra proviene de la población Retrocruce. Resultados obtenidos según tres LODs.

Cuadro 10.1: *Grupos de ligamiento según la metodología bayesiana, respecto a la muestra 6Gr63MarcRetro, y comparativa con los resultados que ofrecen JoinMap y Mapmaker.*

6Gr63MarcF2	modelo bayesiano		
	α=0.05	α=0.01	α=0.005
Agrupación 1	6 grupos de ligamiento con prob= 0.997 grupo 1 grupo 2 grupo 3 grupo 4 grupo 5 grupo 6	6 grupos de ligamiento con prob= 0.982 grupo 1 grupo 2 grupo 3 grupo 4 grupo 5 grupo 6	6 grupos de ligamiento con prob= 0.9815 grupo 1 grupo 2 grupo 3 grupo 4 grupo 5 grupo 6
Agrupación 2	7 grupos prob= 0.002	7 grupos prob= 0.0095	7 grupos prob= 0.0095
Agrupación 3	7 grupos prob= 0.0005	7 grupos prob= 0.0075	7 grupos prob= 0.0075
Agrupación 4	7 grupos prob= 0.0005	7 grupos prob= 0.0005	7 grupos prob= 0.001
Agrupación 5		7 grupos prob= 0.0005	7 grupos prob= 0.0005

(a) Agrupaciones alternativas con la metodología bayesiana; muestra de la población F_2, con no todos los marcadores codominantes. Resultados obtenidos bajo tres niveles de significatividad globales

6Gr63MarcF2	JoinMap		
	LOD 3	LOD 4	LOD 5
Agrupación única	6 grupos de ligamiento grupo 1 grupo 2 grupo 3 grupo 4 grupo 5 grupo 6	6 grupos de ligamiento grupo 1 grupo 2 grupo 3 grupo 4 grupo 5 grupo 6	6 grupos de ligamiento grupo 1 grupo 2 grupo 3 grupo 4 grupo 5 grupo 6

(b) Agrupación por JoinMap; muestra de la población F_2, con no todos los marcadores codominantes. Resultados según tres LODs.

6Gr63MarcF2	MapMaker		
	LOD 3	LOD 4	LOD 5
Agrupación única	6 grupos de ligamiento grupo 1 grupo 2 grupo 3 grupo 4 grupo 5 grupo 6	6 grupos de ligamiento grupo 1 grupo 2 grupo 3 grupo 4 grupo 5 grupo 6	6 grupos de ligamiento grupo 1 grupo 2 grupo 3 grupo 4 grupo 5 grupo 6

(c) Agrupación por MapMaker; muestra de la población F_2, con no todos los marcadores codominantes. Resultados según tres LODs.

Cuadro 10.2: *Grupos de ligamiento según la metodología bayesiana, respecto a la muestra 6Gr63MarcF2, y comparativa con los resultados que ofrecen JoinMap y Mapmaker.*

20Gr290Marc	modelo bayesiano		
	α=0.05	α=0.01	α=0.005
Agrupación 1	18 grupos (prob= 0.269) grupo 1 grupo 2.20 grupo 3 grupo 4.5 grupo 6 grupo 7 grupo 8 grupo 9 grupo 10 grupo 11 grupo 12 grupo 13 grupo 14 grupo 15 grupo 16 grupo 17 grupo 18 grupo 19	19 grupos (prob= 0.974) grupo 1 grupo 2.20 grupo 3 grupo 4 grupo 5 grupo 6 grupo 7 grupo 8 grupo 9 grupo 10 grupo 11 grupo 12 grupo 13 grupo 14 grupo 15 grupo 16 grupo 17 grupo 18 grupo 19	20 grupos (prob= 1) grupo 1 grupo 2 grupo 3 grupo 4 grupo 5 grupo 6 grupo 7 grupo 8 grupo 9 grupo 10 grupo 11 grupo 12 grupo 13 grupo 14 grupo 15 grupo 16 grupo 17 grupo 18 grupo 19 grupo 20
Agrupación 2	19 grupos con prob= 0.2685	20 grupos con prob= 0.026	
Agrupación 3	18 grupos con prob= 0.2275		
Agrupación 4	17 grupos con prob= 0.209		
Agrupación 5	19 grupos con prob= 0.009		
Agrupación 6	20 grupos con prob= 0.007		
Agrupación 7	19 grupos con prob= 0.0065		
Agrupación 8	18 grupos con prob= 0.0035		

20Gr290Marc	JoinMap		
	LOD 3	LOD 4	LOD 5
Agrupación única	4 grupos grupo 1.4.5.16 grupo 2.3.6.7. 8.9.10.12.13.14.15.19.20 grupo 11 grupo 17.18	17 grupos grupo 1 grupo 2.20 grupo 3 grupo 4.5 grupo 6 grupo 7.19 grupo 8 grupo 9 grupo 10 grupo 11 grupo 12 grupo 13 grupo 14 grupo 15 grupo 16 grupo 17 grupo 18	20 grupos grupo 1 grupo 2 grupo 3 grupo 4 grupo 5 grupo 6 grupo 7 grupo 8 grupo 9 grupo 10 grupo 11 grupo 12 grupo 13 grupo 14 grupo 15 grupo 16 grupo 17 grupo 18 grupo 19 grupo 20

20Gr290Marc	MapMaker		
	LOD 3	LOD 4	LOD 5
Agrupación única	14 grupos	18 grupos	20 grupos
	grupo 1	grupo 1	grupo 1
	grupo 2.3.1 3.15.19.20	grupo 2.20	grupo 2
	grupo 4	grupo 3.15	grupo 3
	grupo 5	grupo 4	grupo 4
	grupo 6	grupo 5	grupo 5
	grupo 7	grupo 6	grupo 6
	grupo 8	grupo 7	grupo 7
	grupo 9	grupo 8	grupo 8
	grupo 10	grupo 9	grupo 9
	grupo 11	grupo 10	grupo 10
	grupo 12	grupo 11	grupo 11
	grupo 14	grupo 12	grupo 12
	grupo 16	grupo 13	grupo 13
	grupo 17.18	grupo 14	grupo 14
		grupo 16	grupo 15
		grupo 17	grupo 16
		grupo 18	grupo 17
		grupo 19	grupo 18
			grupo 19
			grupo 20

10.3. Discusión

Según se aprecia en el Cuadro 10.1 (tablas a, b y c) y en el Cuadro 10.2 (tablas d, e y f), la metodología bayesiana es capaz de igualar los resultados ofrecidos por los programas de referencia. Cada uno de los 63 marcadores se clasifica en el grupo correcto. Además, en cada caso, la estructura de grupos de ligamiento más probable es casi inequívoca, obteniendo probabilidades superiores a 0.98, muy alejadas de las siguientes alternativas. No se observan diferencias entre los resultados obtenidos para la muestra que representa a una población Retrocruce y a una población F_2. Tan sólo para esta última población, parece que a medida que disminuye el nivel de significatividad exigido en la construcción de los grupos, disminuye levemente la probabilidad obtenida por la estructura de grupos de ligamiento más probable.

Seguidamente, para el caso referente a la población F_2 definida por 290 marcadores repartidos en 20 grupos de ligamiento, la metodología bayesiana también obtiene resultados satisfactorios. Mientras los programas de referencia varían considerablemente el número de grupos de ligamiento obtenido según el LOD aplicado, con la metodología bayesiana no se aprecia esta variación respecto a los niveles de significatividad considerados. En este caso, para la metodología frecuentista el aumento del LOD supone un aumento en el número de grupos de ligamiento. Parece que el programa Mapmaker [48] se muestra más eficaz y estable en el reparto de marcadores en los diferentes grupos de ligamiento según aumenta el LOD. Con la metodología bayesiana, a medida

que disminuye el nivel de significatividad exigido en la construcción de los grupos, aumenta tanto el número de grupos de ligamiento como la probabilidad obtenida para la estructura de grupos de ligamiento más probable, llegando a ser inequívoca (probabilidad= 1), cuando se exige un nivel de significatividad global de $\alpha = 0.005$. Aunque no se ha profundizado en este trabajo de investigación, se podría explorar cuál sería el nivel de significatividad global recomendable para resolver los test de independencia, en función tanto del tipo de población como del número de marcadores implicados en el mapa genético de la población que se desea estimar.

Según los resultados anteriores, se concluye que la metodología elaborada para determinar grupos de ligamiento es satisfactoria ya que reparte los marcadores de manera correcta en cada cromosoma. Además, permite cuantificar la probabilidad de este reparto ofreciendo distintas estructuras alternativas.

Cuanto mayor es el número de marcadores involucrados en el mapa genético, mayor número de test de independencia se necesitan realizar para llevar a cabo el algoritmo, por lo que el nivel de significatividad global exigido debe ser más pequeño para obtener el reparto de marcadores correcto. Igualmente, si el número cromosómico de la especie en estudio es alto, sería conveniente emplear un nivel de significatividad del orden 0.005. De todos modos, tal como se actúa con métodos frecuentistas tradicionales respecto al LOD, es conveniente emplear un amplio rango de niveles de significatividad para ofrecer información al investigador sobre la estabilidad de los grupos de ligamiento.

Aun así, la metodología bayesiana, a diferencia de los métodos frecuentistas, se comporta de forma estable. El cambio de un nivel de significatividad global de $\alpha = 0.05$ a $\alpha = 0.005$ no supone un aumento excesivo en el número de grupos de ligamiento obtenido. Sin embargo, en los programas de referencia frecuentista el cambio de LOD 3 a LOD 5 sí supone un cambio sustancial en el número de grupos de ligamiento obtenido. Como se puede observar en los resultados anteriores, en el caso de una población con 20 grupos de ligamiento, si las metodologías frecuentistas utilizaran LOD 3, considerado valor estándar equiparable a $\alpha = 0.05$, se obtendrían resultados desastrosos, especialmente en el caso de JoinMap.

Capítulo 11

Efecto de los datos faltantes sobre la estabilidad en la determinación de grupos de ligamiento y sobre la ordenación de marcadores.

Como resultado de los dos capítulos anteriores se ha obtenido una metodología que, en base a una muestra de datos, permite determinar el reparto de marcadores en distintos grupos de ligamiento y además ordenar los marcadores dentro de cada grupo de ligamiento, obteniendo así una estimación del mapa genético de la población de la que procede la muestra de datos. Como ya se ha visto, tanto la estructura de grupos de ligamiento como la ordenación de los marcadores dentro de cada grupo de ligamiento obtienen distintas probabilidades que permiten cuantificar la bondad o fiabilidad de estos resultados.

En este capítulo se desea valorar la sensibilidad del procedimiento descrito, en un entorno más realista como es el de muestras de datos con datos faltantes.

Para llevar a cabo la experimentación se va a ensayar la metodología completa proponiendo distintos porcentajes de datos faltantes sobre la muestra, 20Gr290Marc, procedente del segundo mapa genético considerado en el capítulo anterior, que representa una población F_2, definida por 290 marcadores,

no todos ellos codominantes, repartidos en 20 grupos de ligamiento. La representación gráfica del mapa genético real de la población de la que se ha extraído la muestra aparece en la Figura 1.13

El motivo de esta elección es trabajar en el entorno más desfavorable de todos los que se han tenido en cuenta anteriormente, por el tipo de población y por el número de marcadores implicados.

En primer lugar, se aplica la metodología sobre la muestra completa. Es decir, se determinan los grupos de ligamiento y se ordenan los marcadores dentro de cada grupo de ligamiento en base a la muestra, con un tamaño muestral de 200 individuos, con un 0 % de datos faltantes. A continuación, se eliminan aleatoriamente el 15 % de los datos de la muestra completa y se repite la metodología y para finalizar, se reitera el proceso eliminando aleatoriamente el 25 % de los datos de la muestra completa.

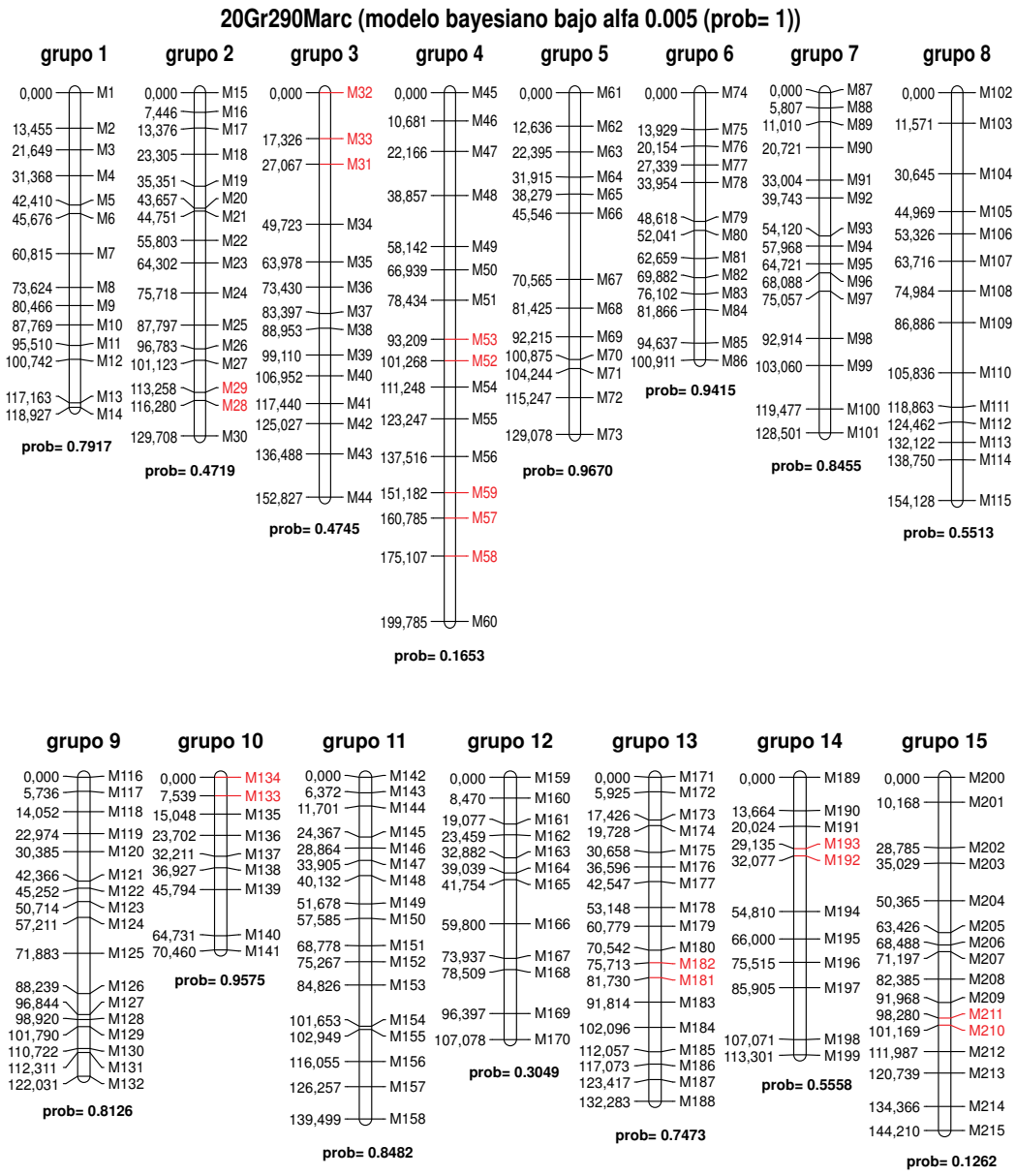
Los resultados de los tres escenarios aparecen en los siguientes apartados. En cada apartado se detalla, mediante tablas, la determinación de los grupos de ligamiento según la metodología bayesiana, para los mismos niveles de significatividad global que se han utilizado en el capítulo anterior (α= 0.05, α= 0.01 y α= 0.005). A continuación, se ofrecen las tablas con los resultados equivalentes obtenidos por JoinMap [85] y Mapmaker [48] a LOD 3, LOD 4 y LOD 5. Nótese que la tabla referente a la metodología bayesiana contiene distintas agrupaciones alternativas acompañadas de su probabilidad. Sin embargo, las tablas obtenidas por los programas de referencia sólo contienen una única agrupación sin ninguna medida que permita cuantificar su bondad. De cada una de las tres tablas, se selecciona la agrupación de marcadores que mejor reproduce el mapa genético real y se representa gráficamente una estimación del mapa genético en función del orden de los marcadores obtenido y de las distancias multipunto estimadas entre los marcadores contiguos. De nuevo, obsérvese que sólo en los resultados obtenidos por la metodología bayesiana se cuantifica la bondad de la ordenación de los marcadores dentro de cada grupo de ligamiento por la probabilidad que aparece bajo de cada grupo, en cada representación gráfica. Para facilitar la comparativa, cuando un grupo de ligamiento real ha sido fragmentado en subgrupos en la estimación, los subgrupos han sido mostrados uno bajo otro en la respectiva representación gráfica.

11.1. Resultados para la muestra con un 0 % de datos faltantes

Como ya se vio en el capítulo anterior, la metodología bayesiana obtiene la agrupación correcta de los marcadores, para la muestra con un 0 % de datos faltantes, de forma inequívoca (probabilidad= 1), bajo un nivel de significatividad global $\alpha = 0.005$. Del mismo modo, tanto JoinMap [85] como Mapmaker [48] obtienen el reparto correcto de los marcadores bajo LOD 5. En base a estas agrupaciones se obtienen las estimaciones del mapa genético de la población de la que procede la muestra 20Gr290Marc (Figuras 11.1, 11.2 y 11.3); con ello finaliza todo el proceso de estimación del mapa genético: estimación de distancias entre marcadores, formación de grupos de ligamiento y ordenación de los marcadores dentro de cada grupo.

Como se puede observar en la Figura 11.3, referente a la ordenación de marcadores que ofrece Mapmaker [48], en casi todos los grupos de ligamiento quedan sin ubicar de forma clara alguno o varios de sus marcadores, por lo que a partir de este resultado preliminar sería necesario ir introduciendo los marcadores manualmente según las recomendaciones maximoverosímiles que ofrece el programa. Es por ello, por lo que nos centramos en la comparativa del modelo bayesiano y el resultado obtenido por JoinMap [85].

En base a las Figuras 11.1 y 11.2 se aprecian resultados similares en la ordenación de los marcadores. En los grupos de ligamiento 1, 5, 10, 11, 14, 15, 16 y 17 la metodología bayesiana obtiene una estimación de la longitud del mapa genético más exacta que el programa JoinMap [85]. Cabe señalar, que aunque la ordenación de marcadores más probable del grupo 16 tiene una probabilidad de 0.4634 y difiere del mapa genético real en la permutación de los marcadores M222 y M223, la segunda ordenación más probable de este mismo grupo de ligamiento tiene una probabilidad prácticamente igual, 0.4585, y coincide con el mapa genético real. Respecto a los grupos de ligamiento 3 y 4, la metodología bayesiana tiene mayor dificultad para obtener la ordenación correcta de sus marcadores que el programa JoinMap [85].


20Gr290Marc (modelo bayesiano bajo alfa 0.005 (prob= 1))
grupo 1
0,000 M1
13,455 M2
21,649 M3
31,368 M4
42,410 M5
45,676 M6
60,815 M7
73,624 M8
80,466 M9
87,769 M10
95,510 M11
100,742 M12
117,163 M13
118,927 M14
prob= 0.7917
grupo 2
0,000 M15
7,446 M16
13,376 M17
23,305 M18
35,351 M19
43,657 M20
44,751 M21
55,803 M22
64,302 M23
75,718 M24
87,797 M25
96,783 M26
101,123 M27
113,258 M29
116,280 M28
129,708 M30
prob= 0.4719
grupo 3
0,000 M32
17,326 M33
27,067 M31
49,723 M34
63,978 M35
73,430 M36
83,397 M37
88,953 M38
99,110 M39
106,952 M40
117,440 M41
125,027 M42
136,488 M43
152,827 M44
prob= 0.4745
grupo 4
0,000 M45
10,681 M46
22,166 M47
38,857 M48
58,142 M49
66,939 M50
78,434 M51
93,209 M53
101,268 M52
111,248 M54
123,247 M55
137,516 M56
151,182 M59
160,785 M57
175,107 M58
199,785 M60
prob= 0.1653
grupo 5
0,000 M61
12,636 M62
22,395 M63
31,915 M64
38,279 M65
45,546 M66
70,565 M67
81,425 M68
92,215 M69
100,875 M70
104,244 M71
115,247 M72
129,078 M73
prob= 0.9670
grupo 6
0,000 M74
13,929 M75
20,154 M76
27,339 M77
33,954 M78
48,618 M79
52,041 M80
62,659 M81
69,882 M82
76,102 M83
81,866 M84
94,637 M85
100,911 M86
prob= 0.9415
grupo 7
0,000 M87
5,807 M88
11,010 M89
20,721 M90
33,004 M91
39,743 M92
54,120 M93
57,968 M94
64,721 M95
68,088 M96
75,057 M97
92,914 M98
103,060 M99
119,477 M100
128,501 M101
prob= 0.8455
grupo 8
0,000 M102
11,571 M103
30,645 M104
44,969 M105
53,326 M106
63,716 M107
74,984 M108
86,886 M109
105,836 M110
118,863 M111
124,462 M112
132,122 M113
138,750 M114
154,128 M115
prob= 0.5513
grupo 9
0,000 M116
5,736 M117
14,052 M118
22,974 M119
30,385 M120
42,366 M121
45,252 M122
50,714 M123
57,211 M124
71,883 M125
88,239 M126
96,844 M127
98,920 M128
101,790 M129
110,722 M130
112,311 M131
122,031 M132
prob= 0.8126
grupo 10
0,000 M134
7,539 M133
15,048 M135
23,702 M136
32,211 M137
36,927 M138
45,794 M139
64,731 M140
70,460 M141
prob= 0.9575
grupo 11
0,000 M142
6,372 M143
11,701 M144
24,367 M145
28,864 M146
33,905 M147
40,132 M148
51,678 M149
57,585 M150
68,778 M151
75,267 M152
84,826 M153
101,653 M154
102,949 M155
116,055 M156
126,257 M157
139,499 M158
prob= 0.8482
grupo 12
0,000 M159
8,470 M160
19,077 M161
23,459 M162
32,882 M163
39,039 M164
41,754 M165
59,800 M166
73,937 M167
78,509 M168
96,397 M169
107,078 M170
prob= 0.3049
grupo 13
0,000 M171
5,925 M172
17,426 M173
19,728 M174
30,658 M175
36,596 M176
42,547 M177
53,148 M178
60,779 M179
70,542 M180
75,713 M182
81,730 M181
91,814 M183
102,096 M184
112,057 M185
117,073 M186
123,417 M187
132,283 M188
prob= 0.7473
grupo 14
0,000 M189
13,664 M190
20,024 M191
29,135 M193
32,077 M192
54,810 M194
66,000 M195
75,515 M196
85,905 M197
107,071 M198
113,301 M199
prob= 0.5558
grupo 15
0,000 M200
10,168 M201
28,785 M202
35,029 M203
50,365 M204
63,426 M205
68,488 M206
71,197 M207
82,385 M208
91,968 M209
98,280 M211
101,169 M210
111,987 M212
120,739 M213
134,366 M214
144,210 M215
prob= 0.1262

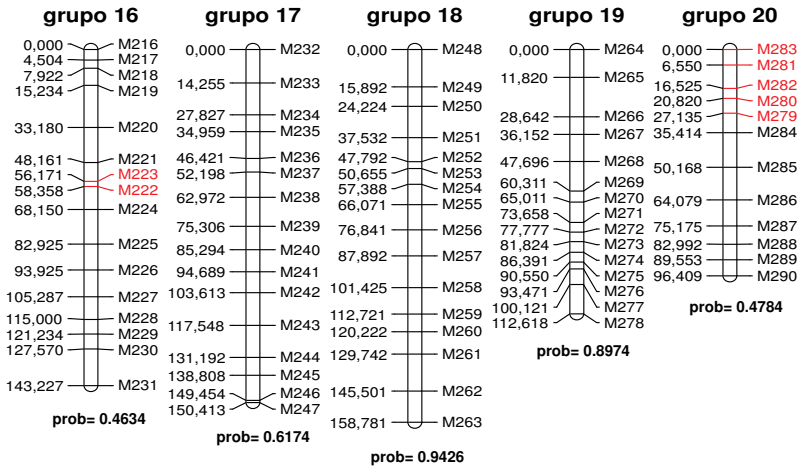
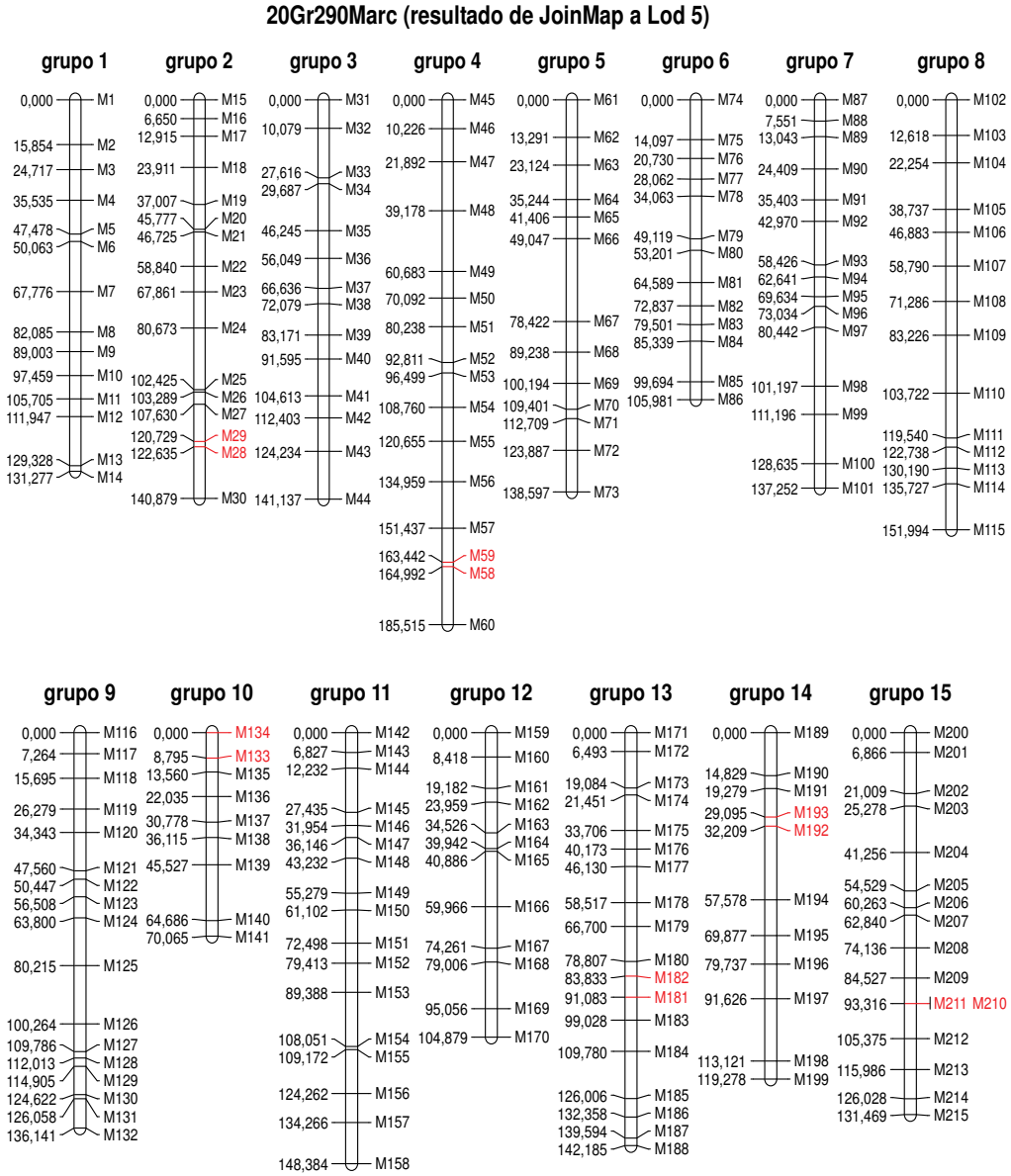


Figura 11.1: *Estimación bayesiana del mapa genético real de la población de la que procede la muestra 20Gr290Marc, con un 0 % de datos faltantes. Estructura obtenida con un nivel de significatividad global* $\alpha = 0.005$


20Gr290Marc (resultado de JoinMap a Lod 5)
grupo 1
0,000 M1
15,854 M2
24,717 M3
35,535 M4
47,478 M5
50,063 M6
67,776 M7
82,085 M8
89,003 M9
97,459 M10
105,705 M11
111,947 M12
129,328 M13
131,277 M14
grupo 2
0,000 M15
6,650 M16
12,915 M17
23,911 M18
37,007 M19
45,777 M20
46,725 M21
58,840 M22
67,861 M23
80,673 M24
102,425 M25
103,289 M26
107,630 M27
120,729 M29
122,635 M28
140,879 M30
grupo 3
0,000 M31
10,079 M32
27,616 M33
29,687 M34
46,245 M35
56,049 M36
66,636 M37
72,079 M38
83,171 M39
91,595 M40
104,613 M41
112,403 M42
124,234 M43
141,137 M44
grupo 4
0,000 M45
10,226 M46
21,892 M47
39,178 M48
60,683 M49
70,092 M50
80,238 M51
92,811 M52
96,499 M53
108,760 M54
120,655 M55
134,959 M56
151,437 M57
163,442 M59
164,992 M58
185,515 M60
grupo 5
0,000 M61
13,291 M62
23,124 M63
35,244 M64
41,406 M65
49,047 M66
78,422 M67
89,238 M68
100,194 M69
109,401 M70
112,709 M71
123,887 M72
138,597 M73
grupo 6
0,000 M74
14,097 M75
20,730 M76
28,062 M77
34,063 M78
49,119 M79
53,201 M80
64,589 M81
72,837 M82
79,501 M83
85,339 M84
99,694 M85
105,981 M86
grupo 7
0,000 M87
7,551 M88
13,043 M89
24,409 M90
35,403 M91
42,970 M92
58,426 M93
62,641 M94
69,634 M95
73,034 M96
80,442 M97
101,197 M98
111,196 M99
128,635 M100
137,252 M101
grupo 8
0,000 M102
12,618 M103
22,254 M104
38,737 M105
46,883 M106
58,790 M107
71,286 M108
83,226 M109
103,722 M110
119,540 M111
122,738 M112
130,190 M113
135,727 M114
151,994 M115
grupo 9
0,000 M116
7,264 M117
15,695 M118
26,279 M119
34,343 M120
47,560 M121
50,447 M122
56,508 M123
63,800 M124
80,215 M125
100,264 M126
109,786 M127
112,013 M128
114,905 M129
124,622 M130
126,058 M131
136,141 M132
grupo 10
0,000 M134
8,795 M133
13,560 M135
22,035 M136
30,778 M137
36,115 M138
45,527 M139
64,686 M140
70,065 M141
grupo 11
0,000 M142
6,827 M143
12,232 M144
27,435 M145
31,954 M146
36,146 M147
43,232 M148
55,279 M149
61,102 M150
72,498 M151
79,413 M152
89,388 M153
108,051 M154
109,172 M155
124,262 M156
134,266 M157
148,384 M158
grupo 12
0,000 M159
8,418 M160
19,182 M161
23,959 M162
34,526 M163
39,942 M164
40,886 M165
59,966 M166
74,261 M167
79,006 M168
95,056 M169
104,879 M170
grupo 13
0,000 M171
6,493 M172
19,084 M173
21,451 M174
33,706 M175
40,173 M176
46,130 M177
58,517 M178
66,700 M179
78,807 M180
83,833 M182
91,083 M181
99,028 M183
109,780 M184
126,006 M185
132,358 M186
139,594 M187
142,185 M188
grupo 14
0,000 M189
14,829 M190
19,279 M191
29,095 M193
32,209 M192
57,578 M194
69,877 M195
79,737 M196
91,626 M197
113,121 M198
119,278 M199
grupo 15
0,000 M200
6,866 M201
21,009 M202
25,278 M203
41,256 M204
54,529 M205
60,263 M206
62,840 M207
74,136 M208
84,527 M209
93,316 M211 M210
105,375 M212
115,986 M213
126,028 M214
131,469 M215

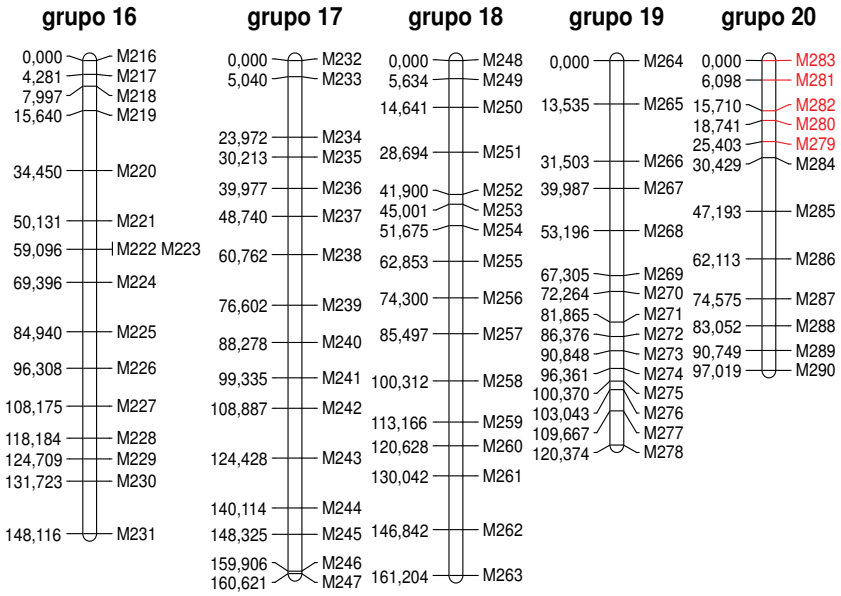
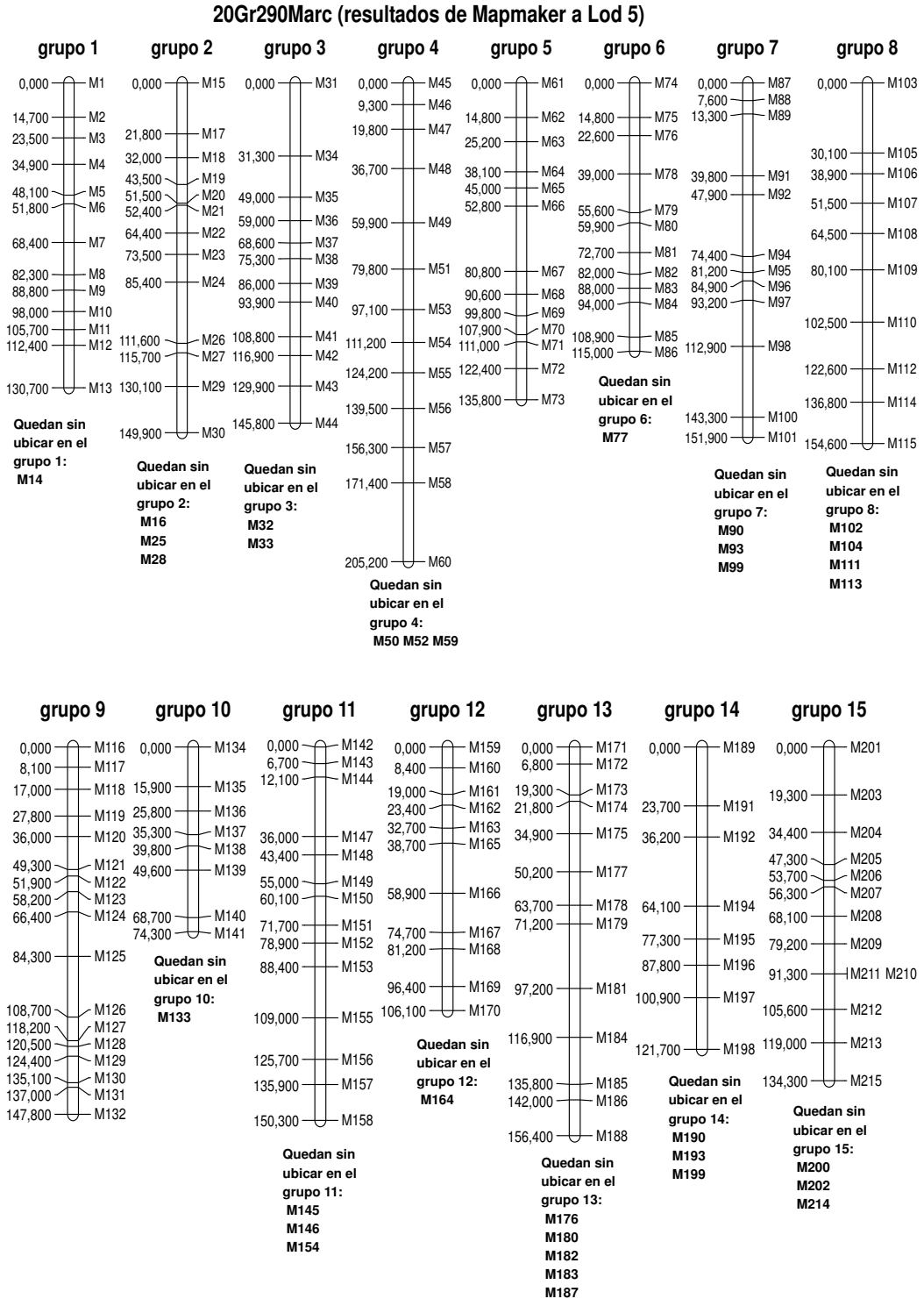


Figura 11.2: *Estimación, según el programa JoinMap, del mapa genético real de la población de la que procede la muestra 20Gr290Marc, con un 0 % de datos faltantes. Estructura obtenida a LOD 5*


20Gr290Marc (resultados de Mapmaker a Lod 5)
grupo 1
0,000 M1
14,700 M2
23,500 M3
34,900 M4
48,100 M5
51,800 M6
68,400 M7
82,300 M8
88,800 M9
98,000 M10
105,700 M11
112,400 M12
130,700 M13
Quedan sin ubicar en el grupo 1: M14
grupo 2
0,000 M15
21,800 M17
32,000 M18
43,500 M19
51,500 M20
52,400 M21
64,400 M22
73,500 M23
85,400 M24
111,600 M26
115,700 M27
130,100 M29
149,900 M30
Quedan sin ubicar en el grupo 2: M16 M25 M28
grupo 3
0,000 M31
31,300 M34
49,000 M35
59,000 M36
68,600 M37
75,300 M38
86,000 M39
93,900 M40
108,800 M41
116,900 M42
129,900 M43
145,800 M44
Quedan sin ubicar en el grupo 3: M32 M33
grupo 4
0,000 M45
9,300 M46
19,800 M47
36,700 M48
59,900 M49
79,800 M51
97,100 M53
111,200 M54
124,200 M55
139,500 M56
156,300 M57
171,400 M58
205,200 M60
Quedan sin ubicar en el grupo 4: M50 M52 M59
grupo 5
0,000 M61
14,800 M62
25,200 M63
38,100 M64
45,000 M65
52,800 M66
80,800 M67
90,600 M68
99,800 M69
107,900 M70
111,000 M71
122,400 M72
135,800 M73
grupo 6
0,000 M74
14,800 M75
22,600 M76
39,000 M78
55,600 M79
59,900 M80
72,700 M81
82,000 M82
88,000 M83
94,000 M84
108,900 M85
115,000 M86
Quedan sin ubicar en el grupo 6: M77
grupo 7
0,000 M87
7,600 M88
13,300 M89
39,800 M91
47,900 M92
74,400 M94
81,200 M95
84,900 M96
93,200 M97
112,900 M98
143,300 M100
151,900 M101
Quedan sin ubicar en el grupo 7: M90 M93 M99
grupo 8
0,000 M103
30,100 M105
38,900 M106
51,500 M107
64,500 M108
80,100 M109
102,500 M110
122,600 M112
136,800 M114
154,600 M115
Quedan sin ubicar en el grupo 8: M102 M104 M111 M113
grupo 9
0,000 M116
8,100 M117
17,000 M118
27,800 M119
36,000 M120
49,300 M121
51,900 M122
58,200 M123
66,400 M124
84,300 M125
108,700 M126
118,200 M127
120,500 M128
124,400 M129
135,100 M130
137,000 M131
147,800 M132
grupo 10
0,000 M134
15,900 M135
25,800 M136
35,300 M137
39,800 M138
49,600 M139
68,700 M140
74,300 M141
Quedan sin ubicar en el grupo 10: M133
grupo 11
0,000 M142
6,700 M143
12,100 M144
36,000 M147
43,400 M148
55,000 M149
60,100 M150
71,700 M151
78,900 M152
88,400 M153
109,000 M155
125,700 M156
135,900 M157
150,300 M158
Quedan sin ubicar en el grupo 11: M145 M146 M154
grupo 12
0,000 M159
8,400 M160
19,000 M161
23,400 M162
32,700 M163
38,700 M165
58,900 M166
74,700 M167
81,200 M168
96,400 M169
106,100 M170
Quedan sin ubicar en el grupo 12: M164
grupo 13
0,000 M171
6,800 M172
19,300 M173
21,800 M174
34,900 M175
50,200 M177
63,700 M178
71,200 M179
97,200 M181
116,900 M184
135,800 M185
142,000 M186
156,400 M188
Quedan sin ubicar en el grupo 13: M176 M180 M182 M183 M187
grupo 14
0,000 M189
23,700 M191
36,200 M192
64,100 M194
77,300 M195
87,800 M196
100,900 M197
121,700 M198
Quedan sin ubicar en el grupo 14: M190 M193 M199
grupo 15
0,000 M201
19,300 M203
34,400 M204
47,300 M205
53,700 M206
56,300 M207
68,100 M208
79,200 M209
91,300 M211 M210
105,600 M212
119,000 M213
134,300 M215
Quedan sin ubicar en el grupo 15: M200 M202 M214

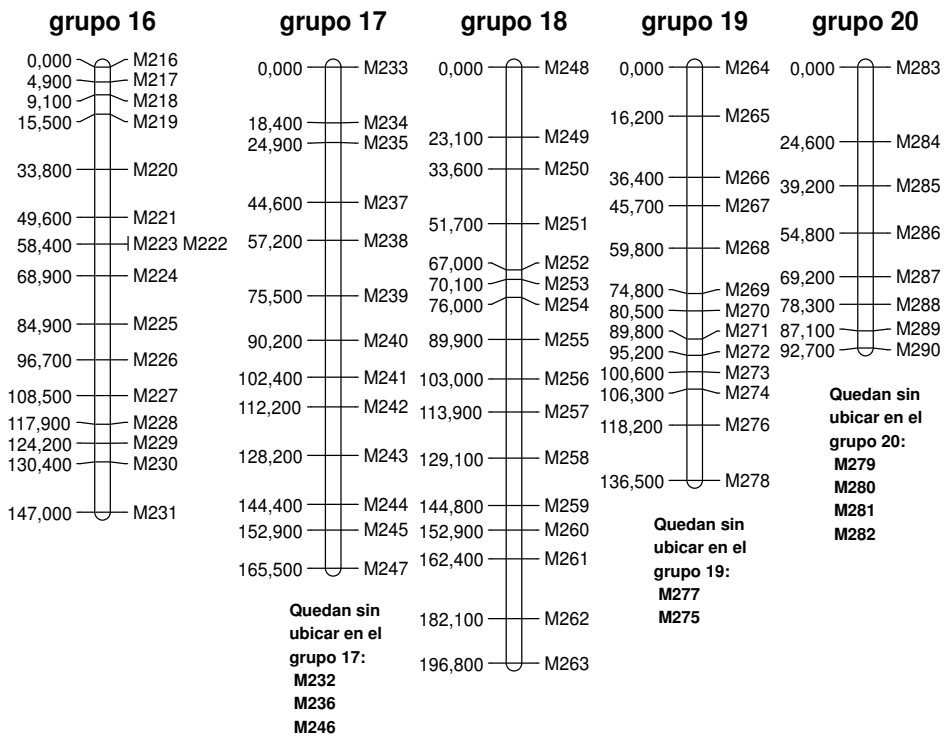


Figura 11.3: *Estimación, según el programa Mapmaker, del mapa genético real de la población de la que procede la muestra 20Gr290Marc, con un 0 % de datos faltantes. Estructura obtenida a LOD 5*

11.2. Resultados para la muestra con un 15 % de datos faltantes

En este apartado se elimina aleatoriamente el 15 % de los datos de la muestra 20Gr290Marc. Esta manipulación produce diferentes porcentajes de individuos faltantes para cada uno de los 290 marcadores. En el siguiente cuadro aparece la descriptiva de individuos faltantes respecto a los 290 marcadores.

Mínimo	Q_1	Mediana	Media	Q_3	Máximo
7 %	13 %	15 %	15 %	16.88 %	22 %

Cuadro 11.1: Porcentajes de individuos faltantes respecto a los 290 marcadores para la muestra con un 15 % de datos faltantes

Por lo tanto, se deduce que todos y cada uno de los marcadores han perdido en mayor o menor medida parte de los individuos. El marcador más desafortunado, ha perdido el 22 % de los individuos. Se puede comprobar que este dato corresponde al marcador 157. El 75 % de los marcadores han perdido un porcentaje inferior al 16.88 % de los individuos.

Además, como ya es sabido, cada pareja de marcadores determina el reparto de los individuos en base a sus genotipos observables, de manera que, además, existe una pérdida de información adicional de individuos que se pierden exclusivamente por uno de los dos marcadores involucrados. Así, en este caso, la metodología se aplica sobre 41905 parejas de marcadores que han experimentado los siguientes porcentajes de individuos perdidos.

Mínimo	Q_1	Mediana	Media	Q_3	Máximo
14.5 %	25.5 %	28 %	27.76 %	30 %	41 %

Cuadro 11.2: Porcentajes de individuos faltantes respecto a las 41905 parejas de marcadores para la muestra con un 15 % de datos faltantes

Por ejemplo, se puede comprobar, que la pareja de marcadores (116, 157) es la que mayor porcentaje de individuos ha perdido. Concretamente el 41 % de los individuos.

Procediendo del mismo modo que en el apartado anterior, como se puede ver en las siguientes tablas, la metodología bayesiana obtiene la agrupación correcta de los marcadores (que no el orden) utilizando niveles de significatividad global $\alpha = 0.05$ y $\alpha = 0.01$. JoinMap [85] y Mapmaker [48] obtiene este mismo resultado bajo LOD 5. Recuérdese que la notación empleada para nombrar los grupos de marcadores que son unión de distintos grupos de ligamiento reales, es la misma que en el capítulo anterior. Además, si un grupo de ligamiento real se estima desglosado en varias agrupaciones diferentes, éstas se nombran con el índice del grupo real seguido de una letra. Por ejemplo, véase los grupos "grupo 2a" y "grupo 2b" de la estructura de marcadores obtenida por la metodología bayesiana bajo un nivel de significatividad $\alpha = 0.005$. La estimación del mapa genético en los escenarios que obtienen la agrupación correcta, aparecen representados en las Figuras 11.4, 11.5 y 11.6

Como cabía esperar, Mapmaker [48] empeora su estimación respecto al resultado obtenido con la muestra completa, hasta el punto que no es capaz de concretar un orden específico para los marcadores pertenecientes a los grupos de ligamiento 3, 4 y 8. Igual que antes, sólo se comenta la comparativa de la metodología bayesiana respecto al resultado obtenido por JoinMap [85].

Las estimaciones del mapa genético que obtienen tanto la metodología bayesiana como el programa JoinMap [85] también se ven afectadas por la pérdida de información de la muestra pero en ambos casos todos los marcadores quedan ubicados en algún grupo de ligamiento.

Se puede ver que, con la metodología bayesiana, aquellos marcadores que con la muestra completa quedaban mal ubicados, continúan estándolo. Excepto las parejas (52, 53) y (222, 223) que obtiene el orden correcto. Adicionalmente, obtienen problemas en el orden las parejas de marcadores (20, 21), (102, 103), (145, 146), (164, 165), (167, 168), (198, 199), (234, 235). Se puede comprobar, que los marcadores contiguos (102, 103), (145, 146), (198, 199) y (234, 235) han perdido entre el 29 % y el 33 % de los individuos. Es decir, por encima de la mediana del total de las parejas. En general, las probabilidades de ordenación de los marcadores, en casi todos los grupos de ligamiento bayesianos, se ven reducidas excepto en el grupo 1 y el grupo 20, que mejoran un poco. Aunque las probabilidades de ordenación se ven reducidas, las ordenaciones en sí no varían en 14 de los grupos de ligamiento o incluso mejoran, como acabamos de decir, en el caso del grupo 4 y grupo 16. En los grupos de ligamiento 1, 5, 7, 8, 10, 15, 16 y 17 la metodología bayesiana obtiene una

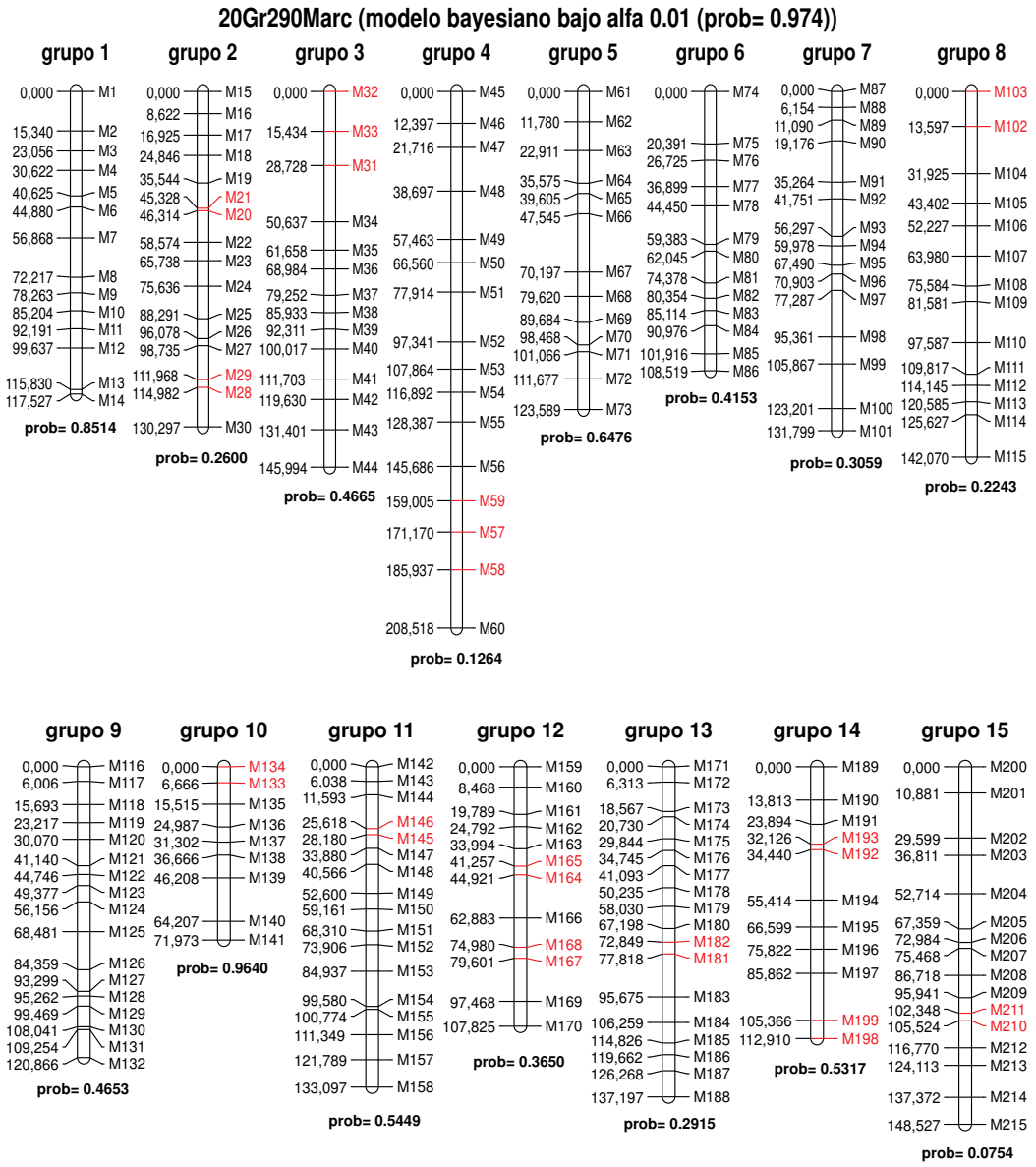
estimación de la longitud del mapa genético más próxima a la real que el programa JoinMap [85]. En los grupos 9 y 16, JoinMap [85] estima una permutación de marcadores contiguos, sin embargo los mismos grupos bayesianos coinciden con los correctos. A la inversa ocurre para el grupo 3 y para el grupo 11. JoinMap mantiene el orden erróneo de los marcadores de la muestra completa y a estos añade los marcadores (111,112,113,114), (116, 117), (164, 165), (222, 223) y (234,235). Curiosamente, todas estas parejas han perdido el 29 % de los individuos o menos.

20Gr290Marc 15 % faltantes	modelo bayesiano α=0.05	α=0.01	α=0.005
Agrupación 1	20 grupos (prob= 1) grupo 1 grupo 2 grupo 3 grupo 4 grupo 5 grupo 6 grupo 7 grupo 8 grupo 9 grupo 10 grupo 11 grupo 12 grupo 13 grupo 14 grupo 15 grupo 16 grupo 17 grupo 18 grupo 19 grupo 20	**20 grupos (prob= 0.974)** grupo 1 grupo 2 grupo 3 grupo 4 grupo 5 grupo 6 grupo 7 grupo 8 grupo 9 grupo 10 grupo 11 grupo 12 grupo 13 grupo 14 grupo 15 grupo 16 grupo 17 grupo 18 grupo 19 grupo 20	21 grupos (prob= 0.9665) marcadores 248 sin agrupar grupo 1 grupo 2a (*) grupo 2b (**) grupo 3 grupo 4 grupo 5 grupo 6 grupo 7 grupo 8 grupo 9 grupo 10 grupo 11 grupo 12 grupo 13 grupo 14 grupo 15 grupo 16 grupo 17 grupo 18 grupo 19 grupo 20
Agrupación 2		21 grupos con prob= 0.023	22 grupos + 1 marc. sin agrupar con prob= 0.0225
Agrupación 3		20 grupos + 1 marc. sin agrupar con prob= 0.0015	22 grupos + 1 marc. sin agrupar con prob= 0.0075
Agrupación 4		21 grupos con prob= 5e-04	21 grupos + 2 marc. sin agrupar con prob= 0.0015
Agrupación 5		20 grupos + 1 marc. sin agrupar con prob= 5e-04	23 grupos + 1 marc. sin agrupar con prob= 5e-04
Agrupación 6		21 grupos con prob= 5e-04	21 grupos + 2 marc. sin agrupar con prob= 5e-04
Agrupación 7			22 grupos + 1 marc. sin agrupar con prob= 5e-04
Agrupación 8			22 grupos + 1 marc. sin agrupar con prob= 5e-04

Nota:
(*) grupo 2a={15 16}
(**) grupo 2b={17 18 19 20 21 22 23 24 25 26 27 28 29 30}
Nótese que grupo 2a $\cup$ grupo 2b= grupo 2

20Gr290Marc 15 % faltantes	JoinMap LOD 3	LOD 4	LOD 5
Agrupación única	7 grupos	19 grupos	**20 grupos**
	grupo 1	grupo 1	grupo 1
	grupo 2.3. 4.5.7.8.9.11.13.15.17.19.20	grupo 2	grupo 2
	grupo 6	grupo 3.20	grupo 3
	grupo 10.14	grupo 4	grupo 4
	grupo 12	grupo 5	grupo 5
	grupo 16	grupo 6	grupo 6
	grupo 18	grupo 7	grupo 7
		grupo 8	grupo 8
		grupo 9	grupo 9
		grupo 10	grupo 10
		grupo 11	grupo 11
		grupo 12	grupo 12
		grupo 13	grupo 13
		grupo 14	grupo 14
		grupo 15	grupo 15
		grupo 16	grupo 16
		grupo 17	grupo 17
		grupo 18	grupo 18
		grupo 19	grupo 19
			grupo 20

20Gr290Marc 15 % faltantes	MapMaker LOD 3	LOD 4	LOD 5
Agrupación única	15 grupos	19 grupos	**20 grupos**
	grupo 1	grupo 1	grupo 1
	grupo 2	grupo 2	grupo 2
	grupo 3.4 .13.15.20	grupo 3.20	grupo 3
	grupo 4	grupo 4	grupo 4
	grupo 5	grupo 5	grupo 5
	grupo 6	grupo 6	grupo 6
	grupo 7.17	grupo 7	grupo 7
	grupo 8	grupo 8	grupo 8
	grupo 9	grupo 9	grupo 9
	grupo 10	grupo 10	grupo 10
	grupo 11	grupo 11	grupo 11
	grupo 12	grupo 12	grupo 12
	grupo 14	grupo 13	grupo 13
	grupo 16	grupo 14	grupo 14
	grupo 18	grupo 15	grupo 15
	grupo 19	grupo 16	grupo 16
		grupo 17	grupo 17
		grupo 18	grupo 18
		grupo 19	grupo 19
			grupo 20


20Gr290Marc (modelo bayesiano bajo alfa 0.01 (prob= 0.974))
grupo 1
0,000 M1
15,340 M2
23,056 M3
30,622 M4
40,625 M5
44,880 M6
56,868 M7
72,217 M8
78,263 M9
85,204 M10
92,191 M11
99,637 M12
115,830 M13
117,527 M14
prob= 0.8514
grupo 2
0,000 M15
8,622 M16
16,925 M17
24,846 M18
35,544 M19
45,328 M21
46,314 M20
58,574 M22
65,738 M23
75,636 M24
88,291 M25
96,078 M26
98,735 M27
111,968 M29
114,982 M28
130,297 M30
prob= 0.2600
grupo 3
0,000 M32
15,434 M33
28,728 M31
50,637 M34
61,658 M35
68,984 M36
79,252 M37
85,933 M38
92,311 M39
100,017 M40
111,703 M41
119,630 M42
131,401 M43
145,994 M44
prob= 0.4665
grupo 4
0,000 M45
12,397 M46
21,716 M47
38,697 M48
57,463 M49
66,560 M50
77,914 M51
97,341 M52
107,864 M53
116,892 M54
128,387 M55
145,686 M56
159,005 M59
171,170 M57
185,937 M58
208,518 M60
prob= 0.1264
grupo 5
0,000 M61
11,780 M62
22,911 M63
35,575 M64
39,605 M65
47,545 M66
70,197 M67
79,620 M68
89,684 M69
98,468 M70
101,066 M71
111,677 M72
123,589 M73
prob= 0.6476
grupo 6
0,000 M74
20,391 M75
26,725 M76
36,899 M77
44,450 M78
59,383 M79
62,045 M80
74,378 M81
80,354 M82
85,114 M83
90,976 M84
101,916 M85
108,519 M86
prob= 0.4153
grupo 7
0,000 M87
6,154 M88
11,090 M89
19,176 M90
35,264 M91
41,751 M92
56,297 M93
59,978 M94
67,490 M95
70,903 M96
77,287 M97
95,361 M98
105,867 M99
123,201 M100
131,799 M101
prob= 0.3059
grupo 8
0,000 M103
13,597 M102
31,925 M104
43,402 M105
52,227 M106
63,980 M107
75,584 M108
81,581 M109
97,587 M110
109,817 M111
114,145 M112
120,585 M113
125,627 M114
142,070 M115
prob= 0.2243
grupo 9
0,000 M116
6,006 M117
15,693 M118
23,217 M119
30,070 M120
41,140 M121
44,746 M122
49,377 M123
56,156 M124
68,481 M125
84,359 M126
93,299 M127
95,262 M128
99,469 M129
108,041 M130
109,254 M131
120,866 M132
prob= 0.4653
grupo 10
0,000 M134
6,666 M133
15,515 M135
24,987 M136
31,302 M137
36,666 M138
46,208 M139
64,207 M140
71,973 M141
prob= 0.9640
grupo 11
0,000 M142
6,038 M143
11,593 M144
25,618 M146
28,180 M145
33,880 M147
40,566 M148
52,600 M149
59,161 M150
68,310 M151
73,906 M152
84,937 M153
99,580 M154
100,774 M155
111,349 M156
121,789 M157
133,097 M158
prob= 0.5449
grupo 12
0,000 M159
8,468 M160
19,789 M161
24,792 M162
33,994 M163
41,257 M165
44,921 M164
62,883 M166
74,980 M168
79,601 M167
97,468 M169
107,825 M170
prob= 0.3650
grupo 13
0,000 M171
6,313 M172
18,567 M173
20,730 M174
29,844 M175
34,745 M176
41,093 M177
50,235 M178
58,030 M179
67,198 M180
72,849 M182
77,818 M181
95,675 M183
106,259 M184
114,826 M185
119,662 M186
126,268 M187
137,197 M188
prob= 0.2915
grupo 14
0,000 M189
13,813 M190
23,894 M191
32,126 M193
34,440 M192
55,414 M194
66,599 M195
75,822 M196
85,862 M197
105,366 M199
112,910 M198
prob= 0.5317
grupo 15
0,000 M200
10,881 M201
29,599 M202
36,811 M203
52,714 M204
67,359 M205
72,984 M206
75,468 M207
86,718 M208
95,941 M209
102,348 M211
105,524 M210
116,770 M212
124,113 M213
137,372 M214
148,527 M215
prob= 0.0754

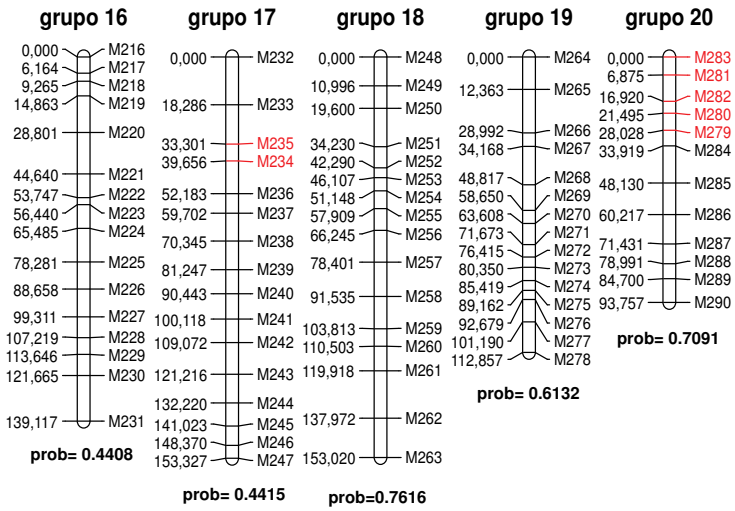
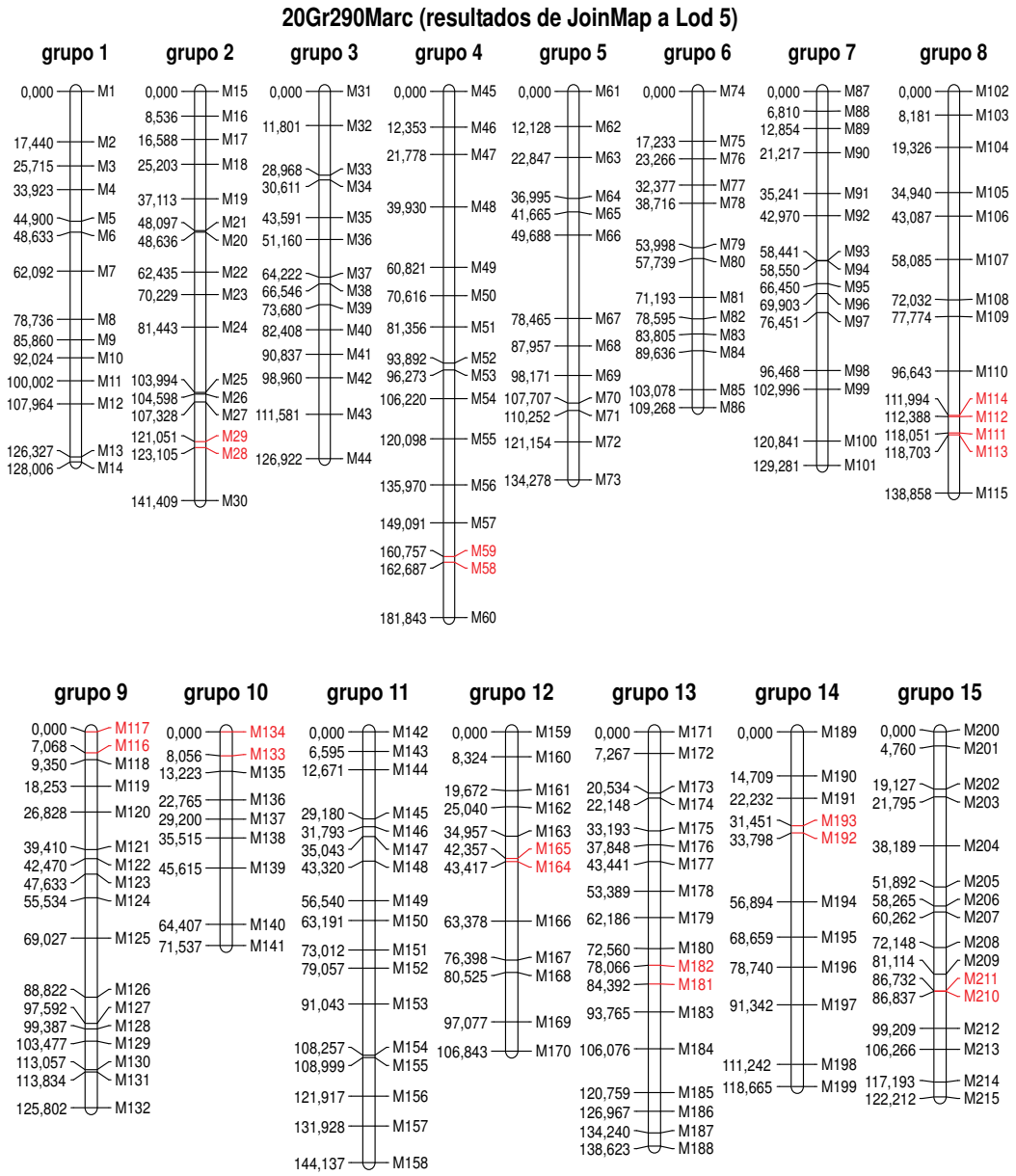


Figura 11.4: *Estimación bayesiana del mapa genético real de la población de la que procede la muestra 20Gr290Marc, con un 15 % de datos faltantes. Estructura obtenida con un nivel de significatividad global* $\alpha = 0.01$.


20Gr290Marc (resultados de JoinMap a Lod 5)
grupo 1
0,000 M1
17,440 M2
25,715 M3
33,923 M4
44,900 M5
48,633 M6
62,092 M7
78,736 M8
85,860 M9
92,024 M10
100,002 M11
107,964 M12
126,327 M13
128,006 M14
grupo 2
0,000 M15
8,536 M16
16,588 M17
25,203 M18
37,113 M19
48,097 M21
48,636 M20
62,435 M22
70,229 M23
81,443 M24
103,994 M25
104,598 M26
107,328 M27
121,051 M29
123,105 M28
141,409 M30
grupo 3
0,000 M31
11,801 M32
28,968 M33
30,611 M34
43,591 M35
51,160 M36
64,222 M37
66,546 M38
73,680 M39
82,408 M40
90,837 M41
98,960 M42
111,581 M43
126,922 M44
grupo 4
0,000 M45
12,353 M46
21,778 M47
39,930 M48
60,821 M49
70,616 M50
81,356 M51
93,892 M52
96,273 M53
106,220 M54
120,098 M55
135,970 M56
149,091 M57
160,757 M59
162,687 M58
181,843 M60
grupo 5
0,000 M61
12,128 M62
22,847 M63
36,995 M64
41,665 M65
49,688 M66
78,465 M67
87,957 M68
98,171 M69
107,707 M70
110,252 M71
121,154 M72
134,278 M73
grupo 6
0,000 M74
17,233 M75
23,266 M76
32,377 M77
38,716 M78
53,998 M79
57,739 M80
71,193 M81
78,595 M82
83,805 M83
89,636 M84
103,078 M85
109,268 M86
grupo 7
0,000 M87
6,810 M88
12,854 M89
21,217 M90
35,241 M91
42,970 M92
58,441 M93
58,550 M94
66,450 M95
69,903 M96
76,451 M97
96,468 M98
102,996 M99
120,841 M100
129,281 M101
grupo 8
0,000 M102
8,181 M103
19,326 M104
34,940 M105
43,087 M106
58,085 M107
72,032 M108
77,774 M109
96,643 M110
111,994 M114
112,388 M112
118,051 M111
118,703 M113
138,858 M115
grupo 9
0,000 M117
7,068 M116
9,350 M118
18,253 M119
26,828 M120
39,410 M121
42,470 M122
47,633 M123
55,534 M124
69,027 M125
88,822 M126
97,592 M127
99,387 M128
103,477 M129
113,057 M130
113,834 M131
125,802 M132
grupo 10
0,000 M134
8,056 M133
13,223 M135
22,765 M136
29,200 M137
35,515 M138
45,615 M139
64,407 M140
71,537 M141
grupo 11
0,000 M142
6,595 M143
12,671 M144
29,180 M145
31,793 M146
35,043 M147
43,320 M148
56,540 M149
63,191 M150
73,012 M151
79,057 M152
91,043 M153
108,257 M154
108,999 M155
121,917 M156
131,928 M157
144,137 M158
grupo 12
0,000 M159
8,324 M160
19,672 M161
25,040 M162
34,957 M163
42,357 M165
43,417 M164
63,378 M166
76,398 M167
80,525 M168
97,077 M169
106,843 M170
grupo 13
0,000 M171
7,267 M172
20,534 M173
22,148 M174
33,193 M175
37,848 M176
43,441 M177
53,389 M178
62,186 M179
72,560 M180
78,066 M182
84,392 M181
93,765 M183
106,076 M184
120,759 M185
126,967 M186
134,240 M187
138,623 M188
grupo 14
0,000 M189
14,709 M190
22,232 M191
31,451 M193
33,798 M192
56,894 M194
68,659 M195
78,740 M196
91,342 M197
111,242 M198
118,665 M199
grupo 15
0,000 M200
4,760 M201
19,127 M202
21,795 M203
38,189 M204
51,892 M205
58,265 M206
60,262 M207
72,148 M208
81,114 M209
86,732 M211
86,837 M210
99,209 M212
106,266 M213
117,193 M214
122,212 M215

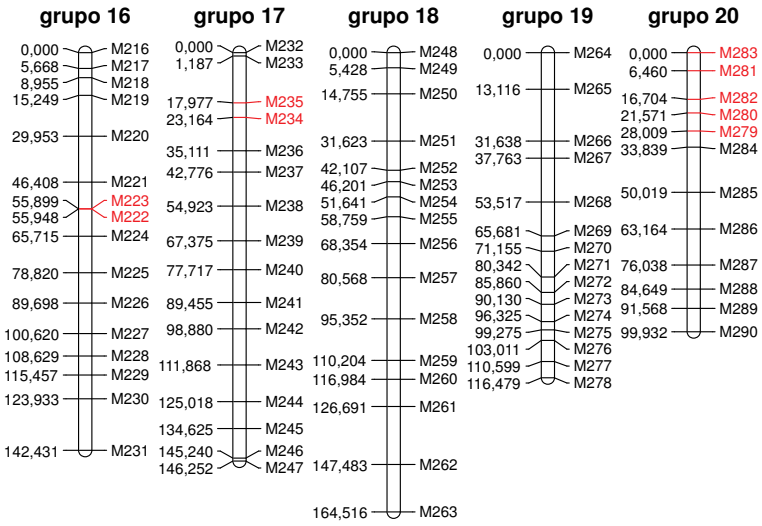
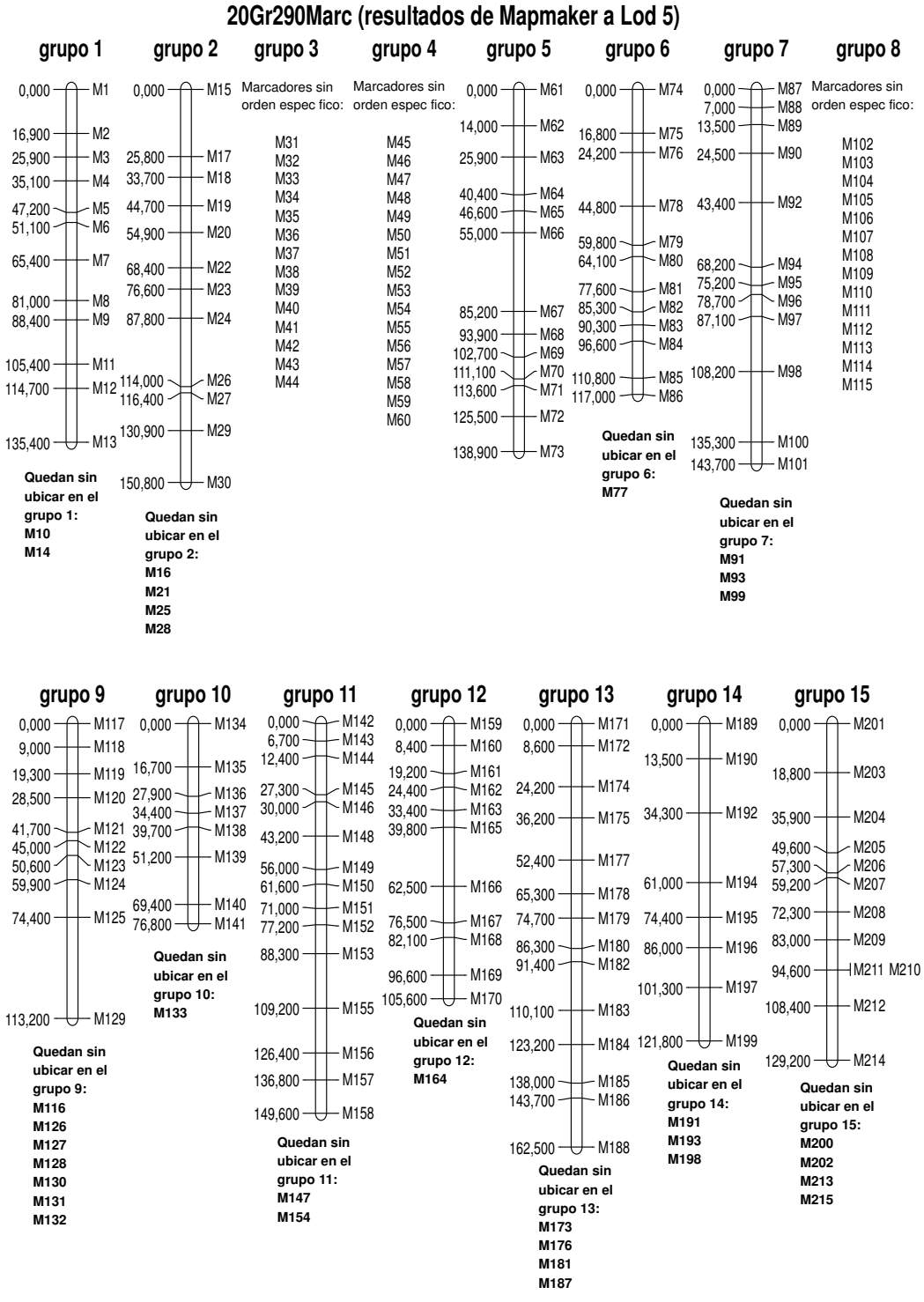


Figura 11.5: *Estimación, según el programa JoinMap, del mapa genético real de la población de la que procede la muestra 20Gr290Marc, con un 15 % de datos faltantes. Estructura obtenida a LOD 5.*


20Gr290Marc (resultados de Mapmaker a Lod 5)
grupo 1
0,000 M1
16,900 M2
25,900 M3
35,100 M4
47,200 M5
51,100 M6
65,400 M7
81,000 M8
88,400 M9
105,400 M11
114,700 M12
135,400 M13
Quedan sin ubicar en el grupo 1: M10 M14
grupo 2
0,000 M15
25,800 M17
33,700 M18
44,700 M19
54,900 M20
68,400 M22
76,600 M23
87,800 M24
114,000 M26
116,400 M27
130,900 M29
150,800 M30
Quedan sin ubicar en el grupo 2: M16 M21 M25 M28
grupo 3
Marcadores sin orden espec fico:
M31 M32 M33 M34 M35 M36 M37 M38 M39 M40 M41 M42 M43 M44
grupo 4
Marcadores sin orden espec fico:
M45 M46 M47 M48 M49 M50 M51 M52 M53 M54 M55 M56 M57 M58 M59 M60
grupo 5
0,000 M61
14,000 M62
25,900 M63
40,400 M64
46,600 M65
55,000 M66
85,200 M67
93,900 M68
102,700 M69
111,100 M70
113,600 M71
125,500 M72
138,900 M73
grupo 6
0,000 M74
16,800 M75
24,200 M76
44,800 M78
59,800 M79
64,100 M80
77,600 M81
85,300 M82
90,300 M83
96,600 M84
110,800 M85
117,000 M86
Quedan sin ubicar en el grupo 6: M77
grupo 7
0,000 M87
7,000 M88
13,500 M89
24,500 M90
43,400 M92
68,200 M94
75,200 M95
78,700 M96
87,100 M97
108,200 M98
135,300 M100
143,700 M101
Quedan sin ubicar en el grupo 7: M91 M93 M99
grupo 8
Marcadores sin orden espec fico:
M102 M103 M104 M105 M106 M107 M108 M109 M110 M111 M112 M113 M114 M115
grupo 9
0,000 M117
9,000 M118
19,300 M119
28,500 M120
41,700 M121
45,000 M122
50,600 M123
59,900 M124
74,400 M125
113,200 M129
Quedan sin ubicar en el grupo 9: M116 M126 M127 M128 M130 M131 M132
grupo 10
0,000 M134
16,700 M135
27,900 M136
34,400 M137
39,700 M138
51,200 M139
69,400 M140
76,800 M141
Quedan sin ubicar en el grupo 10: M133
grupo 11
0,000 M142
6,700 M143
12,400 M144
27,300 M145
30,000 M146
43,200 M148
56,000 M149
61,600 M150
71,000 M151
77,200 M152
88,300 M153
109,200 M155
126,400 M156
136,800 M157
149,600 M158
Quedan sin ubicar en el grupo 11: M147 M154
grupo 12
0,000 M159
8,400 M160
19,200 M161
24,400 M162
33,400 M163
39,800 M165
62,500 M166
76,500 M167
82,100 M168
96,600 M169
105,600 M170
Quedan sin ubicar en el grupo 12: M164
grupo 13
0,000 M171
8,600 M172
24,200 M174
36,200 M175
52,400 M177
65,300 M178
74,700 M179
86,300 M180
91,400 M182
110,100 M183
123,200 M184
138,000 M185
143,700 M186
162,500 M188
Quedan sin ubicar en el grupo 13: M173 M176 M181 M187
grupo 14
0,000 M189
13,500 M190
34,300 M192
61,000 M194
74,400 M195
86,000 M196
101,300 M197
121,800 M199
Quedan sin ubicar en el grupo 14: M191 M193 M198
grupo 15
0,000 M201
18,800 M203
35,900 M204
49,600 M205
57,300 M206
59,200 M207
72,300 M208
83,000 M209
94,600 M211 M210
108,400 M212
129,200 M214
Quedan sin ubicar en el grupo 15: M200 M202 M213 M215

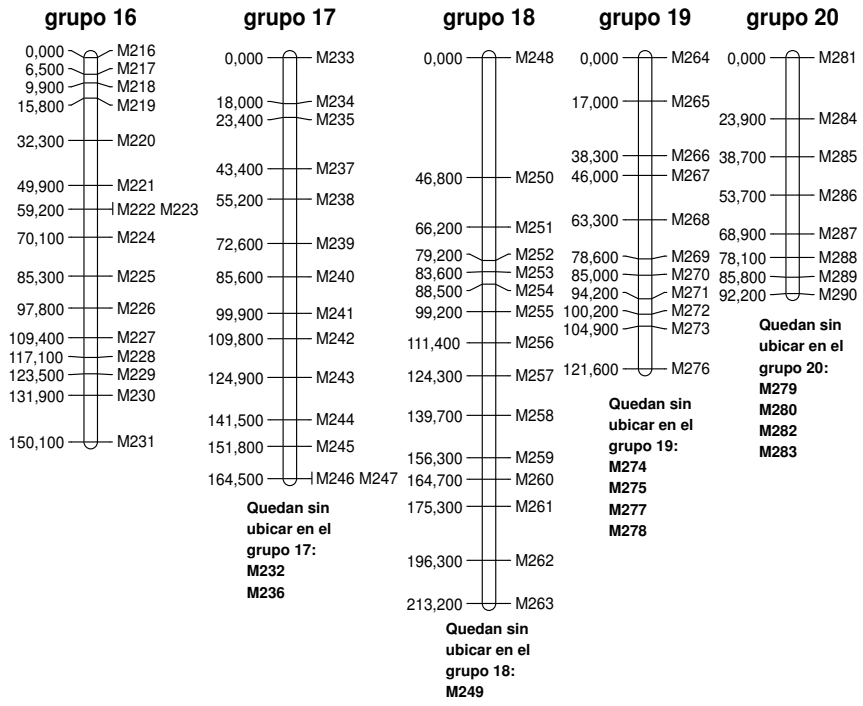


Figura 11.6: *Estimación, según el programa Mapmaker, del mapa genético real de la población de la que procede la muestra 20Gr290Marc, con un 15 % de datos faltantes. Estructura obtenida a LOD 5.*

11.3. Resultados para la muestra con un 25 % de datos faltantes

Para finalizar el capítulo, se elimina el 25 % de los datos de la muestra completa, 20Gr290Marc, y se repite el mismo proceso que en los anteriores apartados. A continuación se especifica el resumen del porcentaje de pérdida de individuos tanto respecto a los 290 marcadores como con respecto a las 41905 parejas de marcadores.

Mínimo	Q_1	Mediana	Media	Q_3	Máximo
17 %	23 %	28 %	25 %	25 %	32 %

Cuadro 11.3: Porcentajes de individuos faltantes respecto a los 290 marcadores para la muestra con un 25 % de datos faltantes

Mínimo	Q_1	Mediana	Media	Q_3	Máximo
29 %	41.5 %	43.5 %	43.75 %	46 %	57.5 %

Cuadro 11.4: Porcentajes de individuos faltantes respecto a las 41905 parejas de marcadores para la muestra con un 25 % de datos faltantes

Los resultados obtenidos por las tres metodologías se presentan de forma análoga al apartado anterior. En las tablas donde se resumen las distintas agrupaciones de los marcadores, aparecen nuevos grupos de ligamiento (que no órdenes) que se detallan a continuación. Nótese que, por ejemplo, el "grupo 3a1" y el "grupo 3a2" representa el desglose del "grupo 3a". Éste último, a su vez, contiene parte de los marcadores del "grupo 3" real, que fueron especificados en el capítulo anterior:

grupo 2a= {15 16}
grupo 2b= {17 18 19 20 21 22 23 24 25 26 27 28 29 30}
grupo 2b1= {17 18 19 20 21 22 23 24}
grupo 2b2= {25 26 27 28 29 30}
grupo 3a= {31 34 35 36 37 38 39 40 41 42 43 44}

grupo 3a1= {31 34 35 36 37 38 39 40}
grupo 3a2= {41 42 43 44}
grupo 3b= {32 33}
grupo 4a= {45 46 47 48 49 50 52}
grupo 4a1= {45 46 47 48 49 50}
grupo 4b= {51 53 54 55 56 57 58 59 60}
grupo 4b1= {51 53 54 55 56 57 58 60}
grupo 5a= {61 62 63}
grupo 5b= {64 65 66 67 68 69 70 71 72 73}
grupo 8a= {102 104 105 106 107 108 109 110 111 112 113 114 115}
grupo 17a= {232 233 234 235 236 237 238}
grupo 17b= {239 240 241 242 243 244 245 246 247}
grupo 18a= {249 250 251 252 253 254 255 256 257 258 259 260 261 262 263}

En este caso, la metodología bayesiana tiene problemas en la determinación correcta de los grupos de ligamiento. Como se observa en la siguiente tabla, el resultado más aproximado al real se obtiene exigiendo un nivel de significatividad global de $\alpha = 0.01$, en la resolución de los múltiples contrastes de independencia. Bajo esta condición se obtienen 25 grupos de ligamiento, 13 de los cuales coinciden con los grupos de ligamiento reales: 1, 6, 7, 9, 10, 11, 12, 13, 14, 15, 16, 19, 20. Los 12 restantes corresponden a fracciones de los otros 7 grupos de ligamiento reales. Quedan sin agrupar los marcadores 52, 59, 103 y 248. Respecto al orden, se mantienen algunos errores que se detectaron con la muestra completa (133, 134), (181,182), (192, 193) y (210,211), se "arreglan" algunas permutaciones de marcadores contiguos, relacionadas con los grupos fraccionados, y aparecen algunas permutaciones nuevas como son (49, 50), (88, 89), (116,117, 118), (246, 247), que han perdido entre el 42 % y el 48 % de los individuos. Como cabía esperar, una reducción tan sustancial de la información de la muestra tiene como consecuencia una reducción en las probabilidades obtenidas sobre las ordenaciones de los marcadores. Sin embargo, especialmente en los grupos de ligamiento que coinciden con los grupos reales, la ordenación estimada de los marcadores prácticamente sigue siendo la misma que cuando la muestra contenía toda la información.

El programa JoinMap [85] tampoco obtiene la agrupación correcta con ninguno de los LOD habituales. Como se detalla en la tabla correspondiente, el resultado más aproximado se determina a LOD 4. De los 20 grupos de

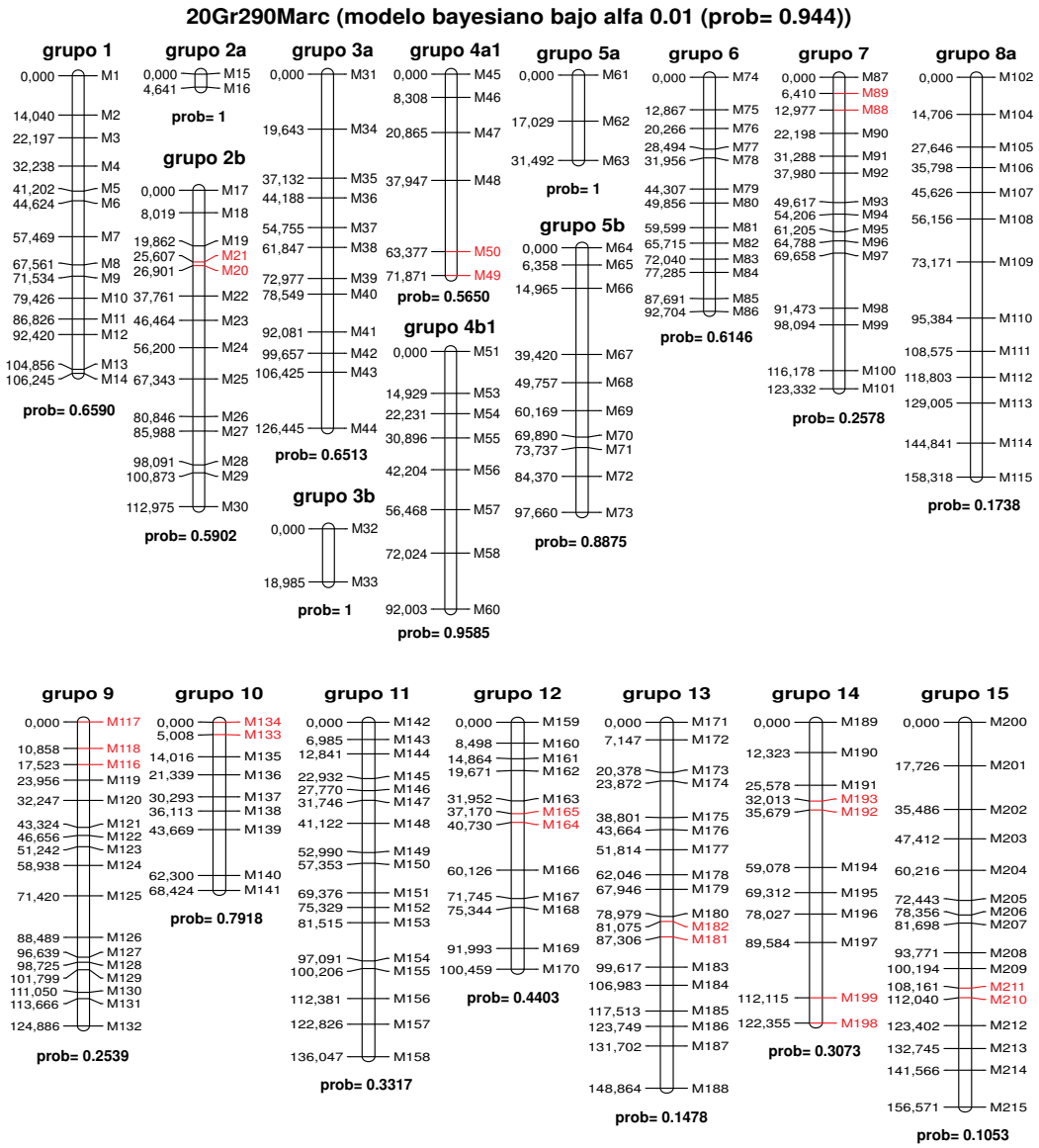
ligamiento, 13 de ellos coinciden con grupos reales: 1, 4, 6, 7, 9, 10, 11, 12, 14, 16, 17, 18 y 19. Por otra parte, 4 de ellos corresponden a fracciones de 2 grupos reales y 2 grupos de ligamiento han agrupado los marcadores de 4 grupos de ligamiento reales. Queda sin agrupar el marcador 103. Con este panorama, la ordenación propuesta tiene gran cantidad de errores excepto en 6 de los grupos de ligamiento, como se puede ver en la Figura 11.8

Con respecto a la determinación de los grupos de ligamiento que obtiene Mapmaker [48], la mejor aproximación se obtiene a LOD 4. De los 22 grupos de ligamiento estimados, 18 coinciden con grupos reales. Sin embargo, como en los dos apartados anteriores, no es capaz de determinar el orden de todos los marcadores en 21 de los 22 grupos de ligamiento. Quedan sin agrupar los marcadores 52, 59 y 103. Por todo ello, aunque la representación gráfica de la estimación del mapa genético se detalla en la Figura 11.9, no se comentará su comparativa con el modelo bayesiano y el modelo estimado por JoinMap [85].

20Gr290Marc 25 % faltantes	α=0.05	modelo bayesiano α=0.01	α=0.005
Agrupación 1	21 grupos (prob= 0.5595) grupo 1 grupo 2a grupo 2b.20 grupo 3a grupo 3b grupo 4a grupo 4b.16 grupo 5 grupo 6 grupo 7 grupo 8 grupo 9 grupo 10 grupo 11 grupo 12 grupo 13 grupo 14 grupo 15 grupo 17 grupo 18 grupo 19	**25 grupos (prob= 0.944)** **marc. sin grupo:** **52, 59, 103, 248** grupo 1 grupo 2a grupo 2b grupo 3a grupo 3b grupo 4a1 grupo 4b1 grupo 5a grupo 5b grupo 6 grupo 7 grupo 8a grupo 9 grupo 10 grupo 11 grupo 12 grupo 13 grupo 14 grupo 15 grupo 16 grupo 17a grupo 17b grupo 18a grupo 19 grupo 20	27 grupos (prob= 0.9815) marc. sin grupo: 52, 59, 103, 248 grupo 1 grupo 2a grupo 2b1 grupo 2b2 grupo 3a1 grupo 3b grupo 3a2 grupo 4a1 grupo 4b1 grupo 5a grupo 5b grupo 6 grupo 7 grupo 8a grupo 9 grupo 10 grupo 11 grupo 12 grupo 13 grupo 14 grupo 15 grupo 16 grupo 17a grupo 17b grupo 18a grupo 19 grupo 20
Agrupación 2	22 grupos con prob= 0.199	27 grupos + 3 marc sin grupo con prob= 0.0355	27 grupos + 5 marc sin grupo con prob= 0.004
Agrupación 3	22 grupos con prob= 0.1075	27 grupos + 3 marc sin grupo con prob= 0.0055	27 grupos + 5 marc sin grupo con prob= 0.0035
Agrupación 4	23 grupos con prob= 0.049	27 grupos + 3 marc sin grupo con prob= 0.004	28 grupos + 4 marc sin grupo con prob= 0.003
Agrupación 5	21 grupos + 1 marc sin grupo con prob= 0.0165	26 grupos + 4 marc sin grupo con prob= 0.003	28 grupos + 4 marc sin grupo con prob= 0.003
Agrupación 6	21 grupos + 1 marc sin grupo con prob= 0.0105	27 grupos + 3 marc sin grupo con prob= 0.003	28 grupos + 4 marc sin grupo con prob= 0.0015
Agrupación 7	22 grupos con prob= 0.009	27 grupos + 3 marc sin grupo con prob= 0.0015	27 grupos + 5 marc sin grupo con prob= 0.001
Agrupación 8	22 grupos con prob= 0.0075	26 grupos + 4 marc sin grupo con prob= 0.001	28 grupos + 4 marc sin grupo con prob= 0.001
Agrupación 9	22 grupos con prob= 0.004	27 grupos + 3 marc sin grupo con prob= 5e-04	28 grupos + 4 marc sin grupo con prob= 0.0015
Agrupación 10	22 grupos + 1 marc sin grupo con prob= 0.004	27 grupos + 3 marc sin grupo con prob= 5e-04	28 grupos + 4 marc sin grupo con prob= 5e-04
Agrupación 11	21 grupos + 1 marc sin grupo con prob= 0.0035	27 grupos + 4 marc sin grupo con prob= 5e-04	28 grupos + 4 marc sin grupo con prob= 5e-04
Agrupación 12	22 grupos + 1 marc sin grupo con prob= 0.0035	27 grupos + 3 marc sin grupo con prob= 5e-04	
Agrupación 13	23 grupos con prob= 0.003	27 grupos + 3 marc sin grupo con prob= 5e-04	
: :	: :		
Agrupación 35	23 grupos con prob= 0.0005		

20Gr290Marc 25 % faltantes	JoinMap		
	LOD 3	LOD 4	LOD 5
Agrupación única	13 grupos de ligamiento	**20 grupos de ligamiento marcador 103 sin grupo**	26 grupos de ligamiento marcadores sin grupo: 52, 59, 103, 201, 248
	grupo 1	grupo 1	grupo 1
	grupo 2.10.13.14.15.20	grupo 2.20	grupo 2a
	grupo 3	grupo 3a	grupo 2b1
	grupo 4.16	grupo 3b	grupo 2b2
	grupo 5	grupo 4	grupo 3c
	grupo 6.8	grupo 5a	grupo 3d
	grupo 7	grupo 5b	grupo 4a1
	grupo 9	grupo 6	grupo 4b1
	grupo 11	grupo 7	grupo 5a
	grupo 12	grupo 8a	grupo 5b
	grupo 17	grupo 9	grupo 6
	grupo 18	grupo 10	grupo 7
	grupo 19	grupo 11	grupo 8a
		grupo 12	grupo 9
		grupo 13.15	grupo 10
		grupo 14	grupo 11
		grupo 16	grupo 12
		grupo 17	grupo 13
		grupo 18	grupo 14
		grupo 19	grupo 15a
			grupo 16
			grupo 17a
			grupo 17b
			grupo 18a
			grupo 19
			grupo 20

20Gr290Marc 25 % faltantes	MapMaker		
	LOD 3	LOD 4	LOD 5
Agrupación única	16 grupos de ligamiento	**22 grupos de ligamiento**	25 grupos de ligamiento marcadores sin grupo: 52, 59 103
	grupo 1	grupo 1	grupo 1
	grupo 2.20	grupo 2	grupo 2a
	grupo 3	grupo 3a	grupo 2b
	grupo 4	grupo 3b	grupo 3a
	grupo 5	grupo 4	grupo 3b
	grupo 6.8	grupo 5	grupo 4a1
	grupo 7	grupo 6	grupo 4b1
	grupo 9	grupo 7	grupo 5a
	grupo 10	grupo 8	grupo 5b
	grupo 11	grupo 9	grupo 6
	grupo 12	grupo 10	grupo 7
	grupo 13.15	grupo 11	grupo 8a
	grupo 14	grupo 12	grupo 9
	grupo 16	grupo 13	grupo 10
	grupo 17.19	grupo 14	grupo 11
	grupo 18	grupo 15	grupo 12
		grupo 16	grupo 13
		grupo 17a	grupo 14
		grupo 17b	grupo 15
		grupo 18	grupo 16
		grupo 19	grupo 17a
		grupo 20	grupo 17b
			grupo 18
			grupo 19
			grupo 20


20Gr290Marc (modelo bayesiano bajo alfa 0.01 (prob= 0.944))
grupo 1
0,000 M1
14,040 M2
22,197 M3
32,238 M4
41,202 M5
44,624 M6
57,469 M7
67,561 M8
71,534 M9
79,426 M10
86,826 M11
92,420 M12
104,856 M13
106,245 M14
prob= 0.6590
grupo 2a
0,000 M15
4,641 M16
prob= 1
grupo 2b
0,000 M17
8,019 M18
19,862 M19
25,607 M21
26,901 M20
37,761 M22
46,464 M23
56,200 M24
67,343 M25
80,846 M26
85,988 M27
98,091 M28
100,873 M29
112,975 M30
prob= 0.5902
grupo 3a
0,000 M31
19,643 M34
37,132 M35
44,188 M36
54,755 M37
61,847 M38
72,977 M39
78,549 M40
92,081 M41
99,657 M42
106,425 M43
126,445 M44
prob= 0.6513
grupo 3b
0,000 M32
18,985 M33
prob= 1
grupo 4a1
0,000 M45
8,308 M46
20,865 M47
37,947 M48
63,377 M50
71,871 M49
prob= 0.5650
grupo 4b1
0,000 M51
14,929 M53
22,231 M54
30,896 M55
42,204 M56
56,468 M57
72,024 M58
92,003 M60
prob= 0.9585
grupo 5a
0,000 M61
17,029 M62
31,492 M63
prob= 1
grupo 5b
0,000 M64
6,358 M65
14,965 M66
39,420 M67
49,757 M68
60,169 M69
69,890 M70
73,737 M71
84,370 M72
97,660 M73
prob= 0.8875
grupo 6
0,000 M74
12,867 M75
20,266 M76
28,494 M77
31,956 M78
44,307 M79
49,856 M80
59,599 M81
65,715 M82
72,040 M83
77,285 M84
87,691 M85
92,704 M86
prob= 0.6146
grupo 7
0,000 M87
6,410 M89
12,977 M88
22,198 M90
31,288 M91
37,980 M92
49,617 M93
54,206 M94
61,205 M95
64,788 M96
69,658 M97
91,473 M98
98,094 M99
116,178 M100
123,332 M101
prob= 0.2578
grupo 8a
0,000 M102
14,706 M104
27,646 M105
35,798 M106
45,626 M107
56,156 M108
73,171 M109
95,384 M110
108,575 M111
118,803 M112
129,005 M113
144,841 M114
158,318 M115
prob= 0.1738
grupo 9
0,000 M117
10,858 M118
17,523 M116
23,956 M119
32,247 M120
43,324 M121
46,656 M122
51,242 M123
58,938 M124
71,420 M125
88,489 M126
96,639 M127
98,725 M128
101,799 M129
111,050 M130
113,666 M131
124,886 M132
prob= 0.2539
grupo 10
0,000 M134
5,008 M133
14,016 M135
21,339 M136
30,293 M137
36,113 M138
43,669 M139
62,300 M140
68,424 M141
prob= 0.7918
grupo 11
0,000 M142
6,985 M143
12,841 M144
22,932 M145
27,770 M146
31,746 M147
41,122 M148
52,990 M149
57,353 M150
69,376 M151
75,329 M152
81,515 M153
97,091 M154
100,206 M155
112,381 M156
122,826 M157
136,047 M158
prob= 0.3317
grupo 12
0,000 M159
8,498 M160
14,864 M161
19,671 M162
31,952 M163
37,170 M165
40,730 M164
60,126 M166
71,745 M167
75,344 M168
91,993 M169
100,459 M170
prob= 0.4403
grupo 13
0,000 M171
7,147 M172
20,378 M173
23,872 M174
38,801 M175
43,664 M176
51,814 M177
62,046 M178
67,946 M179
78,979 M180
81,075 M182
87,306 M181
99,617 M183
106,983 M184
117,513 M185
123,749 M186
131,702 M187
148,864 M188
prob= 0.1478
grupo 14
0,000 M189
12,323 M190
25,578 M191
32,013 M193
35,679 M192
59,078 M194
69,312 M195
78,027 M196
89,584 M197
112,115 M199
122,355 M198
prob= 0.3073
grupo 15
0,000 M200
17,726 M201
35,486 M202
47,412 M203
60,216 M204
72,443 M205
78,356 M206
81,698 M207
93,771 M208
100,194 M209
108,161 M211
112,040 M210
123,402 M212
132,745 M213
141,566 M214
156,571 M215
prob= 0.1053

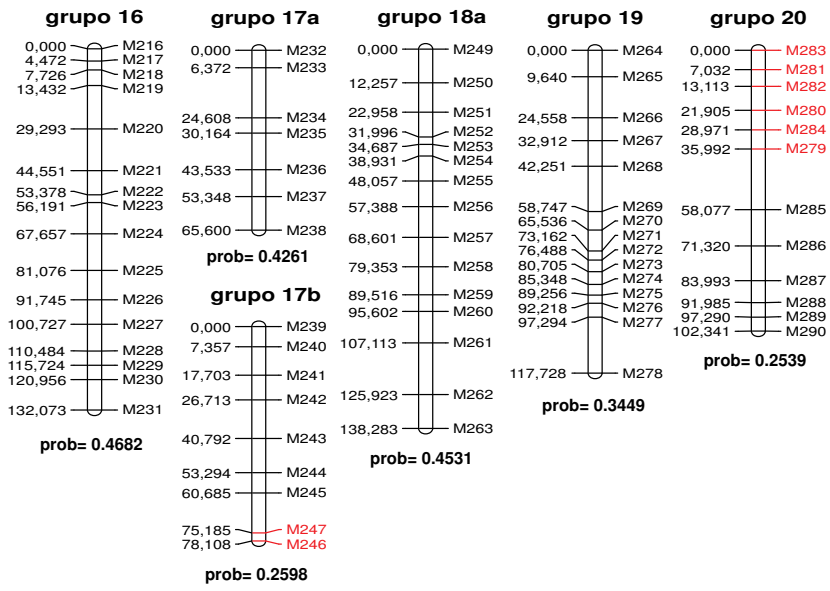
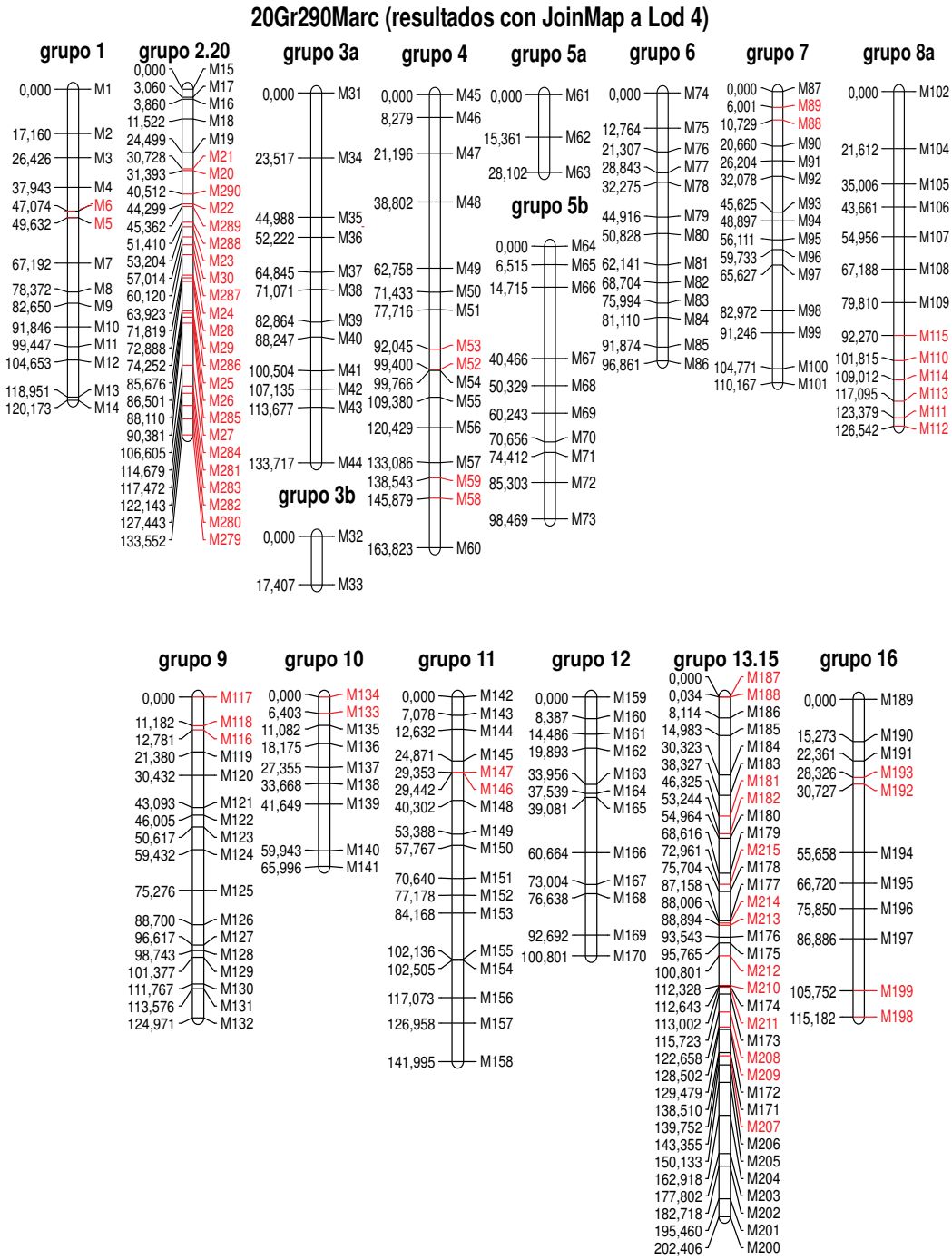


Figura 11.7: *Estimación bayesiana del mapa genético real de la población de la que procede la muestra 20Gr290Marc, con un 25 % de datos faltantes. Estructura obtenida con un nivel de significatividad global* $\alpha = 0.01$. *Marcadores sin agrupar: 52, 59, 103 y 248.*


20Gr290Marc (resultados con JoinMap a Lod 4)
grupo 1
0,000 M1
17,160 M2
26,426 M3
37,943 M4
47,074 M6
49,632 M5
67,192 M7
78,372 M8
82,650 M9
91,846 M10
99,447 M11
104,653 M12
118,951 M13
120,173 M14
grupo 2.20
0,000 M15
3,060 M17
3,860 M16
11,522 M18
24,499 M19
30,728 M21
31,393 M20
40,512 M290
44,299 M22
45,362 M289
51,410 M288
53,204 M23
57,014 M30
60,120 M287
63,923 M24
71,819 M28
72,888 M29
74,252 M286
85,676 M25
86,501 M26
88,110 M285
90,381 M27
106,605 M284
114,679 M281
117,472 M283
122,143 M282
127,443 M280
133,552 M279
grupo 3a
0,000 M31
23,517 M34
44,988 M35
52,222 M36
64,845 M37
71,071 M38
82,864 M39
88,247 M40
100,504 M41
107,135 M42
113,677 M43
133,717 M44
grupo 3b
0,000 M32
17,407 M33
grupo 4
0,000 M45
8,279 M46
21,196 M47
38,802 M48
62,758 M49
71,433 M50
77,716 M51
92,045 M53
99,400 M52
99,766 M54
109,380 M55
120,429 M56
133,086 M57
138,543 M59
145,879 M58
163,823 M60
grupo 5a
0,000 M61
15,361 M62
28,102 M63
grupo 5b
0,000 M64
6,515 M65
14,715 M66
40,466 M67
50,329 M68
60,243 M69
70,656 M70
74,412 M71
85,303 M72
98,469 M73
grupo 6
0,000 M74
12,764 M75
21,307 M76
28,843 M77
32,275 M78
44,916 M79
50,828 M80
62,141 M81
68,704 M82
75,994 M83
81,110 M84
91,874 M85
96,861 M86
grupo 7
0,000 M87
6,001 M89
10,729 M88
20,660 M90
26,204 M91
32,078 M92
45,625 M93
48,897 M94
56,111 M95
59,733 M96
65,627 M97
82,972 M98
91,246 M99
104,771 M100
110,167 M101
grupo 8a
0,000 M102
21,612 M104
35,006 M105
43,661 M106
54,956 M107
67,188 M108
79,810 M109
92,270 M115
101,815 M110
109,012 M114
117,095 M113
123,379 M111
126,542 M112
grupo 9
0,000 M117
11,182 M118
12,781 M116
21,380 M119
30,432 M120
43,093 M121
46,005 M122
50,617 M123
59,432 M124
75,276 M125
88,700 M126
96,617 M127
98,743 M128
101,377 M129
111,767 M130
113,576 M131
124,971 M132
grupo 10
0,000 M134
6,403 M133
11,082 M135
18,175 M136
27,355 M137
33,668 M138
41,649 M139
59,943 M140
65,996 M141
grupo 11
0,000 M142
7,078 M143
12,632 M144
24,871 M145
29,353 M147
29,442 M146
40,302 M148
53,388 M149
57,767 M150
70,640 M151
77,178 M152
84,168 M153
102,136 M155
102,505 M154
117,073 M156
126,958 M157
141,995 M158
grupo 12
0,000 M159
8,387 M160
14,486 M161
19,893 M162
33,956 M163
37,539 M164
39,081 M165
60,664 M166
73,004 M167
76,638 M168
92,692 M169
100,801 M170
grupo 13.15
0,000 M187
0,034 M188
8,114 M186
14,983 M185
30,323 M184
38,327 M183
46,325 M181
53,244 M182
54,964 M180
68,616 M179
72,961 M215
75,704 M178
87,158 M177
88,006 M214
88,894 M213
93,543 M176
95,765 M175
100,801 M212
112,328 M210
112,643 M174
113,002 M211
115,723 M173
122,658 M208
128,502 M209
129,479 M172
138,510 M171
139,752 M207
143,355 M206
150,133 M205
162,918 M204
177,802 M203
182,718 M202
195,460 M201
202,406 M200
grupo 16
0,000 M189
15,273 M190
22,361 M191
28,326 M193
30,727 M192
55,658 M194
66,720 M195
75,850 M196
86,886 M197
105,752 M199
115,182 M198

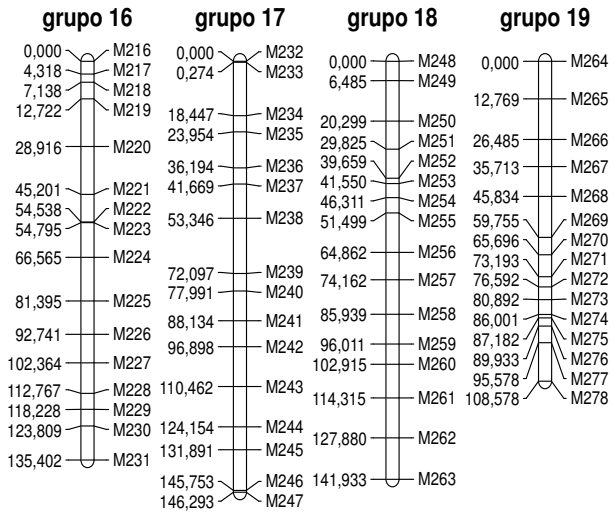
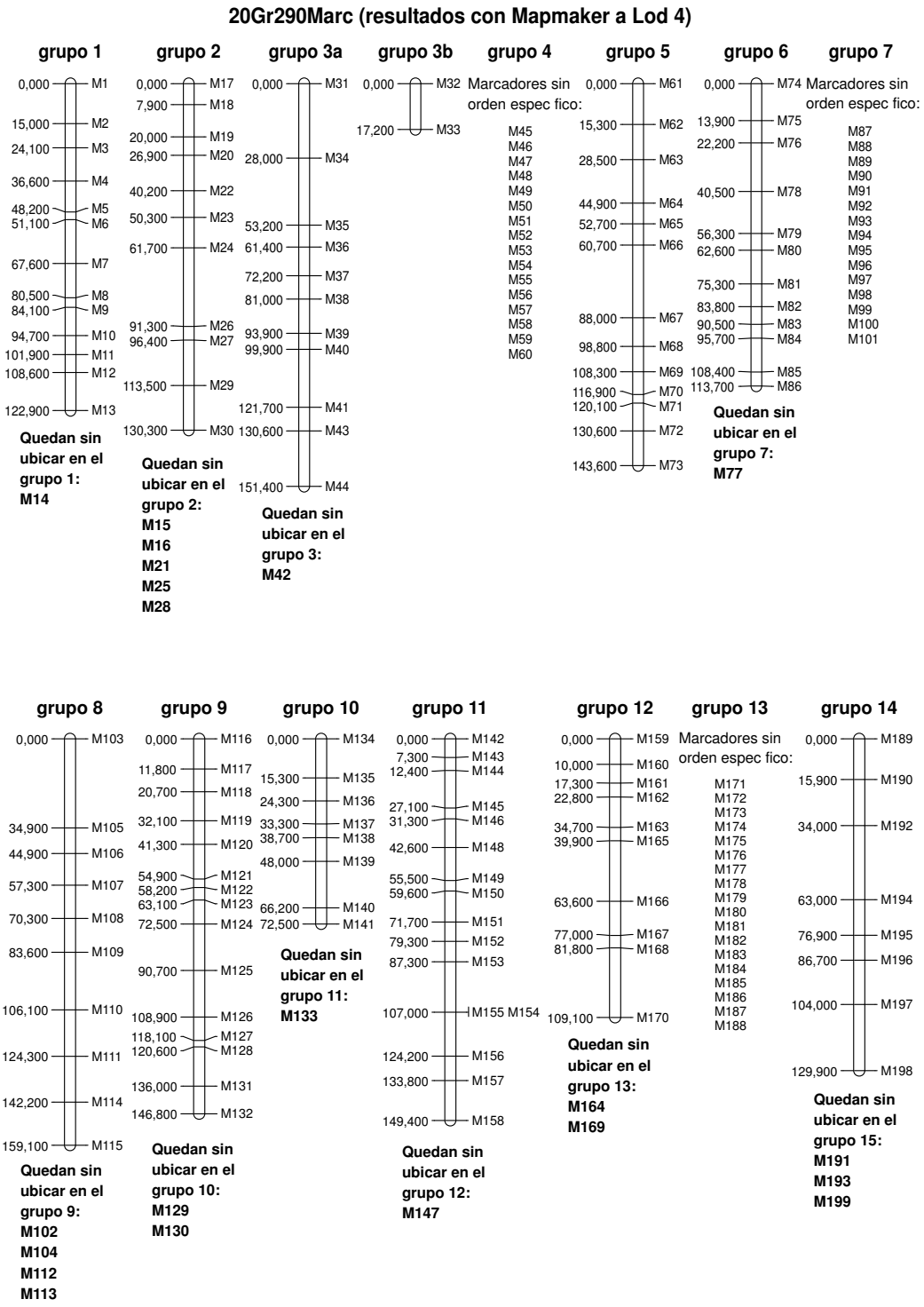


Figura 11.8: *Estimación, según el programa JoinMap, del mapa genético real de la población de la que procede la muestra 20Gr290Marc, con un 25 % de datos faltantes. Estructura obtenida a LOD 4. Marcadores sin agrupar: 103.*


20Gr290Marc (resultados con Mapmaker a Lod 4)
grupo 1
0,000 M1
15,000 M2
24,100 M3
36,600 M4
48,200 M5
51,100 M6
67,600 M7
80,500 M8
84,100 M9
94,700 M10
101,900 M11
108,600 M12
122,900 M13
Quedan sin ubicar en el grupo 1: M14
grupo 2
0,000 M17
7,900 M18
20,000 M19
26,900 M20
40,200 M22
50,300 M23
61,700 M24
91,300 M26
96,400 M27
113,500 M29
130,300 M30
Quedan sin ubicar en el grupo 2: M15 M16 M21 M25 M28
grupo 3a
0,000 M31
28,000 M34
53,200 M35
61,400 M36
72,200 M37
81,000 M38
93,900 M39
99,900 M40
121,700 M41
130,600 M43
151,400 M44
Quedan sin ubicar en el grupo 3: M42
grupo 3b
0,000 M32
17,200 M33
grupo 4
Marcadores sin orden espec fico:
M45 M46 M47 M48 M49 M50 M51 M52 M53 M54 M55 M56 M57 M58 M59 M60
grupo 5
0,000 M61
15,300 M62
28,500 M63
44,900 M64
52,700 M65
60,700 M66
88,000 M67
98,800 M68
108,300 M69
116,900 M70
120,100 M71
130,600 M72
143,600 M73
grupo 6
0,000 M74
13,900 M75
22,200 M76
40,500 M78
56,300 M79
62,600 M80
75,300 M81
83,800 M82
90,500 M83
95,700 M84
108,400 M85
113,700 M86
Quedan sin ubicar en el grupo 7: M77
grupo 7
Marcadores sin orden espec fico:
M87 M88 M89 M90 M91 M92 M93 M94 M95 M96 M97 M98 M99 M100 M101
grupo 8
0,000 M103
34,900 M105
44,900 M106
57,300 M107
70,300 M108
83,600 M109
106,100 M110
124,300 M111
142,200 M114
159,100 M115
Quedan sin ubicar en el grupo 9: M102 M104 M112 M113
grupo 9
0,000 M116
11,800 M117
20,700 M118
32,100 M119
41,300 M120
54,900 M121
58,200 M122
63,100 M123
72,500 M124
90,700 M125
108,900 M126
118,100 M127
120,600 M128
136,000 M131
146,800 M132
Quedan sin ubicar en el grupo 10: M129 M130
grupo 10
0,000 M134
15,300 M135
24,300 M136
33,300 M137
38,700 M138
48,000 M139
66,200 M140
72,500 M141
Quedan sin ubicar en el grupo 11: M133
grupo 11
0,000 M142
7,300 M143
12,400 M144
27,100 M145
31,300 M146
42,600 M148
55,500 M149
59,600 M150
71,700 M151
79,300 M152
87,300 M153
107,000 M155 M154
124,200 M156
133,800 M157
149,400 M158
Quedan sin ubicar en el grupo 12: M147
grupo 12
0,000 M159
10,000 M160
17,300 M161
22,800 M162
34,700 M163
39,900 M165
63,600 M166
77,000 M167
81,800 M168
109,100 M170
Quedan sin ubicar en el grupo 13: M164 M169
grupo 13
Marcadores sin orden espec fico:
M171 M172 M173 M174 M175 M176 M177 M178 M179 M180 M181 M182 M183 M184 M185 M186 M187 M188
grupo 14
0,000 M189
15,900 M190
34,000 M192
63,000 M194
76,900 M195
86,700 M196
104,000 M197
129,900 M198
Quedan sin ubicar en el grupo 15: M191 M193 M199

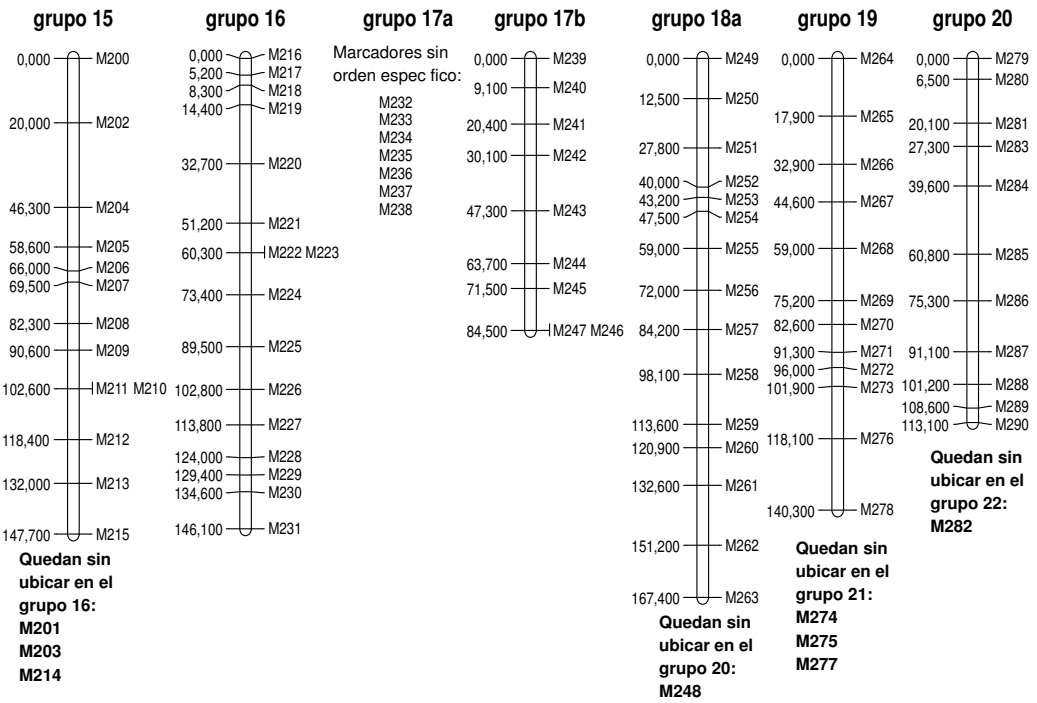


Figura 11.9: *Estimación, según el programa Mapmaker, del mapa genético real de la población de la que procede la muestra 20Gr290Marc, con un 25 % de datos faltantes. Estructura obtenida a LOD 4. Marcadores sin agrupar: 52, 59 y 103.*

11.4. Discusión

En la fase de la determinación de grupos de ligamiento, según aumenta el porcentaje de datos faltantes en la muestra, la metodología bayesiana resulta bastante estable. Eliminando aleatoriamente un 15 % de los datos de la muestra completa, se obtiene el reparto correcto de marcadores con una probabilidad de 0.974 y en el caso extremo en que la muestra tiene un 25 % de los datos faltantes, 7 grupos de ligamiento reales, se ven fraccionados en subgrupos, mientras que los otros 13 grupos de ligamiento reales se estiman correctamente. Esta estructura de grupos de ligamiento obtiene una probabilidad de 0.944. En ambos casos, ha sido necesario un nivel de significatividad global, en la resolución de los múltiples test de independencia, de $\alpha = 0.01$, mayor que el que ha sido preciso para obtener el resultado correcto en el ensayo con la muestra completa ($\alpha = 0.005$). El programa JoinMap [85] se muestra más inestable en la determinación de los grupos de ligamiento que el programa Mapmaker [48]. En ambos programas, se observa el reparto correcto de los marcadores en base a la muestra con un 15 % de datos faltantes a LOD 5. Sin embargo, también en ambos programas, es necesario bajar a LOD 4 para obtener el reparto de marcadores más aproximado al real.

Durante la fase de ordenación de los marcadores de cada grupo de ligamiento, la metodología bayesiana es capaz de obtener el orden correcto en 11 de los 20 grupos de ligamiento para la muestra completa y los otros 9 grupos de ligamiento difieren del mapa genético correcto en la permutación de entre 2 y 5 marcadores contiguos. La pérdida de un 15 % de la información en la muestra repercute en una disminución de las probabilidades obtenidas por cada ordenación. También se observa la disminución a 8 grupos de ligamiento totalmente correctos. Los que no son correctos difieren de los reales en la misma permutación de marcadores contiguos. Tras la pérdida del 25 % de la información de la muestra, 5 grupos de ligamiento coinciden con los reales y el resto se fraccionan en subgrupos. Se aprecia que en los subgrupos, a veces se recupera el orden correcto de los marcadores.

Resultados parecidos produce el programa JoinMap [85] en esta fase de ordenación de los marcadores. Obtiene el orden correcto en 13 de los 20 grupos de ligamiento para la muestra completa. Si se elimina aleatoriamente el 15 % de la información de la muestra, se reduce a 8 grupos de ligamiento correctamente ordenados y si la pérdida es del 25 % de los datos de la muestra, se obtiene

el orden correcto de marcadores en 6 de los grupos reales. Los grupos de ligamiento que no han sido correctamente ordenados difieren de los reales en la permutación de varios marcadores contiguos o en la mezcla de los marcadores de dos grupos de ligamiento reales.

Respecto al programa Mapmaker [48], aunque las ordenaciones de los marcadores que es capaz de incluir en el mapa, son correctas; sin embargo, tiene problemas para ubicar algunos marcadores en casi todos los grupos de ligamiento, incluso para la muestra completa, acentuándose este problema a medida que se aumenta el porcentaje de datos faltantes, hasta el punto que algunos grupos de ligamiento no tienen orden específico.

Tras los resultados, se deduce que la metodología bayesiana para la determinación de grupos de ligamiento y para la ordenación de marcadores dentro de cada grupo de ligamiento son satisfactorias y obtienen resultados razonables equiparables o incluso mejores a los que obtienen los programas de referencia JoinMap [85] y Mapmaker [48]. Comparando con los dos programas de referencia, la metodología bayesiana tiene un comportamiento similar al programa que mejor reparte los marcadores en los grupos de ligamiento (Mapmaker [48]) y después produce un comportamiento similar al programa que mejor ordena los marcadores dentro de cada grupo de ligamiento, ya sea JoinMap [85] o Mapmaker [48].

Capítulo 12

Aplicación a datos reales y comparación de resultados con métodos frecuentistas.

Para finalizar, se ensaya la metodología bayesiana definitiva resultante de todo el estudio para la determinación de grupos de ligamiento y la ordenación de los marcadores dentro de cada grupo de ligamiento, utilizando datos reales que provienen de la construcción de mapas en cítricos. Concretamente las muestras de datos utilizadas han sido generosamente cedidas por la Dra. M. José Asins Cebrián, responsable del Laboratorio de Genética, perteneciente al Centro de Protección Vegetal y Biotecnología del Instituto Valenciano de Investigaciones Agrarias (IVIA). Estas muestras (RetroReal201indv18Falt y F_2Real150indv4Falt) que se describen en el próximo párrafo, se corresponden básicamente, aunque no coinciden en su totalidad, con las muestras que proceden de las poblaciones F x Ch y Ch x F, estudiadas por Bernet et al (2010) [7] y $F_2(R \text{ x } P_r)$, estudiada por Raga et al (2012) [63]. F se refiere a la mandarina Fortune que es un híbrido derivado del cruce *C. clementina* Hort. ex Tan y *C. tangerina* Hort. ex Tan. Ch es el pummelo híbrido Chandler derivado del cruce entre dos variantes de *C. grandis* Osbeck. El parental femenino fue Fortune en 144 individuos (familia F x Ch) y Chandler en 57 individuos (familia Ch x F). Por otra parte, la población $F_2(R \text{ x } P_r)$ proviene de una autofecundación de tres híbridos RxPr, donde R es *Citrus reshni* Hort. ex Tan. (mandarino Cleopatra) y Pr es *Poncirus trifoliata* (L.) Raf. variedad

Flying Daragon.

Con dichos datos se distinguen dos escenarios. El primero de ellos corresponde a una población con 52 loci/ marcadores segregando como en un pseudo-Retrocruce (Grattapaglia y Sederoff 1994 [27]); es decir, la codificación de los marcadores se realizó como un Retrocruce, aunque conceptualmente no lo fuera (cruce entre una F_1 y uno de sus parentales). De esta población se dispone de una muestra de tamaño 201 individuos con un 18 % de datos faltantes, RetroReal201indv18Falt. El segundo escenario corresponde a una población F_2 definida por 53 marcadores. En este caso, se dispone de una muestra de tamaño 150 individuos con un 4 % de datos faltantes, F_2Real150indv4Falt.

Nótese que el hecho de trabajar en el entorno de poblaciones reales implica el desconocimiento del mapa genético real de la población y el porcentaje de datos faltantes de las muestras de que se disponen no es controlado.

A continuación, se detallan los resultados obtenidos por la metodología bayesiana y la comparativa con los programas JoinMap [85] y Mapmaker [48], como en los capítulos anteriores.

12.1. Resultados para datos reales procedentes de un diseño pseudo-Retrocruce

Dada la muestra RetroReal201indv18Falt, extraída de una población Retrocruce real, se aplica la metodología bayesiana descrita en el Capítulo 10, para la determinación de los grupos de ligamiento. Como resultado más probable e independientemente del nivel de significatividad global exigido en la resolución de los múltiples test de independencia ($\alpha = 0.05$, $\alpha = 0.01$ y $\alpha = 0.005$), se obtiene una estructura de 9 grupos de ligamiento en la que dos de los marcadores quedan sin agrupar. Esta estructura alcanza una probabilidad de 0.965 para $\alpha = 0.05$, una probabilidad de 0.759 para $\alpha = 0.01$ y una probabilidad 0.731 para $\alpha = 0.005$. Tras este estable resultado se ordenan los marcadores dentro de cada grupo. En el Cuadro del 12.1 se resumen las distancias multipunto y desviaciones típicas posteriores entre los marcadores contiguos dentro de cada grupo de ligamiento. La representación gráfica de esta información aparece en la Figura 12.1. Como en ocasiones anteriores, debajo de cada grupo de ligamiento se especifica la probabilidad de la ordenación que ha sido valorada como más probable. La forma de nombrar los grupos es

la acordada en los dos capítulos anteriores.

En las Figuras 12.2 y 12.3 aparecen los resultados obtenidos, con los mismos datos, utilizando el programa JoinMap [85]. Como se puede ver, el programa determina los mismos 9 grupos de ligamiento, a LOD 3 y 4, que la metodología bayesiana. Una diferencia es la ubicación del marcador S2AS2 del grupo 5. Igualmente, la longitud total de los grupos es diferente, ya que emplean métodos de estimación diferentes. A LOD 5 se obtienen 10 grupos de ligamiento. Este resultado implica el desglose del grupo 3 en 3a y 3b y la pérdida de un marcador en cada uno de los grupos 5 y 7. El orden de los marcadores dentro de cada grupo no varía.

El programa Mapmaker [48] también obtiene dos estructuras de grupos según el LOD. Para LOD 3 y 4, Figura 12.4, se obtienen 8 grupos de ligamiento pero no coinciden exactamente con los determinados por la metodología bayesiana y por el programa JoinMap [85] a ese mismo LOD. Concuerdan 7 de ellos. El otro contiene los marcadores de los grupos 4 y 5 determinados por las otras metodologías. Es decir, grupo 4.5. De este grupo, 3 marcadores no consiguen una ubicación específica en la fase de ordenación. El grupo 7 pierde uno de sus marcadores, de manera que definitivamente, 3 de los marcadores quedan sin agrupar. Si pasamos a los resultados obtenidos a LOD 5, Figura 12.5, se obtienen 9 grupos de ligamiento. Mismo resultado que a LOD 3 y 4 pero el grupo 3 se desglosa en 3a y 3b. Los marcadores 5R-4R,850 y Aint-Drt900 no se agrupan con ninguno de los tres métodos, independientemente del α o LOD utilizado.

RetroReal201indv18Falt (modelo bayesiano)		
grupo	marcadores contiguos	media ± desv. típ.
1	CT19; CR66	0.09188 ± 0.02781
	CR66; 5F-4R,800	0.01764 ± 0.01237
	5F-4R,800; CTVCH28	0.07792 ± 0.02375
2	5F-4R,950; CTVCH1	0.21019 ± 0.04557
3	CR20 ;ETHREC	0.12686 ± 0.04712
	ETHREC ;CR12,600	0.06793 ± 0.02760
	CR12,600 ;AintCrt1200	0.39695 ± 0.07558
	AintCrt1200 ;AintC8rt320	0.02938 ± 0.03051
	AintC8rt320 ;5F-5R,220	0.06025 ± 0.03354
	5F-5R,220 ;CR16,1080	0.43515 ± 0.08858
	CR16,1080 ;C8intDrt1018	0.14486 ± 0.04262
4	C11,400 ; CR73	0.33469 ± 0.05227
	CR73 ; BintCrt600	0.13559 ± 0.02797
	BintCrt600 ; cNHX1,Bg1II	0.03906 ± 0.01242
	cNHX1,Bg1II ; TAA27	0.06925 ± 0.01609
	TAA27 ; AinC11r1400	0.01785 ± 0.00812
	AinC11r1400 ; AintBint810	0.08490 ± 0.03377
	AintBint810 ; CMS16	0.08904 ± 0.04177
	CMS16 ; CR71	0.01234 ± 0.00875
	CR71 ; CTVCH10	0.16343 ± 0.02671

continuación...

RetroReal201indv18Falt (modelo bayesiano)		
grupo	marcadores contiguos	media ± desv. típ.
5	CR3 ; CMS24	0.28919 ± 0.15170
	CMS24 ; CMS31	0.11945 ± 0.02609
	CMS31 ; CMS48,700	0.15741 ± 0.04001
	CMS48,700 ; CAG01	0.03686 ± 0.01906
	CAG01 ; BintC11rt650	0.09609 ± 0.02824
	BintC11rt650; S2AS2	0.42194 ± 0.23852
	S2AS2 ; BintC8rt900	0.64064 ± 0.45342
6	CR52 ; BintDrt990	0.43593 ± 0.06248
	BintDrt990 ; AintBint300	0.02319 ± 0.01055
7	CR22,1018 ; CTVCH20	0.46916 ± 0.05758
	CTVCH20 ; CR39,900	0.04165 ± 0.01962
	CR39,900 ; C8intCrt110	0.02760 ± 0.01235
	C8intCrt110 ; AintBint700	0.02569 ± 0.01254
	AintBint700 ; CMS20	0.05972 ± 0.02480
8	BintC8rt600 ; AintC11rt900	0.03242 ± 0.01180
	AintC11rt900 ; CR5,1000	0.03984 ± 0.01381
9	CMS4 ; GT03,171	0.08657 ± 0.02360
	GT03,171 ; VIC	0.26043 ± 0.04115
	VIC ; CR16,450	0.28439 ± 0.04757
	CR16,450 ; CTVCH17	0.07475 ± 0.02140
	CTVCH17 ; CMS47,160	0.11871 ± 0.02769

Cuadro 12.1: *Estimación bayesiana de las distancias multipunto y desviaciones típicas posteriores entre marcadores contiguos, para los grupos de ligamiento de la población Retrocruce real. Marcadores sin agrupar: 5R-4R,850 y AintDrt900*

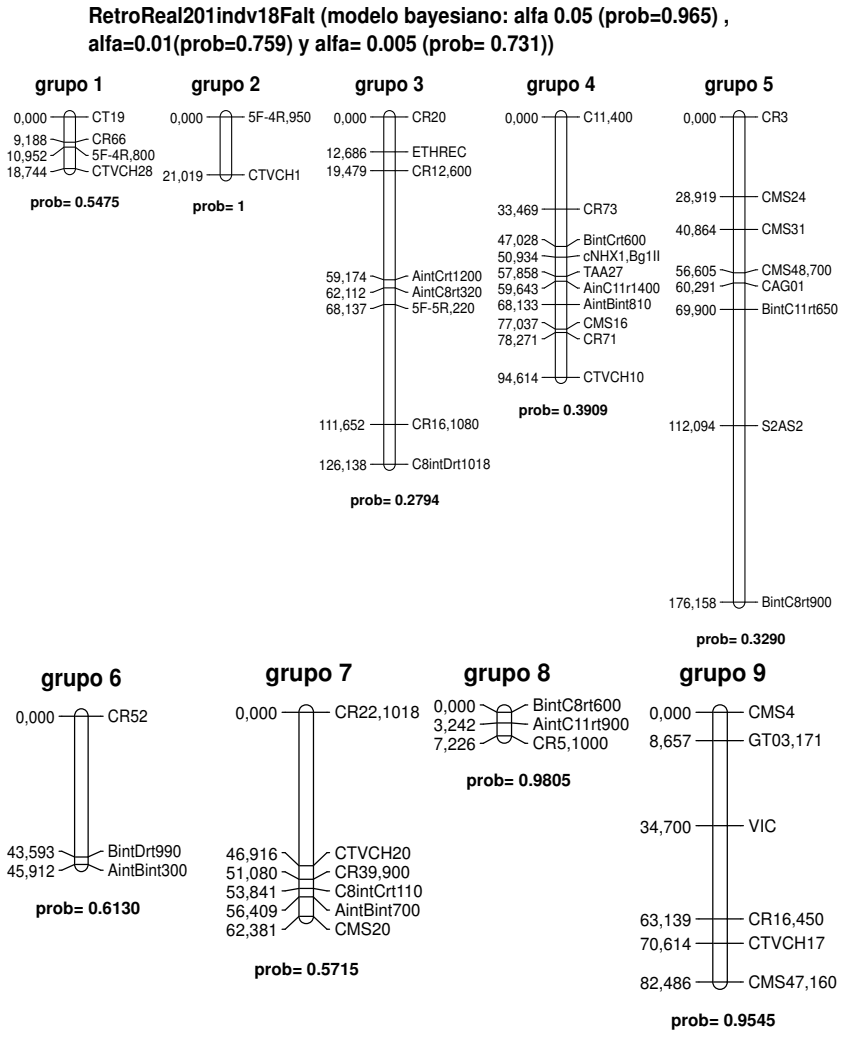


Figura 12.1: *Grupos de ligamiento y ordenaciones de los marcadores dentro de cada grupo de ligamiento obtenidos, según el modelo bayesiano, para el banco de datos RetroReal201indv18Falt, bajo α=0.05, α=0.01 y α=0.005. Marcadores sin agrupar: 5R-4R,850 y AintDrt900*

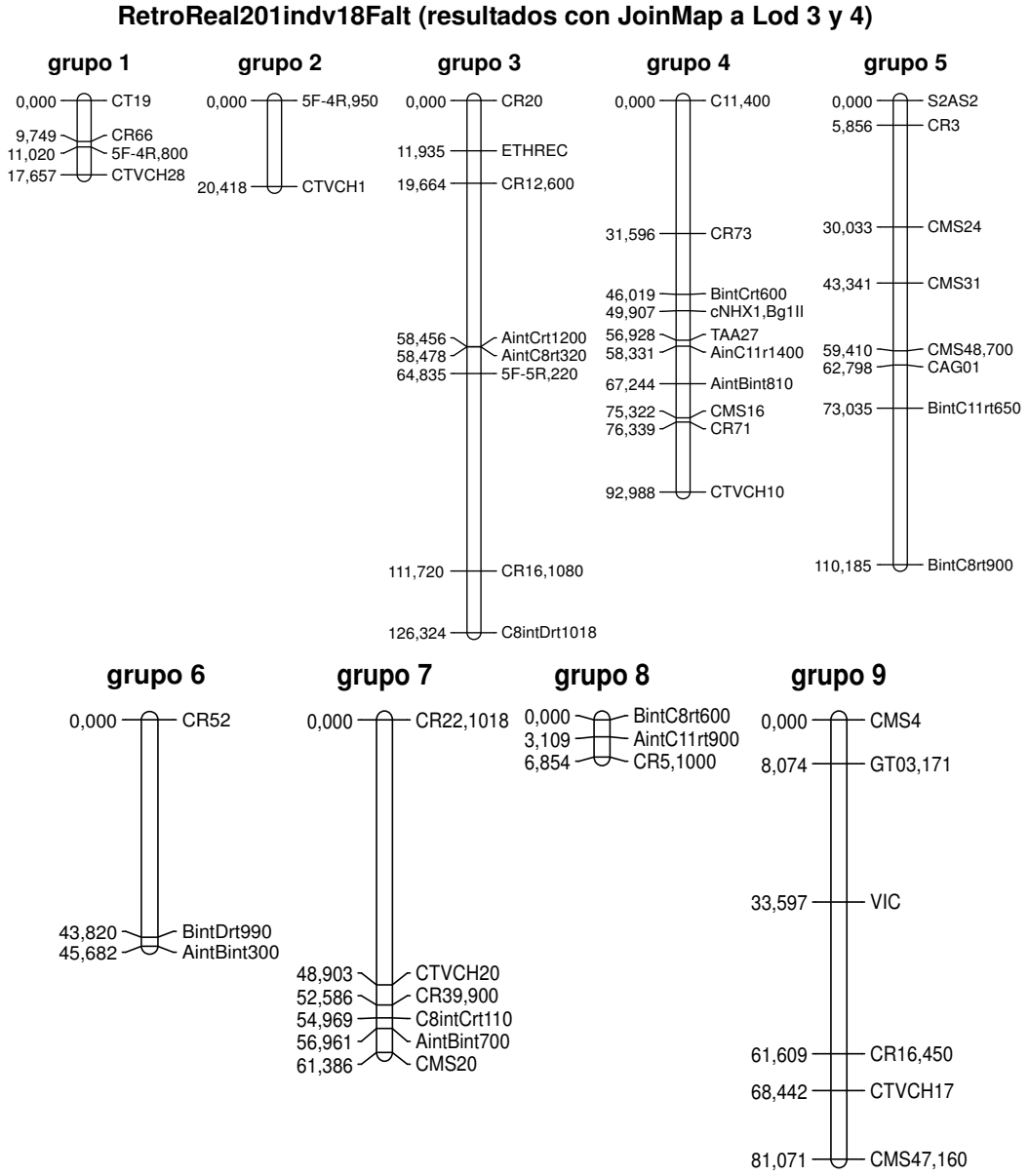


Figura 12.2: *Grupos de ligamiento y ordenaciones de los marcadores dentro de cada grupo de ligamiento obtenidos, según el programa JoinMap, para el banco de datos RetroReal201indv18Falt, bajo LOD 3 y 4. Marcadores sin agrupar: 5R-4R,850 y AintDrt900*

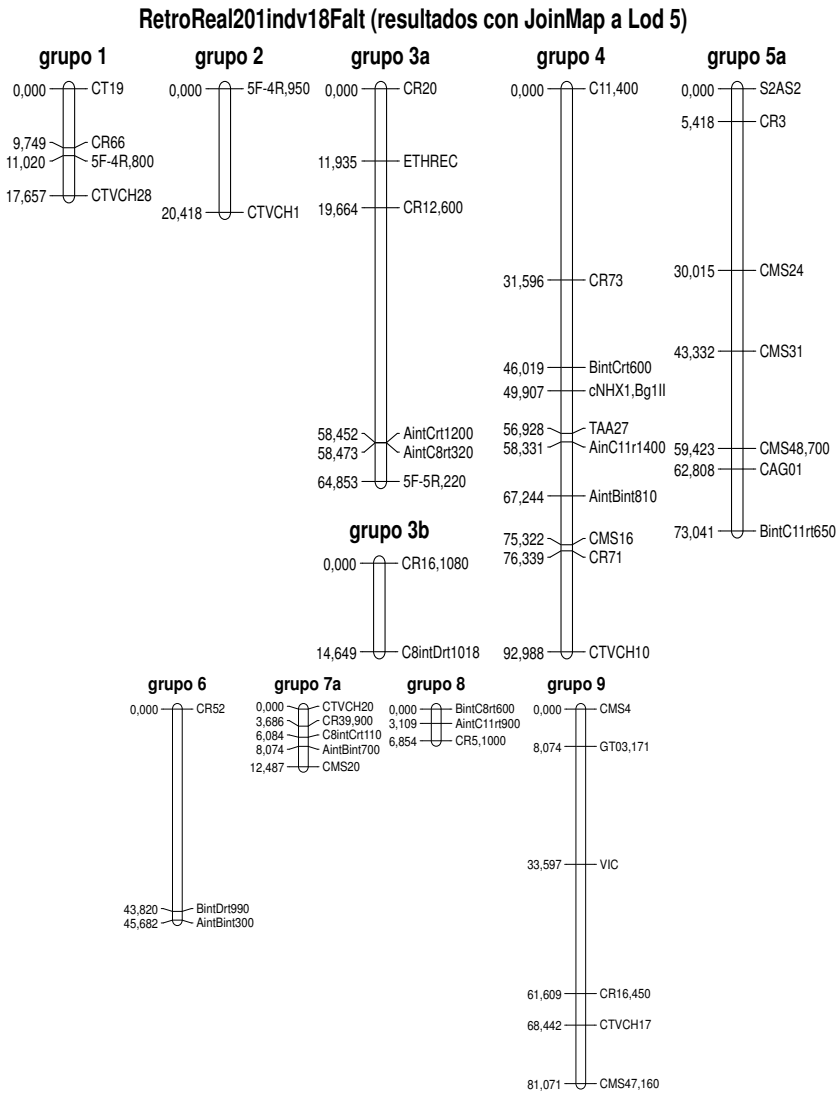


Figura 12.3: *Grupos de ligamiento y ordenaciones de los marcadores dentro de cada grupo de ligamiento obtenidos, según el programa JoinMap, para el banco de datos RetroReal201indv18Falt, bajo LOD 5. Marcadores sin agrupar: 5R-4R,850, AintDrt900, BintC8rt900 y CR22,1018*

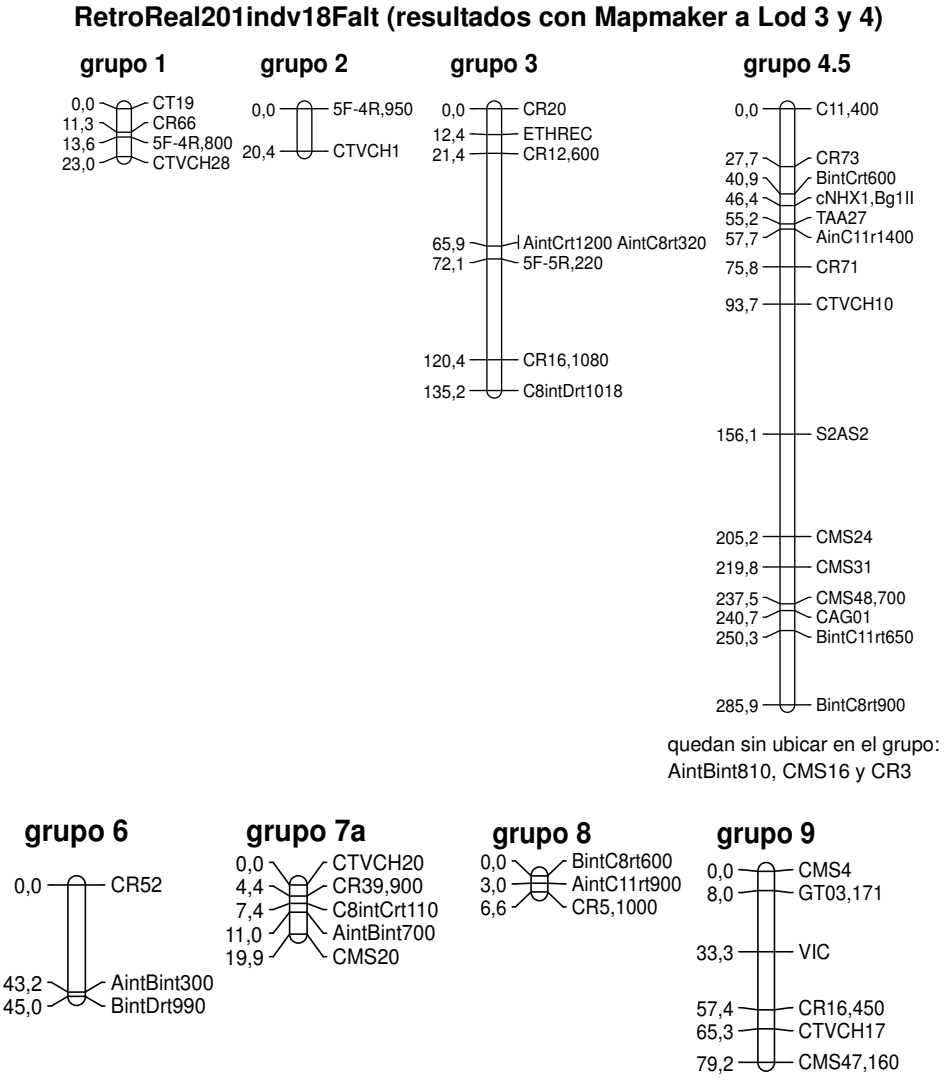


Figura 12.4: *Grupos de ligamiento y ordenaciones de los marcadores dentro de cada grupo de ligamiento obtenidos, según el programa Mapmaker, para el banco de datos RetroReal201indv18Falt, bajo LOD 3 y 4. Marcadores sin agrupar: 5R-4R,850, AintDrt900 y CR22,1018*

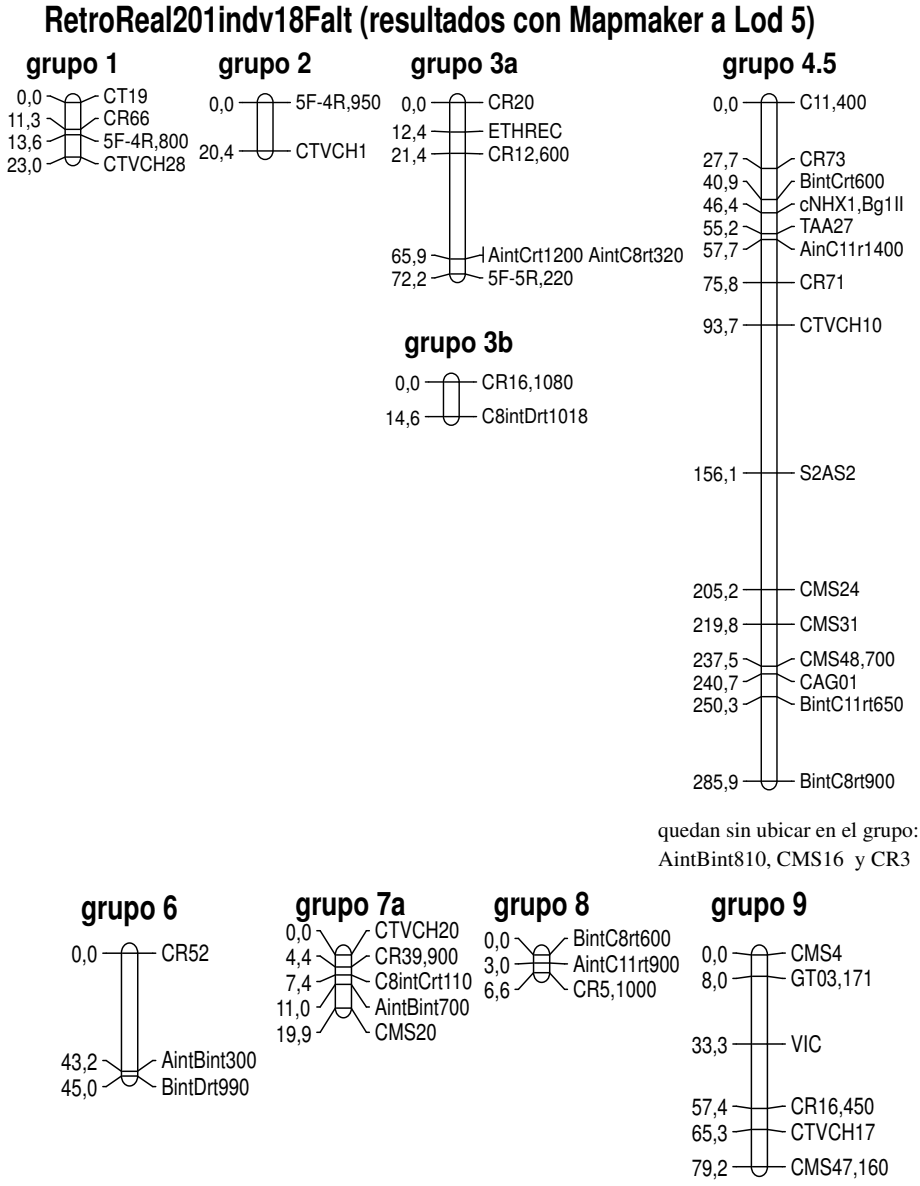


Figura 12.5: *Grupos de ligamiento y ordenaciones de los marcadores dentro de cada grupo de ligamiento obtenidos, según el programa Mapmaker, para el banco de datos RetroReal201indv18Falt, bajo LOD 5. Marcadores sin agrupar: 5R-4R,850, AintDrt900 y CR22,1018*

12.2. Resultados para datos reales procedentes de una población F_2 real.

En este apartado se repite una presentación de resultados similar al del apartado anterior sobre la muestra de datos F_2Real150indv4Falt, que proviene de una población cuyos marcadores segregan como una F_2 real definida por 53 marcadores y que tiene un 4 % de datos faltantes.

Tras la fase de determinación de grupos de ligamiento, con la metodología bayesiana, se obtienen tres estructuras de grupos de ligamiento parecidas dependiendo del nivel de significatividad global exigido en la resolución de los múltiples test de independencia. Para un nivel de significatividad global $\alpha = 0.05$ se obtienen 12 grupos de ligamiento con una probabilidad de 0.6945. Para un nivel de significatividad global $\alpha = 0.01$ se obtienen, de nuevo, 12 grupos de ligamiento, con una probabilidad de 0.937. La diferencia es que el grupo 4 pierde al marcador $CR12\_540$ (grupo 4a) y para un nivel de significatividad global $\alpha = 0.005$ se obtienen 13 grupos de ligamiento, con una probabilidad de 0.952. En este caso, el grupo 4a se desglosa en 4a1 y 4a2. Por lo tanto, el número de grupos se mantiene bastante estable. En la fase de ordenación, no se aprecian variaciones según se pasa de una estructura a otra, ni en los subgrupos fraccionados. En los Cuadros del 12.3 al 12.4 se detallan los nombres de los marcadores contiguos implicados en la ordenación más probable de cada grupo de ligamiento, junto a los tipos de marcador, la media y la desviación típica de la distribución posterior de distancias multipunto. La representación gráfica de estos cuadros, aparecen en las Figuras 12.6, 12.7 y 12.8 y, describen la estimación del mapa genético de la población de la que proviene la muestra, F_2Real150indv4Falt.

Las estimaciones que obtiene el programa JoinMap [85] aparece representadas en las Figuras 12.9, 12.10 y 12.11. Como se puede observar, a LOD 3, se determinan 11 grupos de ligamiento. La ordenación de 8 de los grupos coincide con los homólogos bayesianos. Los marcadores del grupo 1.2 y del grupo 4 no obtienen ningún orden específico, por "ligamiento insuficiente en los datos para completar el mapa" (mensaje especificado por JoinMap). Un par de marcadores del grupo 3 permutan respecto a la ordenación del mismo grupo bayesiano. La estimación del mapa genético a LOD 4, proporciona 11 grupos de ligamiento. En este caso, se mantiene prácticamente el mismo resultado anterior. La única diferencia es que el grupo 4 pierde el marcador $CR2\_540$ y

consigue, de ese modo, elaborar un orden específico que coincide con el estimado de forma bayesiana. A LOD 5, se obtienen 12 grupos de ligamiento. El grupo 4a se desglosa en 4a1 y 4a2. Todo lo demás queda invariante.

Por último, las estimaciones del mapa genético correspondientes al programa Mapmaker [48] se representan en las Figuras 12.12, 12.13 y 12.14. Como se puede ver, a LOD 3 se determinan 8 grupos de ligamiento, 3 de los cuales coinciden en reparto de marcadores y orden con los estimados por la metodología bayesiana. El grupo (1)* contiene los marcadores del grupo 1 más el marcador $CR16\_250$. El grupo (2.3.9.10)** agrupa los marcadores de los 4 grupos correspondientes más 3 marcadores ($C1iC8rt, 350$; $C1iC8rt, 510$ y $CR13, 489$) y no logra emitir un orden específico. El grupo 4, pierde el marcador $CR12\_540$ y el grupo (12)*** contiene los marcadores del grupo 12 más el marcador $28R, 320$. Como resultado de considerar LOD 4, se determinan 11 grupos de ligamiento, 9 de los cuales se corresponden con la agrupación y con el orden de marcadores bayesianos bajo un $\alpha = 0.01$. El grupo 1, continúa sin orden específico para 5 de sus marcadores. A LOD 5, se obtienen 12 grupos de ligamiento. Se repite el resultado anterior a diferencia que que el grupo 6.7 se desglosa en el grupo 6 y el grupo 7; de este modo se logra una coincidencia de 10 grupos con la metodología bayesiana, bajo $\alpha = 0.01$.

F_2Real150indv4Falt (modelo bayesiano con α= 0.05)			
grupo	marcadores contiguos	tipo	media ± desv. típ.
1	CR72,250 ; 520AR,350	C ; D2	0.40045 ± 0.08311
	520AR,350 ; CR3,330	D2 ; D1	0.04975 ± 0.04776
	CR3,330 ; CMS48,700	D1 ; C	0.13216 ± 0.03032
	CMS48,700 ; Py28,230	C ; C	0.03705 ± 0.01235
	Py28,230 ; Py65C,510	C ; D2	0.01463 ± 0.01407
	Py65C,510 ; AGG9,120	D2 ; C	0.17481 ± 0.02537
	AGG9,120 ; ATC09_180	C ; C	0.01778 ± 0.01521
	ATC09_180 ; 1R,700	C ; C	0.11983 ± 0.02151
	1R,700 ; CAT01_160	C ; C	0.20329 ± 0.05444

continuación...

F_2Real150indv4Falt (modelo bayesiano con α= 0.05)			
grupo	marcadores contiguos	tipo	media ± desv. típ.
2	15R,800 ; CR23,750	C ; C	0.08576 ± 0.01914
	CR23,750 ; CR28,270	C ; C	0.11306 ± 0.01675
	CR28,270 ; CR15,1100	C ; C	0.00869 ± 0.00432
	CR15,1100; 27R_1500	C ; C	0.12563 ± 0.01913
3	CR18,200; CR5_180	D2 ; D1	0.11403 ± 0.04031
	CR5_180; CR17,330	D1 ; C	0.03007 ± 0.01349
	CR17,330; 76A_180	C ; C	0.23914 ± 0.03616
	76A_180; Ethrec,1200	C ; C	0.10624 ± 0.02318
4	CR52,340 ; R1010,520	C ; C	0.08934 ± 0.01549
	R1010,520 ; TAA15,170	C ; C	0.03803 ± 0.00893
	TAA15,170 ; CR14,300	C ; D1	0.04448 ± 0.02171
	CR14,300 ; CR74,344	D1 ; C	0.03800 ± 0.01461
	CR74,344 ; 68A_240	C ; C	0.37618 ± 0.10318
	68A_240 ; TAA52,100	C ; D1	0.07995 ± 0.02088
	TAA52,100 ; CR12_540	D1 ; D2	0.37871 ± 0.15420
5	CAC23,240; CR16_650	C ; D2	0.11983 ± 0.03347
6	CR13,506 ; CitSOS1,480	D1 ; C	0.14055 ± 0.04658
7	CR12_250; TAA41,160	C ; C	0.34004 ± 0.06531
8	VIC,380 ; GT03_180	C ; C	0.31067 ± 0.04873
	GT03_180; CMS4,180	C ; C	0.05412 ± 0.01486
9	HLH_400 ; CMS20,180	C ; D1	0.16370 ± 0.04114
	CMS20,180; CR54,310	D1 ; D2	0.15925 ± 0.09908
10	CMS14,160; CR36_310	D1 ; C	0.13394 ± 0.0418
11	CR19,380; TAA1,175	C ; C	0.20267 ± 0.04255
12	TAA27,250; CNHX1,1636	C ; C	0.22088 ± 0.04467

Cuadro 12.2: *Estimación bayesiana de las distancias multipunto y desviaciones típicas posteriores entre marcadores contiguos, para los grupos de ligamiento de la población F_2 real determinados bajo $\alpha = 0.05$. Marcadores sin agrupar: $C1iC8rt, 350$, $DREB\_480$, $CR5\_600$, $CR16\_250$, $C1iC8rt, 510$, $28R, 320$, $CR13, 480$*

F_2Real150indv4Falt (modelo bayesiano con α= 0.01)			
grupo	marcadores contiguos	tipo	media ± desv. típ.
1	CR72,250 ; 520AR,350	C ; D2	0.40045 ± 0.08311
	520AR,350 ; CR3,330	D2 ; D1	0.04975 ± 0.04776
	CR3,330 ; CMS48,700	D1 ; C	0.13216 ± 0.03032
	CMS48,700 ; Py28,230	C ; C	0.03705 ± 0.01235
	Py28,230 ; Py65C,510	C ; D2	0.01463 ± 0.01407
	Py65C,510 ; AGG9,120	D2 ; C	0.17481 ± 0.02537
	AGG9,120 ; ATC09_180	C ; C	0.01778 ± 0.01521
	ATC09_180 ; 1R,700	C ; C	0.11983 ± 0.02151
	1R,700 ; CAT01_160	C ; C	0.20329 ± 0.05444
2	15R,800 ; CR23,750	C ; C	0.08576 ± 0.01914
	CR23,750 ; CR28,270	C ; C	0.11306 ± 0.01675
	CR28,270 ; CR15,1100	C ; C	0.00869 ± 0.00432
	CR15,1100; 27R_1500	C ; C	0.12563 ± 0.01913
3	CR18,200; CR5_180	D2 ; D1	0.11403 ± 0.04031
	CR5_180; CR17,330	D1 ; C	0.03007 ± 0.01349
	CR17,330; 76A_180	C ; C	0.23914 ± 0.03616
	76A_180; Ethrec,1200	C ; C	0.10624 ± 0.02318
4a	CR52,340 ; R1010,520	C ; C	0.08991 ± 0.01540
	R1010,520 ; TAA15,170	C ; C	0.03810 ± 0.00894
	TAA15,170 ; CR14,300	C ; D1	0.04462 ± 0.02180
	CR14,300 ; CR74,344	D1 ; C	0.03803 ± 0.01455
	CR74,344 ; 68A_240	C ; C	0.37667 ± 0.09574
	68A_240 ; TAA52,100	C ; D1	0.08038 ± 0.02145
5	CAC23,240; CR16_650	C ; D2	0.11983 ± 0.03347
6	CR13,506 ; CitSOS1,480	D1 ; C	0.14055 ± 0.04658
7	CR12_250; TAA41,160	C ; C	0.34004 ± 0.06531
8	VIC,380 ; GT03_180	C ; C	0.31067 ± 0.04873
	GT03_180; CMS4,180	C ; C	0.05412 ± 0.01486

continuación...

F_2Real150indv4Falt (modelo bayesiano con α= 0.01)			
grupo	marcadores contiguos	tipo	media ± desv. típ.
9	HLH_400 ; CMS20,180	C ; D1	0.16370 ± 0.04114
	CMS20,180; CR54,310	D1 ; D2	0.15925 ± 0.09908
10	CMS14,160; CR36_310	D1 ; C	0.13394 ± 0.0418
11	CR19,380; TAA1,175	C ; C	0.20267 ± 0.04255
12	TAA27,250; CNHX1,1636	C ; C	0.22088 ± 0.04467

Cuadro 12.3: *Estimación bayesiana de las distancias multipunto y desviaciones típicas posteriores entre marcadores contiguos, para los grupos de ligamiento de la población F_2 real determinados bajo $\alpha = 0.01$. Marcadores sin agrupar: $C1iC8rt, 350$, $DREB\_480$, $CR5\_600$, $CR16\_250$, $C1iC8rt, 510$, $28R, 320$, $CR13, 480$, $CR12\_540$*

F_2Real150indv4Falt (modelo bayesiano con α= 0.005)			
grupo	marcadores contiguos	tipo	media ± desv. típ.
1	CR72,250 ; 520AR,350	C ; D2	0.40045 ± 0.08311
	520AR,350 ; CR3,330	D2 ; D1	0.04975 ± 0.04776
	CR3,330 ; CMS48,700	D1 ; C	0.13216 ± 0.03032
	CMS48,700 ; Py28,230	C ; C	0.03705 ± 0.01235
	Py28,230 ; Py65C,510	C ; D2	0.01463 ± 0.01407
	Py65C,510 ; AGG9,120	D2 ; C	0.17481 ± 0.02537
	AGG9,120 ; ATC09_180	C ; C	0.01778 ± 0.01521
	ATC09_180 ; 1R,700	C ; C	0.11983 ± 0.02151
	1R,700 ; CAT01_160	C ; C	0.20329 ± 0.05444
2	15R,800 ; CR23,750	C ; C	0.08576 ± 0.01914
	CR23,750 ; CR28,270	C ; C	0.11306 ± 0.01675
	CR28,270 ; CR15,1100	C ; C	0.00869 ± 0.00432
	CR15,1100; 27R_1500	C ; C	0.12563 ± 0.01913
3	CR18,200; CR5_180	D2 ; D1	0.11403 ± 0.04031
	CR5_180; CR17,330	D1 ; C	0.03007 ± 0.01349
	CR17,330; 76A_180	C ; C	0.23914 ± 0.03616
	76A_180; Ethrec,1200	C ; C	0.10624 ± 0.02318
4a1	CR52,340 ; R1010,520	C ; C	0.09548 ± 0.01609
	R1010,520 ; TAA15,170	C ; C	0.03892 ± 0.00894
	TAA15,170 ; CR14,300	C ; D1	0.04504 ± 0.02278
	CR14,300 ; CR74,344	D1 ; C	0.03906 ± 0.01499
4a2	68A_240 ; TAA52,100	C ; D1	0.09673 ± 0.0308
5	CAC23,240; CR16_650	C ; D2	0.11983 ± 0.03347
6	CR13,506 ; CitSOS1,480	D1 ; C	0.14055 ± 0.04658
7	CR12_250; TAA41,160	C ; C	0.34004 ± 0.06531
8	VIC,380 ; GT03_180	C ; C	0.31067 ± 0.04873
	GT03_180; CMS4,180	C ; C	0.05412 ± 0.01486

continuación...

F_2Real150indv4Falt (modelo bayesiano con α= 0.005)			
grupo	marcadores contiguos	tipo	media ± desv. típ.
9	HLH_400 ; CMS20,180	C ; D1	0.16370 ± 0.04114
	CMS20,180; CR54,310	D1 ; D2	0.15925 ± 0.09908
10	CMS14,160; CR36_310	D1 ; C	0.13394 ± 0.0418
11	CR19,380; TAA1,175	C ; C	0.20267 ± 0.04255
12	TAA27,250; CNHX1,1636	C ; C	0.22088 ± 0.04467

Cuadro 12.4: *Estimación bayesiana de las distancias multipunto y desviaciones típicas posteriores entre marcadores contiguos, para los grupos de ligamiento de la población F_2 real determinados bajo $\alpha = 0.005$. Marcadores sin agrupar:* $C1iC8rt, 350$, $DREB\_480$, $CR5\_600$, $CR16\_250$, $C1iC8rt, 510$, $28R, 320$, $CR13, 480$, $CR12\_540$

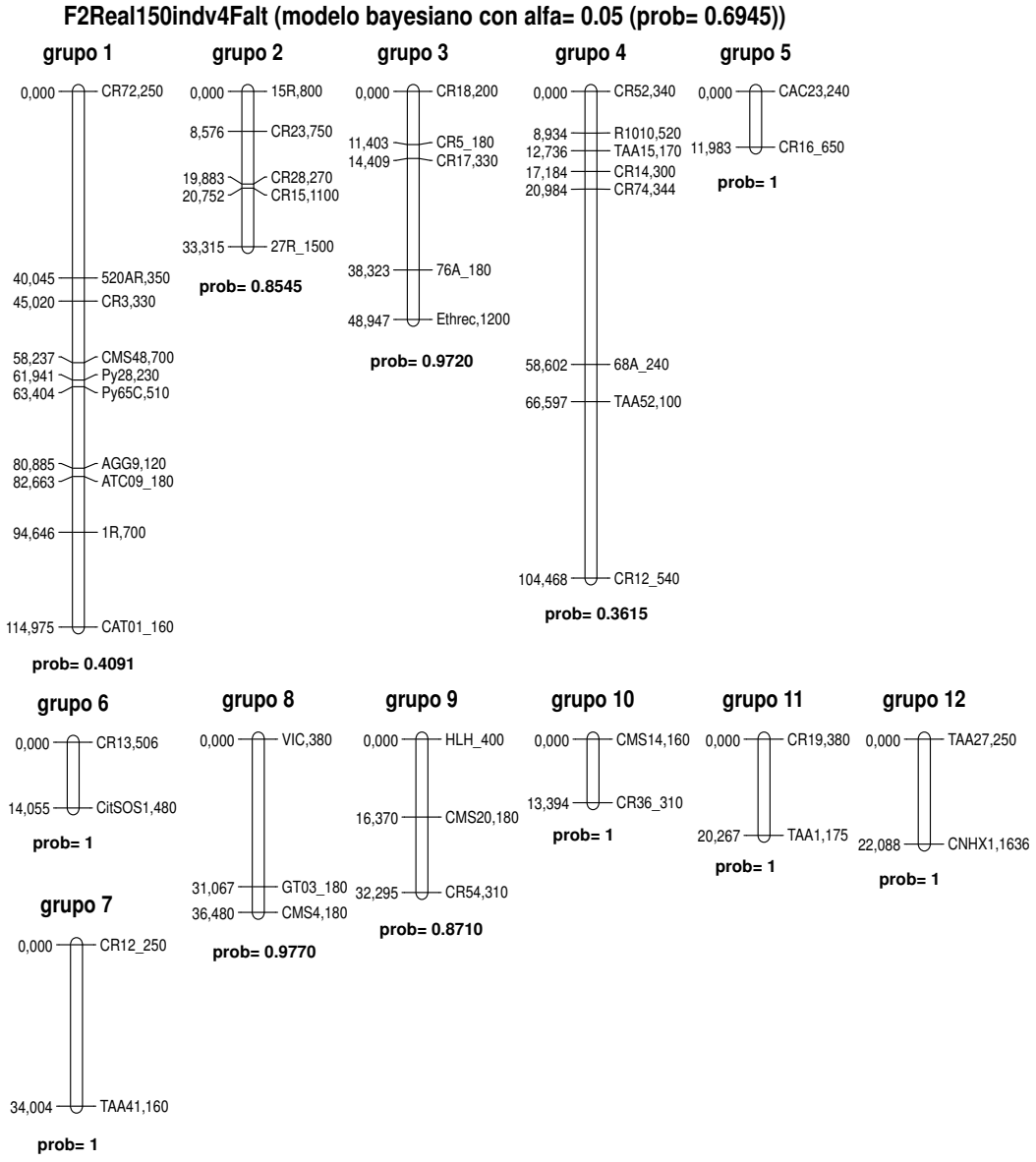


Figura 12.6: *Grupos de ligamiento y ordenaciones de los marcadores dentro de cada grupo de ligamiento obtenidos, según el modelo bayesiano, para el banco de datos* $F_2Real150indv4Falt$*, bajo* α*=0.05. Marcadores sin agrupar:* $C1iC8rt, 350$, $DREB\_480$, $CR5\_600$, $CR16\_250$, $C1iC8rt, 510$, $28R, 320$, $CR13, 480$

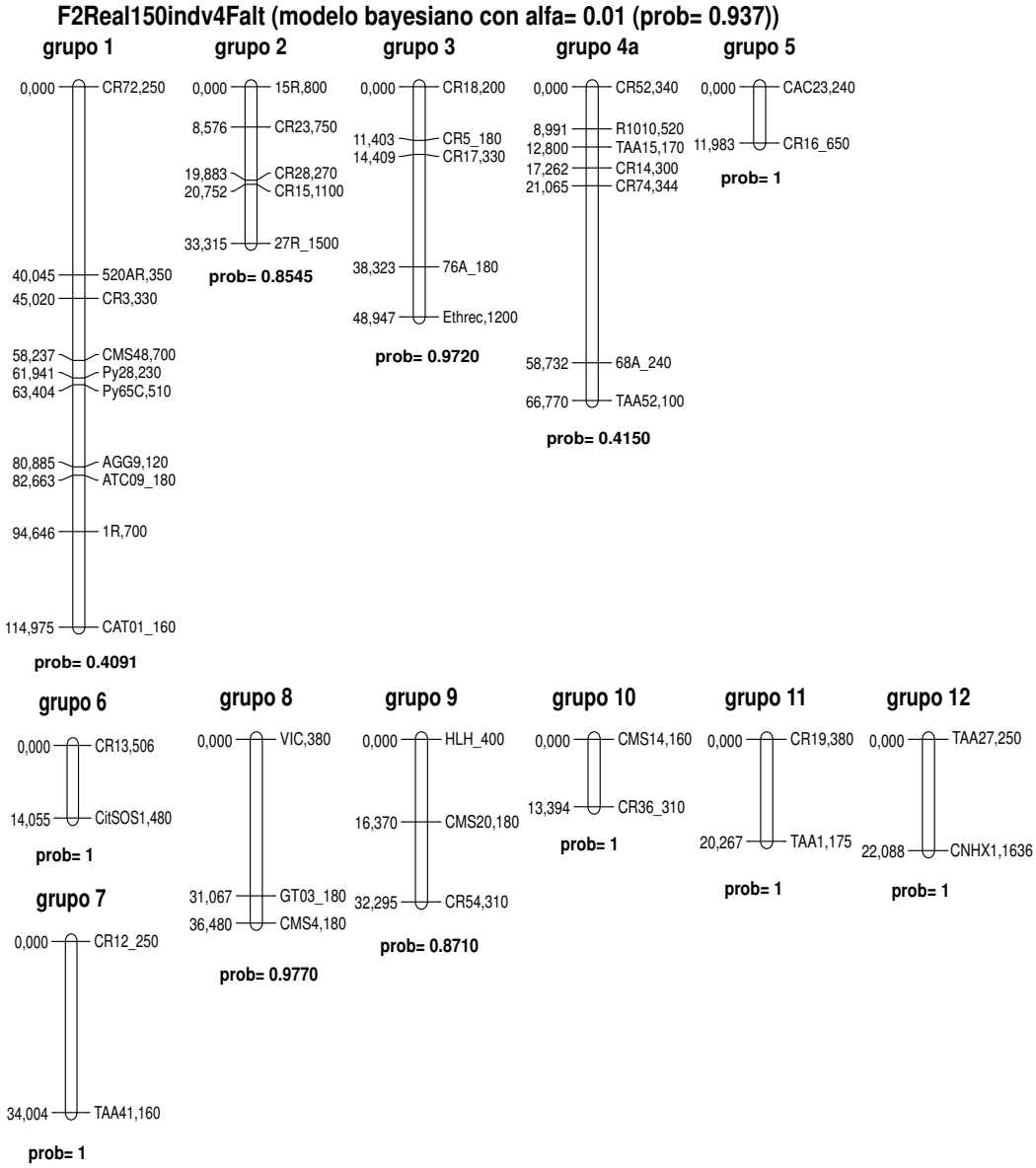


Figura 12.7: *Grupos de ligamiento y ordenaciones de los marcadores dentro de cada grupo de ligamiento obtenidos, según el modelo bayesiano, para el banco de datos F_2Real150indv4Falt, bajo α=0.01. Marcadores sin agrupar: $C1iC8rt, 350$, $DREB\_480$, $CR5\_600$, $CR16\_250$, $C1iC8rt, 510$, $28R, 320$, $CR13, 480$, $CR12\_540$*

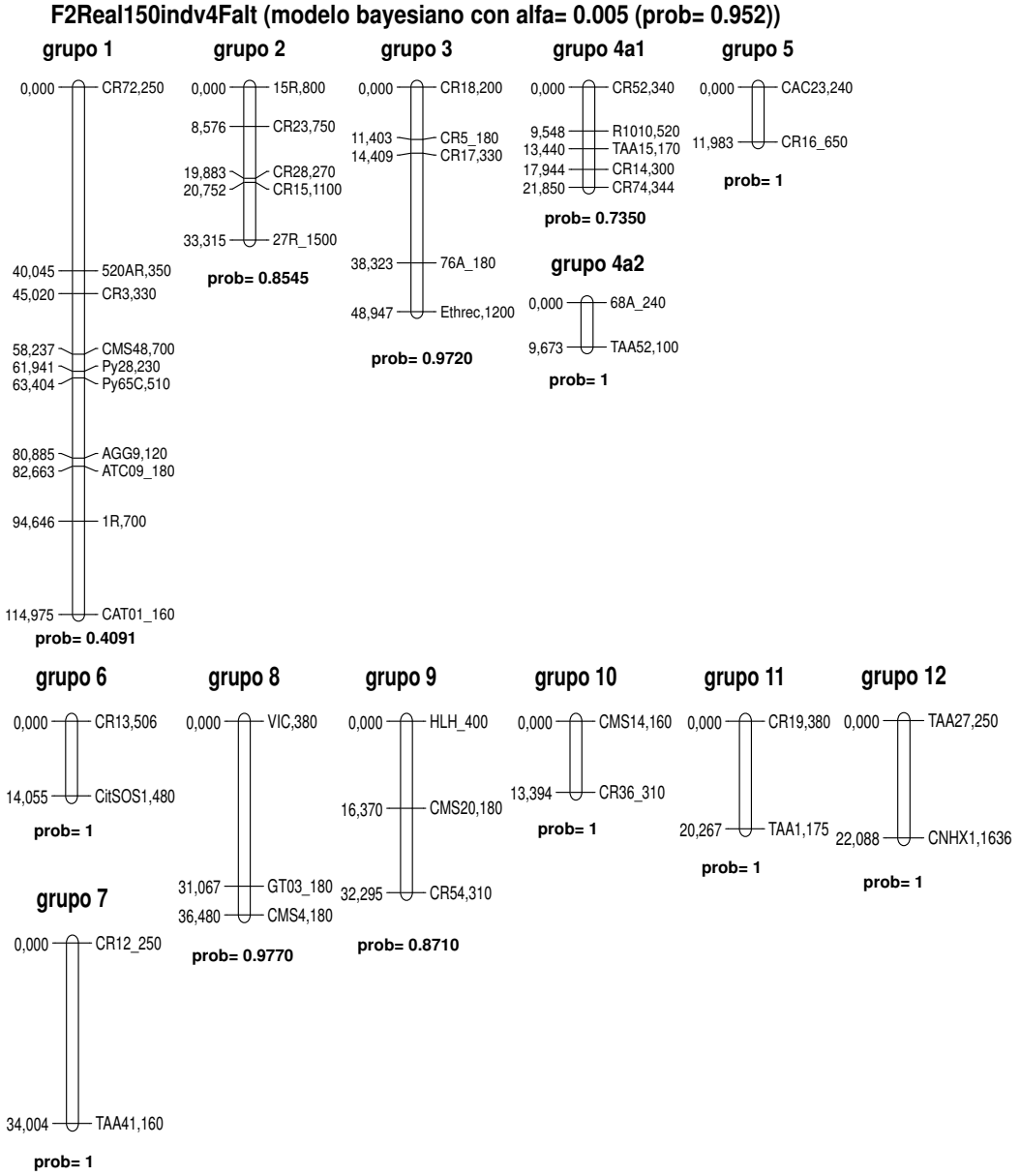


Figura 12.8: *Grupos de ligamiento y ordenaciones de los marcadores dentro de cada grupo de ligamiento obtenidos, según el modelo bayesiano, para el banco de datos F_2Real150indv4Falt, bajo α=0.005. Marcadores sin agrupar:* $C1iC8rt, 350$, $DREB\_480$, $CR5\_600$, $CR16\_250$, $C1iC8rt, 510$, $28R, 320$, $CR13, 480$, $CR12\_540$

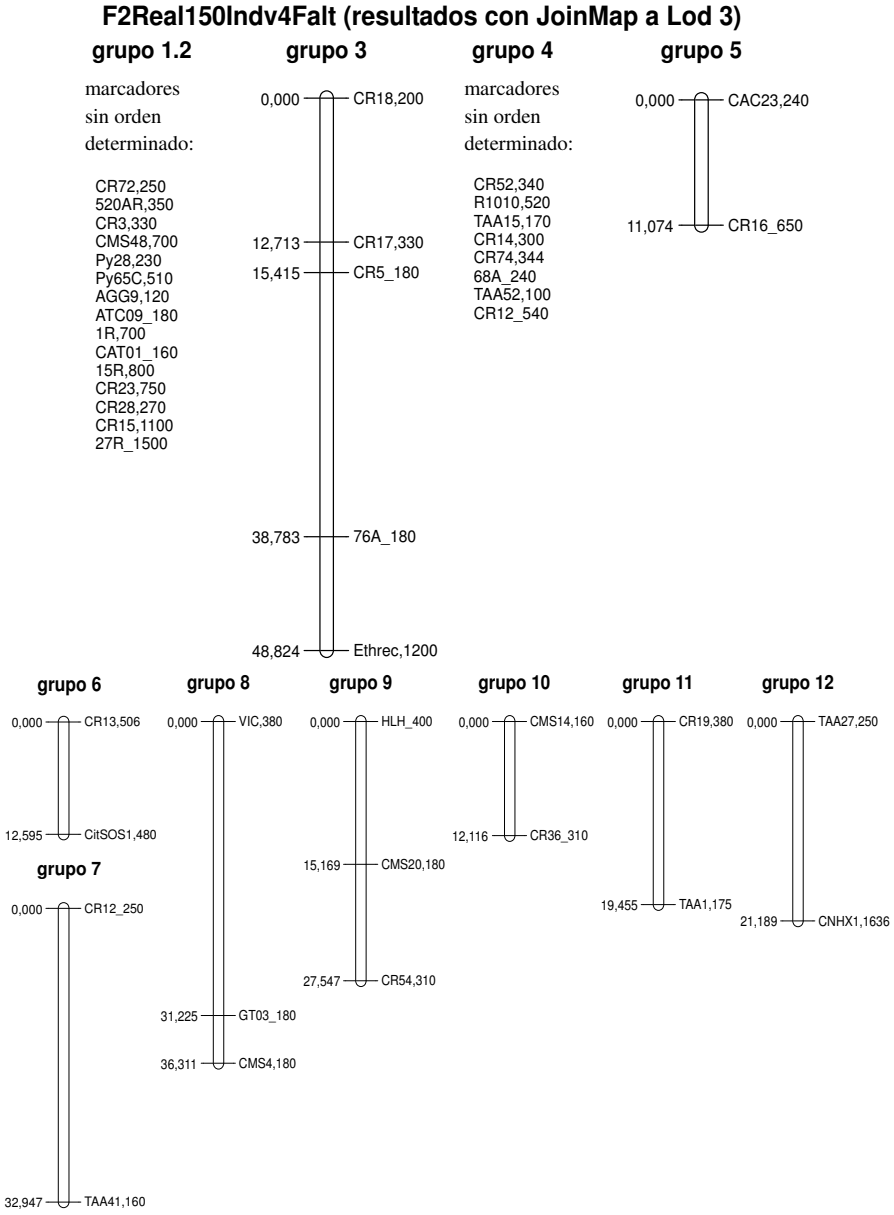


Figura 12.9: *Grupos de ligamiento y ordenaciones de los marcadores dentro de cada grupo de ligamiento obtenidos, según JoinMap, para el banco de datos F_2Real150indv4Falt, bajo LOD 3. Marcadores sin agrupar: $C1iC8rt, 350$, $DREB\_480$, $CR5\_600$, $CR16\_250$, $C1iC8rt, 510$, $28R, 320$, $CR13, 480$.*

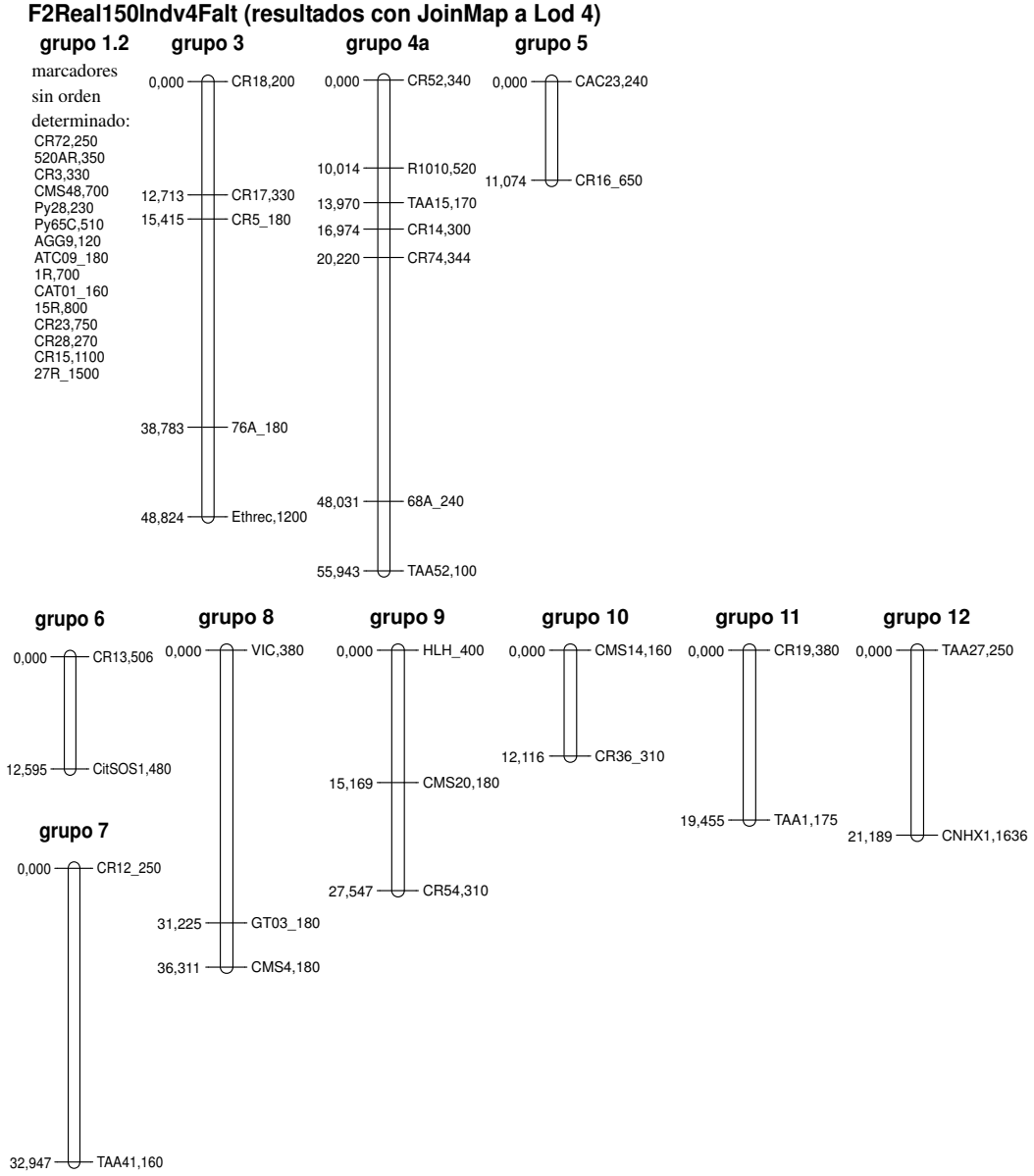


Figura 12.10: *Grupos de ligamiento y ordenaciones de los marcadores dentro de cada grupo de ligamiento obtenidos, según JoinMap, para el banco de datos F_2Real150indv4Falt, bajo LOD 4. Marcadores sin agrupar: $C1iC8rt, 350$, $DREB\_480$, $CR5\_600$, $CR16\_250$, $C1iC8rt, 510$, $28R, 320$, $CR13, 480$, $CR12\_540$*

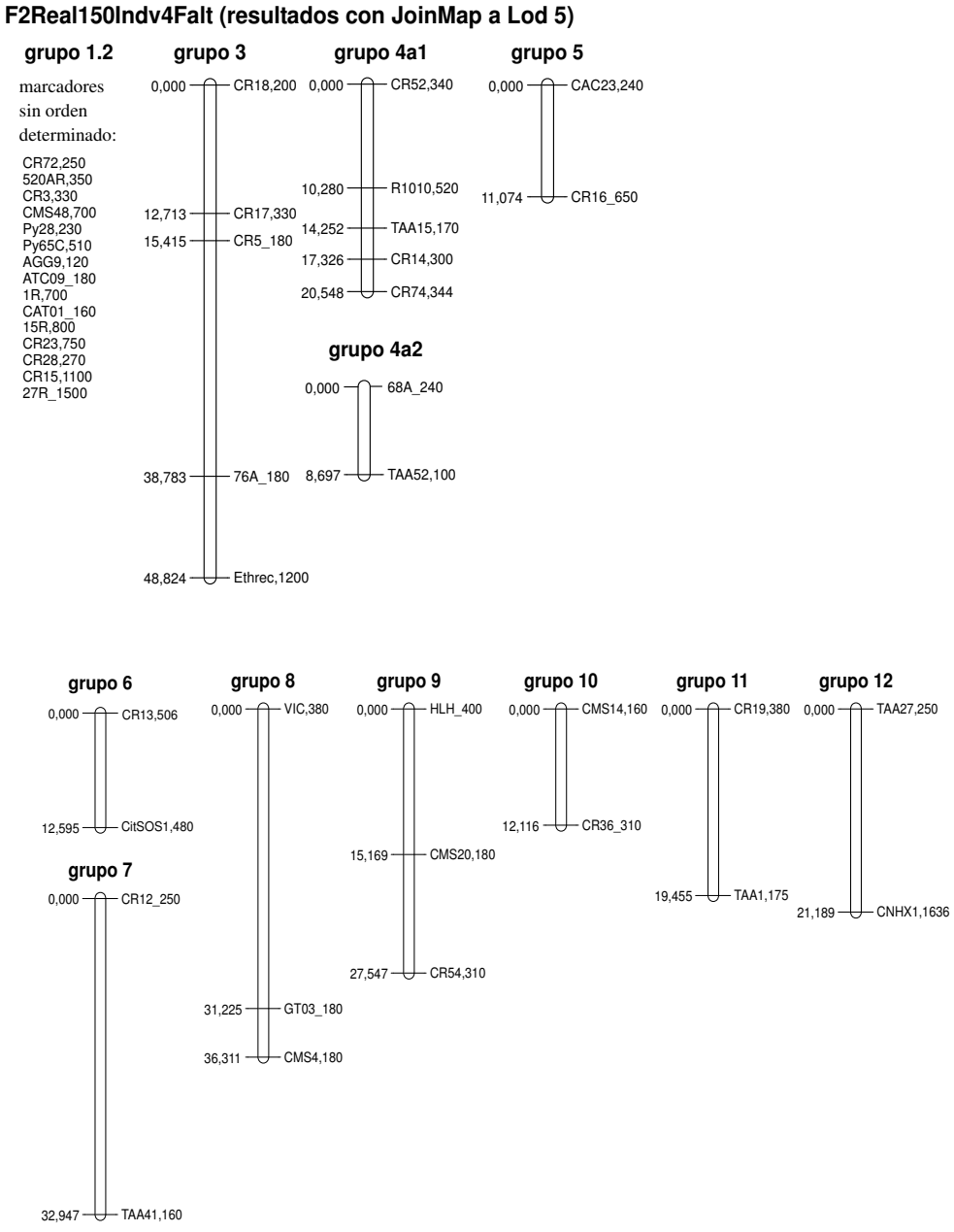


Figura 12.11: *Grupos de ligamiento y ordenaciones de los marcadores dentro de cada grupo de ligamiento obtenidos, según JoinMap, para el banco de datos F_2Real150indv4Falt, bajo LOD 5. Marcadores sin agrupar: $C1iC8rt, 350$, $DREB\_480$, $CR5\_600$, $CR16\_250$, $C1iC8rt, 510$, $28R, 320$, $CR13, 480$, $CR12\_540$*

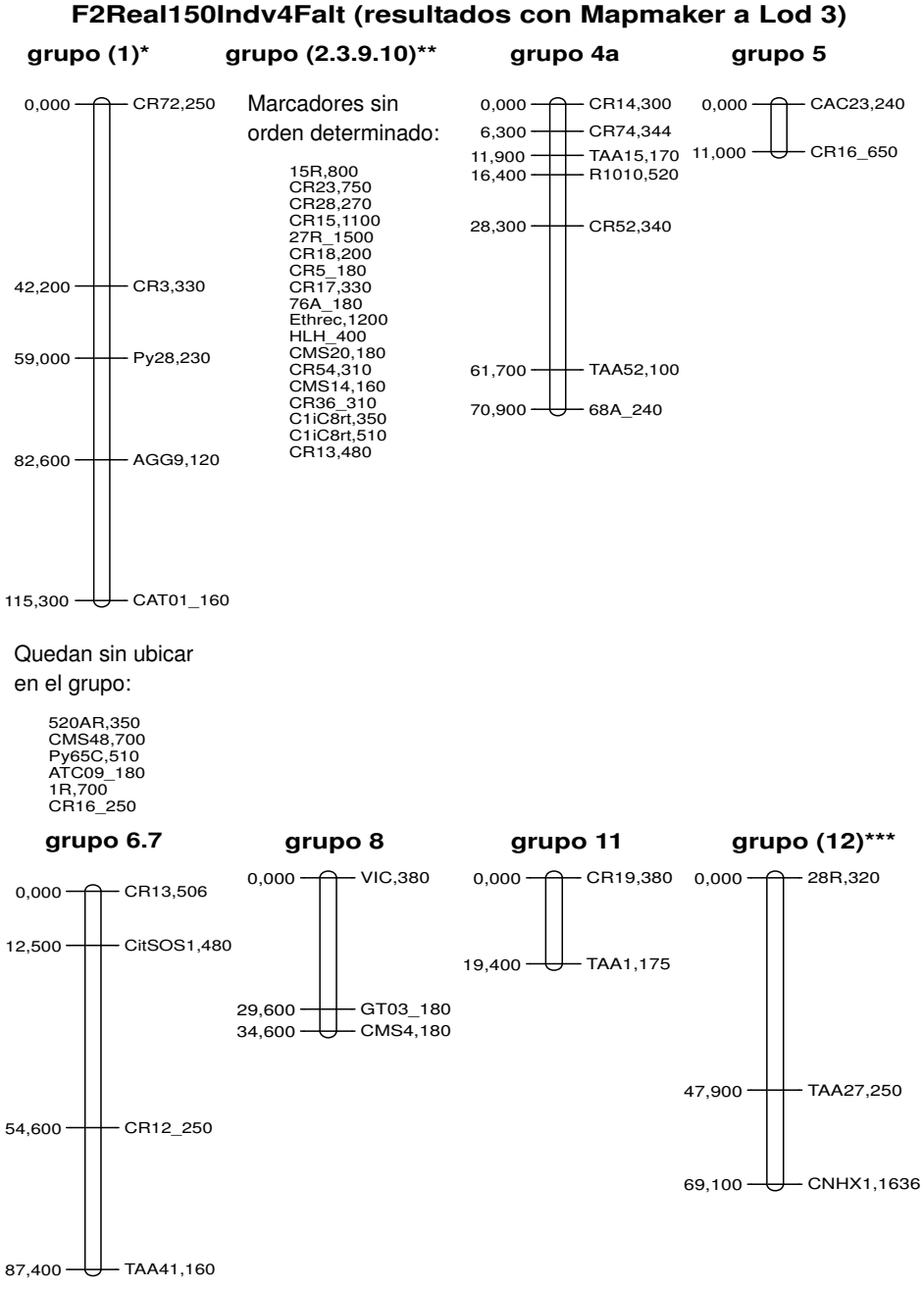


Figura 12.12: *Grupos de ligamiento y ordenaciones de los marcadores dentro de cada grupo de ligamiento obtenidos, según Mapmaker, para el banco de datos F_2Real150indv4Falt, bajo LOD 3. Marcadores sin agrupar: $DREB\_480$, $CR5\_600$, $CR12\_540$*

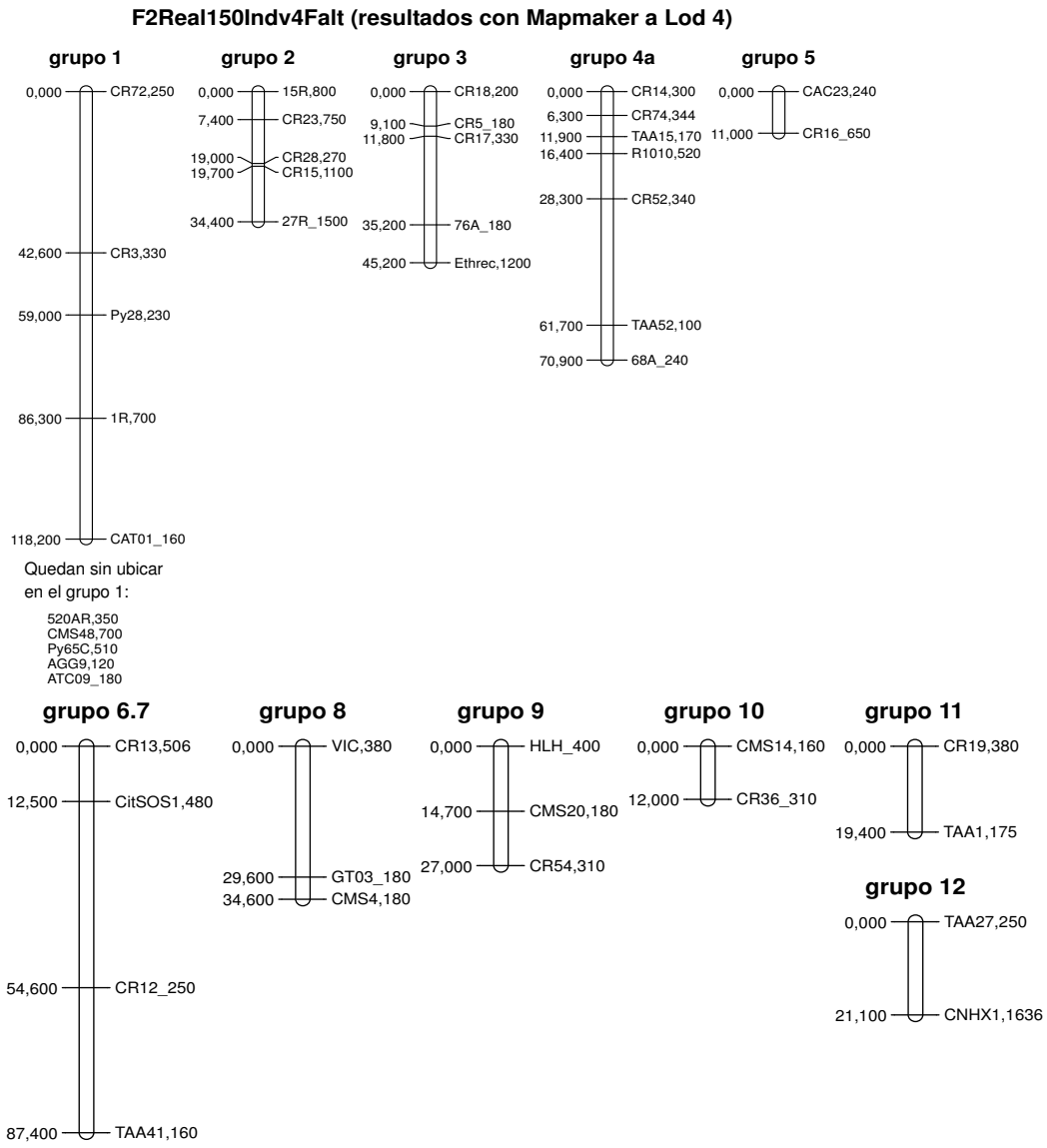


Figura 12.13: *Grupos de ligamiento y ordenaciones de los marcadores dentro de cada grupo de ligamiento obtenidos, según Mapmaker, para el banco de datos F_2Real150indv4Falt, bajo LOD 4. Marcadores sin agrupar: $C1iC8rt, 350$, $DREB\_480$, $CR5\_600$, $CR16\_250$, $C1iC8rt, 510$, $28R, 320$, $CR13, 480$, $CR12\_540$*

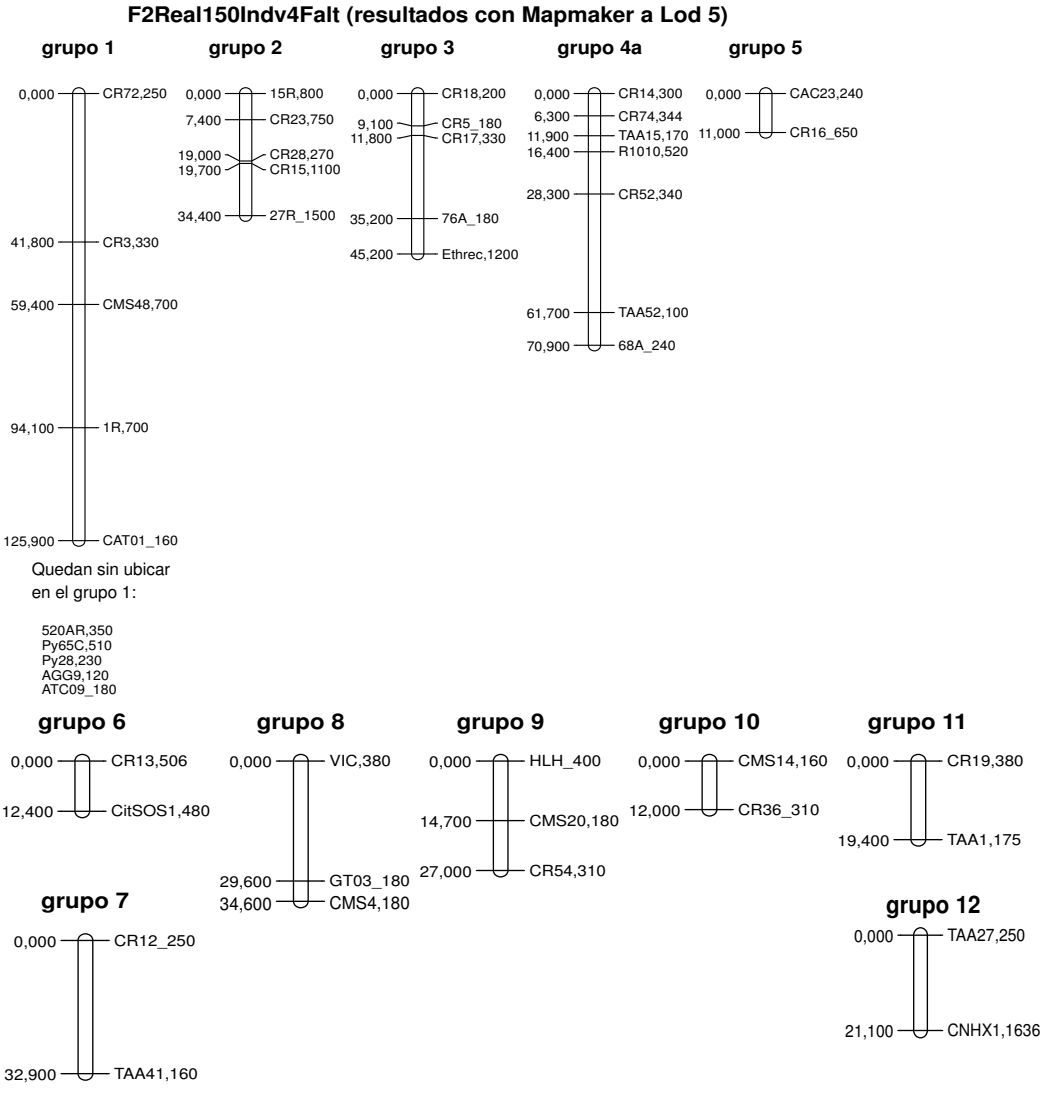


Figura 12.14: *Grupos de ligamiento y ordenaciones de los marcadores dentro de cada grupo de ligamiento obtenidos, según Mapmaker, para el banco de datos F_2Real150indv4Falt, bajo LOD 5. Marcadores sin agrupar:* $C1iC8rt, 350$, $DREB\_480$, $CR5\_600$, $CR16\_250$, $C1iC8rt, 510$, $28R, 320$, $CR13, 480$, $CR12\_540$

12.3. Discusión

La estimación del mapa genético de la población Retrocruce está basada en 52 marcadores. 50 de ellos, quedan distribuidos en 9 grupos de ligamiento, independientemente del α exigido en la resolución de los múltiples test de independencia. Se compara con la estimación obtenida por Bernet et al (2010) [7], mediante JoinMap 3.0 [86], de las poblaciones F x Ch y Ch x F, en que 77 marcadores quedan distribuidos en 11 grupos de ligamiento (Gr1, Gr2, etc.) y 7 marcadores quedan sin agrupar.

Tres de los 4 marcadores del grupo 1 mantienen el mismo orden en Gr10 (LOD 8), que consta de 7 marcadores. En su estudio falta $CTVCH28$ o quizás sea uno de los marcadores no agrupados. Este marcador fue incluido en un mapa posterior (Raga et al. 2012 [63]) El grupo 3, que consta de 8 marcadores se desglosa en Gr7a (LOD 10) y Gr7b (LOD 10). Además, Gr7a intercala 3 marcadores más, que no aparecen en nuestra muestra. Quizás, por ese motivo, el marcador 5F-5R,220 se ubica en una posición diferente a la que estima el modelo bayesiano. El grupo 4, consta de 10 marcadores, 9 de los cuales están contenidos entre los 15 marcadores de Gr3 (LOD 8) y mantienen el mismo orden a excepción de la permutación de $AintBint810$ y $TAA27$. Posiblemente interfieren en este orden, 4 marcadores de los cuales nosotros no disponemos en nuestra muestra. El décimo marcador, $CTVCH10$, del grupo 4, no aparece ubicado en ningún Gr, pues se vio posteriormente que tenía fallos en la codificación (Asins, comunicación personal). El grupo 5 está formado por 8 marcadores, 7 de los cuales pertenecen a Gr4b (LOD 5), que contiene 9 marcadores. $S2AS2$ no está ordenado en la misma posición que en el modelo bayesiano. Sin embargo, los otros 6 marcadores sí. $CMS24$ no aparece en ningún Gr y en contraposición nosotros no disponemos de dos de los marcadores que contiene Gr4b. Los 3 marcadores del grupo 6 están contenidos entre los 5 marcadores de Gr9b (LOD 6). En la ubicación de $CR52$, posiblemente interfieren dos marcadores de los que no disponemos en nuestra muestra. El grupo 7 consta de 6 marcadores, 5 de los cuales están incluidos, en el mismo orden, entre los 8 marcadores que forman Gr12 (LOD 5). $CTVCH20$ no está incluido en ningún Gr. Se puede observar, en el grupo 7, que entre $CR22,1018$ y $CTVCH20$, existe un espacio desierto de marcadores. Sin embargo, este espacio está ocupado por 3 marcadores en Gr12 (LOD 5) de los que no tenemos información en nuestra muestra. El grupo 8 está formado por 3 marcadores

incluidos, en el mismo orden, entre los 9 marcadores de Gr4c. De los otros 6 marcadores no tenemos información. Cinco de los 6 marcadores del grupo 9, guardan el mismo orden en Gr4a (LOD 9). El marcador CTVCH17 no se encuentra en ningún Gr y entre VIC y $CR16,450$ existen 3 marcadores en Gr4a (LOD 9), de los que no disponemos de información. También existe un marcador, desconocido para nosotros, entre $CMS4$ y $GT03,171$. Finalmente, señalar que los marcadores $5R-4R,850$ y $AintDrt900$ quedan sin agrupar en el modelo bayesiano. Sin embargo, $5R-4R,850$ se agrupa con otro marcador, que no aparece en nuestra muestra, para formar Gr5b (LOD 10).

Por otra parte, la estimación bayesiana del mapa genético de la población F_2 se compara con la estimación propuesta por Raga et al (2012) [63], mediante JoinMap 3.0 [86], para la población $F_2(R$ x $P_r)$, Su investigación, según especifican, está basada en 54 marcadores, de los cuales 11 no obtienen agrupación y el resto se distribuyen en 11 grupos de ligamiento a LOD$\geq$ 3. La nomenclatura utilizada para nombrar los marcadores ha sido la misma que la propuesta por Bernet et al (2010) [7].

Nuestra estimación del mapa genético consta de 53 marcadores e incluyen dos marcadores nuevos ($DREB\_480$ y LH_400) que no se estudiaron en el trabajo de Raga et al. (2012) [63]. Quedan sin agrupar los marcadores $C1iC8rt,350$; $DREB\_480$; $CR5\_600$; $CR16\_250$; $C1iC8rt,510$; $28R,320$; $CR13,480$ con $\alpha = 0.05$ y adicionalmente el marcador $CR12\_540$ con $\alpha = 0.01$ y 0.005. Los 45 marcadores restantes, se distribuyen en 12 ($\alpha = 0.05$ y 0.01) y 13 ($\alpha = 0.005$) grupos de ligamiento. Considerando la estructura de grupos de ligamiento según $\alpha = 0.005$ con probabilidad= 0.952, se observa la siguiente coincidencia con el estudio de Raga et al (2012) [63], cuyos grupos se nombran RPr:

El grupo 1 coincide con RPr4b (LOD 7), a excepción de la ubicación del marcador CR72,250, que la metodología bayesiana lo ubica en un extremo del grupo de ligamiento y Raga et al. (2012) [63] lo intercala entre Py65C,506 y AGG9,125. El grupo 3 coincide con RPr7a (LOD 3) a excepción de la permutación de los marcadores CR5,180 y CR17,300. Además, incluyen en un extremo el marcador CR18,180, que no aparece en nuestra muestra. El grupo 4a1 y 4a2 contienen los marcadores de RPr8+6, ordenados del mismo modo. El grupo 9 consta de los marcadores HLH_400, $CMS20,180$ y $CR54,310$. Quizás el marcador HLH_400, que no aparece en el estudio de Raga et al.

(2012) [63], es el necesario para formar el grupo RPr12. Parece que JoinMap 3.0 [86], en ausencia de HLH_400, no es capaz de obtener, a ningún LOD, el mapa conjunto de R12 con Pr12 (llamémoslo RPr12), a los que pertenecen $CMS20,180$ y $CR54,310$, respectivamente. El resto de grupos, es decir 2, 5, 6, 7, 8, 10, 11 y 12 se corresponden con RPr4c (LOD 10), RPr7b (LOD 10), RPr10+5b(b) (LOD 6), RPr10+5b(a) (LOD 7), RPr4a (LOD 8), RPr9 (LOD 9), RPr2 (LOD 8) y RPr3b (LOD 4), respectivamente.

Tanto en el estudio de Raga et al. (2012) [63] como en el de Bernet et al. (2010) [7] se ha utilizado Kosambi como función de mapeo para elaborar las estimaciones de los mapas genéticos, por lo que las longitudes de los grupos de ligamiento no son comparables con las obtenidas por el modelo bayesiano, en las que se ha utilizado Haldane, como siempre.

Salvando las diferencias, que hemos observado, entre los marcadores que aparecen en las muestras utilizadas en los tres estudios (Raga et al. (2012) [63], Bernet et al. (2010) [7] y metodología bayesiana) las estimaciones obtenidas por la metodología bayesiana tienen alta concordancia con las obtenidas de forma frecuentista. Además tienen un comportamiento muy estable con respecto al α utilizado. Esta es una característica que parece propia de la metodología bayesiana presentada en esta tesis frente a la variación de resultados al modificar el LOD de los métodos frecuentistas. Por otra parte, quizás la explicación de que algunos marcadores se queden sistemáticamente sin ubicar en un grupo sea que perteneciendo a un grupo de ligamiento, existe una zona sin marcadores estudiados necesarios para formar estos grupos de ligamiento. Llama la atención que los LODs necesarios para formar los grupos de ligamiento en ambos estudios frecuentistas son, en general, muy superiores a los que han sido necesarios en los Capítulos 10 y 11 para obtener estimaciones correctas, donde se ha trabajado con un mayor número de marcadores y de grupos de ligamiento y posiblemente sea debido al conocimiento previo de la realidad biológica por parte del investigador tal como se apuntaba al final del apartado 1.4.2.

Capítulo 13

Resumen general, futuras líneas de trabajo y conclusiones.

13.1. Resumen general

Como ya se adelantó en el último apartado de la introducción, para llevar a cabo un análisis de QTLs es muy importante una estimación precisa y fiable del mapa genético de la población involucrada. La localización de QTLs y los efectos que producen, se pueden inferir combinando la información de los genotipos y los fenotipos de individuos que provienen de poblaciones en desequilibrio, tales como la de los diseños experimentales controlados, Retrocruce y F_2. Por ese motivo, a lo largo de los capítulos anteriores, se han diseñado y ensayado distintas metodologías bayesianas con el objetivo de estimar el mapa genético en tres escenarios experimentales: poblaciones Retrocruce, poblaciones F_2 con marcadores exclusivamente codominantes y poblaciones F_2 en las que intervienen marcadores codominantes y dominantes conjuntamente.

Nótese, que durante toda la investigación, se ha asumido que los datos han sido previamente depurados mediante un análisis inicial exploratorio. Como es sabido, este es un paso fundamental en cualquier análisis estadístico. Consideramos que esta fase está perfectamente cubierta por algunas de las librerías de R/qtl [13].

En los Capítulos 2 y 3, se elaboran modelizaciones bayesianas equivalentes, para las tres poblaciones de estudio, de la distribución posterior de las fracciones de recombinación entre cada pareja de marcadores y, en base a

ella, se obtiene una cadena de simulación utilizando el algoritmo Metropolis-Hastings. En cada iteración de la cadena, se aplica un algoritmo de ordenación de los marcadores basado en las distancias mínimas entre ellos, SARF. Una vez obtenido el orden de los marcadores, de nuevo en cada iteración, se estima las distancias multipunto entre los marcadores contiguos, según el método de mínimos cuadrados. Tras este proceso, se obtienen distintas estimaciones alternativas del mapa genético (modelos), valoradas por una probabilidad que cuantifica la proporción de veces que se ha obtenido cada modelo. Las distancias entre marcadores contiguos se calculan como media posterior de la distribución de distancias multipunto. Esta metodología bayesiana obtiene resultados favorables tanto en una población Retrocruce como en una población F_2 con todos los marcadores codominantes para tamaños muestrales de al menos 100 individuos. En ambos casos, se detectan problemas de estimación en el caso poco real, para algunos cultivos, de trabajar con muestras de tan sólo 50 individuos. Sin embargo, la metodología no proporciona resultados satisfactorios en el caso en que no son observables todos los genotipos de todos los marcadores. Es decir, en una población F_2 definida por marcadores codominantes y dominantes conjuntamente, ya que en ocasiones, aun trabajando con muestras de 200 individuos, la estimación del mapa genético verdadero no aparece entre los modelos más probables y cuando la metodología se aplica sobre una población densa, los problemas de ubicación de los marcadores se manifiestan por bloques.

Con el objetivo de investigar exhaustivamente el problema de la estimación de las fracciones de recombinación para marcadores fuertemente ligados en poblaciones F_2 con no todos los marcadores codominantes, en los Capítulos 4, 5 y 6 se reduce la población a la mínima expresión que involucra establecer un orden entre sus marcadores (3 marcadores) y se diseña un modelo jerárquico bayesiano de tres niveles, cuya distribución posterior se obtiene simulando, mediante un algoritmo Metropolis-Hastings generalizado, de una mixtura entre el cálculo de la probabilidad del orden de la tripleta dadas las frecuencias genotípicas observadas entre los marcadores y la distribución posterior de las fracciones de recombinación, seleccionado el orden de la tripleta de marcadores. En los tres capítulos se proponen distintas variantes sobre la elección de la distribución pivote a utilizar: una distribución normal truncada en $[0, 0.5]$, una distribución normal bivariada truncada en $[0, 0.5]^2$ o la propia logverosimilitud conjunta de las fracciones de recombinación contiguas. En los tres capítulos se

detectan problemas técnicos de convergencia, provocados por distintos motivos: En el Capítulo 5, se observan problemas de estimación de los parámetros de la distribución binormal que estima la verosimilitud conjunta referente al Orden 2. Problemas en la estimación equiprobable de las posteriores de los Ordenes 1 y 3. En el Capítulo 6, los problemas en la estimación se acentúan en el caso de la combinación de marcadores D1D2. Como ya se ha señalado, para este tipo de pareja, el único genotipo inequívocamente distinguible tiene una frecuencia esperada asociada de $0.25r^2$ y por lo tanto, cuando la distancia entre los dos marcadores es pequeña, la frecuencia esperada es próxima a cero. Se deduce que es necesario un tamaño de muestra elevado para que la frecuencia observada, en ese caso, sea distinta de cero. Es decir, con tamaños de muestra razonables como son 100 o 200 individuos es muy probable que no se observe ningún individuo con ese genotipo. Esto influye negativamente en el cálculo de la verosimilitud y en el desarrollo de la metodología planteada. Además, aunque la metodología hubiera proporcionado resultados satisfactorios, existe el problema añadido de la generalización del proceso a un mapa genético con más de 3 marcadores, ya que uno de los puntos claves de la metodología supone calcular probabilidades mediante integración numérica por cuadratura gaussiana. Aunque los resultados no han sido satisfactorios, tras el estudio realizado en estos tres capítulos, se concluye desde el punto de vista práctico, que quizás se obtendrían mejores resultados si se estimara un mapa genético preliminar, eliminando alguno de los dos marcadores de la pareja estrechamente ligada D1D2 o D2D1 y se diseñara una estrategia para su posterior incorporación.

Sobre esta última idea se trabaja en el Capítulo 7. Se elabora una metodología bayesiana equivalente a la vista en el Capítulo 3 pero esta vez, se simula de la distribución posterior de las fracciones de recombinación para cada pareja de marcadores a través del programa WinBugs [51], optimizando la computación y convergencia. Se diseña un algoritmo de ordenación entre marcadores acorde con los conocimientos adquiridos en los capítulos anteriores. Los resultados son tímidamente mejores que los obtenidos hasta ahora pero no tan satisfactorios como cabía esperar.

En el Capítulo 8 y 9 se continúa simulando de la distribución posterior de las fracciones de recombinación para cada pareja de marcadores a través del programa OpenBugs [75] en los mismos términos que en el capítulo anterior. Para la ordenación de los marcadores, se cambia de estrategia y se ensaya

un algoritmo mucho más exhaustivo y laborioso que los estudiados hasta el momento, ya que la estimación del mapa genético se inicia con una pareja de marcadores y prosigue incluyendo el resto de marcadores uno a uno, valorando su ubicación en el mapa entre todas las posibles. Además, tras la obtención del mapa completo, se proponen todas las posibles permutaciones entre cuatripletas de marcadores contiguos hasta obtener la estimación definitiva del mapa genético. Bajo esta metodología se obtienen resultados satisfactorios incluso en las situaciones que resultaban problemáticas anteriormente. Tras la distribución en el muestreo y la comparativa con los programas de referencia JoinMap [86] y Mapmaker [48], se concluye que la metodología bayesiana prácticamente iguala o mejora los resultados de la mejor de las metodologías frecuentistas en las tres poblaciones estudiadas. Es decir, sobre el conjunto de muestras utilizadas alcanza un porcentaje de aciertos, al estimar el mapa genético de la población, superior al de los programas basados en métodos frecuentistas. Comparando los resultados entre los diseños Retrocruce y F_2 con marcadores dominantes, se puede concluir que estos primeros obtienen modelos con probabilidades asociadas un poco mejores. Además consiguen probabilidades más alejadas entre los primeros mejores modelos y entre los segundos mejores modelos. En definitiva, parece que el diseño Retrocruce estima con más precisión el orden del mapa que el diseño F_2, con marcadores dominantes. Una posible explicación sería que, si bien el número de meiosis informativas es menor en un Retrocruce , ya que uno de los parentales no aporta información sobre la recombinación, parece que el hecho de que las estimas por Retrocruce provengan de una fórmula explícita permiten obtener modelos más fiables. Es importante señalar que se han obtenido resultados razonables incluso en el caso en que el tamaño de la muestra es muy pequeño (50 individuos), ya que en este tipo de experimentos genéticos se suelen emplear tamaños de muestra superiores a 100 individuos. La motivación del estudio sobre un tamaño de muestra tan pequeño es investigar cómo afecta sobre la fiabilidad del mejor modelo obtenido, ya que en el caso de la obtención de buenos resultados podría suponer una ahorro económico importante. Este es el caso de cultivos frutales y arbóreos, que por limitaciones de espacio y tiempo hasta la producción, es normal encontrar muestras de tamaño pequeño. En estos casos, la metodología desarrollada es claramente satisfactoria.

Considerada adecuada la metodología bayesiana propuesta para simular fracciones de recombinación entre parejas de marcadores y el algoritmo de

ordenación de marcadores para la estimación del mapa genético de la población, en el Capítulo 10 se amplía la metodología para cubrir un paso previo al ya investigado. Se trata de trabajar en un entorno más realista, en el que la población está definida por distintos grupos de ligamiento o cromosomas y es necesario determinar el reparto de marcadores en cada uno de los grupos de ligamiento para, posteriormente, estimar el mapa genético de la población. El algoritmo implementado valora, para cada pareja de marcadores, la condición de asociación y la de proximidad, que supone un ejercicio bayesiano de comparaciones múltiples, cuya resolución involucra al nivel de significatividad global empleado. De nuevo, se ensaya la metodología sobre nuevas poblaciones Retrocruce y F_2, prediseñadas para recrear todo tipo de dificultades. Los resultados obtenidos son satisfactorios ya que, en los tres escenarios de estudio habituales, la metodología es capaz de agrupar los marcadores de forma correcta en cada cromosoma. Se observa un comportamiento más estable, según se hace variar el nivel de significatividad global empleado, que el obtenido por los programas JoinMap [86] y Mapmaker [48], según varían el LOD considerado.

En el Capítulo 11, se trabaja en el entorno más desfavorable de todos los planteados con anterioridad, es decir una población F_2 con marcadores codominantes y dominantes definida por 20 grupos de ligamiento. En este escenario, se prueba la metodología completa sobre muestras con distintos porcentajes de datos faltantes. Tras los resultados, se deduce que la metodología bayesiana para la determinación de grupos de ligamiento y para la ordenación de marcadores dentro de cada grupo de ligamiento son satisfactorias y obtienen resultados razonables equiparables o incluso mejores a los que obtienen los programas de referencia JoinMap [86] y Mapmaker [48]. Comparando con los dos programas de referencia, la metodología bayesiana tiene un comportamiento similar al programa que mejor reparte los marcadores en los grupos de ligamiento (Mapmaker [48]) y después produce un comportamiento similar al programa que mejor ordena los marcadores dentro de cada grupo de ligamiento, ya sea JoinMap [86] o Mapmaker [48].

En el Capítulo 12, se prueba la metodología completa sobre datos que provienen de dos poblaciones reales que representan el cultivo de cítricos. Concretamente, una de las muestras (de 201 individuos) ha sido extraída de una población con 52 loci/ marcadores, segregando como en un Retrocruce. La otra muestra (de 150 individuos) proviene de una población F_2 con 53 marcadores, algunos codominantes y otros dominantes. En este caso, el desconocimiento

de los mapas genéticos reales de las poblaciones impide la valoración exacta de los resultados. Sin embargo, se ofrece una comparativa de los resultados bayesianos y los obtenidos por los programas JoinMap [86] y Mapmaker [48], en igualdad de condiciones con respecto a los datos empleados. Mientras la metodología bayesiana cuantifica mediante probabilidades las distintas estructuras de grupos de ligamiento y el orden de los marcadores en cada grupo de ligamiento, los programas comerciales, en algunos grupos de ligamiento, no son capaces de obtener el orden específico de los marcadores. Lógicamente, si se eliminaran algunos marcadores conflictivos podría obtenerse algún tipo de ordenación. Esta práctica dependerá del conocimiento del investigador. Sin embargo, en nuestro caso, hemos preferido ser totalmente neutrales y someter los distintos enfoques a un mismo tratamiento. Además, se han comparado los resultados bayesianos con dos estudios publicados, (Raga et al. (2012) [63] y Bernet et al. (2010) [7]) realizados mediante metodología frecuentista. De la comparación se concluye una alta concordancia y un comportamiento muy estable con respecto al α utilizado en los múltiples test de independencia de la fase de elaboración de los grupos de ligamiento.

13.2. Futuras líneas de investigación

Las líneas inmediatas de futura investigación o desarrollo serían:

1. Desarrollar una estrategia bayesiana que permita imputar información faltante, especialmente para combinaciones de marcadores que tienen genotipos observables cuyas frecuencias esperadas asociadas tienden a cero; por ejemplo, cuando los marcadores implicados están estrechamente ligados.

2. Probar otros algoritmos optimizados, equivalentes a los proporcionados por WinBugs u OpenBugs, como Jags (Just Another Gibbs Sampler), Stan o NIMBLE para la simulación de la distribución posterior de las fracciones de recombinación entre parejas de marcadores.

3. Investigar distintas poblaciones definidas por distinto número de marcadores, para establecer recomendaciones sobre cuál sería el nivel de significatividad global recomendable a emplear en la resolución de las comparaciones múltiples, en la fase de agrupación de marcadores, dependiendo del número de marcadores implicados en el mapa genético, el tamaño muestral disponible y de la población que se desee estimar. Investigar la ampliación a un mayor número de marcadores segregando por grupos de ligamiento.

4. Generalizar la metodología completa para otros diseños experimentales como el de Líneas Recombinantes Puras (RILs), ya que este tipo de diseños es muy adecuado para el análisis de QTLs y diseños "Cross Pollinated" de especies alógamas como son las forestales y los frutales.

5. Implementar la metodología completa a modo de aplicación on-line para su difusión en abierto.

6. Adecuar la metodología al caso en que existe distorsión en la segregación, que como en el caso de los cítricos, suele ocurrir frecuentemente. Los resultados obtenidos en el Capítulo 12 son muy prometedores en este sentido.

13.3. Conclusiones

Las conclusiones más destacadas tras el proceso de investigación son:

C1. En el diseño Retrocruce, la metodología bayesiana de ordenación basada exclusivamente en las distancias mínimas entre sus marcadores (SARF), es satisfactoria especialmente con muestras de al menos 100 individuos. La reducción del tamaño muestral, repercute en una reducción de la probabilidad media del modelo bayesiano más probable y los marcadores más cercanos aumentan las probabilidades de ser ubicados en posiciones alternativas relativamente próximas.

C2. En el diseño F_2 con todos los marcadores codominantes, parece que la metodología bayesiana de ordenación basada exclusivamente en las distancias mínimas entre sus marcadores (SARF) es satisfactoria y además estima mejor el mapa genético de la población que si la muestra procede de un diseño Retrocruce. Sin embargo, en el diseño F_2 con marcadores dominantes y codominantes, dicha metodología no es satisfactoria especialmente en el caso en que el mapa genético de la población es denso. Se observa un problema de falta de información para estimar fracciones de recombinación de parejas de marcadores dominantes en fase de repulsión.

C3. La modelización con mixturas para tripletas de marcadores, no logra satisfacer los problemas detectados con la anterior metodología, bajo ninguna de sus tres variantes. Además, se añade el problema de la generalización de la metodología a mapas genéticos con más de tres marcadores.

C4. La estrategia de estimación del mapa genético de la población basada en un mapa preliminar de marcadores codominantes y la posterior incorporación de los marcadores dominantes, mejora tímidamente las anteriores

propuestas pero no obtiene resultados tan satisfactorios como cabía esperar.

C5. La metodología basada en la información de todos los marcadores, obtiene resultados óptimos como consecuencia de un algoritmo de ordenación más exhaustivo y laborioso que los implementados hasta el momento, que tiende a incorpora, en primer lugar aunque no estrictamente, marcadores codominantes. La metodología es capaz de filtrar aquellos órdenes que son improbables o inconsistentes.

C6. Tras la distribución en el muestreo con la metodología definitiva, se deduce que las muestras que proceden de una población F_2 con todos los marcadores codominantes estiman con mayor fiabilidad el mapa genético, seguidas de las muestras de una población Retrocruce y de una población F_2 con marcadores codominantes y dominantes.

C7. La metodología bayesiana prácticamente iguala o mejora los resultados de la mejor de las metodologías frecuentistas con las que se ha comparado.

C8. La longitud estimada del mapa genético por la metodología bayesiana es similar a la obtenida por JoinMap. Sin embargo, la metodología multipunto utilizada por Mapmaker, basada en el algoritmo EM, acorta la estimación del mapa genético sistemáticamente.

C9. La metodología bayesiana desarrollada para determinar grupos de ligamiento es eficaz, pues reparte los marcadores en cada cromosoma de forma correcta. Además permite cuantificar la probabilidad de este reparto ofreciendo distintas estructuras alternativas. Asimismo, y a diferencia de los métodos frecuentistas con los que se compara, se comporta de forma estable. El cambio de un nivel de significatividad global de $\alpha = 0.05$ a $\alpha = 0.005$ no supone un aumento excesivo en el número de grupos de ligamiento.

C10. En general, en presencia de datos faltantes, la metodología bayesiana tiene un comportamiento similar al programa que mejor reparte los marcadores en los grupos de ligamiento (Mapmaker) y después produce resultados similares al programa que mejor ordena los marcadores dentro de cada grupo de ligamiento, ya sea JoinMap o Mapmaker.

C11. En el ensayo de la metodología bayesiana completa sobre datos reales, se observa alta concordancia con los resultados obtenidos, de forma frecuentista, por los estudios previamente publicados, mostrando un comportamiento muy estable con respecto al α utilizado en los múltiples test de independencia de la fase de elaboración de los grupos de ligamiento.

Apéndice A

Distribuciones y Algoritmos

A.1. Distribuciones de probabilidad

A.1.1. Distribución Beta Truncada

La distribución *Beta* truncada, $BetaT(r|\alpha,\beta;\theta)$, viene definida por la función de densidad:

$$\pi(r) = \frac{1}{Beta(\alpha,\beta)} \frac{r^{\alpha-1}(\theta-r)^{\beta-1}}{\theta^{\alpha+\beta-1}},$$

para $r \in [0,\theta]$ y $Beta(\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$

Nota: *Para simular de una distribución Beta truncada en* $(0,\theta)$ *con parámetros* (α,β)*, basta con simular* $r^* \sim Beta(\alpha,\beta)$ *y tomar como valor simulado* $r = \theta r^*$

A.1.2. Distribución Normal Truncada

La distribución Normal truncada en el intervalo (a, b), con media μ y varianza σ^2, tiene como función de densidad:

$$f(x) = \frac{1}{\sigma(\Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma}))} \phi(\frac{x-\mu}{\sigma}),$$

donde $\Phi()$ y $\phi()$ denotan, respectivamente, las funciones de distribución y de densidad de una distribución normal estándar.

A.1.3. Distribución Multinomial

La distribución multinomial $Multinomial(x|n; p_1, ..., p_k)$ con probabilidades $p_j \in [0, 1]$ y $\sum_{j=1}^{k} p_j = 1$, tiene como distribución de probabilidad:

$$p(x) = (\frac{n!}{x_1! \cdots x_n!}) p_1^{x_1} \cdots p_k^{x_k},$$

con x_j=0, 1, 2,..., n, $\sum_{j=1}^{k} x_j = n$

A.2. Algoritmos de simulación

A.2.1. Aceptación-Rechazo

Si existe una constante $c^* > 0$ y una densidad 'suave' *p(r)*, llamada *función envoltura*, tales que:

$$\pi(r|R) \leq c^* p(r),$$

es posible obtener una muestra aleatoria de $\pi(r|R)$ a partir de simulaciones de *p(r)* según:

1. *Generar* $r^* \sim p(r)$.

2. *Generar* $U \sim Uniforme(0, 1)$.

3. *Si* $q(r^*) = \pi(r^*|R)/p(r^*)$, *aceptar* r^* *si* $U < q(r^*)/c^*$;

en otro caso, rechazar r^* y repetir los pasos 1-3
hasta conseguir el número de simulaciones deseadas.

A.2.2. Metropolis-Hastings

Se simula de una distribución propuesta (proposal) *p(r)*, desde la que vamos 'saltando' a la distribución objeto $\pi(r|R)$:

1. Inicializar $r^{(0)}$ y t=0. Una vez simulado el valor $r^{(t)}$

buscamos la simulación t+1, $r^{(t+1)}$.

2. Simular un candidato $r^* \sim p(r)$.

3. Simular $u \sim Uniforme(0,1)$.

4. Calcular la probabilidad de salto de $r^{(t)}$ a r^* con:

$\alpha(r^{(t)}, r^*) = mín\{1, \frac{\pi(r^*|R)p(r^{(t)})}{\pi(r^{(t)}|R)p(r^*)}\} = mín\{1, \frac{q(r^*)}{q(r^{(t)})}\}$, con q(r)=$\pi(r|R)$/p(r).

Si $u \leq \alpha(r^{(t)}, r^*)$, entonces tomar $r^{(t+1)} = r^*$, y si no, $r^{(t+1)} = r^{(t)}$

5. Repetir los pasos 2-4
hasta completar el número de simulaciones deseadas.

Apéndice B

Apéndice del Capítulo 2

En los Cuadros del B.1 al B.6 aparecen tanto las medias como las desviaciones típicas posteriores de las distancias multipunto entre cada pareja de marcadores para los tres tamaños muestrales: 200, 100 y 50 individuos.

	M2	M3	M4	M5	M6	M7	M8	M9	M10
M1	0.1665	0.2625	0.3357	0.4548	0.6039	0.7000	0.7220	0.9160	0.9245
M2	0	0.0978	0.1701	0.2887	0.4383	0.5349	0.5562	0.7507	0.7592
M3	0	0	0.0749	0.1930	0.3429	0.4397	0.4607	0.6555	0.6639
M4	0	0	0	0.1202	0.2702	0.3672	0.3878	0.5830	0.5915
M5	0	0	0	0	0.1509	0.2477	0.2683	0.4640	0.4725
M6	0	0	0	0	0	0.0984	0.1197	0.3147	0.3231
M7	0	0	0	0	0	0	0.0306	0.2187	0.2272
M8	0	0	0	0	0	0	0	0.1979	0.2064
M9	0	0	0	0	0	0	0	0	0.0285

	M11	M12	M13	M14	M15	M16	M17	M18	M19	M20
M1	1.0242	1.0731	1.1092	1.3141	1.3787	1.4544	1.6266	1.6744	1.7771	1.8059
M2	0.8591	0.9079	0.9441	1.1498	1.2144	1.2900	1.4623	1.5102	1.6129	1.6416
M3	0.7641	0.8128	0.8490	1.0554	1.1199	1.1955	1.3679	1.4158	1.5185	1.5472
M4	0.6917	0.7404	0.7765	0.9834	1.0479	1.1235	1.2959	1.3438	1.4465	1.4753
M5	0.5724	0.6210	0.6571	0.8648	0.9292	1.0048	1.1773	1.2252	1.3279	1.3567
M6	0.4234	0.4720	0.5083	0.7169	0.7810	0.8567	1.0294	1.0772	1.1799	1.2086
M7	0.3271	0.3756	0.4117	0.6210	0.6852	0.7609	0.9335	0.9814	1.0841	1.1128
M8	0.3059	0.3544	0.3907	0.6001	0.6641	0.7398	0.9126	0.9604	1.0631	1.0918
M9	0.1207	0.1710	0.2073	0.4168	0.4809	0.5572	0.7312	0.7793	0.8826	0.9112
M10	0.1137	0.1640	0.2004	0.4091	0.4732	0.5496	0.7234	0.7716	0.8749	0.9035
M11	0	0.0587	0.0926	0.3032	0.3679	0.4444	0.6181	0.6665	0.7699	0.7985
M12	0	0	0.0438	0.2525	0.3165	0.3931	0.5675	0.6157	0.7193	0.7482
M13	0	0	0	0.2159	0.2797	0.3563	0.5307	0.5791	0.6829	0.7118
M14	0	0	0	0	0.0713	0.1496	0.3300	0.3805	0.4889	0.5191
M15	0	0	0	0	0	0.0815	0.2630	0.3135	0.4218	0.4521
M16	0	0	0	0	0	0	0.1839	0.2343	0.3424	0.3727
M17	0	0	0	0	0	0	0	0.0574	0.1656	0.1961
M18	0	0	0	0	0	0	0	0	0.1156	0.1456
M19	0	0	0	0	0	0	0	0	0	0.0528

Cuadro B.1: *Media posterior de la distribución de distancias multipunto con 200 individuos.*

	M2	M3	M4	M5	M6	M7	M8	M9	M10
M1	0.0320	0.0347	0.0361	0.0426	0.0575	0.0661	0.0611	0.1124	0.1137
M2	0	0.0184	0.0217	0.0275	0.0393	0.0437	0.0444	0.0979	0.0996
M3	0	0	0.0156	0.0222	0.0337	0.0352	0.0390	0.0928	0.0946
M4	0	0	0	0.0216	0.0312	0.0324	0.0365	0.0893	0.0911
M5	0	0	0	0	0.0275	0.0281	0.0304	0.0826	0.0848
M6	0	0	0	0	0	0.0250	0.0256	0.0741	0.0762
M7	0	0	0	0	0	0	0.0221	0.0730	0.0749
M8	0	0	0	0	0	0	0	0.0743	0.0764
M9	0	0	0	0	0	0	0	0	0.0073

	M11	M12	M13	M14	M15	M16	M17	M18	M19	M20
M1	0.1016	0.0932	0.0932	0.1264	0.1194	0.1201	0.1418	0.1492	0.1762	0.1836
M2	0.0826	0.0739	0.0738	0.1028	0.0950	0.0963	0.1225	0.1315	0.1611	0.1694
M3	0.0752	0.0664	0.0663	0.0915	0.0836	0.0850	0.1134	0.1232	0.1543	0.1630
M4	0.0707	0.0619	0.0617	0.0846	0.0764	0.0783	0.1086	0.1193	0.1516	0.1604
M5	0.0646	0.0564	0.0562	0.0757	0.0675	0.0702	0.1037	0.1152	0.1483	0.1576
M6	0.0573	0.0502	0.0506	0.0702	0.0642	0.0686	0.1043	0.1170	0.1508	0.1605
M7	0.0570	0.0508	0.0523	0.0723	0.0659	0.0711	0.1083	0.1208	0.1539	0.1632
M8	0.0573	0.0515	0.0517	0.0714	0.0660	0.0708	0.1071	0.1199	0.1534	0.1629
M9	0.0511	0.0470	0.0477	0.0734	0.0663	0.0679	0.0985	0.1101	0.1412	0.1515
M10	0.0514	0.0481	0.0475	0.0735	0.0668	0.0680	0.0985	0.1100	0.1407	0.1512
M11	0	0.0370	0.0330	0.0635	0.0527	0.0511	0.0815	0.0927	0.1249	0.1355
M12	0	0	0.0268	0.0634	0.0504	0.0475	0.0792	0.0912	0.1226	0.1322
M13	0	0	0	0.0657	0.0521	0.0485	0.0796	0.0903	0.1206	0.1301
M14	0	0	0	0	0.0262	0.0281	0.0433	0.0442	0.0579	0.0633
M15	0	0	0	0	0	0.0190	0.0412	0.0408	0.0549	0.0597
M16	0	0	0	0	0	0	0.0393	0.0378	0.0508	0.0560
M17	0	0	0	0	0	0	0	0.0167	0.0324	0.0350
M18	0	0	0	0	0	0	0	0	0.0283	0.0317
M19	0	0	0	0	0	0	0	0	0	0.0133

Cuadro B.2: *Desviación típica posterior de la distribución de distancias multipunto con 200 individuos.*

	M2	M3	M4	M5	M6	M7	M8	M9	M10
M1	0.1704	0.2615	0.3132	0.3924	0.5226	0.5962	0.6255	0.7821	0.7818
M2	0	0.0984	0.1493	0.2269	0.3575	0.4319	0.4616	0.6191	0.6187
M3	0	0	0.0604	0.1356	0.2705	0.3459	0.3756	0.5331	0.5324
M4	0	0	0	0.0844	0.2210	0.2966	0.3265	0.4844	0.4836
M5	0	0	0	0	0.1434	0.2202	0.2497	0.4090	0.4080
M6	0	0	0	0	0	0.0865	0.1154	0.2710	0.2704
M7	0	0	0	0	0	0	0.0435	0.1999	0.1994
M8	0	0	0	0	0	0	0	0.1721	0.1720
M9	0	0	0	0	0	0	0	0	0.0108

	M11	M12	M13	M14	M15	M16	M17	M18	M19	M20
M1	0.8826	0.9264	0.9738	1.1399	1.1963	1.2647	1.3699	1.4240	1.4710	1.4912
M2	0.7199	0.7640	0.8118	0.9815	1.0381	1.1067	1.2121	1.2664	1.3135	1.3337
M3	0.6340	0.6782	0.7262	0.8978	0.9544	1.0230	1.1284	1.1826	1.2299	1.2501
M4	0.5852	0.6296	0.6775	0.8507	0.9073	0.9759	1.0813	1.1356	1.1828	1.2030
M5	0.5097	0.5541	0.6022	0.7773	0.8339	0.9025	1.0079	1.0621	1.1094	1.1296
M6	0.3726	0.4172	0.4654	0.6445	0.7012	0.7700	0.8757	0.9300	0.9770	0.9973
M7	0.3014	0.3456	0.3937	0.5748	0.6315	0.7003	0.8059	0.8602	0.9073	0.9276
M8	0.2736	0.3177	0.3658	0.5479	0.6046	0.6734	0.7791	0.8334	0.8804	0.9007
M9	0.1072	0.1532	0.1999	0.3856	0.4426	0.5117	0.6174	0.6715	0.7187	0.7391
M10	0.1075	0.1529	0.2000	0.3854	0.4424	0.5115	0.6172	0.6713	0.7188	0.7392
M11	0	0.0636	0.1055	0.2909	0.3485	0.4170	0.5229	0.5769	0.6245	0.6448
M12	0	0	0.0538	0.2382	0.2957	0.3642	0.4704	0.5247	0.5720	0.5924
M13	0	0	0	0.1909	0.2476	0.3168	0.4223	0.4764	0.5239	0.5442
M14	0	0	0	0	0.0697	0.1430	0.2667	0.3236	0.3713	0.3924
M15	0	0	0	0	0	0.0800	0.2032	0.2607	0.3082	0.3295
M16	0	0	0	0	0	0	0.1283	0.1859	0.2327	0.2543
M17	0	0	0	0	0	0	0	0.0688	0.1247	0.1445
M18	0	0	0	0	0	0	0	0	0.0841	0.0991
M19	0	0	0	0	0	0	0	0	0	0.0448

Cuadro B.3: *Media posterior de la distribución de distancias multipunto con 100 individuos.*

	M2	M3	M4	M5	M6	M7	M8	M9	M10
M1	0.0642	0.0635	0.0663	0.0712	0.1028	0.1153	0.1180	0.1398	0.1381
M2	0	0.0303	0.0320	0.0397	0.0685	0.0774	0.0801	0.0997	0.0986
M3	0	0	0.0213	0.0292	0.0533	0.0551	0.0583	0.0789	0.0783
M4	0	0	0	0.0279	0.0522	0.0525	0.0537	0.0716	0.0713
M5	0	0	0	0	0.0602	0.0565	0.0571	0.0680	0.0678
M6	0	0	0	0	0	0.0368	0.0392	0.0528	0.0517
M7	0	0	0	0	0	0	0.0235	0.0586	0.0578
M8	0	0	0	0	0	0	0	0.0666	0.0660
M9	0	0	0	0	0	0	0	0	0.0192

	M11	M12	M13	M14	M15	M16	M17	M18	M19	M20
M1	0.1477	0.1469	0.1494	0.2093	0.2033	0.2011	0.2152	0.2258	0.2340	0.2397
M2	0.1099	0.1061	0.1076	0.1618	0.1525	0.1482	0.1642	0.1757	0.1847	0.1907
M3	0.0897	0.0841	0.0856	0.1384	0.1272	0.1218	0.1409	0.1535	0.1633	0.1703
M4	0.0841	0.0776	0.0795	0.1287	0.1163	0.1101	0.1302	0.1432	0.1530	0.1605
M5	0.0796	0.0717	0.0739	0.1175	0.1034	0.0963	0.1183	0.1323	0.1424	0.1505
M6	0.0678	0.0588	0.0615	0.1033	0.0855	0.0759	0.0984	0.1138	0.1270	0.1348
M7	0.0710	0.0646	0.0687	0.1069	0.0896	0.0804	0.1029	0.1175	0.1302	0.1376
M8	0.0771	0.0721	0.0764	0.1118	0.0954	0.0866	0.1081	0.1219	0.1337	0.1409
M9	0.0477	0.0422	0.0443	0.1058	0.0867	0.0757	0.0984	0.1132	0.1248	0.1315
M10	0.0476	0.0409	0.0436	0.1053	0.0864	0.0752	0.0982	0.1131	0.1242	0.1308
M11	0	0.0447	0.0496	0.1225	0.1063	0.1003	0.1198	0.1335	0.1430	0.1490
M12	0	0	0.0375	0.1135	0.0948	0.0868	0.1039	0.1170	0.1293	0.1356
M13	0	0	0	0.1156	0.0969	0.0867	0.1031	0.1163	0.1271	0.1335
M14	0	0	0	0	0.0265	0.0332	0.0511	0.0603	0.0680	0.0731
M15	0	0	0	0	0	0.0281	0.0469	0.0518	0.0596	0.0645
M16	0	0	0	0	0	0	0.0451	0.0469	0.0528	0.0564
M17	0	0	0	0	0	0	0	0.0415	0.0396	0.0427
M18	0	0	0	0	0	0	0	0	0.0317	0.0367
M19	0	0	0	0	0	0	0	0	0	0.0212

Cuadro B.4: *Desviación típica posterior de la distribución de distancias multipunto con 100 individuos.*

	M2	M3	M4	M5	M6	M7	M8	M9	M10
M1	0.1392	0.1770	0.2429	0.3157	0.4412	0.4612	0.4907	0.6332	0.6326
M2	0	0.0547	0.1225	0.2028	0.3413	0.3595	0.3911	0.5625	0.5621
M3	0	0	0.0871	0.1658	0.3103	0.3281	0.3601	0.5367	0.5363
M4	0	0	0	0.1040	0.2538	0.2704	0.3025	0.4891	0.4887
M5	0	0	0	0	0.1797	0.1925	0.2262	0.4291	0.4286
M6	0	0	0	0	0	0.0649	0.1002	0.2902	0.2899
M7	0	0	0	0	0	0	0.0662	0.2876	0.2873
M8	0	0	0	0	0	0	0	0.2802	0.2798
M9	0	0	0	0	0	0	0	0	0.0142

	M11	M12	M13	M14	M15	M16	M17	M18	M19	M20
M1	0.6579	0.6742	0.7084	0.8035	0.8135	0.8586	0.9142	0.9542	0.9677	0.9799
M2	0.5906	0.6073	0.6446	0.7671	0.7779	0.8260	0.8899	0.9318	0.9447	0.9574
M3	0.5647	0.5815	0.6187	0.7495	0.7603	0.8087	0.8740	0.9159	0.9288	0.9415
M4	0.5172	0.5339	0.5712	0.7169	0.7277	0.7766	0.8438	0.8855	0.8985	0.9111
M5	0.4575	0.4747	0.5129	0.6811	0.6919	0.7423	0.8129	0.8555	0.8680	0.8804
M6	0.3220	0.3403	0.3820	0.5898	0.6018	0.6561	0.7349	0.7792	0.7921	0.8052
M7	0.3192	0.3376	0.3797	0.5918	0.6038	0.6582	0.7378	0.7819	0.7949	0.8078
M8	0.3130	0.3311	0.3728	0.5858	0.5978	0.6523	0.7319	0.7761	0.7889	0.8020
M9	0.1018	0.1160	0.1545	0.4227	0.4356	0.4963	0.5962	0.6424	0.6556	0.6696
M10	0.1015	0.1167	0.1549	0.4232	0.4361	0.4968	0.5968	0.6430	0.6562	0.6701
M11	0	0.0784	0.1134	0.3899	0.4023	0.4640	0.5664	0.6125	0.6262	0.6399
M12	0	0	0.0662	0.3474	0.3602	0.4220	0.5265	0.5723	0.5859	0.5997
M13	0	0	0	0.3008	0.3133	0.3760	0.4818	0.5279	0.5414	0.5552
M14	0	0	0	0	0.0493	0.1263	0.2959	0.3481	0.3550	0.3701
M15	0	0	0	0	0	0.1015	0.2738	0.3268	0.3336	0.3488
M16	0	0	0	0	0	0	0.2025	0.2571	0.2645	0.2802
M17	0	0	0	0	0	0	0	0.0914	0.1445	0.1566
M18	0	0	0	0	0	0	0	0	0.1338	0.1346
M19	0	0	0	0	0	0	0	0	0	0.0675

Cuadro B.5: *Media posterior de la distribución de distancias multipunto con 50 individuos.*

	M2	M3	M4	M5	M6	M7	M8	M9	M10
M1	0.1720	0.1665	0.1631	0.1647	0.1959	0.1939	0.2091	0.2395	0.2387
M2	0	0.0315	0.0466	0.0785	0.1469	0.1453	0.1603	0.1814	0.1800
M3	0	0	0.0459	0.0750	0.1460	0.1445	0.1588	0.1681	0.1667
M4	0	0	0	0.0827	0.1530	0.1494	0.1639	0.1569	0.1558
M5	0	0	0	0	0.1600	0.1545	0.1666	0.1557	0.1553
M6	0	0	0	0	0	0.0657	0.0676	0.1428	0.1431
M7	0	0	0	0	0	0	0.0504	0.1511	0.1517
M8	0	0	0	0	0	0	0	0.1742	0.1749
M9	0	0	0	0	0	0	0	0	0.0214

	M11	M12	M13	M14	M15	M16	M17	M18	M19	M20
M1	0.2448	0.2334	0.2418	0.3487	0.3512	0.3695	0.4544	0.4827	0.4867	0.4935
M2	0.1861	0.1742	0.1800	0.2777	0.2775	0.2930	0.3753	0.4031	0.4078	0.4137
M3	0.1722	0.1585	0.1639	0.2559	0.2547	0.2695	0.3502	0.3788	0.3837	0.3895
M4	0.1582	0.1431	0.1482	0.2265	0.2236	0.2376	0.3145	0.3444	0.3487	0.3551
M5	0.1537	0.1388	0.1414	0.1993	0.1935	0.2024	0.2714	0.3009	0.3067	0.3136
M6	0.1404	0.1243	0.1269	0.1817	0.1707	0.1653	0.2131	0.2382	0.2427	0.2496
M7	0.1470	0.1360	0.1396	0.1892	0.1785	0.1744	0.2174	0.2432	0.2477	0.2548
M8	0.1676	0.1580	0.1600	0.2078	0.1973	0.1923	0.2278	0.2508	0.2551	0.2627
M9	0.0603	0.0579	0.0615	0.2504	0.2379	0.2213	0.2023	0.2111	0.2160	0.2184
M10	0.0614	0.0572	0.0614	0.2507	0.2382	0.2217	0.2022	0.2110	0.2161	0.2184
M11	0	0.0573	0.0639	0.2623	0.2502	0.2345	0.2113	0.2186	0.2222	0.2242
M12	0	0	0.0590	0.2565	0.2434	0.2271	0.2032	0.2114	0.2168	0.2191
M13	0	0	0	0.2721	0.2593	0.2419	0.2149	0.2203	0.2239	0.2257
M14	0	0	0	0	0.0320	0.0836	0.1533	0.1593	0.1636	0.1672
M15	0	0	0	0	0	0.0832	0.1511	0.1554	0.1607	0.1644
M16	0	0	0	0	0	0	0.1598	0.1576	0.1590	0.1600
M17	0	0	0	0	0	0	0	0.0701	0.0645	0.0622
M18	0	0	0	0	0	0	0	0	0.0543	0.0610
M19	0	0	0	0	0	0	0	0	0	0.0481

Cuadro B.6: *Desviación típica posterior de la distribución de distancias multipunto con 50 individuos.*

Apéndice C

Apéndice del Capítulo 3

Este Apéndice está organizado en 4 secciones (población F_2 con mapa menos denso con todos los marcadores codominantes, población F_2 con mapa más denso con todos los marcadores codominantes, población F_2 con mapa menos denso con marcadores codominantes y dominantes y población F_2 con mapa más denso con marcadores codominantes y dominantes) que se corresponden con las organizadas en el Capítulo 3. En cada una de las secciones se muestran unos Cuadros donde aparecen las medias posteriores de las distribuciones de las distancias multipunto entre cada pareja de marcadores y, como medida de error, las desviaciones típicas de la distribución posterior de las distancias multipunto de las muestras representadas en el Capítulo 3. En las dos primeras secciones los resultados provienen de los tres tamaños muestrales (200, 100 y 50 individuos). Sin embargo, en las dos últimas secciones los resultados provienen de muestras únicamente de 200 individuos, como ya se ha explicado en el Capítulo 3.

C.1. Población F_2, con mapa menos denso, con todos los marcadores codominantes

	M2	M3	M4	M5	M6	M7	M8	M9	M10
M1	0.2014	0.3944	0.5056	0.6354	0.8449	0.9170	0.9404	1.1620	1.2163
M2	0	0.1930	0.3042	0.4340	0.6435	0.7155	0.7390	0.9606	1.0149
M3	0	0	0.1112	0.2409	0.4505	0.5225	0.5459	0.7676	0.8219
M4	0	0	0	0.1298	0.3393	0.4114	0.4348	0.6564	0.7107
M5	0	0	0	0	0.2095	0.2816	0.3050	0.5266	0.5809
M6	0	0	0	0	0	0.0720	0.0954	0.3171	0.3714
M7	0	0	0	0	0	0	0.0257	0.2450	0.2993
M8	0	0	0	0	0	0	0	0.2216	0.2759
M9	0	0	0	0	0	0	0	0	0.0544

	M11	M12	M13	M14	M15	M16	M17	M18	M19	M20
M1	1.3398	1.3891	1.4282	1.7107	1.8047	1.9150	2.0855	2.1561	2.2636	2.3274
M2	1.1384	1.1877	1.2267	1.5093	1.6033	1.7136	1.8841	1.9547	2.0621	2.1259
M3	0.9454	0.9946	1.0337	1.3163	1.4102	1.5205	1.6911	1.7617	1.8691	1.9329
M4	0.8342	0.8835	0.9226	1.2051	1.2991	1.4094	1.5800	1.6506	1.7580	1.8218
M5	0.7044	0.7537	0.7928	1.0754	1.1693	1.2796	1.4502	1.5208	1.6282	1.6920
M6	0.4949	0.5441	0.5832	0.8658	0.9597	1.0701	1.2406	1.3112	1.4186	1.4824
M7	0.4229	0.4721	0.5112	0.7938	0.8877	0.9980	1.1686	1.2392	1.3466	1.4104
M8	0.3994	0.4487	0.4878	0.7704	0.8643	0.9746	1.1452	1.2158	1.3232	1.3870
M9	0.1778	0.2271	0.2662	0.5487	0.6427	0.7530	0.9235	0.9941	1.1016	1.1654
M10	0.1235	0.1728	0.2119	0.4944	0.5884	0.6987	0.8692	0.9398	1.0473	1.1111
M11	0	0.0493	0.0883	0.3709	0.4648	0.5752	0.7457	0.8163	0.9237	0.9875
M12	0	0	0.0392	0.3217	0.4156	0.5259	0.6965	0.7671	0.8745	0.9383
M13	0	0	0	0.2826	0.3765	0.4868	0.6574	0.7280	0.8354	0.8992
M14	0	0	0	0	0.0939	0.2042	0.3748	0.4454	0.5528	0.6166
M15	0	0	0	0	0	0.1103	0.2809	0.3515	0.4589	0.5227
M16	0	0	0	0	0	0	0.1706	0.2412	0.3486	0.4124
M17	0	0	0	0	0	0	0	0.0706	0.1780	0.2418
M18	0	0	0	0	0	0	0	0	0.1074	0.1712
M19	0	0	0	0	0	0	0	0	0	0.0664

Cuadro C.1: *Media posterior de la distribución de distancias multipunto con 200 individuos, que proviene de una población F_2, con mapa menos denso, con todos los marcadores codominantes.*

	M2	M3	M4	M5	M6	M7	M8	M9	M10
M1	0.0158	0.0176	0.0188	0.0198	0.0210	0.0218	0.0219	0.0227	0.0229
M2	0	0.0128	0.0144	0.0159	0.0172	0.0181	0.0182	0.0193	0.0195
M3	0	0	0.0096	0.0115	0.0135	0.0146	0.0149	0.0159	0.0159
M4	0	0	0	0.0100	0.0123	0.0136	0.0139	0.0148	0.0148
M5	0	0	0	0	0.0110	0.0125	0.0128	0.0138	0.0139
M6	0	0	0	0	0	0.0082	0.0084	0.0095	0.0096
M7	0	0	0	0	0	0	0.0046	0.0101	0.0102
M8	0	0	0	0	0	0	0	0.0102	0.0104
M9	0	0	0	0	0	0	0	0	0.0054

	M11	M12	M13	M14	M15	M16	M17	M18	M19	M20
M1	0.0236	0.0236	0.0239	0.0271	0.0277	0.0286	0.0297	0.0301	0.0321	0.0330
M2	0.0202	0.0203	0.0205	0.0243	0.0247	0.0257	0.0270	0.0274	0.0297	0.0303
M3	0.0168	0.0170	0.0174	0.0214	0.0219	0.0230	0.0244	0.0248	0.0270	0.0279
M4	0.0157	0.0159	0.0163	0.0204	0.0209	0.0220	0.0234	0.0239	0.0262	0.0271
M5	0.0149	0.0151	0.0156	0.0198	0.0202	0.0214	0.0230	0.0235	0.0259	0.0268
M6	0.0106	0.0108	0.0113	0.0158	0.0164	0.0175	0.0192	0.0197	0.0225	0.0235
M7	0.0111	0.0114	0.0118	0.0162	0.0168	0.0179	0.0194	0.0199	0.0226	0.0235
M8	0.0115	0.0117	0.0120	0.0162	0.0169	0.0180	0.0196	0.0201	0.0228	0.0237
M9	0.0075	0.0079	0.0085	0.0133	0.0138	0.0149	0.0165	0.0170	0.0201	0.0211
M10	0.0067	0.0072	0.0077	0.0129	0.0134	0.0145	0.0160	0.0165	0.0197	0.0207
M11	0	0.0051	0.0062	0.0123	0.0127	0.0139	0.0154	0.0159	0.0191	0.0201
M12	0	0	0.0050	0.0121	0.0127	0.0138	0.0153	0.0159	0.0191	0.0202
M13	0	0	0	0.0121	0.0127	0.0138	0.0154	0.0159	0.0191	0.0202
M14	0	0	0	0	0.0083	0.0095	0.0111	0.0117	0.0157	0.0171
M15	0	0	0	0	0	0.0086	0.0105	0.0110	0.0153	0.0166
M16	0	0	0	0	0	0	0.0092	0.0098	0.0143	0.0159
M17	0	0	0	0	0	0	0	0.0061	0.0116	0.0138
M18	0	0	0	0	0	0	0	0	0.0111	0.0130
M19	0	0	0	0	0	0	0	0	0	0.0069

Cuadro C.2: *Desviación típica de la distribución de distancias multipunto con 200 individuos, que proviene de una población F_2, con mapa menos denso, con todos los marcadores codominantes.*

	M2	M3	M4	M5	M6	M7	M8	M9	M10
M1	0.2007	0.3873	0.4668	0.5980	0.8094	0.8648	0.8965	1.0711	1.0995
M2	0	0.1866	0.2661	0.3974	0.6089	0.6641	0.6958	0.8712	0.8995
M3	0	0	0.0798	0.2109	0.4224	0.4775	0.5092	0.6852	0.7135
M4	0	0	0	0.1314	0.3429	0.3980	0.4297	0.6060	0.6343
M5	0	0	0	0	0.2118	0.2668	0.2986	0.4749	0.5032
M6	0	0	0	0	0	0.0580	0.0894	0.2638	0.2921
M7	0	0	0	0	0	0	0.0347	0.2089	0.2372
M8	0	0	0	0	0	0	0	0.1773	0.2056
M9	0	0	0	0	0	0	0	0	0.0300

	M11	M12	M13	M14	M15	M16	M17	M18	M19	M20
M1	1.1920	1.2379	1.2771	1.4757	1.5515	1.6270	1.7169	1.7608	1.8433	1.8835
M2	0.9920	1.0380	1.0772	1.2757	1.3516	1.4270	1.5170	1.5609	1.6434	1.6836
M3	0.8060	0.8520	0.8912	1.0897	1.1656	1.2410	1.3310	1.3749	1.4574	1.4976
M4	0.7268	0.7728	0.8120	1.0105	1.0864	1.1618	1.2518	1.2956	1.3781	1.4184
M5	0.5957	0.6417	0.6809	0.8794	0.9553	1.0307	1.1207	1.1646	1.2470	1.2873
M6	0.3846	0.4305	0.4698	0.6683	0.7441	0.8196	0.9095	0.9534	1.0359	1.0762
M7	0.3297	0.3757	0.4149	0.6134	0.6893	0.7647	0.8547	0.8986	0.9810	1.0213
M8	0.2981	0.3440	0.3833	0.5818	0.6577	0.7331	0.8231	0.8669	0.9494	0.9897
M9	0.1219	0.1682	0.2075	0.4057	0.4820	0.5575	0.6480	0.6921	0.7747	0.8152
M10	0.0933	0.1397	0.1789	0.3772	0.4535	0.5291	0.6195	0.6637	0.7463	0.7867
M11	0	0.0468	0.0859	0.2844	0.3608	0.4363	0.5268	0.5709	0.6535	0.6940
M12	0	0	0.0396	0.2383	0.3146	0.3902	0.4806	0.5247	0.6074	0.6478
M13	0	0	0	0.1989	0.2753	0.3508	0.4412	0.4854	0.5680	0.6085
M14	0	0	0	0	0.0806	0.1617	0.2576	0.3036	0.3896	0.4301
M15	0	0	0	0	0	0.0814	0.1772	0.2232	0.3092	0.3497
M16	0	0	0	0	0	0	0.0962	0.1426	0.2294	0.2703
M17	0	0	0	0	0	0	0	0.0468	0.1340	0.1748
M18	0	0	0	0	0	0	0	0	0.0875	0.1283
M19	0	0	0	0	0	0	0	0	0	0.0588

Cuadro C.3: *Media posterior de la distribución de distancias multipunto con 100 individuos, que proviene de una población F_2, con mapa menos denso, con todos los marcadores codominantes.*

	M2	M3	M4	M5	M6	M7	M8	M9	M10
M1	0.0220	0.0251	0.0259	0.0301	0.0366	0.0327	0.0340	0.0632	0.0619
M2	0	0.0174	0.0185	0.0216	0.0285	0.0269	0.0287	0.0477	0.0460
M3	0	0	0.0108	0.0165	0.0244	0.0236	0.0253	0.0390	0.0368
M4	0	0	0	0.0154	0.0230	0.0221	0.0238	0.0360	0.0336
M5	0	0	0	0	0.0225	0.0207	0.0226	0.0353	0.0330
M6	0	0	0	0	0	0.0133	0.0157	0.0343	0.0320
M7	0	0	0	0	0	0	0.0080	0.0345	0.0324
M8	0	0	0	0	0	0	0	0.0374	0.0353
M9	0	0	0	0	0	0	0	0	0.0061

	M11	M12	M13	M14	M15	M16	M17	M18	M19	M20
M1	0.0592	0.0584	0.0579	0.0883	0.0730	0.0650	0.0713	0.0787	0.0972	0.1043
M2	0.0426	0.0416	0.0409	0.0791	0.0617	0.0518	0.0596	0.0681	0.0888	0.0965
M3	0.0324	0.0312	0.0304	0.0748	0.0558	0.0447	0.0531	0.0623	0.0842	0.0926
M4	0.0292	0.0279	0.0271	0.0735	0.0543	0.0432	0.0522	0.0615	0.0837	0.0922
M5	0.0285	0.0273	0.0267	0.0738	0.0546	0.0436	0.0525	0.0618	0.0838	0.0925
M6	0.0281	0.0275	0.0272	0.0729	0.0545	0.0447	0.0548	0.0640	0.0862	0.0947
M7	0.0288	0.0285	0.0281	0.0731	0.0550	0.0456	0.0558	0.0650	0.0868	0.0957
M8	0.0317	0.0312	0.0309	0.0741	0.0563	0.0472	0.0570	0.0659	0.0875	0.0963
M9	0.0100	0.0115	0.0125	0.0622	0.0402	0.0296	0.0427	0.0530	0.0782	0.0864
M10	0.0094	0.0097	0.0117	0.0621	0.0397	0.0284	0.0413	0.0517	0.0770	0.0852
M11	0	0.0061	0.0084	0.0614	0.0383	0.0257	0.0386	0.0492	0.0752	0.0829
M12	0	0	0.0069	0.0633	0.0402	0.0266	0.0374	0.0476	0.0734	0.0809
M13	0	0	0	0.0636	0.0404	0.0265	0.0368	0.0469	0.0726	0.0803
M14	0	0	0	0	0.0087	0.0148	0.0157	0.0161	0.0306	0.0363
M15	0	0	0	0	0	0.0128	0.0130	0.0141	0.0303	0.0362
M16	0	0	0	0	0	0	0.0124	0.0124	0.0246	0.0277
M17	0	0	0	0	0	0	0	0.0060	0.0210	0.0264
M18	0	0	0	0	0	0	0	0	0.0214	0.0256
M19	0	0	0	0	0	0	0	0	0	0.0109

Cuadro C.4: *Desviación típica de la distribución de distancias multipunto con 100 individuos, que proviene de una población F_2, con mapa menos denso, con todos los marcadores codominantes.*

	M2	M3	M4	M5	M6	M7	M8	M9	M10
M1	0.1556	0.2813	0.3694	0.5005	0.7354	0.7733	0.8238	1.0394	1.0639
M2	0	0.1261	0.2143	0.3452	0.5811	0.6187	0.6691	0.8856	0.9100
M3	0	0	0.0883	0.2194	0.4564	0.4938	0.5442	0.7612	0.7856
M4	0	0	0	0.1315	0.3690	0.4063	0.4568	0.6741	0.6986
M5	0	0	0	0	0.2385	0.2757	0.3260	0.5440	0.5684
M6	0	0	0	0	0	0.0509	0.0987	0.3077	0.3322
M7	0	0	0	0	0	0	0.0595	0.2710	0.2954
M8	0	0	0	0	0	0	0	0.2217	0.2461
M9	0	0	0	0	0	0	0	0	0.0310

	M11	M12	M13	M14	M15	M16	M17	M18	M19	M20
M1	1.1919	1.2292	1.2812	1.5610	1.5902	1.6652	1.7597	1.8155	1.9240	1.9202
M2	1.0380	1.0753	1.1273	1.4076	1.4367	1.5118	1.6063	1.6620	1.7705	1.7667
M3	0.9137	0.9510	1.0030	1.2836	1.3128	1.3878	1.4823	1.5380	1.6465	1.6428
M4	0.8266	0.8639	0.9159	1.1968	1.2260	1.3010	1.3955	1.4512	1.5597	1.5559
M5	0.6965	0.7337	0.7857	1.0671	1.0963	1.1713	1.2658	1.3215	1.4300	1.4262
M6	0.4601	0.4974	0.5494	0.8311	0.8602	0.9353	1.0298	1.0855	1.1940	1.1902
M7	0.4234	0.4607	0.5127	0.7944	0.8236	0.8986	0.9931	1.0489	1.1574	1.1536
M8	0.3740	0.4113	0.4632	0.7453	0.7745	0.8495	0.9440	0.9997	1.1082	1.1044
M9	0.1559	0.1938	0.2457	0.5263	0.5555	0.6307	0.7253	0.7812	0.8897	0.8860
M10	0.1309	0.1687	0.2206	0.5017	0.5309	0.6061	0.7007	0.7566	0.8651	0.8614
M11	0	0.0431	0.0922	0.3735	0.4027	0.4781	0.5730	0.6290	0.7378	0.7340
M12	0	0	0.0538	0.3361	0.3654	0.4408	0.5356	0.5917	0.7005	0.6966
M13	0	0	0	0.2840	0.3133	0.3887	0.4835	0.5396	0.6484	0.6445
M14	0	0	0	0	0.0585	0.1771	0.2843	0.3472	0.4663	0.4576
M15	0	0	0	0	0	0.1284	0.2362	0.2991	0.4182	0.4096
M16	0	0	0	0	0	0	0.1167	0.1861	0.3187	0.3089
M17	0	0	0	0	0	0	0	0.0714	0.2081	0.1983
M18	0	0	0	0	0	0	0	0	0.1375	0.1275
M19	0	0	0	0	0	0	0	0	0	0.0683

Cuadro C.5: *Media posterior de la distribución de distancias multipunto con 50 individuos, que proviene de una población F_2, con mapa menos denso, con todos los marcadores codominantes.*

	M2	M3	M4	M5	M6	M7	M8	M9	M10
M1	0.0272	0.0314	0.0349	0.0429	0.0752	0.0717	0.0761	0.0985	0.0973
M2	0	0.0213	0.0247	0.0348	0.0601	0.0577	0.0632	0.0804	0.0790
M3	0	0	0.0165	0.0254	0.0459	0.0443	0.0507	0.0638	0.0617
M4	0	0	0	0.0212	0.0418	0.0392	0.0461	0.0563	0.0541
M5	0	0	0	0	0.0397	0.0352	0.0432	0.0496	0.0471
M6	0	0	0	0	0	0.0282	0.0311	0.0492	0.0469
M7	0	0	0	0	0	0	0.0236	0.0480	0.0453
M8	0	0	0	0	0	0	0	0.0584	0.0561
M9	0	0	0	0	0	0	0	0	0.0097

	M11	M12	M13	M14	M15	M16	M17	M18	M19	M20
M1	0.0986	0.0991	0.1004	0.1796	0.1531	0.1289	0.1434	0.1634	0.2208	0.2139
M2	0.0807	0.0815	0.0830	0.1676	0.1384	0.1109	0.1268	0.1486	0.2096	0.2022
M3	0.0646	0.0654	0.0672	0.1586	0.1267	0.0956	0.1129	0.1365	0.2008	0.1930
M4	0.0578	0.0585	0.0603	0.1551	0.1221	0.0889	0.1066	0.1310	0.1967	0.1885
M5	0.0520	0.0526	0.0543	0.1525	0.1181	0.0820	0.1002	0.1253	0.1924	0.1841
M6	0.0522	0.0526	0.0541	0.1508	0.1160	0.0806	0.0984	0.1239	0.1908	0.1829
M7	0.0509	0.0512	0.0528	0.1503	0.1151	0.0804	0.0987	0.1243	0.1915	0.1829
M8	0.0592	0.0591	0.0606	0.1530	0.1183	0.0854	0.1029	0.1274	0.1937	0.1849
M9	0.0259	0.0249	0.0266	0.1381	0.1034	0.0763	0.1021	0.1293	0.1988	0.1900
M10	0.0261	0.0243	0.0257	0.1373	0.1021	0.0741	0.1000	0.1275	0.1975	0.1884
M11	0	0.0099	0.0172	0.1348	0.0993	0.0726	0.0988	0.1261	0.1954	0.1865
M12	0	0	0.0148	0.1365	0.1010	0.0736	0.0977	0.1245	0.1936	0.1847
M13	0	0	0	0.1390	0.1037	0.0754	0.0973	0.1233	0.1916	0.1828
M14	0	0	0	0	0.0133	0.0511	0.0338	0.0449	0.1067	0.1038
M15	0	0	0	0	0	0.0455	0.0305	0.0440	0.1073	0.1051
M16	0	0	0	0	0	0	0.0360	0.0321	0.0540	0.0573
M17	0	0	0	0	0	0	0	0.0113	0.0360	0.0418
M18	0	0	0	0	0	0	0	0	0.0357	0.0408
M19	0	0	0	0	0	0	0	0	0	0.0158

Cuadro C.6: *Desviación típica de la distribución de distancias multipunto con 50 individuos, que proviene de una población F_2, con mapa menos denso, con todos los marcadores codominantes.*

C.2. Población F_2, con mapa más denso, con todos los marcadores codominantes

	M2	M3	M4	M5	M6	M7	M8	M9	M10
M1	0.0248	0.0459	0.2251	0.2613	0.2895	0.3038	0.3755	0.3903	0.4256
M2	0	0.0252	0.2044	0.2405	0.2688	0.2831	0.3547	0.3695	0.4048
M3	0	0	0.1859	0.2218	0.2502	0.2644	0.3361	0.3509	0.3862
M4	0	0	0	0.0367	0.0650	0.0798	0.1509	0.1660	0.2015
M5	0	0	0	0	0.0296	0.0440	0.1151	0.1301	0.1656
M6	0	0	0	0	0	0.0173	0.0867	0.1017	0.1371
M7	0	0	0	0	0	0	0.0725	0.0875	0.1228
M8	0	0	0	0	0	0	0	0.0173	0.0515
M9	0	0	0	0	0	0	0	0	0.0369

	M11	M12	M13	M14	M15	M16	M17	M18	M19	M20
M1	0.4555	0.5002	0.5572	0.5762	0.6067	0.6831	0.7176	0.7219	0.7326	0.8302
M2	0.4348	0.4795	0.5364	0.5554	0.5859	0.6625	0.6969	0.7012	0.7119	0.8095
M3	0.4162	0.4608	0.5177	0.5369	0.5673	0.6437	0.6782	0.6825	0.6932	0.7908
M4	0.2314	0.2761	0.3331	0.3522	0.3830	0.4586	0.4931	0.4973	0.5080	0.6057
M5	0.1957	0.2403	0.2973	0.3163	0.3472	0.4227	0.4572	0.4615	0.4722	0.5698
M6	0.1671	0.2121	0.2691	0.2880	0.3190	0.3942	0.4287	0.4329	0.4437	0.5413
M7	0.1530	0.1979	0.2547	0.2735	0.3044	0.3798	0.4144	0.4186	0.4294	0.5269
M8	0.0818	0.1265	0.1831	0.2024	0.2333	0.3085	0.3431	0.3473	0.3581	0.4556
M9	0.0672	0.1114	0.1683	0.1876	0.2183	0.2937	0.3283	0.3326	0.3433	0.4409
M10	0.0315	0.0762	0.1331	0.1526	0.1830	0.2585	0.2931	0.2973	0.3081	0.4056
M11	0	0.0466	0.1033	0.1225	0.1531	0.2286	0.2632	0.2674	0.2782	0.3757
M12	0	0	0.0582	0.0774	0.1084	0.1837	0.2182	0.2225	0.2332	0.3308
M13	0	0	0	0.0217	0.0511	0.1271	0.1617	0.1659	0.1766	0.2742
M14	0	0	0	0	0.0324	0.1076	0.1423	0.1465	0.1572	0.2547
M15	0	0	0	0	0	0.0778	0.1124	0.1166	0.1273	0.2249
M16	0	0	0	0	0	0	0.0351	0.0392	0.0500	0.1472
M17	0	0	0	0	0	0	0	0.0084	0.0174	0.1127
M18	0	0	0	0	0	0	0	0	0.0133	0.1085
M19	0	0	0	0	0	0	0	0	0	0.0977

Cuadro C.7: *Media posterior de la distribución de distancias multipunto con 200 individuos, que proviene de una población F_2, con mapa más denso, con todos los marcadores codominantes.*

	M2	M3	M4	M5	M6	M7	M8	M9	M10
M1	0.0084	0.0112	0.0271	0.0273	0.0280	0.0290	0.0319	0.0341	0.0367
M2	0	0.0089	0.0238	0.0244	0.0246	0.0253	0.0285	0.0309	0.0332
M3	0	0	0.0253	0.0262	0.0255	0.0264	0.0288	0.0309	0.0330
M4	0	0	0	0.0109	0.0111	0.0141	0.0165	0.0167	0.0195
M5	0	0	0	0	0.0101	0.0112	0.0132	0.0137	0.0157
M6	0	0	0	0	0	0.0104	0.0116	0.0120	0.0142
M7	0	0	0	0	0	0	0.0127	0.0125	0.0141
M8	0	0	0	0	0	0	0	0.0093	0.0119
M9	0	0	0	0	0	0	0	0	0.0129

	M11	M12	M13	M14	M15	M16	M17	M18	M19	M20
M1	0.0397	0.0426	0.0467	0.0500	0.0536	0.0459	0.0469	0.0467	0.0469	0.0485
M2	0.0361	0.0394	0.0438	0.0468	0.0509	0.0418	0.0429	0.0428	0.0429	0.0447
M3	0.0358	0.0394	0.0436	0.0458	0.0496	0.0425	0.0434	0.0432	0.0435	0.0451
M4	0.0235	0.0246	0.0300	0.0321	0.0338	0.0322	0.0335	0.0331	0.0335	0.0355
M5	0.0193	0.0207	0.0263	0.0284	0.0297	0.0291	0.0298	0.0298	0.0301	0.0325
M6	0.0179	0.0174	0.0225	0.0253	0.0259	0.0276	0.0287	0.0283	0.0288	0.0317
M7	0.0164	0.0160	0.0217	0.0257	0.0263	0.0267	0.0271	0.0272	0.0276	0.0316
M8	0.0135	0.0133	0.0171	0.0179	0.0187	0.0202	0.0211	0.0210	0.0215	0.0252
M9	0.0141	0.0141	0.0165	0.0175	0.0187	0.0190	0.0197	0.0195	0.0202	0.0241
M10	0.0150	0.0142	0.0149	0.0150	0.0169	0.0177	0.0183	0.0185	0.0192	0.0233
M11	0	0.0175	0.0163	0.0161	0.0162	0.0194	0.0199	0.0201	0.0204	0.0241
M12	0	0	0.0150	0.0135	0.0157	0.0159	0.0170	0.0166	0.0177	0.0203
M13	0	0	0	0.0191	0.0158	0.0195	0.0200	0.0199	0.0207	0.0237
M14	0	0	0	0	0.0176	0.0191	0.0192	0.0193	0.0200	0.0236
M15	0	0	0	0	0	0.0262	0.0264	0.0265	0.0271	0.0292
M16	0	0	0	0	0	0	0.0085	0.0083	0.0092	0.0130
M17	0	0	0	0	0	0	0	0.0048	0.0061	0.0112
M18	0	0	0	0	0	0	0	0	0.0060	0.0112
M19	0	0	0	0	0	0	0	0	0	0.0118

Cuadro C.8: *Desviación Típica posterior de la distribución de distancias multipunto con 200 individuos, que proviene de una población F_2, con mapa más denso, con todos los marcadores codominantes.*

	M2	M3	M4	M5	M6	M7	M8	M9	M10
M1	0.0078	0.0255	0.1416	0.1710	0.1941	0.2460	0.3256	0.3510	0.3852
M2	0	0.0264	0.1423	0.1717	0.1948	0.2467	0.3263	0.3517	0.3858
M3	0	0	0.1312	0.1606	0.1837	0.2357	0.3154	0.3409	0.3749
M4	0	0	0	0.0345	0.0607	0.1152	0.1987	0.2256	0.2608
M5	0	0	0	0	0.0309	0.0840	0.1677	0.1948	0.2301
M6	0	0	0	0	0	0.0587	0.1419	0.1698	0.2050
M7	0	0	0	0	0	0	0.0874	0.1154	0.1508
M8	0	0	0	0	0	0	0	0.0345	0.0722
M9	0	0	0	0	0	0	0	0	0.0446

	M11	M12	M13	M14	M15	M16	M17	M18	M19	M20
M1	0.4025	0.4234	0.4610	0.4712	0.4881	0.5866	0.5962	0.6036	0.6284	0.6876
M2	0.4032	0.4241	0.4617	0.4719	0.4888	0.5874	0.5969	0.6043	0.6292	0.6883
M3	0.3923	0.4132	0.4508	0.4610	0.4779	0.5765	0.5861	0.5935	0.6183	0.6775
M4	0.2780	0.2982	0.3361	0.3473	0.3635	0.4627	0.4723	0.4798	0.5047	0.5637
M5	0.2469	0.2674	0.3050	0.3161	0.3323	0.4316	0.4412	0.4487	0.4735	0.5327
M6	0.2222	0.2428	0.2802	0.2914	0.3075	0.4069	0.4165	0.4240	0.4487	0.5079
M7	0.1676	0.1880	0.2262	0.2372	0.2536	0.3527	0.3623	0.3699	0.3946	0.4539
M8	0.0901	0.1113	0.1488	0.1597	0.1762	0.2745	0.2839	0.2913	0.3163	0.3758
M9	0.0622	0.0836	0.1207	0.1323	0.1482	0.2467	0.2561	0.2636	0.2884	0.3477
M10	0.0262	0.0463	0.0836	0.0941	0.1109	0.2099	0.2193	0.2270	0.2515	0.3109
M11	0	0.0303	0.0659	0.0772	0.0931	0.1908	0.2002	0.2077	0.2327	0.2921
M12	0	0	0.0438	0.0538	0.0709	0.1716	0.1811	0.1889	0.2137	0.2733
M13	0	0	0	0.0194	0.0334	0.1349	0.1442	0.1518	0.1766	0.2358
M14	0	0	0	0	0.0221	0.1228	0.1322	0.1400	0.1649	0.2244
M15	0	0	0	0	0	0.1064	0.1158	0.1235	0.1479	0.2075
M16	0	0	0	0	0	0	0.0143	0.0240	0.0490	0.1085
M17	0	0	0	0	0	0	0	0.0148	0.0397	0.0988
M18	0	0	0	0	0	0	0	0	0.0311	0.0900
M19	0	0	0	0	0	0	0	0	0	0.0624

Cuadro C.9: *Media posterior de la distribución de distancias multipunto con 100 individuos, que proviene de una población F_2, con mapa más denso, con todos los marcadores codominantes.*

	M2	M3	M4	M5	M6	M7	M8	M9	M10
M1	0.0063	0.0087	0.0792	0.0676	0.0583	0.0474	0.0580	0.0577	0.0552
M2	0	0.0103	0.0795	0.0680	0.0588	0.0478	0.0583	0.0580	0.0555
M3	0	0	0.0829	0.0715	0.0618	0.0503	0.0591	0.0572	0.0548
M4	0	0	0	0.0212	0.0259	0.0271	0.0488	0.0473	0.0457
M5	0	0	0	0	0.0272	0.0240	0.0462	0.0434	0.0407
M6	0	0	0	0	0	0.0242	0.0454	0.0413	0.0376
M7	0	0	0	0	0	0	0.0439	0.0378	0.0322
M8	0	0	0	0	0	0	0	0.0283	0.0289
M9	0	0	0	0	0	0	0	0	0.0299

	M11	M12	M13	M14	M15	M16	M17	M18	M19	M20
M1	0.0558	0.0577	0.0608	0.0634	0.0641	0.0805	0.0818	0.0837	0.0882	0.1025
M2	0.0560	0.0578	0.0607	0.0635	0.0642	0.0804	0.0817	0.0836	0.0881	0.1025
M3	0.0552	0.0574	0.0605	0.0625	0.0636	0.0794	0.0807	0.0826	0.0871	0.1014
M4	0.0450	0.0490	0.0517	0.0516	0.0553	0.0674	0.0683	0.0700	0.0748	0.0895
M5	0.0406	0.0437	0.0481	0.0478	0.0517	0.0638	0.0647	0.0664	0.0718	0.0868
M6	0.0365	0.0393	0.0440	0.0435	0.0474	0.0599	0.0609	0.0626	0.0686	0.0836
M7	0.0313	0.0337	0.0359	0.0359	0.0381	0.0531	0.0538	0.0552	0.0613	0.0768
M8	0.0274	0.0271	0.0280	0.0279	0.0289	0.0453	0.0464	0.0483	0.0540	0.0699
M9	0.0282	0.0278	0.0292	0.0275	0.0290	0.0436	0.0449	0.0469	0.0534	0.0703
M10	0.0298	0.0278	0.0283	0.0272	0.0272	0.0418	0.0429	0.0440	0.0510	0.0681
M11	0	0.0322	0.0288	0.0268	0.0269	0.0400	0.0408	0.0427	0.0484	0.0654
M12	0	0	0.0290	0.0262	0.0262	0.0397	0.0397	0.0409	0.0467	0.0624
M13	0	0	0	0.0214	0.0207	0.0353	0.0354	0.0357	0.0414	0.0570
M14	0	0	0	0	0.0197	0.0345	0.0337	0.0348	0.0395	0.0554
M15	0	0	0	0	0	0.0347	0.0338	0.0347	0.0397	0.0545
M16	0	0	0	0	0	0	0.0088	0.0122	0.0143	0.0259
M17	0	0	0	0	0	0	0	0.0108	0.0121	0.0237
M18	0	0	0	0	0	0	0	0	0.0151	0.0228
M19	0	0	0	0	0	0	0	0	0	0.0232

Cuadro C.10: *Desviación Típica posterior de la distribución de distancias multipunto con 100 individuos, que proviene de una población F_2, con mapa más denso, con todos los marcadores codominantes.*

	M2	M3	M4	M5	M6	M7	M8	M9	M10
M1	0.0262	0.0460	0.2032	0.2422	0.2676	0.2877	0.3449	0.3491	0.3553
M2	0	0.0373	0.1941	0.2330	0.2584	0.2787	0.3358	0.3399	0.3461
M3	0	0	0.1954	0.2341	0.2597	0.2798	0.3369	0.3411	0.3472
M4	0	0	0	0.0446	0.0708	0.0920	0.1556	0.1603	0.1667
M5	0	0	0	0	0.0326	0.0526	0.1160	0.1206	0.1270
M6	0	0	0	0	0	0.0283	0.0898	0.0944	0.1009
M7	0	0	0	0	0	0	0.0714	0.0756	0.0821
M8	0	0	0	0	0	0	0	0.0117	0.0181
M9	0	0	0	0	0	0	0	0	0.0135

	M11	M12	M13	M14	M15	M16	M17	M18	M19	M20
M1	0.3933	0.3932	0.4210	0.4607	0.4674	0.5648	0.5800	0.5877	0.6103	0.6588
M2	0.3842	0.3841	0.4120	0.4517	0.4583	0.5558	0.5710	0.5787	0.6013	0.6498
M3	0.3853	0.3852	0.4132	0.4529	0.4595	0.5570	0.5722	0.5800	0.6025	0.6510
M4	0.2041	0.2047	0.2313	0.2723	0.2791	0.3792	0.3952	0.4031	0.4264	0.4767
M5	0.1637	0.1646	0.1910	0.2319	0.2389	0.3389	0.3550	0.3627	0.3861	0.4365
M6	0.1377	0.1385	0.1650	0.2056	0.2125	0.3125	0.3284	0.3363	0.3597	0.4099
M7	0.1169	0.1178	0.1443	0.1850	0.1916	0.2919	0.3079	0.3156	0.3389	0.3894
M8	0.0594	0.0608	0.0864	0.1283	0.1349	0.2342	0.2502	0.2581	0.2811	0.3314
M9	0.0548	0.0562	0.0823	0.1239	0.1304	0.2298	0.2457	0.2534	0.2766	0.3268
M10	0.0488	0.0500	0.0759	0.1171	0.1240	0.2233	0.2393	0.2470	0.2700	0.3203
M11	0	0.0118	0.0385	0.0791	0.0853	0.1846	0.2004	0.2081	0.2311	0.2812
M12	0	0	0.0401	0.0801	0.0863	0.1856	0.2013	0.2089	0.2320	0.2823
M13	0	0	0	0.0494	0.0553	0.1537	0.1694	0.1775	0.2006	0.2509
M14	0	0	0	0	0.0143	0.1117	0.1274	0.1349	0.1581	0.2082
M15	0	0	0	0	0	0.1049	0.1207	0.1286	0.1512	0.2013
M16	0	0	0	0	0	0	0.0240	0.0317	0.0535	0.1036
M17	0	0	0	0	0	0	0	0.0166	0.0384	0.0868
M18	0	0	0	0	0	0	0	0	0.0310	0.0786
M19	0	0	0	0	0	0	0	0	0	0.0564

Cuadro C.11: *Media posterior de la distribución de distancias multipunto con 50 individuos, que proviene de una población F_2, con mapa más denso, con todos los marcadores codominantes.*

	M2	M3	M4	M5	M6	M7	M8	M9	M10
M1	0.0206	0.0200	0.0759	0.0655	0.0616	0.0616	0.0599	0.0596	0.0594
M2	0	0.0179	0.0752	0.0640	0.0593	0.0586	0.0565	0.0563	0.0561
M3	0	0	0.0800	0.0696	0.0644	0.0641	0.0615	0.0609	0.0609
M4	0	0	0	0.0252	0.0294	0.0364	0.0501	0.0493	0.0492
M5	0	0	0	0	0.0198	0.0261	0.0404	0.0392	0.0389
M6	0	0	0	0	0	0.0238	0.0367	0.0357	0.0352
M7	0	0	0	0	0	0	0.0370	0.0359	0.0351
M8	0	0	0	0	0	0	0	0.0153	0.0183
M9	0	0	0	0	0	0	0	0	0.0165

	M11	M12	M13	M14	M15	M16	M17	M18	M19	M20
M1	0.0621	0.0650	0.0629	0.0672	0.0682	0.0887	0.0923	0.0942	0.1007	0.1167
M2	0.0590	0.0618	0.0593	0.0638	0.0650	0.0858	0.0895	0.0916	0.0984	0.1146
M3	0.0635	0.0665	0.0633	0.0668	0.0680	0.0876	0.0912	0.0928	0.0998	0.1154
M4	0.0519	0.0531	0.0533	0.0538	0.0536	0.0649	0.0650	0.0653	0.0697	0.0796
M5	0.0421	0.0434	0.0439	0.0452	0.0451	0.0583	0.0583	0.0593	0.0640	0.0749
M6	0.0374	0.0394	0.0385	0.0406	0.0409	0.0551	0.0558	0.0566	0.0618	0.0736
M7	0.0383	0.0400	0.0396	0.0417	0.0426	0.0555	0.0562	0.0574	0.0628	0.0742
M8	0.0288	0.0314	0.0304	0.0320	0.0321	0.0467	0.0472	0.0486	0.0552	0.0679
M9	0.0289	0.0307	0.0299	0.0308	0.0313	0.0455	0.0465	0.0481	0.0546	0.0676
M10	0.0298	0.0315	0.0306	0.0321	0.0320	0.0462	0.0472	0.0485	0.0554	0.0684
M11	0	0.0203	0.0292	0.0302	0.0306	0.0437	0.0448	0.0466	0.0535	0.0674
M12	0	0	0.0325	0.0319	0.0318	0.0444	0.0458	0.0478	0.0543	0.0675
M13	0	0	0	0.0298	0.0296	0.0407	0.0415	0.0422	0.0483	0.0617
M14	0	0	0	0	0.0181	0.0337	0.0328	0.0346	0.0412	0.0553
M15	0	0	0	0	0	0.0343	0.0334	0.0347	0.0416	0.0559
M16	0	0	0	0	0	0	0.0181	0.0199	0.0232	0.0334
M17	0	0	0	0	0	0	0	0.0146	0.0218	0.0307
M18	0	0	0	0	0	0	0	0	0.0214	0.0291
M19	0	0	0	0	0	0	0	0	0	0.0280

Cuadro C.12: *Desviación Típica posterior de la distribución de distancias multipunto con 50 individuos, que proviene de una población F_2, con mapa más denso, con todos los marcadores codominantes.*

C.3. Población F_2, con mapa menos denso, con marcadores codominantes y dominantes

	M2	M3	M4	M5	M6	M7	M8	M9	M10
M1	0.1924	0.3362	0.4463	0.5419	0.7378	0.7970	0.7853	1.0262	1.0630
M2	0	0.1608	0.2693	0.3648	0.5604	0.6196	0.6079	0.8485	0.8853
M3	0	0	0.1359	0.2288	0.4242	0.4836	0.4718	0.7124	0.7491
M4	0	0	0	0.1063	0.2942	0.3531	0.3409	0.5820	0.6188
M5	0	0	0	0	0.1992	0.2581	0.2456	0.4870	0.5238
M6	0	0	0	0	0	0.0673	0.0745	0.2916	0.3284
M7	0	0	0	0	0	0	0.0324	0.2326	0.2694
M8	0	0	0	0	0	0	0	0.2444	0.2812
M9	0	0	0	0	0	0	0	0	0.0520

	M11	M12	M13	M14	M15	M16	M17	M18	M19	M20
M1	1.1663	1.1886	1.2231	1.4522	1.5170	1.6097	1.7310	1.7996	1.8705	1.9322
M2	0.9888	1.0109	1.0455	1.2765	1.3413	1.4340	1.5553	1.6240	1.6948	1.7565
M3	0.8525	0.8746	0.9092	1.1406	1.2053	1.2981	1.4194	1.4880	1.5589	1.6205
M4	0.7220	0.7442	0.7788	1.0130	1.0777	1.1704	1.2915	1.3601	1.4310	1.4926
M5	0.6268	0.6490	0.6836	0.9200	0.9848	1.0774	1.1985	1.2671	1.3380	1.3997
M6	0.4328	0.4548	0.4896	0.7291	0.7939	0.8866	1.0078	1.0765	1.1474	1.2091
M7	0.3739	0.3961	0.4307	0.6714	0.7362	0.8289	0.9501	1.0188	1.0897	1.1514
M8	0.3824	0.4048	0.4393	0.6807	0.7455	0.8382	0.9594	1.0281	1.0990	1.1607
M9	0.1768	0.2120	0.2333	0.4623	0.5296	0.6232	0.7466	0.8159	0.8882	0.9497
M10	0.1388	0.1746	0.1954	0.4250	0.4923	0.5859	0.7093	0.7786	0.8509	0.9124
M11	0	0.0474	0.0809	0.3119	0.3792	0.4729	0.5956	0.6651	0.7376	0.7993
M12	0	0	0.0505	0.2863	0.3498	0.4439	0.5655	0.6340	0.7053	0.7673
M13	0	0	0	0.2527	0.3207	0.4144	0.5381	0.6077	0.6803	0.7419
M14	0	0	0	0	0.0771	0.1755	0.3094	0.3846	0.4651	0.5277
M15	0	0	0	0	0	0.1028	0.2370	0.3127	0.3937	0.4571
M16	0	0	0	0	0	0	0.1379	0.2145	0.2969	0.3592
M17	0	0	0	0	0	0	0	0.0811	0.1691	0.2323
M18	0	0	0	0	0	0	0	0	0.0903	0.1536
M19	0	0	0	0	0	0	0	0	0	0.0724

Cuadro C.13: *Media posterior de la distribución de distancias multipunto con 200 individuos, que proviene de una población F_2, con mapa menos denso, con marcadores codominantes y dominantes.*

	M2	M3	M4	M5	M6	M7	M8	M9	M10
M1	0.0340	0.0861	0.1062	0.1064	0.1207	0.1238	0.1279	0.1741	0.1655
M2	0	0.0403	0.0599	0.0546	0.0740	0.0776	0.0835	0.1431	0.1327
M3	0	0	0.0764	0.0694	0.0796	0.0798	0.0873	0.1417	0.1310
M4	0	0	0	0.0340	0.0574	0.0599	0.0684	0.1276	0.1162
M5	0	0	0	0	0.0411	0.0420	0.0524	0.1145	0.1022
M6	0	0	0	0	0	0.0249	0.0371	0.1082	0.0966
M7	0	0	0	0	0	0	0.0507	0.1044	0.0923
M8	0	0	0	0	0	0	0	0.1034	0.0926
M9	0	0	0	0	0	0	0	0	0.0095

	M11	M12	M13	M14	M15	M16	M17	M18	M19	M20
M1	0.1594	0.1584	0.1575	0.2021	0.1995	0.2019	0.2209	0.2342	0.2599	0.2740
M2	0.1214	0.1215	0.1191	0.1597	0.1557	0.1587	0.1815	0.1972	0.2270	0.2429
M3	0.1179	0.1178	0.1156	0.1533	0.1496	0.1522	0.1752	0.1913	0.2218	0.2382
M4	0.1030	0.1016	0.1014	0.1277	0.1231	0.1279	0.1572	0.1747	0.2076	0.2246
M5	0.0898	0.0919	0.0878	0.1097	0.1054	0.1110	0.1436	0.1636	0.1990	0.2166
M6	0.0821	0.0875	0.0836	0.1021	0.0985	0.1052	0.1405	0.1606	0.1961	0.2144
M7	0.0812	0.0890	0.0832	0.1015	0.0988	0.1061	0.1422	0.1625	0.1978	0.2160
M8	0.0872	0.0938	0.0877	0.1013	0.0987	0.1068	0.1430	0.1639	0.1999	0.2178
M9	0.0629	0.0718	0.0601	0.1128	0.1085	0.1146	0.1446	0.1633	0.1958	0.2133
M10	0.0681	0.0741	0.0585	0.1131	0.1070	0.1107	0.1392	0.1573	0.1896	0.2071
M11	0	0.0642	0.0401	0.1135	0.0982	0.0930	0.1140	0.1283	0.1596	0.1761
M12	0	0	0.0716	0.1121	0.0979	0.0891	0.1107	0.1277	0.1628	0.1779
M13	0	0	0	0.1295	0.1124	0.1041	0.1154	0.1251	0.1538	0.1690
M14	0	0	0	0	0.0326	0.0294	0.0483	0.0470	0.0685	0.0807
M15	0	0	0	0	0	0.0308	0.0475	0.0428	0.0635	0.0709
M16	0	0	0	0	0	0	0.0412	0.0277	0.0418	0.0524
M17	0	0	0	0	0	0	0	0.0184	0.0291	0.0313
M18	0	0	0	0	0	0	0	0	0.0277	0.0227
M19	0	0	0	0	0	0	0	0	0	0.0229

Cuadro C.14: *Desviación Típica posterior de la distribución de distancias multipunto con 200 individuos, que proviene de una población F_2, con mapa menos denso, con marcadores codominantes y dominantes.*

C.4. Población F_2, con mapa más denso, con marcadores codominantes y dominantes

	M2	M3	M4	M5	M6	M7	M8	M9	M10
M1	0.0344	0.0388	0.2625	0.2531	0.2723	0.2432	0.3828	0.3865	0.4251
M2	0	0.0089	0.2583	0.2506	0.2699	0.2431	0.3797	0.3834	0.4222
M3	0	0	0.2548	0.2472	0.2665	0.2397	0.3764	0.3800	0.4189
M4	0	0	0	0.0202	0.0365	0.0297	0.1281	0.1318	0.1692
M5	0	0	0	0	0.0276	0.0152	0.1357	0.1393	0.1765
M6	0	0	0	0	0	0.0304	0.1158	0.1194	0.1563
M7	0	0	0	0	0	0	0.1406	0.1442	0.1822
M8	0	0	0	0	0	0	0	0.0082	0.0464
M9	0	0	0	0	0	0	0	0	0.0432

	M11	M12	M13	M14	M15	M16	M17	M18	M19	M20
M1	0.4530	0.4949	0.5567	0.5543	0.5969	0.6684	0.7014	0.7033	0.7111	0.8017
M2	0.4491	0.4919	0.5539	0.5513	0.5943	0.6656	0.6986	0.7006	0.7083	0.7986
M3	0.4458	0.4886	0.5505	0.5480	0.5910	0.6624	0.6953	0.6973	0.7050	0.7953
M4	0.1962	0.2409	0.3019	0.2996	0.3416	0.4141	0.4468	0.4489	0.4568	0.5469
M5	0.2039	0.2484	0.3093	0.3069	0.3487	0.4214	0.4542	0.4561	0.4640	0.5544
M6	0.1835	0.2282	0.2890	0.2867	0.3283	0.4010	0.4338	0.4358	0.4437	0.5341
M7	0.2103	0.2529	0.3144	0.3124	0.3544	0.4265	0.4594	0.4613	0.4692	0.5597
M8	0.0796	0.1271	0.1864	0.1854	0.2274	0.3005	0.3330	0.3350	0.3431	0.4339
M9	0.0765	0.1238	0.1831	0.1820	0.2240	0.2971	0.3297	0.3317	0.3398	0.4306
M10	0.0343	0.0817	0.1410	0.1399	0.1818	0.2550	0.2875	0.2894	0.2976	0.3885
M11	0	0.0515	0.1111	0.1099	0.1520	0.2251	0.2577	0.2599	0.2680	0.3584
M12	0	0	0.0688	0.0616	0.1058	0.1782	0.2115	0.2134	0.2214	0.3125
M13	0	0	0	0.0139	0.0448	0.1183	0.1510	0.1526	0.1605	0.2511
M14	0	0	0	0	0.0446	0.1186	0.1524	0.1541	0.1621	0.2532
M15	0	0	0	0	0	0.0771	0.1112	0.1125	0.1205	0.2120
M16	0	0	0	0	0	0	0.0351	0.0369	0.0449	0.1371
M17	0	0	0	0	0	0	0	0.0066	0.0160	0.1024
M18	0	0	0	0	0	0	0	0	0.0116	0.1015
M19	0	0	0	0	0	0	0	0	0	0.0935

Cuadro C.15: *Media posterior de la distribución de distancias multipunto con 200 individuos, que proviene de una población F_2, con mapa más denso, con marcadores codominantes y dominantes.*

	M2	M3	M4	M5	M6	M7	M8	M9	M10
M1	0.0339	0.0335	0.0474	0.0488	0.0568	0.0542	0.0701	0.0704	0.0615
M2	0	0.0041	0.0429	0.0398	0.0478	0.0399	0.0642	0.0644	0.0532
M3	0	0	0.0429	0.0406	0.0486	0.0412	0.0642	0.0644	0.0531
M4	0	0	0	0.0196	0.0244	0.0380	0.0521	0.0523	0.0407
M5	0	0	0	0	0.0253	0.0309	0.0472	0.0473	0.0353
M6	0	0	0	0	0	0.0307	0.0502	0.0501	0.0367
M7	0	0	0	0	0	0	0.0491	0.0490	0.0364
M8	0	0	0	0	0	0	0	0.0056	0.0199
M9	0	0	0	0	0	0	0	0	0.0204

	M11	M12	M13	M14	M15	M16	M17	M18	M19	M20
M1	0.0578	0.0719	0.0745	0.0742	0.0770	0.0883	0.0927	0.0923	0.0943	0.1099
M2	0.0526	0.0637	0.0649	0.0654	0.0657	0.0788	0.0835	0.0828	0.0854	0.1031
M3	0.0526	0.0633	0.0648	0.0650	0.0655	0.0784	0.0832	0.0825	0.0851	0.1028
M4	0.0361	0.0389	0.0426	0.0420	0.0487	0.0596	0.0666	0.0658	0.0679	0.0900
M5	0.0303	0.0357	0.0396	0.0396	0.0478	0.0588	0.0659	0.0656	0.0675	0.0891
M6	0.0310	0.0350	0.0388	0.0381	0.0468	0.0577	0.0647	0.0643	0.0662	0.0878
M7	0.0290	0.0407	0.0412	0.0395	0.0466	0.0595	0.0660	0.0656	0.0676	0.0882
M8	0.0259	0.0324	0.0339	0.0317	0.0363	0.0459	0.0531	0.0525	0.0539	0.0736
M9	0.0265	0.0325	0.0334	0.0317	0.0359	0.0454	0.0523	0.0517	0.0531	0.0725
M10	0.0200	0.0282	0.0279	0.0260	0.0298	0.0406	0.0485	0.0479	0.0493	0.0690
M11	0	0.0299	0.0265	0.0255	0.0268	0.0368	0.0439	0.0423	0.0436	0.0655
M12	0	0	0.0289	0.0268	0.0299	0.0339	0.0393	0.0379	0.0400	0.0580
M13	0	0	0	0.0286	0.0212	0.0287	0.0340	0.0343	0.0352	0.0553
M14	0	0	0	0	0.0215	0.0279	0.0290	0.0289	0.0311	0.0488
M15	0	0	0	0	0	0.0281	0.0267	0.0277	0.0300	0.0427
M16	0	0	0	0	0	0	0.0160	0.0134	0.0116	0.0258
M17	0	0	0	0	0	0	0	0.0083	0.0120	0.0225
M18	0	0	0	0	0	0	0	0	0.0098	0.0175
M19	0	0	0	0	0	0	0	0	0	0.0197

Cuadro C.16: *Desviación Típica posterior de la distribución de distancias multipunto con 200 individuos, que proviene de una población F_2, con mapa más denso, con marcadores codominantes y dominantes.*

Apéndice D

Apéndice del Capítulo 4

D.1. Cálculo de $\pi(O_k|R)$ por integración numérica, mediante cuadratura gaussiana

Los pasos básicos para resolver la doble integral 4.11, aproximando por integración numérica, mediante cuadratura gaussiana, son los siguientes:

Considerando como objetivo calcular la integral

$$\int_{r_{k_1}}\int_{r_{k_2}} h(r_{k_1}, r_{k_2})dr_{k_1}dr_{k_2} = \int_{a=0}^{b=0.5}\int_{c=0}^{d=0.5} h(r_{k_1}, r_{k_2})dr_{k_1}dr_{k_2}, \quad \text{siendo}$$

$h(r_{k_1}, r_{k_2}) = Multinomial(R_{k_1}|n; \rho_{k_1}(r_{k_1})) * Multinomial(R_{k_2}|n; \rho_{k_2}(r_{k_2}))*$
$Multinomial(R_{k_3}|n; \rho_{k_3}(g(r_{k_1}, r_{k_2})))$

Para poder utilizar los polinomios de Legendre, es necesario hacer un cambio de variable, de forma que los límites de integración varíen entre (-1,1):

$$\int_{a=0}^{b=0.5}\int_{c=0}^{d=0.5} h(r_{k_1}, r_{k_2})dr_{k_1}dr_{k_2} =$$

$$= (\frac{b-a}{2})(\frac{d-c}{2})\int_{-1}^{1}\int_{-1}^{1} h(\frac{(b-a)r_{k_1}+(b+a)}{2}, \frac{(d-c)r_{k_2}+(d+c)}{2})dr_{k_1}dr_{k_2}.$$

Considerando

$$h1 = \frac{b-a}{2},\ h2 = \frac{b+a}{2},\ k1 = \frac{d-c}{2},\ k2 = \frac{d-c}{2}$$

la integral tiene la siguiente expresión:

$$h1 * k1 \int_{-1}^{1} \int_{-1}^{1} h((h1r_{k_1} + h2), (k1r_{k_2} + k2)) dr_{k_1} dr_{k_2}.$$

Considerando x_{wi} y c_{wi} las raíces y pesos del polinomio Legendre de grado w, respectivamente:

$$h1 * k1 \int_{-1}^{1} \int_{-1}^{1} h((h1r_{k_1} + h2), (k1r_{k_2} + k2)) dr_{k_1} dr_{k_2} =$$

$$= h1 * k1 \sum_{i=1}^{m} c_{mi} \sum_{j=1}^{n} c_{nj}\ h((h1x_{mi} + h2), (k1x_{nj} + k2))$$

Apéndice E

Apéndice del Capítulo 9

E.1. Influencia de la tolerancia del algoritmo EM

E.1.1. Algoritmo EM

Como se ha comentado en el Capítulo 1, el algoritmo EM es una propuesta iterativa que se utiliza, en presencia de datos faltantes, para obtener el estimador máximo verosímil. En el estudio de mapas genéticos existen incertidumbres en los datos que pueden tratarse como datos faltantes. Así, en el caso de un doble heterocigoto, $AaBb$, no se conoce qué genotipo concreto es recombinante pues puede proceder de la unión de un gameto AB con otro ab, que no son recombinantes, o alternativamente, de la unión de Ab con aB en donde ambos gametos lo son. Por otro lado, en marcadores dominantes el genotipo $A\_$ no se sabe si es AA o Aa. Por ello, el algoritmo EM es empleado ampliamente en este tipo de datos.

El algoritmo EM suele estar implementado, con enfoque frecuentista, en algunos programas informáticos (Mapmaker [48] (Lander et al. 1987 [45])) para el cálculo de mapas genéticos. Uno de los parámetros que es necesario aportar al programa es la tolerancia que actúa como elemento de parada en las sucesivas iteraciones. Sin embargo, no se suelen encontrar indicaciones de los valores recomendables a emplear. A continuación, se evalúa la influencia de la tolerancia definida en el algoritmo EM en su utilidad para la estimación de fracciones de recombinación entre pares de marcadores genéticos dominantes

y codominantes y se establecen recomendaciones sobre la tolerancia óptima en función de la distancia genética y el tipo de marcadores.

Adecuando el algoritmo para la estima de fracciones de recombinación entre marcadores, este costa de los siguientes pasos:

1. Inicializar r

2.**Esperanza**: Cálculo del valor esperado de recombinantes, suponiendo cierta la fracción de recombinación r.

3.**Maximización**: Cálculo de la fracción de recombinación máximo verosímil, rnueva, dados los valores esperados anteriores:

$$rnueva = \sum_{i=1}^{ngeno} P(G_i) * P(Rc|G_i)/n \tag{E.1}$$

donde Rc significa recombinante, G_i cada uno de los genotipos posibles, $ngeno$ el número de genotipos y n el tamaño muestral.

4. Condición de parada:|r-rnueva| <**tolerancia**. Iterar volviendo al paso 1. en caso de que no se cumpla la condición de parada.

Nótese que si tenemos una secuencia genética que consta de m marcadores, deberemos estimar $m(m\text{-}1)/2$ fracciones de recombinación, o lo que es lo mismo, ejecutar el algoritmo EM ese mismo número de veces.

E.1.2. Metodología

En la ejecución del algoritmo es imprescindible distinguir todas las posibilidades según el carácter codominante o dominante, de los dos marcadores que estudiamos en cada momento. Recordamos que un marcador es codominante (C), si los dos alelos que lo forman son observables. Es decir, AA, Aa, aa, y es dominante si alguno de sus alelos no es observable. Dependiendo del parental del que provenga distinguiremos dominante (D1), con genotipos observables $A\_$, aa y dominante (D2), con genotipos observables AA, $a\_$.

Consideremos la estimación de la fracción de recombinación entre pares de marcadores en los que, al menos uno de ellos es dominante, que es el caso en el que es posible que aparezcan problemas computacionales: D1-D1, D1-D2, D2-D2, C-D1 y C-D2. Para cada pareja, se ensayaron 4 distancias de separación (1, 2, 5 y 10 centimorgan) y 4 tolerancias (10^{-3}, 10^{-4}, 10^{-5} y 10^{-6}), obteniendo así 80 combinaciones. Se calculó la distribución en el muestreo del estimador que proporciona el algoritmo EM, basada en una muestra de tamaño 5000,

obtenida mediante simulación de repartos multinomiales de generaciones de 200 individuos. Las probabilidades condicionadas para la recombinación dado el genotipo observado, $P(Rc|G_i)$, así como la frecuencia relativa esperada de cada uno de los individuos genotípicamente observables, $P(G_i)$, se extraen del Cuadro E.1 que aparece a continuación, en función de cuál sea el tipo de los marcadores involucrados.

C-D1	$P(G_i)$	$P(Rc\|G_i)$
$AAB\_$	$0.25(1-r^2)$	$\frac{r}{1+r}$
$AAbb$	$0.25r^2$	1
$AaB\_$	$0.5(1-r+r^2)$	$\frac{r(1+r)}{2(1-r+r^2)}$
$Aabb$	$0.5r(1-r)$	$\frac{1}{2}$
$aaB\_$	$0.25r(2-r)$	$\frac{1}{2-r}$
$aabb$	$0.25(1-r)^2$	0

C-D2	$P(G_i)$	$P(Rc\|G_i)$
$AABB$	$0.25(1-r)^2$	0
$AAb\_$	$0.25r(2-r)$	$\frac{1}{2-r}$
$AaBB$	$0.5r(1-r)$	$\frac{1}{2}$
$Aab\_$	$0.5(1-r+r^2)$	$\frac{r(1+r)}{2(1-r+r^2)}$
$aaBB$	$0.25r^2$	1
$aab\_$	$0.25(1-r^2)$	$\frac{r}{1+r}$

D1-D1	$P(G_i)$	$P(Rc\|G_i)$
$A\_B\_$	$0.25(3-2r+r^2)$	$\frac{2r}{3-2r+r^2}$
A_bb	$0.25r(2-r)$	$\frac{1}{2-r}$
$aaB\_$	$0.25r(2-r)$	$\frac{1}{2-r}$
$aabb$	$0.25(1-r)^2$	0

D1-D2	$P(G_i)$	$P(Rc\|G_i)$
A_BB	$0.25(1-r^2)$	$\frac{r}{1+r}$
$A\_b\_$	$0.25r(2+r^2)$	$\frac{2r+r^2}{2+r^2}$
$aaBB$	$0.25r^2$	1
$aab\_$	$0.25(1-r^2)$	$\frac{r}{1+r}$

D2-D2	$P(G_i)$	$P(Rc\|G_i)$
$AABB$	$0.25(1-r)^2$	0
$AAb\_$	$0.25r(2-r)$	$\frac{1}{2-r}$
a_BB	$0.25r(2-r)$	$\frac{1}{2-r}$
$a\_b\_$	$0.25(3-2r+r^2)$	$\frac{2r}{3-2r+r^2}$

Cuadro E.1: Probabilidades condicionadas de recombinación dado un genotipo y frecuencias esperadas de cada uno de los genotipos.

E.1.3. Resultados y discusión

Los resultados se presentan en las Figuras de la E.1 a la E.4. En general, las estimas de las fracciones de recombinación resultan poco sensibles, tanto respecto a las variaciones en la tolerancia empleada, como a la distancia entre marcadores, aun en el caso en que dichas distancias sean pequeñas. De hecho, una tolerancia de 10^{-3} es suficiente (valor utilizado por defecto según especifica el manual de Mapmaker [48]). Sin embargo, para parejas de marcadores del tipo D1 – D2 el valor de la tolerancia tiene una gran influencia en la estimación de la fracción de recombinación incluso a distancias moderadas entre marcadores. La tolerancia de 10^{-3}, que hemos recomendado para la mayor parte de los casos, puede multiplicar por más de 6 el verdadero valor de la fracción de recombinación si la distancia entre marcadores es pequeña. En este caso, si se emplean tolerancias del orden de 10^{-6}, la estimación de la fracción de recombinación tiende a cero. Los resultados parecen indicar que a distancias pequeñas, entre marcadores, sería recomendable utilizar tolerancias algo mayores a 10^{-5} mientras que a distancias iguales o superiores a 10 cM sería suficiente tomar como tolerancia 10^{-3} o superior.

En definitiva, concluimos que en general la recomendación es emplear tolerancias de 10^{-3}. Sin embargo, para el caso de marcadores dominantes en repulsión, es necesario hacer una estima previa de modo que si la fracción de recombinación estimada es menor que 10 cM, es conveniente modificar la tolerancia a 10^{-4} o como máximo a 10^{-5} y revisar el valor de la estima de la fracción de recombinación.

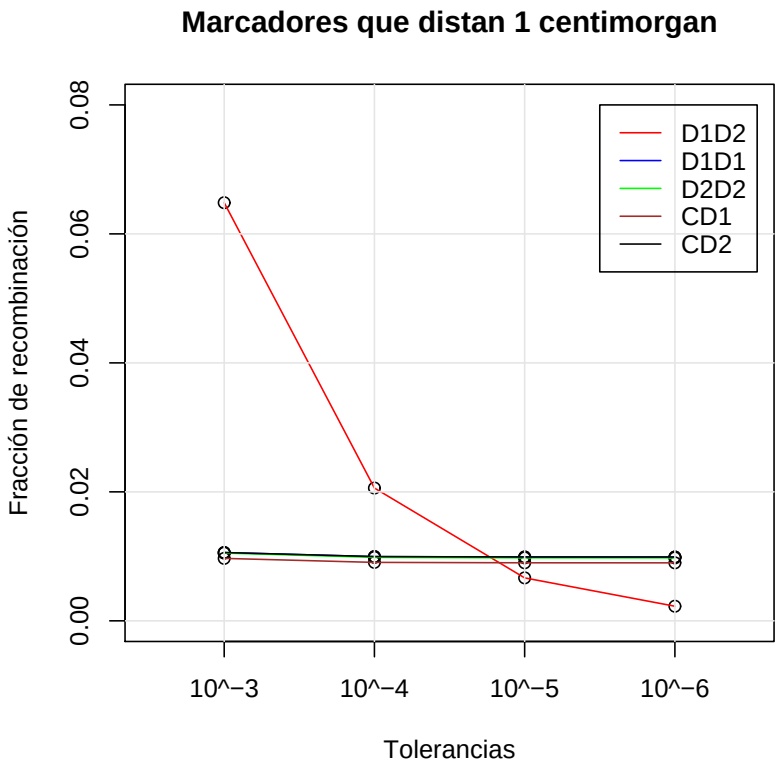


Figura E.1: Algoritmo EM. Marcadores que distan 1 centimorgan.

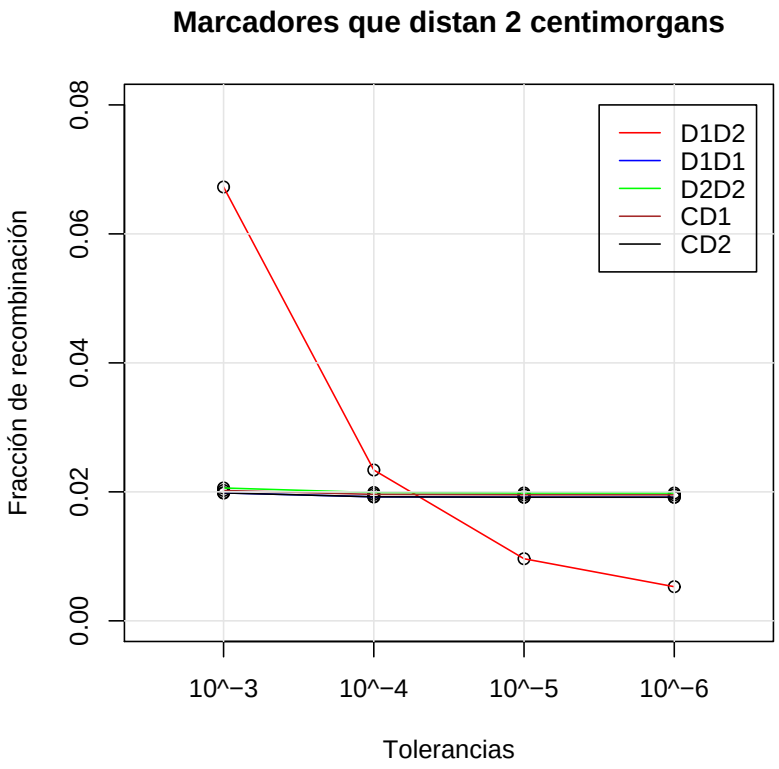


Figura E.2: Algoritmo EM. Marcadores que distan 2 centimorgan.

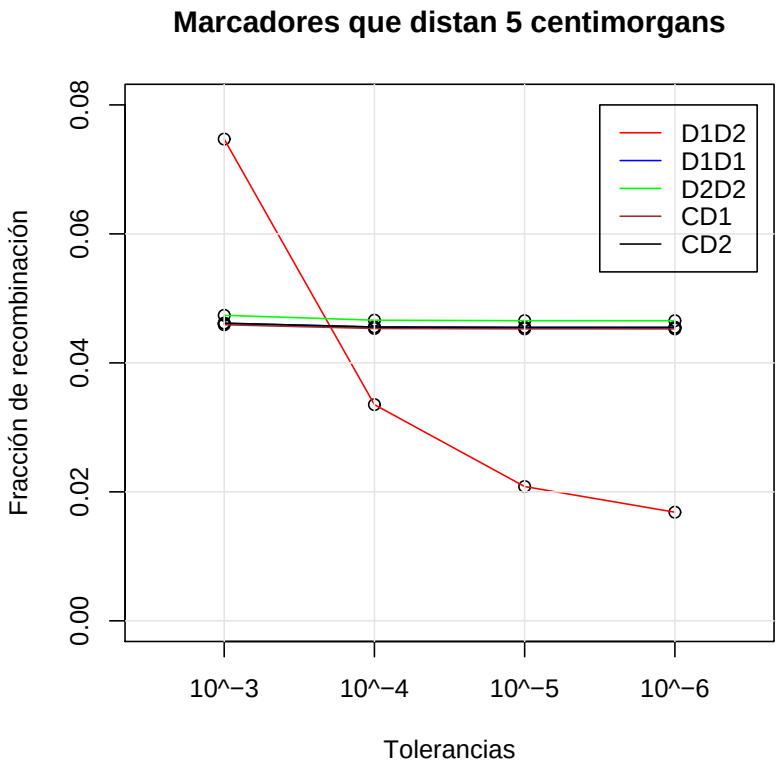


Figura E.3: Algoritmo EM. Marcadores que distan 5 centimorgan.

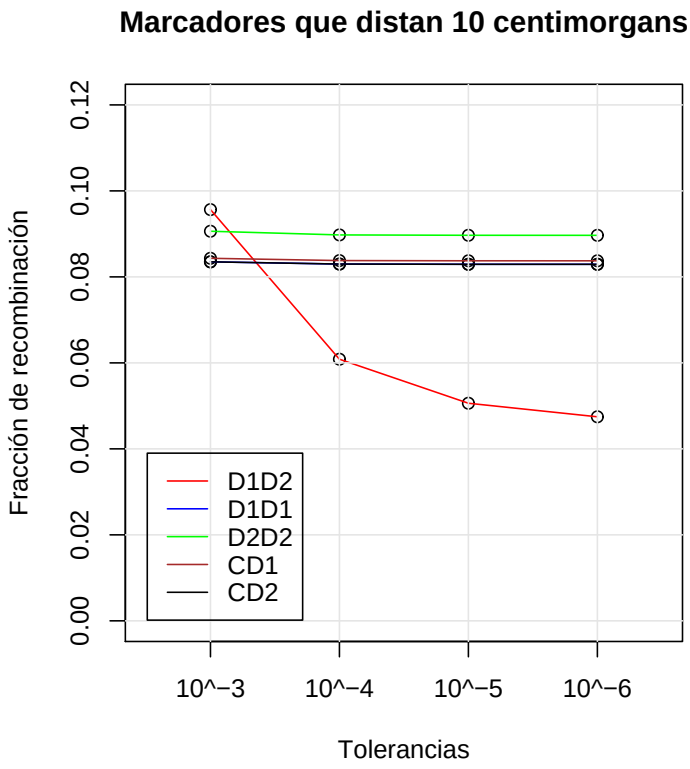


Figura E.4: Algoritmo EM. Marcadores que distan 10 centimorgan.

Bibliografía

[1] Aarts, E., y Korst, J. *Simulated Annealing and Boltzman Machines*. John Wiley and Son, 1989.

[2] Ahfock, D., Wood, I., Stephen, S., Cavanagh, C. R., y Huang, B. E. Characterizing uncertainty in high-density maps from multiparental populations. *Genetics 198*, 1 (2014), 117–128.

[3] Albert, J. *LearnBayes: Functions for Learning Bayesian Inference*, 2012. R package version 2.12.

[4] Beaumont, M. A., y Rannala, B. The bayesian revolution in genetics. *Nature Reviews Genetics 5*, 4 (2004), 251–261.

[5] Benjamini, Y., y Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. 57* (1995), 289–300.

[6] Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., Evers, D. J., Barnes, C. L., Bignell, H. R., et al. Accurate whole human genome sequencing using reversible terminator chemistry. *nature 456*, 7218 (2008), 53–59.

[7] Bernet, G. P., Fernandez-Ribacoba, J., Carbonell, E. A., y Asins, M. J. Comparative genome-wide segregation analysis and map construction using a reciprocal cross design to facilitate citrus germplasm utilization. *Molecular breeding 25*, 4 (2010), 659–673.

[8] Bink, M., Jansen, J., Madduri, M., Voorrips, R., Durel, C.-E., Kouassi, A., Laurens, F., Mathis, F., Gessler, C., Gobbin, D.,

ET AL. Bayesian qtl analyses using pedigreed families of an outcrossing species, with application to fruit firmness in apple. *Theoretical and applied genetics 127*, 5 (2014), 1073–1090.

[9] BLASCO, A. The bayesian controversy in animal breeding. *Journal of animal science 79*, 8 (2001), 2023–2046.

[10] BLASCO, A. The use of bayesian statistics in meat quality analyses: a review. *Meat science 69*, 1 (2005), 115–122.

[11] BLASCO, A. La significación es irrelevante y los p-valores engañosos.¿ qué hacer?, 2011.

[12] BOTSTEIN, D., WHITE, R., SKOLNICK, M., Y DAVIS, R. Construction of a genetic linkage map in man using restriction fragment length polyorphisms. *Am J Hum Genet. 32* (1980), 314–331.

[13] BROMAN, K. W., WU, H., SEN, Ś., Y CHURCHILL, G. A. R/qtl: Qtl mapping in experimental crosses. *Bioinformatics 19*, 7 (2003), 889–890.

[14] BUETOW, K. Multipoint gene mapping using seriation. ii. analysis of simulated and empirical data. *Am J Hum Genet. 41* (1987), 189–201.

[15] BUETOW, K., Y CHAKRAVARTI, A. Multipoint gene mapping using seriation i general methods. *Am J Hum Genet. 41* (1987), 180–188.

[16] BYRD, R., LU, P., Y AD C. ZHU, J. N. A limited memory algorithm for bound constrained optimization. *SIAM J. Scientific Computing 16* (1995), 1190–1208.

[17] DEMPSTER, A., LAIRD, N., Y RUBIN, D. Maximum likelihood from incomplete data via the em algorithm. *JR Statist. Soc. B 39* (1977), 1–38.

[18] DESCHAMPS, S., LLACA, V., Y MAY, G. D. Genotyping-by-sequencing in plants. *Biology 1*, 3 (2012), 460–483.

[19] EDWARDS, J. The locus ordering problem. *Ann Hum Genet 51* (1987), 253–275.

[20] Falk, C. Construction of multilocus genetic linkage maps in human. *Proc Natl Acad Sci USA 84* (1987), 2363–2367.

[21] Foolad, M. R., y Panthee, D. R. Marker-assisted selection in tomato breeding. *Critical reviews in plant sciences 31*, 2 (2012), 93–123.

[22] Gelman, A., Carlin, J. B., Stern, H. S., y Rubin, D. B. *Bayesian data analysis*, vol. 2. Taylor & Francis, 2014.

[23] George, A., Mengersen, K., y Davis, G. A bayesian approach to ordering gene markers. *Biometrics 55* (1999), 419–429.

[24] George, A. W. A novel markov chain monte carlo approach for constructing accurate meiotic maps. *Genetics 171*, 2 (2005), 791–801.

[25] Goldberg, D. *Genetic Algorithms in Search Optimization and Machine Learning*. Addison-Wesley, 1989.

[26] Golding, B., Hühn, M., y Piepho, H.-P. A note on the bias of genetic distances in linkage maps based on small samples for backcrosses and intercrosses with complete dominance. *Genome 51*, 12 (2008), 1054–1061.

[27] Grattapaglia, D., y Sederoff, R. Genetic linkage maps of eucalyptus grandis and eucalyptus urophylla using a pseudo-testcross: mapping strategy and rapd markers. *Genetics 137*, 4 (1994), 1121–1137.

[28] Green, P. J. Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika 82*, 4 (1995), 711–732.

[29] Gupta, P. K., Langridge, P., y Mir, R. R. Marker-assisted wheat breeding: present status and future possibilities. *Molecular Breeding 26*, 2 (2010), 145–161.

[30] He, J., Zhao, X., Laroche, A., Lu, Z.-X., Liu, H., y Li, Z. Genotyping-by-sequencing (gbs), an ultimate marker-assisted selection (mas) tool to accelerate plant breeding. *Frontiers in plant science 5* (2014).

[31] Holland, J. *Adaptation in Natural and Artificial Systems.* University of Michigan Press, 1975.

[32] Hospital, F. Challenges for effective marker-assisted selection in plants. *Genetica 136* (2009), 303–310.

[33] Jansen, J. Ordering dominant markers in f2 populations. *Euphytica 165*, 2 (2009), 401–417.

[34] Jansen, J., de Jong, A., y Ooijen, J. V. Constructing dense genetic linkage maps. *Theoretical and Applied Genetics 102* (2001), 1113–1122.

[35] Jensen, J., y Jorgensen, J. The barley chromosome 5 linkage map. *Hereditas 80* (1975), 5–16.

[36] Jensen, J., y Jorgensen, J. The barley chromosome 5 linkage map ii. *Hereditas 80* (1975), 17–26.

[37] Keats, B., Sherman, S., Morton, N., Robson, E., Buetow, K., Cann, H., Cartwright, P., Chakravarti, A., Franvke, U., Green, P., y Ott, J. Guideline for human linkage maps: an international system for human linkage maps (islm 1990). *Genomics 9* (1991), 557–560.

[38] Kirkpatrick, S., Gelatt, C., y Vecchi, M. Optimization by simulated annealing. *Science. 220* (1983), 671–680.

[39] Knapp, S., Holloway, J., Bridges, W., y Liu, B. Mapping dominant markers using f2 matings. *Theoretical and applied genetics 91*, 1 (1995), 74–81.

[40] Kumar, J., Choudhary, A. K., Solanki, R. K., y Pratap, A. Towards marker-assisted selection in pulses: a review. *Plant Breeding 130*, 3 (2011), 297–313.

[41] Lalouel, J. Linkage mapping from pair-wise recombination data. *Heredity 38* (1977), 61–77.

[42] Lande, R., y Thompson, R. Efficiency of marker-assisted selection in the improvement of quantitative traits. *Genetics 124*, 3 (1990), 743–756.

[43] Lander, E., y Green, P. Construction of multilocus genetic linkage maps in human. *Proc Natl Acad Sci USA 84* (1987), 2363–2367.

[44] Lander, E., y Green, P. Counting algorithms for linkage: correction to morton and collins. *Ann Hum Genet 55* (1991), 33–38.

[45] Lander, E., Green, P., Abrahamson, J., Barlow, A., Daly, M., Lincoln, S., y Newburg, L. Mapmaker: An interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. *Genomics 1* (2005), 174–181.

[46] Lander, E. S., y Green, P. Construction of multilocus genetic linkage maps in humans. *Proceedings of the National Academy of Sciences 84*, 8 (1987), 2363–2367.

[47] Lathrop, G., Chotai, J., Ott, J., y Lalouel, J. Tests of gene order from three-locus linkage data. *Ann Hum Genet 51* (1987), 235–249.

[48] Lincoln, S. E., Daly, M. J., y Lander, E. S. *Mapmaker 3.0.: Linkage analysis package designed to help construct primary linkage maps of markers segregating in both experimental crosses as well as simple natural populations*. Whitehead Institute for Biomedical Research and the M.I.T Center for Genome Research, Cambridge, Massachusetts.

[49] Liu, B. *Statistical genomics: linkage mapping and QTL analysis*. CRC Press LLC, 1998.

[50] Lunn, D. J., Best, N., y Whittaker, J. C. Generic reversible jump mcmc using graphical models. *Statistics and Computing 19*, 4 (2009), 395–408.

[51] Lunn, D. J., Thomas, A., Best, N., y Spiegelhalter, D. Winbugs-a bayesian modelling framework: concepts, structure, and extensibility. *Statistics and computing 10*, 4 (2000), 325–337.

[52] Martin, O. C., et al. Two-and three-locus tests for linkage analysis using recombinant inbred lines. *Genetics 173*, 1 (2006), 451–459.

[53] Mester, D., Ronin, Y., Hu, Y., Peng, J., Nevo, E., y Korol, A. Efficient multipoint mapping: making use of dominant repulsion-phase markers. *Theoretical and Applied Genetics 107*, 6 (2003), 1102–1112.

[54] Miller, D., y Pekny, J. Exact solution of large asymmetric traveling salesman problems. *Science 251* (1991), 754–761.

[55] Muller, P., Parmigiani, G., y Rice, K. Fdr and bayesian multiple comparisons rules. *Proc. Valencia / ISBA 8th World Meeting on Bayesian Statistics* (2006).

[56] Neal, R. M. Slice sampling. *Annals of statistics* (2003), 705–741.

[57] Olson, J., y Boehnke, M. Monte carlo comparison of preliminary methods for ordering multiple genetic loci. *Am J Hum Genet. 47* (1990), 470–482.

[58] Ott, J. Counting methods (em algorithm) in human pedigree analysis: Linkage and segregation analysis. *Ann Hum Genet 40* (1977), 443–454.

[59] Ott, J. *Analysis of Human Genetic Linkage*. The Johns Hopkins University Press, 1999.

[60] Plummer, M., Best, N., Cowles, K., y Vines, K. Coda: Convergence diagnosis and output analysis for mcmc. *R News 6*, 1 (2006), 7–11.

[61] Quail, M. A., Smith, M., Coupland, P., Otto, T. D., Harris, S. R., Connor, T. R., Bertoni, A., Swerdlow, H. P., y Gu, Y. A tale of three next generation sequencing platforms: comparison of ion torrent, pacific biosciences and illumina miseq sequencers. *BMC genomics 13*, 1 (2012), 341.

[62] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2012. ISBN 3-900051-07-0.

[63] Raga, V., Bernet, G. P., Carbonell, E. A., y Asins, M. J. Segregation and linkage analyses in two complex populations derived from

the citrus rootstock cleopatra mandarin. inheritance of seed reproductive traits. *Tree Genetics and Genomes 8*, 5 (2012), 1061–1071.

[64] Robert, C., y Casella, G. *Monte Carlo statistical methods*. Springer Science & Business Media, 2013.

[65] Rogatko, A., y Zacks, S. Ordering genes: Controlling the decision-error probabilities. *American Journal of Human Genetics 52* (1993), 947–957.

[66] Rosa, G., Yandell, B., y Gianola, D. A bayesian approach for constructing genetic maps when markers are miscoded. *Genet. Sel. Evol. 34* (2002), 353–369.

[67] Rothberg, J. M., Hinz, W., Rearick, T. M., Schultz, J., Mileski, W., Davey, M., Leamon, J. H., Johnson, K., Milgrew, M. J., Edwards, M., et al. An integrated semiconductor device enabling non-optical genome sequencing. *Nature 475*, 7356 (2011), 348–352.

[68] Ruiz, C., y Asins, M. Comparison between poncirus and citrus genetic linkage maps. *Theoretical and Applied Genetics 106*, 5 (2003), 826–836.

[69] Säll, T., y Nilsson, N.-O. The robustness of recombination frequency estimates in intercrosses with dominant markers. *Genetics 137*, 2 (1994), 589–596.

[70] Shendure, J., y Ji, H. Next-generation dna sequencing. *Nature biotechnology 26*, 10 (2008), 1135–1145.

[71] Shoemaker, J. S., Painter, I. S., y Weir, B. S. Bayesian statistics in genetics: a guide for the uninitiated. *Trends in Genetics 15*, 9 (1999), 354–358.

[72] Sillanpää, M. J., y Arjas, E. Bayesian mapping of multiple quantitative trait loci from incomplete inbred line cross data. *Genetics 148*, 3 (1998), 1373–1388.

[73] Smith, C. Probabilities of orders in linkage calculations. *Amm Hum Genet 54* (1990), 339–363.

[74] Sorensen, D., y Gianola, D. *Likelihood, Bayesian, and MCMC methods in quantitative genetics*. Springer Science & Business Media, 2002.

[75] Spiegelhalter, D., Thomas, A., Best, N., y Lunn, D. Openbugs user manual, version 3.0. 2. *MRC Biostatistics Unit, Cambridge* (2007).

[76] Stam, P. Construction of integrated genetic linkage maps by means of a new computer package: Joinmap. *The Plant Journal 3* (1993), 739–744.

[77] Stam, P., y Van Ooijen, J. Joinmap tm version 2.0: software for the calculation of genetic maps. *CPRODLO, Wageningen* (1995).

[78] Stephens, D., y Smith, A. Bayesian inference in multipoint gene mapping. *Annals of Human Genetics 57* (1993), 65–82.

[79] Sturtevant, A. The linear arrangement of six sex-linked factors in drosophila as shown by their mode of association. *J Exp Zool 14* (1913), 43–59.

[80] Tan, Y.-D., y Fu, Y.-X. A new strategy for estimating recombination fractions between dominant markers from an f2 population. *Genetics 175*, 2 (2007), 923–931.

[81] Tanner, M. *Tools for Statistical Inference*. Springer-Verlag,NY, 1996.

[82] Thompson, E. Crossover counts and likelihood in multipoint linkage analysis. *IMA Journal of Mathematics Applied in Medicine and biology 4* (1987), 98–108.

[83] Thudi, M., Li, Y., Jackson, S. A., May, G. D., y Varshney, R. K. Current state-of-art of sequencing technologies for plant genomics research. *Briefings in Functional Genomics 11*, 1 (2012), 3–11.

[84] Van den Berg, I., Fritz, S., y Boichard, D. Qtl fine mapping with bayes c(π): a simulation study. *Genetics Selection Evolution 45*, 19 (2013), 10.1186.

[85] Van Ooijen, J. Joinmap 4. *Software for the calculation of genetic linkage maps in experimental populations. Kyazma BV, Wageningen, Netherlands* (2006).

[86] Van Ooijen, J., y Voorrips, R. *JoinMap 3.0.: Software for the calculation of genetic linkage maps*. Plant Research International B.V., Wageningen, The Netherlands, 2001.

[87] Weeks, D., y Lange, K. Preliminary ranking procedures for multilocus ordering. *Genomics. 1* (1987), 236–242.

[88] Weeks, D., Lathrop, G., y Ott, J. Multipoint mapping under genetic interference. *Hum Hered 43* (1993), 86–97.

[89] Weir, B. *Genetic Data Analysis II*. Sinauer, 1996.

[90] Wilson, S. A major simplification in the preliminary ordering of linked loci. *Gent. Epidemiol. 5* (1988), 75–80.

[91] York, T., Durrett, R., Tanksley, S., y Nielsen, R. Bayesian and maximum likelihood estimation of genetic maps. *Genetical Research 85* (2005), 159–168.

[92] Yue, G. H. Recent advances of genome mapping and marker-assisted selection in aquaculture. *Fish and Fisheries 15*, 3 (2014), 376–396.

VNIVERSITAT DE VALÈNCIA