eman ta zabal zazu

UPV EHU

Presiones selectivas en la región HLA

David Sánchez Cuenca

2019

eman ta zabal zazu
Universidad del País Vasco
Euskal Herriko Unibertsitatea
ZIENTZIA ETA TEKNOLOGIA FAKULTATEA
FACULTAD DE CIENCIA Y TECNOLOGÍA

Presiones selectivas en la región HLA

Memoria de Tesis Doctoral dirigida por:

Dr. José Ángel Peña García

para la obtención del grado “Doctor en Biología” de:

David Sánchez Cuenca

Departamento de Genética, Antropología Física y Fisiología Animal

Leioa, 2019

AGRADECIMIENTOS

En primer lugar, mi más profundo agradecimiento al Dr. Jose Ángel Peña García, primero por aceptarme como doctorando y durante estos largos 5 años por su infinita paciencia, ayuda y colaboración como director, sin las que esta tesis doctoral nunca habría podido ser empezada, ni mucho menos acabada.

A Manuel Quintero, tanto por su ayuda inestimable en los análisis de las cuasiinfinitas hojas de Excel como por las napolitanas de chocolate que tantas neuronas ayudaron a hacer funcionar.

A Leire, porque es la persona más cercana que tengo que sabe cuan jodido es esto de hacer una tesis, porque sabe la frustración que se siente a veces, porque siempre está ahí con un abrazo (o unas vacaciones) cuando es necesario, porque has creído que esta tesis era posible aun cuando ni yo lo he tenido claro, porque la quiero y por tantos porque que es imposible enumerarlos todos.

A la gente del departamento de Genética, Antropología Física y Fisiología Animal que me ha echado un cable cuando lo he necesitado.

A mi madre y mi hermana. Sé que han sido 5 años muy difíciles, y que mi estado anímico no siempre ha sido el mejor, pero ellas siempre han sabido perdonarme por eso. Os quiero.

A mi abuelo, mis tíos, mis tías, mis primas y mis primos por estar ahí cuando se les ha necesitado.

A Vieju, Maiki, Oscar, Jokin, Lenny, Marta, Maider, Bea, María, Mirella, Maitane, Iskander, Jose y Lucía. Porque La Familia es, y siempre será, la familia.

A Casti, Gon, JC, Lexuri, López y Marina, por los maxibones, flashes y porque sin ellos la vida sería mucho más aburrida y monótona.

A Aitziber, Carlos, Humberto, Aitor y Oier, porque me habéis dado muy buenos momentos.

A la gente de Patos en una rotonda, porque nos vemos 3 días al año, ¡¡¡pero que 3 días!!!.

A la gente de LDCP, FON (¡¡¡SALVE EMPERADOR!!!) y Port Royal porque son unos cachondos irredentos y lo mejor-peor de cada casa. Mención especial para Tío

Parches por crear los 11 meses (y subiendo) de emocionante contenido que FDEV no ha sido capaz de crear.

A la Fundación Jesús de Gangoiti Barrera, por la beca que me dieron el primer año de esta tesis.

A Jinkros, Stoyka, Carlton, Viruzz, Pedro, Ezagututa, Wolfie, DarkSkeletor y Albóndigas por sus inestimables conocimientos informáticos y chorradas varias.

ÍNDICE

ÍNDICE

INTRODUCCIÓN 1
Conceptos básicos 3
Diversidad genética 6
Causas de las diferencias entre individuos 7
Intercambio génico 8
Selección natural 11
Tipos de selección natural 12
Importancia de la selección natural 13
Deriva génica 14
Principio de Hardy-Weinberg 15
Deriva y selección 16
Mecanismos de actuación de la deriva génica 16
Bases de datos en el estudio de la variación genética 19
Tipos de bases de datos 20
Complejo Mayor de Histocompatibilidad (CMH) 23
Organización del CMH 23
Funciones del CMH 24
Emparejamiento selectivo y CMH 25
Proceso de colonización de *Homo sapiens* 27
Orígenes africanos de *Homo sapiens* 28
Teoría de la "ruta de la costa" 28
Dispersión temprana por el norte de África 29
Datación: pre- o post-Toba 30
Entrecruzamiento entre humanos arcaicos y modernos 30
Introgresión 31
Situación de la población gitana en el País Vasco 40

HIPÓTESIS Y OBJETIVOS 41

MATERIALES Y MÉTODOS 45
Bases de datos 47
1000Genomes 47
HapMap 53
Genoma Neandertal 53
Programas para minería de datos 54
FstMap 55
MEGA 75
Haploview 81
Muestras de ADN 81
Análisis genéticos 83
SNPs analizados 83
Análisis estadísticos 84
Test Chi-2 84
Test exacto de Fisher 85
Tablas de contingencia 85
Análisis de componentes principales 85

RESULTADOS 87
Análisis de control de datos 89
Desequilibrio de Hardy-Weinberg 95
Diferencias entre genes HLA 95
Diferencias entre grupos continentales 108
Diferencias intracontinentales 111
Poblaciones desplazadas 113
Análisis de Componentes Principales 114
Varianza de Wahlund 116
Diferencias en la varianza de Wahlund 117
Relación entre desequilibrio Hardy-Weinberg y varianza de Wahlund 118
Marcadores de ancestralidad 137
Diferencias en los procesos de selección 138
Análisis de una muestra de SNPs en Gitanos del País Vasco 154

Relación entre genes HLA de Neandertal y humanos anatómicamente modernos ..161

DISCUSIÓN 185
Indels 187
Hardy- Weinberg y selección 189
Varianza de Wahlund 191
Marcadores de ancestralidad 192
Análisis de una muestra de SNPs en Gitanos del País Vasco 192
Relación entre genes HLA de Neandertal y humanos modernos 193

CONCLUSIONES 195

BIBLIOGRAFÍA 199

INTRODUCCIÓN

INTRODUCCIÓN

Conceptos básicos

El concepto de variación genética humana se refiere a las diferencias observables tanto entre como dentro de las poblaciones humanas a nivel genético. El hecho de puedan existir múltiples variantes (alelos) de cada gen da lugar al polimorfismo, si bien no todos los genes los presentan. Un alelo es cada una de las formas alternativas que puede tener un mismo gen que se diferencian en su secuencia y que se puede manifestar en modificaciones concretas de la función de ese gen. Dado que la mayoría de los mamíferos son diploides, poseen dos juegos de cromosomas, uno de ellos procedente del padre y el otro de la madre. Cada par de alelos se ubica en igual locus o lugar del cromosoma (Mattei 2002). En ese caso, en el que solo hay un alelo presente en la población, se dice que el gen se ha fijado. Como valor medio, y en lo referente a las secuencias de ADN, todos los humanos somos iguales en un 99,5% a cualquier otro miembro de nuestra especie (Levy et al. 2007). Cuando se habla de variación genética en humanos se suele hacer una distinción entre alelos comunes y raros, con el fin de distinguir la frecuencia del alelo menos frecuente en la población. Los alelos comunes están relacionadas con el concepto de polimorfismo, que se define como aquellos casos en los que el alelo menos frecuente de un gen tiene una frecuencia alélica de al menos el 1% en la población, mientras que las variantes raras tienen una frecuencia alélica menor del 1%. También se pueden diferenciar los alelos según su composición nucleotídica. En términos generales, se pueden dividir en dos grupos: alelos de un solo nucleótido y variantes estructurales (Kidd et al. 2008). Los alelos de un solo nucleótido, más conocidos como polimorfismos de un solo nucleótido (SNP por sus siglas en inglés) son el tipo de variación genética más común entre distintos individuos. Se ha estimado que en el genoma humano hay al menos 11 millones de SNP, con alrededor de 7 millones de estos presentando una frecuencia alélica de más del 5%, y el resto de entre el 1 y el 5% (Kruglyak & Nickerson 2001). Además de los SNP hay innumerables variaciones de un solo nucleótido que son raras y nuevas ("*de novo*"), segregando en algunos casos solo en una familia nuclear o un solo individuo. Por ejemplo, cualquier par de bases que, aún alterado, es compatible con la vida es probable que se encuentre en al menos una persona de entre toda la población mundial. Sin embargo, es importante remarcar que en cualquier individuo la mayoría de las variantes alélicas que estén presentes, serán aquellas comunes a la población en su conjunto (Kruglyak & Nickerson 2001). Incluso cuando se comparan los genomas de dos individuos concretos, la mayor parte de las pares de bases que difieren se encuentran en loci donde hay variación poblacional.

Los alelos de SNPs localizados en la misma zona del genoma están, a menudo, relacionados. Esta estructura de correlación, llamada desequilibrio de ligamiento (Slatkin 2008), varía de una compleja e impredecible forma a lo largo del genoma y entre las distintas poblaciones. Los esfuerzos realizados durante la Fase 1 del Proyecto Internacional HapMap (Altshuler et al. 2005), junto con los de grupos asociados (Hinds et al. 2005), marcaron el punto de inicio en el camino para reducir el genoma a grupos de SNPs altamente correlacionados que, generalmente, se heredan juntos, conocidos como grupos, agrupaciones o asociaciones LD, o más comúnmente

haplotipos. A partir de la Fase 2 del Proyecto Internacional HapMap (Frazer et al. 2007) se determinó que la mayor parte de los SNP en los que la frecuencia del alelo menor es de al menos un 5% podrían reducirse a unos 550.000 haplotipos para individuos de ancestría asiática o europea, y a 1.100.000 haplotipos para individuos de ancestría africana ($r^2 \geq 0.8$). Mediante el genotipado de la muestra de ADN de un individuo con un SNP representativo (*SNP "tagging"*) para cada haplotipo, se consiguió mapear más del 80% de los SNPs presentes a lo largo del genoma a frecuencias mayores del 5% (Barrett J. C. & Cardon L. R. 2006, Pe'er I. et al. 2006, Clark A. G. & Li J. 2007, Eberle M. A. et al. 2007).

La variación estructural es, en un sentido amplio, todo conjunto de pares de bases que es diferente entre individuos y que no son variantes de un solo nucleótido. Este tipo de variación incluye las inserciones-deleciones (indels), sustituciones en bloque, inversiones de secuencias de ADN y variación en el número de copias (CNV). En contraposición a lo ocurrido con las variantes de un solo nucleótido, la capacidad tecnológica para detectar variantes estructurales en el genoma humano ha aparecido más recientemente (Tuzun et al. 2005, Khaja et al. 2006, Redon et al. 2006, Eichler et al. 2007, Korbel & et al. 2007, Cooper et al. 2008). Por eso, nuestra compresión de la localización y las frecuencias de las distintas variantes estructurales aún está desarrollándose (Conrad et al. 2006, Cooper, Nickerson & Eichler 2007, McCarroll & Altshuler 2007, Sebat 2007, Barnes et al. 2008, Korn et al. 2008). Estos estudios sugieren que la variación estructural suma al menos el 20% de todas las variantes genéticas en humanos, y subyace o se relaciona de alguna forma en más del 70% del total. En conjunto, para cualquier individuo dado, las variantes estructurales constituyen entre 9 y 25 Mb del genoma (entre el 0.5% y el 1%, aproximadamente), subrayando los importantes papeles de esta clase de variación en la evolución del genoma y en la salud humana y la enfermedad.

No hay dos humanos que sean genéticamente idénticos. Incluso los gemelos monocigóticos, es decir, los gemelos que se desarrollan a partir de un único óvulo fecundado, tienen diferencias genéticas debido a mutaciones ocurridas durante el desarrollo y variaciones en el número de copias dentro de los genes (Bruder et al. 2008).

Además, los procesos epigenéticos contribuyen a aumentar más estas diferencias. La epigenética es el campo de estudio de los cambios fenotípicos heredables que no se relacionan con cambios en la secuencia de ADN (Dupont et al. 2009). Normalmente, se asocia el término epigenética con cambios que afectan a la actividad génica o su expresión, pero también puede usarse para describir cualquier cambio fenotípico heredable. Estos efectos en características fenotípicas celulares o fisiológicas pueden ser el resultado de factores ambientales o externos, así como del normal proceso de desarrollo. La definición de epigenética aceptada de forma general requiere que estos cambios sean heredables, bien en la progenie celular o del organismo (Ledford 2008, Berger et al. 2009). Algunos ejemplos de los mecanismos que contribuyen a estos cambios son la metilación del ADN y la modificación histónica, cada uno de los cuales cambia la expresión génica sin alterar la secuencia de DNA subyacente.

La expresión génica puede ser controlada mediante la acción de proteínas represoras (un tipo de factores reguladores de transcripción) que se unen a regiones silenciadoras del ADN. Estos cambios epigenéticos pueden durar todo el ciclo de vida celular, durante los procesos de división celular, e incluso durante múltiples generaciones aunque no influyan en la secuencia de ADN del organismo (Bird 2007); dado que son factores no genéticos los que hacen que los genes se expresen de manera diferencial (Hunter 2008). Un ejemplo de un cambio epigenético es el proceso de diferenciación celular de los organismos eucariotas. Durante la morfogénesis, las células madre totipotenciales se transforman en las distintas líneas celulares pluripotenciales del embrión, lo que hace que se conviertan en células completamente diferenciadas. Esto se consigue mediante la activación de ciertos genes y la inhibición de la expresión de otros (Reik 2007).

Históricamente, se han descrito como epigenéticos algunos procesos que no son necesariamente heredables. Un ejemplo sería cualquier modificación en las regiones cromosómicas, especialmente las relacionadas con la modificación histónica, indiferentemente de si estos cambios eran heredables o asociados a un fenotipo. Hoy en día, la definición consensuada requiere que una característica sea heredable para ser consideraba epigenética (Berger et al. 2009).

Como ya se ha explicado, los cambios epigenéticos modifican la activación de ciertos genes, pero no código de la secuencia del ADN. La microestructura del propio ADN o las proteínas de cromatina asociadas pueden ser modificadas, provocando la activación o inhibición. Este mecanismo permite a las células diferenciadas de un organismo multicelular expresar solo los genes necesarios para su propia actividad. Los cambios epigenéticos son preservados en la división celular. Si bien la mayoría de estos cambios ocurren únicamente durante la vida de un organismo individual, pueden ser transmitidos a la descendencia mediante un proceso denominado herencia epigenética transgeneracional. Además, si la modificación epigenética del gen ocurre en el óvulo o el espermatozoide que participan en la fertilización, esta modificación puede ser transmitida a la siguiente generación (Chandler 2007). Algunos procesos epigenéticos concretos son la paramutación, la impronta genética, el silenciamiento génico, la inactivación del cromosoma X, los efectos de posición debidos a translocaciones, la reprogramación genética debida a metilación, la transvección, los efectos maternos, el proceso de la carcinogénesis, la regulación de las modificaciones en las histonas y la heterocromatina, y las limitaciones técnicas que afectan a procesos de partenogénesis o clonación.

El daño sufrido por el ADN también puede causar cambios epigenéticos (Kovalchuk & Baulch 2008, Ilnytskyy & Kovalchuk 2011, Friedl et al. 2012). Este tipo de daño es muy frecuente, ocurriendo varias decenas de miles de veces por cada célula del cuerpo humano al día. Estos daños son normalmente reparados, pero en el lugar de la reparación pueden mantenerse cambios epigenéticos (Cuozzo et al. 2007). Por ejemplo, una rotura de la doble hebra de ADN puede dar lugar a una inhibición génica por metilación de ADN así como por modificaciones histónicas (O'Hagan, Mohammad & Baylin 2008). Además la enzima PARP1 (poli ADP ribosa polimerasa 1) y su producto PAR (poli ADP ribosa) se acumulan en los lugares donde ha ocurrido el daño al ADN como parte del proceso de reparación (Malanga & Althaus 2005). Esta acumulación, genera

además la activación de la proteína remodeladora de cromatina ALC1 que puede causar la remodelación del nucleosoma (Gottschalk et al. 2009), lo cual puede dar lugar al silenciamiento epigenético del gen reparador de DNA MLH1 (Riggs, Russo & Martienssen 1996, Lin et al. 2007).

Algunos productos químicos dañinos para el ADN, como el benceno, la hidroquinona, el estireno, el tetracloruro de carbono o el tricloroetileno, causan una considerable hipometilación del ADN, algunos mediante la activación de rutas de stress oxidativo (Tabish et al. 2012).

El daño sufrido por el ADN es claramente distinto a una mutación, aunque ambos son tipos de error en el ADN. El daño en el ADN es una estructura química anormal en el ADN, mientras que una mutación es un cambio en la secuencia de pares de bases estándar. El daño en el ADN causa cambios en la estructura del material genético y previene que el mecanismo de replicación funcione y cumpla su función de manera adecuada (Köhler et al. 2016). Asimismo, el daño en el ADN y la mutación tienen distintas consecuencias biológicas. Mientras que la mayor parte de los cambios en el ADN pueden repararse, dicha reparación no es 100% eficiente. Los daños en el ADN que no se repararan se acumulan en células no replicantes, como las células cerebrales o musculares en mamíferos adultos y pueden ser causa de envejecimiento (Bernstein et al. 2008, Hoeijmakers 2009, Freitas & de Magalhães 2011). En células replicantes, como las células que recubren el colon, se producen errores tras la replicación de daños pasados en la hebra molde de ADN, o durante la reparación de daños en el ADN. Estos errores pueden dar lugar a mutaciones o alteraciones epigenéticas (O'Hagan, Mohammad & Baylin 2008). Ambos tipos de alteraciones pueden ser replicadas y transmitidas a generaciones celulares subsiguientes, pudiendo cambiar la función del gen o la regulación de la expresión del gen y, posiblemente, contribuir a la progresión del desarrollo del cáncer.

Diversidad genética

En una misma población pueden darse a un mismo tiempo varias frecuencias genotípicas, y cada una de estas podría identificar una subpoblación. Esto da lugar al concepto de distancia genética. La distancia genética es una medida de la divergencia genética entre especies o entre poblaciones dentro de una especie, ya sea que la distancia mida el tiempo desde el ancestro común o el grado de diferenciación (Nei 1987). Las poblaciones con muchos alelos similares tienen pequeñas distancias genéticas. Esto indica que están estrechamente relacionados y tienen un ancestro común reciente. Es decir, es una medida de frecuencia de recombinación promedio sobre una muestra. La distancia genética es útil para reconstruir la historia de las poblaciones. Por ejemplo, la evidencia de la distancia genética sugiere que las personas de África Subsahariana y Eurasia divergieron hace unos 100.000 años (Nei & Roychoudhury 1974). Las variaciones alélicas en cada locus causan variación fenotípica dentro de las especies (por ejemplo, color de cabello, color de ojos). Sin embargo, la mayoría de los alelos no tienen un impacto observable en el fenotipo. Dentro de una población, los alelos nuevos generados por la mutación o bien desaparecen o se diseminan por toda la población. Cuando una población se divide en diferentes poblaciones aisladas (ya sea por factores geográficos o ecológicos), las mutaciones que ocurren

después de la división estarán presentes solo en la población aislada. La fluctuación aleatoria de las frecuencias de los alelos también produce una diferenciación genética entre las poblaciones. Este proceso se conoce como deriva genética.

Las diferencias entre las poblaciones representan una pequeña proporción de la variación genética humana en general. Las poblaciones también difieren en la cantidad de variación entre sus miembros. La mayor divergencia entre las poblaciones se encuentra en el África subsahariana, en consonancia con el reciente origen africano de las poblaciones no africanas. Las poblaciones también varían en la proporción y los loci exactos que ocupan de los genes introgresados que recibieron por mezcla arcaica tanto dentro como fuera de África. El estudio de la variación genética humana tiene una importancia evolutiva y aplicaciones médicas. Puede ayudar a los científicos a entender las antiguas migraciones de poblaciones humanas y cómo los grupos humanos están biológicamente relacionados entre sí. El estudio de la variación genética humana tiene tanto significación evolutiva como aplicaciones médicas. Puede ayudar a entender migraciones de poblaciones humanas antiguas así como las relaciones existentes entre los distintos grupos humanos. Para la medicina, el estudio de la variación humana puede ser de importancia debido a que algunos alelos causantes o relacionados con enfermedades son más frecuentes en regiones geográficas específicas.

Causas de las diferencias entre individuos

Las causas de las diferencias entre individuos incluyen la segregación independiente de los caracteres de la generación parental a la filial (Segunda Ley de Mendel), el intercambio génico (entrecruzamiento cromosómico y recombinación) durante la meiosis y diversos eventos mutacionales.

La Segunda Ley de Mendel (1866) o Ley de la Segregación Independiente establece que los alelos para rasgos separados se transmiten independientemente el uno del otro. Es decir, la selección biológica de un alelo para un rasgo no tiene nada que ver con la selección de un alelo por ningún otro rasgo. La segregación independiente ocurre en organismos eucarióticos durante la profase meiótica I, y produce un gameto con una mezcla de cromosomas del organismo. La base física de la segregación independiente de cromosomas es la orientación aleatoria de cada cromosoma bivalente a lo largo de la placa de metafase con respecto a los otros cromosomas bivalentes. Junto con el entrecruzamiento cromosómico, la segregación independiente aumenta la diversidad genética al producir nuevas combinaciones genéticas, aunque los fenómenos de ligamiento génico actúan en su contra. De los 46 cromosomas en una célula humana diploide normal, la mitad son derivados de la madre (del óvulo) y la otra mitad son derivados del padre (del esperma). Esto ocurre porque la reproducción sexual implica la fusión de dos gametos haploides (el óvulo y el esperma) para producir un nuevo organismo con el conjunto completo de cromosomas. Durante la gametogénesis, la producción de nuevos gametos por parte de un adulto, el conjunto normal de 46 cromosomas debe reducirse a la mitad para garantizar que el gameto haploide resultante pueda unirse a otro gameto haploide para producir un organismo diploide. En

la segregación independiente, los cromosomas resultantes son elegidos aleatoriamente de entre todos los posibles cromosomas maternos y paternos. Dado que los zigotos terminan con una mezcla aleatoria en lugar de un conjunto predefinido de cualquiera de los padres, se consideran segregados de forma independiente. Por lo tanto, el cigoto puede terminar con cualquier combinación de cromosomas paternos o maternos. Cualquiera de las posibles variantes de un cigoto formado a partir de cromosomas maternos y paternos se producirá con la misma frecuencia. El cigoto terminará con 23 pares de cromosomas, pero el origen de cualquier cromosoma particular se seleccionará al azar de los cromosomas paternos o maternos. Esto es lo que contribuye a la variabilidad genética de la descendencia.

Intercambio génico

El proceso de intercambio génico por entrecruzamiento cromosómico y recombinación es el intercambio de material genético entre cromosomas homólogos que da como resultado cromosomas recombinantes durante la reproducción sexual. Ocurre en la etapa de paquiteno de la profase I de la meiosis durante un proceso llamado sinapsis. La sinapsis comienza antes de que se desarrolle el complejo sinaptonémico y no se completa hasta casi el final de la profase I. El cruce generalmente ocurre cuando las regiones coincidentes en los cromosomas enfrentados se rompen y luego se vuelven a conectar al otro cromosoma. La base física del cruce fue demostrada por primera vez por Harriet Creighton y Barbara McClintock (1931). La frecuencia vinculada al cruce entre dos loci cualesquiera se denomina valor de cruce. Para un conjunto fijo de condiciones genéticas y ambientales, la recombinación en una región particular tiende a ser constante.

Existen dos teorías que explican los orígenes del proceso de entrecruzamiento, cada una derivada de las diferentes teorías sobre el origen de la meiosis. La primera teoría proviene de la idea de que la meiosis evolucionó a partir de la transformación bacteriana, con la función de propagar la diversidad (Bernstein & Bernstein 2010). La segunda teoría se basa en la idea de que la meiosis evolucionó como otro método de reparación del ADN, y así el cruce es una forma novedosa de reemplazar secciones de ADN posiblemente dañadas (Bernstein, Bernstein & Michod 2011).

El cruce y la reparación del ADN son procesos muy similares, que utilizan muchos de los mismos complejos de proteínas (Saponaro et al. 2010; Dangel, Knoll & Puchta 2014). McClintock (1984) estudió el maíz para mostrar cómo el genoma del maíz cambiaría para superar las amenazas a su supervivencia. Usó 450 plantas autopolinizadas que recibieron de cada padre un cromosoma con un extremo roto, y patrones modificados de expresión genética en diferentes sectores de hojas de maíz dado que los elementos transponibles ("elementos controladores") se esconden en el genoma, y su movilidad les permite alterar la acción de los genes en diferentes loci. Estos elementos también pueden reestructurar el genoma, desde unos pocos nucleótidos hasta segmentos completos del cromosoma. Las recombinasas y las primasas son las encargadas de sentar las bases de los nucleótidos a lo largo de la secuencia de ADN. Uno de estos complejos de proteínas particulares que se conserva entre los procesos es RAD51, una proteína recombinasa

bien conservada que se ha demostrado que es crucial en la reparación del ADN y en el cruzamiento (Esposito 1978, Shinohara et al. 1993). Otros genes en *Drosophila melanogaster* se han relacionado también con ambos procesos, al mostrar que los mutantes en estos loci específicos no pueden someterse a la reparación del ADN o al cruce (Bernstein, Bernstein & Michod 2011). Este gran grupo de genes conservados entre procesos apoya la teoría de una estrecha relación evolutiva. Además, se ha encontrado que la reparación del ADN y el cruce favorecen regiones similares en los cromosomas. En un experimento que utilizó el mapeo híbrido de radiación en el cromosoma 3B del trigo (*Triticum aestivum L*.), se encontró que el cruzamiento y la reparación del ADN se producen predominantemente en las mismas regiones (Kumar, Bassi & Paux 2012). Además, se ha correlacionado que el cruce se produce en respuesta a condiciones estresantes, y probablemente dañinas para el ADN (Nedelcu, Marcu & Michod 2004, Steinboeck 2010).

El proceso de transformación bacteriana también comparte muchas similitudes con el proceso de entrecruzamiento cromosómico, particularmente en la formación de salientes a los lados de la cadena de ADN rota, lo que permite la elaboración de una nueva cadena. La transformación bacteriana en sí misma se ha relacionado con la reparación del ADN muchas veces.

Por lo tanto, la evidencia sugiere que la pregunta trata de si el cruce está más relacionado con la reparación del ADN o la transformación bacteriana, ya que los dos no parecen ser mutuamente excluyentes. Es probable que el cruce haya evolucionado a partir de la transformación bacteriana, que a su vez se desarrolló a partir de la reparación del ADN, lo que explica los vínculos entre los tres procesos.

En cuanto a la bioquímica del proceso, cabe destacar que la recombinación meiótica puede iniciarse mediante rupturas bicatenarias que se producen en el ADN mediante la exposición a agentes que dañan el ADN (Bernstein, Bernstein & Michod 2011) o la proteína Spo11 (Keeny, Giroux & Kleckner 1997). A continuación, una o más exonucleasas digieren los extremos 5' generados por las roturas bicatenarias para producir colas de ADN monocatenarias 3'. La recombinasa específica de meiosis DMC1 y la recombinasa general Rad51 recubren el ADN monocatenario para formar filamentos de nucleoproteína (Sauvageau et al. 2005). Las recombinasas catalizan la unión de la cromátida opuesta y el ADN monocatenario desde un extremo de la ruptura. A continuación, el extremo 3' del ADN monocatenario que se une ceba la síntesis de ADN, provocando el desplazamiento de la cadena complementaria, que posteriormente se empareja con el ADN monocatenario generado a partir del otro extremo de la rotura bicatenaria inicial. La estructura que resulta es un tetrámero en forma de cruz, también conocido como un cruce Holliday. El cruce Holliday es una estructura tetraédrica que puede ser “arrastrada” por otras recombinasas, moviéndola a lo largo de la estructura de cuatro cadenas. El contacto entre dos cromátidas que pronto sufrirá un cruce se conoce como quiasma. Church y Wimber (1969) midieron la frecuencia del quiasma en las diferentes etapas de la meiosis del saltamontes *Melanoplus femurrubrum* mediante irradiación con rayos X. Se encontró que la irradiación durante las etapas de leptoteno - cigoteno de la meiosis (es decir, antes del período de paquiteno en el que se produce la recombinación de cruce) aumenta la frecuencia de quiasma posterior. De manera similar, en el saltamontes *Chorthippus brunneus*, la exposición a la irradiación con rayos X durante

las etapas de cigoteno - paquiteno temprano causó un aumento significativo en la frecuencia media de quiasma celular (Westerman 1971). La frecuencia de quiasma se puntuó en las etapas posteriores de diploteno - diaquinesis de la meiosis. Estos resultados sugieren que los rayos X inducen daños en el ADN que se reparan mediante una vía cruzada que conduce a la formación de quiasma.

Típicamente, los entrecruzamientos se producen entre regiones homólogas de cromosomas coincidentes, pero las similitudes en la secuencia y otros factores pueden dar como resultado alineamientos no coincidentes, proceso denominado entrecruzamiento no homólogo, o más raramente, recombinación desigual. Durante la replicación del ADN, cada cadena de ADN se utiliza como plantilla para la creación de nuevas cadenas utilizando un mecanismo parcialmente conservado, cuyo funcionamiento adecuado da como resultado dos cromosomas idénticos emparejados. Se sabe que los eventos de cruce de cromátidas hermanas ocurren varias veces por célula en división en eucariotas (Smith 1976). La mayoría de estos eventos implican un intercambio de cantidades iguales de información genética, pero pueden ocurrir intercambios desiguales debido al desajuste de secuencia que dan como resultado una inserción o eliminación de información genética en el cromosoma. Aunque son raros en comparación con los eventos de cruce homólogos, estas mutaciones son drásticas y afectan a muchos loci al mismo tiempo. Se los considera el principal impulsor de la generación de duplicaciones génicas y son una fuente general de mutación dentro del genoma. Se desconocen las causas específicas de los eventos de entrecruzamientos no homólogos, pero se sabe que varios factores aumentan la probabilidad de un cruce desigual. Un factor común es la reparación de roturas de doble cadena (Puchta 2005). Estas roturas se reparan a menudo utilizando uniones finales no homólogas, un proceso que implica la inserción de un filamento molde en el filamento roto. Las regiones homólogas cercanas de la cadena molde se usan a menudo para la reparación, lo que puede dar lugar a inserciones o deleciones en el genoma si se utiliza una parte no homóloga pero complementaria de la cadena molde. La similitud entre secuencias es un factor importante en el proceso de entrecruzamiento: es más probable que los eventos de cruce ocurran en regiones largas de secuencia similar en un gen (Metzenberg et al. 1991). Esto implica que cualquier sección del genoma con largas secciones de ADN repetitivo es propensa a eventos cruzados.

La presencia de elementos transponibles es otro elemento influyente del entrecruzamiento no homólogo. Debido a que las regiones cromosómicas compuestas de transposones tienen grandes cantidades de código idéntico y repetitivo en un espacio condensado, se cree que las regiones de transposones que se someten a un evento de cruce son más propensas a un emparejamiento complementario erróneo (Robberecht et al. 2012); es decir, una sección de un cromosoma que contiene muchas secuencias idénticas, si se somete a un evento de cruce, es menos seguro que coincida con una sección perfectamente homóloga de código complementario y más propenso a unirse con una sección de código en una parte ligeramente diferente del cromosoma. Esto da como resultado una recombinación desequilibrada, ya que la información genética puede insertarse o eliminarse en el nuevo cromosoma, dependiendo de dónde se produjo la recombinación. Si bien los factores causantes de la recombinación desigual siguen sin esclarecerse del todo, se han descubierto algunos elementos del mecanismo físico del proceso. Las

proteínas de reparación de desajustes (MMR, *MisMatch Repair*), por ejemplo, son una familia reguladora de proteínas responsable de regular las secuencias de ADN no coincidentes durante la replicación (Kunkel & Erie 2005). El objetivo de las proteínas MMR es la restauración del genotipo parental. Se sabe que una clase de proteína MMR en particular, MSH3, forma el heterodímero MutSβ con MSH2 para corregir los largos bucles de inserción / deleción y el apareamiento erróneo base-base en los microsatélites durante la síntesis de ADN. La funcionalidad deficiente en las proteínas MMR se encuentra en aproximadamente el 15% de los cánceres colorrectales, y las mutaciones somáticas en el gen MSH3 se pueden encontrar en casi el 50% de los cánceres colorrectales deficientes en actividad MMR (Gao et al. 2013). Existen múltiples vías de proteínas MMR implicadas en el mantenimiento de la estabilidad del genoma de un organismo complejo, y cualquiera de los muchos posibles errores de funcionamiento da como resultado errores de edición y corrección del ADN (Surtees, Argueso & Alani 2004). Por lo tanto, si bien se desconoce con total certeza qué mecanismos conducen a errores de entrecruzamiento no homólogo, es muy probable que la vía MMR esté implicada.

Selección natural

La selección natural puede conferir una ventaja adaptativa a los individuos en un ambiente determinado si un alelo de la población les dota de una ventaja competitiva (Darwin 1859). Los alelos que presentan selección es más probable que se encuentren presentes en aquellas regiones en las que confieren una ventaja. La selección natural actúa sobre el fenotipo (las características observables de un organismo), pero es el componente genético heredable el que, si confiere una ventaja reproductiva, puede volverse más común en una población. Con el paso del tiempo, este proceso puede dar como resultado poblaciones especialmente adaptadas a determinados nichos ecológicos (microevolución) y, eventualmente, el surgimiento de nuevas especies (especiación, macroevolución). Dicho de otro modo, la selección natural es un proceso clave en la evolución de una población. La selección natural actúa mediante diversos mecanismos:

- Variación heredable: como se ha explicado más arriba, la segregación independiente de los caracteres de la generación parental a la filial (Segunda Ley de Mendel), que codifican para cada característica, durante la producción de gametos mediante una división celular meiótica. Esto significa que cada gameto va a contener un solo alelo para cada gen. Lo cual permite que los alelos materno y paterno se combinen en el descendiente, asegurando la variación.

- Fitness: en biología evolutiva, es la representación cuantitativa de la selección natural y sexual. Se puede definir respecto a un genotipo o un fenotipo, para un entorno determinado. En cualquier caso, describe el éxito reproductivo y es igual a la contribución media al *pool* genético de la siguiente generación resultante de dicho genotipo o fenotipo. El fitness de un genotipo es manifestado a través de su fenotipo, el cual también se ve afectado por factores ambientales. Por eso el fitness de un fenotipo dado puede ser distinto en diferentes ambientes selectivos (Orr 2009). La selección natural tiende a hacer que los alelos con mayor fitness sean más comunes con el tiempo, dando como resultado procesos de evolución. Se puede obtener una cantidad de

información importante sobre los distintos tipos de evolución considerando un modelo de locus simple. Si suponemos un locus inicial A_1, y que en algún punto una mutación introduce el alelo A_2, los tres posibles genotipos resultantes serían A_1A_1, A_1A_2 y A_2A_2. Cada uno de estos genotipos tendría asociado un fitness exclusivo (w_{11}, w_{12} y w_{22}), que si bien es en realidad algo más complejo como hemos visto más arriba, en este ejemplo podríamos asumir que representan las probabilidades de supervivencia de cada individuo. En la literatura científica, los valores absolutos de fitness (w_{11}, w_{12} y w_{22}) son a menudo convertidos en valores de fitness relativo respecto a uno de los genotipos. En nuestro ejemplo, 1, 1+hs y 1+s, respectivamente. El fitness de los genotipos A_1A_2 y A_2A_2 es expresado de forma relativa al del genotipo A_1A_1. Así si asumimos que el fitness de A_1A_1 es igual a 1, 1+hs = w_{12}/w_{11}, y 1+s = w_{22}/w_{11}. Los parámetros *s* y *h* son el coeficiente de selección y el efecto del heterocigoto, respectivamente (Akey 2009). Se pueden definir diversos tipos de selección en términos de fitness relativo.

- Competencia: es la interacción entre organismos en la cual el fitness de uno es reducido por la presencia del otro. Esto puede darse por la necesidad de ambos organismos de un mismo recurso que esté limitado, como es el agua, el alimento o el territorio (Begon, Harper & Townsend 1996). La competencia puede darse entre especies o dentro de una misma especie, y puede ser directa o indirecta (Sahney, Benton & Ferry 2010). Las especies menos preparadas para competir deben adaptarse o desaparecer, teniendo en cuenta el importante papel que juega la competencia en la selección natural.

- Selección sexual: El concepto de selección sexual se refiere específicamente a la competencia por la pareja (Andersson 1994); la cual puede ser intrasexual, entre individuos del mismo sexo, como por ejemplo la competencia entre machos por el derecho a aparearse con las hembras; o intersexual, donde uno de los géneros es el que elige la pareja, más comúnmente las hembras eligen entre un conjunto de machos que se exhiben (Hosken & House 2011). Diversos rasgos fenotípicos pueden asociarse a la selección sexual, siendo mostrados por uno de los sexos y deseado por el otro, dando lugar a un *feedback* positivo en el que se retroalimenta la evolución exagerada de un determinado carácter. Este proceso se conoce como selección autorreforzante de Fisher (Greenfield et al. 2014). La agresión entre miembros del mismo sexo a veces se asocia a características intersexuales muy distintivas, como las cornamentas de los ciervos, que se usan en combates por el derecho reproductivo. Además, las selección intrasexual se asocia generalmente con el dimorfismo sexual, incluyendo las diferencias de tamaño corporal entre machos y hembras de una misma especie (Hosken & House 2011).

Tipos de selección natural

Cuando no hay diferencias en el fitness entre genotipos (s = 0), se dicen que las frecuencias alélicas y genotípicas evolucionan de manera neutral, y si no es el caso, hay selección natural. El tipo concreto de selección es dependiente de si el valor del coeficiente de selección es positivo o negativo, y de la relación de dominancia entre alelos reflejada en el valor del efecto del heterocigoto. Por ejemplo, la selección direccional ocurre con dominancia incompleta ($0 < h < 1$). Si

$s < 0$, entonces el nuevo alelo A_2 es deletéreo, y los individuos portadores del mismo son menos aptos, y la selección purificadora (también conocida como selección negativa) actúa para purgar A_2 de la población. Si $s > 0$, el nuevo alelo A_2 es ventajoso, y los individuos portadores estarán mejor adaptados, y A_2 será fijado en la población. La selección direccional tiene como resultado una pérdida de variación genética y, en general, la selección direccional se corresponde con la forma de selección descrita por Darwin.

Otro tipo de selección que actúa sobre alelos ventajosos se denomina selección sobredominante, y ocurre cuando el heterocigoto A_1A_2 tiene el fitness relativo más alto ($s > 0$ y $h > 1$). La selección sobredominante (también llamada ventaja del heterocigoto) es una de las formas específicas en las que se manifiesta la selección balanceadora, la cual actúa de tal modo que se conserva la variación genética de la población. Gillespie (1991) señaló que la selección balanceadora puede ocurrir en ausencia de sobredominancia. Tanto Nielsen (2005) como Akey (2009) se refieren a cualquier tipo de selección que actúa sobre un alelo ventajoso como selección positiva. Sin embargo, Wang et al. (2006) identifican la selección direccional como la fuente primaria de selección positiva en estudios a nivel genómico.

Importancia de la selección natural

En lo referente al genoma humano, la selección natural ha jugado un papel clave en el desarrollo de la especie. Cuando las poblaciones se ven sometidas a condiciones ambientales muy distintas, o sufren presiones por fenómenos como enfermedades, la selección natural puede llegar a cambiar de manera drástica la frecuencia alélica de una población a otra. Por ello, cambios acentuados en las frecuencias alélicas entre poblaciones pueden señalizar loci del genoma que han sufrido procesos de selección. Otras señales de selección pueden ser haplotipos característicos (que al fin y al cabo son conjuntos de alelos característicos de una población o subpoblación) y una variación alélica reducidas en las regiones próximas a las variantes seleccionadas (al encontrarse estas ligadas).

Las inferencias sobre la selección natural generalmente se basan en la detección de sus efectos sobre los patrones de variación neutral vinculada a dicha selección. El autostop genético, también llamado "*draft* genético" (Gillespie 2000) o efecto de autostop (Smith & Haigh 1974), se refiere a la influencia que la selección sobre los alelos ventajosos tiene en los patrones de variación ligada. Es decir, se da cuando un alelo cambia su frecuencia, no porque el mismo se encuentre bajo los efectos de la selección natural, sino porque se encuentra cerca de otro gen que está bajo los efectos de un barrido selectivo y que está en la misma cadena de ADN. Cuando un gen sufre un barrido selectivo, cualquier polimorfismo que se encuentre cerca y esté en desequilibrio de ligamiento, tenderá a cambiar también sus frecuencias alélicas (Futuyma 2013). Los barridos ocurren cuando mutaciones *de novo* (y por tanto, aún raras) son ventajosas y aumentan su frecuencia. Los alelos neutrales o mínimamente deletéreos que se encuentren cerca en el cromosoma son arrastrados junto con el alelo bajo selección, proceso conocido como barrido selectivo. Los barridos que se encuentran en proceso o están incompletos son aquellos que se

encuentran en cualquier estadio anterior a la fijación del alelo ventajoso. Una vez que se ha fijado, se dice que el barrido está completo. En contraste, los efectos sobre un locus neutral debidos a desequilibrio de ligamiento con mutaciones deletéreas *de novo* se denominan selección de fondo. Tanto el autostop genético como la selección de fondo son fuerzas evolutivas estocásticas (aleatorias), como la deriva genética (Gillespie 2001).

El modelo clásico de selección positiva dice que la selección actúa sobre mutaciones ventajosas *de novo*. De manera alternativa, la selección podría actuar sobre la variación genética preexistente que fuera bien neutral o deletérea, pero que haya pasado a ser una adaptación debido a cambios en el ambiente o en el trasfondo genético. Hermisson & Pennings (2005) se refieren a la selección asociada a variación preexistente como "barrido suave", a fin de distinguirla del modelo clásico, o "barrido fuerte". Patrones de variación genética derivados de la selección de mutaciones *de novo* pueden diferir enormemente entre ambos modelos (Hemisson & Pennings 2005, Przeworski et al. 2005).

La caracterización de las huellas que deja la selección direccional positiva en los genes que son de importancia adaptativa en humanos puede tener una importante relevancia a nivel médico, ayudando a identificar variantes funcionales que juegan un rol en la salud de las poblaciones.

Deriva genética

La segunda causa de la variación genética es debida a la deriva genética (o génica) y al alto grado de neutralidad de la mayor parte de las mutaciones. La mayoría de las mutaciones no parecen tener un efecto selectivo positivo ni negativo en el organismo, es decir, son selectivamente neutras (Kimura 1968, King & Jukes 1969). La causa principal es la deriva genética, que es el cambio aleatorio de las frecuencias alélicas de una población debida al muestreo aleatorio de individuos (Masel 2011). Los alelos presentes en la descendencia son una muestra de los alelos presentes en los padres, y la casualidad tiene un papel en determinar si un individuo dado sobrevive y se reproduce. La deriva genética puede hacer que las variantes genéticas desaparezcan por completo y, por lo tanto, reduzcan la variación genética o causen que los alelos inicialmente raros se vuelvan mucho más frecuentes o incluso lleguen a fijarse. Los efectos de la deriva genética son más fuertes en poblaciones pequeñas y aisladas y pueden conducir a frecuencias de alelos significativamente modificadas en un período de tiempo relativamente corto en dichas poblaciones (Star & Spencer 2013; LaBar & Adami 2017). Cuando hay pocas copias de un alelo, el efecto de la deriva genética es mayor, y cuando hay muchas copias, el efecto es menor. A principios del siglo XX, se produjeron vigorosos debates sobre la importancia relativa de la selección natural frente a los procesos neutros, incluida la deriva genética. Ronald Fisher (1922) sostuvo que la deriva genética juega un papel menor en la evolución, y esta siguió siendo la visión dominante durante varias décadas. El genetista de poblaciones Motoo Kimura (1968) reavivó el debate con su teoría neutral de la evolución molecular, que afirma que la mayoría de los casos donde un cambio genético se propaga a través de una población (aunque no necesariamente

cambios en los fenotipos) son causados por la deriva genética actuando sobre mutaciones neutrales.

Los modelos matemáticos de deriva genética pueden diseñarse usando procesos de ramificación o una ecuación de difusión que describe los cambios en la frecuencia de alelos en una población idealizada (Wahl 2011).

Los modelos más importantes son el modelo de Wright-Fisher y el modelo de Moran (Moran 1958). Sin embargo, existen diversos problemas a tener en cuenta al usar estos modelos. Si la varianza en el número de descendientes es mucho mayor que la dada por la distribución binomial asumida por el modelo de Wright-Fisher, dada la misma velocidad general de deriva genética (el tamaño de población efectivo de varianza), la deriva genética será una fuerza menos poderosa en comparación con la selección (Charlesworth 2009). Incluso para la misma varianza, si los momentos superiores de la distribución numérica de la descendencia exceden los de la distribución binomial, la fuerza de la deriva genética se debilita sustancialmente (Der, Epstein & Plotkin 2011). Los cambios aleatorios en las frecuencias de los alelos también pueden ser causados por efectos distintos al error de muestreo, por ejemplo, cambios aleatorios en la presión de selección. Como se ha comentado, una fuente alternativa importante de estocasticidad, quizás más importante que la deriva genética, es el autoestop genético (Gillespie 2001). Las propiedades matemáticas del autostop genético son diferentes de las de la deriva genética (Neher & Shraiman 2011).

Principio de Hardy-Weinberg

El principio de Hardy-Weinberg establece que dentro de poblaciones suficientemente grandes, las frecuencias de los alelos permanecen constantes de una generación a la siguiente a menos que el equilibrio se vea alterado por la migración, las mutaciones genéticas o la selección. Sin embargo, en poblaciones finitas, no se obtienen nuevos alelos a partir del muestreo aleatorio de los alelos pasados a la siguiente generación, pero el muestreo puede causar la desaparición de un alelo existente. Debido a que el muestreo aleatorio puede eliminar, pero no reemplazar, un alelo, y debido a disminuciones o aumentos aleatorios en la frecuencia de alelos, la deriva genética conduce a la población hacia la uniformidad genética a lo largo del tiempo. Cuando un alelo alcanza una frecuencia de 1 (100%) se dice que se ha fijado en la población y cuando un alelo alcanza una frecuencia de 0 (0%) se dice que se ha perdido en dicha población. Las poblaciones más pequeñas logran una fijación más rápida, mientras que en el límite de una población infinita, la fijación no se logra. Una vez que un alelo se fija, la deriva genética se detiene, y la frecuencia de los alelos no puede cambiar a menos que se introduzca un nuevo alelo en la población a través de la mutación o el flujo de genes. Por lo tanto, incluso si la deriva genética es un proceso aleatorio y sin dirección, actúa para eliminar la variación genética a lo largo del tiempo.

Deriva y selección

En las poblaciones naturales, la deriva genética y la selección natural no actúan aisladamente; ambas fuerzas están siempre en juego, junto con la mutación y la migración. La evolución neutra es el producto tanto de la mutación como de la deriva, no solo de la deriva. De manera similar, incluso cuando la selección supera la deriva genética, solo puede actuar sobre la variación que proporciona la mutación. Si bien la selección natural tiene una dirección que guía la evolución hacia adaptaciones hereditarias al entorno actual, la deriva genética no tiene sentido y está guiada solo por las matemáticas del azar. Como resultado, la deriva actúa sobre las frecuencias genotípicas dentro de una población sin tener en cuenta sus efectos fenotípicos. Por el contrario, la selección favorece la propagación de alelos cuyos efectos fenotípicos aumentan la supervivencia y / o la reproducción de sus portadores, disminuye las frecuencias de los alelos que causan rasgos desfavorables e ignora los que son neutros. La ley de los grandes números aplicada a la deriva predice que cuando el número absoluto de copias del alelo es pequeño (por ejemplo, en poblaciones pequeñas), la magnitud de la deriva en las frecuencias de los alelos por generación es mayor. La magnitud de la deriva es lo suficientemente grande como para hacer insignificante a la selección en cualquier frecuencia de alelos cuando el coeficiente de selección es inferior a 1 dividido por el tamaño efectivo de la población. Por esto se considera que la evolución no adaptativa resultante del producto de la mutación y la deriva genética es un mecanismo consecuente de cambio evolutivo principalmente en poblaciones pequeñas aisladas. El ligamiento genético con otros genes que están bajo selección puede reducir el tamaño efectivo de la población experimentado por un alelo neutral. Con una tasa de recombinación más alta, el ligamiento disminuye y con él, este efecto local sobre el tamaño efectivo de la población (Charlesworth, Morgan & Charleswoth 1993). Este efecto es visible en los datos moleculares como una correlación entre la tasa de recombinación local y la diversidad genética (Presgraves 2005), y la correlación negativa entre la densidad de genes y la diversidad en regiones de ADN no codificantes (Nordborg et al. 2005). Cuando la frecuencia de los alelos es muy pequeña, la deriva también puede dominar la selección incluso en grandes poblaciones. Por ejemplo, aunque las mutaciones desfavorables generalmente se eliminan rápidamente en grandes poblaciones, las nuevas mutaciones ventajosas son casi tan vulnerables a la pérdida por deriva genética como las mutaciones neutrales. Hasta que la frecuencia de los alelos para la mutación ventajosa no alcance un cierto umbral, la deriva genética no tendrá efecto.

Mecanismos de actuación de la deriva génica

El principal mecanismo por el cual actúa la deriva génica se conoce como efecto “cuello de botella”. Se caracteriza por una reducción drástica del tamaño poblacional en un corto periodo de tiempo asociado cambios aleatorios en el medio que derivan en una mortalidad elevada en una población (Robinson 2003). En un proceso cuello de botella puro, las probabilidades de sobrevivir de cualquier miembro de la población son puramente aleatorias, y no se ven aumentadas por ninguna característica genética inherente. El proceso puede dar como resultado cambios radicales en las frecuencias alélicas, de manera completamente independiente de la selección natural.

El impacto del cuello de botella en una población puede ser mantenido, incluso cuando el cuello de botella es causado por un único evento fortuito, como un desastre natural. Después de un cuello de botella, la endogamia aumenta. Esto hace que se incremente el daño realizado a la población por las mutaciones recesivas deletéreas, en un proceso conocido como depresión endogámica. La selección contra mutaciones deletéreas puede conllevar la pérdida de otros alelos ligados genéticamente a ellas en un proceso denominado selección de fondo (Masel 2011), y puede verse incrementado su efecto por un proceso de purga genética (García-Dorado 2015). Esto puede conllevar una pérdida incrementada de diversidad genética. Además, una reducción poblacional sostenida en el tiempo puede incrementar las probabilidades de fluctuaciones alélicas debidas a la deriva en generaciones venideras. La variación genética de una población puede verse muy reducida por un cuello de botella, e incluso adaptaciones beneficiosas pueden ser permanentemente eliminadas (Futuyma 1998). La pérdida de variación hace que la población superviviente sea vulnerable a cualquier nueva presión selectiva, dado que la adaptación a cambios ambientales necesita de suficiente variación genética en la población para que la selección natural pueda tener efecto (O'Corry-Crowe 2008, Cornuet & Luikart 1996). Empíricamente, se ha estudiado la deriva genética aleatoria en humanos, así como en poblaciones naturales y experimentales de otros muchos organismos. Por ejemplo, Helgason et al. (2003) mostraron por medio de diversos análisis que los patrones de variación genética en los islandeses, como la baja diversidad genética, son consistentes con altos niveles de deriva genética en el pasado.

El efecto fundador es un caso especial de efecto cuello de botella, que ocurre cuando un pequeño grupo de una población se separa de la población original y forma una nueva población. El muestreo aleatorio de alelos en la población recién formada es de esperar que no describa de manera fehaciente la población original en algunos aspectos (Campbell 1996). Es incluso posible que el número de alelos para algunos genes en la población original sea mayor que el número de copias génicas en la población fundadora, haciendo que la representación completa sea totalmente imposible. Por eso, cuando el tamaño poblacional de la población fundadora es muy pequeño, sus fundadores pueden afectar fuertemente la composición genética de la población en el futuro. La idea de que la importancia de la deriva genética acabe superando la de la selección natural en poblaciones pequeñas ha sido validada en múltiple estudios de evolución molecular en humanos. Por ejemplo, frecuencias altas para determinadas enfermedades en grupos étnicos relativamente cerrados son generalmente atribuibles a valores altos de deriva genética, bien durante o después de la fundación de estas poblaciones (Ostrer 2001). La alta frecuencia de acromatopsia en la población del atolón Pingelap es un ejemplo ampliamente citado de como los procesos extremos de cuello de botella pueden afectar a la frecuencia de los alelos deletéreos (Hussels & Morton 1972). Un buen ejemplo es la migración Amish a Pennsylvania en 1744. Dos miembros de la colonia original compartían el alelo recesivo para el síndrome de Ellis-van Creveld. Como resultado de muchas generaciones de endogamia debida al aislamiento de la población, el síndrome de Ellis-van Creveld presenta mucha más prevalencia entre los Amish que entre la población general (Cavalli-Sforza et al. 1996).

La diferencia en las frecuencias génicas entre la población original y la fundadora puede también provocar que los dos grupos se diferencien significativamente en el curso de muchas generaciones. Según la diferencia (distancia genética) se incrementa, las dos poblaciones separadas pueden llegar a ser distintas, tanto a nivel genético como fenotípico, aunque no solo la deriva genética sino también la selección natural, el flujo génico y la mutación contribuyen a esta divergencia. Este potencial para provocar cambios relativamente rápidos en la frecuencia génica de la población fundadora, hizo que la mayoría de científicos consideraran el efecto fundador (y por extensión, la deriva genética) una fuerza significativa en la evolución de nuevas especies. Sewall Wright fue el primero en relacionar esta significancia a la deriva aleatoria y a poblaciones pequeñas recientemente formadas, en su teoría del equilibrio cambiante en la especiación (Wolf, Brodie & Wade 2000). Después de Wright, Ernst Mayr creo muchos modelos que explicaban como la reducción en la variación genética y el pequeño tamaño poblacional debido al efecto fundador, eran de importancia crítica para el desarrollo de nuevas especies (Hey, Fitch & Ayala 2005). Sin embargo, hoy en día este punto de vista cuenta con mucho menos apoyo dado que los resultados experimentales son equívocos (Howard & Berlocher 1998).

En humanos, el efecto fundador asociado a un tamaño poblacional original pequeño (lo cual se asocia a un aumento de probabilidad de que ocurran procesos de deriva genética), puede haber tenido una importante influencia en las diferencias alélicas neutrales entre poblaciones. Los modelos neutrales de deriva genética y mutación han sido aplicados también a caracteres cuantitativos (Lande 1976, Orr 1998). Usando este tipo de análisis, muchos estudios han sugerido que muchos aspectos en la evolución de la morfología facial de los primeros representantes del género *Homo* podrían ser más consistentes con la deriva genética que con la selección natural (Ackerman & Cheverud 2004, Weaver et al. 2007). A menudo, el tamaño efectivo de una población se infiere de su composición genética, basándose en modelos teóricos que incluyen la deriva genética (Hey 2005). Aunque hay métodos para estimar cambios en el tamaño poblacional efectivo a lo largo de toda una genealogía, han sido aplicados principalmente a patógenos más que a poblaciones humanas (Shackelton et al. 2006).

Las inferencias moleculares relativas a la historia de las poblaciones humanas son objeto de investigación, dando como resultado datos que no pueden ser obtenidos fácilmente de otras fuentes como fósiles o lingüísticas. Sin embargo, dado que tanto la deriva como la mutación son procesos estocásticos, la inferencias relativas a la historia poblacional solo tienen robustez cuando se analizan múltiples genes en un marco de referencia que tenga en cuenta dicha estocasticidad. Basándose en el análisis de 25 regiones génicas, Templeton (2005) ha planteado numerosas expansiones durante los últimos 2 millones de años de evolución humana. De manera similar, Gherman et al. (2007) sugieren que una explosión inusualmente alta de cambios a gran escala en el genoma humano ocurrió hace 54 millones de años, atribuibles a altas tasas de deriva en el ancestro común de los monos del Nuevo Mundo (Platyrrhini) y el clado que incluye a los monos del Viejo Mundo y a Hominoidea (Catarrhini).

Bases de datos en el estudio de la variación genética

Las bases de datos son esenciales para la investigación y aplicaciones de bioinformática. Existen muchas bases de datos que cubren diversos tipos de información: por ejemplo, secuencias de ADN y proteínas, estructuras moleculares, fenotipos y biodiversidad. Las bases de datos pueden contener datos empíricos (obtenidos directamente de los experimentos), datos pronosticados (obtenidos del análisis) o, con mayor frecuencia, ambos. Pueden ser específicos de un organismo, ruta o molécula de interés en particular. Contienen información de áreas de investigación que incluyen genómica, proteómica, metabolómica, expresión génica de *microarrays* y filogenia (Altman 2004). La información contenida en las bases de datos biológicas incluye la función del gen, la estructura, la localización (tanto celular como cromosómica), los efectos clínicos de las mutaciones, así como las similitudes de las secuencias y estructuras biológicas. También pueden incorporar datos compilados de muchas otras bases de datos. Estas bases de datos varían en su formato, mecanismo de acceso y si son públicas o no. Las bases de datos biológicas se pueden clasificar en bases de datos secuenciales, estructurales y funcionales. Las secuencias de ácidos nucleicos y proteínas se almacenan en bases de datos de secuencias y bases de datos de estructuras almacenan estructuras resueltas de ARN y proteínas. Las bases de datos funcionales proporcionan información sobre la función fisiológica de los productos génicos, por ejemplo, actividades enzimáticas, fenotipos mutantes o vías biológicas. Las bases de datos de organismos modelo son bases de datos funcionales que proporcionan datos específicos de cada especie. Las bases de datos son herramientas importantes para ayudar a los científicos a analizar y explicar una serie de fenómenos biológicos de la estructura de las biomoléculas y su interacción, a todo el metabolismo de los organismos y a la comprensión de la evolución de las especies. Este conocimiento ayuda a facilitar la lucha contra las enfermedades, ayuda al desarrollo de medicamentos, a predecir ciertas enfermedades genéticas y a descubrir las relaciones básicas entre las especies.

El conocimiento biológico se distribuye entre muchas bases de datos generales y especializadas diferentes. Esto a veces hace que sea difícil garantizar la coherencia de la información. La bioinformática integrativa es un campo que intenta abordar este problema proporcionando acceso unificado. Una solución es cómo las bases de datos biológicas hacen uso de referencias cruzadas a otras bases de datos con códigos de acceso estandarizados para vincular sus conocimientos relacionados entre sí.

Los conceptos de base de datos relacional en ciencias informáticas y búsqueda y recuperación de información en bibliotecas digitales son importantes para comprender las bases de datos biológicos. El diseño de bases de datos biológicos, el desarrollo y la gestión a largo plazo es un área central de la bioinformática (Bourne 2005). El contenido de los datos incluye secuencias de genes, descripciones textuales, atributos y clasificaciones ontológicas, citas y datos tabulares. A menudo se describen como datos semiestructurados y se pueden representar como tablas, registros delimitados por claves y estructuras XML.

Tipos de bases de datos

Hay dos tipos comunes de bases de datos biológicas: bases de datos primarias y bases de datos secundarias. Estos dos difieren en su estructura de archivo. Las bases de datos primarias a menudo contienen solo un tipo de datos específicos que se almacenan en su propio archivo. Cargan datos nuevos analizados en experimentos y actualizan sus entradas para garantizar la calidad de los datos. Las bases de datos secundarias son bases de datos que utilizan otras bases de datos como fuente de información, por lo que obtienen sus datos solicitando otras bases de datos. A menudo procesan o analizan los datos que coinciden con la solicitud correspondiente para obtener nuevos resultados.

La mayoría de las bases de datos biológicas están disponibles a través de sitios web que organizan los datos de modo que los usuarios puedan navegar por los datos en línea. La revista Nucleic Acids Research regularmente publica números especiales sobre bases de datos biológicos y tiene una lista de tales bases de datos. El número de 2018 tiene una lista de cerca de 180 de tales bases de datos y actualizaciones de las bases de datos descritas anteriormente (Rigden & Fernández 2018).

Las metabases de datos son bases de datos de bases de datos que recopilan información de bases de datos para generar nuevos datos. Son capaces de fusionar información de diferentes fuentes y ponerla a disposición de una forma nueva y más conveniente, o con énfasis en una enfermedad u organismo en particular, como hace Entrez (NCBI Resource Coordinators 2012), por ejemplo. Las bases de datos de ácidos nucleicos pueden clasificarse en base de datos de ADN y bases de datos de expresión génica.

A su vez, las bases de datos de ADN pueden separarse en primarias y secundarias, según sean bases de datos con repositorios propios o consigan datos de terceros. Entre las bases de datos de ADN primarias cabe destacar DDBJ (*Dna Data Bank of Japan*) del *National Institute of Genetics* en Japón, GenBank del *National Center for Biotechnology Information* (NCBI) en EE.UU y *European Nucleotide Archive* del EMBL-EBI (European Molecular Biology Laboratory - European Bioinformatics Institute) en Europa son repositorios de datos de secuencias de nucleótidos de todos los organismos. Los tres aceptan envíos de secuencias de nucleótidos y luego intercambian datos nuevos y actualizados diariamente para lograr una sincronización óptima entre ellos. Estas tres bases de datos son bases de datos primarias, ya que contienen datos de secuencias originales. Colaboran con *Sequence Read Archive* (SRA), que archiva lecturas sin procesar de instrumentos de secuenciación de alto rendimiento. Entre las bases de datos secundarias destacan:

- HapMap: El *International HapMap Project* fue una organización que tuvo como objetivo desarrollar un mapa de haplotipos (HapMap) del genoma humano, para describir los patrones comunes de la variación genética humana. HapMap se usa para encontrar variantes genéticas que afectan la salud, la enfermedad y las respuestas a las drogas y los factores ambientales.

- OMIM: es un catálogo continuamente actualizado de genes humanos y trastornos y rasgos genéticos, con un enfoque particular en la relación gen-fenotipo. OMIM es producido y actualizado en la Escuela de Medicina Johns Hopkins (JHUSOM).
- RefSeq: La base de datos RefSeq es una colección de acceso público de secuencias de nucleótidos (ADN y ARN) y sus productos proteínicos. Esta base de datos está construida por el *National Center for Biotechnology Information* (NCBI) y, a diferencia de GenBank, proporciona solo un registro para cada molécula biológica natural (es decir, ADN, ARN o proteína) para organismos principales que van desde virus hasta bacterias y eucariotas.

Además, los datos subyacentes generalmente están disponibles para su descarga en una variedad de formatos. Los datos biológicos vienen en muchos formatos. Estos formatos incluyen texto (PubMed, OMIM,...), datos de secuencia (GenBank, UniProt,...), estructura de proteínas (PDB, SCOP,...) y enlaces.

Una base de datos de expresión génica, generalmente analizada a partir de datos obtenidos de micromatrices de ADN, es un repositorio cuyos usos clave son almacenar los datos de medición de micromatrices, administrar un índice de búsqueda y poner los datos a disposición de otras aplicaciones para su análisis e interpretación. Estas bases de datos recopilan secuencias del genoma, las anotan y las analizan, y permiten su acceso al público. Estas bases de datos pueden contener muchos genomas de especies, o un genoma de organismo modelo único. Las más conocidas que incluyen datos de *Homo sapiens* son:

1000 Genomes Project: Lanzado en enero de 2008, fue un esfuerzo de un consorcio internacional de investigación para establecer el catálogo más detallado posible de la variación genética humana. Los científicos planearon secuenciar los genomas de al menos mil participantes anónimos de una serie de diferentes grupos étnicos dentro de los siguientes tres años, usando tecnologías recientemente desarrolladas que eran más rápidas y menos costosas (Durbin et al. 2010). McVean et al. (2012), anunciaron la secuenciación de 1092 genomas y su integración en el proyecto. En 2015, se informó de los resultados y la finalización del proyecto y de las oportunidades que el mismo podía ofrecer para futuras investigaciones (Auton et al. 2015). Se logró identificar muchas variaciones raras, restringidas a grupos estrechamente relacionados, y se analizaron ocho clases de variación estructural.

El proyecto une equipos de investigación multidisciplinarios de institutos de todo el mundo, incluidos China, Italia, Japón, Kenia, Nigeria, Perú, el Reino Unido y los Estados Unidos. Cada uno contribuye al conjunto de datos de secuencias y al mapa del genoma humano, que es de libre acceso a través de bases de datos públicas para la comunidad científica y el público en general (McVean et al. 2012). Al proporcionar una visión general de toda la variación genética humana, el consorcio generará una herramienta valiosa para todos los campos de la ciencia biológica, especialmente en las disciplinas de genética, medicina, farmacología, bioquímica y bioinformática.

El objetivo principal de este proyecto es crear un catálogo completo y detallado de las variaciones genéticas humanas, que a su vez se puede utilizar para los estudios de asociación que

relacionan la variación genética con la enfermedad. Al hacerlo, el consorcio busca descubrir más del 95% de las variantes (NP, CNV, indels) con frecuencias alélicas menores, de hasta el 1%, en todo el genoma y de entre 0,1% y 0,5% en regiones génicas, así como para estimar las frecuencias poblacionales, el *background* de los haplotipos, y los patrones de desequilibrio de ligamiento de alelos variantes.

Ensembl: es un proyecto científico conjunto entre el Instituto Europeo de Bioinformática (EBI) y el Instituto Wellcome Trust Sanger, que se lanzó en 1999 en respuesta a la finalización del Proyecto del Genoma Humano (Flicek, Amode, Barrell et al. 2010). Ensembl tiene como objetivo proporcionar un recurso centralizado para genetistas, biólogos moleculares y otros investigadores que estudian los genomas de nuestra propia especie y otros vertebrados y organismos modelo (Flicek, Aken, Ballester et al. 2010).

En el proyecto Ensembl, los datos de secuencia se introducen en el sistema de anotación de genes que crea un conjunto de ubicaciones pronosticadas de genes y los guarda en una base de datos MySQL para su posterior análisis y visualización. Estos datos son de acceso libre para la comunidad investigadora mundial. Todos los datos y códigos producidos por el proyecto Ensembl están disponibles para descargar (Ruffier et al. 2017), y también hay un servidor de base de datos de acceso público que permite el acceso remoto. Además, el sitio web de Ensembl proporciona visualizaciones generadas por ordenador de gran parte de los datos. Con el tiempo, el proyecto se ha ampliado para incluir especies adicionales (incluidos organismos modelo clave como el ratón, la mosca de la fruta y el pez cebra), así como una gama más amplia de datos genómicos, incluidas las variaciones genéticas y las características reglamentarias. Desde abril de 2009, un proyecto hermano, Ensembl Genomes, ha ampliado el alcance de Ensembl en metazoos, plantas, hongos, bacterias y protistas de invertebrados, mientras que el proyecto original continúa centrándose en los vertebrados.

SNPedia: es un sitio web bioinformático basado en concepto wiki (una comunidad virtual, cuyas páginas pueden ser editadas directamente desde el navegador, donde los mismos usuarios crean, modifican o eliminan contenidos que, generalmente, comparten), que sirve como base de datos de polimorfismos de nucleótido único (SNP). Cada artículo sobre un SNP proporciona una breve descripción, enlaces a artículos científicos y sitios web de genómica personal, así como información de micromatrices sobre ese SNP (Cariaso & Lennon 2012). A fecha de 8 de agosto de 2018, hay 109238 SNPs registrados en la base de datos (SNPedia 2018). l número de SNP en SNPedia se ha duplicado aproximadamente una vez cada 14 meses desde agosto de 2007 (Cariaso & Lennon 2012). Un programa informático asociado llamado Promethease, también desarrollado por el equipo SNPedia, permite a los usuarios comparar los resultados de genética personal con la base de datos SNPedia, generando un informe con información sobre los atributos de una persona, como la propensión a enfermedades, en función de la presencia de SNP específicos dentro su genoma.

ENCODE: es un proyecto público de investigación que tiene como objetivo identificar elementos funcionales en el genoma humano (Hong et al. 2016). Encode fue lanzado por el

Instituto Nacional de Investigación del Genoma Humano de los EE. UU. (NHGRI) en septiembre de 2003. Diseñado como continuación del Proyecto Genoma Humano, el proyecto ENCODE tiene como objetivo identificar todos los elementos funcionales en el genoma humano. El proyecto involucra a un consorcio mundial de grupos de investigación, y se puede acceder a los datos generados a partir de este proyecto a través de bases de datos públicas. El Instituto Nacional de Investigación del Genoma Humano (NHGRI) ha categorizado a ENCODE como un "proyecto de recursos comunitarios", es decir, como un proyecto de investigación específicamente diseñado e implementado para crear un conjunto de datos, reactivos u otro material cuya utilidad principal sea cumplir como un recurso para el conjunto de la comunidad científica. En consecuencia, la política de publicación de datos ENCODE estipula que los datos, una vez verificados, se depositarán en bases de datos públicas y se pondrán a disposición de todos para su uso sin restricciones (Raney et al. 2010).

Complejo Mayor de Histocompatibilidad (CMH)

El Complejo Mayor de Histocompatibilidad (CMH) es un conjunto de proteínas presentes en la superficie celular, esenciales para el reconocimiento de moléculas y agentes externos por parte del sistema inmunitario específico en los vertebrados, lo cual determina la histocompatibilidad. La función principal de las moléculas del CMH es unirse a los antígenos propios de patógenos, y señalizarlos en la superficie celular para que sean reconocidos por los linfocitos T (Janeway et al. 2001). Las moléculas de CMH median en la relación entre los leucocitos (también llamados glóbulos blancos), un tipo de células del sistema inmunitario, con otros leucocitos o células del organismo. El CMH determina la compatibilidad de los donantes en los trasplantes de órganos, así como la susceptibilidad frente a enfermedades autoinmunes. En humanos, el CMH se suele llamar HLA (acrónimo inglés de *Human Leukocyte Antigen)*.

En la célula, las moléculas de proteína del propio fenotipo del huésped o de otras entidades biológicas son continuamente sintetizadas y degradadas. Cada molécula del CMH en la superficie celular señaliza una fracción de una proteína, llamada epítopo. El antígeno presentado puede ser propio o ajeno, evitando así que el sistema inmune de un organismo apunte a sus propias células.

Organización de CMH

La familia de genes del CMH se divide en 3 subgrupos: clase I, clase II, y clase III. Cada subgrupo cuenta con unas subunidades específicas que son reconocidas por distintos tipos de correceptores. Las proteínas del subgrupo clase I tienen subunidades β2 que solo pueden ser reconocidas por correceptores CD8. Las de clase II tienen β1 y β2 y pueden ser reconocidas por correceptores CD4. De esta forma, las chaperonas asociadas a proteínas del CMH regulan que tipo de linfocitos se pueden unir a un antígeno dado con alta afinidad, dado que diferentes linfocitos expresan diferentes correceptores.

La diversidad antigénica, mediada por las clases I y II del CMH, se obtiene de al menos tres formas: (1) el repertorio proteico del CMH de un organismo es poligénico, es decir, se consigue a partir de múltiples genes interactuantes; (2) la expresión génica del CMH es codominante (de ambos conjuntos de alelos heredados de la generación parental); (3) las variantes génicas del CMH son altamente polimórficas, por lo que son altamente variables incluso intraespecíficamente (Janeway et al. 2001).

Las primeras descripciones del CMH fueron hechas por Peter Gorer (1936). Los genes del CMH fueron identificados por primera vez en líneas de ratones consanguíneos. Se trasplantaron tejidos tumorales entre distintas líneas y se vio que había rechazo de los tejidos trasplantados del donante al receptor (Little 1941). Snell y Higgings (1951) consiguieron identificar un locus de un gen del CMH mediante entrecruzamiento de ratones.

De las tres clases identificadas del CMH, la atención se suele centrar en las clases I y II. Mediante la interacción con las glicoproteínas CD4 de la superficie de los linfocitos T colaboradores (también llamados Linfocitos T CD4+), la clase II del CMH media en el establecimiento de la inmunidad específica (también llamada inmunidad adquirida o inmunidad adaptativa). Mediante la interacción con la glicoproteína transmembrana CD8 de la superficie de los linfocitos T citotóxicos (o CTL, por sus siglas en inglés Citolytic T Lymphocyte), la clase I del CMH, media en la destrucción de células del huésped infectadas o que sean identificadas como malignas, esto es, el aspecto de inmunidad específica denominada inmunidad celular.

Funciones del CMH

El CMH y especialmente los genes HLA se encuentran implicados en la respuesta inmune. Actualmente se encuentran registrados más de 12.000 alelos de clase I y más de 4.000 alelos de clase II. Este nivel de hiperpolimorfismo se ha atribuido a un modelo de selección balanceada inducida por parásitos (Garamszegi y Nunn, 2011) e influenciado por la selección sexual. En efecto, los agentes infecciosos han constituido probablemente una de las presiones selectivas más intensas en el modelado de nuestro patrimonio genético en el pasado (Karlsson, Kwiatkowski & Sabeti 2014), lo cual proporciona un mayor *fitness* y una mejor defensa contra patógenos con un genoma altamente mutable, produciendo una progenie con CMH heterocigoto (Singh 2001), pero todavía en la actualidad suponen el 64% de las muertes de menores de 5 años (Liu et al. 2012). Está por ver, no obstante, si en las poblaciones con una asistencia médica eficaz las presiones selectivas continúan actuando de forma perceptible. El CMH es el sistema antígeno-tejido que permite al sistema inmune (más específicamente a las células T) unirse, reconocerse y tolerarse (autoreconocerse). El CMH interactúa con los receptores de los linfocitos T (TCR) y sus co-receptores para optimizar las condiciones de unión para la interacción TCR-antígeno, en términos de afinidad y especificidad de unión a antígeno, y eficacia de transducción de señales. Esencialmente, el complejo CMH-péptido es un complejo de autoantígeno / aloantígeno. Tras la unión, los linfocitos T deben, en principio, tolerar el auto-antígeno, pero activarse cuando se

exponen al aloantígeno. Los estados de enfermedad ocurren cuando este principio es interrumpido.

Presentación del antígeno: Las moléculas de CMH se unen tanto al receptor de células T como a los co-receptores CD4 / CD8 en linfocitos T, y el epítopo antigénico contenido en el surco de unión al péptido de la molécula del CMH interactúa con el dominio de la inmunoglobulina del TCR para desencadenar la activación de los linfocitos T.

Reacción autoinmune: Tener algunas moléculas del CMH aumenta el riesgo de enfermedades autoinmunes más que tener otras. El antígeno de clase I HLA-B27 es un ejemplo. No está claro exactamente cómo tener HLA-B27 aumenta el riesgo de espondilitis anquilosante y otras enfermedades inflamatorias asociadas (Kataria y Brent 2004), pero se han planteado hipótesis de mecanismos que implican presentación de antígenos aberrantes o activación de células T (Hacquard-Bouder, Ittah & Breban 2005).

Aloreconocimiento de tejidos: Las moléculas de CMH, en complejo con epítopos peptídicos, son esencialmente ligandos para los TCR. Las células T se activan por su unión a péptidos de cualquier molécula de CMH, las cuales no fueran “entrenadas” para reconocer durante la selección positiva (también llamado reconocimiento restrictivo del antígeno) del timo.

Emparejamiento selectivo y CMH

El término emparejamiento selectivo hace referencia al fenómeno por el cual un sujeto tiende a emparejarse con otros individuos que se asemejan a él en algún aspecto. El emparejamiento selectivo en humanos ha sido ampliamente observado y estudiado, y dentro de los tipos que pueden observarse se encuentra el emparejamiento selectivo basado en el genotipo y la expresión fenotípica.

Pearson et al. (1903) observaron una alta correlación entre la altura, la longitud de los brazos, y del antebrazo izquierdo en 1000 parejas. Más recientemente (Kocsor et al. 2011) han descrito como los hombres prefieren mujeres cuya apariencia facial se parece a la suya propia, cuando se les presentan tres opciones a elegir entre las que se encuentra una modificada para que se parezca a la suya propia. Sin embargo, no ocurrió lo mismo con las mujeres. Pero Hedrick (1992) había observado, a través de un modelo con ratones, que la elección de pareja llevada a cabo por las hembras podía reducir la proporción observada de heterocigotos, contribuir al mantenimiento del polimorfismo, influir en las frecuencias del tipo de apareamiento y generar desequilibrio gamético.

Estudios realizados en EEUU concluyeron que el emparejamiento selectivo basado en similitudes genéticas juega un papel importante. Se observó que las parejas estudiadas establecidas con anterioridad eran más similares genéticamente que cualquier pareja de individuos escogidos al azar (Guo el al. 2014). Algunos investigadores argumentan que este

emparejamiento selectivo es debido exclusivamente a la estratificación poblacional, el hecho de que las personas se emparejen de manera más probable con gente de su mismo grupo étnico (Abdellaoui, Verweij & Zietsch 2014).

Hay una serie de evidencias que parecen sustentar el papel de la selección sexual sobre el mantenimiento de los altos niveles de heterocigosidad en los genes del CMH (Ejsmond, Radwan & Wilson 2014). El mecanismo impulsado por la reproducción sexual funciona a través de las preferencias de elección de la pareja dependiente del CMH antes del apareamiento y, después del apareamiento, influye en el resultado exitoso del embarazo. Este efecto no se aplica solo tras la implantación y durante el desarrollo embrionario, sino que también opera controlando el procesa de implantación embrionario (Capittini, Martinetti & Cuccia 2008). La proporción de resultados fallidos alcanza el 80% del total de embarazos si consideramos los abortos desde las primeras etapas después de la fecundación (Apanius et al. 1997, Gilbert et al. 1998, Racowsky 2002).

En ratones, se ha observado un comportamiento selectivo particular determinado por el CMH, el "efecto Bruce": una hembra preñada, que tenga la oportunidad de aparearse con una segunda pareja que difiera en los loci del CMH, abortará en favor del macho que sea más diferente en cuanto al CMH (Yamazaki et al. 1983). Así, el bloqueo del embarazo también puede incorporarse al esquema del polimorfismo en el CMH determinado por la presión selectiva de los patógenos propuesto por Potts y Wakeland (1993), ya que serviría para evitar la endogamia, aumentando la heterocigosidad del CMH. Se han propuesto 3 mecanismos de actuación de la búsqueda selectiva del cónyuge ligada al genotipo CMH (Piertney y Oliver, 2006): a) Preferencia por una pareja con alto grado de heterocigosis (y por tanto de diversidad), b) Preferencia por individuos portadores de genotipos específicos con protección contra determinados agentes infecciosos y c) Preferencia por un cónyuge de genotipo diferente. En un meta análisis reciente (Kamiya et al. 2014) se han encontrado evidencias de selección de pareja por diversidad MHC en varias especies no humanas. En humanos, sin embargo, los resultados han sido contradictorios (Winternitz y Abbate, 2015), aunque entre los diferentes modelos, también parece encontrarse una significativa tendencia a seleccionar parejas con un alto grado de diversidad (Winternitz et al. 2013, Winternitz et al. 2017). Chaix, Cao & Donnelly (2008) observaron en los individuos que estudiaron mostraban emparejamiento selectivo negativo (también llamado heterogamia) para genes de la región del CMH en el cromosoma 6. Los individuos se sienten más atraídos por los olores de otros individuos que son genéticamente diferentes en esta región (Wedekind et al. 1995, Wedekind et al. 1997). Los investigadores concluyen que esto explica la abundancia de heterocigosis en los niños, haciéndoles menos vulnerables a los patógenos. Sin embargo, estos estudios fueron muy criticados en base a sus procedimientos metodológicos (Hedrick & Loeschcke 1996). Sin embargo, Winking et al. (2014) concluyeron que los patrones de emparejamiento selectivo negativo asociados a olores o preferencias por la pareja, no se desvían significativamente del equilibrio Hardy-Weinberg.

Se han desarrollado diversos test para estudiar la relación entre el emparejamiento selectivo y los genes HLA (Jin, Speed & Thompson 1995; Génin et al. 2000), aunque son escasos y a veces contradictorios, y únicamente para grupos poblacionales pequeños y endogámicos, o para

poblaciones mixtas (Ober et al. 1997). En este estudio se encontraron evidencias de que entre los hutteritas (comunidad religiosa de EEUU) se evitaban los matrimonios entre personas que presentaban el mismo genotipo. Sin embargo, en un estudio similar realizado por Hedrick y Black (1997) en nativos sudamericanos, no se encontró evidencia de este emparejamiento no aleatorio. En relación a los escasos resultados significativos observados en nuestra especie en relación a la selección sexual asociada al CMH, se han propuesto varios factores que podrían afectar a la validez de los resultados. Entre ellos, destaca la tendencia al mestizaje entre individuos de diferentes poblaciones, lo que podría generar artefactos en la detección de modelos de búsqueda selectiva del cónyuge por estratificación poblacional, así como diferentes comportamientos asociados al patrimonio cultural, como la higiene o la cosmética (Winternitz et al. 2017).

Proceso de colonización de *Homo sapiens*

Como se ha señalado anteriormente, el efecto fundador asociado a un tamaño poblacional original pequeño (lo cual se asocia a un aumento de probabilidad de que ocurran procesos de deriva genética), puede haber tenido una importante influencia en las diferencias alélicas neutrales entre poblaciones. El modelo explicativo del origen geográfico de los humanos anatómicamente modernos denominado “Out of Africa” apoya esta idea.

El origen reciente de los humanos modernos (también llamado teoría del “Out of Africa”, hipótesis del origen único reciente, hipótesis de la sustitución o modelo del origen africano reciente) es, en paleoantropología, el modelo dominante sobre el origen y migración temprana de los humanos anatómicamente modernos (*Homo sapiens*), el cual propone un área de origen única para los humanos modernos. De acuerdo con este modelo, los humanos habrían evolucionado en el este de África y empezaron a expandirse por el mundo hace entre 50000 y 100000 años (Stringer 2003, Liu et al. 2006).

La principal hipótesis que se opone a la del “*Out of Africa*” es la hipótesis multirregional del origen de los humanos modernos, que postula una migración temprana de *Homo sapiens* africanos que se habrían cruzado con poblaciones locales de *Homo erectus* en diversas regiones del mundo (Wolpoff et al. 2000, Jurmain et al. 2008).

Sin embargo, la situación es más complicada. Prüfer et al. (2014) exponen evidencias de que hubo entrecruzamiento entre humanos modernos fuera de África con Neandertales y Denisovanos. Varios estudios estipulan que hubo hasta dos procesos migratorios (Macaulay et al. 2005, Beyin 2011) La primera habría tenido lugar entre hace 130000 y 115000 años a través del norte de África (Smith 2007, Armitage et al. 2011, Balter 2011, Cruciani et al. 2011), y al parecer los individuos habrían muerto o se habrían retirado, aunque hay pruebas de la presencia de humanos modernos en China hace 80000 años (Liu et al. 20015). Una segunda migración habría tenido lugar por la llamada ruta del Sur, siguiendo la línea de la costa sur de Asia, que llevó a la colonización de Eurasia y Australia hace alrededor de 50000 años. La evidencia más temprana de la presencia de humanos en Australia es de al menos hace 65000 años (Clarkson et al. 2017). De acuerdo con esta

teoría, Europa fue colonizada bien por una migración posterior desde la India, que fue repoblada desde el sudeste de Asia, o por una rama temprana de esta segunda migración que pobló Oriente Próximo y Europa (Macaulay et al. 2005, Posth et al. 2016). Aunque recientemente, se ha sugerido que *H. sapiens* podría haber migrado de África hace 270000 años (Posth et al. 2017).

Orígenes africanos de Homo sapiens

La cronología del origen de los humanos anatómicamente modernos es un tema complicado. Hublin et al. (2017) han descrito restos de *Homo sapiens* modernos de entre hace 350000 y 280000 años, que les ha servido para apoyar una hipótesis del origen panafricano de *H. sapiens*. En general, se acepta que los seres humanos anatómicamente modernos surgieron en África alrededor de hace 200000 años. Delson et al. (2000) expusieron que la tendencia en expansión craneal y la tecnología lítica achelense que ocurrió entre hace 400000 años y el segundo periodo interglaciar en el Pleistoceno Medio (hace aproximadamente 250000 años), proporcionan pruebas de una transición de *Homo erectus* a *H. sapiens*. En la teoría del Out of Africa, la migración de *H. sapiens*, tanto dentro como fuera de África, eventualmente reemplazó a las poblaciones dispersas de *H. erectus* que estaban presentes con anterioridad. Hace 100000 años, empiezan a surgir evidencias de una tecnología más sofisticada, y hace 50000 años pruebas de comportamiento inequívocamente asociable a H. sapiens modernos se convierten en dominantes. Las herramientas líticas muestran patrones regulares que son reproducidos o duplicados con más precisión que antes, y las herramientas hechas de hueso y asta aparecen por primera vez (Hoffecker 2009, Tattersall 2009).

Teoría de la “ruta de la costa”

La teoría de la “ruta de la costa” es principalmente usada para describir el poblamiento inicial de la Península Arábiga, India, el sudeste de Asia, Nueva Guinea, Australia, parte de Oceanía, la costa de China, y Japón (Wells 2003, Pope & Terrel 2007). Está ligado a la presencia y dispersión de los haplogrupos M y N del ADN mitocondrial, así como los patrones de distribución específicos de los haplogrupos C y D del ADN del cromosoma Y en estas regiones (Macaulay et al. 2005, Mirazón Lahr et al. 2012). La teoría propone que los primero humanos, de manera similar a los negritos del sudeste de Asia o los proto-Australoides actuales, presentes ya en las regiones costeras del sur de la India continental, se expandieron a las islas Andaman e Indonesia, y más tarde se ramificaron al sur hacia Australia, y al norte hacia Japón (Wells 2003). Hace unos 70000 años, parte de los portadores del haplogrupo mitocondrial L3 migraron del este de África a Oriente Próximo. Se ha estimado que de una población de entre 2000 y 5000 individuos en África, solo un pequeño grupo de entre 150 y 1000 personas, cruzaron el Mar Rojo (Zhivotovsky et al. 2003, Stix 2008). Este grupo viajó a lo largo de la ruta de costa de Arabia y Persia hacia la India, que parece ser el primer punto de asentamiento importante (Metspalu et al. 2004). Wells (2003) argumenta que la ruta se extendía 250 kilómetros por la ruta de la costa sur de Asia, llegando a Australia alrededor de hace 50000 años.

Hoy en día, en el estrecho de Bab-el-Mandeb, el Mar Rojo tiene una anchura de 20 kilómetros, pero hace 50000 años el nivel del mar era 70 metros menor, debido a la glaciación, y el mar era más angosto. Aunque el estrecho nunca se llegó a cerrar por completo, la separación fue lo suficientemente pequeña, y la posibilidad de la existencia de islas o islotes en el camino, habrían permitido el paso mediante el uso de balsas simples (Beyin 2011, Fernandes et. al. 2006). El hallazgo de depósitos de restos de conchas en Eritrea datados en 125000 años (Walter 2000), indica que la dieta de los primeros humanos habría incluido marisco obtenido de la zona intermareal.

Dispersión temprana por el norte de África

La migración temprana por el norte de África tuvo lugar hace entre 130000 y 115000 años. El descubrimiento de herramientas líticas en los Emiratos Árabes Unidos sirvió para probar la presencia de humanos modernos hace entre 100000 y 125000 años (Armitage et al. 2011), volviendo a poner de actualidad la anteriormente controvertida ruta del norte de África (Smith 2007, Balter 2011, Cruciani et al. 2011, Scerri et al. 2014).

Se han encontrado fósiles de *H. sapiens* en Qafzeh (Israel) datados hace entre 80000 y 100000 años. Parece ser que estos humanos se extinguieron o se retiraron de vuelta a África hace entre 70000 y 80000 años, siendo posiblemente reemplazados por los Neandertales ubicados más al sur, que trataban de escapar del clima más frío de Europa (Finlayson 2009). Liu et al. (2006) analizaron marcadores autosómicos microsatélites que dataron en 56000 años. Llegaron a la conclusión de que los fósiles de Qafzeh son una ramificación aislada temprana que se retiró de vuelta a África.

De acuerdo a Kuhlwilm et al. (2016) los Neandertales recibieron flujo genético hace alrededor de 100000 años, de un grupo de humanos que se separó de otros humanos modernos hace alrededor de 200000 años. Argumentan que los ancestros de los Neandertales de las montañas Altai y los humanos modernos tempranos se encontraron y entrecruzaron, posiblemente en Oriente Próximo, varios miles de años antes de lo anteriormente pensado, y que esto complementa la evidencia arqueológica sobre la presencia de humanos modernos fuera de África hace más de 100000 años, proporcionando la primera prueba genética de dichas poblaciones.

Shen et al. (2002) cuestionan la extinción de esta dispersión temprana. Datan los restos del hombre de Liujang hace entre 111000 y 139000 años. Liu et al. (2015) argumentan la presencia de humanos modernos en China por el hallazgo de dientes de humanos modernos.

Datación: pre- o post- Toba

La datación de la cronología asociada a la ruta del Sur está en disputa (Appenzeller 2012). Puede haber ocurrido antes o después de la catástrofe de Toba. La catástrofe de Toba fue una supererupción volcánica que ocurrió hace entre 69000 y 77000 años en lo que actualmente es el lago Toba (Sumatra, Indonesia), que se formó a raíz de la propia erupción. La erupción causó un invierno volcánico global que duró entre 6 y 10 años, y conllevo un periodo de enfriamiento de 1000 años. En 1993, Gibbons sugirió la relación entre la erupción y un proceso de cuello de botella en la población humana, e investigadores como Rampino & Self (1993) o Ambrose (1998) apoyaron la idea. Sin embargo, tanto la teoría del invierno volcánico como su relación con el proceso de deriva en la población humana es altamente controvertida y ha sido discutida intensivamente (Oppenheimer 2002, Choi 2013). Las herramientas líticas descubiertas bajo capas de ceniza depositadas en la India parecen apuntar a que la dispersión fue pre-Toba, pero el origen de las herramientas está en disputa (Appenzeller 2012). Por otro lado, el haplogrupo L3 sugiere una expansión post-Toba, dado que este haplogrupo se originó antes de la expansión de los humanos fuera de África, y puede ser datado hasta hace 60000-70000 años, sugiriendo que la expansión ocurrió algunos miles de años después de la catástrofe de Toba (Appenzeller 2012). Brahic (2012) publicó un artículo sobre una nueva investigación que mostraba que algunas mutaciones en el ADN humano habrían ocurrido más lentamente de lo ocurrido, e incluyendo una revisión en la datación para la migración fuera de África, ubicándola en una horquilla de hace 90000 y 130000 años.

Entrecruzamiento entre humanos arcaicos y modernos

Existe evidencia de mestizaje entre humanos arcaicos y modernos durante el Paleolítico Medio y el Paleolítico Superior temprano. El entrecruzamiento ocurrió en varios eventos independientes que incluyeron neandertales, denisovanos, así como varios homínidos no identificados. En Eurasia, el entrecruzamiento entre neandertales y denisovanos con humanos modernos tuvo lugar varias veces hace entre 100.000 y 40.000 años, y tanto antes como después de la reciente migración fuera de África hace 70.000 años. A través de la secuenciación del genoma completo de tres neandertales de Vindija (Green et al. 2010) se reveló que los neandertales compartían más alelos con poblaciones de Eurasia que con las poblaciones de África subsahariana. El exceso observado de similitud genética se explica mejor por el flujo génico reciente de los neandertales a los humanos modernos después de la migración fuera de África. Se ha hallado ADN derivado de neandertal en el genoma de las poblaciones contemporáneas de Europa y Asia, y se estima que representa entre el 1% y el 6% de los genomas modernos. El genoma de la mayoría de europeos y asiáticos tienen entre el 1% y el 4% de ADN neandertal (Green et al. 2010). Prüfer et al. (2014) estimaron una proporción de entre el 1,5% y un 2,1%, pero luego fue revisado a un intervalo entre el 1,8% y el 2,6% (Prüfer et al. 2017). El mismo estudio observó que los asiáticos orientales tienen más ADN neandertal (2,3-2,6%) que los euroasiáticos occidentales (1,8-2,4%). Lohse & Frantz (2014) infieren una tasa aún mayor de 3,4-7,3%. Las tasas más altas de mezcla arcaica en se han encontrado en las poblaciones indígenas del sudeste asiático y Oceanía, con un

4-7% estimado del genoma de los melanesios (Meyer et al. 2012) o los aborígenes australianos (Rasmuseen et al. 2011) modernos derivado de denisovanos. La ancestría derivada de neandertales y denisovanos está significativamente ausente de la mayoría de las poblaciones modernas en el África subsahariana. Sin embargo, se han encontrado alelos arcaicos consistentes con varios eventos de mezcla independientes. En ciertas poblaciones de África occidental, las tasas de mezcla parecen ser significativamente más altas que en Eurasia. La mezcla de un linaje basal de humanos arcaicos de África occidental en los Mende de Sierra Leona se ha estimado en un 13% (Skoglund et al. 2017).

Estudiando una secuencia de genoma de alta calidad de una mujer neandertal de Altai (Siberia), se ha encontrado que el componente de neandertal en humanos modernos no africanos está más relacionado con el neandertal de Mezmaiskaya (Cáucaso), seguido por los neandertales de Vindija (Croacia) y, por último, que con el neandertal de Altai (Siberia). Estos resultados sugieren que la mayoría de la mezcla en humanos modernos provino de poblaciones de Neanderthal que se habían separado (alrededor de 80-100 kya) de los linajes neandertal Vindija y Mezmaiskaya antes de que las dos últimas divergieran entre sí (Prüfer et al. 2014).

Al analizar el cromosoma 21 de los neandertales de Altai (Siberia), El Sidrón (España) y Vindija (Croacia), se determina que, de estos tres linajes, solo los neandertales de El Sidrón y Vindija muestran tasas significativas de flujo génico (0,3-2,6 %) hacia los humanos modernos, lo que sugiere que los neandertales de El Sidrón y Vindija están más estrechamente relacionados que los neandertales de Altai con los neandertales que se cruzaron con los humanos modernos hace unos 47000-65000 años (Kuhlwilm et al. 2016). Y de forma contraria, también se ha determinado que se dieron tasas significativas de flujo génico de humanos modernos a neandertales solo para el neandertal de Altai (0,1-2,1%), lo que sugiere que el flujo de genes humanos modernos en los neandertales se produjo principalmente después de la separación de los neandertales de Altai de los neandertales El Sidrón y Vindija hace aproximadamente 110000 años. Los resultados muestran que la fuente del flujo génico de humanos modernos hacia los neandertales se originó a partir de una población de humanos modernos hace unos 100.000 años, anterior a la migración fuera de África de los antepasados humanos modernos de los no africanos actuales.

Introgresión

La introgresión, también conocida como hibridación introgresiva, es el movimiento de un gen (flujo génico) de una especie a otra mediante el retrocruzamiento repetido de un híbrido interespecífico con una de sus especies parentales. Es un proceso a largo plazo, ya que pueden ser necesarias muchas generaciones híbridas antes de que ocurra el retrocruzamiento. La introgresión difiere de la hibridación simple. La introgresión da como resultado una mezcla compleja de genes parentales, mientras que la hibridación simple da como resultado una mezcla más uniforme, que en la primera generación será una mezcla uniforme de dos especies parentales. La introgresión es una fuente importante de variación genética en poblaciones naturales y puede contribuir a la adaptación e incluso a la radiación adaptativa (Grant, Grant & Petren 2005). Existe evidencia de

que la introgresión es un fenómeno omnipresente en plantas, animales (Bullini 1994; Dowling & Secor 1997) e incluso en humanos (Holliday 2003) en la que puede haber introducido el haplogrupo D de la microcefalina (Evans et al. 2006). Existe una fuerte evidencia de la introgresión de los genes neandertal (Wills 2011) y los genes de denisovanos (Huerta-Sánchez et al. 2014) en partes del *pool* genético humanos modernos. Alrededor del 20% del genoma de neandertal se ha encontrado introgresado en la población humana moderna de asiáticos del este y europeos (Vernot & Akey 2014). En los indígenas papúes de Nueva Guinea, los alelos de neandertal introgresados se encuentran en mayor frecuencia en los genes expresados en el cerebro, mientras que los alelos de denisovanos tienen la mayor frecuencia en los genes expresados en los huesos y otros tejidos (Akkuratov, Gelfand & Khrameeva 2018).

El hecho de que se haya detectado una mezcla más alta de neandertales en los asiáticos orientales que en los europeos (Wall et al. 2013; Sankararaman et al. 2014; Nielsen et al. 2017), posiblemente podría explicarse por la ocurrencia de más eventos de mezcla en los antepasados tempranos de los asiáticos orientales después de la separación de los europeos y asiáticos orientales, dilución de la ascendencia neandertal en los europeos por mezcla con poblaciones con baja ascendencia neandertal a través de migraciones posteriores, o selección natural que pudo haber sido relativamente menor en los asiáticos orientales que en los europeos. Los estudios indican que una eficacia reducida de la selección purificadora contra alelos de Neanderthal en asiáticos orientales no podría explicar la mayor proporción de ascendencia neandertal de los asiáticos orientales, favoreciendo modelos más complejos que implican pulsos adicionales de introgresión neandertal en asiáticos orientales (Vernot & Akey 2015; Kim & Lohmueller 2015). También se ha observado que existe una variación pequeña pero significativa de las tasas de mezcla de neandertal dentro de las poblaciones europeas, pero ninguna variación significativa dentro de las poblaciones de Asia oriental (Vernot & Akey 2014).

El análisis genómico sugiere que existe una división global en la introgresión de neandertal entre las poblaciones de África subsahariana y otros grupos humanos modernos, incluidos los norteafricanos, en lugar de entre poblaciones africanas y no africanas (Sánchez-Quinto et al. 2012). Los grupos norteafricanos comparten un exceso similar de alelos derivados con los neandertales, al igual que las poblaciones no africanas, mientras que los grupos del África subsahariana son las únicas poblaciones humanas modernas que generalmente no experimentaron la mezcla de neandertales. Se encontró que la señal genética neandertal entre las poblaciones del norte de África varía en función de la cantidad relativa de ascendencia autóctona del norte de África, europea, de Oriente Próximo y subsahariana. Se observó que la mezcla inferida con neandertal era la más alta entre las poblaciones norteafricanas con ascendencia autóctona norteafricana máxima, como los bereberes tunecinos, donde estaba al mismo nivel o incluso más alto que las poblaciones de Eurasia (100 -138%); alto entre las poblaciones de África del Norte con una mayor mezcla europea o de Oriente Próximo, como grupos en el norte de Marruecos y Egipto (60-70%); y la más baja entre las poblaciones del norte de África con una mayor mezcla subsahariana, como en el sur de Marruecos (20%) (Sánchez-Quinto et al. 2012). Estos autores postulan que la presencia de esta señal genética neandertal en África no se debe al flujo génico reciente de poblaciones europeas o de Oriente Próximo, ya que es más alta entre las poblaciones que tienen ascendencia preneolítica

indígena del norte de África. El haplotipo B006 ligado a Neanderthal del gen de la distrofina también se ha encontrado entre los grupos de pastores nómadas en el Sahel, Etiopía y Burkina-Faso, que están asociados con las poblaciones del norte. En consecuencia, la presencia de este haplotipo B006 en la zona norte y noreste del África Subsahariana se atribuye al flujo de genes desde un punto de origen no africano (Yotova et al. 2011). También se han observado tasas bajas pero significativas de mezcla de neandertal para los masai de África Oriental. Después de identificar ascendencia africana y no africana entre los masai, se puede concluir que el reciente flujo genético humano no africano (post-Neandertal) fue la fuente de la contribución ya que alrededor del 30% del genoma masai se puede rastrear hasta introgresión no africana de hace aproximadamente 100 generaciones (Wall et al. 2013).

No se ha encontrado evidencia de ADN mitocondrial neandertal en humanos modernos (Krings et al. 1997; Serre et al. 2004; Wall & Hammer 2006). Esto sugeriría que la mezcla exitosa con los neandertales sucedió por vía paterna en vez de por vía materna neandertal (Mason & Short 2011; Wang, Farina & Li 2013). Posibles hipótesis son que el ADN mitocondrial de neandertal tuvo mutaciones perjudiciales que llevaron a la extinción de portadores, que la descendencia híbrida de las madres de neandertal se criaron en grupos de neandertal y se extinguieron con ellas, o que las hembras neandertales y machos sapiens no produjeron descendencia fértil (Mason & Short 2011).

Como se muestra en un modelo de mestizaje producido por Neves & Serva (2012), el grado de mezcla con neandertales observado en los humanos modernos puede haber sido causada aún con una tasa muy baja de cruzamiento entre humanos modernos y neandertales, con el intercambio de un par de individuos entre las dos poblaciones en aproximadamente 77 generaciones. Esta baja tasa de cruzamiento explicaría la ausencia de ADN mitocondrial de neandertal en el *pool* genético de humanos modernos mencionado anteriormente, ya que el modelo estima una probabilidad de solo un 7% para un origen neandertal tanto del ADN mitocondrial como del cromosoma Y en humanos modernos. Cabrera et al. (2018) sugieren que los genes neandertal observados en ciertas poblaciones modernas en África pueden haber sido traídos de Eurasia alrededor de 70 kya por machos que portan el haplogrupo E paterno y hembras portadoras del haplogrupo materno L3.

Se ha descubierto la existencia de grandes regiones genómicas con una contribución neandertal muy reducida en los humanos modernos debido a la selección negativa (Vernot & Akey 2014) causada en parte por la infertilidad masculina híbrida (Sankararaman et al. 2014). Estas grandes regiones de baja contribución de neandertal son más comunes en el cromosoma X, dado que su ascendencia neandertal es cinco veces menor que en los cromosomas autosómicos, y contenían un número relativamente alto de genes específicos para los testículos (Sankararaman et al. 2014). Esto significa que los humanos modernos tienen relativamente pocos genes de Neanderthal que se encuentran en el cromosoma X o se expresan en los testículos, lo que concuerda con el hecho de que la infertilidad masculina se ve afectada por una cantidad desproporcionadamente grande de genes en los cromosomas X. También se ha demostrado que la ascendencia neandertal se ha seleccionado en rutas biológicas conservadas, como el

procesamiento de ARN (Sankararaman et al. 2014). De acuerdo con la hipótesis de que la selección purificadora ha reducido la contribución de los neandertales en los genomas humanos modernos actuales, los humanos modernos eurasiáticos del Alto Paleolítico, como los restos del hombre de Tianyuan, portan más ADN de Neandertal (alrededor del 4-5%) que los humanos modernos de Eurasia (aproximadamente 1-2%) (Yang et al. 2017).

Se descubrió que los genes que afectan a la queratina han sido introgresados desde los neandertales a los humanos modernos, sugiriendo que estos genes dieron una adaptación morfológica en la piel y el cabello a los humanos modernos para hacer frente a ambientes no africanos (Vernot & Akey 2014). Esto también ocurre con varios genes implicados en el lupus eritematoso sistémico, la cirrosis biliar primaria, la enfermedad de Crohn, el tamaño del disco óptico, los niveles de interleucina 18 y la diabetes mellitus tipo 2 (Sankararaman et al. 2014). Ding et al. (2014) investigaron la introgresión neandertal en 18 genes, varios de los cuales están relacionados con la adaptación a la luz ultravioleta, dentro de la región del cromosoma 3p21.31 (región HYAL) de asiáticos orientales. Los haplotipos introgresivos se seleccionaron positivamente solo en poblaciones de Asia oriental, aumentando constantemente desde hace unos 45,000 años hasta un aumento repentino de la tasa de crecimiento hace entre 5,000 y 3,500 años. Los hallazgos también sugieren que esta introgresión de Neanderthal ocurrió dentro de la población ancestral compartida por asiáticos orientales y nativos americanos.

Evans et al. (2006) habían sugerido previamente que el haplogrupo D de microcefalina, un gen regulador crítico para el volumen cerebral, se originó a partir de una población humana arcaica. Basándose en la edad de coalescencia de los alelos D derivados, los resultados muestran que el haplogrupo D introgresó hace 37,000 años en humanos modernos a partir de una población humana arcaica que se separó hace 1,1 millones de años (basado en el tiempo de separación entre alelos D y no D), coherente con el período en que los neandertales y los humanos modernos coexistieron y divergieron respectivamente. La alta frecuencia del haplogrupo D (70%) sugiere que fue seleccionado positivamente en humanos modernos. La distribución del alelo D de la microcefalina es alta fuera de África, pero es baja en el África subsahariana, lo que sugiere que el evento de mezcla ocurrió en poblaciones arcaicas de Eurasia. Según Lari et al. (2010), esta diferencia de distribución entre África y Eurasia sugiere que el alelo D se originó de los neandertales, pero se descubrió que un individuo neandertal de la cueva de Mezzena, en Monti Lessini (Italia) era homocigoto para un alelo ancestral de microcefalina, por lo que no proporcionaba ningún apoyo a que los neandertales contribuyesen con el alelo D a los humanos modernos y tampoco excluía la posibilidad de un origen de neandertal del alelo D. Green et al. (2010), habiendo analizado a los neandertales de Vindija, tampoco pudieron confirmar un origen neandertal del haplogrupo D del gen de la microcefalina.

Se ha encontrado que los alelos HLA-A*02, A*26, A*66, B*07, B*51, C*07:02 y C*16:02 del sistema inmune introgresaron desde los neandertales a los humanos modernos (Abi-Rached et al. 2011). Después de emigrar de África, los humanos modernos se encontraron y se cruzaron con humanos arcaicos, lo que fue una ventaja para los humanos modernos al restaurar rápidamente la diversidad de HLA y adquirir nuevas variantes de HLA que se adapten mejor a los patógenos

locales. Los genes neandertal introgresados exhiben efectos reguladores en cis (regiones no codificantes del ADN que regulan la transcripción de los genes cercanos) en los humanos modernos, lo que contribuye a la complejidad genómica y la variación del fenotipo de los humanos modernos (McCoy, Wakefield & Akey 2017). Al observar individuos heterocigotos (portadores de Neandertal y versiones humanas modernas de un gen), se encontró que la expresión alelo específica de los alelos de neandertal introgresados era significativamente menor en el cerebro y los testículos en relación con otros tejidos. En el cerebro, esto fue más pronunciado en el cerebelo y los ganglios basales. Esta regulación negativa sugiere que los humanos modernos y los neandertales posiblemente experimentaron una tasa de divergencia relativamente más alta en estos tejidos específicos.

Estudiando el genoma de una hembra de neandertal de Vindija, Prüfer et al. (2017) identificaron diversas variantes genéticas derivadas de neandertal, incluidas aquellas que afectan los niveles de colesterol LDL y vitamina D, y tienen influencia sobre los trastornos alimentarios, la acumulación de grasa visceral, la artritis reumatoide, la esquizofrenia y la respuesta a fármacos antipsicóticos.

Examinando a los humanos modernos europeos en relación con el genoma del neandertal de Altai, los resultados muestran que la mezcla con neandertal se asocia con varios cambios en el cráneo y la morfología cerebral subyacente, sugiriendo cambios en la función neurológica a través de la variación genética derivada del neandertal (Gregory et al. 2017). La mezcla de neandertales se asocia con una expansión del área posterolateral del cráneo humano moderno, que se extiende desde los huesos parietales occipitales e inferiores a los locales temporales bilaterales. Con respecto a la morfología moderna del cerebro humano, la mezcla de Neanderthal se correlaciona positivamente con un aumento en la profundidad del surco intraparietal derecho y un aumento en la complejidad cortical para la corteza visual temprana del hemisferio izquierdo. La mezcla con neandertales también se correlaciona positivamente con un aumento en el volumen de materia blanca y gris localizado en la región parietal derecha adyacente al surco intraparietal derecho. En el área que se superpone a la circunvolución de la corteza visual primaria en el hemisferio izquierdo, la mezcla de neandertales se correlaciona positivamente con el volumen de materia gris. Los resultados también muestran evidencia de una correlación negativa entre la mezcla con neandertal y el volumen de materia blanca en la corteza orbitofrontal.

Aunque es menos parsimonioso que la idea de que se debe al flujo génico reciente, la observación de diferencias en la cantidad de ancestría neandertal entre Eurasia y África pudo deberse a la subestructura de la población antigua en África, causando homogenización genética incompleta en los humanos modernos cuando los neandertales divergieron, mientras que los ancestros primitivos de Eurasia estaban aún más relacionados con los neandertales que los africanos (Green et al. 2010). Sobre la base del espectro de frecuencias de alelos, se demostró que el modelo de mezcla reciente se ajustaba mejor a los resultados, mientras que el modelo de subestructura de poblaciones antiguas no tenía cabida, lo que demuestra que el mejor modelo fue un evento de mezcla reciente precedido por un evento de cuello de botella entre los humanos modernos, lo que confirma la mezcla reciente como la explicación más parsimoniosa y plausible

para el exceso observado de similitudes genéticas entre los humanos modernos no africanos y los neandertales (Yang et al. 2012). En base a los patrones de desequilibrio de ligamiento, los datos confirman un reciente evento de mezcla (Sankararaman et al. 2012). Teniendo en cuenta los patrones observados en el desequilibrio de ligamiento, se estimó que el último flujo de genes de neandertal hacia los ancestros primitivos de los europeos ocurrió hace entre 49000 y 67000 años. Junto con la evidencia arqueológica y fósil, se cree que el flujo de genes se produjo en algún lugar de Eurasia occidental, posiblemente en Oriente Medio. Utilizando un genoma de neandertal, euroasiático, africano y chimpancé (como *outgroup*) y dividiéndolo en bloques de secuencias cortas no recombinantes, para estimar la máxima verosimilitud del genoma en diferentes modelos, se descartó la idea de una subestructura poblacional antigua en África y se confirmó un evento de mezcla neandertal (Lohse & Frantz 2014).

Los restos de enterramientos del Paleolítico superior de un niño humano moderno de Abrigo do Lagar Velho (Portugal) presentan rasgos que indican el cruce de los neandertales con los humanos modernos que poblaron la península ibérica (Duarte et al. 1999). Teniendo en cuenta la fecha de los restos funerarios, hace unos 24500 años, y la persistencia de los rasgos de neandertal mucho después del período de transición de una población neandertal a una población humana moderna en la península ibérica, hace entre 28000 y 30000, el niño puede haber sido descendiente de una población ya muy mezclada.

Los restos de un humano moderno del Paleolítico superior de Peştera Muierilor (Rumania) de hace 35000 años muestran un patrón morfológico de los primeros humanos modernos europeos, pero posee características arcaicas o neandertales, sugiriendo que los humanos modernos europeos se entrecruzan con los neandertales (Soficaru, Dobos & Trinkaus 2006). Estas características incluyen una gran amplitud interorbital, arcos superciliares relativamente planos, un prominente moño occipital, una muesca mandibular asimétrica y poco profunda, un proceso coronoideo mandibular alto, el cóndilo perpendicular a la posición de la muesca y una cavidad glenoidea escapular estrecha. La mandíbula Oase 1, perteneciente a un individuo humano moderno temprano, de Peştera cu Oase (Rumania) de entre 34000 y 36000 presenta un mosaico de características modernas, arcaicas y posibles de neandertal (Trinkaus et al. 2003). Muestra un puente lingual del foramen mandibular, no presente en humanos anteriores, excepto los neandertales de finales del Pleistoceno medio y tardío, lo que sugiere la afinidad con los neandertales. A partir de la mandíbula de Oase 1 se concluye que, aparentemente, hubo un cambio craneofacial significativo de los primeros humanos modernos, al menos de Europa, posiblemente debido a cierto grado de mezcla con los neandertales. Los primeros humanos modernos europeos y los subsecuentes del Paleolítico superior medio, que se relacionan anatómicamente en gran medida con los primeros humanos africanos modernos del Paleolítico medio, también muestran rasgos distintivamente neandertales, lo que sugiere que solamente la ascendencia humana moderna del Paleolítico medio era poco probable para los humanos europeos modernos tempranos (Trinkaus 2007). Una mandíbula de neandertal de la cueva de Mezzena en Monti Lessin (Italia) muestra indicios de un posible cruce entre los últimos neandertales italianos (Condemi et al. 2013). La mandíbula se encuentra dentro del rango morfológico de los humanos modernos, pero también muestra fuertes similitudes con algunos de

los otros especímenes de neandertal, lo que indica un cambio en la morfología tardía de neandertal debido al posible entrecruzamiento con los humanos modernos. Manot 1, una calota parcial de un ser humano moderno descubierta en la Cueva de Manot (Galilea occidental, Israel) y fechada en alrededor de 54700 años, representa la primera evidencia fósil del período en que los humanos modernos emigraron con éxito de África y colonizaron Eurasia (Hershkovitz et al. 2015). También proporciona la primera evidencia fósil de que los humanos modernos habitaron el sur de Levante durante la interfaz paleolítica media-superior, contemporáneamente con los neandertales y cerca del probable evento de cruzamiento. Las características morfológicas sugieren que la población Manot puede estar estrechamente relacionada o dar lugar a los primeros humanos modernos que más tarde colonizaron con éxito Europa para establecer poblaciones del Paleolítico superior temprano.

El ADN de denisovanos se ha encontrado en humanos modernos. Se ha demostrado que los melanesios comparten relativamente más alelos con los denisovanos en comparación con otros eurasiáticos y africanos (Reich et al. 2010). Se estima que del 4% al 6% del genoma en melanesios se deriva de denisovanos, mientras que otros eurasiáticos o africanos no muestran contribuciones de los genes de denisovanos. Se ha observado que los denisovanos aportaron genes a los melanesios pero no a los asiáticos orientales, lo que indica que hubo interacción entre los antepasados primitivos de los melanesios con los denisovanos, pero que esta interacción no tuvo lugar en las regiones cercanas al sur de Siberia, donde se han encontrado los únicos restos de denisovanos hasta la fecha. Además, los aborígenes australianos también muestran un relativo aumento de alelos compartidos con los denisovanos, en comparación con otras poblaciones de Eurasia y África, de acuerdo con la hipótesis de una mayor mezcla entre los denisovanos y los melanesios (Rasmussen et al. 2011).

Reich et al. (2011) postulan que la mayor presencia de mezcla con denisovanos está en las poblaciones de Oceanía, seguidas por muchas poblaciones del sudeste asiático y ninguna en las poblaciones de Asia oriental. La presencia de ADN denisova es significativa en las poblaciones orientales del sudeste asiático y de Oceanía como aborígenes australianos, polinesios, fijianos, indonesios orientales, filipina mamanwa y manobo; pero no en ciertas poblaciones occidentales y continentales del sudeste asiático como indonesios occidentales, los Jehai de Malasia, los Onge de las islas Andaman y los asiáticos continentales, lo que indica que el evento de mezcla con los denisova sucedió en el sudeste de Asia en lugar de en Eurasia continental. La observación de la alta mezcla con denisovanos en Oceanía y su falta en Asia continental sugiere que los humanos modernos y los denisovanos se habían cruzado al este de la línea de Wallace que divide el sudeste asiático de acuerdo con Cooper & Stringer (2013).

Skoglund y Jakobsson (2011) observaron que, de forma particular, los pobladores de Oceanía, seguidos por los del sudeste de Asia, poseen una alta mezcla con denisova respecto a otras poblaciones. Además, hallaron bajas tasas de mezcla con denisovanos en el este de Asia, pero ninguna en las poblaciones nativas de América. Por el contrario, Prüfer et al. (2014) encontraron que los asiáticos continentales y los nativos americanos poseen un 0,2% de

contribución de denisovanos. Wall et al. (2013) exponen que no encuentran evidencia de mezcla con población denisova en los pobladores actuales del este de Asia.

Los hallazgos indican que el evento de flujo génico de denisova sucedió a los ancestros comunes de los aborígenes filipinos, australianos y guineanos (Reich et al. 2011). Los nativos de Nueva Guinea y los australianos tienen tasas similares de mezcla con denisova, lo que indica que el entrecruzamiento tuvo lugar antes de la entrada de sus ancestros comunes en Nueva Guinea y Australia, durante el Pleistoceno, hace al menos 44000 años. También se ha observado que la cantidad de ascendencia de Oceanía cercana en el sudeste asiático es proporcional a la mezcla con denisova, excepto en Filipinas. Reich et al. (2011) sugirieron también un posible modelo de ola de migración temprana hacia el este de humanos modernos, algunos que fueron antepasados de las poblaciones mencionadas anteriormente, y que se cruzaron con los denisovanos, seguidos respectivamente por la divergencia de los antepasados filipinos, el mestizaje entre los ancestros tempranos de los nativos de Nueva Guinea y los australianos con una parte de la misma población de inmigrantes tempranos que no experimentaron el flujo génico con denisovanos, y el mestizaje entre los antepasados tempranos filipinos con una parte de la población de una ola migratoria tardía hacia el este, de la cual descenderían los asiáticos orientales. Se ha demostrado que los eurasiáticos tienen material genético de origen arcaico, aunque en proporción significativamente menor, que se superpone con el de los denisovanos, debido al hecho de que los denisovanos están relacionados con los neandertales, que contribuyeron al acervo genético eurasiático, más que con el mestizaje de los denisovanos con los antepasados primitivos de esos euroasiáticos (Reich et al. 2010; Meyer et al. 2012). Los restos óseos de un humano moderno temprano de la cueva de Tianyuan (cerca de Zhoukoudian, China) de hace 4,000 años mostraron una contribución de Neanderthal dentro del rango de los humanos modernos eurasiáticos de hoy en día, pero no tenía ninguna contribución discernible de denisova (Fu et al. 2013). Es un pariente lejano de los antepasados de muchas poblaciones asiáticas e indígenas americanas, pero posterior a la divergencia entre asiáticos y europeos. La falta de un componente de denisova en el individuo de Tianyuan sugiere que la contribución genética ha sido siempre escasa en el continente (Prüfer et al. 2014).

Estudiando los alelos HLA del sistema inmune, se ha sugerido que HLA-B*73 introgresó de los denisovanos a humanos modernos en Asia occidental debido al patrón de distribución y la divergencia de HLA-B*73 respecto a otros alelos HLA (Abi-Rached et al. 2011). En los humanos modernos, HLA-B*73 se concentra en el oeste de Asia, pero es raro o está ausente en otros lugares. Aunque HLA-B*73 no está presente en el genoma secuenciado de denisova, el estudio del desequilibrio de ligamiento observó que estaba asociado al alelo HLA-C*15:05 derivado de denisova, en consonancia con la estimación de que el 98% de los humanos modernos que portan B*73 también poseen C*15:05. Los alelos de HLA-A*02, HLA-A*11, HLA-C*15 y HLA-C*12:02 de denisova corresponden a alelos comunes en humanos modernos, y se cree que deben haber sido aportados por los denisovanos a los humanos modernos, ya que es poco probable que se hayan conservado independientemente en ambos durante tanto tiempo debido a la alta tasa de mutación de los alelos HLA (Abi-Rached et al. 2011).

Se ha encontrado que una variante del gen EPAS1 (un factor de transcripción que responde a la falta de oxígeno a nivel celular) se introdujo de denisovanos a los humanos modernos (Huerta-Sánchez et al. 2014). La variante ancestral regula al alza los niveles de hemoglobina para compensar los bajos niveles de oxígeno, como en altitudes elevadas, pero esto también tiene la contraprestación del aumento de la viscosidad sanguínea. La variante derivada de denisova limita este aumento de los niveles de hemoglobina, lo que resulta en una mejor adaptación de la altitud. La variante del gen EPAS1 derivado de denisova es común en los tibetanos y fue seleccionada positivamente en sus ancestros después de que colonizaron la meseta tibetana (Huerta-Sánchez et al. 2014).

La rápida descomposición de los fósiles en los entornos del África subsahariana hace que actualmente no sea factible comparar la mezcla humana moderna con muestras de referencia de homínidos arcaicos del África subsahariana (Lachance et al. 2012). De las tres regiones candidatas con introgresión encontradas buscando patrones inusuales de variación (como por ejemplo, una profunda divergencia de haplotipos, patrones inusuales de desequilibrio de ligamiento y tamaño de clado basal pequeño) en 61 regiones no codificantes de dos grupos de cazadores-recolectores (los Aka y los San, quienes tienen una mezcla significativa) y un grupo agrícola de África Occidental (Mandinka, que no tiene una mezcla significativa), se concluye que aproximadamente el 2% del material genético encontrado en estas poblaciones del África Subsahariana se insertó en el genoma humano hace aproximadamente 35,000 años desde los homínidos arcaicos que se separaron del linaje humano moderno hace unos 700,000 años (Hammer et al. 2011). Además se sugiere que este evento de mezcla ocurrió con los homínidos arcaicos que alguna vez habitaron África Central. Investigando secuencias de genoma completo de alta cobertura de quince individuos varones subsaharianos cazadores-recolectores de tres grupos; cinco pigmeos (tres Baka, un Bedzan y un Bakola) de Camerún, cinco Hadza de Tanzania y cinco Sandawe de Tanzania; se han hallado signos de que los antepasados de los cazadores-recolectores se cruzaron con una o más poblaciones humanas arcaicas (Lachance et al. 2012) probablemente hace más de 40,000 años (Callaway 2012). El análisis de los supuestos haplotipos introgresivos en las quince muestras de cazadores-recolectores sugiere que la población africana arcaica y los humanos modernos divergieron hace alrededor de 1.2 a 1.3 millones de años. Xu et al. (2017) analizaron la evolución de la proteína Mucin-7 en la saliva de ciertas poblaciones africanas y encontraron evidencia de que una especie de humanos arcaicos puede haber contribuido con ADN en su reserva genética. Esta especie no fue identificada y fue referida como una población fantasma de humanos. Skoglund et al. (2017) examinaron los genomas de varias poblaciones antiguas y recientes en África y también identificaron pruebas que apuntaban a un grupo extinto de humanos arcaicos, un "linaje basal de población de África occidental", con contribuciones al conjunto genético de las poblaciones de África occidental, estimadas en 13% para Mende y en 9% para Yoruba. De manera similar, un estudio de 2018 ha calculado la mezcla arcaica de las poblaciones en Yoruba al 8% (Durvasula & Sankararaman 2018).

Situación de la población gitana en el País Vasco

El hecho de que la población gitana practique cierto grado de selección sexual mediante emparejamiento intraétnico, la convierte en un grupo poblacional muy interesante para el estudio de la deriva y la selección genética. Asimismo, el origen de esta población en Punjab (norte de la India) y su historia como población nómada hace que sea igualmente interesante el estudio de sus posibles entrecruzamientos con poblaciones euroasiáticas. En la UPV se han realizado estudios con poblaciones gitanas para diversos campos como la determinación de fenotipos de obesidad y estudio de la influencia de los factores genéticos y ambientales sobre el desarrollo de la adiposidad en población caso-control y en familias de etnia gitana de la Comunidad Autónoma Vasca por parte del grupo de investigación de la Dra. Esther Rebato Ochoa, o como el trabajo de Fin de Grado de Iñigo Marcos Sarobe (2016) donde se estudian inserciones Alu de la población gitana del País Vasco. Por ello resulta interesante estudiar la situación de esta población respecto a los genes HLA.

HIPÓTESIS Y OBJETIVOS

HIPÓTESIS Y OBJETIVOS

En la presente tesis doctoral se ha trabajado con la hipótesis de que existe un conjunto de procesos y presiones selectivas que afectan a los genes HLA clase I y II del Complejo Mayor de Histocompatibilidad de manera diferencial, tanto entre regiones de los propios genes, como entre los genes al completo y entre distintas poblaciones humanas. Además se ha considerado la idea de que, a lo largo del proceso evolutivo (donde se cree tuvo gran importancia la relación entre *Homo sapiens* y humanos arcaicos como Neandertales y Denisovanos) y colonizador humano (desde la salida de África hasta la actualidad), estas presiones selectivas han dejado su marca en el genoma, y que estas marcas pueden usarse para estudiar la historia evolutiva de nuestra especie.

Podemos resumir la hipótesis de trabajo de esta tesis doctoral en la siguiente frase: **Existen presiones selectivas que afectan a los genes HLA clase I y II del Complejo Mayor de Histocompatibilidad**.

Para evaluar esta hipótesis se ha trabajado con los siguientes objetivos:

1. Conocer el grado de presencia de desequilibrios Hardy-Weinberg en el conjunto de los genes HLA clase I y II de del Complejo Mayor de Histocompatibilidad.

2. Establecer el rango de acción de la varianza de Wahlund en los genes estudiados, a fin de valorar la posible acción de las diversas presiones selectivas.

3. Valorar el alcance y presencia de marcadores informativos de ancestría (AIMs) en los genes estudiados, con el fin de saber si estos genes pueden usarse para estudiar movimientos migratorios humanos.

4. Caracterizar la población gitana de la comunidad autónoma del País Vasco teniendo en cuenta los genes objeto de estudio.

5. Evaluar la presencia de introgresiones de ADN Neandertal en humanos actuales, así como valorar las relaciones geográficas que pudieran surgir.

MATERIALES Y MÉTODOS

MATERIALES Y METODOS

Bases de datos

1000Genomes

Una parte de los datos genéticos usados en esta tesis corresponden a datos genotípicos de SNPs de la región del Complejo Mayor de Histocompatibilidad obtenidos de la base de datos del proyecto 1000 Genomes Fase 3 (https://www.ncbi.nlm.nih.gov/variation/tools/1000genomes/). Se han usado las secuencias de 2504 individuos de las 26 poblaciones de todo el mundo que forman parte del proyecto (Tabla M.1), obtenidos de la página web del proyecto, en los formatos VCF (datos individuales) y popVCF (datos poblacionales).

La herramienta "1000 Genomes Browser" (Figura M.1) permite seleccionar regiones concretas (cromosomas, genes, exones, regiones intergénicas,...) dentro del conjunto del genoma humano. Los datos pueden ser descargados desde la propia herramienta en los formatos VCF y popVCF anteriormente mencionados.

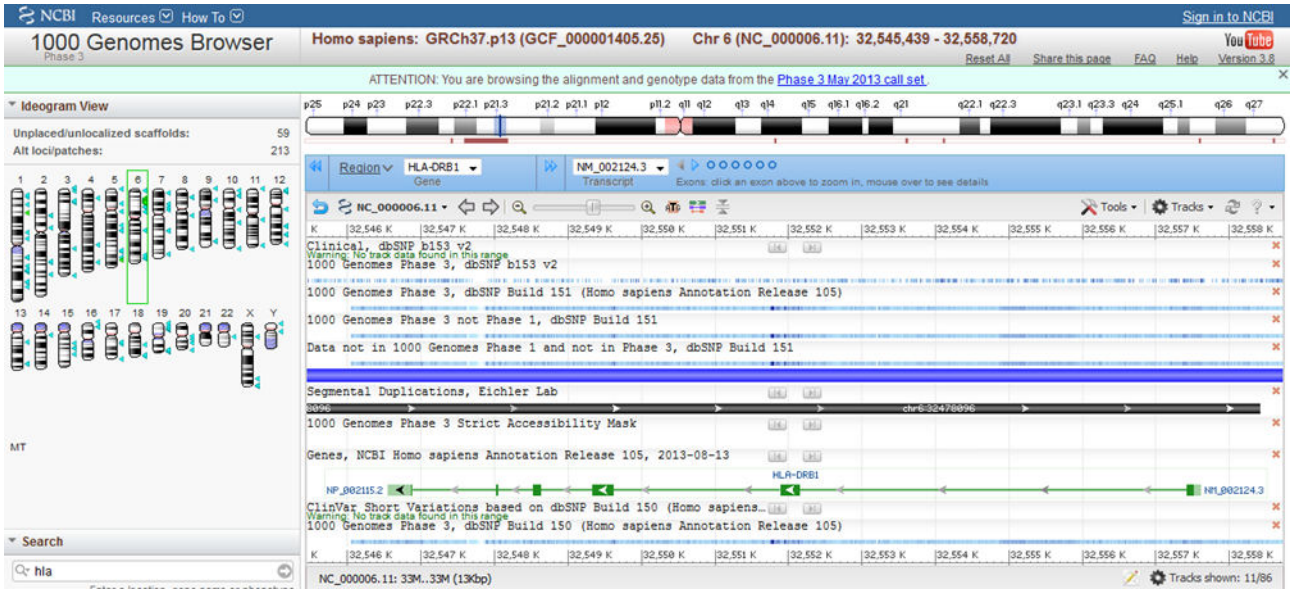


Figura M.1: Ventana de navegación de la herramienta 1000 Genomes Browser del NCBI. En este ejemplo, podemos visualizar el gen HLA-DRB1.

Código población	Descripción	Procedencia	Número de individuos
ACB	Caribeños de origen africano	Barbados	96
ASW	Afroamericanos	Suroeste de EEUU	61
BEB	Bengalíes	Bangladesh	86
CDX	Chinos Dai	Xishuangbanna, China	93
CEU	Ancestría del norte y oeste de Europa	Utah, EEUU	99
CHB	Chinos Han	Pekín, China	103
CHS	Chinos Han	Sur de China	105
CLM	Colombianos	Medellín, Colombia	94
ESN	Esan	Nigeria	99
FIN	Fineses	Finlandia	99
GBR	Británicos	Inglaterra y Escocia	91
GIH	Indios guyaratíes	Houston, Texas	103
GWD	Gambianos	División Occidental, Gambia	113
IBS	Población ibérica	España	107
ITU	Indios telugu	Reino Unido	102
JPT	Japoneses	Tokio, Japón	104
KHV	Kinh	Ciudad Ho Chi Minh, Vietnam	99
LWK	Luhya	Webuye, Kenia	99
MSL	Mende	Sierra Leona	85
MXL	Ancestría mexicana	Los Ángeles, EEUU	64
PEL	Peruanos	Lima, Perú	85
PJL	Punyabíes	Lahore, Pakistán	96
PUR	Puertorriqueños	Puerto Rico	104
STU	Tamiles de Sri Lanka	Reino Unido	102
TSI	Toscanos	Región de Toscana, Italia	107
YRI	Yoruba	Ibadán, Nigeria	108

Tabla M.1: Poblaciones del proyecto 1000 Genomes incluidas en los estudios realizados en esta tesis. Se incluyen datos sobre el origen geográfico, así como el número de individuos de cada población.

En la Figura M.2 podemos ver un mapa con las localizaciones de cada una de las poblaciones estudiadas en este trabajo.

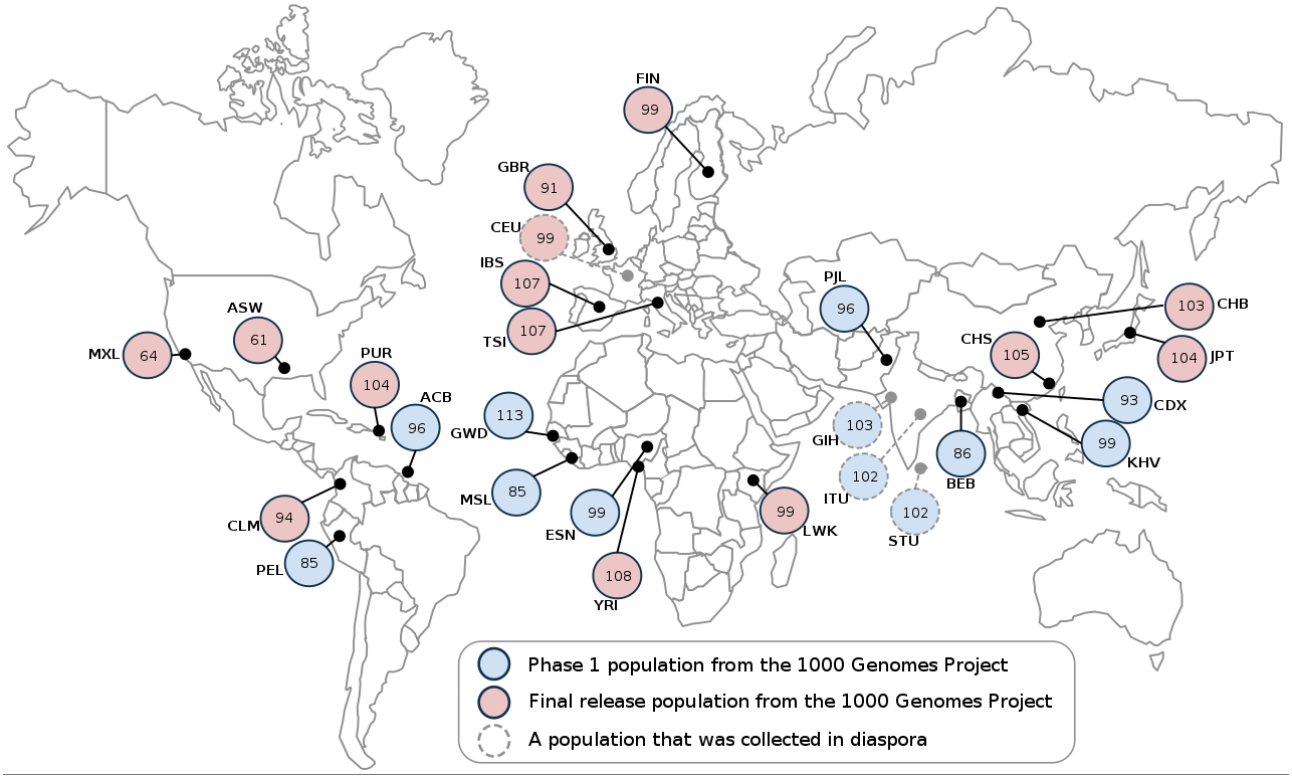


Figura M.2: Mapamundi con las localizaciones de las 26 poblaciones estudiadas. La cifra dentro de cada círculo indica el número de individuos estudiados de dicha población. Se incluye una leyenda en la que se explica en qué fase del proyecto 1000 Genomes se incluyó dicha población, así como si la muestra fue recogida en una localización diferente al lugar de origen ancestral de la población.

Hemos considerado que las poblaciones LWK, YRI, ESN, MSL, GWD, ACB y ASW entran dentro del bloque de poblaciones africanas. En el caso de las dos últimas, si bien se han recogido las muestras en la isla de Barbados y el suroeste de EEUU respectivamente, se ha decidido incluirlas dentro del grupo de poblaciones africanas porque en origen fueron poblaciones esclavizadas de origen africano que han mantenido los rasgos de la población original.

Los Luhya (LWK) de Webuye son un grupo étnico bantú de Kenia. Son aproximadamente 5,3 millones de personas según el censo de 2009, divididos en unas 19 tribus distintas cada una con su propio dialecto. Representan alrededor del 16% de la población total de Kenia de 38,5 millones, siendo el segundo grupo étnico más grande de Kenia.

Los Yoruba de Ibadán (YRI) son alrededor de 105 millones de personas en total. La mayoría de esta población es de Nigeria, donde los Yoruba representan el 21% de la población del país, convirtiéndolos en uno de los grupos étnicos más grandes de África. Según Schlebusch et al (2017) los Yoruba tienen ~ 31% de mezcla prehistórica de lo que los autores consideran “humanos basales”.

Los Esan (ESN) son un grupo étnico de alrededor de 1,5 millones de personas del sur de Nigeria. Los Esan son tradicionalmente agricultores, practicantes de medicina tradicional, guerreros mercenarios y cazadores. Hay una fuerte diáspora Esan, sobre todo hacia Reino Unido, en parte debido al pasado colonial.

Los Mende (MSL) son un grupo étnico de alrededor de 2 millones de personas, y son uno de los grupos étnicos más grandes de Sierra Leona. Su lenguaje, también llamado mende, es hablado por alrededor del 46% de la población de Sierra Leona, y es usado como lengua franca por grupos étnicos menores que habitan en la misma región.

Los Gambianos de la zona conocida como División Oeste (GWD) de Gambia son un grupo de alrededor de 700.000 personas que habitan el extremos suroeste del país, en la zona de la desembocadura del río Gambia.

Como hemos dicho, tanto los caribeños de origen africano de Barbados (ACB) como la población afroamericana del suroeste de EEUU (ASW), son población de origen africano pero que habitan zonas del continente americano. Barbados cuenta con una población de alrededor de 300.000 personas, de las que el 93% son población de origen africano o mulatos, esto hace pensar que ha habido cierto grado de mestizaje. En el suroeste de EEUU hay una zona conocida como “Black Belt” o Cinturón Negro en la que la población afroamericana es predominante. El censo de los Estados Unidos informó que en 2000 los Estados Unidos tenían 96 condados con un porcentaje de población negra de más del 50%, de los que 95 estaban ubicados a lo largo de la costa y las tierras bajas del sur en una zona relacionada con las áreas tradicionales de la agricultura de plantación, incluido el Delta del Mississippi.

En el grupo de poblaciones europeas hemos incluido a IBS, TSI, GBR, FIN y CEU. Si bien la muestra de esta última se ha muestreado en Utah (EEUU), mantiene el pool genético de las poblaciones del centro y oeste de Europa de las que proviene en origen.

La población ibérica española (IBS) está compuesta por casi 47 millones de personas residentes en la Península Ibérica.

Los toscanos (TSI) son un grupo de unos 3,5 millones de personas que habitan la región de la Toscana en Italia.

Los británicos (GBR) son un grupo de unos 65 millones de personas que habitan Reino Unido.

Los fineses (FIN) son un grupo de unos 5 millones de personas que habitan Finlandia. Los fineses muestran muy poco o ningún gen mediterráneo y africano, pero por otro lado, casi el 10% de los genes finlandeses parecen compartirse con las poblaciones siberianas. Sin embargo, más del 80% de los genes finlandeses provienen de una sola población antigua del noreste de Europa, mientras que la mayoría de los europeos son una mezcla de 3 o más componentes principales.

La población con ancestría del norte y oeste de Europa (CEU) de Utah (EEUU) es un grupo de unos 2,8 millones de personas.

En cuanto a las englobadas en el grupo de poblaciones asiáticas, cabe destacar primero que son el grupo más numeroso. Así, pertenecen a este grupo PJL, BEB, CHS, KHV, CDX, JPT, CHB, GIH, ITU y STU. Además es el grupo continental que más poblaciones tiene en diáspora: GIH ha sido muestreada en Houston (Texas, EEUU), mientras que ITU y STU han sido muestreadas en Reino Unido.

Los punyabíes o punjabis (PJL) de Pakistán son el grupo étnico más numeroso del país, que con casi 92 millones de personas, representan casi el 45% de la población total del país. El pueblo punjabi ha emigrado en gran número a muchas partes del mundo. A principios del siglo XX, muchos Punjabis comenzaron a establecerse en los Estados Unidos. El Reino Unido tiene un número significativo de Punjabis de Pakistán e India.

Los bengalíes de Bangladesh (BEB) son un grupo de casi 163 millones de personas. Son un grupo étnico indo-ario nativo de la región de Bengala en el sur de Asia, específicamente en la parte oriental del subcontinente indio, actualmente dividido entre Bangladesh y los estados indios de Bengala Occidental, Tripura y el valle de Assam Barak, que hablan bengalí, un idioma de la familia de lenguas indo-arias.

Los chinos han (CHS del sur de China, CHB de Pekín) son el grupo étnico más numeroso del mundo con alrededor de 1300 millones de personas, un 18% del total de la población mundial. La mayoría (1200 millones aproximadamente) viven en la China continental, y el resto se dividen entre las regiones administrativas especiales de Hong Kong y Macao, así como Taiwán, regiones del sudeste de Asia y diversos países del todo el mundo. Por ejemplo, en 2016, residían en España algo más de 200.000 ciudadanos chinos, la mayoría procedentes de la región de Zhejiang, al este de China.

El pueblo vietnamita o el pueblo Kinh (KHV), son un grupo étnico de unos 86 millones de personas del sudeste asiático, originario de Vietnam. Hablan vietnamita, el idioma austroasiático más común. En el censo de 1999 constituían el 86% de la población del país, y se conocen oficialmente como Kinh para distinguirlos de otros grupos étnicos en Vietnam. Los análisis genómicos de KHV junto con otras poblaciones asiáticas sirvieron para constatar que KHV y otras poblaciones del sudeste de Asia derivaron principalmente de la misma población original del sudeste asiático. Los resultados de diferentes análisis genómicos son generalmente consistentes y respaldan la hipótesis de la migración de la población de África a Asia siguiendo la ruta Sur.

Los chinos Dai (CDX) son una minoría étnica de alrededor de 1,2 millones de personas de la prefectura autónoma Xishuangbanna Dai y de la prefectura autónoma Dehong Dai-Kachin en la provincia de Yunnan (China). También se encuentran grupos Dai en Laos, Birmania, Tailandia, Vietnam e India. Aunque oficialmente forman una única etnia, el pueblo Dai está compuesto por diferentes grupos culturales y lingüísticos.

Los japoneses (JPT) son un grupo étnico de unos 129 millones de personas, nativo del archipiélago japonés donde constituyen el 98,5% de la población total (unos 125 millones de personas). El término japonés étnico se usa a menudo para referirse a los integrantes de la etnia Yamato o Wajin. Los japoneses son uno de los grupos étnicos más grandes del mundo.

Los guyaratíes o gujaratíes (GIH) son un grupo étnico de unos 60 millones de personas cuyo origen se sitúa en el estado indio de Gujarat. Los gujaratíes tienen una larga tradición de navegación marítima y una historia de emigración a países como Yemen, Omán, Bahréin, Kuwait y otros países del Golfo Pérsico. Los países con las mayores poblaciones de gujaratíes son Pakistán, Reino Unido, Estados Unidos, Canadá y varios países de África meridional y oriental. A nivel mundial, se estima que los gujaratíes comprenden alrededor del 33% de la diáspora india en todo el mundo y se pueden encontrar en 129 de 190 países listados como naciones soberanas por las Naciones Unidas.

Los telugu (ITU) son un grupo étnico dravídico de unos 85 millones de personas que habla telugu como lengua materna y remontan su ascendencia a los estados indios de Andhra Pradesh y Telangana. Unos 4 millones de personas del total de población telugu se encuentra repartido por diversos países de todo el mundo, principalmente Estados Unidos, Arabia Saudí y la Unión Europea.

Los tamiles (STU) son un grupo étnico de unos 3 millones de personas, de las que aproximadamente 2,3 millones son nativos del estado insular de Sri Lanka en el sur de Asia, y el resto se encuentran repartidos en diversos países, principalmente en Canadá y reino Unido.

En lo referente a las poblaciones americanas (CLM, MXL, PEL y PUR) es importante destacar el alto grado de mestizaje que presentan dado que, en general, son el resultado del entrecruzamiento de la población indígena original y población inmigrante del suroeste europeo a partir del siglo XVI.

Los colombianos (CLM) son un grupo de unos 52 millones de personas que forman una sociedad multiétnica. La mayoría de la población colombiana está compuesta por inmigrantes del Viejo Mundo y sus descendientes. Después del período inicial de conquista e inmigración españolas, se produjeron diferentes oleadas de inmigración y asentamiento de pueblos no indígenas en el transcurso de casi seis siglos, proceso que aún continúa hoy. No sorprende pues que la población mestiza de Colombia represente el 49% de la población total.

La población de ancestría mejicana de Los Ángeles (MXL) es un grupo de unos 1,3 millones de personas que representa casi el 32% de la población total de la ciudad de Los Ángeles de unos 4 millones de personas.

La población de Lima en Perú (PEL) es de unos 9 millones de personas, y está fuertemente influenciada por procesos de mestizaje con poblaciones inmigrantes como en el caso de CLM. Así los mestizos componen el 47% de la población.

Los puertorriqueños (PUR) son un grupo de unos 3 millones de personas, aunque si incluimos a los individuos con ancestría puertorriqueña la cifra aumenta hasta los 6 millones de personas aproximadamente. Al igual que ocurre con el resto de poblaciones americanas, es una población fuertemente mestizada.

HapMap

Además, se han usado datos de 550 individuos agrupados en 4 poblaciones del proyecto HapMap (Fase 3), a fin de estudiar patrones de selección (Tabla M.2). En dichas poblaciones se han seleccionado los individuos que forman tríos (madre, padre y descendiente). Se han descartado los individuos que no forman tríos válidos o cuyo genotipo no esté incluido en la última actualización emitida por el proyecto HapMap.

Código	**Descripción**	**Individuos analizados**	**Número de tríos**
ASW	Afroamericanos (SO EEUU)	87	10
CEU	Ancestría del norte y oeste de Europa (Utah, EEUU)	174	52
MXL	Ancestría mexicana (Los Ángeles, EEUU)	86	24
YRI	Yoruba (Ibadán, Nigeria)	203	57

Tabla M.2: Poblaciones del proyecto HapMap incluidas en los estudios de selección realizados en esta tesis. Se incluyen el número de individuos analizados, así como el número de tríos estudiados.

Genoma Neandertal

Para estudiar la relación de los genes del CMH de Neandertal con los de *Homo sapiens* se han usado datos de la secuencia de los genes HLA de Neandertal obtenidos del proyecto Genoma Neandertal (https://www.eva.mpg.de/genetics/genome-projects/neandertal/index.html, Figura M.3) a través de la base de datos de Ensembl. Para las poblaciones humanas actuales se han usado los datos de las secuencias de los genes HLA obtenidos del servidor FTP del European Bioinformatics Institute (EMBL-EBI) (https://www.ebi.ac.uk/, Figura M.4).

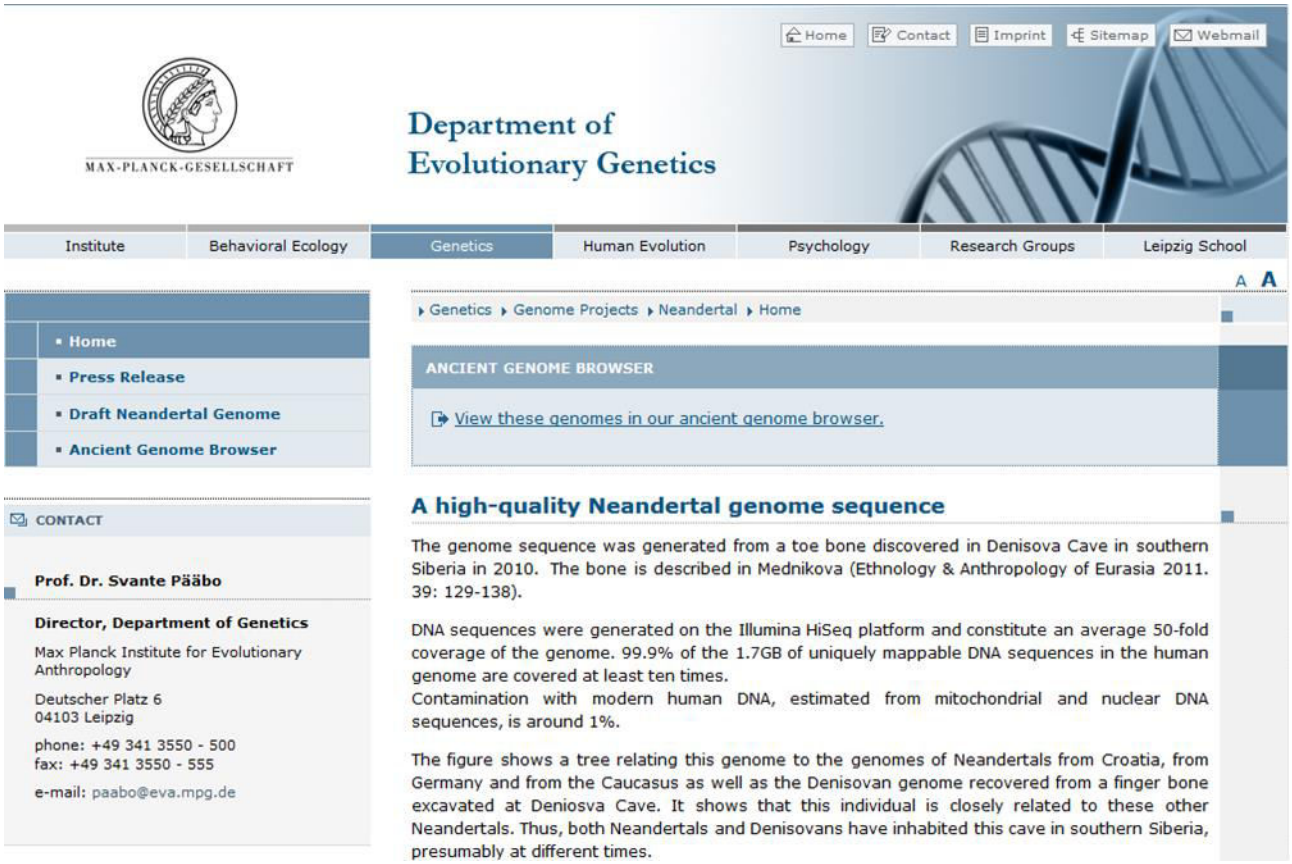


Figura M.3: Portada de la web del proyecto Genoma Neandertal.

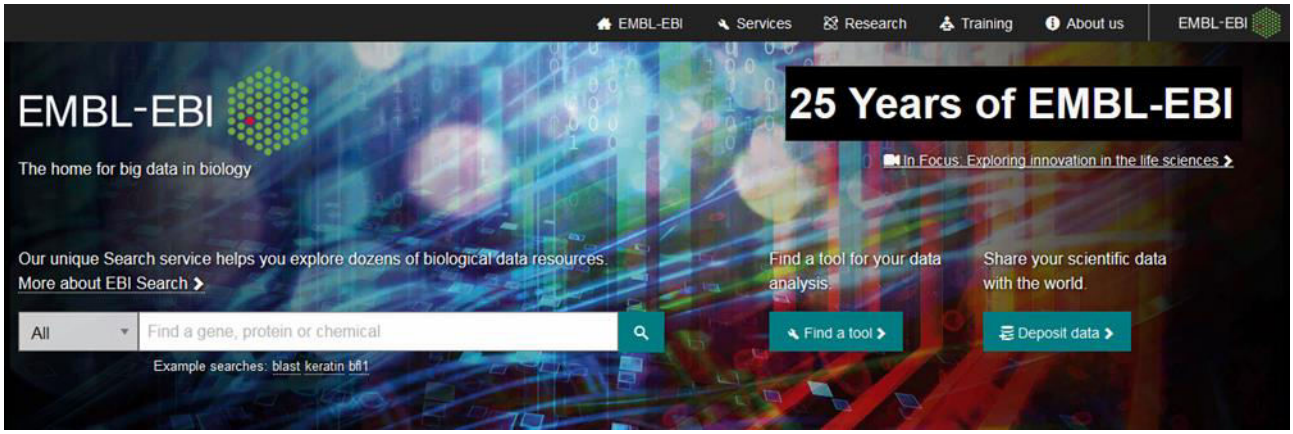


Figura M.4: Portada de la web del European Bioinformatics Institute *(EMBL-EBI).*

Programas para minería de datos

Si bien existen diversos programas capaces de hacer diferentes análisis estadísticos con datos genéticos (como Arlequin, Fstat o módulos de Python como BioPython), la casi totalidad de los mismos carecen de la capacidad de tratar datos brutos de manera masiva, requiriendo la fragmentación del conjunto de datos para poder analizarlos o su conversión a formatos propios y exclusivos, lo que dificulta aún más el poder analizarlos con programas que no sean para el cual están diseñados esos formatos. Ante la necesidad que se presentó de analizar los datos del CMH

de 2504 individuos distintos, se hizo obligado el desarrollo de una herramienta capaz de tratar con la gran cantidad de datos necesarios, evitando la pérdida por compartimentación de los datos para poder tratarlos, y que al mismo tiempo fuera capaz de tratar con los datos en bruto de los archivos.

FstMap

Nuestro equipo ha desarrollado la herramienta FstMap con el objetivo de poder analizar gran cantidad de datos en bruto a partir de archivos VCF y popVCF. Ha sido desarrollado en lenguaje Java. Por ello, no precisa de compilación, sino que las instrucciones son interpretadas por el propio ordenador donde se ejecuta el programa, generando así un entorno de trabajo más claro y atractivo, lo cual teniendo en cuenta la cantidad de datos generado en los *outputs*, es indispensable. FstMap presenta una interfaz de usuario simple y fácil de usar (Figura M.5).

Figura M.5: Interfaz de FstMap al iniciar el programa.

El punto de partida para comenzar con los análisis de FstMap son los archivos VCF y popVCF. Es importante que ambos archivos contengan datos para la misma región objetivo, puesto

que son archivos interrelacionados, y si fueran de distintas regiones, los resultados carecerían muy posiblemente de sentido alguno, y con toda seguridad, de validez. Sin embargo, la herramienta "*1000 Genomes Browser*" del proyecto *1000 Genomes* permite descargar ambos archivos para una misma región de manera sencilla y rápida.

Debemos clicar en la pestaña "*Input*", y después en la primera opción "*Read vcf populations data*". Se abrirá una ventana donde veremos el explorador de archivos. Deberemos localizar el archivo popVCF descargado de la región que queramos analizar y seleccionarlo. Una vez hecho esto, la interfaz general de FstMap mostrará una información básica sobre la región de interés: número de poblaciones y SNPs que incluye, cromosoma donde se sitúa, y posiciones del primer y del último locus que la definen (Figura M.6). Igualmente nos informará que la lectura de los datos de poblaciones ha sido completada, y que procedamos a incluir el archivo de datos VCF de los individuos.

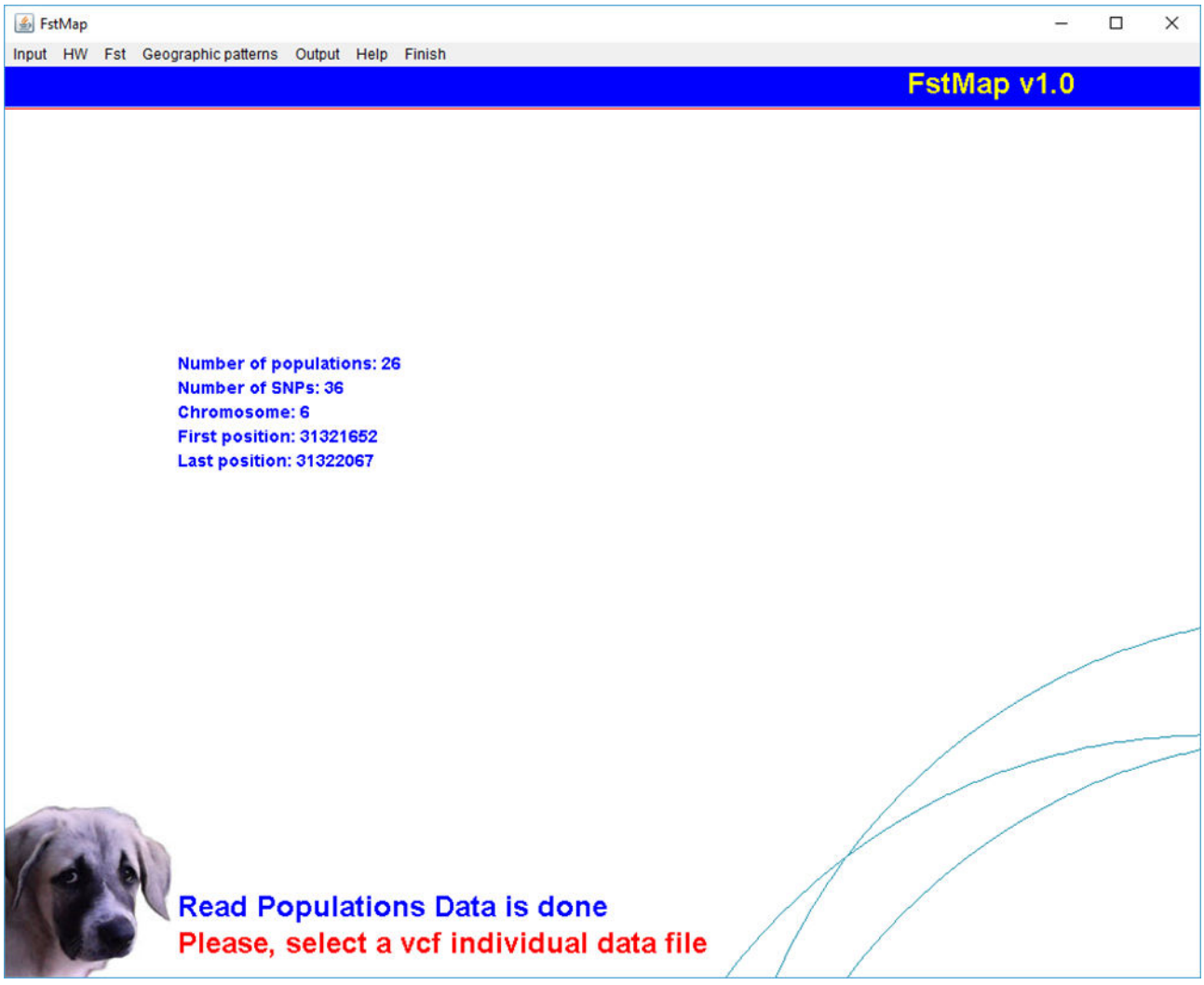


Figura M.6: Interfaz de FstMap tras leer el archivo popVCF.

A continuación, volvemos a seleccionar la pestaña "*Input*", pero esta vez elegimos "*Read vcf individual data*". Se abrirá una ventana pequeña con el título "*Work in progress*" donde se mostrará el número total de SNPs que se van a analizar. De igual manera que con el archivo

popVCF anterior, se abre una ventana de explorador de archivo donde deberemos buscar y seleccionar el archivo VCF descargado de la región que queramos analizar. Una vez hecho esto, veremos que en la ventana "*Work in progress*" va mostrando el avance en el análisis (Figura M.7), hasta completar el total de SNPs indicados con anterioridad. Una vez completado, la interfaz general de FstMap mostrará, además de la información básica que mostró al incluir el archivo popVCF, el número de individuos que ha detectado en el archivo VCF de datos de individuos. Igualmente informará que la lectura de dicho archivo ha sido completada (Figura M.8).

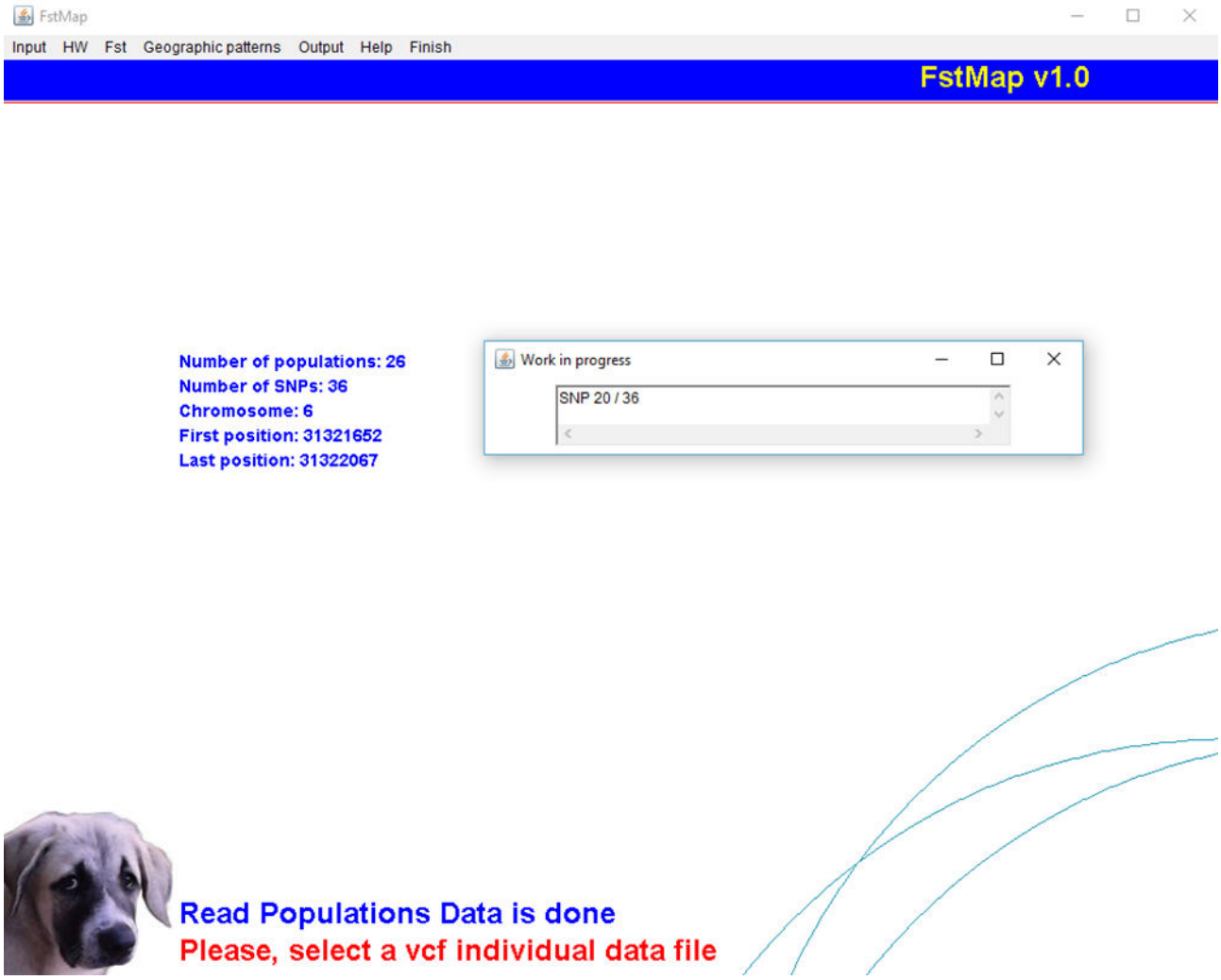


Figura M.7: Interfaz de FstMap durante el análisis de los SNPs de los archivos introducidos.

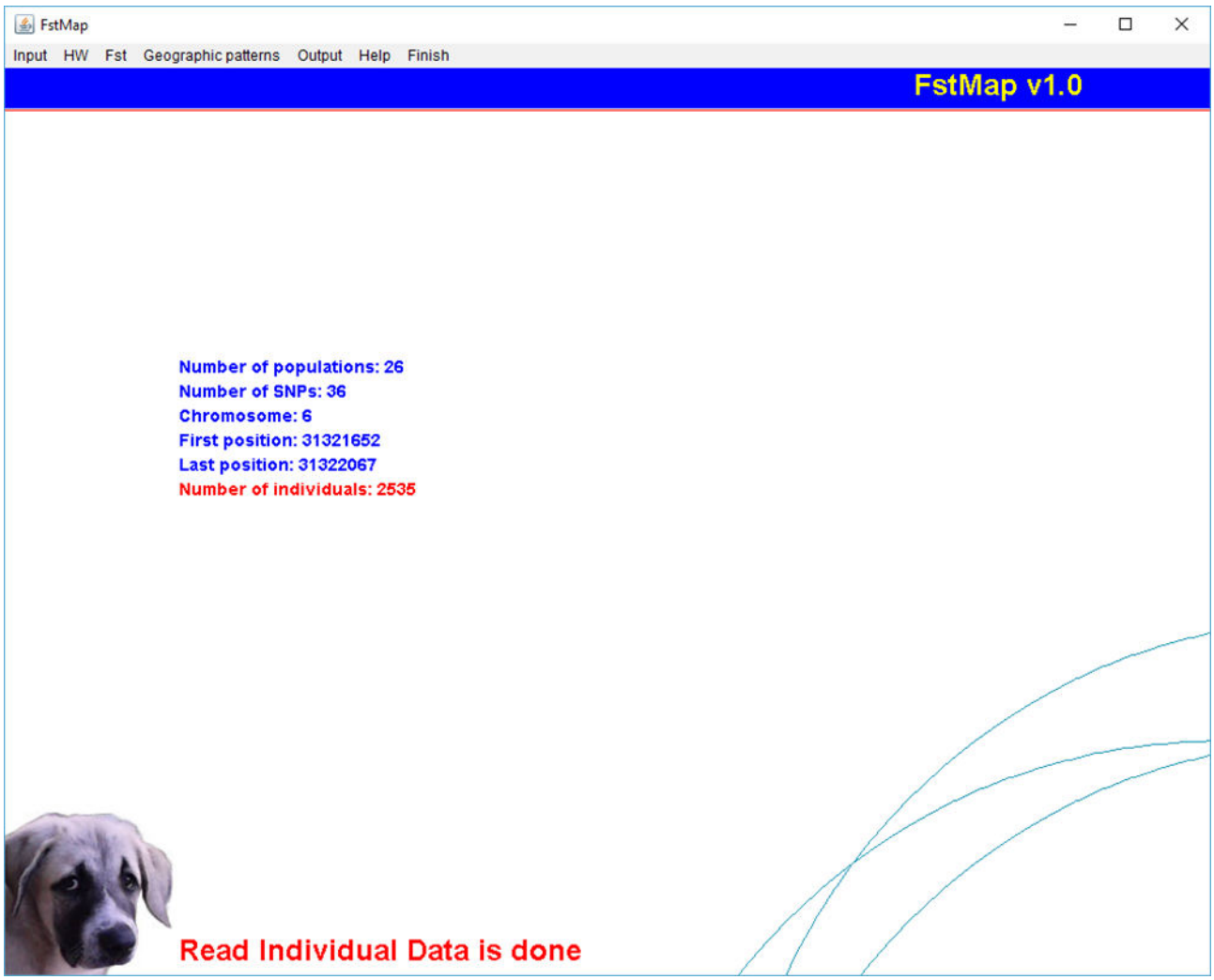


Figura M.8: Interfaz de FstMap una vez completado el análisis de la región objetivo.

Es importante destacar 3 factores que influyen en la velocidad del análisis de FstMap:

- El procesador del ordenador: Si bien el programa no requiere de unas especificaciones técnicas fuera de lo común hoy en día, y de hecho puede usarse en la mayoría de los ordenadores con procesadores de generaciones anteriores, en los ordenadores más antiguos puede suponer la ralentización del ordenador. En los portátiles más antiguos puede provocar una ligera subida de temperatura, pero nada fuera de lo común cuando se usan programas que cargan el procesador de este tipo de aparatos.

- El propio programa FstMap: Al ser un programa interpretado y no compilado, el tiempo que necesita para ser interpretado es superior. Al tener que ser traducido a lenguaje máquina con cada ejecución, este proceso es más lento que en los lenguajes compilados. Además, otra desventaja de un lenguaje interpretado es que, para ser ejecutado, se debe tener instalado el interpretador. Sin embargo, algunos lenguajes de programación poseen una máquina virtual que hace una traducción a lenguaje intermedio con lo cual el traducirlo a lenguaje de bajo nivel toma menos tiempo.

- El número de SNPs: El número de SNPs que se encuentren en la región objeto de estudio será el mayor factor limitante en cuanto a la velocidad de ejecución del análisis. Cuanto más grande, y por regla general, cuantos más SNPs tenga la región objetivo, más tardará en completarse el análisis. Una sección que contenga un par de cientos de SNPs tardará entre menos de 1 minuto y 3 minutos (dependiendo de otros factores como los expuestos anteriormente), mientras que un análisis de una región cromosómica más amplia, como por ejemplo la región completa del CMH, puede llegar a tardar varias horas.

Una vez completada la lectura de ambos archivos (popVCF y VCF), FstMap permitirá obtener una gran variedad de resultados.

Si seleccionamos la pestaña "*HW*" podremos acceder a datos y cálculos relativos al equilibrio de Hardy-Weinberg.

Una vez elegida la primera opción, "*HW equilibrium*", mostrará un plano de coordenadas cartesianas con el eje horizontal x representando la posición de cada SNP que el programa ha leído, y con el eje vertical y representando un valor de Fst, o índice de fijación, para cada posición (Figura M.9). Señalará los valores máximo y mínimo encontrados para Fst, así como la primera y última posición del fragmento analizado. Veremos una nube de puntos de color azul, que representan cada uno de los SNPs analizados con su posición y valor de Fst asociados. En el caso de que un SNP presente desequilibrio HW en al menos una población, el punto será de color rojo. Esta opción calcula el número total de poblaciones que están en desequilibrio Hardy-Weinberg ($p<0,05$) para cada SNP, así como el valor de p del desequilibrio para cada SNP en cada población. Los resultados se muestran en el fichero Results.xls. FstMap diferencia si el desequilibrio produce exceso de homocigotos o de heterocigotos. A fin de diferenciar estas dos situaciones, los valores de p para cada población y SNP será menor de 0,05 si se observan menos heterocigotos de los esperados (defecto de heterocigotos o pérdida de heterocigosidad), y será mayor o igual a 99 si hay más heterocigotos de los esperados (exceso de heterocigotos o ganancia de heterocigosidad). El "99" no tiene significado real, es una solución que el programa añade a posteriori, a fin de diferenciar ambos casos donde una población presente desequilibrio para un SNP. Así, un valor de p de 0,035 significa que hay un 3,5% de probabilidad de que las diferencias de genotipo se deban al azar y un 96,5% de probabilidad de que no se deban al azar. Este valor p es significativo ($p<0,05$), la hipótesis nula es rechazada, y decimos que la población no está en equilibrio Hardy-Weinberg con exceso de homocigotos o pérdida de heterocigosidad. Un valor de p de 99,02 significa que hay un 2% de probabilidad de que las diferencias en las frecuencias genotípicas se deban al azar y un 98% de probabilidad de que no se deban al azar. Este valor p es significativo, la hipótesis nula es rechazada, y decimos que la población no está en equilibrio Hardy-Weinberg con exceso de heterocigotos y ganancia de heterocigosidad.

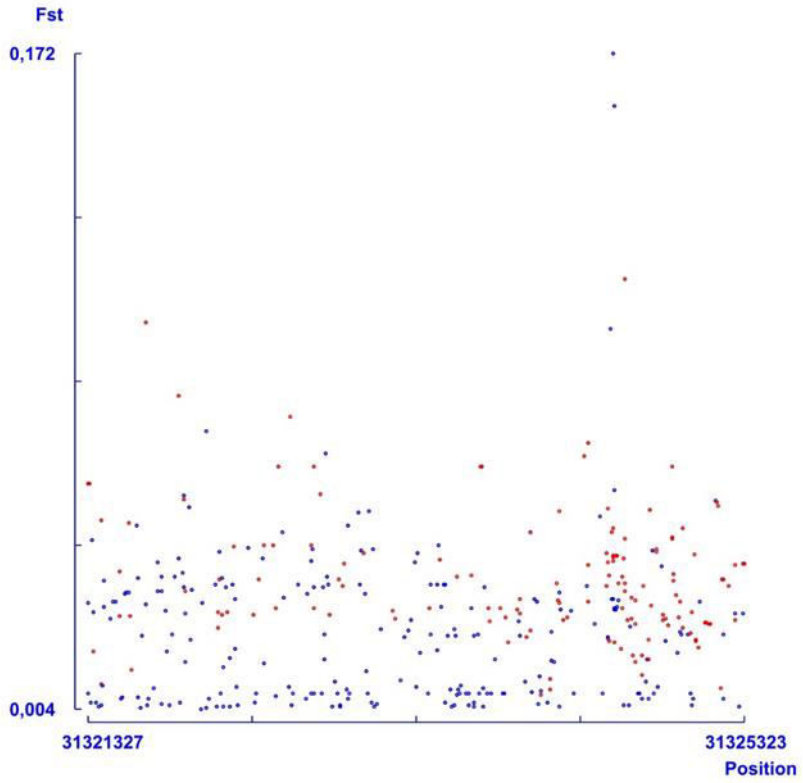


Figura M.9: Resultado de FstMap al seleccionar la opción "HW equilibrium". Se observan numerosos SNPs en desequilibrio HW.

La opción "*Searching for selection*" sirve para efectuar un análisis pormenorizado de potenciales efectos de la selección bajo diferentes modelos en los SNPs de la región estudiada. Realiza el análisis para cada marcador (Figura M.10) en cada una de las 4 poblaciones de la base de datos de HapMap (Tabla M.2). Una vez terminado el análisis, informará de ello y se generará un archivo XLS con el nombre "*Selection*" donde el programa guardará toda la información de este análisis (Figura M.11). Los resultados incluyen, para cada SNP en cada una de las cuatro poblaciones, el código del SNP, su posición, frecuencias alélicas, frecuencias genotípicas observadas y esperadas de acuerdo al equilibrio HW, el valor de p del test Chi cuadrado de equilibrio HW y el valor del test exacto de p, las frecuencias alélicas y genotípicas entre los progenitores, entre los no progenitores (como población control) y entre los descendientes, con las correspondientes comparaciones entre frecuencias genotípicas observadas y esperadas en cada caso. Además, compara mediante un test exacto de p las frecuencias alélicas entre cada par de grupos (población general, progenitores, no progenitores y descendientes). Y por último, establece las frecuencias observadas y esperadas (de acuerdo a un modelo panmíctico) de cruzamientos genotípicos en las parejas de progenitores, con su correspondiente Chi cuadrado de contingencia.

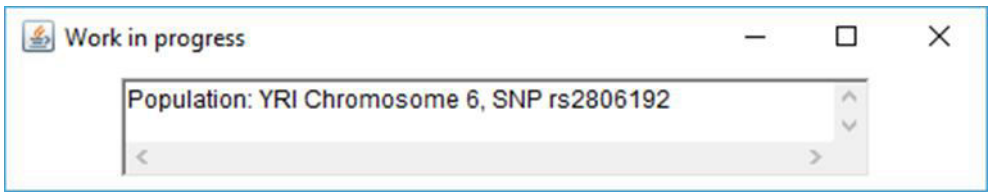
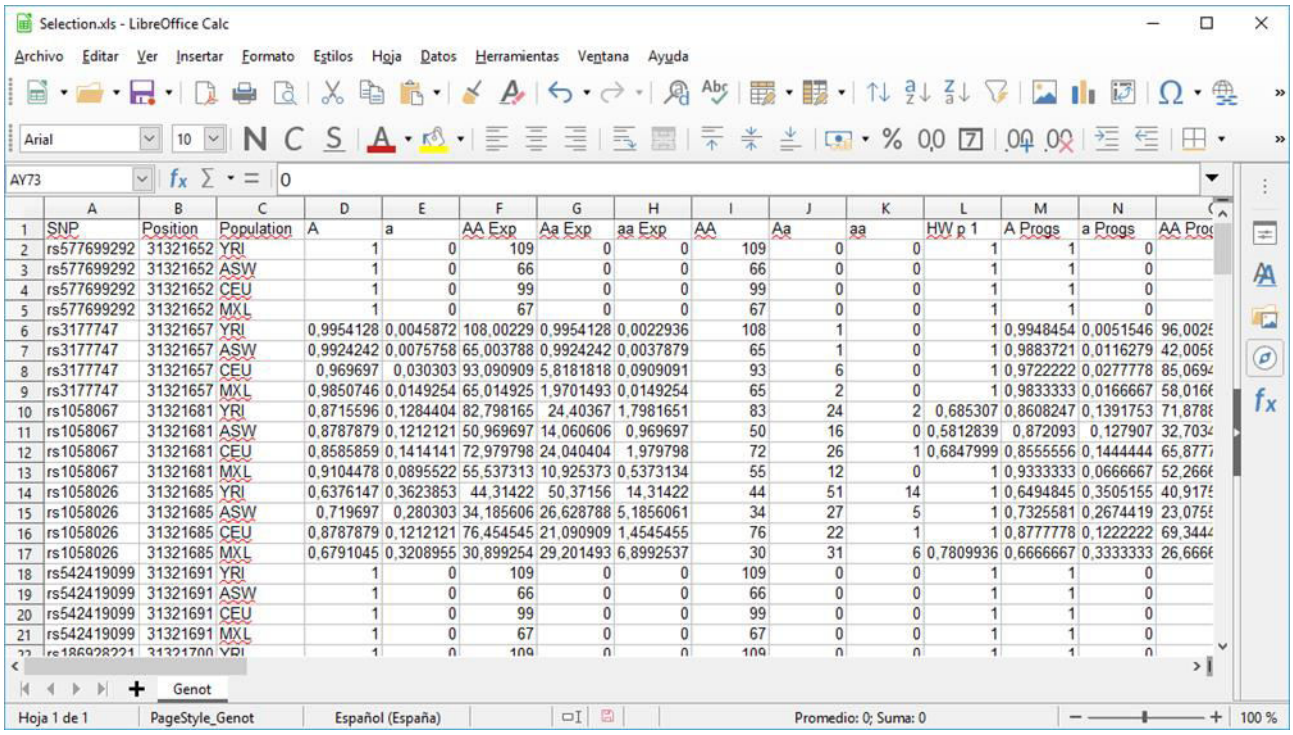


Figura M.10: Interfaz de FstMap durante el análisis de selección. En el instante capturado en esta imagen el programa está analizando el SNP rs2806192 para la población YRI (Yoruba de Ibadán, Nigeria).



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	SNP	Position	Population	A	a	AA Exp	Aa Exp	aa Exp	AA	Aa	aa	HW p 1	A Progs	a Progs	AA Prog
2	rs577699292	31321652	YRI	1	0	109	0	0	109	0	0	1	1	0	
3	rs577699292	31321652	ASW	1	0	66	0	0	66	0	0	1	1	0	
4	rs577699292	31321652	CEU	1	0	99	0	0	99	0	0	1	1	0	
5	rs577699292	31321652	MXL	1	0	67	0	0	67	0	0	1	1	0	
6	rs3177747	31321657	YRI	0,9954128	0,0045872	108,00229	0,9954128	0,0022936	108	1	0	1	0,9948454	0,0051546	96,0025
7	rs3177747	31321657	ASW	0,9924242	0,0075758	65,003788	0,9924242	0,0037879	65	1	0	1	0,9883721	0,0116279	42,0058
8	rs3177747	31321657	CEU	0,969697	0,030303	93,090909	5,8181818	0,0909091	93	6	0	1	0,9722222	0,0277778	85,0694
9	rs3177747	31321657	MXL	0,9850746	0,0149254	65,014925	1,9701493	0,0149254	65	2	0	1	0,9833333	0,0166667	58,0166
10	rs1058067	31321681	YRI	0,8715596	0,1284404	82,798165	24,40367	1,7981651	83	24	2	0,685307	0,8608247	0,1391753	71,8788
11	rs1058067	31321681	ASW	0,8787879	0,1212121	50,969697	14,060606	0,969697	50	16	0	0,5812839	0,872093	0,127907	32,7034
12	rs1058067	31321681	CEU	0,8585859	0,1414141	72,979798	24,040404	1,979798	72	26	1	0,6847999	0,8555556	0,1444444	65,8777
13	rs1058067	31321681	MXL	0,9104478	0,0895522	55,537313	10,925373	0,5373134	55	12	0	1	0,9333333	0,0666667	52,2666
14	rs1058026	31321685	YRI	0,6376147	0,3623853	44,31422	50,37156	14,31422	44	51	14	1	0,6494845	0,3505155	40,9175
15	rs1058026	31321685	ASW	0,719697	0,280303	34,185606	26,628788	5,1856061	34	27	5	1	0,7325581	0,2674419	23,0755
16	rs1058026	31321685	CEU	0,8787879	0,1212121	76,454545	21,090909	1,4545455	76	22	1	1	0,8777778	0,1222222	69,3444
17	rs1058026	31321685	MXL	0,6791045	0,3208955	30,899254	29,201493	6,8992537	30	31	6	0,7809936	0,6666667	0,3333333	26,6666
18	rs542419099	31321691	YRI	1	0	109	0	0	109	0	0	1	1	0	
19	rs542419099	31321691	ASW	1	0	66	0	0	66	0	0	1	1	0	
20	rs542419099	31321691	CEU	1	0	99	0	0	99	0	0	1	1	0	
21	rs542419099	31321691	MXL	1	0	67	0	0	67	0	0	1	1	0	

*Figura M.11: Vista parcial de una sección del archivo "*Selection*" generado por FstMap. Este archivo en concreto contaba con 145 filas y 59 columnas de resultados.*

Las opciones "*Inconsistencies*" y "*Simulating illegitimacy*" revisa los tríos putativos de las poblaciones HapMap para verificar que las combinaciones genotípicas son compatibles, permitiendo eliminar aquellos casos con una inconsistencia fruto de una ilegitimidad, error de genotipado o de otro tipo. Esto permite verificar la validez de los datos, y excluir aquellos inadecuados, que pudieran afectar a los análisis posteriores.

En la opción del menú Fst se realizan diferentes análisis en relación al índice de fijación.

En la primera opción, "*Fst Map*", mostrará un plano de coordenadas cartesianas con el eje horizontal (x) representando la posición de cada SNP, y con el eje vertical (y) representando un valor de Fst, o índice de fijación, para cada posición (Figura M.12). Señalará los valores máximo y mínimo encontrados para Fst, así como la primera y última posición del fragmento analizado. Veremos una nube de puntos, donde cada punto representa uno de los SNPs analizados con su posición y valor de Fst asociados. Estos valores corresponden con la medida de la varianza del

efecto Wahlund para cada marcador en todo el conjunto de poblaciones de la base de datos de 1000 Genomes. Este gráfico es similar al obtenido con la opción "*HW equilibrium*", pero los análisis subyacentes permitirán realizar una serie de análisis diferenciados en esta sección en relación a la distribución geográfica de los valores del índice.

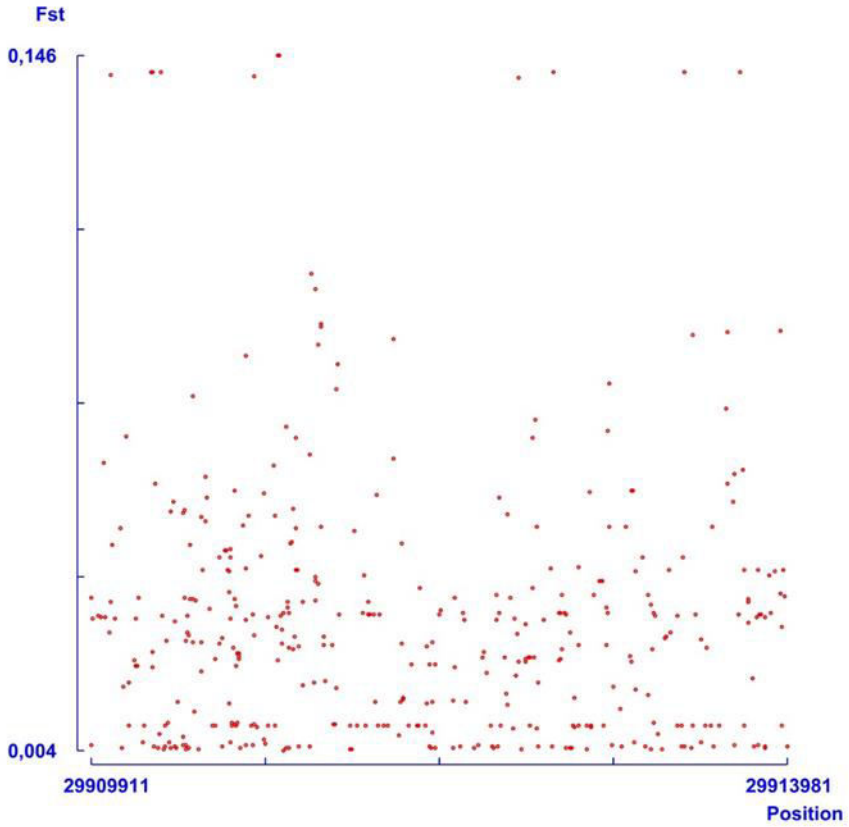


Figura M.12: Resultado de FstMap para el análisis de Fst o índice de fijación.

La siguiente opción, "*Fst Map > 1 pop*", calculará la varianza de Wahlund sólo en aquellos marcadores que presenten polimorfismo en más de una población en el conjunto de poblaciones de la base de datos de 1000 Genomes (Figura M.13).

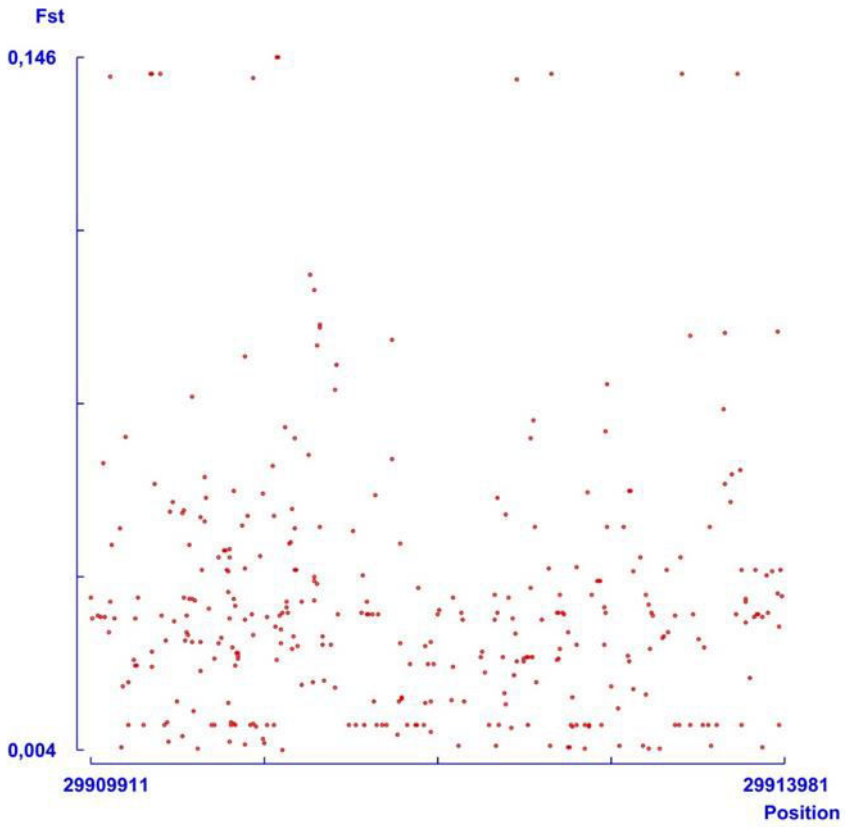


Figura M.13: Resultado de FstMap para el análisis de Fst Map >1 pop. Podemos observar como varios marcadores con valores de Fst bajos desaparecen en comparación con el gráfico de la Figura M.12.

La opción “Fst Map AIMs” permitirá calcular la varianza de Wahlund para el conjunto de poblaciones de los marcadores que correspondan con las características de los AIMs (Ancestry Informative Marker), es decir, marcadores que exhiben frecuencias sustancialmente diferentes entre diferentes poblaciones, con al menos una diferencia de 0,3 entre dos de ellas (Figura M.14). Se puede usar un conjunto de AIMs para estimar la proporción de ancestros de unas poblaciones de referencia en una población determinada, así como para categorizar diferentes individuos de una población mestiza.

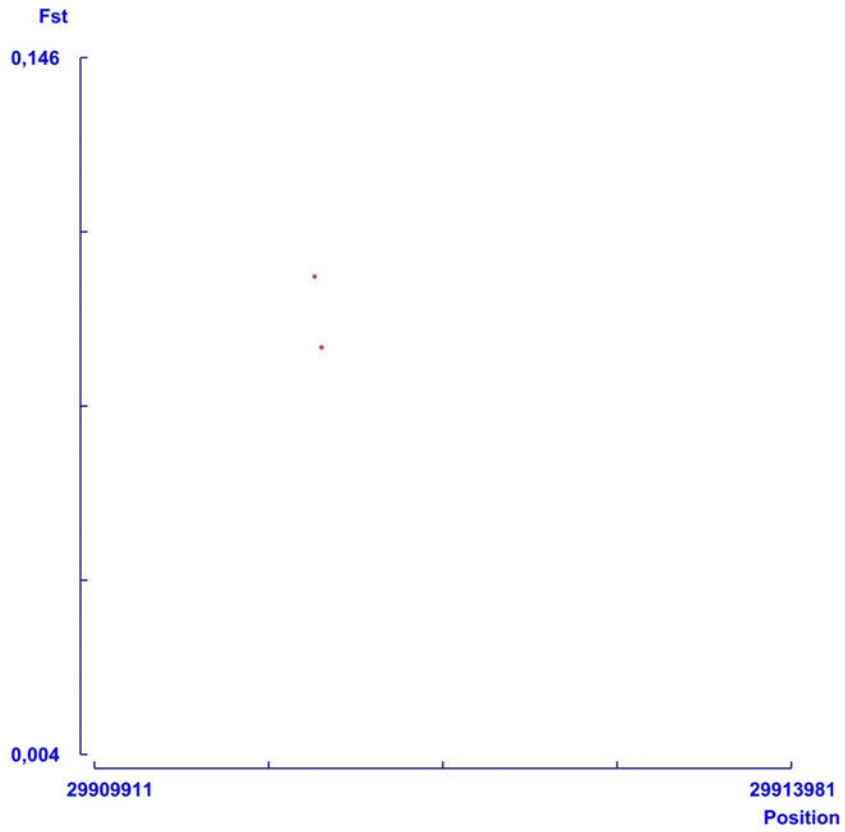


Figura M.14: Resultado de FstMap para el análisis de Fst Map AIMs. Podemos observar como casi todos los marcadores desaparecen en comparación con los gráficos de las figuras anteriores.

En la pestaña "*Geographic patterns*" se realiza un análisis de búsqueda automática de patrones de distribución geográfica, para eventualmente determinar el posible origen geográfico de cada patrón y marcador.

La opción "*Detect patterns*" es la primera de esta serie. Analiza todos los patrones de distribución de frecuencias por poblaciones para cada SNPs y busca similitudes entre ellos. Mediante un análisis de mínimos cuadrados, agrupa las distribuciones en una serie de patrones en función de su similitud. Al acabar muestra un plano de coordenadas cartesianas con el eje horizontal (x) representando la posición de cada marcador, y con el eje vertical (y) representando el patrón al cual se asocia el marcador en cuestión según el programa. Se observa una nube de puntos, donde cada punto representa un SNP con su posición y el patrón al que pertenece (Figura M.15).

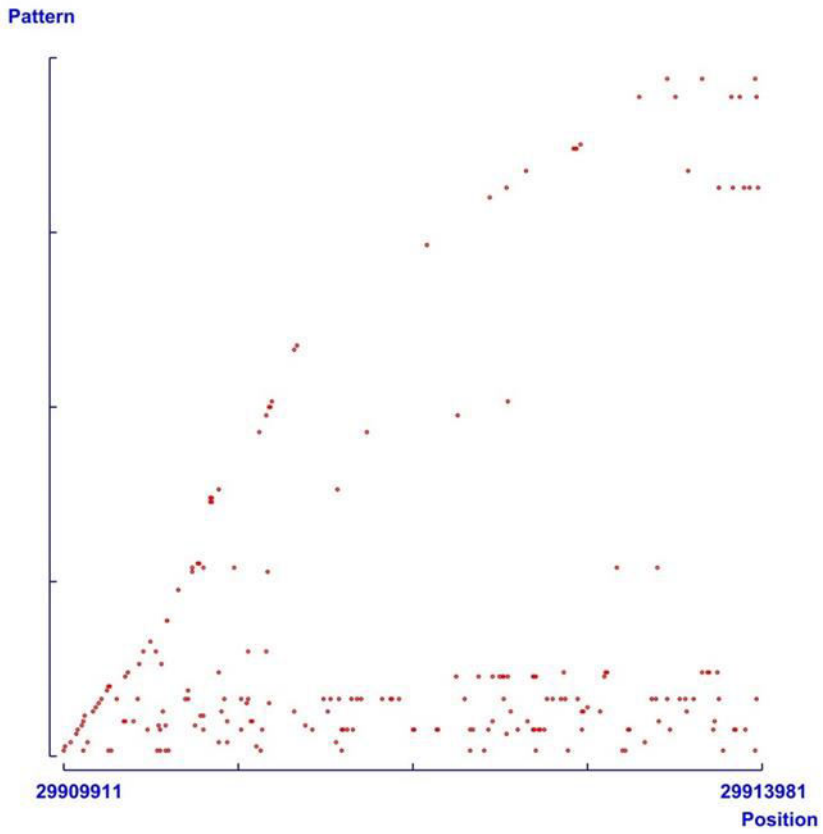


Figura M.15: Resultado de FstMap para el análisis de patrones de distribución de los marcadores. Se puede observar que hay marcadores que comparten el mismo patrón de distribución.

La opción "*Show patterns*" muestra una ventana donde puede seleccionarse un patrón concreto. El programa muestra un gráfico de frecuencias donde en el eje horizontal (x) se representan las 26 poblaciones de la base de datos del proyecto 1000 Genomes y en el eje vertical (y) se representa la frecuencia alélica para los marcadores que comparten ese patrón (Figura M.16). Las poblaciones están divididas en grupos según su origen geográfico: africanas (LWK, YRI, ESN, GWD, MSL, ACB y ASW), europeas (IBS, TSI, CEU, GBR y FIN), sur de Asia (PJL, BEB, GIH, ITU y STU), este de Asia (CHB, CDX, CHS, KHV y JPT) o americanas (CLM, MXL, PEL y PUR). Además, se informa del número de SNPs que presentan este patrón de distribución, y para cada SNP se muestra su propia línea representando sus frecuencias alélicas.

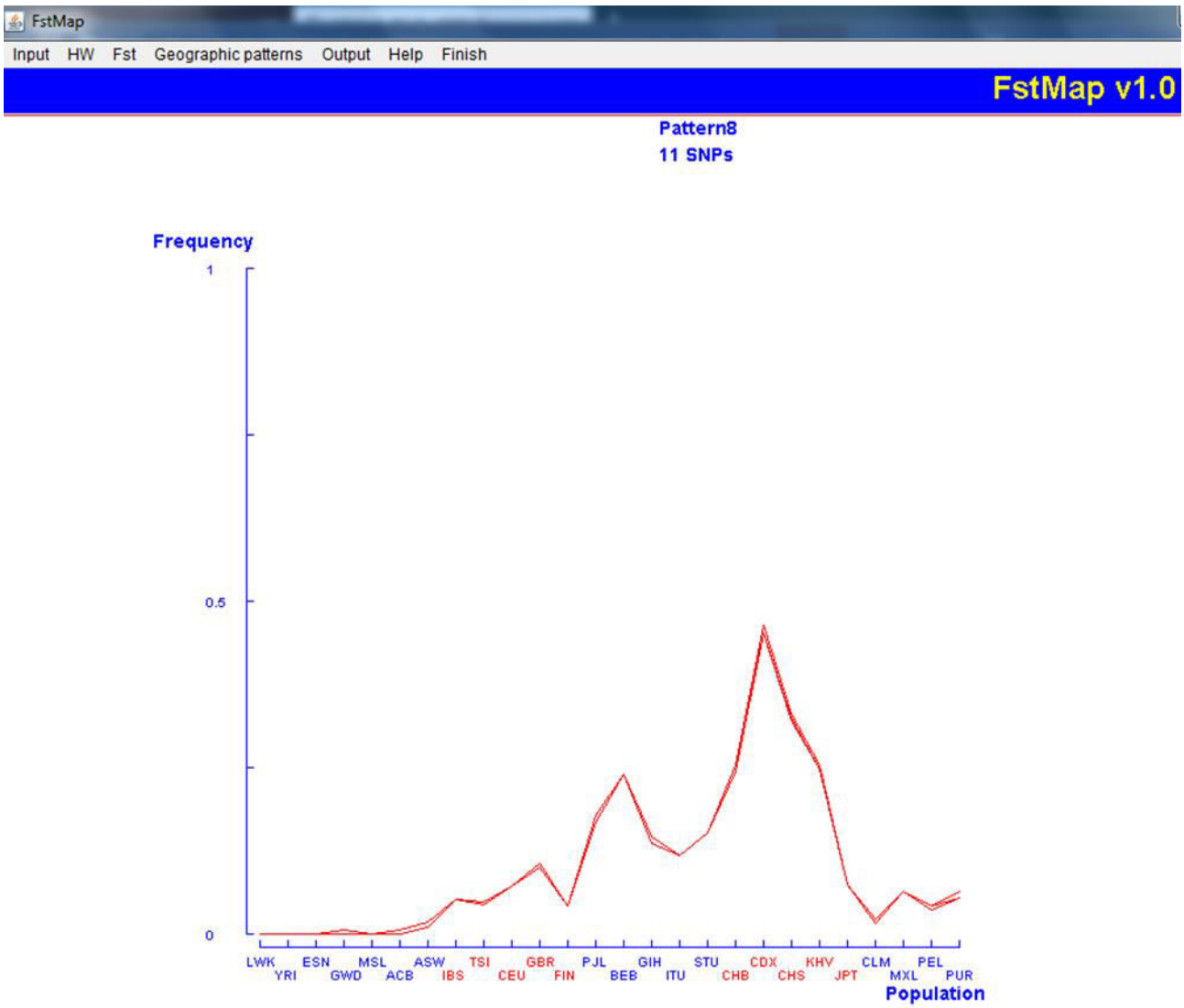


Figura M.16: Interfaz de FstMap para el análisis de un patrón determinado. Se informa que hay 11 SNPs que comparten este patrón y se observa que las frecuencias de dichos SNPs son casi totalmente coincidentes en todas las poblaciones.

La opción "Locate patterns" muestra de nuevo la nube de puntos generada al seleccionar la opción "Detect patterns", pero señalando en este caso con un círculo azul la posición de cada SNP que se ajusta al patrón seleccionado (Figura M.17). Esto permite visualizar a la vez las posiciones que ocupan los loci asociados a ese patrón concreto.

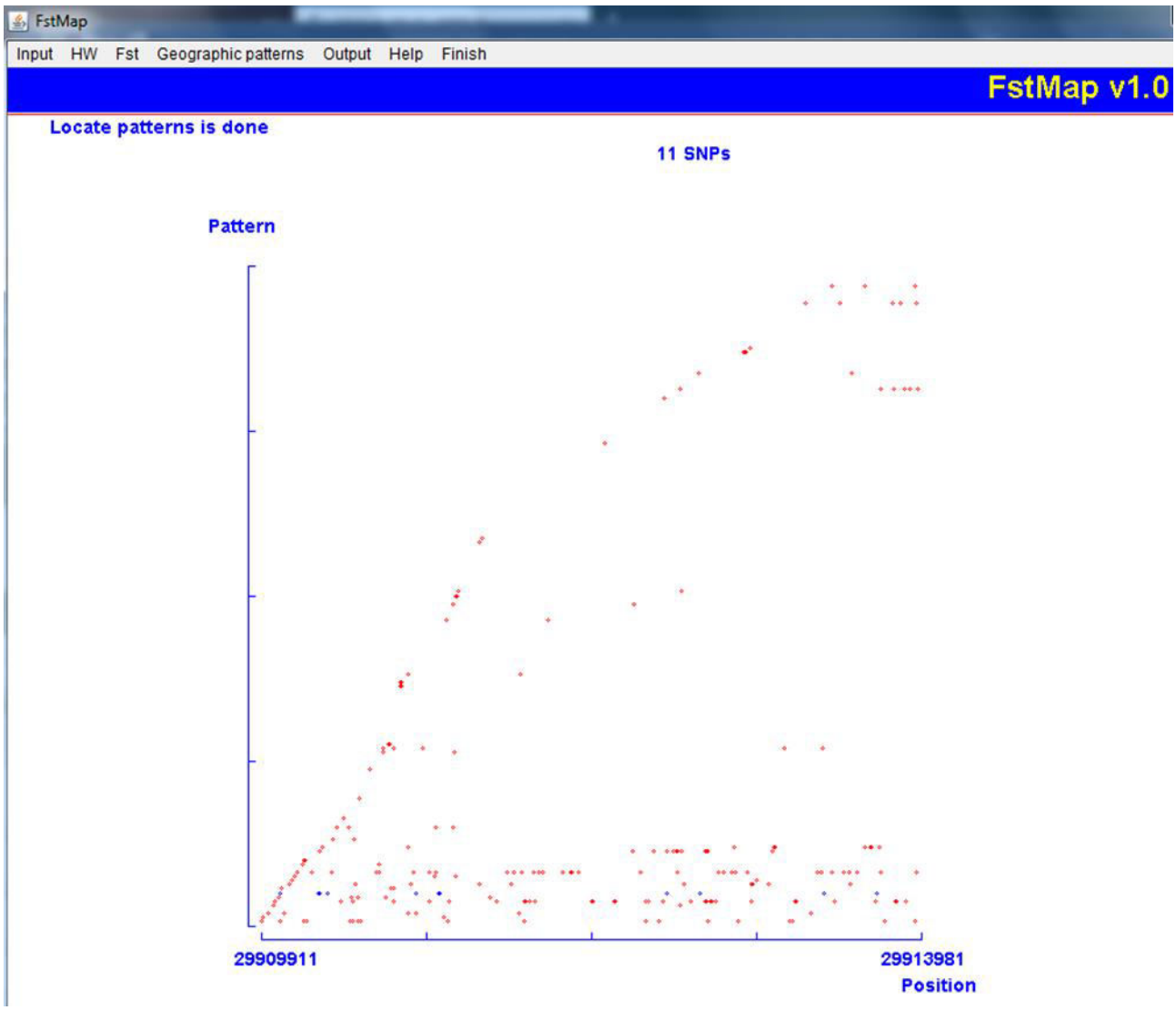


Figura M.17: Interfaz de FstMap para la localización de un patrón determinado. Podemos observar en la zona inferior del gráfico 11 círculos azules indicando los SNPs que comparten dicho patrón de distribución.

Las siguientes tres opciones permiten identificar marcadores que se asocian con los diferentes grupos continentales (África, Europa y el este de Asia), teniendo en cuenta sus valores de Fst y los patrones analizados por el programa. "*Identify patterns: Africa*" mostrará una nube de puntos con los SNPs que haya asociado con un patrón de distribución africano en un gráfico con el eje horizontal (x) representando la posición de cada SNP, y con el eje vertical (y) representando un valor de Fst para cada posición (Figura M.18). También mostrará un gráfico con las frecuencias en las diferentes poblaciones de los SNPs asociados a este patrón (Figura M.19). Se entiende como patrón africano aquél en el que las frecuencias del alelo de referencia son, en la mayor parte de las poblaciones africanas, más altas o más bajas que en el resto de poblaciones. "*Identify patterns: Europe*" (Figuras II.20 y II.21) e "*Identify patterns: East Asia*" (Figuras II.22 y II.23) mostrarán los mismos gráficos asociados a Europa y el Este de Asia respectivamente. No se ha desarrollado la rutina correspondiente al continente americano, dado que las poblaciones que lo representan son eminentemente mestizas. Tampoco se ha estimado la opción del sur de Asia, por presentar características intermedias entre Europa y el Este de Asia.

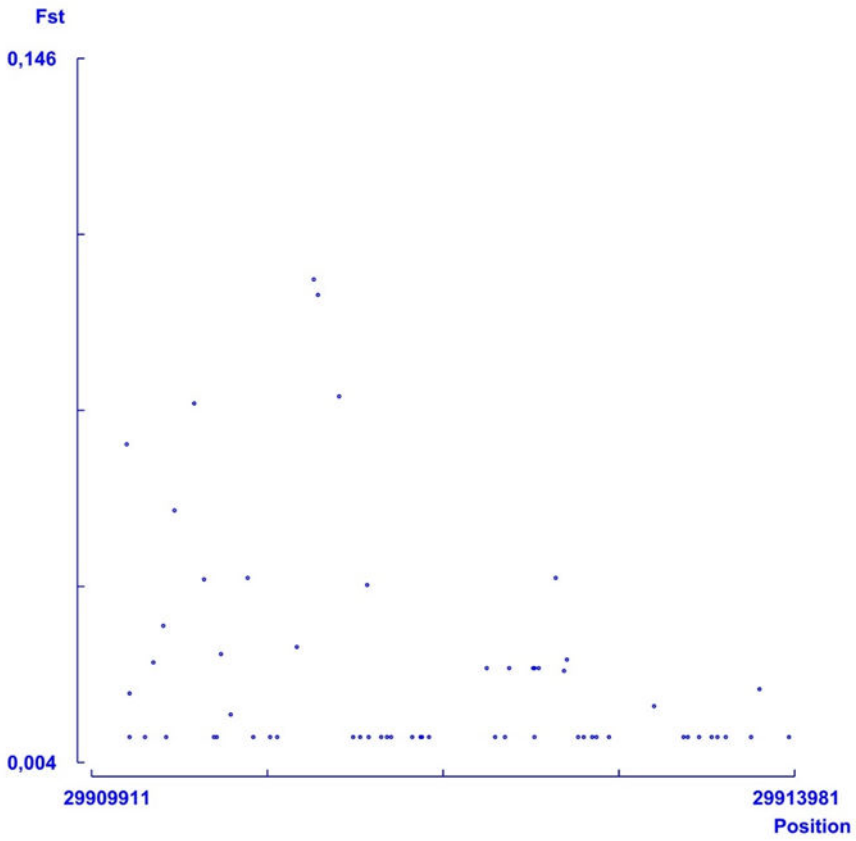


Figura M.18: Resultado de FstMap para los marcadores con patrón de distribución africano. Los marcadores están representados con puntos azules.

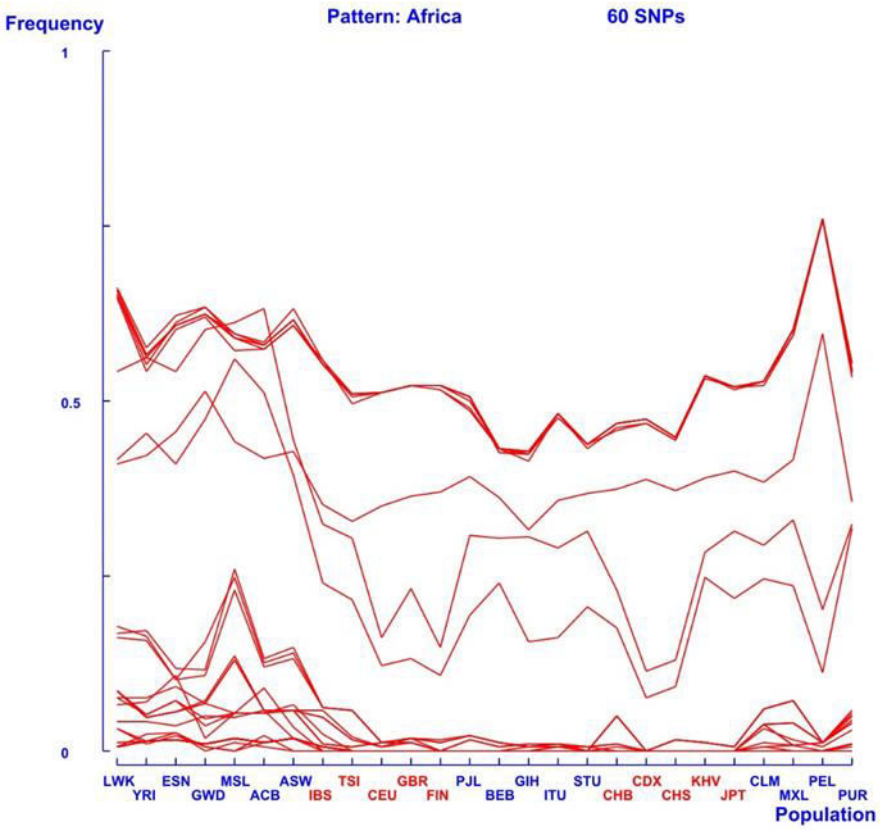


Figura M.19: Distribución geográfica de frecuencias del SNP asociado al patrón africano.

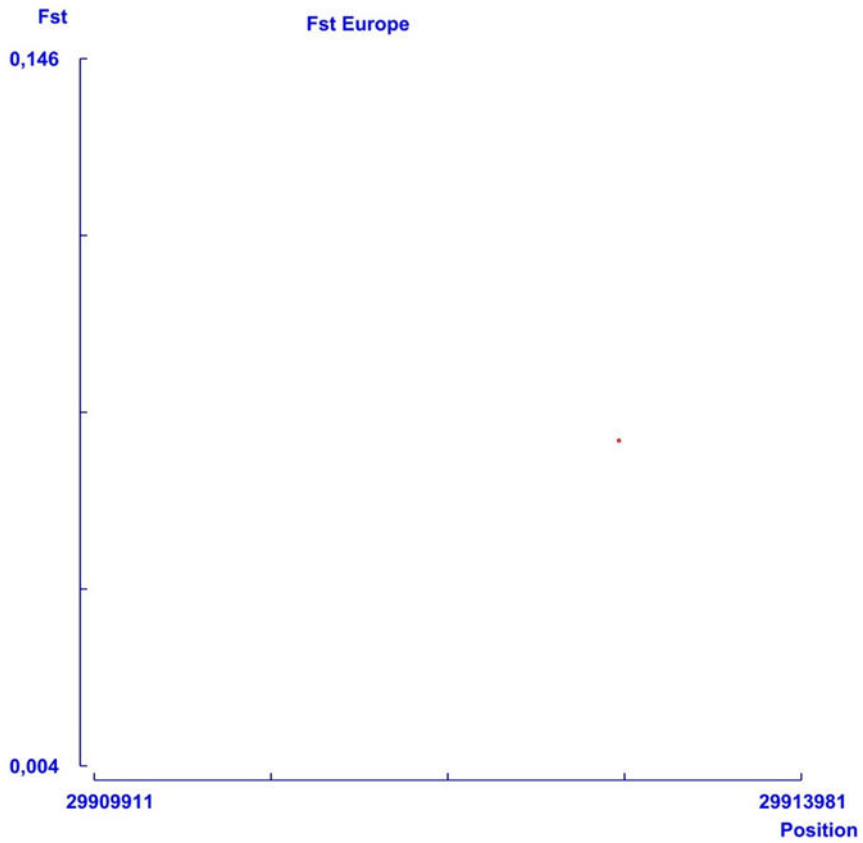


Figura M.20: Resultado de FstMap para los marcadores con patrón de distribución europeo. El marcador está representado con un punto rojo.

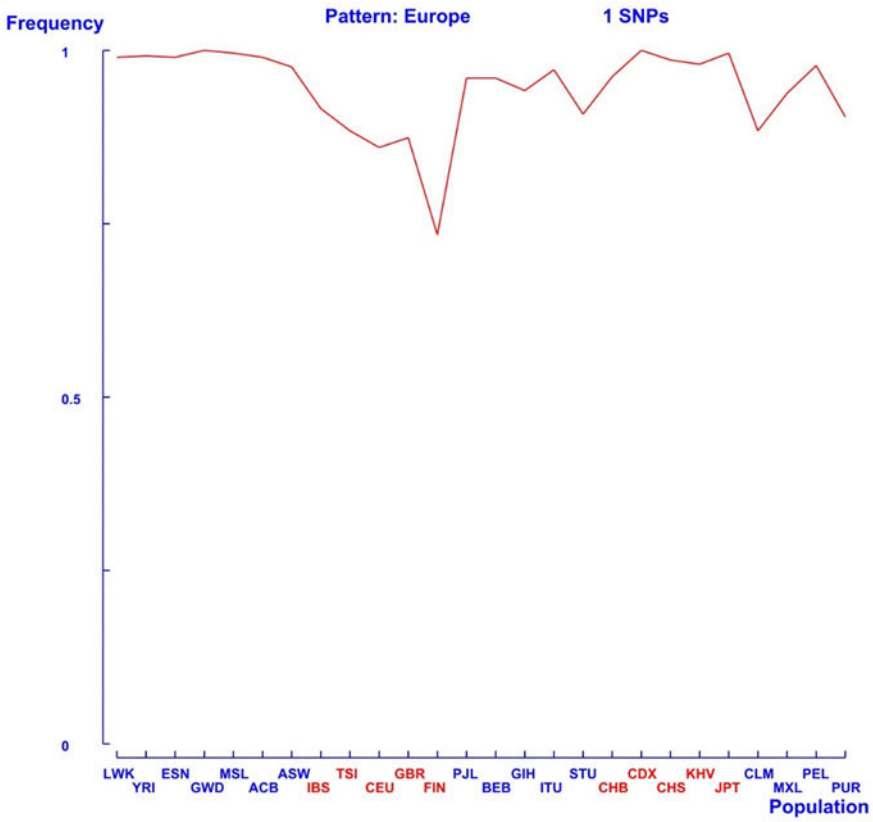


Figura M.21: Distribución geográfica de frecuencias del SNP asociado al patrón europeo.

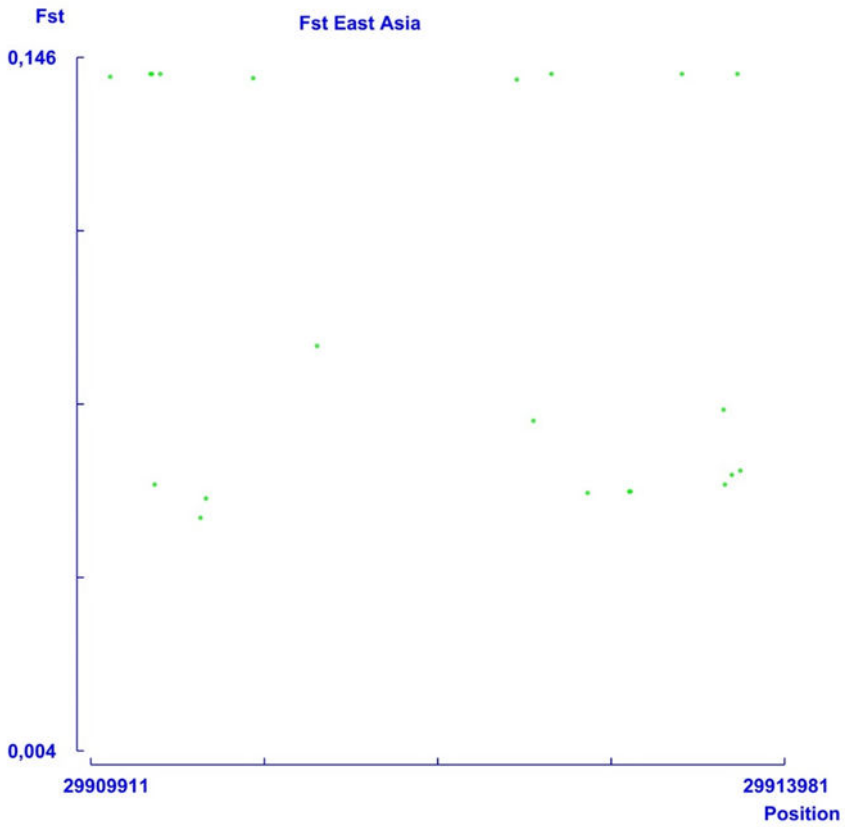


Figura M.22: Resultado de FstMap para los marcadores con patrón de distribución del Este de Asia. Los marcadores están representados con puntos verdes.

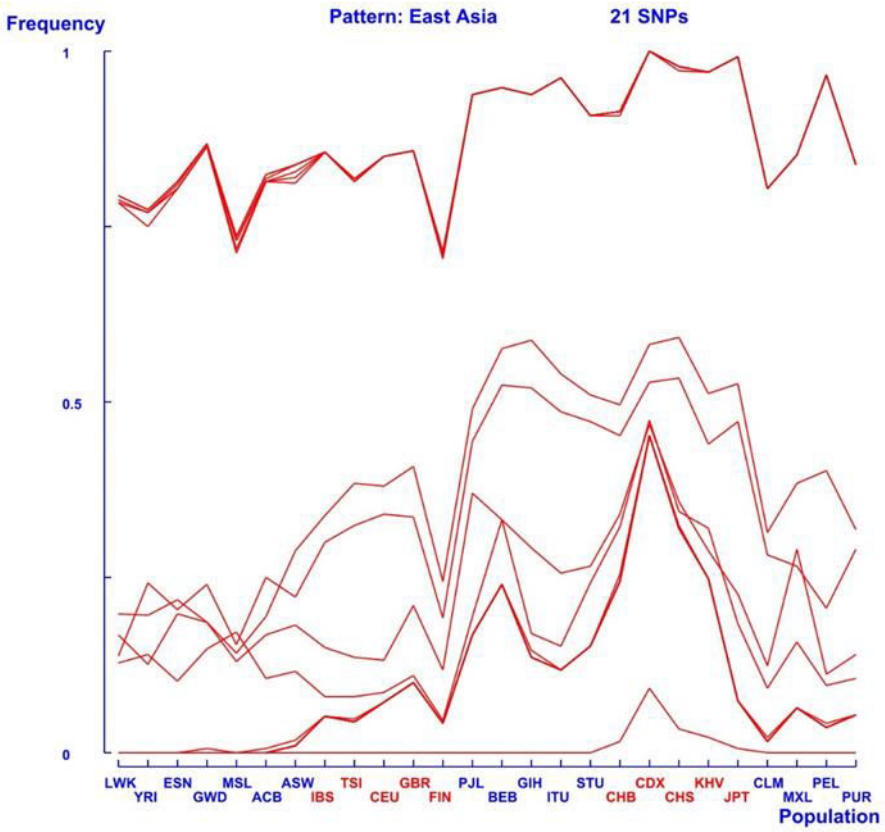


Figura M.23: Distribución geográfica de frecuencias del SNP asociado al patrón del Este de Asia.

La opción "*Fst Map by continents*" muestra una superposición de las nubes de puntos de las tres opciones anteriores ("*Identify patterns*") a fin de poder visualizar la distribución y variabilidad continental de los marcadores en conjunto. Los puntos azules, rojos y verdes representan a los marcadores con patrón de distribución africana, europea y del este asiático respectivamente. Los puntos grises representan los marcadores que no pueden ser incluidos en una de estas 3 categorías, al carecer de una distribución de frecuencias y valores de Fst que puedan asociarse de manera inequívoca y exclusiva a uno de los 3 grupos continentales (Figura M.24).

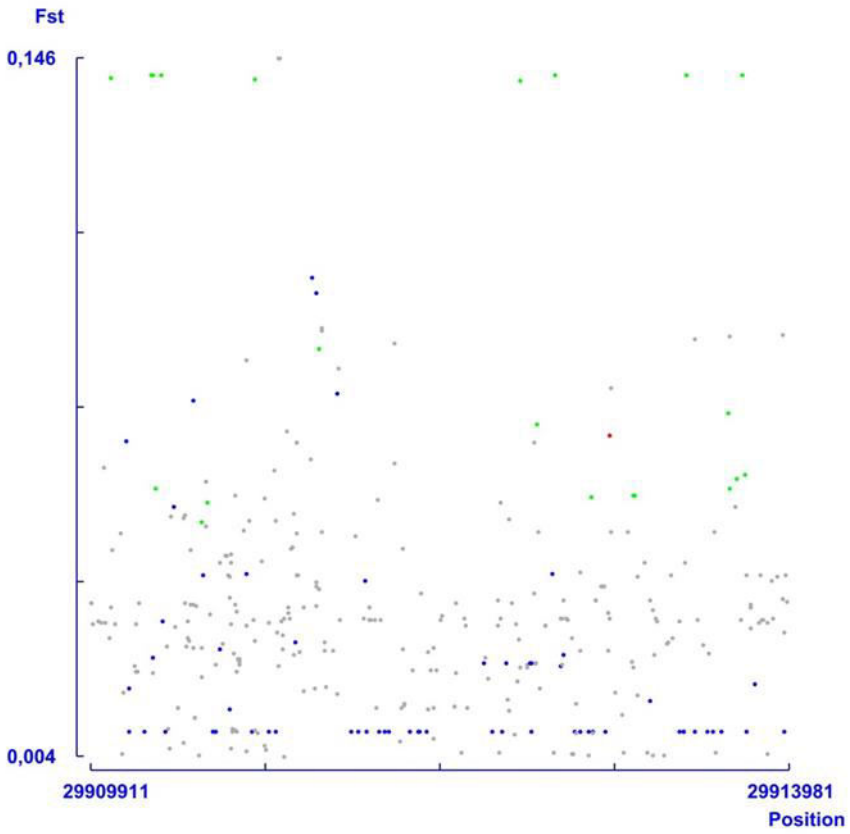


Figura M.24: Resultado de FstMap para los marcadores según su distribución continental.

La pestaña "*Output*", con la opción "*Save results*", permite generar un archivo XLS (Results.xls), que contiene, para cada SNP, las frecuencias alélicas promedio y por población, test de equilibrio Hardy-Weinberg por población (incluyendo los valores de p) y valores de Fst (Figura M.25).

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	FstMap v.1.0 @Jose A. Pena 2017														
2	Population	26	Chromoso	6											
3	SNPs	420								N of popul	Frequencies				
4	Allele	Position	Mean freq	Wahlund v	Wahlund >1 pop		AIMs	Group	Continent	disequilibri	LWK	YRI	ESN	GWD	MSL
5	rs1828643	29909911	0,000205	0,005116				0		0	0	0	0	0	0
6	rs9260111	29909916	0,681731	0,035172	0,035172			1		0	0,717172	0,634259	0,722222	0,769912	0,658824
7	rs9260112	29909918	0,681731	0,035172	0,035172			1		0	0,717172	0,634259	0,722222	0,769912	0,658824
8	rs2734903	29909924	0,363671	0,030873	0,030873			2		3	0,444444	0,324074	0,40404	0,376106	0,317647
9	rs7923133	29909957	0,088711	0,031504	0,031504			3		0	0,035354	0,083333	0,060606	0,154867	0,123529
10	rs1163951	29909970	0,068301	0,03117	0,03117			0		0	0,111111	0,12037	0,146465	0,110619	0,088235
11	rs9260114	29909988	0,732192	0,062711	0,062711			5		1	0,868687	0,773148	0,808081	0,867257	0,870588
12	rs3515822	29909999	0,151854	0,031385	0,031385			6		1	0,070707	0,115741	0,121212	0,146018	0,117647
13	rs2735115	29910022	0,02789	0,028226	0,028226			7		1	0,065657	0,009259	0,005051	0,035398	0,011765
14	rs9260115	29910027	0,675107	0,03437	0,03437			1		0	0,707071	0,611111	0,707071	0,761062	0,629412
15	rs4154122	29910033	0,10101	0,141866	0,141866			8	3	0	0	0	0	0,004425	0
16	rs2735114	29910034	0,39801	0,045933	0,045933			9		2	0,308081	0,25463	0,313131	0,300885	0,311765
17	rs4127254	29910057	0,088526	0,031055	0,031055			3		0	0,035354	0,083333	0,060606	0,154867	0,123529
18	rs3132688	29910083	0,925851	0,049358	0,049358			10		0	0,944444	0,912037	0,90404	0,969027	0,929412
19	rs5657000	29910094	0,000372	0,004479	0,004479			0		0	0,005051	0,00463	0	0	0
20	rs9260116	29910105	0,040314	0,017137	0,017137			11		0	0,060606	0,027778	0,035354	0,070796	0
21	rs1196425	29910122	0,056677	0,068009	0,068009			12	1	0	0,161616	0,157407	0,10101	0,106195	0,229412
22	rs1875141	29910132	0,003911	0,009197	0,009197			13	1	0	0,005051	0,013889	0,015152	0,00885	0,017647
23	rs1157431	29910135	0,003027	0,018024	0,018024			0	1	0	0,005051	0,023148	0,025253	0,004425	0
24	rs2735113	29910167	0,067236	0,022564	0,022564			15		0	0,151515	0,083333	0,116162	0,075221	0,064706
25	rs5698751	29910175	0,023817	0,021263	0,021263			16		0	0,025253	0,037037	0,025253	0,022124	0,082353
26	rs9260118	29910176	0,666816	0,03109	0,03109			1		0	0,69697	0,606481	0,712121	0,743363	0,611765
27	rs5556148	29910179	0,023817	0,021263	0,021263			16		0	0,025253	0,037037	0,025253	0,022124	0,082353
28	rs5723850	29910180	0,023817	0,021263	0,021263			16		0	0,025253	0,037037	0,025253	0,022124	0,082353
29	rs5411034	29910182	0,023817	0,021263	0,021263			16		0	0,025253	0,037037	0,025253	0,022124	0,082353
30	rs9260119	29910189	0,68462	0,035167	0,035167			1		1	0,712121	0,666667	0,732323	0,769912	0,688235
31	rs5780343	29910217	0,000226	0,005657				0		0	0	0	0	0	0

*Figura M.25: Vista parcial de una sección del archivo "*Results*" generado por FstMap. Este archivo en concreto contaba con 40 filas y 62 columnas de resultados.*

Cabe señalar que cada una de los gráficos generados por el programa, ya sea una nube de puntos o una distribución de frecuencias, se guardará automáticamente en un archivo PostScript encapsulado (EPS, *Encapsulated PostScript*). Tanto estos archivos como los de *Selection* y *Results* se guardarán en la carpeta donde está el ejecutable del programa FstMap. Para cada región cromosómica analizada se generará una carpeta de manera automática cuyo nombre será el código de tiempo (año/mes/día/hora/minuto/segundo) del momento en que se inició el análisis. El renombrado de esta carpeta no generará incompatibilidades de ningún tipo una vez acabados los análisis.

MEGA

El programa MEGA (versión 7.0.26) ha sido utilizado para realizar los alineamientos de las secuencias y generar los árboles que explican las relaciones evolutivas de los distintos alelos de los genes HLA (Figura M.26). La historia evolutiva representada por cada árbol se infirió utilizando el método Neighbor-Joining con un *bootstrap* de 1000 réplicas. Las distancias evolutivas se calcularon utilizando el método Tamura de 3 parámetros. Se eliminaron todas las posiciones que contenían huecos o datos desaparecidos.

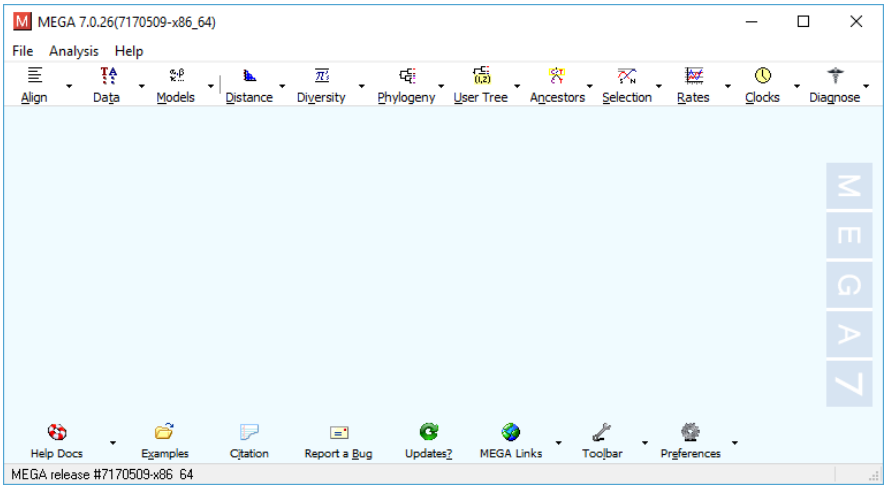


Figura M.26: Interfaz de MEGA.

Una vez obtenidas las secuencias tanto de poblaciones humanas actuales como de Neandertal de sus respectivas bases de datos, procederemos a introducirlas en MEGA. Para ello seleccionamos la pestaña *"File"* y elegimos la opción *"Open A File/Session"*. Seleccionamos el archivo que contiene las secuencias de humanos modernos. Preguntará como deseamos abrir este archivo FASTA (Figura M.27).

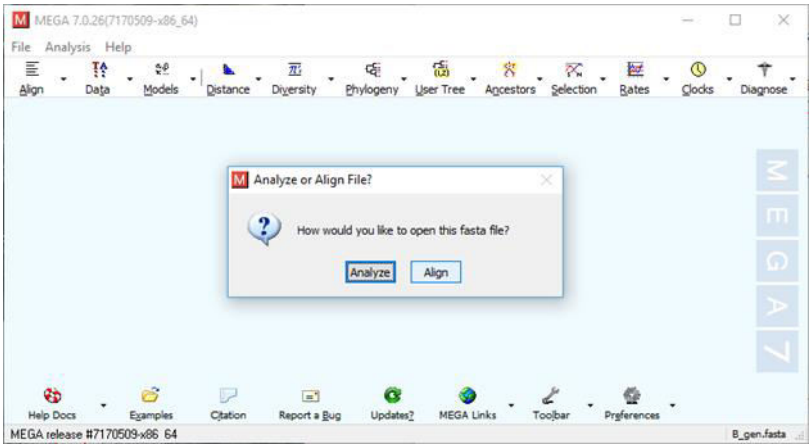


Figura M.27: Interfaz de MEGA después de seleccionar el archivo que deseamos introducir al programa.

Elegimos "*Align*" y el programa abrirá una ventana donde veremos los distintos haplotipos y sus secuencias (Figura M.28)

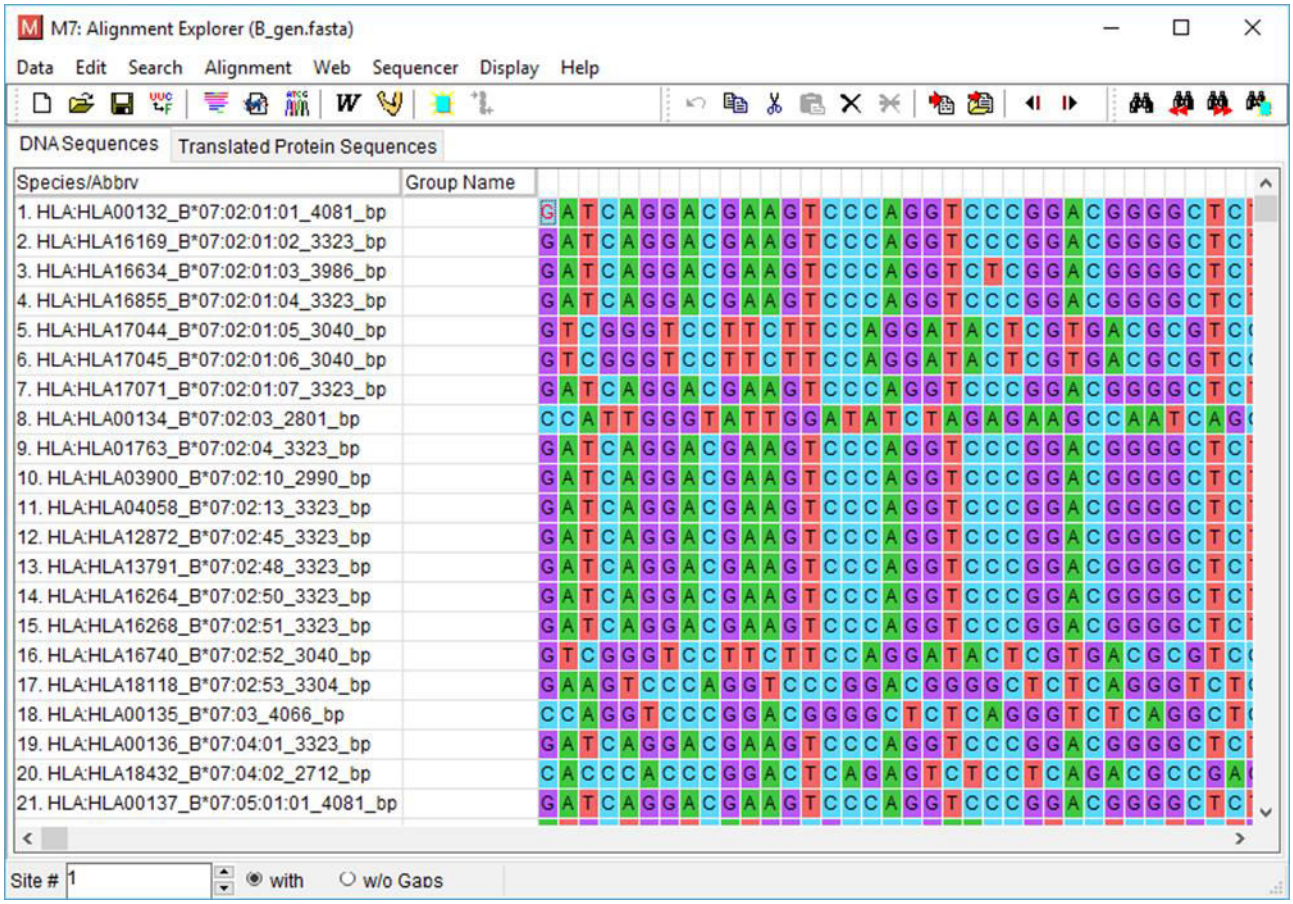


Figura M.28: Vista parcial de una sección de la ventana de MEGA después de introducir el archivo de las secuencias. En este caso, cada fila corresponde a un alelo del gen HLA-B. A la derecha podemos observar parte de las secuencias de cada alelo.

A continuación, debemos incluir la secuencia de ADN Neandertal que deseamos comparar. Para ello seleccionamos la pestaña "*Edit*" y la opción "*Insert sequence from file*". En la ventana que se abre, buscamos la secuencia de ADN Neandertal y la seleccionamos. Dicha secuencia se cargará, y añadirá a la lista anterior de secuencias (Figura M.29).

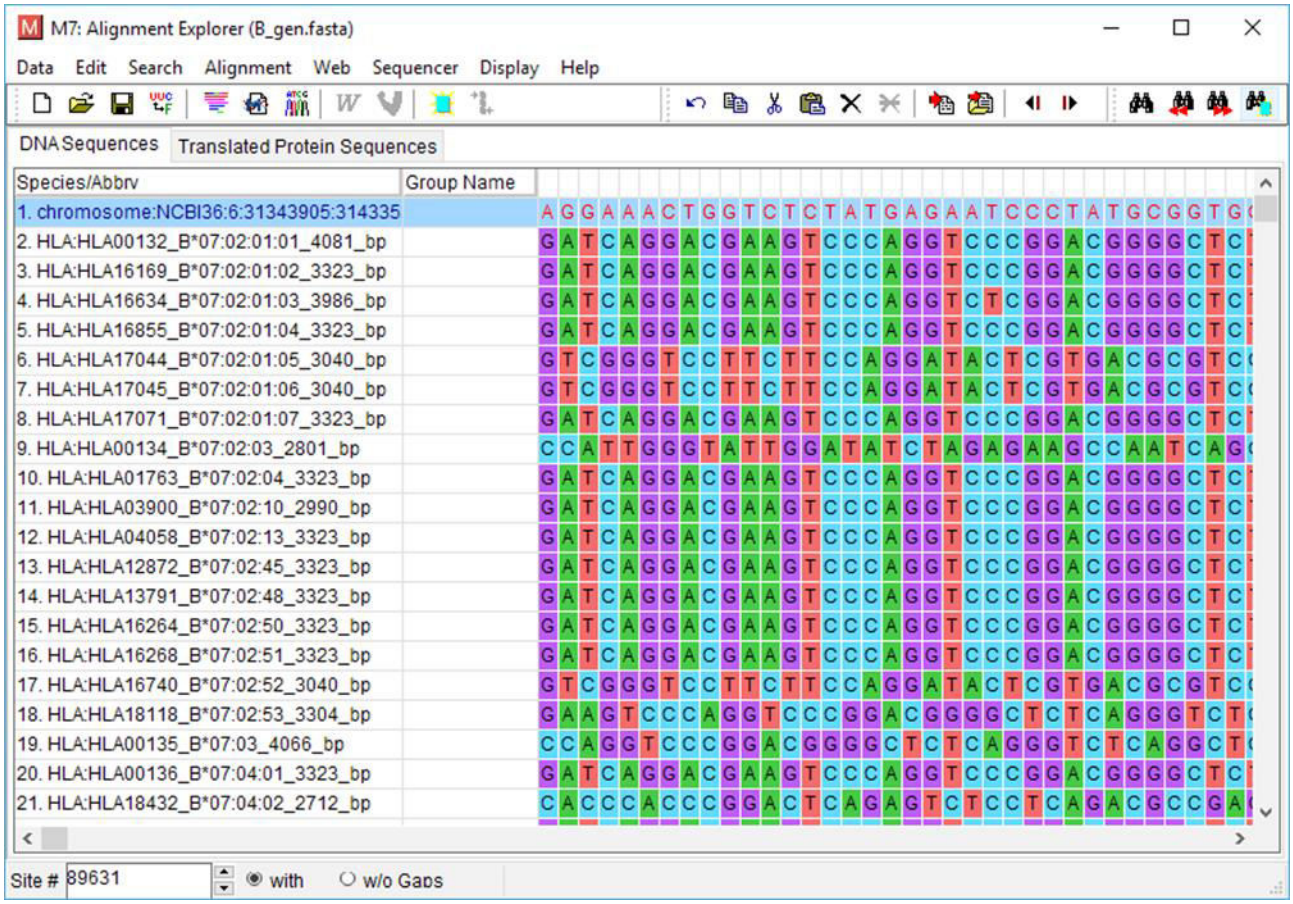


Figura M.29: Vista parcial de una sección de la ventana de MEGA después de introducir el archivo de la secuencia de ADN Neandertal. La fila resaltada corresponde con la secuencia de HLA-B Neandertal.

El siguiente paso es alinear las distintas secuencias. Para ello debemos seleccionar todas las secuencias a alinear, y luego en la pestaña "*Alignment*" elegir la opción "*Align By Muscle*". Se abrirá una ventana donde podremos elegir entre distintas configuraciones para realizar el alineamiento (Figura M.30). Una vez decididas, seleccionamos "*Compute*" para que el programa alinee las secuencias.

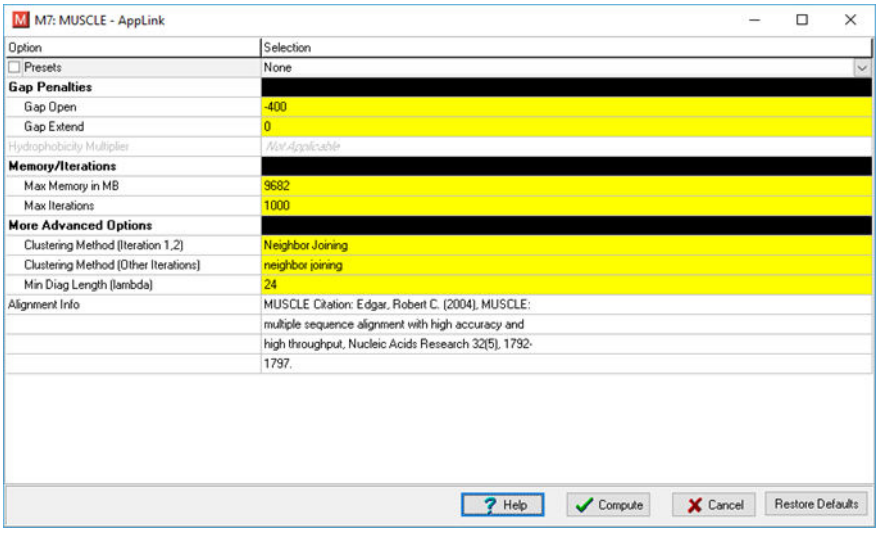


Figura M.30: Ventana de MEGA para la configuración del alineamiento de secuencias.

Una vez que el programa termina de realizar los cálculos, ofrece una ventana en la que vemos las secuencias ya alineadas (Figura M.31). Deberemos guardar el archivo de alineamiento.

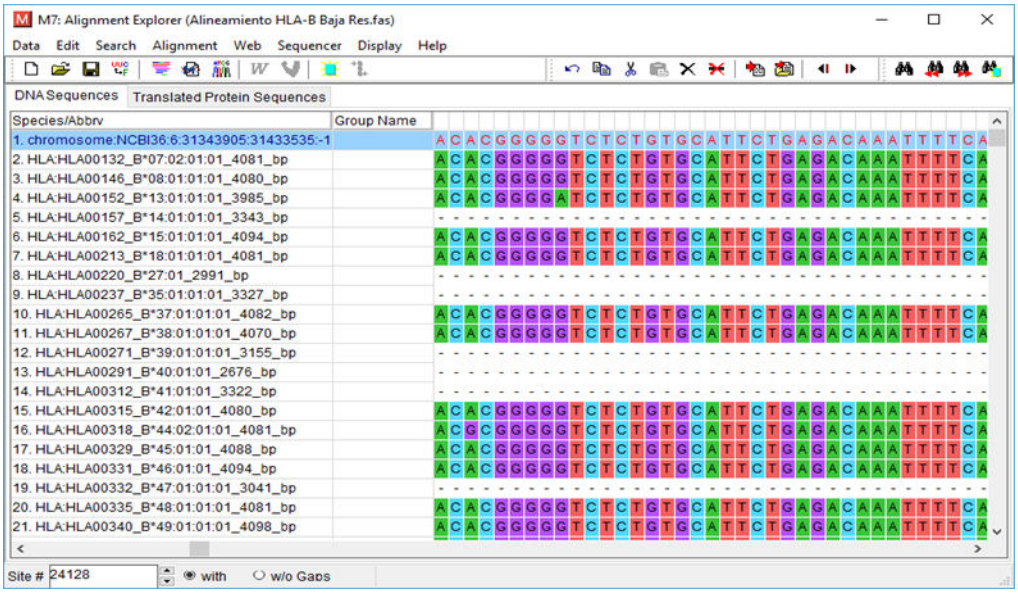


Figura M.31: Vista parcial de una sección de la ventana de MEGA después de realizar el alineamiento de secuencias. La secuencia de HLA-B Neandertal aparece resaltada. El programa marca los loci desaparecidos en alguna de las secuencias con un guion.

Entonces procederemos a realizar el árbol de relaciones evolutivas. Para ello, en la ventana de gráficos seleccionamos la pestaña "*Phylogeny*" y dentro, la opción "*Construct/Test Neighbor-Joining Tree*" (Figura M.32).

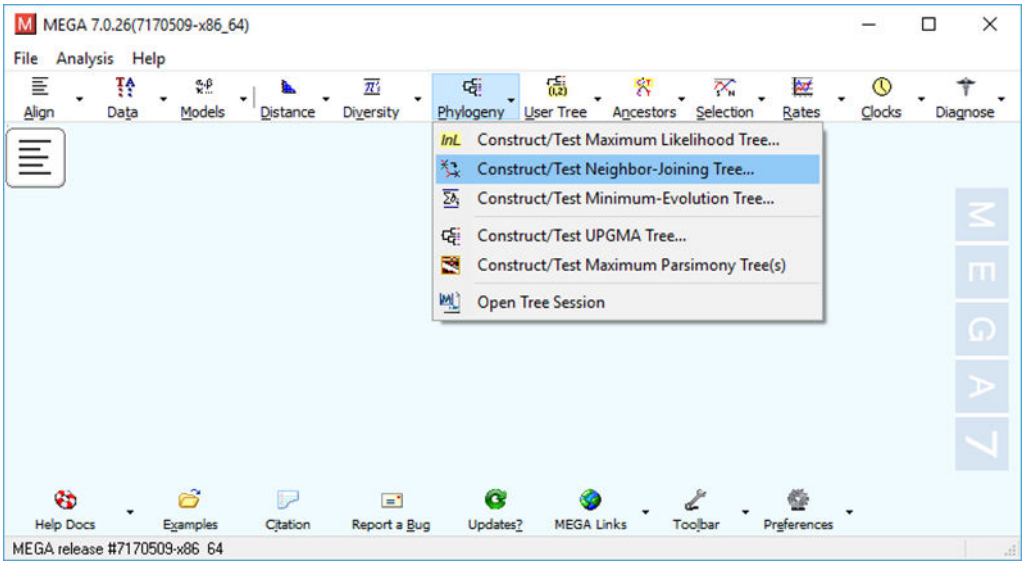


Figura M.32: Ventana de gráficos de MEGA.

Se abrirá una ventana en la que deberemos buscar el archivo del alineamiento que hemos guardado antes, y una vez seleccionado, pedirá que digamos qué tipo de secuencia es la que hemos introducido. En nuestro caso son secuencias nucleotídicas, por lo que seleccionamos esa

opción. Se abrirá otra ventana donde podremos elegir las opciones de configuración a la hora de realizar arboles evolutivos (Figura M.33). Una vez elegidos los parámetros de configuración, pinchamos en "*Compute*".

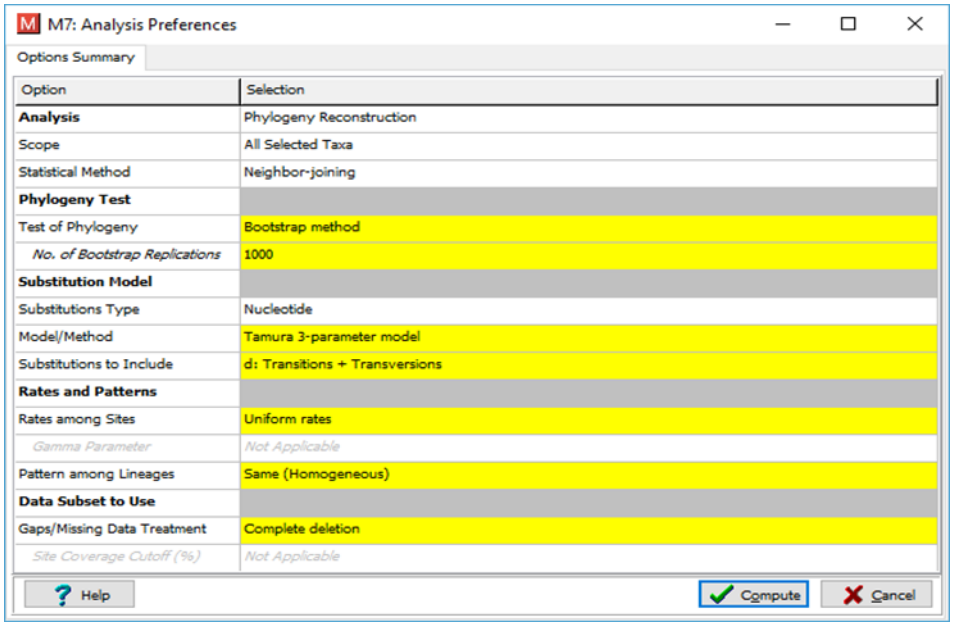


Figura M.33: Ventana de MEGA para la configuración de los árboles de relaciones evolutivas.

Una vez acabados los cálculos, el programa abrirá una nueva ventana con el árbol ya construido e información sobre el mismo (Figura M.34)

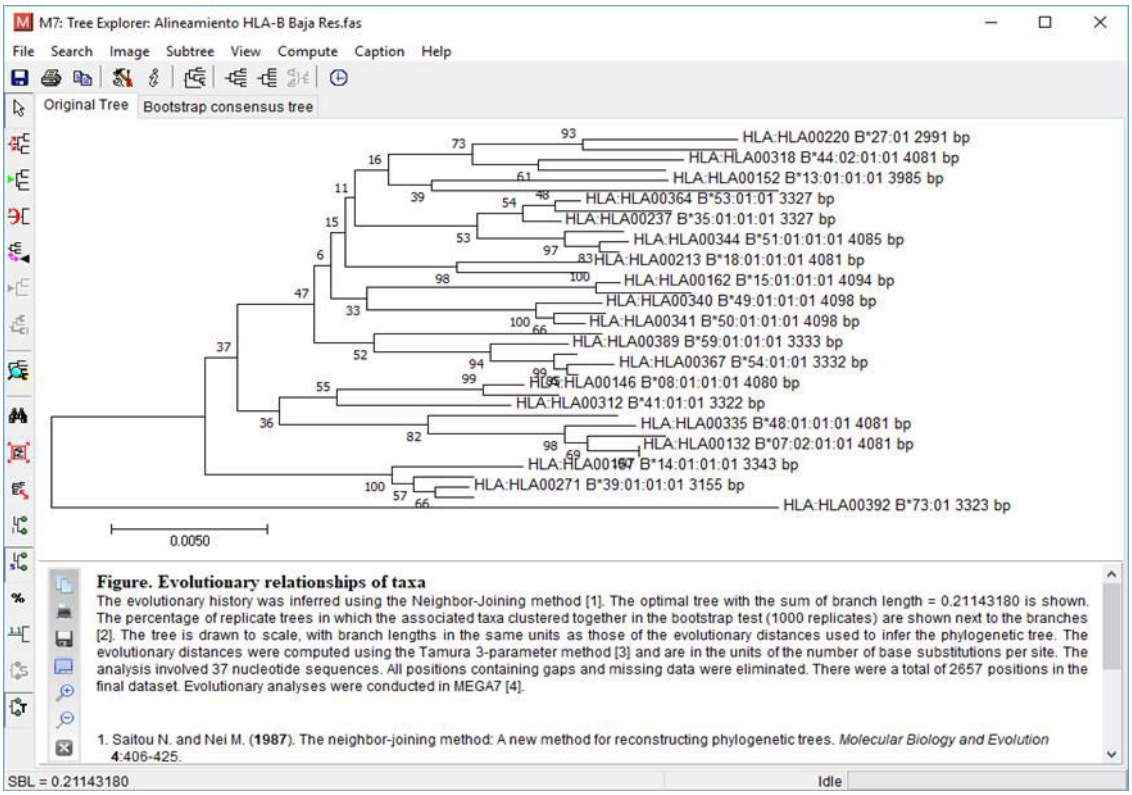


Figura M.29: Ejemplo de árbol de relaciones evolutivas para secuencias de HLA-B de humanos modernos y Neandertal.

Este árbol podremos guardarlo en formato MTSX para poder abrirlo directamente con MEGA o como una imagen que se podrá añadir a un documento.

Haploview

Haploview es un software bioinformático que está diseñado para analizar y visualizar patrones de desequilibrio de ligamiento en datos genéticos, así como para simplificar y acelerar el proceso de análisis de haplotipos al proporcionar una interfaz común para varias tareas relacionadas con dichos análisis. Está desarrollado y mantenido por el laboratorio del Dr. Mark Daly en el MIT/ Harvard Broad Institute. Actualmente, Haploview posee las siguientes funcionalidades:

1. Análisis de desequilibrios de ligamiento y de bloques de haplotipos.
2. Estimación de frecuencias haplotípicas en una población.
3. Pruebas de asociación entre SNPs y haplotipos concretos, y pruebas de permutación para la significación de la asociación.
4. Implementación del algoritmo de selección tagSNPs.
5. Descarga automática de datos genotípicos de HapMap.
6. Visualización y trazado de resultados de asociación de genoma completo de PLINK, incluyendo opciones de filtrado avanzadas.

Haploview es totalmente compatible con los volcados de datos del proyecto HapMap y el Perlegen Genotype Browser. Puede analizar miles de SNP (decenas de miles en modo de línea de comando) en miles de personas.

Muestras de ADN

Con el fin de estudiar posibles presiones selectivas en la población gitana se han analizado 94 muestras de individuos del banco de muestras de la Dra. Esther Rebato Ochoa. Dichas muestras han sido obtenidas del conjunto de la población gitana residente en el País Vasco. El procedimiento, denominado “Conservación de muestras de saliva (congeladas a -20°C) y extracción posterior de DNA” con número de registro CEIAB/16/2010/REBATO OCHOA de la Comisión Universitaria de Ética de la Investigación y la Docencia (CUEID) de la Universidad del País Vasco (UPV/EHU), presentado por la Dra. Rebato para su evaluación por el Comité de evaluación ética CEIAB, se adecuó a las exigencias metodológicas, éticas y jurídicas vigentes, y recibió el informe de certificado favorable. A los voluntarios sanos, debidamente informados y habiendo dado su consentimiento, se les sometió a un proceso de extracción de muestras de saliva no invasivo. Una parte de estas muestras corresponden a parientes (tríos madre/padre/descendiente) y otra parte configuran una muestra de no parientes. Una vez seleccionadas las muestras, se procedió a la realización de la extracción de material genético. Se ha seguido el protocolo preIT para la purificación manual de ADN a partir de una muestra completa, de la empresa DNA

Genotek. Este protocolo se basa en la precipitación por etanol y el uso del reactivo purificador prepIT-L2P. El proceso seguido es el siguiente:

1) Mezclamos la muestra del vial de recogida por inversión y agitación suave durante unos segundos, a fin de asegurar que las muestras viscosas se mezclen adecuadamente.
2) Transferimos 500µl de muestra a un tubo de centrífuga de 1,5ml.
3) Las muestras se deben incubar en agua a 50°C durante una hora, asegurándonos de que la parte de tubo que contiene la muestra esté sumergida completamente. Este tratamiento térmico es esencial para maximizar el rendimiento del ADN y garantizar que las nucleasas se inactiven permanentemente.
4) Añadimos 20µl de reactivo purificador prepIT-L2P al tubo y lo pasamos por el vórtex unos segundos para asegurar que el conjunto se mezcle adecuadamente. La muestra se enturbiará dado que las impurezas e inhibidores empezarán a precipitar.
5) Incubar la muestra en hielo durante 9 minutos.
6) Centrifugar durante 10 minutos a temperatura ambiente (20°C) a 13000 rpm.
7) Con cuidado de no tocar el pellet de impurezas depositado en el fondo del tubo, pipetearemos el sobrenadante a un nuevo tubo de centrífuga de 1,5ml, y descartamos el pellet. Es recomendable dejar una pequeña cantidad de sobrenadante a fin de no arrastrar parte del pellet de impurezas al intentar apurarlo todo.
8) Añadimos 500µl de etanol al 95% al sobrenadante y mezclamos por inversión suave un mínimo de 10 veces. Durante la mezcla con etanol, el ADN precipitará. El ADN precipitado puede aparecer como un coágulo de fibras de ADN. Incluso si no se ve coágulo, el ADN se recuperará en los siguientes pasos.
9) Incubar 10 minutos a temperatura ambiente (20°C) para permitir que el ADN precipite completamente.
10) Centrifugar durante 10 minutos a temperatura ambiente (20°C) a 13000 rpm.
11) Descartar el sobrenadante de etanol, teniendo cuidado de no tocar el pellet de ADN del fondo. El ADN precipitado también puede aparecer como como una mancha en el costado del tubo más alejado del centro de la centrífuga.
12) Limpiaremos la muestra con etanol. Añadir 250µl de etanol al 70% sin dañar el pellet o la mancha de ADN. Esperar un minuto a temperatura ambiente (20°C) y retirar el etanol apurando lo máximo posible, pero teniendo cuidado de no coger el pellet o mancha de ADN.
13) Rehidratamos la muestra añadiendo 100µl de agua. Pasar por vórtex de forma vigorosa para resuspender la muestra.
14) Para asegurar la completa rehidratación de la muestra de ADN, incubaremos a temperatura ambiente 1 ó 2 días. La rehidratación incompleta del ADN es una causa de imprecisión en la estimación de la concentración de ADN y del fallo potencial de aplicaciones posteriores como la PCR.
15) Almacenaje a -20°C por tiempo indefinido.

Análisis genéticos

Los ADNs se procesaron en los Servicios Generales de la Facultad de Ciencia y Tecnología para el análisis de un grupo de 16 SNPs mediante la técnica RT-PCR Biomark de Fluidigm.

SNPs analizados

Los SNPs incluidos en este trabajo se muestran en la Tabla M.3, con sus principales características.

SNP	Posición	Alelo de referencia
rs3823324	29912766	A
rs79244404	29913483	C
rs2770	31321807	C
rs9273352	32625928	G
rs369150	32975720	A
rs9277332	33028638	A
rs200789833	33028653	G
rs72873921	33028685	C
rs72873922	33028686	A
rs2071350	33043526	C
rs9277413	33051720	A
rs9277418	33051749	C
rs116818505	33051900	G
rs9277498	33054141	C
rs72500564	33055197	C
rs9374640	117207488	A

Tabla M.3: SNPs incluidos en el análisis. Se especifica su posición en el cromosoma 6 y el alelo de referencia.

Hubo además una serie de SNPs que no pudieron ser analizados por diferentes problemas de la plataforma tecnológica. Se muestran en la Tabla M-4.

SNP	Posición
rs9260122	29910419
rs3132687	29911356
rs2770	31321807
rs3179865	31324194
rs1050556	31324586
rs9273352	32625928
rs200160235	32628070
rs28746821	32633348
rs558793154	32899695
rs576193745	32965189
rs375256	32975869
rs9277498	33054141
rs9277522	33054353
rs9374640	117207488

Tabla M.4: SNPs propuestos inicialmente para su análisis, pero desechados por problemas metodológicos de la plataforma.

Análisis estadísticos

En los siguientes párrafos se enumeran los análisis estadísticos que se realizaron sobre los diferentes tipos de datos.

Test Chi-2

El test chi-cuadrado es una prueba de hipótesis estadística en la que la distribución de muestreo del estadístico de prueba es una distribución de chi-cuadrado cuando la hipótesis nula es verdadera. La prueba de chi-cuadrado a menudo se usa como abreviatura para la prueba de chi-cuadrado de Pearson. La prueba de chi cuadrado se usa para determinar si hay una diferencia significativa entre las frecuencias esperadas y las frecuencias observadas en una o más categorías.

En las aplicaciones estándar de esta prueba, las observaciones se clasifican en clases mutuamente excluyentes, y existe una hipótesis nula que da la probabilidad de que cualquier observación se encuentre en la clase correspondiente. El propósito de la prueba es evaluar la probabilidad de ocurrencia de las observaciones que se hagan, suponiendo que la hipótesis nula sea verdadera.

La hipótesis alternativa o de investigación (H1) es que existe una diferencia en la distribución de respuestas a la variable de resultado entre los grupos de comparación (es decir, que la distribución de respuestas "depende" del grupo). Para probar la hipótesis, medimos la

variable de resultado discreta en cada participante en cada grupo de comparación. Los datos de interés son las frecuencias observadas.

Test exacto de Fisher

El test exacto de Fisher es una prueba de significación estadística utilizada para el análisis de tablas de contingencia. Se utiliza sobre todo en tablas 2x2, para las que fue diseñado originalmente, aunque puede extenderse a tablas de mayor tamaño. Como indica su nombre, en esta prueba el significado de la desviación de la hipótesis nula se puede calcular con exactitud.

Tablas de contingencia

Las tablas de contingencia se emplean generalmente para analizar la asociación entre dos variables, habitualmente de naturaleza cualitativa.

El test Chi-2 que se utilizará en este tipo de tablas contrastará la hipótesis nula de que los valores de la tabla se distribuyen al azar.

Análisis de componentes principales

El análisis de componentes principales es una técnica de análisis multivariante utilizada para reducir la dimensionalidad de un conjunto de datos, describiéndolo mediante unas nuevas variables o componentes ortogonales, es decir, no correlacionadas. Los componente o vectores propios se ordenan por la cantidad de varianza que describen. La eficacia del método radica en su capacidad para concentrar la mayor parte de la cantidad de información disponible en la matriz original de datos en las primeras componentes. De este modo, mediante la representación e interpretación de los primeros eigenvectores, se obtendrá la mayor parte de la información, con una pérdida residual.

El método parte de una matriz de correlaciones (o covarianzas) entre las variables introducidas, por lo que es recomendable que sigan una distribución normal.

RESULTADOS

RESULTADOS

En la presente tesis doctoral se han generado un total de 4,41 GB de datos de resultados divididos en 11.867 archivos repartidos en 3.862 carpetas. Solo los análisis básicos de los genes HLA obtenidos con el programa FstMap ocupan 2,08 GB divididos entre 11.433 archivos repartidos en 3.838 carpetas. Dada esta enorme cantidad de información, se hace necesario priorizar el análisis de resultados en función de los objetivos de la tesis doctoral, e imposible incluir todos los datos y tablas generados en anexos. En este capítulo se va a dividir la presentación de los resultados en 5 grandes apartados: Desequilibrio de Hardy-Weinberg, Varianza de Wahlund, Marcadores de ancestralidad, Procesos de selección y Relación de los genes HLA de humanos actuales con los genes HLA de Neandertal, con un análisis previo de control de calidad de los datos.

En lo referente a las presiones evolutivas observadas sobre los genes HLA, vamos a utilizar datos de desequilibrio Hardy-Weinberg y de varianza de Wahlund, con el fin de valorar posibles efectos de la selección y su interacción con el equilibrio flujo génico – deriva. Para ello, analizaremos las grandes diferencias entre continentes y entre genes HLA, así como las diferencias intracontinentales, analizando las variaciones entre los distintos genes y las poblaciones. Finalmente, realizaremos un estudio de la relación de los genes HLA de humanos actuales con los genes de Neandertal, tanto a nivel de conjunto, como pormenorizando gen por gen, mediante árboles evolutivos que nos permitan dilucidar la posición evolutiva del gen HLA neandertal respecto a diversas poblaciones humanas actuales concretas.

Análisis de control de los datos

Al considerar una gran cantidad de datos, su volumen dificulta el análisis pormenorizado, por lo que es difícil detectar factores que puedan afectar a su calidad. Por ello es importante hacer un estudio previo de los datos que se van a usar, a fin de comprobar que no haya anomalías que desvirtúen los análisis y hagan que los resultados sean erróneos o equívocos.

Con este fin, se ha realizado una búsqueda de artefactos que pudieran introducir ruido en la muestra, como variaciones en el número de copias de un fragmento de ADN (CNV), pequeñas inserciones-delecciones (indels) cerca de SNPs que pudieran afectar a su presencia (Tabla R-1).

Para el caso de los CNV se revisó si cada uno de los genes se incluía dentro de algún CNV descrito en la base de 1000Genomas, y en caso afirmativo se descartaron únicamente los SNPs afectados por el mismo, esto es, un gen podía tener parte de sus SNPs afectados por la presencia del CNV y parte de los SNPs no afectados, pasando estos últimos el filtro. En el caso de los indels, se estableció un rango de seguridad *upstream* de 30 bases en el que la presencia de una inserción o deleción podría afectar al patrón de herencia de cada SNP y por tanto a los resultados de los análisis.

	HLA-A	HLA-B	HLA-C	HLA-DRA	HLA-DMA	HLA-DMB	HLA-DOA	HLA-DPA1	HLA-DPB1	HLA-DQA1	HLA-DQB1	MEDIA
% CNV	0	43,13	100	0	0	0	0	0	32,6	100	25	27,3391
% indel	54,55	43,13	50,88	50	0	100	25	25	28,3	24,4	32,5	39,4327
% polialélicos	31,82	41,18	31,58	50	0	100	25	8,33	8,7	4,9	0	27,41
% SNPs no afectados	18,18	7,84	0	0	100	0	75	66,67	34,78	0	50	32,0427
Wahlund min.	0,0042	0	0	0	0	0,0043	0	0	0	0,0043	0	0,0012
Wahlund media	0,0306	0,0364	0,0271	0,0542	0,0158	0,0152	0,0233	0,0622	0,0673	0,0423	0,0479	0,0384
Wahlund max.	0,1460	0,1918	0,1309	0,8504	0,1276	0,1215	0,1778	0,3713	0,3713	0,1201	0,1912	0,2545
Wahlund rango	0,1418	0,1918	0,1309	0,8504	0,1276	0,1172	0,1778	0,3713	0,3713	0,1158	0,1912	0,2534

Tabla R-1: Resumen de los análisis de bondad de la muestra para cada uno de los 11 genes HLA estudiados. Se representa el porcentaje de SNPs de cada gen para cada tipo de variación estructural. Además se presentan datos de los valores de la varianza de Wahlund para facilitar una visión de conjunto.

Lo primero que llama la atención es que no son pocos los SNPs que están afectados por más de un evento estructural: si sumamos los porcentajes de SNPs afectados por CNVs, indels, cualidad polialélica y los no afectados, en muchos genes nos da más de un 100%. También llama la atención la baja proporción de SNPs que no están afectados por un gran indel en forma de CNV, un pequeño indel *upstream* o una caracterización errónea en las distintas bases de datos: únicamente un 32,04% de los SNPs serían válidos para realizar análisis de selección. Cabe destacar que en los casos de HLA-C, HLA-DRA, HLA-DMB y HLA-DQA1 no hay ningún SNP que no esté afectado por alguno de los artefactos tenidos en cuenta. En conjunto, los pequeños indels son los que más SNPs involucran, seguidos de la cualidad polialélica y la presencia de CNVs. Cabe señalar que HLA-DMA es el único gen en el que la totalidad de los SNPs no están afectados por ningún artefacto.

Resulta llamativo que tanto HLA-C como HLA-DQA1 estén afectados al 100% por un CNV. En el caso de HLA-C, esv3608531 (en la posición 31131451) es el nombre que recibe el CNV que lo afecta (Figura R-1). Este CNV ocupa una región de 140857 pares de bases, se han identificado 346 variantes de este CNV y afecta a 10 genes: HLA-C, TCF19, HCG27, LINC02571, WASF5P, PSORS1C3, POU5F1, LOC112267902, RPL3P2 y USP8P1.

HLA-C pertenece al grupo de parálogos de cadenas pesadas de la clase I de los genes HLA del CMH. Es un heterodímero que consiste en una cadena pesada y una cadena ligera (microglobulina beta-2). La cadena pesada está anclada en la membrana. Las moléculas de clase I

juegan un papel clave en el sistema inmunitario. Se expresan en casi todas las células. La cadena pesada es de aproximadamente 45 kDa y su gen contiene 8 exones. El exón 1 codifica el péptido líder, los exones 2 y 3 codifican el dominio alfa1 y alfa2, que se unen al péptido, el exón 4 codifica el dominio alfa3, el exón 5 codifica la región transmembrana y los exones 6 y 7 codifican la cola citoplasmática. Los polimorfismos dentro del exón 2 y el exón 3 son responsables de la especificidad de unión al péptido de cada molécula de clase I.

TCF19 codifica una proteína que contiene un dominio de "dedo de zinc" de tipo PHD y probablemente funciona como un factor de transcripción. La proteína codificada desempeña un papel de proliferación y apoptosis de las células beta pancreáticas. Tiene una expresión ubicua en los ganglios linfáticos, el apéndice y otros 25 tejidos.

HCG27 es una secuencia de ARN no codificante que forma parte del grupo de genes del complejo de histocompatibilidad relacionado con los genes HLA. Se ha observado una amplia expresión en bazo, apéndice y otros 25 tejidos.

LINC02571 es un ARN no codificante intergénico largo (en inglés *Long Intergenic Non-Coding RNA*). Los LINC son ARN no codificantes transcritos de forma autónoma de más de 200 nucleótidos que no se superponen a los genes codificadores anotados. Comparten características con las otras transcripciones de la LNC-RNA y constituyen más de la mitad de las transcripciones de LNC-RNA en humanos. Los LINC tienen diversas características que los distinguen de los genes que codifican ARNm y ejercen funciones como la remodelación de la arquitectura de la cromatina, la estabilización del ARN y la regulación de la transcripción. Los LINC pueden servir, de forma amplia, para ajustar la expresión de genes vecinos con especificidad de tejido a través de una diversidad de mecanismos.

WASF5P es un pseudogen perteneciente a la familia de genes que codifican las proteínas del síndrome de Wiskott-Aldrich, que participan en la transmisión de señales al citoesqueleto de actina. El síndrome de Wiskott-Aldrich es una rara enfermedad recesiva del sistema inmune ligada al cromosoma X caracterizada por eccema, trombocitopenia (recuento bajo de plaquetas), inmunodeficiencia y diarrea con sangre (derivada de la trombocitopenia). Este pseudogen, que aparentemente no se transcribe, se asemeja al gen que codifica el miembro 3 de la familia de proteínas WAS, que se encuentra en el cromosoma 13.

PSORS1C3 (Psoriasis Susceptibility 1 Candidate 3) es un gen de ARN y está afiliado a la clase de ARN no codificante. Como su nombre indica está relacionado con la psoriasis, enfermedad inflamatoria crónica de la piel de origen autoinmune, que produce lesiones escamosas engrosadas e inflamadas, con una amplia variabilidad clínica y evolutiva. Este gen presente una amplia expresión en duodeno, vesícula biliar y otros 16 tejidos.

POU5F1 codifica un factor de transcripción que contiene un homeodominio POU que desempeña un papel clave en el desarrollo embrionario y la pluripotencia de células madre. La expresión aberrante de este gen en tejidos adultos se asocia con tumorigénesis. Este gen puede participar en una translocación con el gen del sarcoma de Ewing en el cromosoma 21, lo que también conduce a la formación de tumores. Uno de los codones de inicio AUG es polimórfico en poblaciones humanas. Se han identificado pseudogenes relacionados en los cromosomas 1, 3, 8, 10 y 12. Este gen presenta una amplia expresión en el pulmón, el intestino delgado y otros 21 tejidos.

LOC112267902 es un fragmento de ARN no codificante que actualmente está sin caracterizar.

RPL3P2 (*Ribosomal Protein L3 Pseudogene 2*) es un pseudogen que se relaciona con el gen RPL3. El gen RPL3 codifica una proteína ribosómica que es un componente de la subunidad 60S. La proteína pertenece a la familia L3P de proteínas ribosómicas y se encuentra en el citoplasma. Puede unirse al ARNm de TAR del VIH-1, y se ha sugerido que la proteína contribuye a la transactivación mediada por la proteína TAT. Este gen se cotranscribe con varios genes de ARN nucleolar pequeños, que se encuentran en varios de los intrones de este gen. Se han caracterizado variantes de *splicing* transcripcional alternativas que codifican diferentes isoformas. Como es típico para los genes que codifican proteínas ribosómicas, existen múltiples pseudogenes de este gen dispersos a través del genoma, como es el caso de RPL3P2.

USP8P1 es un pseudogen que se relaciona con el gen de la peptidasa específica de ubiquitina 8 (USP8). La proteína USP8 pertenece al grupo de peptidasas deubiquitinizantes, que son un gran grupo de proteasas que escinden la ubiquitina de las proteínas y otras moléculas. La ubiquitina se une a las proteínas para regular la degradación de las proteínas a través del proteosoma y el lisosoma; coordinar la localización celular de proteínas; activar e inactivar proteínas y modular las interacciones proteína-proteína.

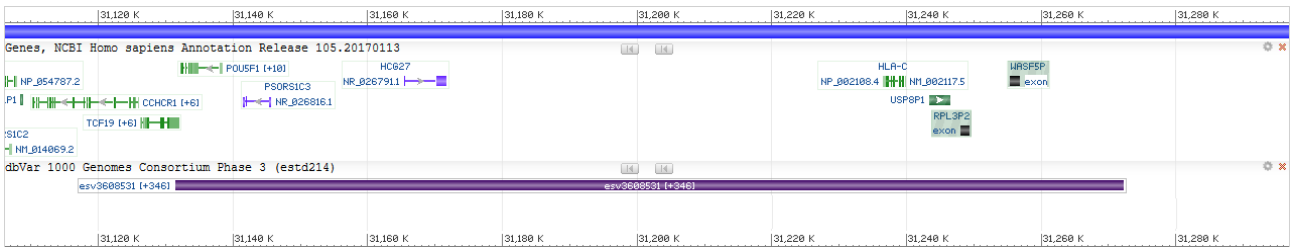


Figura R-1: Posición del CNV esv3608531 (en morado) y genes a los que afecta (en verde).

El caso del CNV esv3608602 (en la posición 32604936) es algo más sencillo (Figura R-2). Ocupa 11197 pares de bases, se han identificado 3480 variantes y afecta únicamente a dos genes: HLA-DQA1 y LOC107986589.

HLA-DQA1 pertenece al grupo de genes MHC clase II. La proteína producida a partir del gen HLA-DQA1 se une a la proteína producida a partir de otro gen MHC de clase II, HLA-DQB1. Juntos, forman un complejo proteico funcional llamado heterodímero DQαβ de unión a antígeno. Este complejo muestra péptidos extraños al sistema inmune para activar la respuesta inmune del cuerpo.

LOC107986589 es un fragmento de ARN no codificante que actualmente está sin caracterizar.

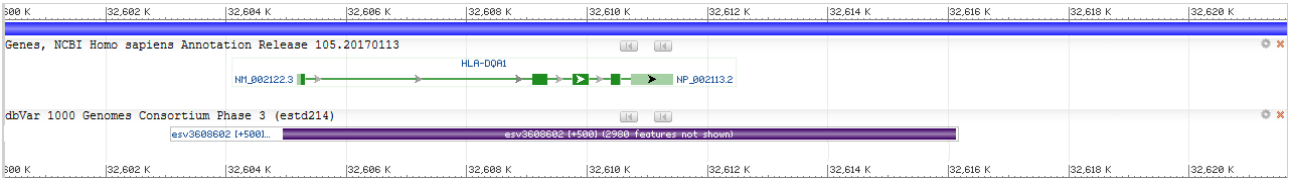


Figura R-2: Posición del CNV esv3608602 (en morado) y gen al que afecta (en verde). No se muestra LOC107986589.

Dada la importancia de muchos de los genes afectados por estos dos CNVs se hace necesario analizar en profundidad las frecuencias y posibles implicaciones en las distintas poblaciones estudiadas (Tabla R-2).

En primer lugar, llaman la atención las diferencias que se observan entre ambos CNVs. Por ejemplo, en el caso de esv3608531, únicamente en la población afrocaribeña de Barbados (ACB) hay duplicación del genoma, si bien la frecuencia alélica apenas supera el 1,6%. Sin embargo, en el caso de esv3608602, todas las poblaciones presentan alelos que caracterizan una duplicación de esa región genómica, con frecuencias que van del 13,5% de los peruanos de Lima (PEL) al 41,9% de los chinos Dai de Xishuangbanna (CDX), con una frecuencia alélica media del 31,9%.

Además, en el caso de esv3608531, las poblaciones LWK, ESN, ASW, TSI, CEU, GBR, FIN, CDX, CHS y CLM son monomórficas para el alelo normal. En el resto de poblaciones el alelo que indica una deleción (CN0) está presente en frecuencia variable: desde el 0,4% de la población de la División Oeste de Gambia (GWD) hasta el 40,2% de la población Telugu (ITU), con una media de frecuencia alélica del 6,7%. En el caso de esv3608602 el alelo que indica una deleción (CN0) tiene una frecuencia alélica media del 38% entre las poblaciones, yendo desde un mínimo de 28,8% en los Telugu (ITU) hasta un máximo del 49,2% en la población de ancestría mejicana de Los Ángeles (MXL).

CNV	VAR.	POBLACIONES																								
		LWK	YRI	ESN	GWD	MSL	ACB	ASW	IBS	TSI	CEU	GBR	FIN	PJL	BEB	GIH	ITU	STU	CHB	CDX	CHS	KHV	JPT	CLM	MXL	PEL	PUR
esv3608531	CN0	0	0,005	0	0,004	0,012	0,094	0	0,009	0	0	0	0	0,370	0,006	0,005	0,402	0,373	0,005	0	0	0,157	0,014	0	0,016	0,124	0,149
	C	1	0,995	1	0,996	0,988	0,891	1	0,991	1	1	1	1	0,630	0,994	0,995	0,598	0,627	0,995	1	1	0,843	0,986	1	0,984	0,876	0,851
	CN2	0	0	0	0	0	0,016	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
esv3608602	CN0	0,313	0,398	0,298	0,385	0,418	0,297	0,377	0,402	0,430	0,404	0,423	0,369	0,354	0,297	0,393	0,289	0,319	0,379	0,306	0,390	0,465	0,418	0,367	0,492	0,488	0,413
	A	0,288	0,218	0,374	0,336	0,341	0,328	0,246	0,313	0,252	0,283	0,258	0,278	0,297	0,366	0,233	0,402	0,368	0,330	0,274	0,286	0,197	0,202	0,335	0,289	0,377	0,351
	CN2	0,399	0,384	0,328	0,279	0,241	0,375	0,377	0,285	0,318	0,313	0,319	0,354	0,349	0,337	0,374	0,309	0,314	0,291	0,419	0,324	0,338	0,380	0,298	0,219	0,135	0,236

Tabla R-2: Frecuencias alélicas de cada variante de los CNVs estudiados en cada una de las poblaciones. La columna "VAR." indica la variante del CNV: CN0 indica una gran deleción comparado con el genoma normal, CN2 es una gran duplicación del genoma, y C y A indican la base del alelo en el genoma normal en esa posición respectivamente, para cada uno de los CNVs.

Desequilibrio de Hardy-Weinberg

Existen diferentes métodos diseñados con el objetivo de valorar la posible acción de selección. A menudo, su uso depende del tipo de datos disponible. Así, la pérdida de diversidad genética en un grupo de SNPs es muy útil en especies domésticas, cuando se dispone de la referencia de una población originaria. Puesto que no es el caso, hemos optado por analizar de forma combinada la existencia de desequilibrio para el test de Hardy-Weinberg y la presencia de un alto valor de la varianza de Wahlund. Particularmente, un alto nivel de varianza se considera muy buen indicio, si bien no es una condición suficiente.

Presumiblemente, la región estudiada es objeto de intensas presiones evolutivas, puesto que es responsable de una parte de la respuesta inmune y en consecuencia es una región genómica hipervariable. Se han considerado 11 genes de las regiones HLA clase I y HLA clase II: HLA-A, HLA-B, HLA-C, HLA-DRA, HLA-DMA, HLA-DMB, HLA-DOA, HLA-DPA1, HLA-DPB1, HLA-DQA1 y HLA-DQB1. No se ha considerado HLA-DRB1 debido a que los datos de este gen no se encuentran disponibles en la base de datos de 1000 Genomas y no fue posible recabar información acerca de la razón para ello. En total se han hallado 80647 puntos de presión evolutiva potencial (entendidos como marcadores en desequilibrio Hardy-Weinberg) distribuidos a lo largo de los 11 genes y 26 poblaciones estudiadas. Se han hallado diferencias significativas en las distribuciones alélicas de multitud de marcadores, tanto entre poblaciones de un mismo grupo continental como entre grandes grupos continentales. Igualmente, estas diferencias son independientes para cada gen HLA estudiado. Es decir, que la variación observada en cada gen no sigue una relación lineal con su tamaño, siendo posible que un gen con más marcadores presente, en cuanto a porcentaje, una menor variación en algunas poblaciones o grupos continentales. Asimismo, se han hallado distribuciones alélicas coincidentes con las explicaciones aceptadas actualmente sobre el proceso de expansión de *Homo sapiens* por los diferentes continentes.

Diferencias entre genes HLA

Se observan diferencias sustanciales entre los distintos genes (Tabla R-3) en número de SNPs, lo cual no es extraño teniendo en cuenta las diferencias de tamaño entre ellos. De igual manera se han hallado diferencias en la existencia de desequilibrios Hardy-Weinberg en cada gen para el conjunto de poblaciones. De manera análoga que con el número de SNPs para cada gen (a mayor tamaño del gen, mayor número de SNPs), podemos observar una relación entre el número de SNPs del gen y el total de SNPs en desequilibrio Hardy-Weinberg para el conjunto de poblaciones. Sin embargo, la relación observada no es tan lineal como en el caso anterior.

GENES	SNP GEN	SNP HW	HW/GEN	SUMA GEN	MEDIA GEN	MEDIANA GEN
HLA-A	475	173	0,36	479	18,88	13
HLA-B	4702	2431	0,52	18248	701,85	684
HLA-C	3139	1026	0,33	2547	97,96	76,5
HLA-DRA	1277	303	0,24	573	22,04	12,5
HLA-DMA	731	61	0,08	84	3,23	2,5
HLA-DMB	1046	172	0,16	241	9,27	2,5
HLA-DOA	894	113	0,13	169	6,5	5
HLA-DPA1	2885	963	0,33	2073	79,73	42,5
HLA-DPB1	782	367	0,47	958	36,85	24,5
HLA-DQA1	917	799	0,87	11130	428,08	427,5
HLA-DQB1	4388	3264	0,74	44133	1697,42	1656

Tabla R-3: Resumen de las diferencias entre los genes HLA estudiados. La columna "SNP GEN" hace referencia al número de polimorfismos de un solo nucleótido (SNP) que se han encontrado para cada gen. "SNP HW" representa el número de SNPs que se encuentran en desequilibrio Hardy-Weinberg en al menos una de las poblaciones analizadas. La columna "HW/GEN" es el cociente entre las columnas "SNP HW" y "SNP GEN". "SUMA GEN" representa el total de SNPs en desequilibrio Hardy-Weinberg para ese gen teniendo en cuenta todas las poblaciones estudiadas. "MEDIA GEN" hace referencia al valor de la media para la variable "número de SNPs en desequilibrio Hardy-Weinberg" entre el conjunto de poblaciones para cada gen. La columna "MEDIANA GEN" representa el valor en la posición central de la variable "número de SNPs en desequilibrio Hardy-Weinberg" en el conjunto de poblaciones para cada gen.

El gen HLA-DQB1 presenta un total de SNPs en desequilibrio para el conjunto poblacional casi 2,5 veces mayor que el gen HLA-B aun cuando este último gen tiene 314 SNPs más. Otro caso curioso es el del binomio formado por HLA-A y HLA-DOA. Mientras que HLA-A tiene casi la mitad de SNPs que HLA-DOA, presenta un total de SNPs en desequilibrio para el conjunto poblacional casi 3 veces mayor que HLA-DOA. Es preciso destacar también el caso de HLA-DMA: si bien presenta un número de SNPs cercano a la mediana entre el conjunto de genes estudiados, su número total de SNPs en desequilibrio es el menor de todos. En cuanto a los valores de las medianas, ocurre algo similar que en el caso del total de SNPs en desequilibrio Hardy-Weinberg para el conjunto de poblaciones. Los genes HLA-A, HLA-DRA, HLA-DMA, HLA-DMB y HLA-DOA presenta valores de SNPs en desequilibrio de ligamiento significativamente más bajos que los del resto de genes. Mención especial para el trinomio HLA-DMA, HLA-DMB y HLA-DOA, que presenta valores de la mediana de SNPs en desequilibrio muy bajos. En cuanto a la proporción de SNPs en desequilibrio frente a SNPs presentes en cada gen, HLA-DQA1 es el que presente el valor más alto (un 87% de sus SNPs están en desequilibrio HW), mientras que HLA-DMA presenta el valor más bajo (únicamente el 8% de sus SNPs están en desequilibrio HW).

Hemos utilizado la prueba Chi-2 con el fin de comparar A) el número de SNPs en desequilibrio en cada gen frente al número de SNPs en equilibrio, B) el número de SNPs total frente a la media de la variable “número de SNPs en desequilibrio Hardy-Weinberg" entre el conjunto de poblaciones para cada gen”; y C) el número de SNPs total frente a la mediana de la variable "número de SNPs en desequilibrio Hardy-Weinberg" en el conjunto de poblaciones para cada gen (Tabla R-4)

	Chi-2	p	p Monte Carlo
A) SNP EQ vs SNP HW	3981,3	0,0000	0,0001
B) SNP GEN vs MEDIA GEN	2829,6	0,0000	0,0001
C) SNP GEN vs MEDIANA GEN	3008,8	0,0000	0,0001

Tabla R-4: Resultados de los test Chi cuadrado. El valor de significación viene dado por el valor de p: si es menor de 0,05 se rechaza la hipótesis nula de homogeneidad de muestras y concluimos que hay diferencias. La prueba de Monte Carlo usa 10,000 permutaciones.

Los resultados muestran que hay diferencias en las 3 pruebas realizadas, esto es, hay diferencias significativas entre el número de SNPs en desequilibrio HW entre los diferentes genes HLA.

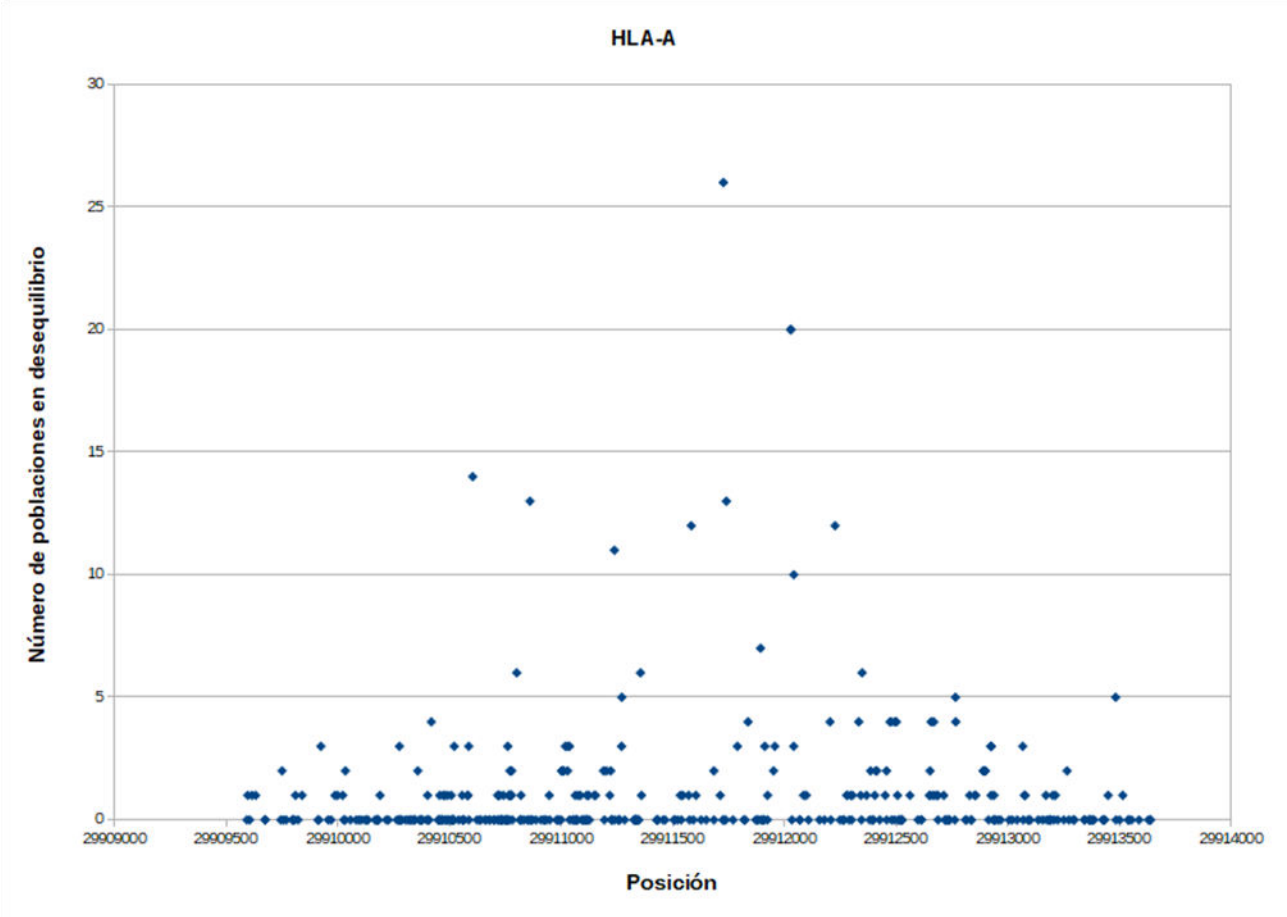


Figura R-3: Gráfico de puntos con la posición de los SNPs presentes en el gen HLA-A, según el número de poblaciones en las que se encuentran en desequilibrio Hardy-Weinberg.

En la Figura R-3 se muestra la distribución de SNPs en el gen HLA-A de acuerdo a su posición y el número de poblaciones en las que se encuentran en desequilibrio. Se observan valores más altos en la zona media del gen. Dicho esto, cabe destacar la baja media (1,008421) del valor "número de poblaciones en desequilibrio". Este hecho, junto con que 302 de los 475 SNPs del gen HLA no estén en desequilibrio en ninguna población, nos permite aventurar que los valores de desequilibrio Hardy-Weinberg son, en general, bajos. De los 475 SNPs representados en la Figura R-1, únicamente 13 SNPs están en desequilibrio en más de 5 poblaciones, únicamente 2 SNPs se encuentran en desequilibrio en más de 15 poblaciones y solamente 1 SNP (rs29028878, en la posición 29911727) está en desequilibrio en todas las poblaciones. Se aprecian también zonas de alta densidad de SNPs que no están presentes en desequilibrio en ninguna población en los extremos del gen, concentradas particularmente entre las posiciones 29910000 y 29911000.

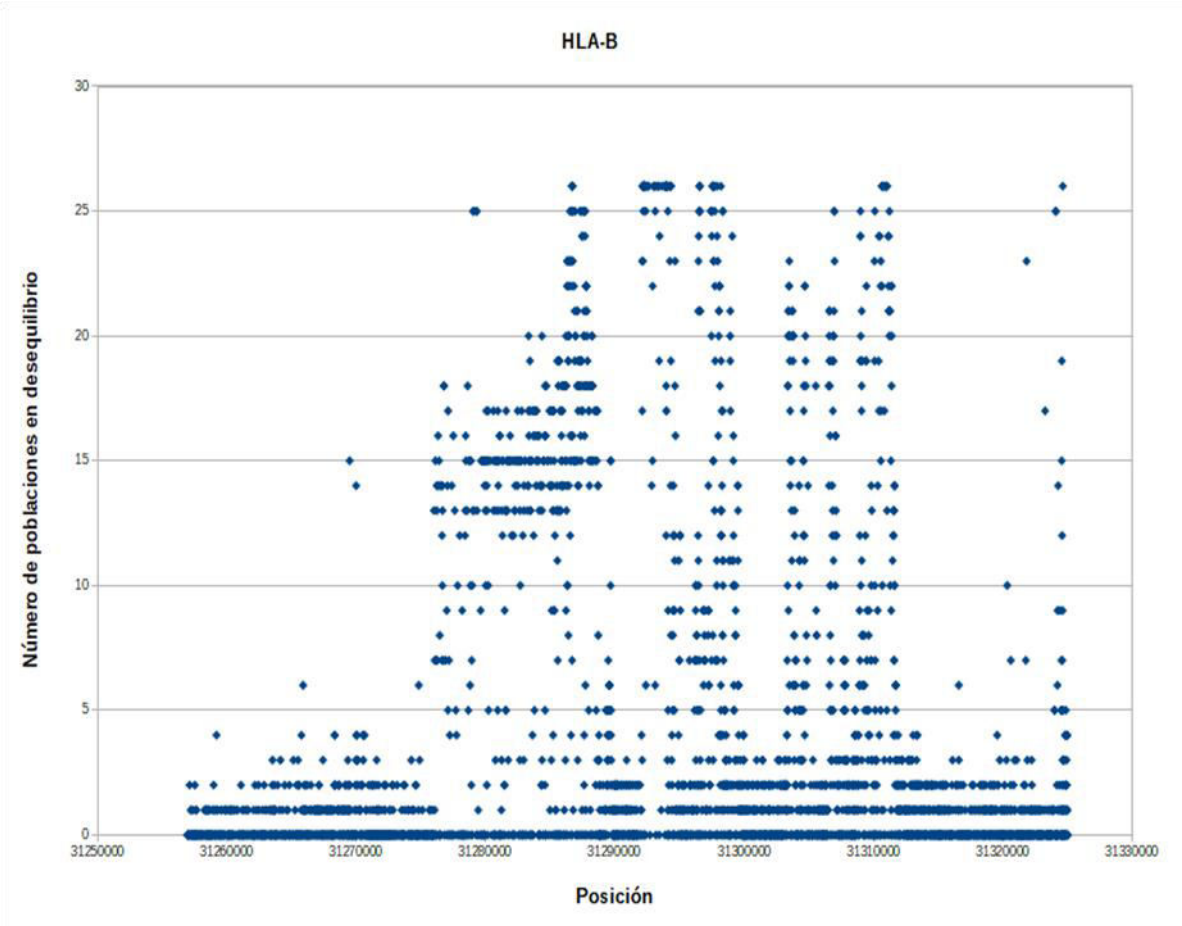


Figura R-4: Gráfico de puntos con la posición de los SNPs presentes en el gen HLA-B, según el número de poblaciones en las que se encuentran en desequilibrio Hardy-Weinberg.

Se muestra en la Figura R-4 la distribución de SNPs en el gen HLA-B según su posición y el número de poblaciones en las que se encuentran en desequilibrio. Se observa una distribución de valores en forma de meseta: valores bajos a ambos extremos, mientras que en la zona central, se observa una zona de alta densidad de marcadores en desequilibrio Hardy-Weinberg en al menos 5 poblaciones. El hecho de que la media del valor "número de poblaciones en desequilibrio" sea

3,881727, junto con que sólo 2271 de los 4702 SNPs del gen no estén en desequilibrio en ninguna población, nos permite aventurar que, en general, el gen HLA-B está bastante afectado por factores de desequilibrio Hardy-Weinberg. De los 4702 SNPs representados en la Figura R-2, 962 SNPs están en desequilibrio en más de 5 poblaciones, 470 SNPs se encuentran en desequilibrio en más de 15 poblaciones y 102 SNPs está en desequilibrio en todas las poblaciones. Es preciso destacar también el hecho que de que en la zona central, aproximadamente entre las posiciones 31296418 y 31299053, hay un pico de alta densidad de valores altos con 52 SNPs en desequilibrio en 20 o más poblaciones.

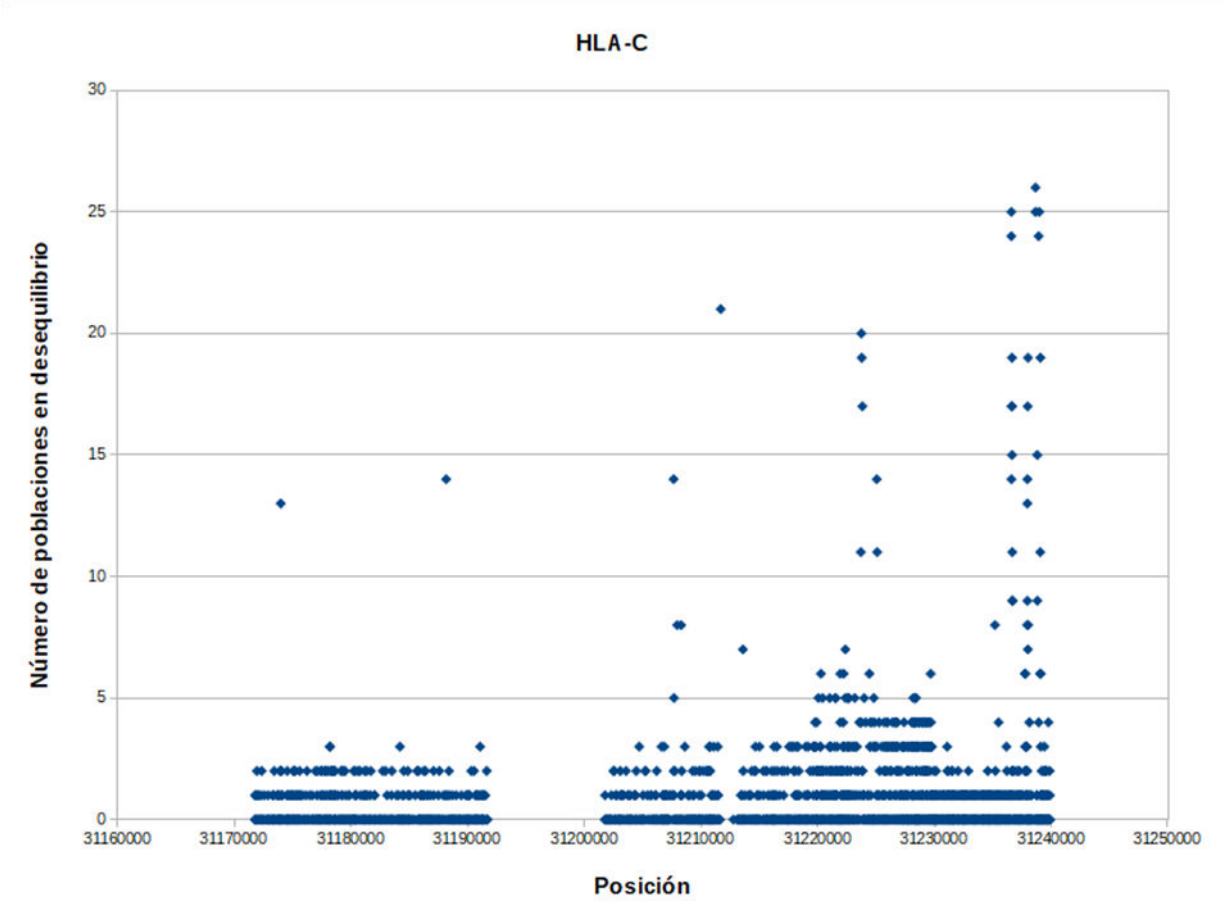


Figura R-5: Gráfico de puntos con la posición de los SNPs presentes en el gen HLA-C, según el número de poblaciones en las que se encuentran en desequilibrio Hardy-Weinberg.

En la Figura R-5 se muestra la distribución de SNPs en el gen HLA-C según el número de poblaciones en las que están presentes en desequilibrio. Se observa que aparecen valores más altos en la zona final del gen, con algunos valores dispersos entre el resto de las posiciones. Hay que destacar la bajísima media (0,811405) del valor "número de poblaciones en desequilibrio". Este hecho, junto con que 2113 de los 3139 SNPs del gen HLA-C no estén en desequilibrio en ninguna población, permite aventurar que los valores de desequilibrio Hardy-Weinberg son muy bajos a lo largo del gen HLA-C. De los 3139 SNPs representados en la Figura R-3, únicamente 67 SNPs están en desequilibrio en más de 5 poblaciones, sólo 21 SNPs se encuentran en desequilibrio en más de 15 poblaciones y solamente 2 SNP (rs9264648 en la posición 31238661, y rs9264649 en

la posición 31238662) está en desequilibrio en todas las poblaciones. Se aprecia una zona entre las posiciones 31191744 y 31201763 en las que no hay ningún SNP, es decir, no se ha observado variación polimórfica.

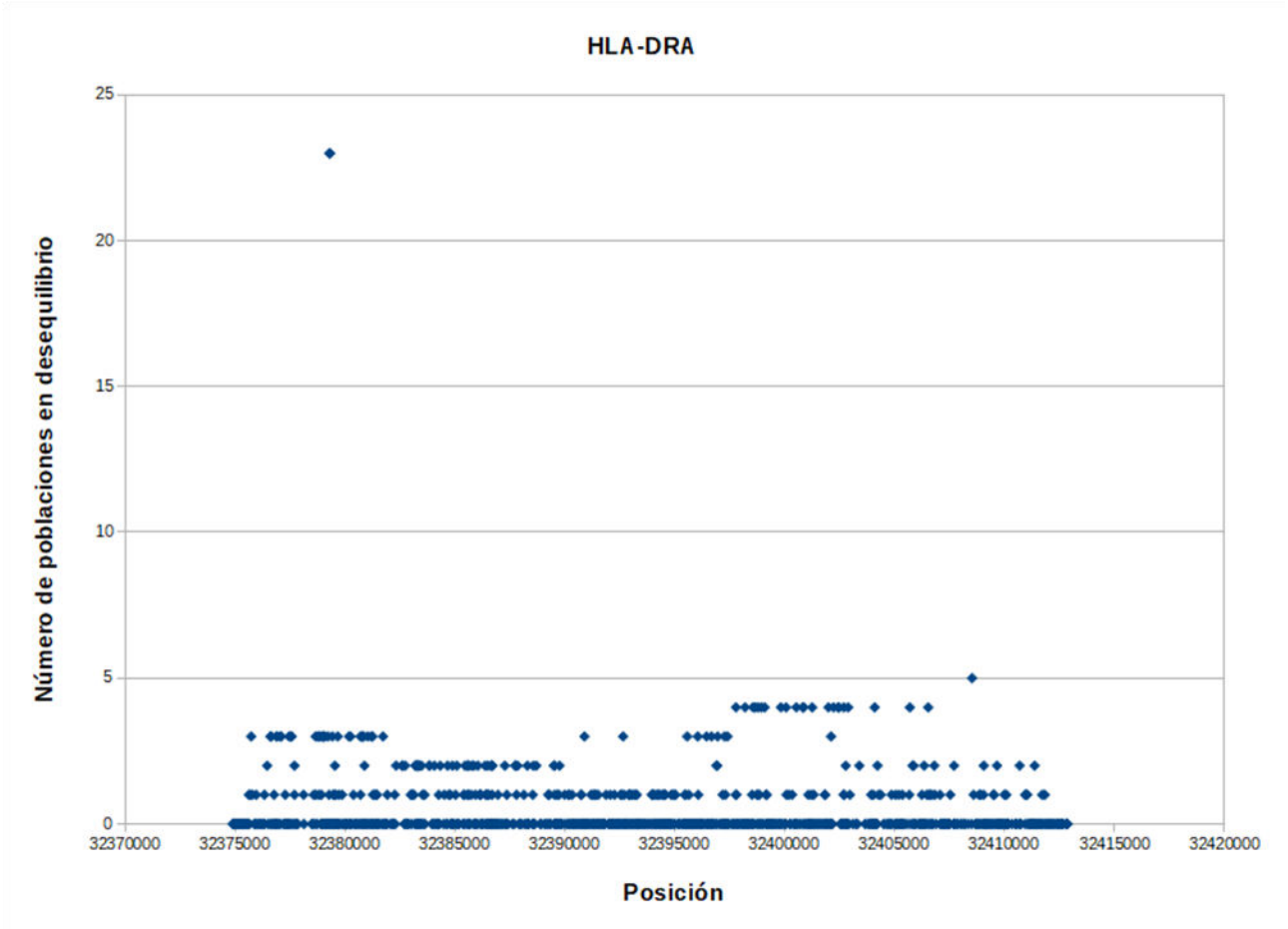


Figura R-6: Gráfico de puntos con la posición de los SNPs presentes en el gen HLA-DRA, según el número de poblaciones en las que se encuentran en desequilibrio Hardy-Weinberg.

En la Figura R-6 se muestra la distribución de SNPs en el gen HLA-DRA con el número de poblaciones en las que se encuentran en desequilibrio. Se observan valores muy bajos de desequilibrio a lo largo de todo el gen. HLA-DRA tiene un valor medio del número de poblaciones en desequilibrio de 0,044871, lo que junto a que 974 de los 1277 SNPs del gen no estén en desequilibrio en ninguna población, refleja unos valores de desequilibrio realmente bajos. De los 1277 SNPs representados en la Figura R-6, solo 3 se encuentran en desequilibrio en 5 o más poblaciones, uno (rs9268645 en la posición 32408527) en 5 poblaciones y 2 SNPs (rs574027387 en la posición 32379313, y rs78348693 en la posición 32379319) en 23 poblaciones. El resto de marcadores que se encuentran en desequilibrio, lo están en 4 poblaciones (23 SNPs), 3 (45 SNPs), 2 (63 SNPs) o 1 población (169 SNPs).

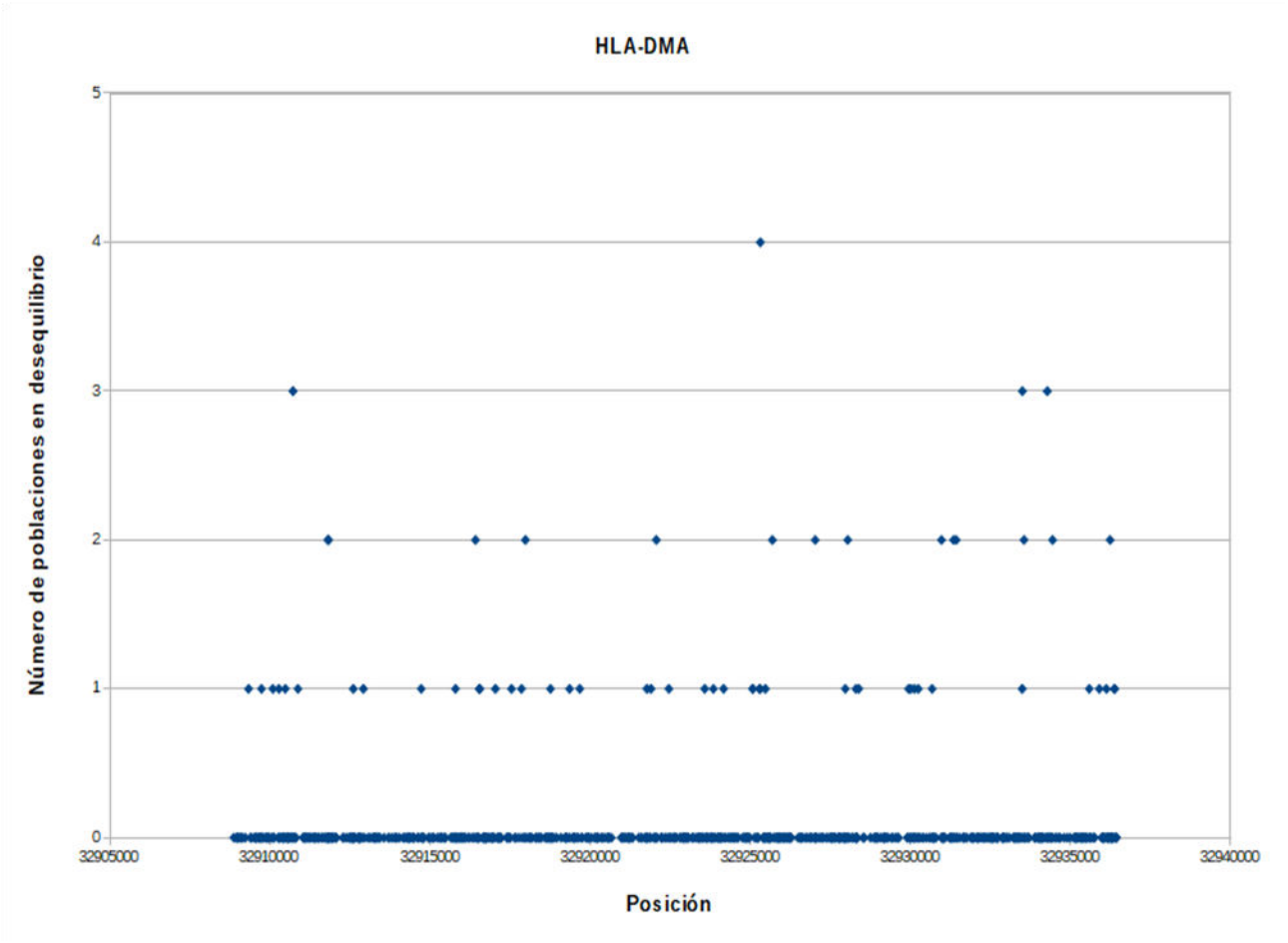


Figura R-7: Gráfico de puntos con la posición de los SNPs presentes en el gen HLA-DMA, según el número de poblaciones en las que se encuentran en desequilibrio Hardy-Weinberg.

Podemos observar en la Figura R-7 que la distribución de SNPs en el gen HLA-DMA según el número de poblaciones en las que se encuentran en desequilibrio muestra valores de 0 en la mayor parte de las posiciones, esto es, HLA-DMA tiene muy pocos marcadores que estén en desequilibrio en alguna población. El valor de la media (0,11491) del valor "número de poblaciones en desequilibrio" si bien no es el más bajo de los observados en el conjunto de genes, sigue siendo un valor extremadamente bajo. De los 731 SNPs del gen HLA-DMA, 670 SNPs no estén en desequilibrio en ninguna población. Del total de SNPs representados en la Figura R-7, 43 SNPs están en desequilibrio en 1 población, 14 SNPs se encuentran en desequilibrio en 2 poblaciones, 3 SNP están en desequilibrio en 3 poblaciones, y únicamente 1 SNP (rs373900326 en la posición 32925315) se encuentra en desequilibrio en 4 poblaciones. No es posible discernir zonas de concentración de valores, más allá de la alta cantidad de SNPs que no se encuentran en desequilibrio en ninguna población.

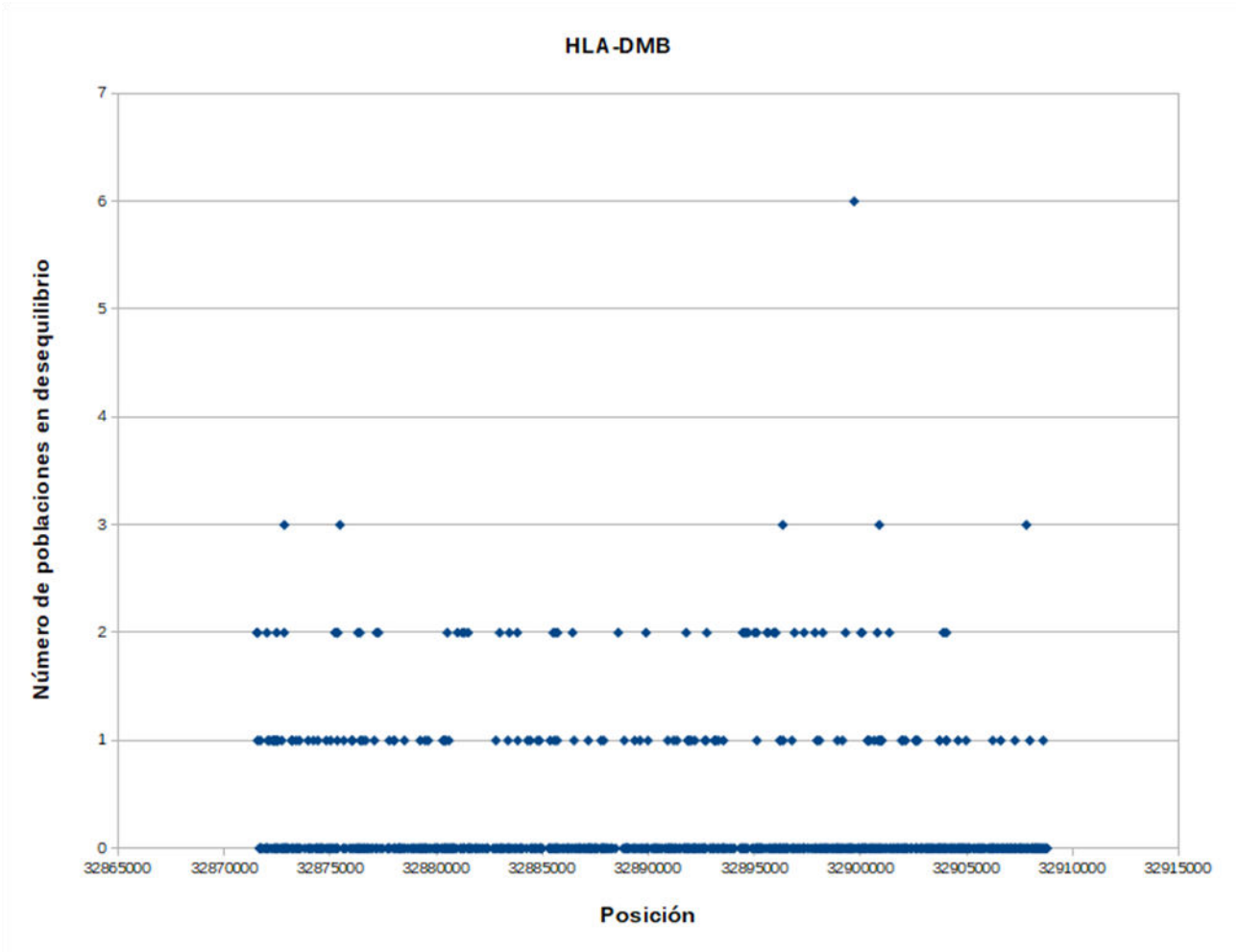


Figura R-8: Gráfico de puntos con la posición de los SNPs presentes en el gen HLA-DMB, según el número de poblaciones en las que se encuentran en desequilibrio Hardy-Weinberg.

En la Figura R-8 se muestra la distribución de SNPs en el gen HLA-DMB de acuerdo al número de poblaciones en las que se encuentran en desequilibrio. Se observa una distribución de valores similar al de HLA-DMA (Figura R-5). Sin embargo, el valor de la media para la variable "número de poblaciones en desequilibrio" es mayor (0,2304). De los 1047 SNPs del gen HLA-DMB no estén en desequilibrio en ninguna población 874 SNPs. Del total de SNPs representados en la Figura R-8, 112 SNPs están en desequilibrio en 1 población, 54 SNPs se encuentran en desequilibrio en 2 poblaciones, 5 SNPs están en desequilibrio en 3 poblaciones, y únicamente 1 SNP (rs558793154 en la posición 32899695) se encuentra en desequilibrio en 6 poblaciones. Si bien se aprecia una mayor concentración de valores para SNPs que presentan desequilibrio en al menos una población, del mismo modo que en HLA-DMA, no es posible discernir zonas de concentración de valores, más allá de la alta cantidad de SNPs que no se encuentran en desequilibrio en ninguna población.

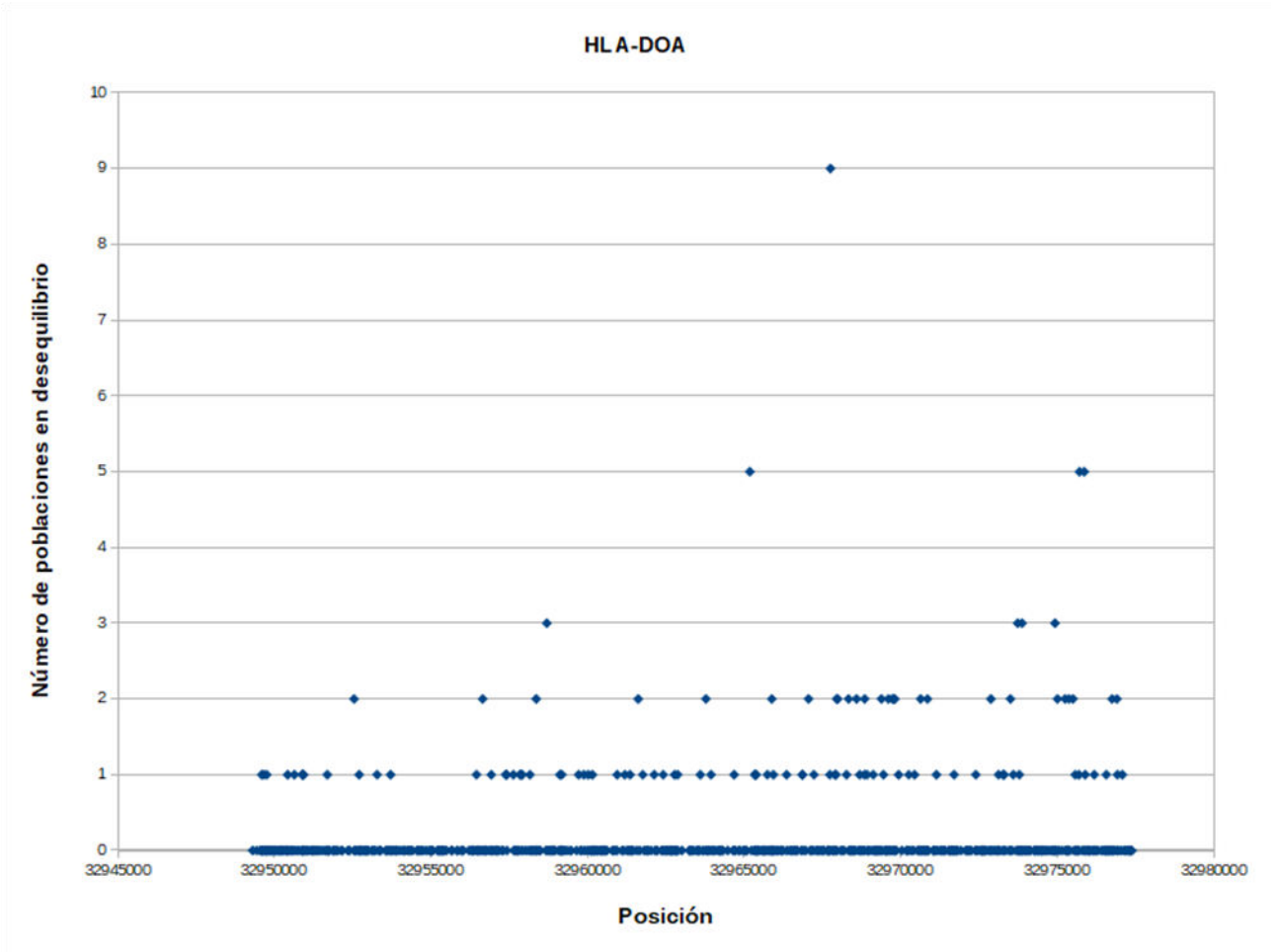


Figura R-9: Gráfico de puntos con la posición de los SNPs presentes en el gen HLA-DOA, según el número de poblaciones en las que se encuentran en desequilibrio Hardy-Weinberg.

Se observa en la Figura R-9 que la distribución de SNPs en el gen HLA-DOA en función del número de poblaciones en las que están en desequilibrio tiene una distribución de valores similar al de HLA-DMA (Figura R-5) y HLA-DMB (Figura R-6), en tanto que los tres genes presentan muchos SNPs que no se encuentran en desequilibrio Hardy-Weinberg, y comparativamente, pocos SNPs que estén en desequilibrio en alguna población. Sin embargo, el valor de la media para la variable "número de poblaciones en desequilibrio" se sitúa entre las de HLA-DMA y HLA-DMB (0,189038). De los 894 SNPs del gen HLA-DOA 781 SNPs no están en desequilibrio en ninguna población. Del total de SNPs representados en la Figura R-9, 77 SNPs están en desequilibrio en 1 población, 28 SNPs se encuentran en desequilibrio en 2 poblaciones, 4 SNPs están en desequilibrio en 3 poblaciones, 3 SNPs se encuentran en desequilibrio en 5 poblaciones, y únicamente 1 SNP (rs67808968 en la posición 32967762) se encuentra en desequilibrio en 9 poblaciones. Del mismo modo que con el caso de los dos genes anteriores, no es posible discernir zonas de concentración de valores, más allá de la alta cantidad de SNPs que no se encuentran en desequilibrio en ninguna población, y de una mayor concentración de SNPs con desequilibrio en la segunda mitad del gen.

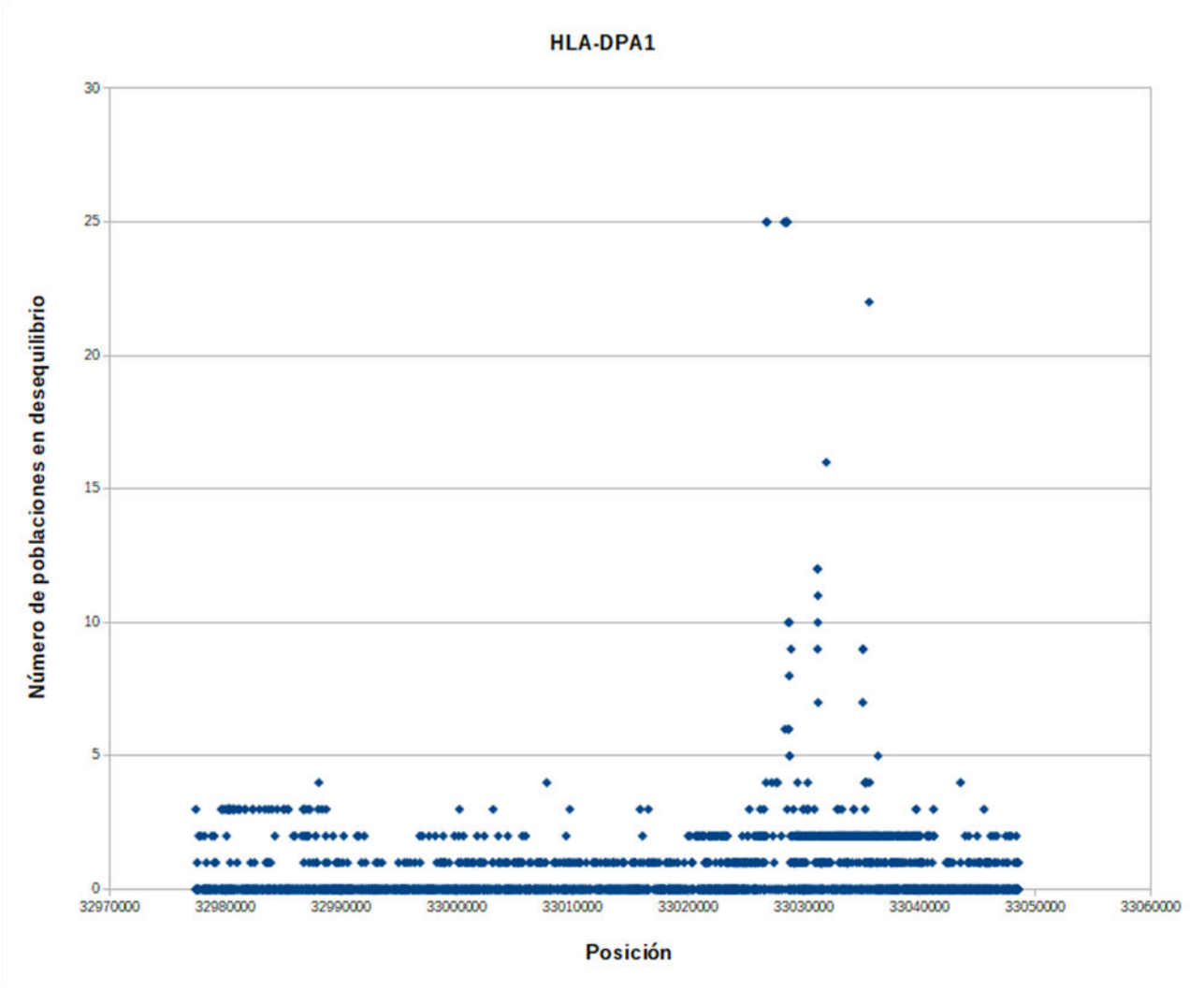


Figura R-10: Gráfico de puntos con la posición de los SNPs presentes en el gen HLA-DPA1, según el número de poblaciones en las que se encuentran en desequilibrio Hardy-Weinberg.

En la Figura R-10 se muestra la distribución de SNPs en el gen HLA-DPA1 según el número de poblaciones en las que están en desequilibrio. Se observan unos valores relativamente bajos a lo largo de todo el gen, excepto en el último tercio. HLA-DPA1 tiene un valor de media del número de poblaciones en desequilibrio de 0,718544, lo que junto con el hecho de que 1922 de los 2885 SNPs del gen no estén en desequilibrio en ninguna población, cataloga los valores de desequilibrio como relativamente bajos. De los 2885 SNPs representados en la Figura R-10, 930 SNPs están en desequilibrio en 1 a 5 poblaciones, 20 SNPs se encuentran en desequilibrio en 6 a 15 poblaciones, y 13 SNPs están en desequilibrio en 16 a 25 poblaciones. Es importante destacar que los 11 SNPs que están en desequilibrio para 25 poblaciones se encuentran en una región relativamente estrecha del gen, entre los loci 33026742 y 33028543, donde hay un total de 36 SNPs y 19 presentan desequilibrio para alguna población.

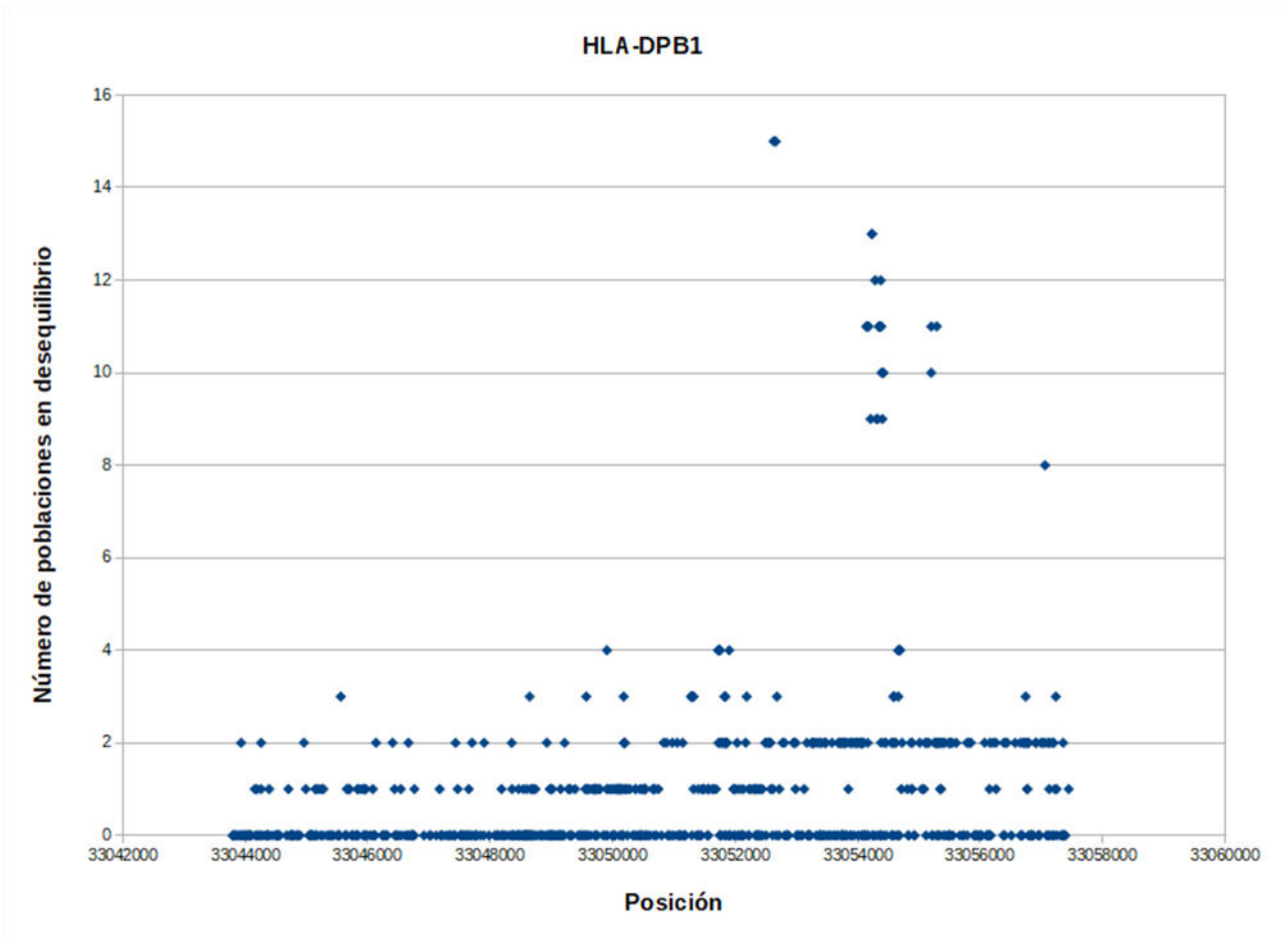


Figura R-11: Gráfico de puntos con la posición de los SNPs presentes en el gen HLA-DPB1, según el número de poblaciones en las que se encuentran en desequilibrio Hardy-Weinberg.

Podemos observar en la Figura R-11 la distribución de SNPs en el gen HLA-DPB1 según el número de poblaciones en las que están en desequilibrio. El gráfico muestra, de forma similar a HLA-DPA1, valores comparativamente más bajos a lo largo de todo el gen que en el último tercio. HLA-DPB1 tiene un valor de media del número de poblaciones en desequilibrio de 1,225064, uno de los valores más altos para esta variable entre los genes HLA. De los 782 SNPs del gen HLA-DPB1, 415 no estén en desequilibrio en ninguna población, 329 SNPs están en desequilibrio en 1 a 5 poblaciones, 13 SNPs se encuentran en desequilibrio en 6 a 10 poblaciones, y 25 SNPs están en desequilibrio en 11 a 15 poblaciones. Es importante destacar que los 2 SNPs que están en desequilibrio para 15 poblaciones (rs549336625 y rs560721463) y los 3 SNPs en desequilibrio en 13 poblaciones (rs9277512, rs9277513 y rs9277514) se encuentran ligados entre sí, dado que los valores de sus frecuencias en cada población son iguales, y se encuentran dispuestos muy próximos y algunos de forma consecutiva en el cromosoma. Así, en un análisis mediante Haploview, el programa detecta un bloque de ligamiento que incluye, entre otros, estos SNPs (Figura R-12).

Figura R-1: Detección mediante Haploview de un bloque de ligamiento (en la parte central del gráfico) en la región 6:33054000-33055000, correspondiente a un fragmento del gen HLA-DPB1.

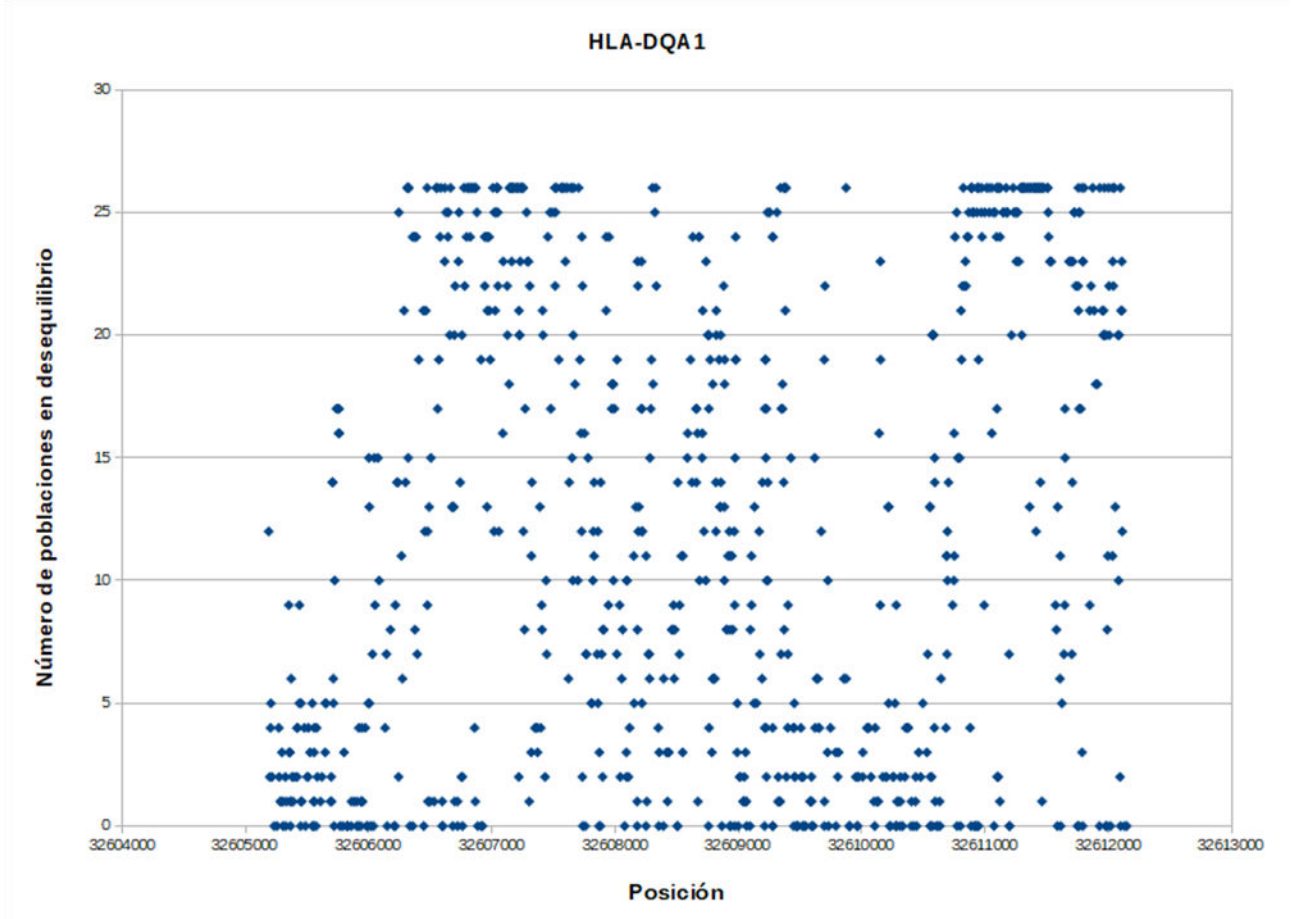


Figura R-2: Gráfico de puntos con la posición de los SNPs presentes en el gen HLA-DQA1, según el número de poblaciones en las que se encuentran en desequilibrio Hardy-Weinberg.

Se muestra en la Figura R-13 la distribución de SNPs en el gen HLA-DQA1 de acuerdo al número de poblaciones en las que se encuentran en desequilibrio. Se observa que presenta una distribución de valores en forma de cuadrado irregular: hay una distribución relativamente homogénea de frecuencias a lo largo del gen. El hecho de que la media del valor "número de poblaciones en desequilibrio" para el gen HLA-DQA1 sea el más alto entre los genes estudiados (12,1374), junto con el que únicamente 118 de los 917 SNPs del gen HLA no estén en desequilibrio en ninguna población, permite aventurar que el gen HLA-DQA1 está bastante afectado por desequilibrios Hardy-Weinberg. De los 917 SNPs representados en la Figura R-13, 570 SNPs están en desequilibrio en más de 5 poblaciones, 365 SNPs se encuentran en desequilibrio en más de 15 poblaciones, y 122 SNPs está en desequilibrio en todas las poblaciones. Pueden destacarse también dos hechos: primero, los marcadores que están en desequilibrio para todas las poblaciones son mayoría, incluso por encima de los marcadores que no están en desequilibrio para ninguna población; y segundo, dichos SNPs están concentrados en las regiones inicial y final del gen, habiendo un cierto vacío en la zona central.

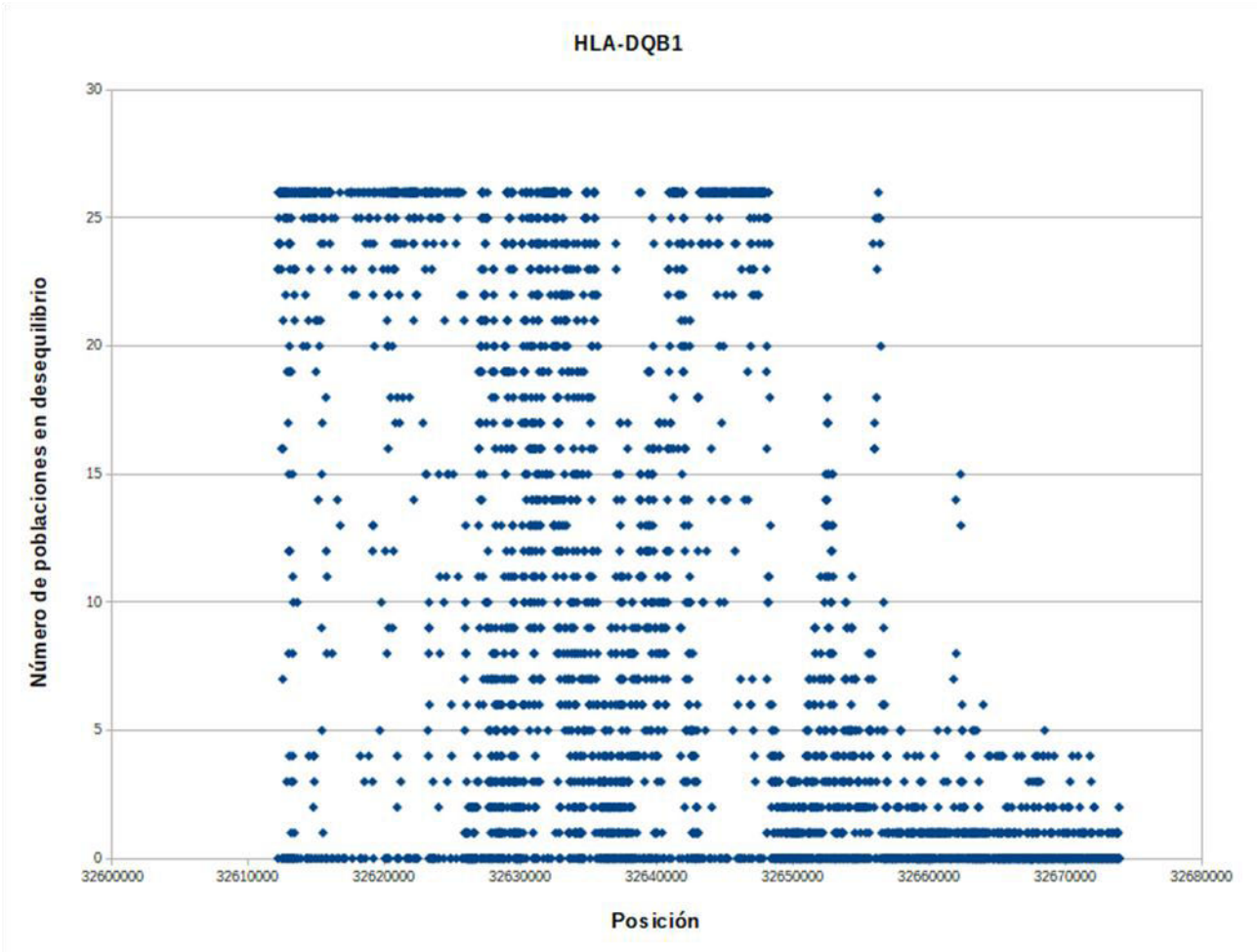


Figura R-3: Gráfico de puntos con la posición de los SNPs presentes en el gen HLA-DQB1, según el número de poblaciones en las que se encuentran en desequilibrio Hardy-Weinberg.

Como se observa en la Figura R-14, la distribución de SNPs en el gen HLA-DQB1 en función del número de poblaciones con desequilibrio tiene una distribución de valores similar a la del gen HLA-DQA1, esto es, en forma de cuadrado irregular. El hecho de que la media del valor "número de poblaciones en desequilibrio" para el gen HLA-DQB1 sea el segundo valor más alto (10,05766), junto con el que solo 1124 de los 4388 SNPs del gen HLA no estén en desequilibrio en ninguna población, permite aventurar que el gen HLA-DQA1 se encuentra notablemente afectado por desequilibrios Hardy-Weinberg. De los 4388 SNPs representados en la Figura R-14, 2119 SNPs están en desequilibrio en más de 5 poblaciones, 1451 SNPs se encuentran en desequilibrio en más de 15 poblaciones, y 804 SNPs está en desequilibrio en todas las poblaciones. Este último valor convierte a HLA-DQB1 en el gen con más SNPs en desequilibrio para todas las poblaciones entre los genes estudiados. Es interesante destacar que estos SNPs se concentran sobre todo en la primera mitad del gen. También cabe destacar la baja concentración de SNPs con un alto número de poblaciones en desequilibrio hacia el final del gen, así como la menor concentración de SNPs con pocas poblaciones en desequilibrio al inicio.

Diferencias entre grupos continentales

Podemos observar diferencias entre los distintos grupos continentales (Tabla R-5). Se han realizado análisis de Chi-cuadrado para cada par de poblaciones a fin de comparar el número de SNPs en desequilibrio entre cada una de ellas (Tabla R-6), así como entre cada par de grupos continentales (Tabla R-7). Si bien Asia es el continente que más marcadores en desequilibrio presenta (31386), seguida de África (19852), Europa (16430) y América (12979), el hecho de que el número de poblaciones sea distinto entre estos grupos continentales desvirtúa el análisis de los datos, por lo que debemos comparar las medias de cada grupo continental. Así, Europa (3286) es el continente que tiene una media mayor en el número de SNPs en desequilibrio en el conjunto de genes, seguido de América (3245), Asia (3139) y por último África (2836). También hemos realizado los cálculos para la mediana a fin de evitar los efectos de los valores muestrales extremos sobra la media. Así, América es el continente con una mediana del valor de desequilibrio Hardy-Weinberg más alta entre sus poblaciones (3403), seguida de Asia (3244), Europa (3235) y África (2832). Si tenemos en cuenta que el principio de equilibrio Hardy-Weinberg establece que la composición genética de una población permanece en equilibrio mientras no actúe la selección natural, entre otros factores, diríase que se observa aproximadamente un aumento de dichos factores a lo largo del proceso de expansión de *Homo sapiens* (*Out of Africa*, colonización de Europa y Asia, y más tarde, de América).

	POBLACIONES																								
	ÁFRICA							EUROPA					ASIA										AMÉRICA			
GENES	LWK	YRI	ESN	GWD	MSL	ACB	ASW	IBS	TSI	CEU	GBR	FIN	PJL	BEB	GIH	ITU	STU	CHB	CDX	CHS	KHV	JPT	CLM	MXL	PEL	PUR
HLA-A	13	51	31	35	21	16	10	24	10	6	10	8	9	33	13	8	12	57	5	12	17	16	20	6	9	39
HLA-B	1034	635	495	638	529	799	457	605	1239	705	1026	791	571	662	749	887	750	774	487	837	673	675	791	431	315	693
HLA-C	35	213	68	255	50	69	87	118	49	141	110	31	75	73	26	212	84	78	34	82	106	303	43	43	111	51
HLA-DRA	5	36	14	8	5	82	9	11	39	36	1	70	4	7	21	3	45	24	8	52	2	19	11	6	15	40
HLA-DMA	1	1	1	5	0	0	0	13	6	0	7	8	3	3	5	6	4	4	0	1	5	1	2	1	6	1
HLA-DMB	25	1	0	2	2	1	4	19	4	3	6	2	1	1	27	11	36	1	2	2	1	4	0	7	66	13
HLA-DOA	10	2	5	3	3	1	10	6	11	5	14	15	5	2	4	3	2	6	8	5	11	0	13	19	4	2
HLA-DPA1	75	40	70	19	101	59	87	64	21	51	34	67	37	23	27	25	40	29	90	45	36	425	472	71	31	34
HLA-DPB1	42	44	29	13	14	26	7	21	15	37	6	2	21	239	27	4	21	23	31	72	39	77	98	3	27	20
HLA-DQA1	448	372	412	553	428	334	306	499	355	453	463	519	426	454	427	546	413	515	314	404	447	324	368	333	524	493
HLA-DQB1	1435	1535	1693	1606	1501	1445	1381	1855	1816	1645	1460	1898	1672	1738	1928	1862	1518	2061	1526	1874	1504	1433	1667	1513	2214	2353
SUMA P.	3123	2930	2818	3137	2654	2832	2358	3235	3565	3082	3137	3411	2824	3235	3254	3567	2925	3572	2505	3386	2841	3277	3485	2433	3322	3739
SUMA C.	19852							16430					31386										12979			
MEDIA C.	2836							3286					3139										3245			
MEDIANA C.	2832							3235					3244										3403			

Tabla R-5: Resumen de las diferencias de variabilidad entre poblaciones para cada uno de los genes estudiados. Cada valor indica el número de SNPs que se encuentran en desequilibrio Hardy-Weinberg para una población en un gen determinado. Los celdas han sido coloreadas con un gradiente de 3 colores para visualizar los valores menores (rojo), los valores medios (amarillo) y los valores más altos (verde) para cada población en cada gen. La fila "SUMA P." indica el total de SNPs en desequilibrio Hardy-Weinberg para una población determinada, con los valores más bajos en rojo y los más altos en blanco. La fila "SUMA C." representa el total de SNPs en desequilibrio para cada grupo continental del conjunto de poblaciones y genes. La fila "MEDIA C." representa el valor medio de la variable "número de SNPs en desequilibrio Hardy-Weinberg" de entre el conjunto de poblaciones para cada grupo continental, mientras que la fila "MEDIANA C." representa el valor en la posición central de la variable "número de SNPs en desequilibrio Hardy-Weinberg" de entre el conjunto de poblaciones para cada grupo continental.

	LWK	YRI	ESN	GWD	MSL	ACB	ASW	IBS	TSI	CEU	GBR	FIN	PJL	BEB	GIH	ITU	STU	CHB	CDX	CHS	KHV	JPT	CLM	MXL	PEL	PUR
LWK		311,91	249,21	366,74	174,91	148,68	214,27	237,27	136,28	191,03	101,55	210,22	197,76	325,29	158,04	245,53	121,60	232,40	185,55	165,23	141,20	593,44	379,67	235,45	641,36	314,74
YRI	2,7E-60		127,92	82,16	181,16	141,25	132,30	127,12	321,73	67,01	237,01	283,73	142,66	238,69	248,37	153,37	136,01	129,07	204,21	124,37	99,66	360,56	512,51	229,80	376,12	217,62
ESN	4,1E-47	4,6E-22		174,99	31,37	157,87	44,43	43,80	310,90	87,77	263,51	137,68	37,23	193,08	113,10	202,96	134,74	52,18	41,70	103,26	79,31	467,66	352,08	57,01	190,06	64,07
GWD	6,9E-72	5,6E-13	1,1E-31		193,77	271,61	162,04	117,52	402,31	116,27	198,61	316,11	129,32	319,16	282,89	67,39	188,05	157,01	273,21	248,56	95,58	506,66	686,24	237,29	344,41	270,24
MSL	1,1E-31	5,9E-33	9,6E-04	1,5E-35		149,75	36,03	58,55	288,50	103,69	203,56	112,89	50,42	237,95	124,30	183,64	131,19	91,35	38,10	135,71	85,75	410,78	274,02	49,12	270,02	110,19
ACB	2,8E-26	9,1E-25	3,7E-28	8,2E-52	4,2E-27		141,18	186,63	112,52	79,72	164,92	108,50	149,80	292,17	152,38	220,83	64,66	136,83	150,21	49,89	133,85	473,04	391,20	199,84	515,86	179,30
ASW	8,1E-40	6,0E-23	6,1E-06	5,1E-29	8,3E-05	2,4E-25		44,85	267,18	70,98	186,32	126,86	55,08	255,96	147,91	139,28	117,59	109,86	44,93	120,49	87,21	319,12	283,96	31,08	225,57	125,60
IBS	1,3E-44	6,7E-22	7,9E-06	5,7E-20	1,7E-08	4,4E-34	5,2E-06		313,94	75,55	199,28	157,47	37,96	247,97	108,58	113,20	89,85	75,97	88,15	136,63	60,39	504,55	471,30	68,62	171,40	98,27
TSI	9,3E-24	2,3E-62	4,4E-60	2,0E-79	2,3E-55	5,7E-19	7,0E-51	1,0E-60		208,89	113,37	176,78	220,41	419,88	154,34	246,78	138,44	202,11	258,51	140,31	200,84	815,96	603,76	282,94	711,79	277,51
CEU	5,4E-35	4,5E-10	4,6E-14	1,0E-19	9,9E-18	5,7E-13	2,9E-11	1,1E-11	1,1E-38		146,13	139,50	60,98	223,91	138,92	97,05	60,67	94,53	106,61	48,15	48,90	414,02	438,86	131,95	312,51	148,88
GBR	8,8E-17	1,5E-44	4,1E-50	1,4E-36	1,4E-37	1,3E-29	5,1E-34	1,0E-36	3,9E-19	9,2E-26		206,04	155,89	363,49	211,13	82,30	125,43	190,76	254,49	186,85	90,59	614,95	548,53	233,19	612,13	323,86
FIN	5,6E-39	2,3E-54	4,8E-24	3,5E-61	4,8E-19	3,7E-18	7,5E-22	4,4E-28	4,7E-32	2,1E-24	4,2E-38		113,88	369,31	101,89	233,82	113,67	121,18	116,25	123,15	162,78	723,01	492,22	94,22	410,45	123,52
PJL	2,2E-36	4,7E-25	1,1E-04	2,4E-22	5,3E-07	1,6E-26	7,5E-08	8,0E-05	4,2E-41	6,1E-09	9,3E-28	3,1E-19		183,75	75,29	97,78	100,05	43,26	57,36	84,14	34,31	525,13	417,56	57,68	199,42	71,20
BEB	4,1E-63	6,5E-45	2,0E-35	8,0E-62	9,3E-45	3,9E-56	1,6E-48	7,4E-47	3,7E-83	7,9E-42	3,4E-71	2,0E-72	1,7E-33		248,59	351,57	266,32	220,69	218,45	169,17	159,62	651,64	500,69	275,48	443,85	286,00
GIH	3,4E-28	6,1E-47	4,4E-19	3,5E-54	2,5E-21	4,9E-27	4,0E-26	3,5E-18	1,9E-27	2,7E-24	3,6E-39	7,6E-17	1,2E-11	5,5E-47		197,56	76,40	77,32	96,31	95,23	125,08	722,65	496,16	109,40	277,18	52,83
ITU	2,4E-46	3,1E-27	1,8E-37	3,8E-10	1,8E-33	3,5E-41	2,3E-24	4,2E-19	1,3E-46	6,9E-16	5,3E-13	6,8E-44	4,9E-16	1,1E-68	2,4E-39		124,47	157,27	237,72	196,93	78,30	591,62	679,70	189,48	391,61	246,63
STU	8,6E-21	1,1E-23	1,9E-23	2,2E-34	1,0E-22	1,2E-09	5,5E-20	1,8E-14	3,4E-24	7,0E-09	1,5E-21	3,4E-19	1,7E-16	1,1E-50	7,3E-12	2,3E-21		103,48	146,89	71,24	92,18	514,05	460,18	153,20	327,67	127,67
CHB	1,3E-43	2,7E-22	2,5E-07	5,5E-28	9,1E-15	7,2E-24	2,0E-18	8,8E-12	2,7E-37	2,1E-15	6,1E-35	1,0E-20	9,8E-06	3,7E-41	4,9E-12	4,8E-28	3,6E-17		116,06	90,57	67,40	692,35	540,49	117,19	297,34	48,20
CDX	7,3E-34	9,9E-38	1,8E-05	3,8E-52	3,6E-05	3,4E-27	2,2E-06	3,8E-14	4,6E-49	2,6E-18	3,2E-48	1,0E-19	2,9E-08	1,1E-40	9,6E-16	1,0E-44	6,4E-26	1,1E-19		97,88	109,58	409,86	239,05	37,98	247,16	103,48
CHS	1,1E-29	2,4E-21	4,0E-17	5,6E-47	1,2E-23	6,5E-07	1,4E-20	7,9E-24	1,4E-24	1,3E-06	3,9E-34	4,2E-21	2,3E-13	1,8E-30	1,6E-15	3,2E-36	7,1E-11	1,3E-14	4,7E-16		79,71	539,16	417,12	154,28	391,92	122,84
KHV	9,3E-25	2,1E-16	2,0E-12	1,3E-15	1,1E-13	2,9E-23	5,9E-14	7,8E-09	5,0E-37	9,9E-07	1,3E-14	3,6E-29	3,2E-04	1,6E-28	1,7E-21	3,1E-12	6,2E-15	3,8E-10	2,2E-18	1,7E-12		457,43	399,30	124,42	317,29	157,45
JPT	3,5E-120	1,4E-70	2,5E-93	1,2E-101	3,1E-81	1,8E-94	8,1E-62	3,4E-101	7,1E-168	6,4E-82	8,9E-125	6,3E-148	1,4E-105	1,2E-132	7,5E-148	8,7E-120	3,2E-103	2,4E-141	4,9E-81	1,4E-108	3,8E-91		243,74	489,67	863,49	801,63
CLM	1,3E-74	6,9E-103	1,4E-69	4,8E-140	2,6E-52	4,5E-77	2,1E-54	4,2E-94	2,2E-122	3,4E-87	1,4E-110	1,5E-98	1,1E-82	2,3E-100	2,1E-99	1,2E-138	9,8E-92	7,3E-109	5,5E-45	1,4E-82	8,6E-79	5,7E-46		342,02	840,37	608,89
MXL	3,1E-44	4,7E-43	3,3E-08	1,3E-44	9,0E-07	8,0E-37	1,1E-03	2,2E-10	3,4E-54	7,0E-23	9,2E-44	2,5E-15	2,5E-08	1,3E-52	2,4E-18	1,1E-34	3,3E-27	6,7E-20	7,9E-05	2,0E-27	2,3E-21	5,1E-98	1,2E-66		190,07	93,60
PEL	2,0E-130	7,1E-74	8,5E-35	3,7E-67	1,8E-51	1,3E-103	3,6E-42	6,1E-31	1,6E-145	2,0E-60	3,6E-124	3,7E-81	9,8E-37	2,9E-88	5,5E-53	3,7E-77	1,3E-63	3,2E-57	1,1E-46	3,2E-77	2,0E-61	4,4E-178	4,1E-173	8,5E-35		216,37
PUR	6,8E-61	1,6E-40	1,6E-09	1,6E-51	1,7E-18	1,4E-32	1,3E-21	3,9E-16	4,7E-53	2,5E-26	8,1E-63	3,5E-21	7,2E-11	7,8E-55	1,9E-07	1,4E-46	5,2E-22	1,3E-06	3,6E-17	4,9E-21	4,4E-28	9,2E-166	1,8E-123	3,3E-15	3,0E-40	

Tabla R-6: Pruebas de Chi-2 para cada par de poblaciones estudiadas para las diferencias en el número de SNPs. Sobre la diagonal se presenta el valor del estadístico Chi-2. Bajo la diagonal, se encuentra el p-valor: si es menor de 0,01 se rechaza la hipótesis nula de homogeneidad de muestras y concluimos que hay diferencias, esto es, se trata de muestras independientes.

Diferencias intracontinentales

Se han analizado las diferencias entre las poblaciones de cada grupo continental.

En África, la población LWK tiene un número de SNPs en desequilibrio Hardy-Weinberg para los genes HLA-B y HLA-DMB bastante más alto que las demás poblaciones africanas. La población YRI posee el valor de SNPs en desequilibrio más alto para el gen HLA-A. La población ESN posee el valor más alto en el gen HLA-DQB1. La población GWD tiene muy pocos SNPs en desequilibrio en el gen HLA-DPA1 en comparación con el resto de poblaciones, y si bien solo son 5 SNPs en desequilibrio los que tiene en el gen HLA-DMA, es el valor más alto entre las poblaciones africanas. Además, GWD es la población que presenta más SNPs en desequilibrio para el conjunto de genes HLA, de entre todas las poblaciones africanas. La población MSL posee 101 SNPs en desequilibrio para el gen HLA-DPA1, siendo el valor más alto registrado entre las poblaciones africanas, y, de manera análoga, en los genes HLA-DRA, HLA-DMA, HLA-DMB, y HLA-DOA presenta valores de SNPs en desequilibrio por debajo de las respectivas medias para el conjunto poblacional de África. La población ACB presenta 82 SNPs en desequilibrio para el gen HLA-DRA, un valor 4 veces mayor que la media, y el valor más alto para ese gen entre las poblaciones africanas. La población ASW destaca por poseer junto con LWK el valor más alto de SNPs en desequilibrio para el gen HLA-DOA, y por tener el valor más bajo para el gen HLA-DPB1. Además, ASW es la población que presenta menos SNPs en desequilibrio para el conjunto de genes HLA, de entre todas las poblaciones, tanto a nivel continental como global. Por último cabe destacar que las poblaciones MSL, ACB y ASW no poseen ningún SNP en desequilibrio para el gen HLA-DMA, lo mismo que ocurre con la población ESN y el gen HLA-DMB.

En Europa, la población IBS presenta los valores más altos de SNPs en desequilibrio de entre las poblaciones europeas para los genes HLA-A, HLA-DMA y HLA-DMB, y el más bajo para HLA-B. La población TSI posee el valor más alto para HLA-B, el más bajo para HLA-DQA1, y además es la población que en el conjunto de genes HLA para las poblaciones europeas presenta más SNPs con desequilibrio. La población CEU presenta los valores más bajos en HLA-A, HLA-DMA y HLA-DOA, los más altos en HLA-C y HLA-DPB1, y además es la población con menos SNPs en desequilibrio dentro del conjunto europeo. La población GBR presenta el valor más bajo para los genes HLA-DRA y HLA-DQB1. En cuanto a la población FIN, es destacable el gran número de SNPs en desequilibrio que presenta en los genes HLA-DRA, HLA-DPA1, HLA-DQA1 y HLA-DQB1; y los valores tan bajos que presenta en HLA-DMB y HLA-DPB1. A diferencia de lo ocurrido en la muestra de poblaciones africanas donde se han hallado varios casos de poblaciones que no tenían ningún SNP en desequilibrio para determinados genes, en el caso de Europa solamente se da en el caso de CEU y el gen HLA-DMA.

Dentro del conjunto de poblaciones asiáticas, la población PJL posee el valor más bajo en el gen HLA-DMB junto con las poblaciones BEB, CHB y KHV. La población BEB posee además el valor más alto para el gen HLA-DPB1, superando en 4 veces la media de todas las poblaciones asiáticas.

La población GIH posee el valor más bajo para SNPs en desequilibrio en el gen HLA-C. La población ITU posee el valor más alto en los genes HLA-B, HLA- DMA y HLA-DQA1, y el más bajo en el gen HLA-DPB1 siendo casi 14 veces menor que la media de SNPs en desequilibrio para el gen HLA-DPB1 para las poblaciones asiáticas. La población STU destaca por ser la que tiene el valor más alto en el gen HLA-DMB, algo más de 4 veces la media asiática. La población CHB es la que posee más SNP en desequilibrio en el gen HLA-A, y además es la que presenta un mayor número total de SNPs en desequilibrio para el conjunto de todos los genes HLA estudiados de entre las poblaciones asiáticas. La población CDX posee el valor más bajo en los genes HLA-A, HLA-B y HLA-DQA1, no posee ningún SNP en desequilibrio para el gen HLA-DMA, y además posee el menor número total de SNPs en desequilibrio para el conjunto de todos los genes HLA estudiados entre las poblaciones asiáticas. La población CHS es la que posee el valor más alto en el gen HLA-DRA. La población KHV posee el valor más bajo del gen HLA-DRA y el más alto para el gen HLA-DOA de entre las poblaciones de Asia. El caso de JPT es especial dado el aislamiento por insularidad que ha sufrido respecto al resto de las poblaciones asiáticas. Así, llama la atención los valores tan altos de SNP en desequilibrio que presenta en los genes HLA-C y, en especial, en el gen HLA-DPA1. Cabe igualmente destacar que no posee ningún SNP en desequilibrio en el gen HLA-DOA y que tiene el valor más bajo en el gen HLA-DQB1 en las poblaciones de Asia.

En América destaca la variabilidad observada entre las distintas poblaciones. Por ejemplo, incluso entre los genes HLA-DMB y HLA-DOA que presentan valores generales de SNP en desequilibrio bajos en el conjunto global de poblaciones estudiadas, en el caso de las poblaciones americanas presentan un rango mayor que en el resto de grupos continentales. De igual manera hay una gran dispersión en el total de SNPs en desequilibrio que presentan las poblaciones americanas respectos al resto de continentes. La población CLM tiene el valor más alto de HLA-B, HLA-DPA1 (más de 3 veces mayor que la media americana) y HLA-DPB1 (más de 2,5 veces mayor); mientras que no presenta ningún SNP en desequilibrio en el gen HLA-DMB; y junto con MXL tiene el valor más bajo en HLA-C. La población MXL tiene los valores más bajos en HLA-A, HLA-C, HLA-DRA, HLA-DPB1, HLA-DQA1, HLA-DQB1, y junto con PUR, en HLA-DMA. Además es una de las poblaciones con menos SNPs en desequilibrio a nivel global, solo superada por ASW. La población PEL destaca por tener los valores más altos de SNPs en desequilibrio Hardy-Weinberg en los genes HLA-C, HLA-DMA, HLA-DMB y HLA-DQA1; y al mismo tiempo ser la población que menos SNPs en desequilibrio posee en los genes HLA-B y HLA-DPA1. Por último, la población PUR es la que más SNPs en desequilibrio presenta a nivel americano en los genes HLA-A, HLA-DRA y HLA-DQB1; y además es la población que más SNPs en desequilibrio presenta a nivel americano y global. Por otro lado, PUR es la población que menos SNPs en desequilibrio presenta en los genes HLA-DMA (junto con MXL) y HLA-DOA.

Considerando la Tabla R-6 podemos observar que en todas las comparaciones entre pares de poblaciones se descarta la homogeneidad de muestras, y por tanto podemos decir que observamos diferencias, por lo que las muestras son independientes entre sí, esto es, en todos los casos hay menos de un 0,01 de probabilidad de que los resultados puedan ser explicados por el

azar. De hecho, todas las comparaciones realizadas superan el test con un valor de significación de 0,01.

Por otra parte, dado el elevado número de comparaciones realizado, cabe aplicar la corrección de Bonferroni, con el fin de evitar rechazar incorrectamente una hipótesis nula por la realización de comparaciones múltiples. En este caso, la significación debería ser inferior a $3{,}08.10^{-5}$.

En este caso, tan sólo unos pocos emparejamientos no mostrarían significación. Es el caso de parejas formadas por ESN/CDX, ESN/MSL, ESN/PJL, ASW/MXL y PJL/KHV. Teniendo en cuenta la procedencia de estas muestras, podemos concluir que la proporción de SNPs con desequilibrios de ligamiento no sigue un patrón geográfico determinado y se distribuye entre poblaciones al azar.

En la Tabla R-7 se muestra la comparación entre continentes. Salvo en el caso de África/Asia, todos los pares de poblaciones se descarta la homogeneidad de muestras, y por tanto podemos decir que observamos diferencias y que las muestras son independientes entre sí, esto es, en todos los casos hay menos de un 0,01 de probabilidad de que las diferencias puedan ser explicadas por el azar. En el caso de la pareja formada por África y Asia el p-valor es de 0,159 por lo que no podemos decir que observamos diferencias entre este par de continentes, ni que las muestras son independientes entre sí.

En todo caso, podemos afirmar que no hay un patrón geográfico de distribución de las proporciones de SNPs con desequilibrios entre continentes.

	ÁFRICA	**EUROPA**	**ASIA**	**AMÉRICA**
ÁFRICA		35,53	15,55	98,81
EUROPA	2,03E-04		54,24	164,38
ASIA	1,59E-01	1,07E-07		76,87
AMÉRICA	3,07E-16	1,69E-29	5,93E-12	

Tabla R-7: Pruebas de Chi-cuadrado para cada par de agrupaciones continentales estudiadas. *Sobre la diagonal se presenta el valor del estadístico Chi-2. Bajo la diagonal, se encuentra el valor de p: si es menor de 0,01 se rechaza la hipótesis nula de homogeneidad de muestras y concluimos que hay diferencias, esto es, se trata de muestras independientes.*

Poblaciones desplazadas

Dado que algunos muestreos de poblaciones se han hecho en grupos emigrantes en países diferentes, y en muchos casos también en continentes distintos, es interesante comparar dichas poblaciones inmigrantes con las poblaciones originarias, con el fin de conocer sus afinidades genéticas.

En Norteamérica, se han muestreado 4 poblaciones: ASW (de origen africano), CEU (de origen europeo central, GIH (de origen asiático) y MXL (autóctona americana). La población GIH, muestreada de Houston (Texas), es con diferencia la que menos se parece al resto, seguida por la población CEU. Las poblaciones MXL y ASW, si bien difieren en el número de SNPs en desequilibrio de ligamiento en todos los genes, dichas diferencias son menores que en comparación con el resto de poblaciones, salvo con los genes HLA-A en el que MXL y CEU presentan el mismo número de SNPs en desequilibrio, y el gen HLA-DMA para el que ASW y CEU no presentan ningún SNP en desequilibrio.

En el resto de América, concretamente Sudamérica y la zona del Caribe se han muestreado 3 poblaciones: CLM (de origen sudamericano), PUR (de origen caribeño) y ACB (de ancestría africana). Dado que GIH se ha muestreado en Houston, ciudad a orillas del Golfo de México, se ha incluido también en esta comparativa. A diferencia de Norteamérica, en este caso es más difícil subdividir las poblaciones por diferencias puntuales. Sin embargo, las poblaciones GIH (por distancia) y PUR (por aislamiento insular) son las que más difieren del conjunto de poblaciones. Podemos observar múltiples diferencias entre distintas poblaciones para cada gen muestreado, no siendo siempre las mismas poblaciones las que difieren entre sí, lo que vendría a remarcar el aislamiento poblacional que se da en esta zona.

En Europa, se han muestreado 3 poblaciones: GBR (autóctona de Inglaterra y Escocia), e ITU y STU, ambas de origen asiático. La población STU muestra un número mayor de SNPs en desequilibrio en los genes que poseen menos de forma general (HLA-DRA, HLA-DMB y HLA-DPB1). En general, para el conjunto de todos los genes estudiados, podemos afirmar que ITU se asemeja más a GBR que a STU si comparamos el número de SNPs en desequilibrio Hardy-Weinberg para cada gen.

Análisis de Componentes Principales

Un análisis multivariante es un método estadístico utilizado para determinar la contribución de varios factores en un simple evento o resultado, y ayuda a sintetizar la información y desvelar tendencias subyacentes.

El análisis de componentes principales (ACP) se utiliza para describir un conjunto de datos en términos de nuevas variables no correlacionadas a las que se les da el nombre de componentes. Los componentes se ordenan por la cantidad de varianza original que describen, por lo que el ACP resulta de gran utilidad para reducir la dimensionalidad de un conjunto de datos. El ACP se emplea sobre todo en análisis exploratorio de datos y para construir modelos predictivos. Comporta el cálculo de la descomposición en autovalores de la matriz de covarianza, normalmente tras centrar los datos en la media de cada atributo.

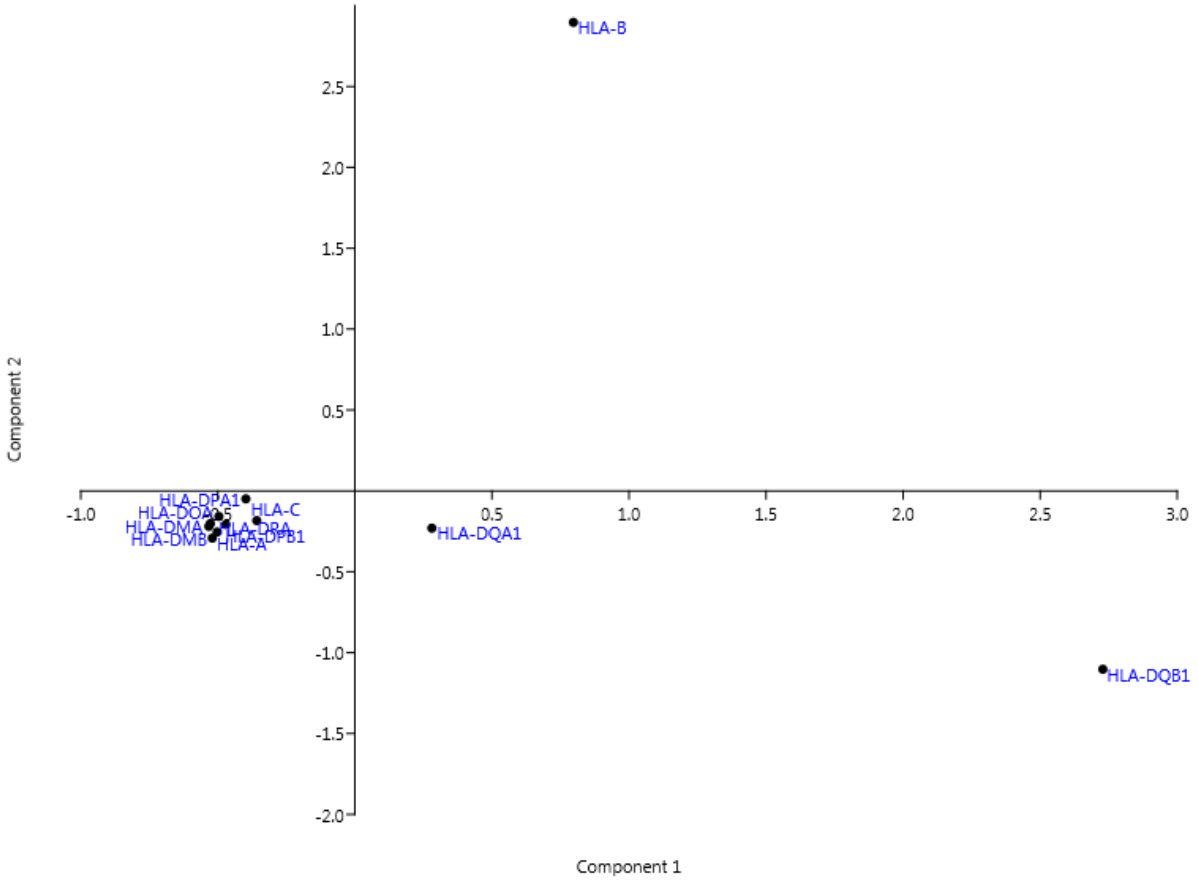


Figura R-15: Análisis de Componentes Principales de los genes, realizado sobre el número de SNPs con desequilibrio Hardy-Weinberg por gen y población. La varianza explicada es del 97,76% para el primer eje y del 1,48% para el segundo.

En un ACP realizado sobre el número de SNPs en desequilibrio Hardy-Weinberg por gen y población (Figura R-15), se observó un eje predominante, con HLA-DQB1 en el extremo positivo, seguido de HLA-B, HLA-DQA1 y finalmente en el otro extremo el resto de genes. Es claramente un eje de tamaño, ya que ordena los genes por número de poblaciones en desequilibrio. El segundo eje se caracteriza por HLA-B, que parece mostrar un patrón de distribución geográfico diferente al resto de genes. De hecho, en HLA-B se observa un número más alto de poblaciones en desequilibrio en África y Europa, en tanto que para el resto de genes África no muestra valores máximos.

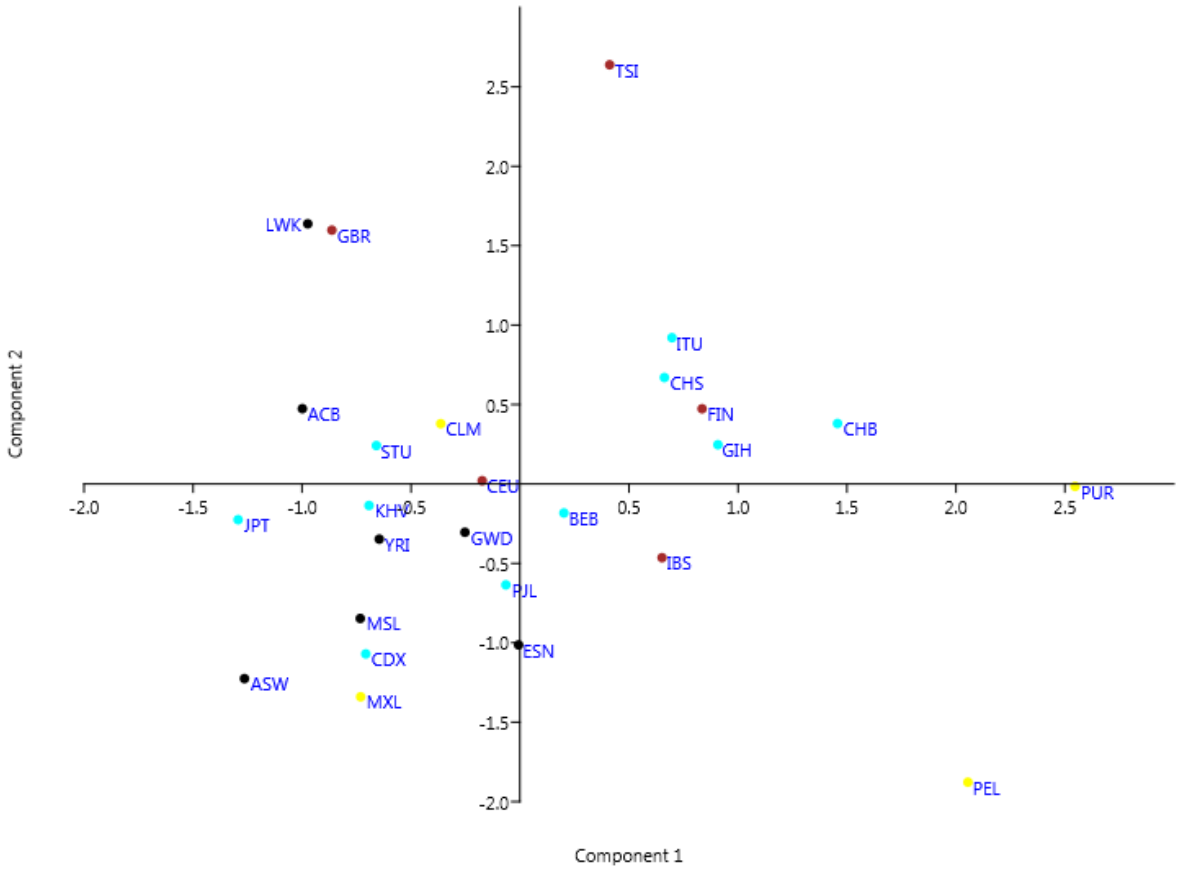


Figura R-16: Análisis de Componentes Principales de las poblaciones, realizado sobre el número de SNPs con desequilibrio Hardy-Weinberg por gen y población. La varianza explicada es del 51,02% para el primer eje y del 31,58% para el segundo. Las poblaciones africanas están representadas por puntos negros, las europeas por puntos rojos, las asiáticas por puntos azules y las americanas por puntos amarillos.

Al realizar un análisis equivalente, considerando las poblaciones como variables (Figura R-16), se observa una disposición de éstas poco estructurada. No obstante, las poblaciones africanas se disponen en el lado negativo del eje 1. El resto de continentes presentan sus poblaciones distribuidas a ambos lados, tanto del eje 1 como del eje 2.

Varianza de Wahlund

Como ya se ha explicado, en genética de poblaciones el efecto Wahlund se refiere a la reducción de la heterocigosidad en una población, causada por su estructuración en subpoblaciones. Es decir, si dos o más subpoblaciones muestran diferentes frecuencias de alelos, entonces la heterocigosidad general se reduce, incluso si las subpoblaciones por separado están en

equilibrio de Hardy-Weinberg. Las causas subyacentes de esta subdivisión de la población podrían ser las barreras (geográficas o de otro tipo) al flujo génico seguidas por la deriva genética entre las subpoblaciones. Dicho esto, la varianza de Wahlund mide la varianza observada en relación con el valor máximo de varianza que podríamos observar si las subpoblaciones estuvieran completamente diferenciadas para un locus dado. Este valor puede usarse a modo de reloj molecular relativo para estimar tiempos de divergencia.

Diferencias en la varianza de Wahlund

En líneas generales, en todos los genes estudiados se aprecian múltiples regiones cuyos valores de varianza de Wahlund son exactamente iguales. Esto es debido a que son regiones que se heredan ligadas, esto es, que no segregan durante la meiosis. Así, por ejemplo, en el gen HLA-A hay una región con varianzas de Wahlund idénticas entre las posiciones 29909674 y 29913588 que afecta a 34 SNPs que presentan variabilidad en alguna de las poblaciones; en el gen HLA-B hay una región entre las posiciones 31257685 y 31320559 que afecta a 96 SNPs, etc.

Además se aprecia una tendencia a la concentración de estas regiones de ligamiento en zonas donde la varianza de Wahlund es más baja. Esto está en concordancia con el hecho de que la presencia de desequilibrio de ligamiento en una región genómica tiene una relación inversa con la heterocigosidad en los loci de dicha región.

Se observan diferencias sustanciales entre los distintos genes (Tabla R-8) en los valores de la varianza de Wahlund.

Podemos apreciar que los valores mínimos de varianza de Wahlund son muy similares, oscilando entre 0 y 0,0043.

Es importante destacar que todos los genes estudiados (exceptuando HLA-A, HLA-DMA y HLA-DQA1) tienen algunos SNP cuya varianza de Wahlund es igual a cero. Esto tan sólo quiere decir que la plataforma del proyecto 1000 Genomas caracteriza el SNP como polimórfico, quizá a partir de trabajos previos, pero en la propia base de datos del proyecto no lo es.

El caso de HLA-DRA es de especial interés ya que tiene un valor más alto de varianza de Wahlund en el conjunto de genes. Un valor alto de la varianza de Wahlund indica que ha ocurrido algo que ha hecho que las diferencias para ese gen entre distintos grupos poblacionales o subpoblacionales sean grandes. Como ya hemos visto, esto puede deberse a una acción de la deriva genética que no ha sido correspondida con la acción contrapuesta del flujo génico, como por ejemplo en casos de insularidad, pero la razón más plausible, cuando como en este caso, todos los genes se refieren al mismo grupo de poblaciones, es que algún proceso diversificador haya actuado con especial intensidad en este gen.

GENES	WAHLUND MIN.	WAHLUND MAX.	MEDIA	MEDIANA
HLA-A	0,0042	0,1460	0,0306	0,0259
HLA-B	0	0,1918	0,0364	0,0339
HLA-C	0	0,1309	0,0234	0,0225
HLA-DRA	0	0,8504	0,0542	0,0212
HLA-DMA	0,0043	0,1215	0,0152	0,0079
HLA-DMB	0	0,1276	0,0158	0,0089
HLA-DOA	0	0,1778	0,0233	0,0097
HLA-DPA1	0	0,3713	0,0533	0,0286
HLA-DPB1	0	0,3713	0,0622	0,0391
HLA-DQA1	0,0043	0,1201	0,0423	0,0417
HLA-DQB1	0	0,1912	0,0479	0,0468

Tabla R-8: Resumen de los valores de la varianza de Wahlund de los distintos genes HLA estudiados. Se representan los valores mínimo, máximo, medio y el valor de la mediana para cada gen.

Relación entre desequilibrio Hardy-Weinberg y varianza de Wahlund

Un criterio utilizado habitualmente para establecer indicios acerca de posibles presiones selectivas es el valor de la varianza de Wahlund. En efecto, tradicionalmente se ha considerado que un elevado valor de la varianza podría ser el resultado de algún tipo de selección. Por ello, se han analizado los valores de la varianza de Wahlund para cada SNP y se han contrastado con el número de poblaciones en desequilibrio de Hardy-Weinberg. A continuación se muestran los resultados para todos los genes analizados. Entre todos los que muestran valores muy elevados de la varianza de Wahlund o un gran número de poblaciones en desequilibrio, sólo unos pocos se encuentran en regiones codificantes o promotoras. Son estos los que se comentarán, puesto que es muy difícil conocer los factores que afectan a los SNPs en los intrones o regiones intergénicas.

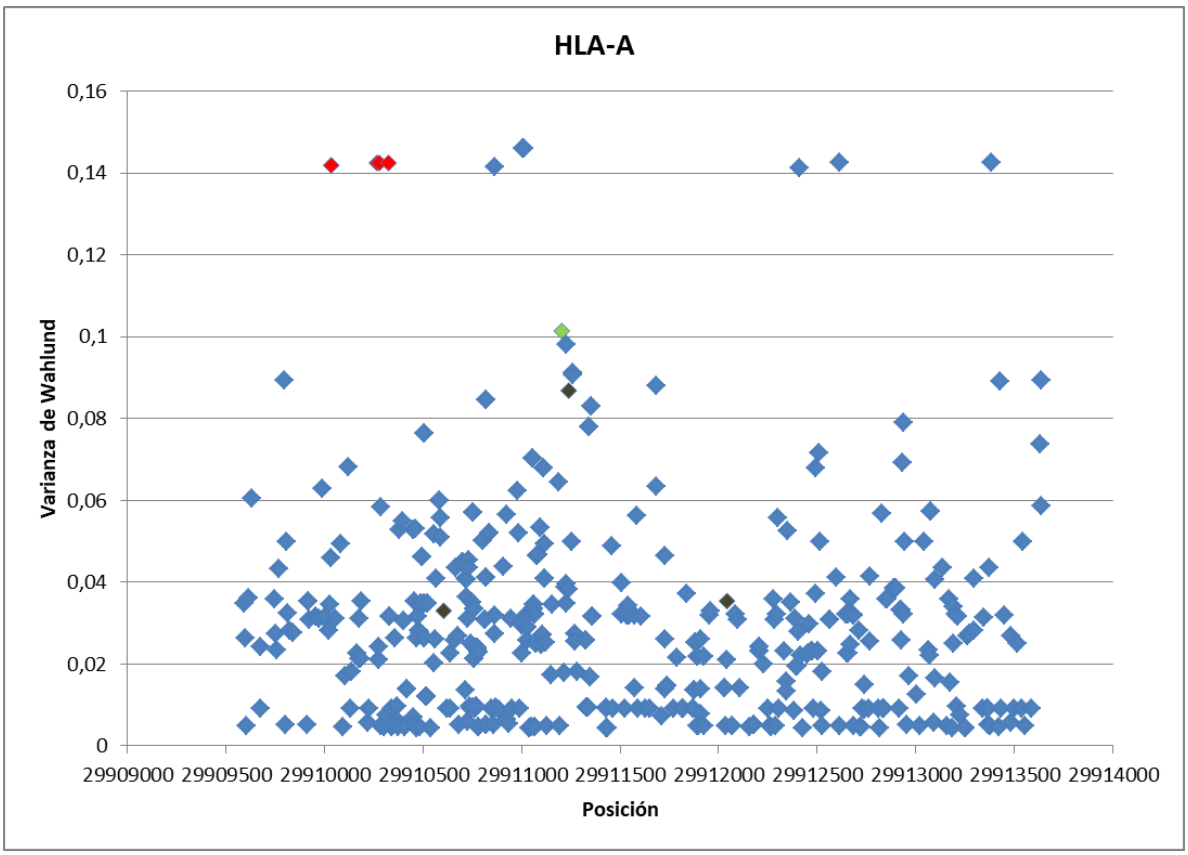


Figura R-17: Valores de la varianza de Wahlund de los SNPs del gen HLA-A en relación a su posición.

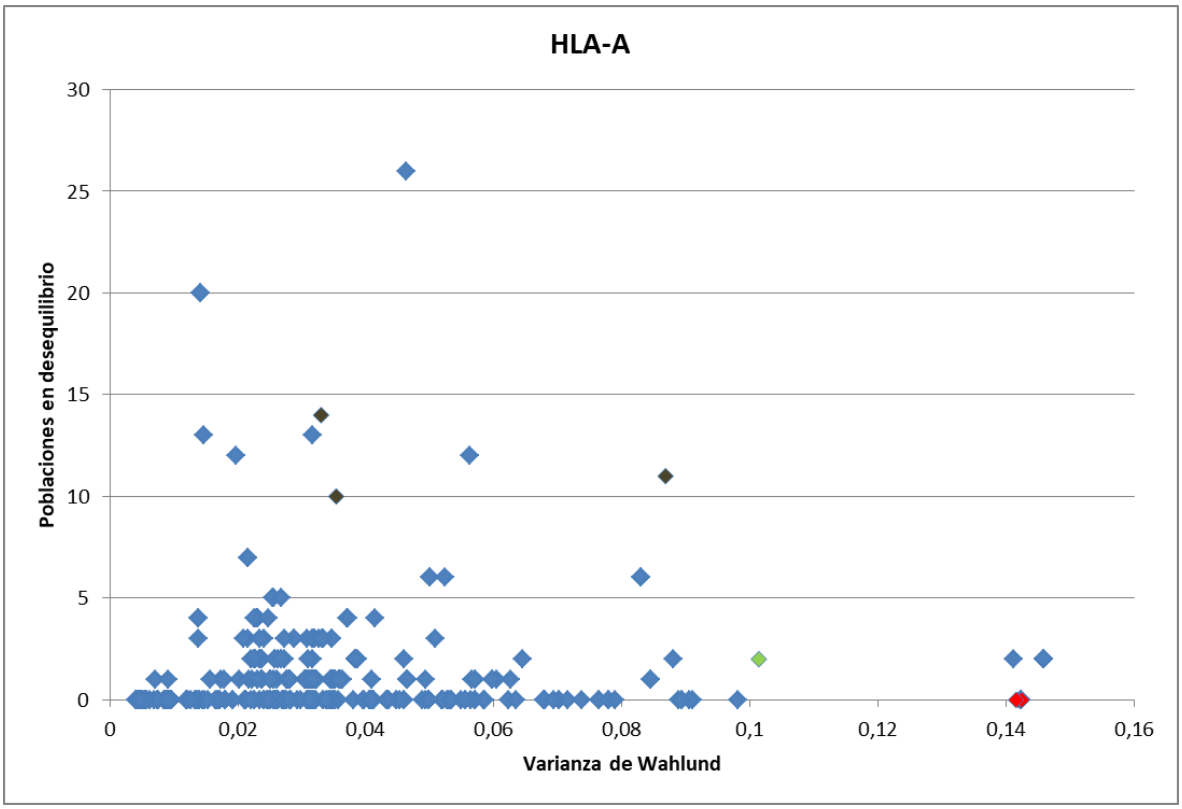


Figura R-18: Relación entre los valores de la varianza de Wahlund y el número de poblaciones en desequilibrio de Hardy-Weinberg para los SNPs del gen HLA-A.

En las figuras R-17 y R-18 se muestran los valores de la varianza de Wahlund de cada SNP en relación a su posición y el número de poblaciones en desequilibrio para el gen HLA-A.

Entre los SNPs con un valor de la varianza de Wahlund por encima de 0,1, se han encontrado 3 en la región promotora (rs41545520, rs9260120 y rs114945359, marcados en rojo) y uno en el exón 3 del gen HLA-A (rs1059517, marcado en verde). Ninguno de estos SNP con un valor de la varianza de Wahlund por encima de 0,1 ha mostrado desequilibrio Hardy-Weinberg en más de 3 poblaciones.

Entre los SNPs que muestran un número de poblaciones con desequilibrio de Hardy-Weinberg en más de 10 poblaciones, se encuentra rs12721675. Este SNP aparece en desequilibrio en 14 poblaciones repartidas por los diferentes continentes, siempre con exceso de heterocigotos y muestra un valor de varianza de Wahlund moderado (0,033, marcado en marrón). Se encuentra en el segundo exón de HLA-A. El SNP rs9260155 se encuentra en el exón 3 de HLA-A. Su varianza es de 0,087 y se encuentra en desequilibrio en 11 poblaciones, con exceso de heterocigotos. Se ha marcado también en color marrón.

Los SNPs rs558831267 y rs145046067, a dos nucleótidos de distancia, se encuentran en desequilibrio en 20 poblaciones, pero este hecho se debe a que se ven afectados por sendos indels. Lo curioso es que se encuentran en un exón, el cuarto de HLA-A. Hemos de suponer, puesto que son ambos polimórficos, que corresponden a la variabilidad genética de este gen, configurando alelos con diferente longitud. A 12 nucleótidos de distancia, se encuentra el SNP rs150028516, con una varianza de 0,035 y desequilibrio en 10 poblaciones, con exceso de heterocigotos. Se ha coloreado también en marrón.

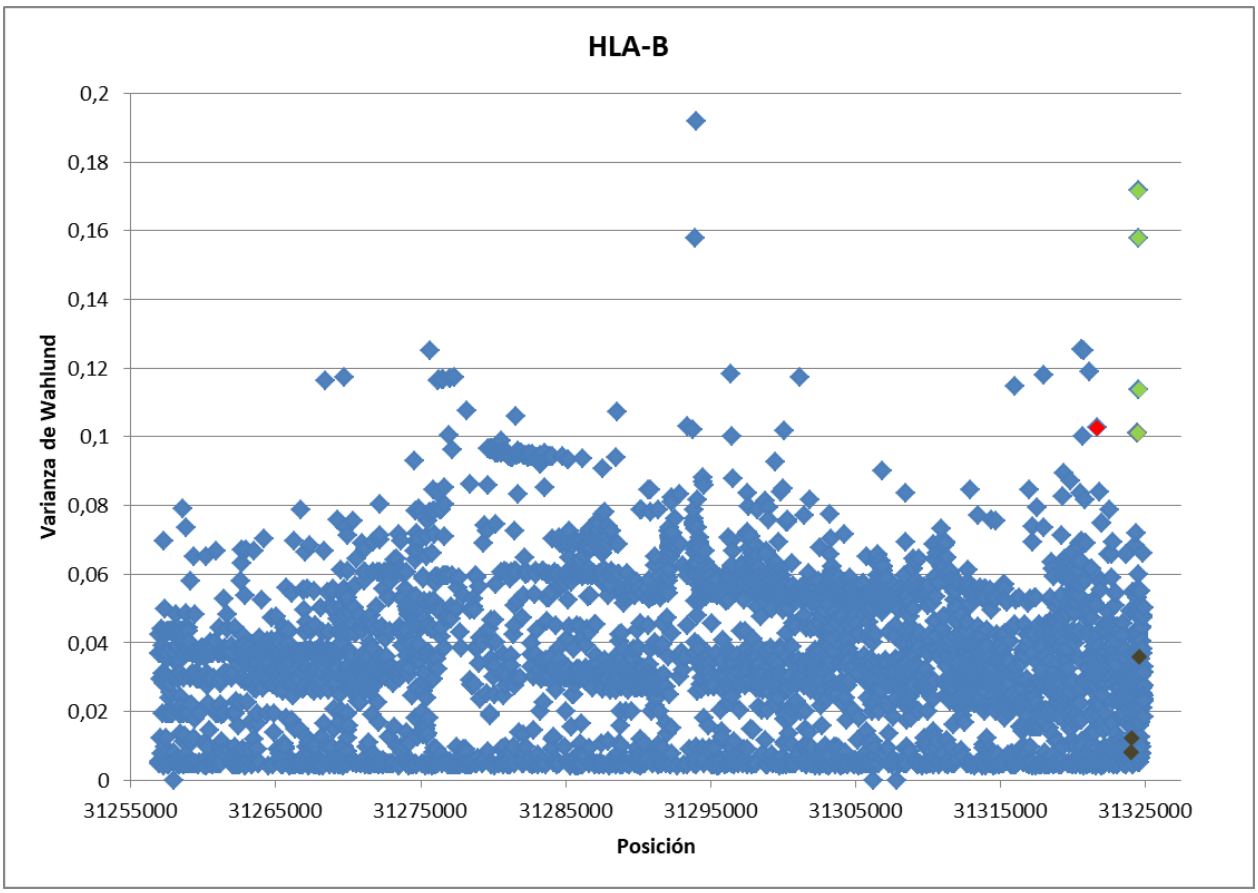


Figura R-19: Valores de la varianza de Wahlund de los SNPs del gen HLA-B en relación a su posición.

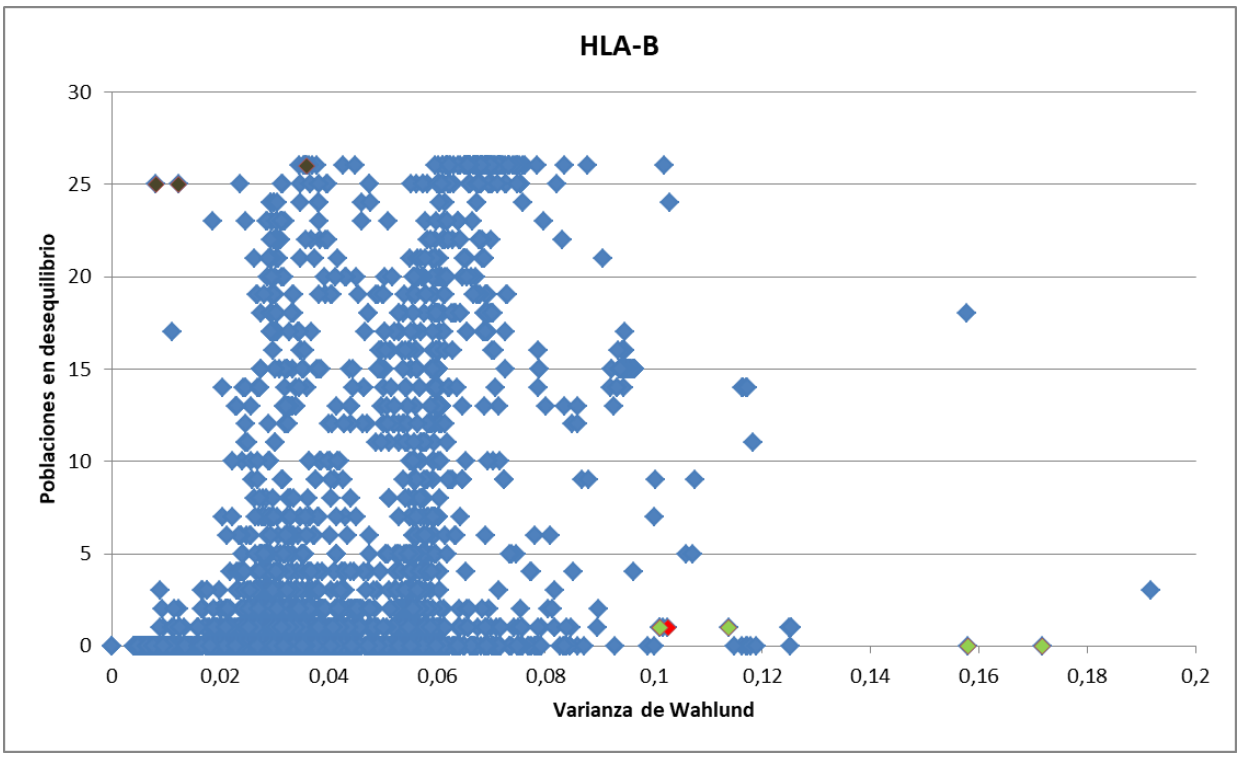


Figura R-20: Relación entre los valores de la varianza de Wahlund y el número de poblaciones en desequilibrio de Hardy-Weinberg para los SNPs del gen HLA-B.

En las figuras R-19 y R-20 se muestran los valores de la varianza de Wahlund de cada SNP en relación a su posición y el número de poblaciones en desequilibrio para el gen HLA-B.

Entre los SNPs con un valor de la varianza de Wahlund por encima de 0,1, se han encontrado uno en la región reguladora (rs1058067, marcado en rojo) y 4 en el exón 2 del gen HLA-B (rs41553715, rs41546313, rs707909 y rs1050538, marcados en verde). Ninguno de estos SNP con un valor de la varianza de Wahlund por encima de 0,1 ha mostrado desequilibrio Hardy-Weinberg en más de 1 población.

Entre los SNPs que muestran un número de poblaciones con desequilibrio de Hardy-Weinberg en 25 poblaciones o más, se encuentran rs151341293 y rs709053 con 25 poblaciones en desequilibrio. Se encuentran en ambos casos con exceso de heterocigotos y un valor relativamente bajo de varianza de Wahlund (0,0081 y 0,0123 respectivamente, marcados en marrón), y se localizan en el tercer exón de HLA-B. El SNP rs9266178, localizado en el exón 2 de HLA-B, con 26 poblaciones en desequilibrio, en las 26 poblaciones se da exceso de homocigotos.

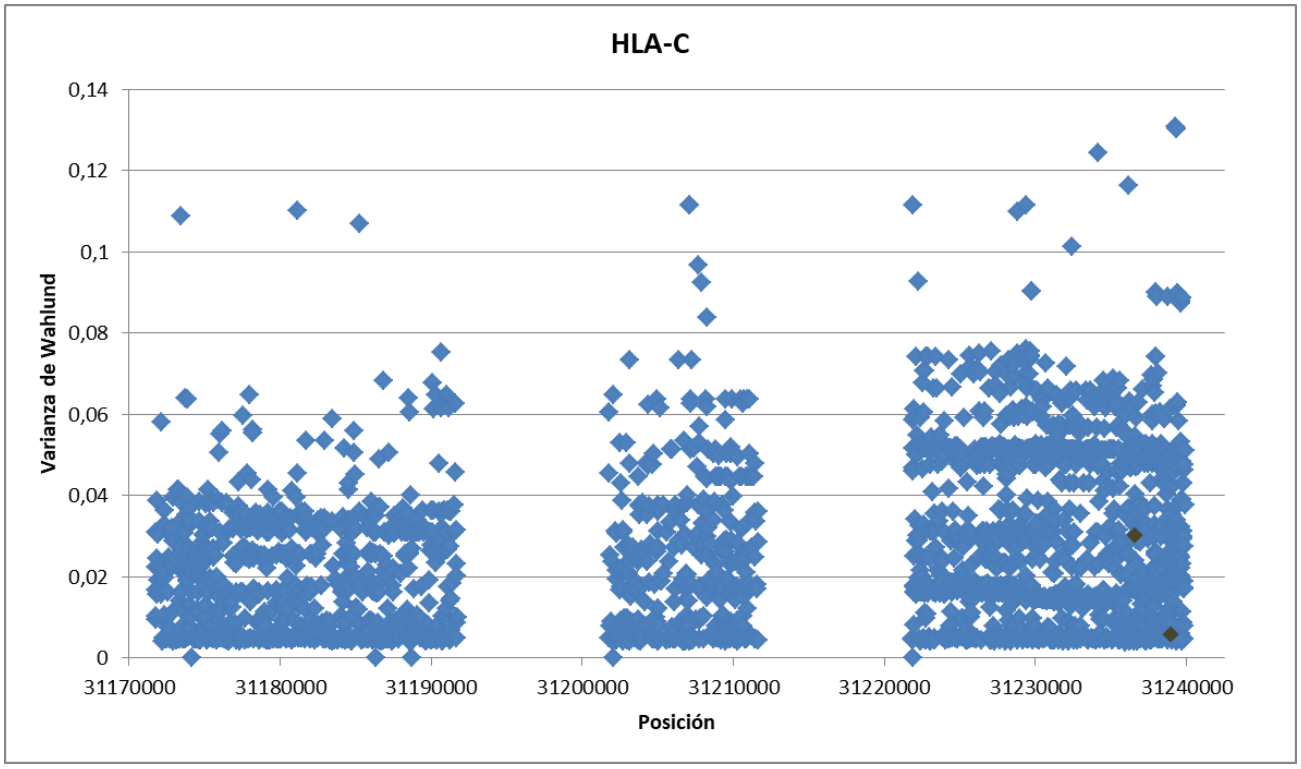


Figura R-21: Valores de la varianza de Wahlund de los SNPs del gen HLA-C en relación a su posición.

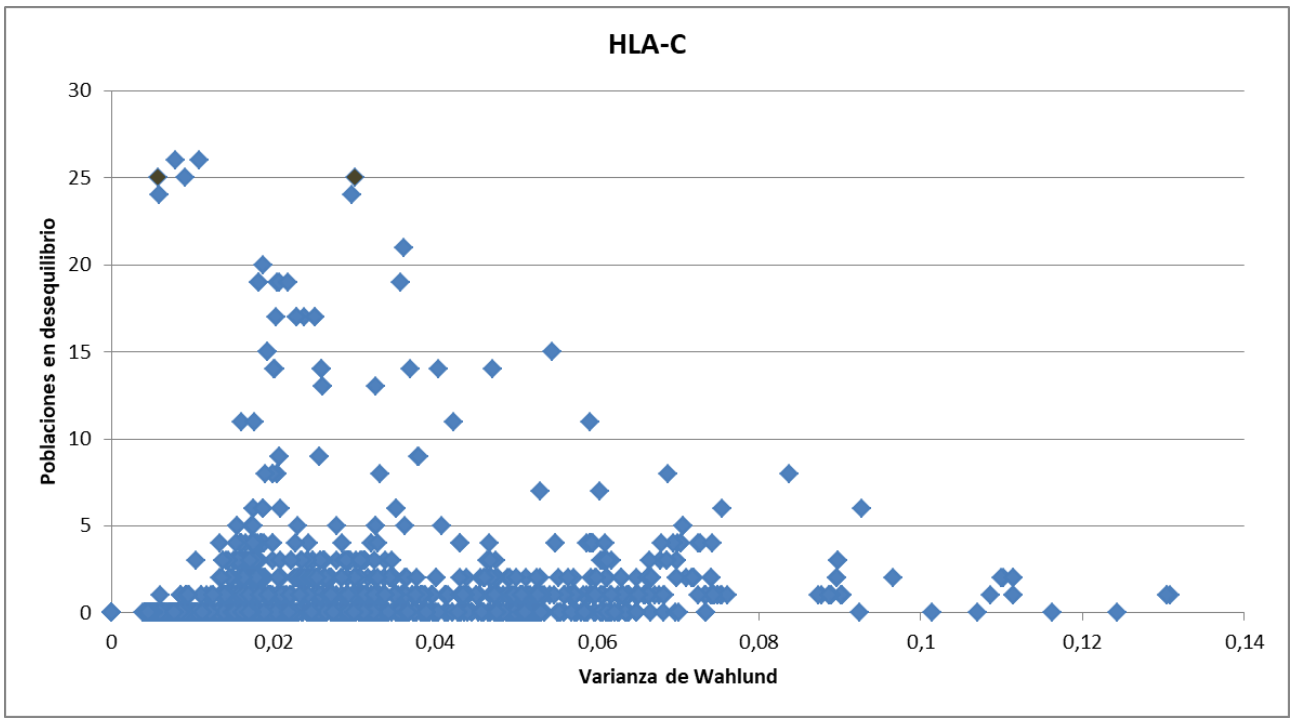


Figura R-22: Relación entre los valores de la varianza de Wahlund y el número de poblaciones en desequilibrio de Hardy-Weinberg para los SNPs del gen HLA-C.

En las figuras R-21 y R-22 se muestran los valores de la varianza de Wahlund de cada SNP en relación a su posición y el número de poblaciones en desequilibrio para el gen HLA-C.

No se han hallado SNPs con un valor de la varianza de Wahlund por encima de 0,1 ni en la región reguladora ni en ninguno de los 8 exones de HLA-C.

Entre los SNPs que muestran un número de poblaciones con desequilibrio de Hardy-Weinberg en 25 poblaciones o más, se encuentran rs3189472 y rs2308585 con 25 poblaciones en desequilibrio. En el caso de rs3189472, que presenta un valor de varianza de Wahlund moderado (0,0301, marcado en marrón) y está localizado en el exón 8, las 25 poblaciones se encuentran en desequilibrio con exceso de homocigotos. En el caso de rs2308585, se encuentra con exceso de heterocigotos y un valor relativamente bajo de varianza de Wahlund (0,0058, marcado en marrón), y se localiza en el tercer exón de HLA-C.

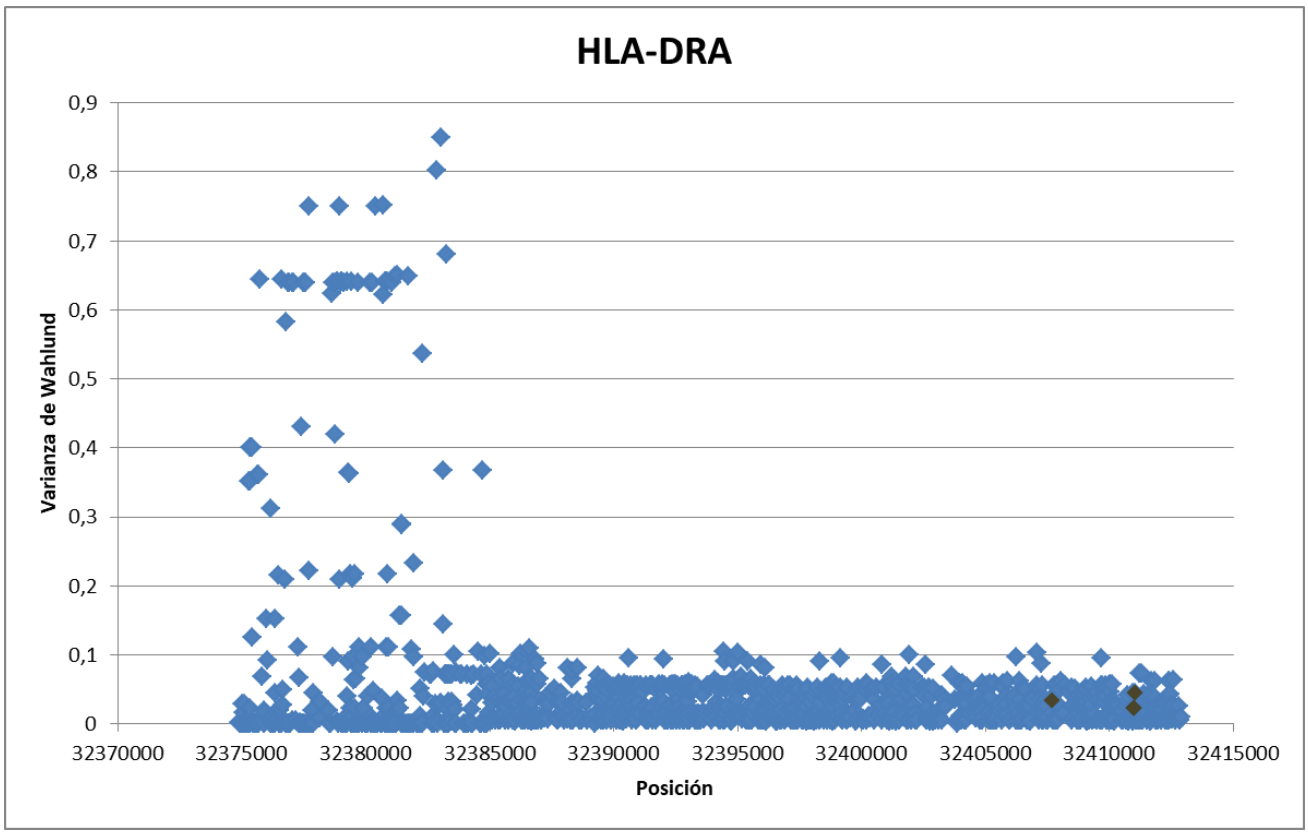


Figura R-23: Valores de la varianza de Wahlund de los SNPs del gen HLA-DRA en relación a su posición.

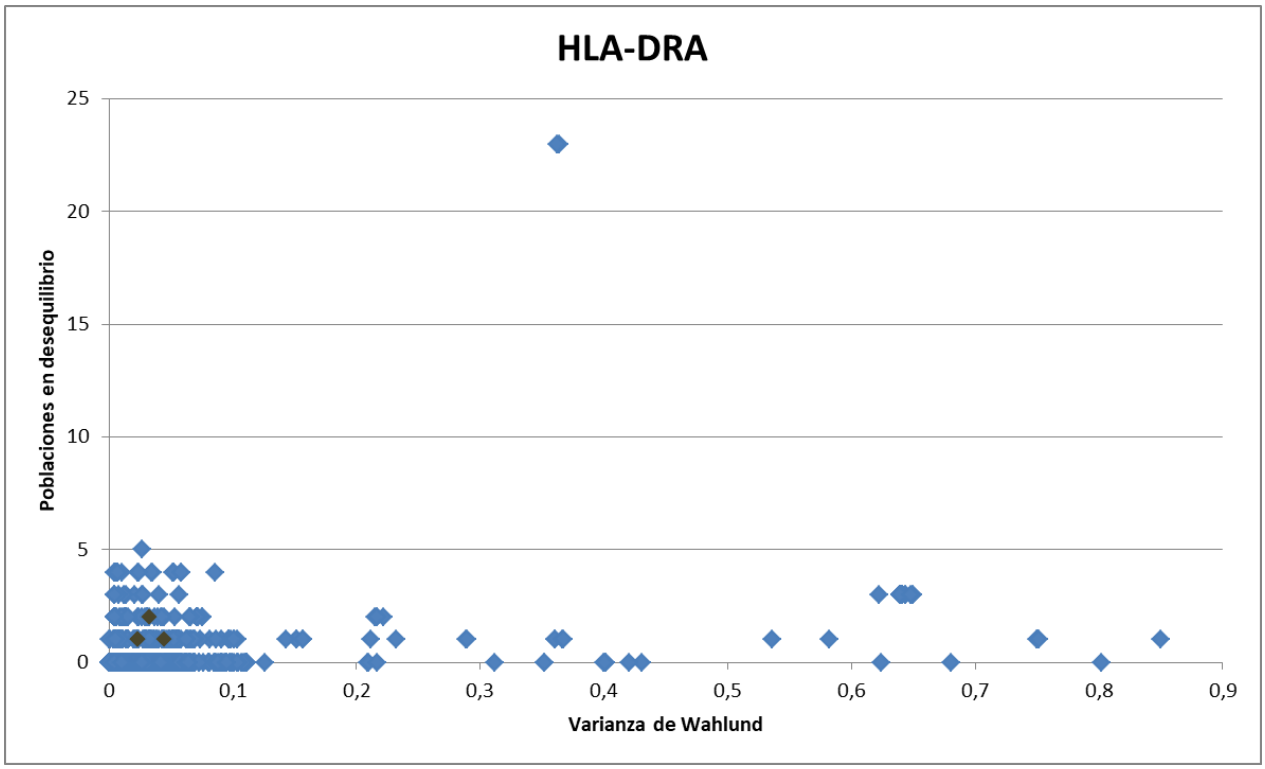


Figura R-24: Relación entre los valores de la varianza de Wahlund y el número de poblaciones en desequilibrio de Hardy-Weinberg para los SNPs del gen HLA-DRA.

En las figuras R-23 y R-24 se muestran los valores de la varianza de Wahlund de cada SNP en relación a su posición y el número de poblaciones en desequilibrio para el gen HLA-DRA.

No se han hallado SNPs con un valor de la varianza de Wahlund por encima de 0,1 ni en la región reguladora ni en ninguno de los 5 exones de HLA-DRA.

Entre los SNPs que muestran un número de poblaciones con desequilibrio de Hardy-Weinberg (marcados en marrón), se encuentran en la región reguladora 5´ el SNP rs14004 con 2 poblaciones en desequilibrio, amabas con exceso de homocigotos, y una varianza de Wahlund de 0,0331. En el exón 3 tenemos los SNPs rs3135391 y rs8084 con una población en desequilibrio, en ambos casos es la población FIN (Fineses de Finlandia) y en ambos casos se da un desequilibrio con exceso de homocigotos. Sus varianzas de Wahlund son 0,0232 y 0,0444, respectivamente.

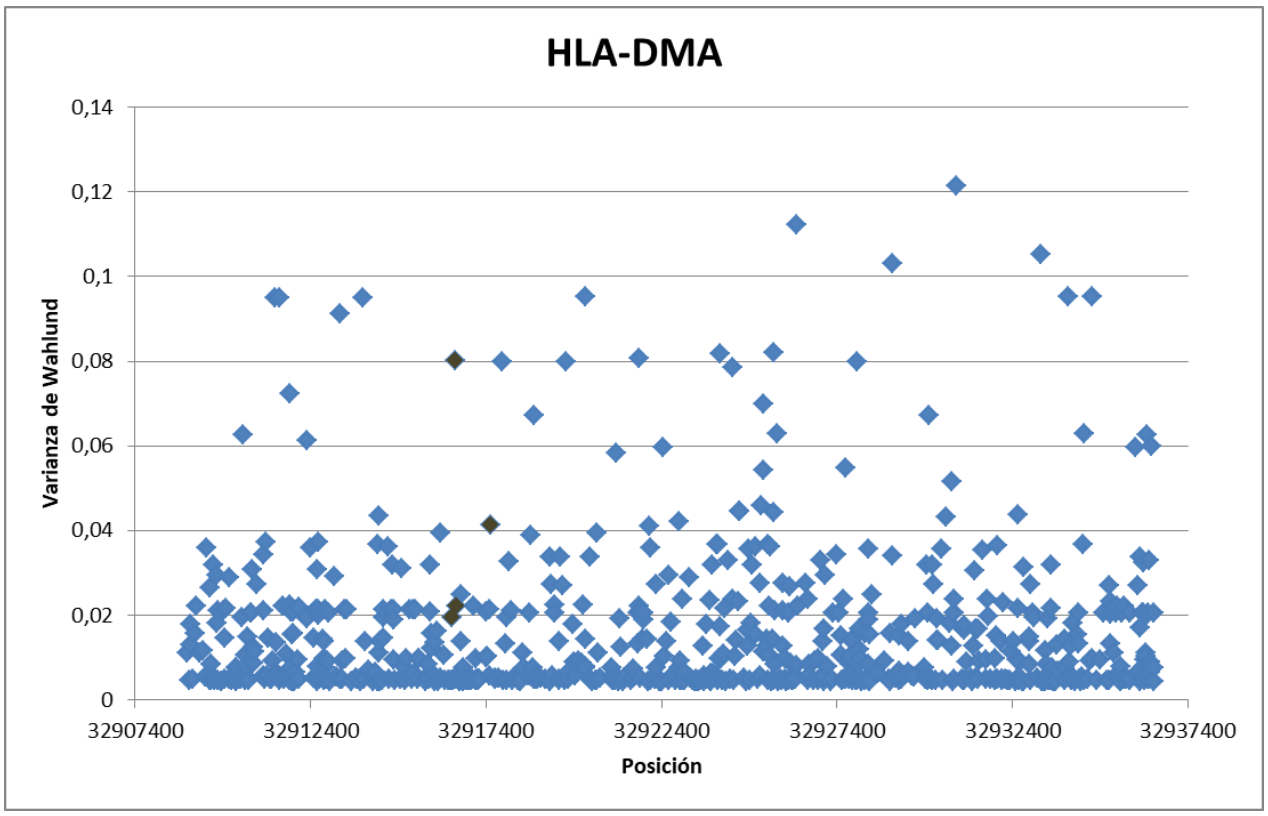


Figura R-25: Valores de la varianza de Wahlund de los SNPs del gen HLA-DMA en relación a su posición.

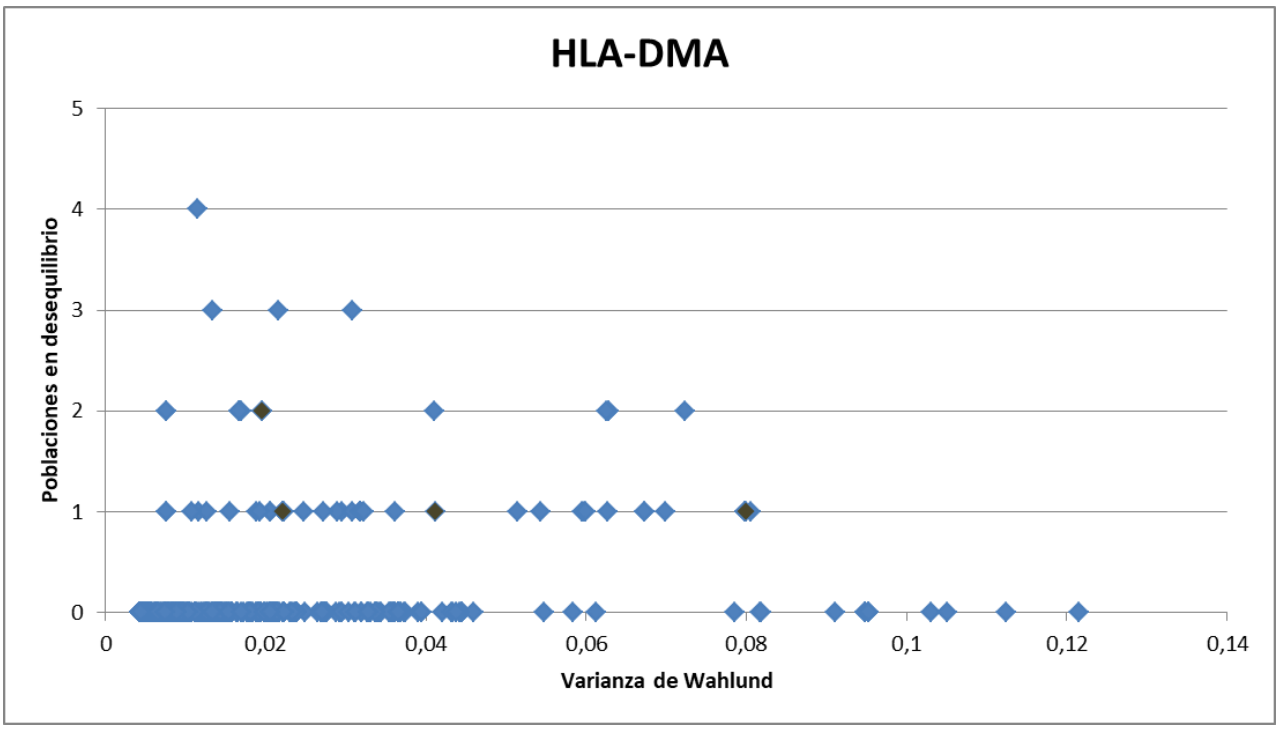


Figura R-26: Relación entre los valores de la varianza de Wahlund y el número de poblaciones en desequilibrio de Hardy-Weinberg para los SNPs del gen HLA-DMA.

En las figuras R-25 y R-26 se muestran los valores de la varianza de Wahlund de cada SNP en relación a su posición y el número de poblaciones en desequilibrio para el gen HLA-DMA.

No se han hallado SNPs con un valor de la varianza de Wahlund por encima de 0,1 ni en la región reguladora ni en ninguno de los 5 exones de HLA-DMA.

Entre los SNPs que muestran un número de poblaciones con desequilibrio de Hardy-Weinberg (marcados en marrón), se encuentran en la región reguladora 5´ el SNP rs116611693 con 2 poblaciones en desequilibrio, ambas con exceso de homocigotos, y una varianza de Wahlund de 0,0195; y los SNPs rs528263420 y rs10679 con una población en desequilibrio con exceso de homocigotos (FIN e IBS respectivamente). Sus valores de varianza de Wahlund son 0,0801 y 0,0221. En el exón 3 tenemos el SNP rs1063478 con una población en desequilibrio con exceso de homocigotos (ITU, indios telugu) y una varianza de Wahlund de 0,0412.

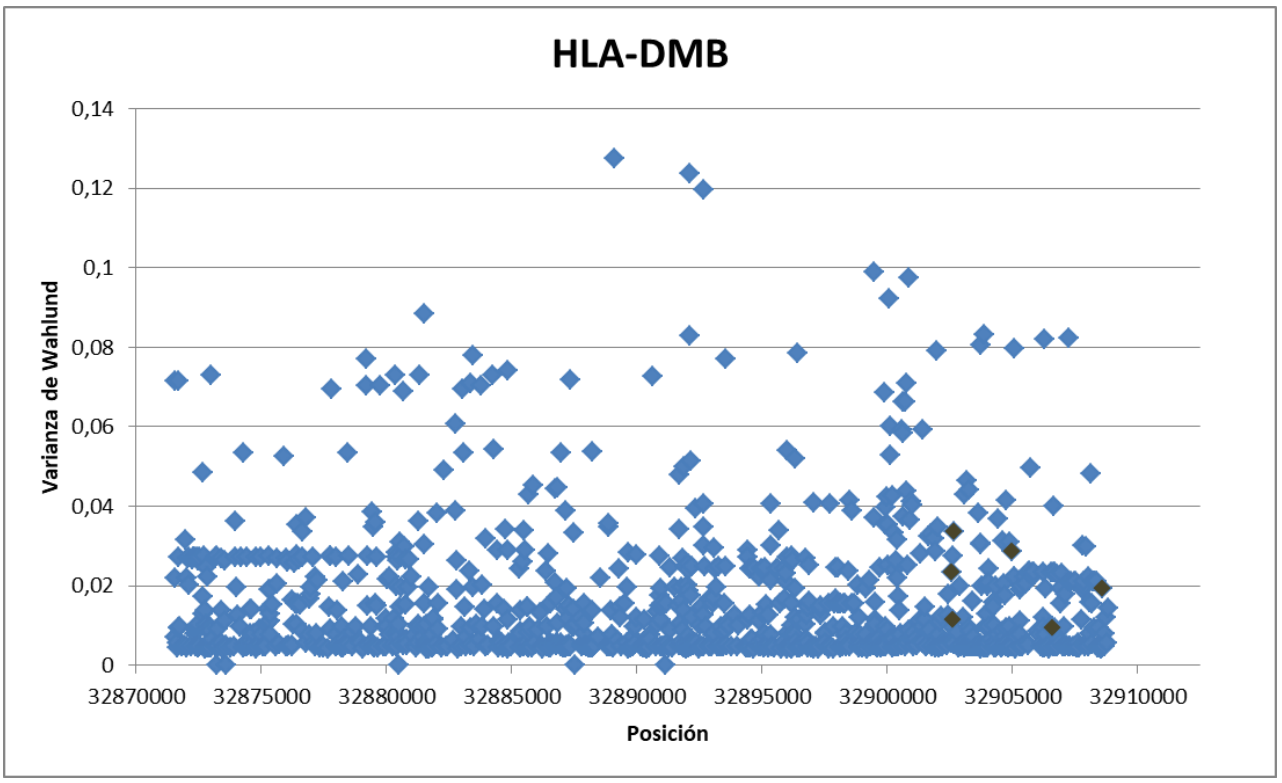


Figura R-27: Valores de la varianza de Wahlund de los SNPs del gen HLA-DMB en relación a su posición.

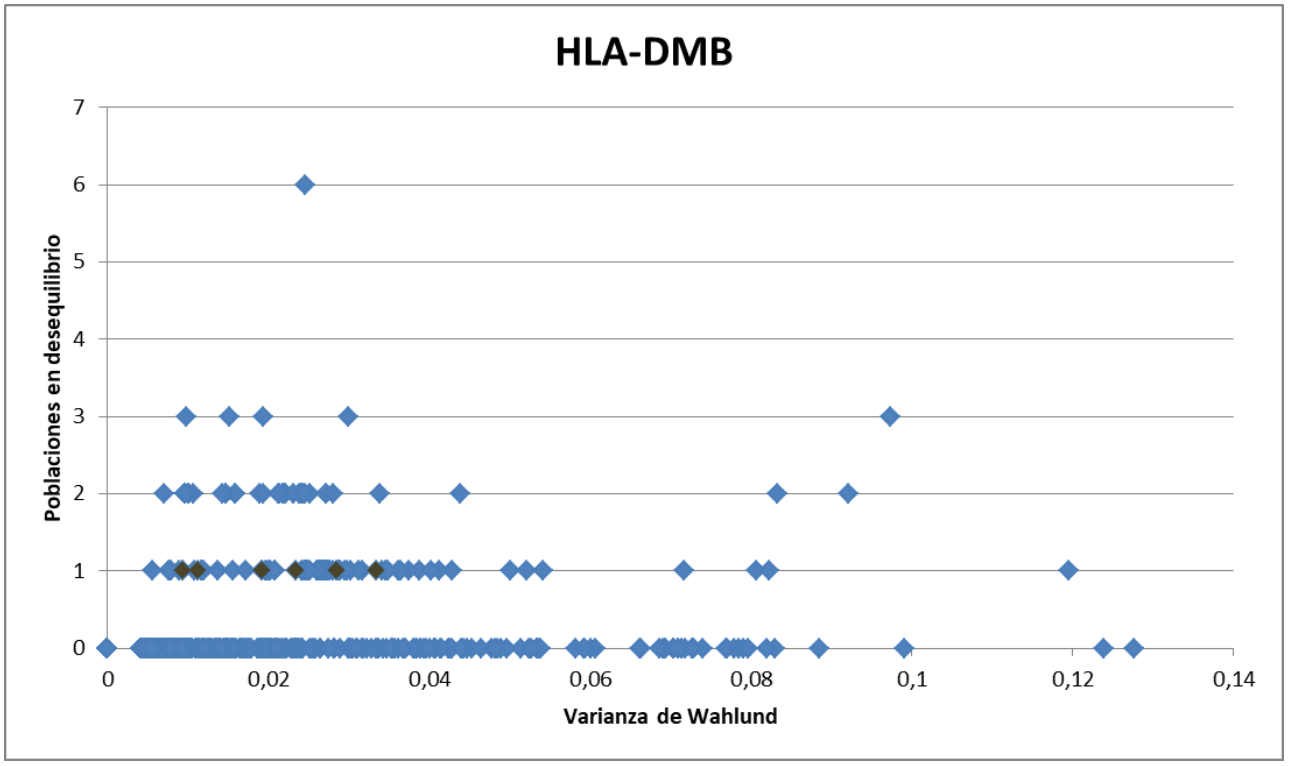


Figura R-28: Relación entre los valores de la varianza de Wahlund y el número de poblaciones en desequilibrio de Hardy-Weinberg para los SNPs del gen HLA-DMB.

En las figuras R-27 y R-28 se muestran los valores de la varianza de Wahlund de cada SNP en relación a su posición y el número de poblaciones en desequilibrio para el gen HLA-DMB.

No se han hallado SNPs con un valor de la varianza de Wahlund por encima de 0,1 ni en la región reguladora ni en ninguno de los 6 exones de HLA-DMB.

Entre los SNPs que muestran un número de poblaciones con desequilibrio de Hardy-Weinberg (marcados en marrón), se encuentran en la región reguladora 3´ los SNPs rs10751, rs12206426 y rs11540148 con una población en desequilibrio, todos con exceso de homocigotos, y una varianza de Wahlund de 0,0234, 0,0113 y 0,0334 respectivamente. En el exón 3 tenemos el SNP rs1042337 con una población en desequilibrio con exceso de heterocigotos (MXL, población de ancestría mexicana de Los Ángeles) y una varianza de Wahlund de 0,0285. En el exón 2 tenemos el SNP rs142107957 con una población en desequilibrio con exceso de homocigotos (PEL, peruanos de Lima) y una varianza de Wahlund de 0,0093. Y en la región reguladora 5´ tenemos el SNP rs149457045 con una población en desequilibrio con exceso de homocigotos (ITU, indios telugu) y una varianza de Wahlund de 0,0192.

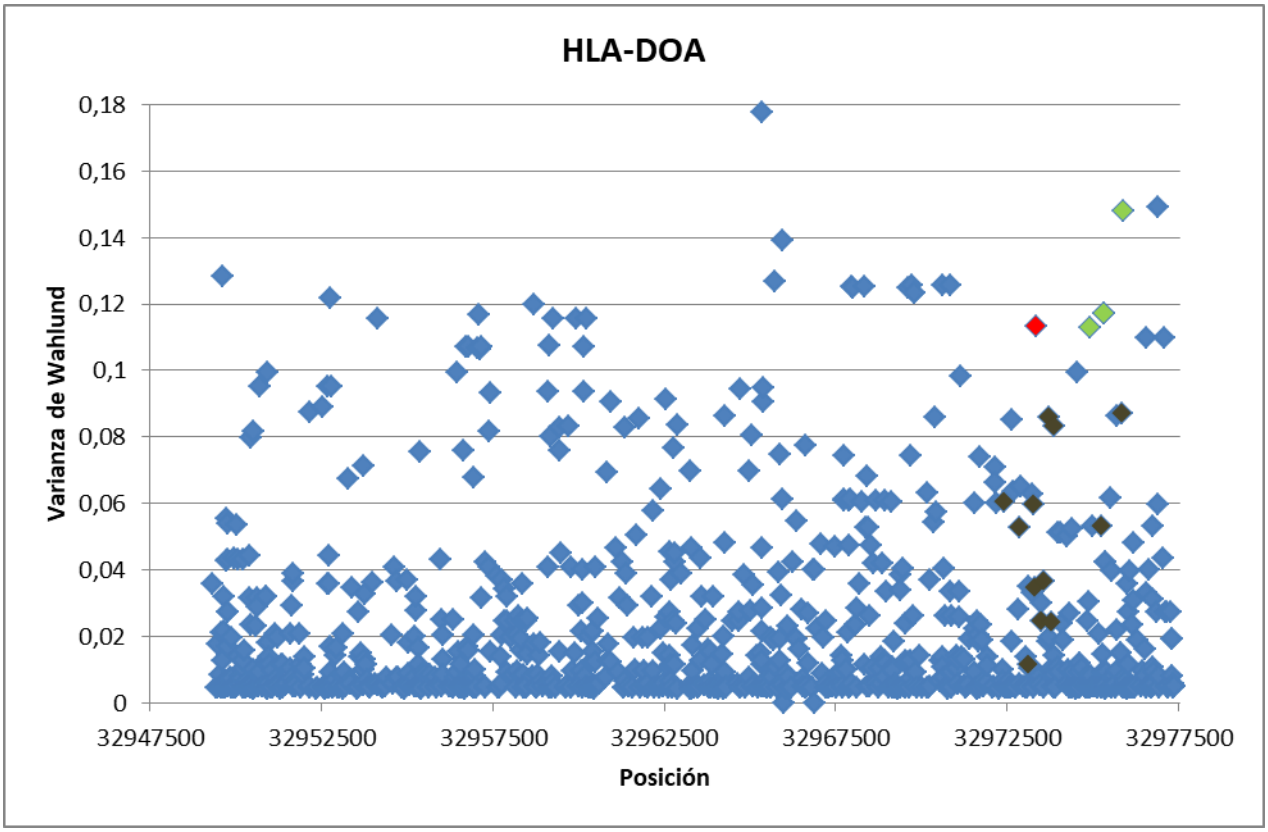


Figura R-29: Valores de la varianza de Wahlund de los SNPs del gen HLA-DOA en relación a su posición.

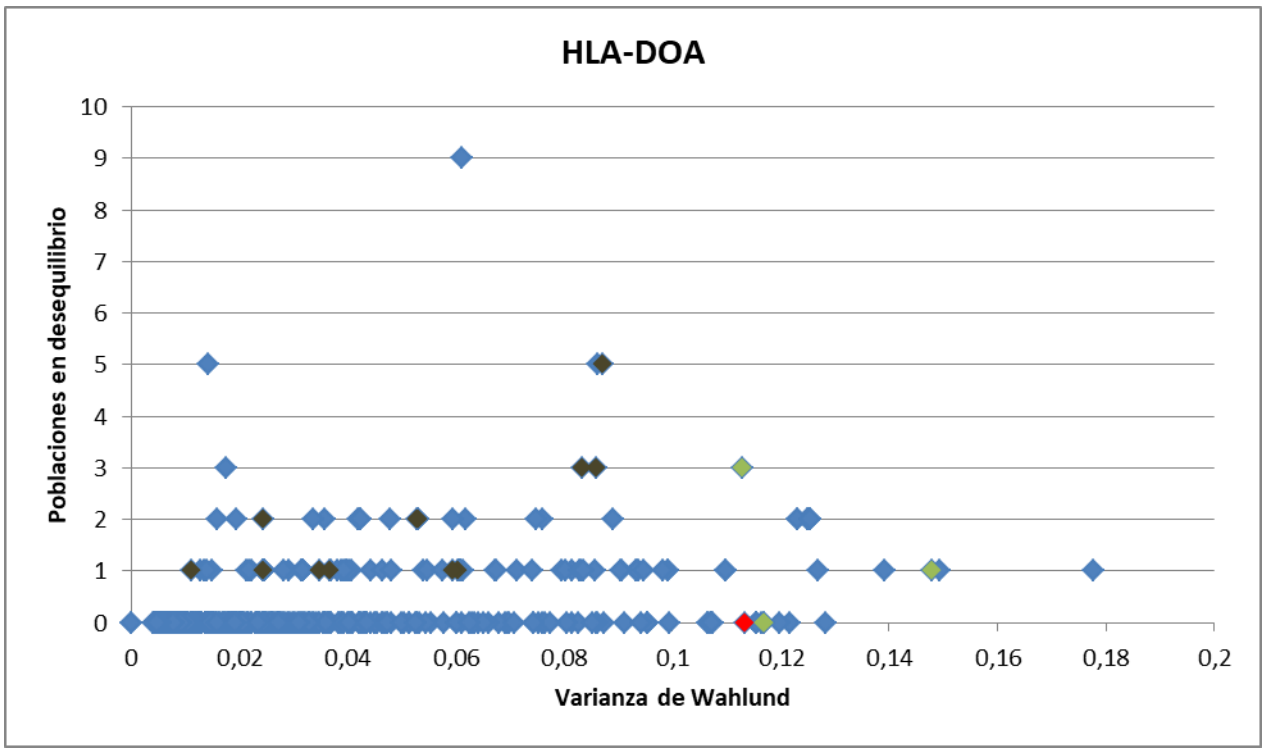


Figura R-30: Relación entre los valores de la varianza de Wahlund y el número de poblaciones en desequilibrio de Hardy-Weinberg para los SNPs del gen HLA-DOA.

En las figuras R-29 y R-30 se muestran los valores de la varianza de Wahlund de cada SNP en relación a su posición y el número de poblaciones en desequilibrio para el gen HLA-DOA.

Entre los SNPs con un valor de la varianza de Wahlund por encima de 0,1, se han encontrado uno en la región reguladora (rs142850513, marcado en rojo) y 3 (rs364950, rs10947368 y rs378352) en los exones 2, 3 y 4 del gen HLA-DOA respectivamente (marcados en verde). Ninguno de estos SNP con un valor de la varianza de Wahlund por encima de 0,1 ha mostrado desequilibrio Hardy-Weinberg en más de 3 poblaciones.

Entre los SNPs que muestran un número de poblaciones con desequilibrio de Hardy-Weinberg (marcados en marrón), se encuentran en la región reguladora 3´ los rs3128935, rs79804075, rs416622y rs71565360 con una población en desequilibrio, todos con exceso de homocigotos y unas varianzas de Wahlund de 0,0603, 0,0112, 0,0595 y 0,0347; y rs9276975 y rs9276976 con una población en desequilibrio con exceso de heterocigotos, y unas varianzas de Wahlund de 0,0366 y 0,0243. Igualmente en la región reguladora 3´ tenemos los SNPs rs376892 y rs410168 con 2 poblaciones en desequilibrio. Mientras que rs376892 tiene ambas poblaciones en desequilibrio con exceso de heterocigotos, en el caso de rs410168, una de la poblaciones (CDX, chinos Dai de Xishuangbanna) esta con exceso de homocigotos y otra (CLM, colombianos de Medellín) con exceso de heterocigotos. Sus varianzas de Wahlund son 0,0527 y 0,0245 respectivamente. Además en la misma región reguladora tenemos dos SNPs (rs3129304 y rs3129303) con 3 poblaciones en desequilibrio. En ambos casos dos de la poblaciones, ASW

(población afroamericana del suroeste de EEUU) y MXL (población de ancestría mexicana de Los Ángeles), están en desequilibrio con exceso de homocigotos y la tercera, CLM (colombianos de Medellín), está en desequilibrio con exceso de heterocigotos. Sus varianzas de Wahlund son 0,0859 y 0,0832 respectivamente. En el exón 4 tenemos el SNP rs378352 con 3 poblaciones en desequilibrio con exceso de homocigotos y una varianza de Wahlund de 0,1129. En el exón 3 tenemos el SNP rs365066 con dos poblaciones en desequilibrio: una por exceso de heterocigotos (BEB, bengalíes de Bangladesh) y otra por exceso de homocigotos (GIH, indios guyaratíes). Este SNP posee una varianza de Wahlund de 0,0531. Finalmente, en el exón 2, tenemos el SNP rs375256 con una varianza de Wahlund de 0,0872, y en desequilibrio en 5 poblaciones: ESN (Esan de Nigeria), CDX (chinos Dai de Xishuangbanna), CHS (chinos Han del sur de China) y MXL (población de ancestría mexicana de Los Ángeles) por exceso de homocigotos; y FIN (Fineses de Finlandia) por exceso de heterocigotos. También está el SNP rs364950 con una varianza de Wahlund de 0,148 y desequilibrio en una población (ASW, población afroamericana del suroeste de EEUU) por exceso de homocigotos.

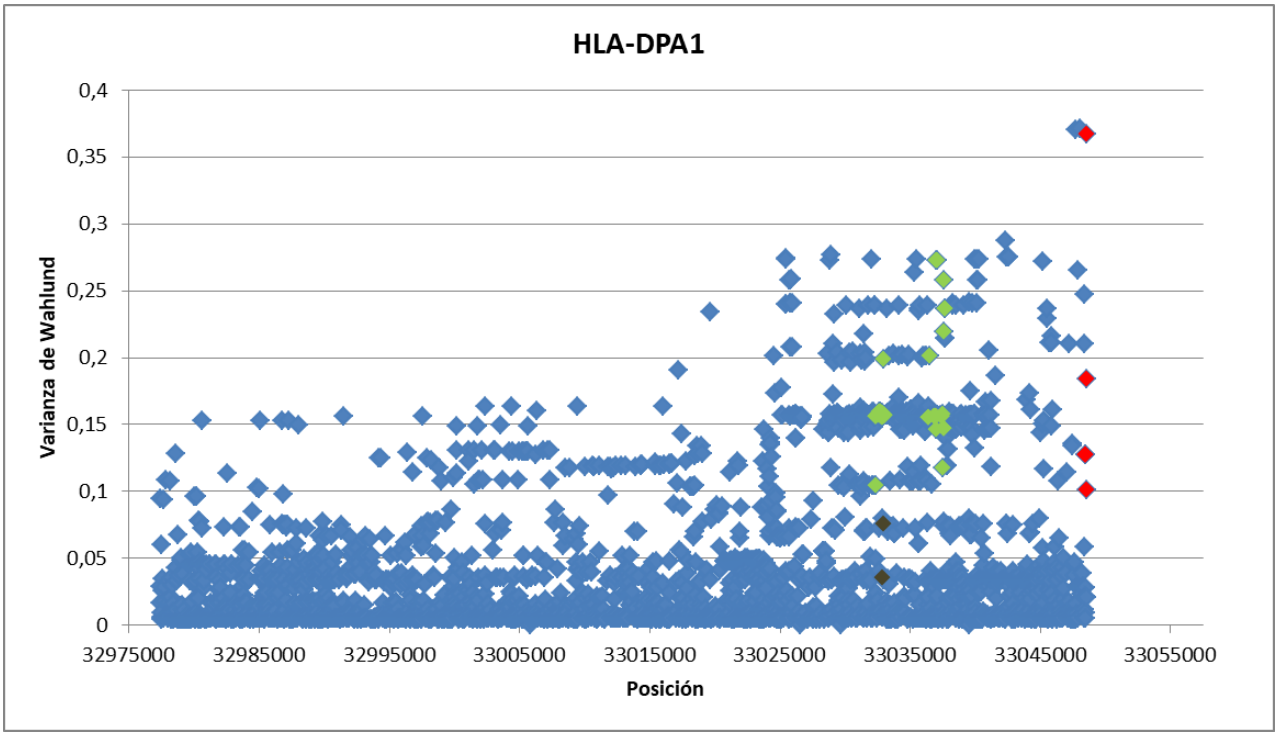


Figura R-31: Valores de la varianza de Wahlund de los SNPs del gen HLA-DPA1 en relación a su posición.

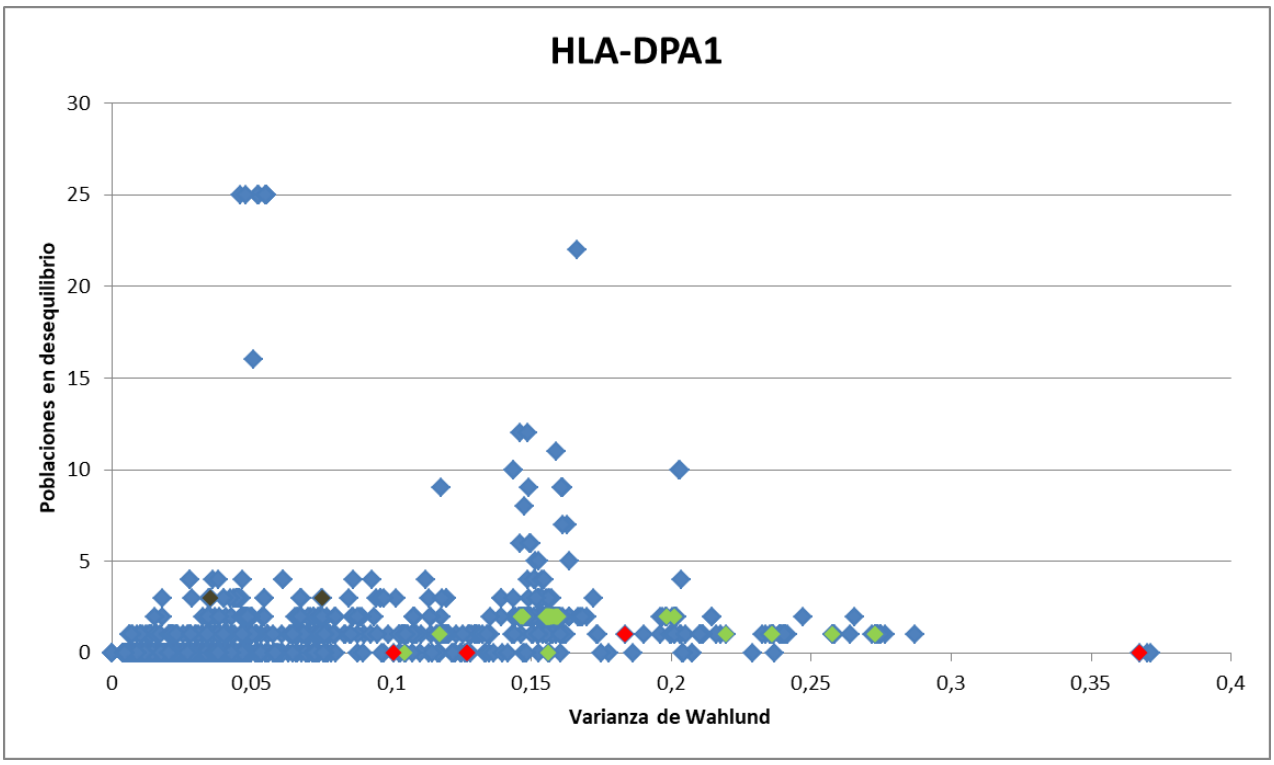


Figura R-32: Relación entre los valores de la varianza de Wahlund y el número de poblaciones en desequilibrio de Hardy-Weinberg para los SNPs del gen HLA-DPA1.

En las figuras R-31 y R-32 se muestran los valores de la varianza de Wahlund de cada SNP en relación a su posición y el número de poblaciones en desequilibrio para el gen HLA-DPA1.

Entre los SNPs con un valor de la varianza de Wahlund por encima de 0,1, se han encontrado 6 en la región reguladora (rs1126504, rs1126506, rs1126509, rs9277348, rs1042117 y rs1042121, marcados en rojo), 10 en el exón 2 (rs1126543, rs1126542, rs2308917, rs1042178, rs2308912, rs2308911, rs1062481, rs1042174, rs1126534 y rs1126533), 7 en el exón 3 (rs1042308, rs2308930, rs2308929, rs2308928, rs2308927, rs1042190 y rs1126544), 2 en el exón 2 (rs1126769 y rs1042434) y 26 en el exón 1 (rs17220927, rs17220934, rs6920294, rs374129814, rs17214555, rs17214562, rs17214567, rs2071366, rs17220961, rs17220968, rs17214573, rs17214580, rs17214587, rs17214594, rs2071365, rs1042926, rs1042920, rs1042908, rs1042901, rs1042872, rs1042866, rs361527, rs8486, rs3077, rs1042688 y rs3211475). Como en casos anteriores, los marcadores con varianza de Wahlund mayor que 0,1 que se encuentren en exones se han marcado en verde en los gráficos. Ninguno de estos SNP con un valor de la varianza de Wahlund por encima de 0,1 ha mostrado desequilibrio Hardy-Weinberg en más de 2 poblaciones.

Entre los SNPs que muestran un número de poblaciones con desequilibrio de Hardy-Weinberg (marcados en marrón), se encuentran en exón 1 los SNPs rs8807 y rs7905, con 3 poblaciones en desequilibrio, todos con exceso de homocigotos en el primer caso y con dos poblaciones en exceso de homocigotos (ESN, Esan de Nigeria, y CEU, población con ancestría del

norte y oeste de Europa) y una con exceso de heterocigotos (CLM, colombianos de Medellín). Sus varianzas de Wahlund son 0,0352 y 0,0752, respectivamente.

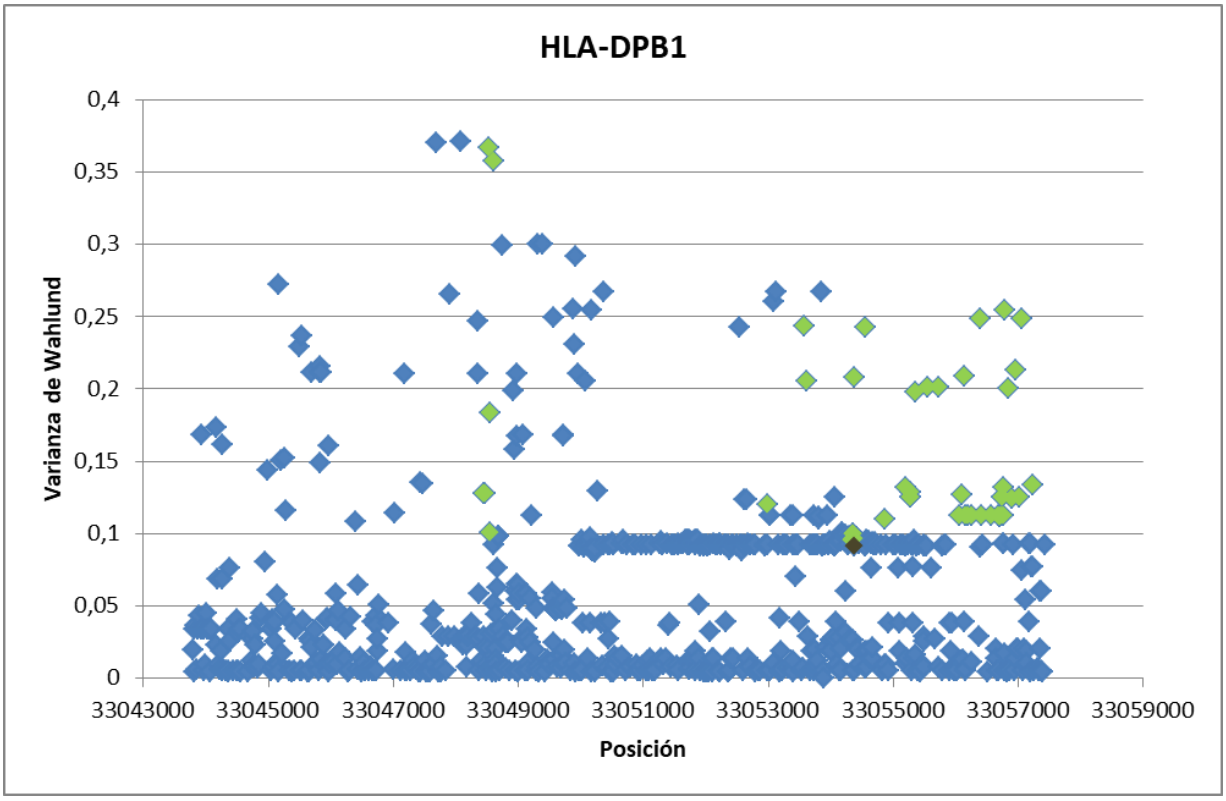


Figura R-33: Valores de la varianza de Wahlund de los SNPs del gen HLA-DPB1 en relación a su posición.

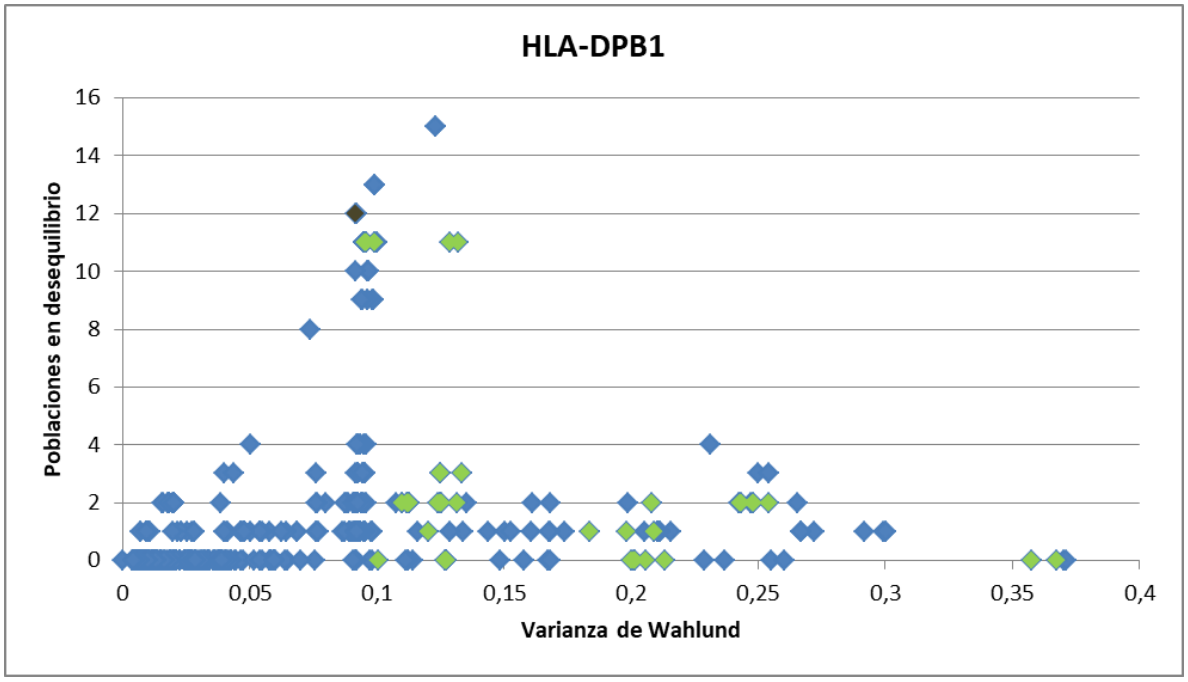


Figura R-34: Relación entre los valores de la varianza de Wahlund y el número de poblaciones en desequilibrio de Hardy-Weinberg para los SNPs del gen HLA-DPB1.

En las figuras R-33 y R-34 se muestran los valores de la varianza de Wahlund de cada SNP en relación a su posición y el número de poblaciones en desequilibrio para el gen HLA-DPB1.

Entre los SNPs con un valor de la varianza de Wahlund por encima de 0,1, se han encontrado 7 en el exón 2 (rs1126504, rs1126506, rs1126509, rs9277348, rs1042117, rs1042121 y rs9277351), uno en el exón 3 (rs14362), 2 en el exón 4 (rs9276 y rs11551421) y 33 en el exón 6 (rs6760, rs1042634, rs9277535, rs72500564, rs540086241, rs541802787, rs73740309, rs9501257, rs9501259, rs112170964, rs3117229, rs3128967, rs58649023, rs566978547, rs3130186, rs3128968, rs9461832, rs3130187, rs3091281, rs9277557, rs9277558, rs9277559, rs9277560, rs9277561, rs9277562, rs3117227, rs9296075, rs9296076, rs9277565, rs3097649, rs9277567, rs66953188 y rs3128970). Como en casos anteriores, los marcadores con varianza de Wahlund mayor que 0,1 que se encuentren en exones se han marcado en verde en los gráficos. Ninguno de estos SNP con un valor de la varianza de Wahlund por encima de 0,1 ha mostrado desequilibrio Hardy-Weinberg en más de 11 poblaciones, si bien solo rs72500564 y rs73740309 están en desequilibrio en 11 poblaciones, mientras que el resto de SNPs están en desequilibrio en 3 poblaciones o menos.

Entre los SNPs que muestran un número de poblaciones con desequilibrio de Hardy-Weinberg (marcados en marrón), se encuentra en exón 6 el SNP rs1042467, con 12 poblaciones en desequilibrio, todos con exceso de homocigotos, siendo su varianza de Wahlund igual a 0,0916.

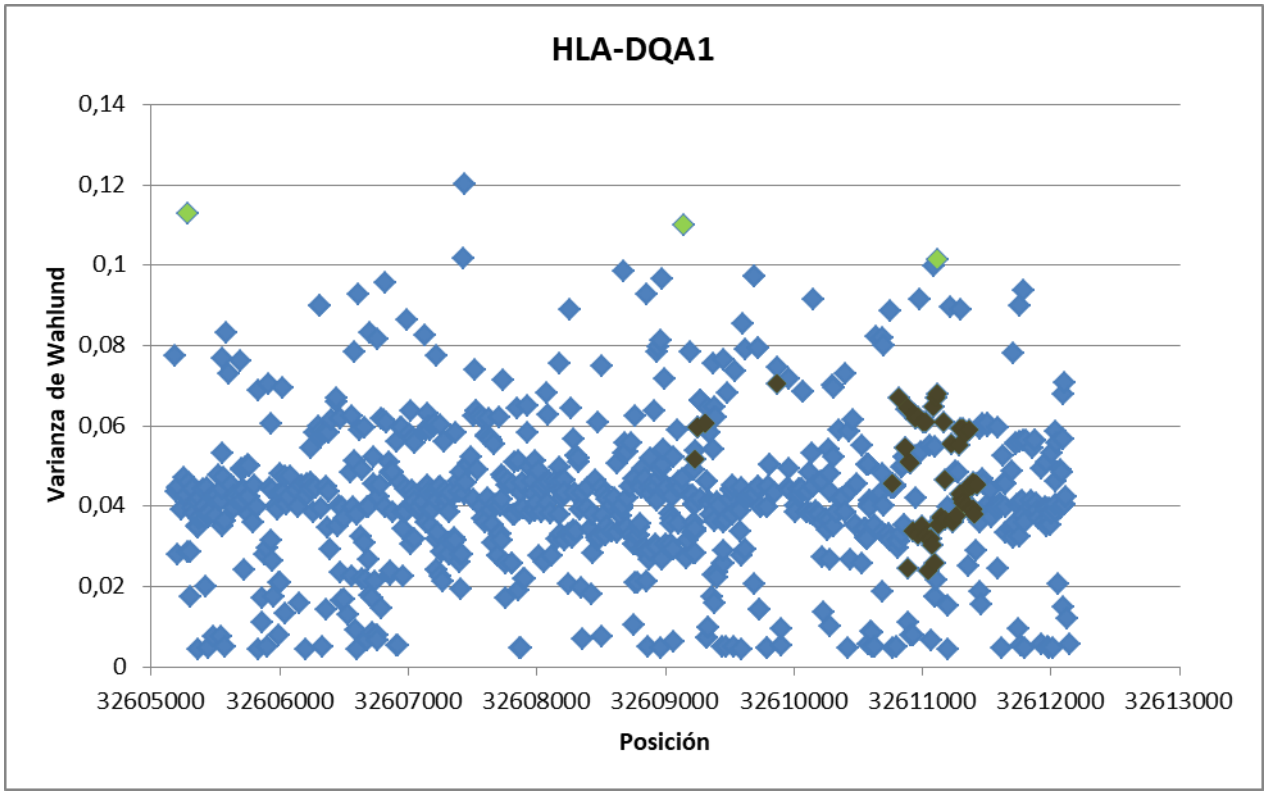


Figura R-35: Valores de la varianza de Wahlund de los SNPs del gen HLA-DQA1 en relación a su posición.

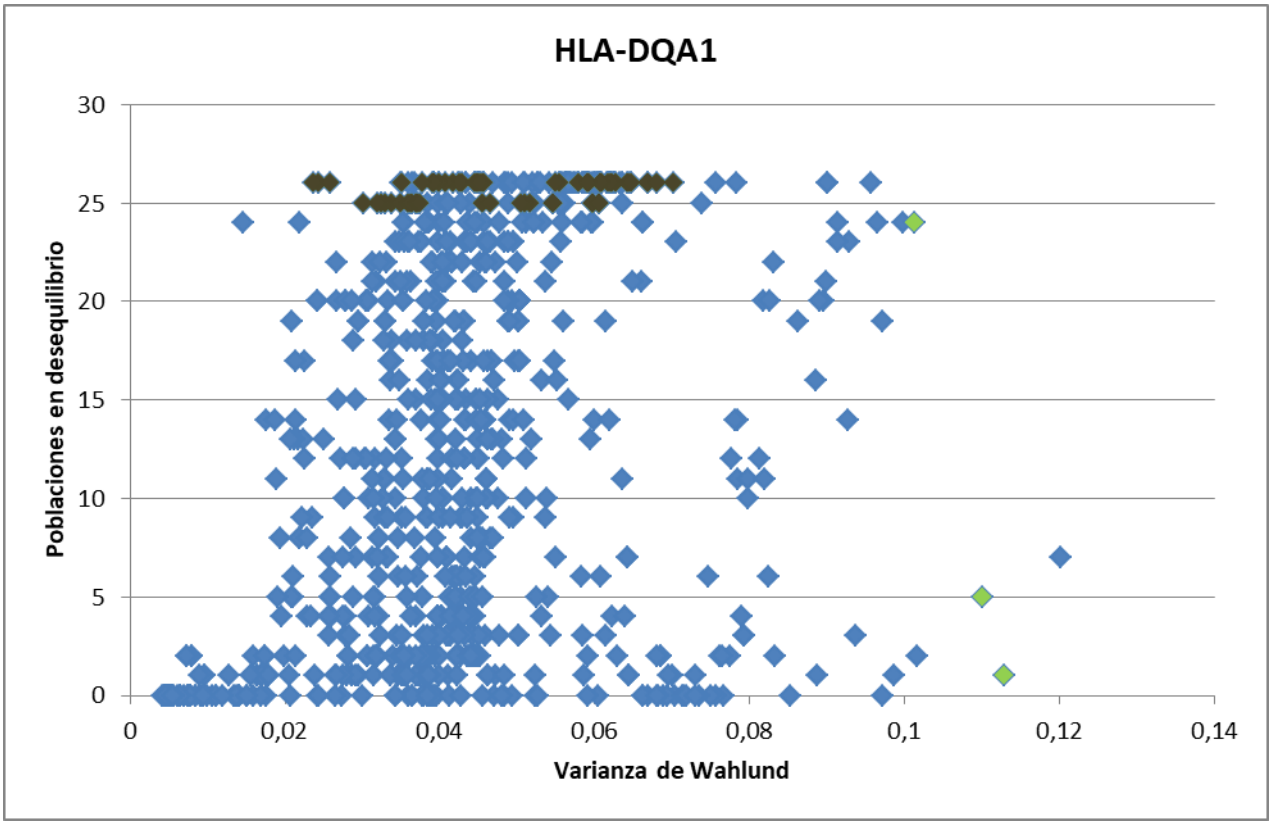


Figura R-36: Relación entre los valores de la varianza de Wahlund y el número de poblaciones en desequilibrio de Hardy-Weinberg para los SNPs del gen HLA-DQA1.

En las figuras R-35 y R-36 se muestran los valores de la varianza de Wahlund de cada SNP en relación a su posición y el número de poblaciones en desequilibrio para el gen HLA-DQA1.

Entre los SNPs con un valor de la varianza de Wahlund por encima de 0,1, se han encontrado 1 en el exón 1 (rs11545686), uno en el exón 2 (rs12722051) y uno en el exón 5 (rs62404112). Como en casos anteriores, los marcadores con varianza de Wahlund mayor que 0,1 que se encuentren en exones se han marcado en verde en los gráficos. Estos SNPs están en desequilibrio en 1, 5 y 24 poblaciones respectivamente.

Entre los SNPs que muestran un número de poblaciones con desequilibrio de Hardy-Weinberg en 25 poblaciones o más (marcados en marrón), se encuentran en el exón 2, rs4193, rs9272702, rs1142331, rs1142332 y rs1129808 con 25 poblaciones en desequilibrio. Se encuentran todos con exceso de homocigotos y un valor medio de varianza de Wahlund: 0,0517 para rs4193; 0,0597 para rs9272702, rs1142331 y rs1142332 (al tener el mismo valor exacto cabe esperar que se hereden ligados); y 0,0605 para rs1129808. En el exón 3 se encuentra rs707950 con 26 poblaciones en desequilibrio, todas con exceso de homocigotos, y un valor de varianza de Wahlund de 0,0702. En el exón 5 se encuentran 19 marcadores en desequilibrio en 25 poblaciones: rs7600, rs1130148, rs1142414, rs1142422, rs1142429, rs3667, rs1130156, rs1130162, rs1064985, rs707947, rs9272940, rs9272955, rs9272957, rs1064991, rs1064993, rs1065043, rs1065044, rs1065047 y rs1065048. Presentan valores de varianza de Wahlund variados, entre un mínimo de

0,0301 (correspondiente a rs9272940) y un máximo de 0,0545 (correspondiente a rs1130148 y rs1142414, que al tener el mismo valor exacto cabe esperar que se hereden ligados). El valor medio de la varianza de Wahlund para estos 19 SNPs es de 0,0398, y todos presentan exceso de homocigotos para las 25 poblaciones en las que están en desequilibrio. En el exón 5 también se encuentran 38 SNPs en desequilibrio en 26 poblaciones: rs7142, rs7143, rs9272917, rs9272918, rs1130151, rs1130152, rs1130155, rs1130158, rs1048726, rs1065036, rs28376734, rs9272943, rs28538060, rs9272948, rs9272950, rs9272951, rs9272958, rs9272962, rs9272969, rs9272970, rs9272971, rs9272972, rs9272973, rs9272974, rs9272975, rs9272976, rs9272978, rs9272980, rs9272981, rs9272982, rs9272983, rs9272984, rs9272985, rs9272986, rs9272987, rs9272988, rs9272989 y rs9272990. Presentan valores de varianza de Wahlund variados, entre un mínimo de 0,0237 (correspondiente a rs28376734) y un máximo de 0,0679 (correspondiente a rs9272950). El valor medio de la varianza de Wahlund para estos 19 SNPs es de 0,0504, y todos presentan exceso de homocigotos para las 26 poblaciones en las que están en desequilibrio.

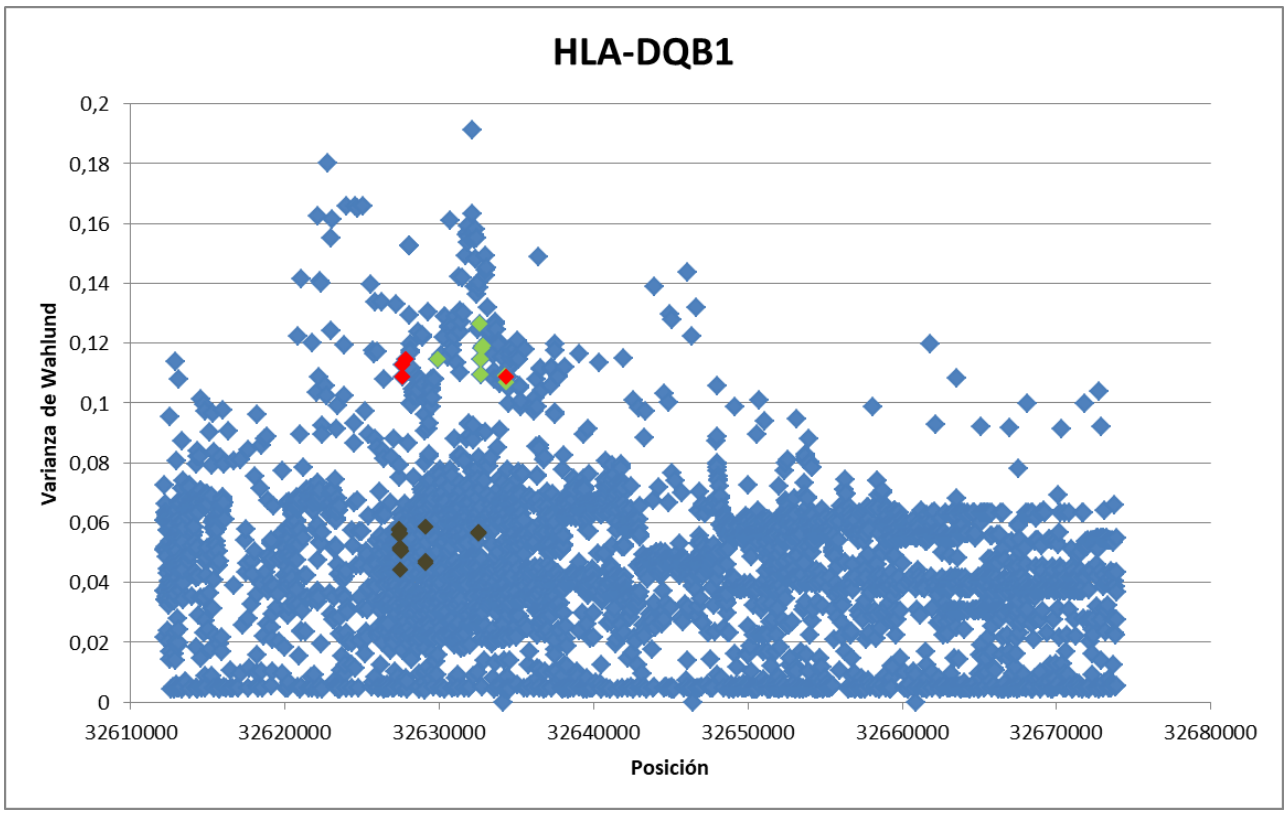


Figura R-37: Valores de la varianza de Wahlund de los SNPs del gen HLA-DQB1 en relación a su posición.

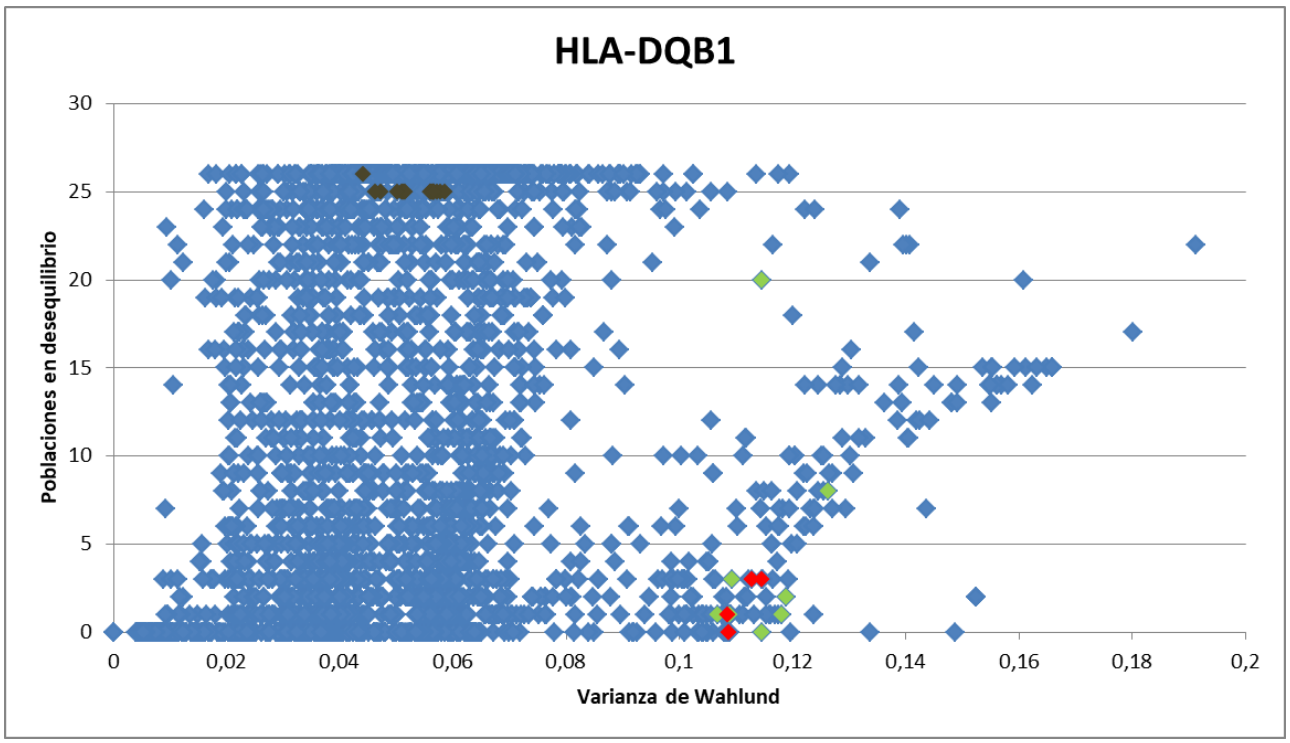


Figura R-38: Relación entre los valores de la varianza de Wahlund y el número de poblaciones en desequilibrio de Hardy-Weinberg para los SNPs del gen HLA-DQB1.

En las figuras R-37 y R-38 se muestran los valores de la varianza de Wahlund de cada SNP en relación a su posición y el número de poblaciones en desequilibrio para el gen HLA-DQB1.

Entre los SNPs con un valor de la varianza de Wahlund por encima de 0,1, se han encontrado 3 en la región reguladora 5´ (rs117789582, rs117067420 y rs113664950), uno en el exón 3 (rs41542812), 5 en el exón 2 (rs3204379, rs41552812, rs1049073, rs3204373 y rs12722107), 2 en el exón 1 (rs12722106 y rs36222416) y uno en la región reguladora 3´ (rs34236112). Como en casos anteriores, los marcadores con varianza de Wahlund mayor que 0,1 que se encuentren en exones se han marcado en verde en los gráficos, y lo que se encuentran en regiones reguladoras están marcados en rojo. Ninguno de estos SNPs está en desequilibrio en más de 3 poblaciones, salvo los casos de rs3204379 (8 poblaciones) y rs1049073 (20 poblaciones).

Entre los SNPs que muestran un número de poblaciones con desequilibrio de Hardy-Weinberg en 25 poblaciones o más (marcados en marrón), se encuentran en la región reguladora 5´ 10 SNPs (rs9273424, rs9273425, rs9273426, rs9273427, rs9273430, rs9273432, rs9273433, rs9273436, rs9273438 y rs28724234) con 25 poblaciones en desequilibrio y un SNP (rs1130451) con 26 poblaciones en desequilibrio. Se encuentran todos con exceso de homocigotos. Entre los 10 SNPs con desequilibrio en 25 poblaciones destaca el hecho de que en todas la única población que no está en desequilibrio es KHV (Kinh de Vietnam), y que presentan valores de varianza de Wahlund variados aunque moderados para el conjunto de SNPs de HLA-DQB1, entre un mínimo de 0,0502 (correspondiente a rs28724234) y un máximo de 0,0577 (correspondiente a rs9273427). El valor medio de la varianza de Wahlund para estos 10 SNPs es de 0,0538. Para el caso de rs1130451

el valor de la varianza de Wahlund es algo más bajo: 0,0441. En el exón 4 se encuentran rs1130432, rs1130431 y rs1130430 con 25 poblaciones en desequilibrio, todas con exceso de homocigotos, y un valor de varianza de Wahlund de 0,0472, 0,0463 y 0,0585 respectivamente. En el exón 5 se encuentran 5 marcadores en desequilibrio en 25 poblaciones: rs1140322, rs1140321, rs1140320, rs1140319 y rs1140318. Presentan valores de varianza de Wahlund variados aunque en un rango relativamente estrecho, entre un mínimo de 0,05639 (correspondiente a rs1140320) y un máximo de 0,05656 (correspondiente a rs1140319 y rs1140318, que al tener el mismo valor exacto cabe esperar que se hereden ligados). El valor medio de la varianza de Wahlund para estos 5 SNPs es de 0,05650, y todos presentan exceso de homocigotos para las 25 poblaciones en las que están en desequilibrio, siendo la población JPT (Japoneses de Tokio) la única para la que no están en desequilibrio.

Marcadores de ancestralidad

Se han encontrado AIMs (*Ancestry Informative Marker*, Marcadores de ancestralidad) en todos los genes y, en muchos casos, ha sido posible asignarlos a un patrón de distribución continental (Tabla R-9).

GENES	SNPs	AIMs	% AIMs	ÁFRICA	EUROPA	ASIA	SIN PATRÓN
HLA-A	475	2	0,421	1	0	1	0
HLA-B	4702	120	2,552	11	0	32	77
HLA-C	3139	11	0,350	3	0	4	4
HLA-DRA	1277	3	0,235	1	2	0	0
HLA-DMA	731	7	0,958	0	0	2	5
HLA-DMB	1046	6	0,574	0	0	1	5
HLA-DOA	894	27	3,020	17	0	0	10
HLA-DPA1	2885	626	21,698	36	438	15	137
HLA-DPB1	782	334	42,711	29	245	36	24
HLA-DQA1	917	4	0,436	0	0	0	4
HLA-DQB1	4388	42	0,957	0	0	20	22

Tabla R-9: Resumen de la presencia de AIMs en los genes estudiados. Se incluyen los SNPs totales, el número de AIMs, el porcentaje de SNPs que son AIMs, así como los AIMs que presentan patrón de distribución continental, y los que no lo presentan para cada uno de los genes estudiados.

Del total de 21236 SNPs encontrados en los 11 genes estudiados, 1182 son marcadores informativos de ancestralidad, es decir, marcadores que exhiben frecuencias sustancialmente diferentes entre diferentes poblaciones continentales, con al menos una diferencia de 0,3 entre dos de ellas. De estos AIMs, 98 son representativos de ancestría africana, 685 de ancestría europea y 111 de ancestría asiática. 288 AIMs no han podido ser asociados a un patrón de distribución continental.

De los grupos continentales, Asia es el que tiene AIMs en más genes (todos salvo HLA-DRA, HLA-DOA y HLA-DQA1), seguido de África (todos salvo HLA-DMA, HLA-DMB, HLA-DQA1 y HLA-DQB1). En lo referente a Europa llama la atención la dicotomía que presenta: únicamente tiene AIMs para HLA-DRA, HLA-DPA1 y HLA-DPB1, y sin embargo, tanto en HLA-DPA1 como en HLA-DPB1, presenta varios cientos de AIMs, agrupando más de la mitad de los mismos en cada uno de esos genes.

En cuanto a genes, cabe destacar que HLA-DPA1 y HLA-DPB1 son los únicos que presentan AIMs con los 3 patrones de distribución continental, y además son los genes que más AIMs presentan, seguidos de HLA-B. Únicamente HLA-DQA1 y HLA-DQB1 no tienen ningún AIM asociable a algún patrón continental. En cuanto al porcentaje de SNPs que son AIMs entre el total de SNPs en cada gen, HLA-DPB1 (42,711%) y HLA-DPA1 (21,698%) son los que más proporción de AIMs presentan entre sus SNPs, seguidos de lejos por HLA-DOA (3,02%) y HLA-B (2,552%).

Diferencias en los procesos de selección

En cuanto a los patrones de selección se han realizado varios análisis teniendo en cuenta las frecuencias alélicas de los tríos de las poblaciones HapMap. Así, para cada uno de los 11 genes estudiados se han realizado una tabla donde se indica el número de SNPs que presentan desequilibrio Hardy-Weinberg para las frecuencias alélicas del total de la población (HW p1), para las frecuencias alélicas de los individuos adultos progenitores (HW p2), para los individuos adultos no progenitores (HW p3), para los descendientes que forman parte de tríos (HW p4); y aparte se han realizado varios test de Fisher a fin de determinar si existe diferentes significativas en las frecuencias alélicas en cada marcador para los distintos cruces de las subpoblaciones mencionadas: total vs progenitores (Fisher 12), total vs no progenitores (Fisher 13), progenitores vs no progenitores (Fisher 23), total vs descendientes (Fisher 14), progenitores vs descendientes (Fisher 24) y no progenitores vs descendientes (Fisher 34). Además se ha incluido un análisis de Chi2, a fin de averiguar si existen diferencias significativas entre el número de genotipos observados y los esperados por azar.

TEST	POBLACIONES			
	YRI	ASW	CEU	MXL
HW p1	3	2	2	2
HW p2	3	2	2	2
HW p3	0	5	0	0
HW p4	0	0	0	0
Fisher p12	0	0	0	0
Fisher p13	0	0	0	0
Fisher p23	1	0	0	0
Fisher p14	0	0	0	0
Fisher p24	0	0	0	0
Fisher p34	0	0	0	0
Chi2	1	0	0	1

Tabla R-10: Número de SNPs que muestran significación para los test estadísticos realizados en el gen HLA-A.

En la tabla R-10 podemos observar la información obtenida para el gen HLA-A. Se han hallado 4 SNPs distintos que presentan desequilibrio Hardy-Weinberg en el conjunto total de las poblaciones de la base de datos HapMap. En la población YRI (Yoruba de Ibadán, Nigeria) se han identificado rs139109856, rs29028878 y rs41561516, todos en el intrón 3. Estos SNPs presentan valores de varianza de Wahlund de 0,0562, 0,0463 y 0,0258 respectivamente. Para la población ASW (población afroamericana de suroeste de EEUU) se han identificado rs29028878 y rs2508038, ambos en el intrón 3, con varianzas de Wahlund de 0,0463 y 0,0146. Para las poblaciones CEU (población con ancestría del norte y oeste de Europa) y MXL (ancestría mexicana de Los Ángeles, EEUU) se han identificado los SNPs rs29028878 y rs41561516 (ambas en el intrón 3) como en desequilibrio Hardy-Weinberg para el total de la población, siendo sus varianzas de Wahlund 0,0463 y 0,0258. Como podemos observar, hay SNPs que muestran desequilibrio en varias poblaciones: rs29028878 en todas, y rs41561516 en YRI, CEU y MXL. En ningún caso ha sido posible asociar ningún marcador a algún patrón de distribución continental.

Para el conjunto poblacional de los adultos progenitores se han identificado 9 SNPs en desequilibrio Hardy-Weinberg. Para las poblaciones YRI, ASW y CEU los SNPs identificados son los mismos que en conjunto total de la población. Para la población MXL se han identificado los SNPs rs139109856 y rs29028878. Del mismo modo que en el caso anterior, hay SNPs que están en desequilibrio en varias poblaciones, y en ningún caso ha sido posible asociar ningún marcador a algún patrón de distribución continental.

En el caso de los adultos no progenitores, se han identificado 5 SNPs en desequilibrio y únicamente en la población ASW: rs41559412, rs41553212, rs29028878, rs2508038 y rs41551518. Los SNPs rs41559412, rs41553212 y rs41551518 presentan unos valores de varianza de Wahlund de 0,0411, 0,0398 y 0,0412 respectivamente, y en los tres casos se ha asociado a un patrón de distribución africano.

No se han hallado resultados significativos en los test de Hardy-Weinberg para el conjunto poblacional de los descendientes, ni para los test de Fisher, salvo en el caso de Fisher p23 para la población YRI, donde el SNP rs560554218 presenta diferencias alélicas significativas entre el conjunto de individuos adultos progenitores y no progenitores. Este SNP presenta una varianza de Wahlund de 0,0088, y no ha sido posible asociarlo a un patrón de distribución continental.

En cuanto a los test de Chi2, se ha determinado que los SNPs rs12721675 para la población YRI y rs572407231 para MXL presentan diferencias significativas entre los genotipos observados y los esperados por azar. Estos SNPs presentan valores de varianza de Wahlund de 0,0329 y 0,0273, y no ha sido posible asignar su patrón de distribución a ningún continente.

TEST	POBLACIONES			
	YRI	ASW	CEU	MXL
HW p1	18	9	25	3
HW p2	7	2	27	4
HW p3	3	3	0	0
HW p4	1	0	0	0
Fisher p12	0	0	0	0
Fisher p13	2	0	0	7
Fisher p23	3	0	5	12
Fisher p14	0	0	0	0
Fisher p24	0	0	0	0
Fisher p34	0	0	0	0
Chi2	0	0	1	0

Tabla R-11: Número de SNPs que muestran significación para los test estadísticos realizados en el gen HLA-B.

En la tabla R-11 podemos observar la información obtenida para el gen HLA-B. Se han hallado 36 SNPs distintos que presentan desequilibrio Hardy-Weinberg en el conjunto total de las poblaciones de la base de datos HapMap, localizados en exones (como por ejemplo rs1140546 en el exón 5 o rs3180379 en el 2), intrones (como rs16899066 en el intrón 3) y la región reguladora 5´(como rs9266207); y 32 SNPs distintos que presentan desequilibrio Hardy-Weinberg en el conjunto de adultos progenitores, localizados en exones (como por ejemplo rs1140546 en el exón 5 o rs1050543 en el 2), intrones (como rs16899066 en el intrón 3) y la región reguladora 5´(como rs9266207).

En el caso de los adultos no progenitores, se han identificado 6 SNPs en desequilibrio: rs1131212, rs41562914 y rs9266178 para la población YRI localizados los tres en el exón 2 de HLA-B; y rs1140412, rs1065386 y rs1050570 para la población ASW, localizados en el exón 3 en el primer caso y en el 2 en los dos últimos. Sus varianzas de Wahlund son 0,0501, 0,0474 y 0,0359; y 0,0311, 0,0429 y 0,0429. Para el caso de los SNPs de la población YRI no ha sido posible asignar un

patrón continental, pero en el caso del SNP rs1140412 se ha establecido que posee un patrón de distribución asiático.

Para el conjunto de descendientes, se ha hallado un SNP en la población YRI en desequilibrio: rs2596494, con una varianza de Wahlund de 0,0296 y localizado en el intrón 3 de HLA-B.

No se han hallado resultados significativos para los test de Fisher 12, 14, 24 y 34. En el test conjunto total frente a no progenitores se han identificado 2 SNPs en la población YRI: rs3819292 (localizado en el intrón 5) y rs4999718 (localizado en el intrón 2). Sus varianzas de Wahlund son 0,0324 y 0,0314, y en el primer caso es posible asignarle un patrón continental europeo. También se han identificado 7 SNPs en la población MXL: rs1058067 (en la región reguladora 3´), rs3819294 (en el intrón 5), rs17193012 (en el intrón 5), rs12526858 (en el intrón 3), rs12528645 (en el intrón 3), rs3179865 (en el exón 3) y rs4999717 (en el intrón 2). Salvo en el caso de rs3179865, todos estos SNPs poseen un patrón de distribución continental asiático. Las varianzas de Wahlund para el primer, sexto y séptimo SNP son 0,1026, 0,0315 y 0,0684. Para el resto, el hecho de que posean una varianza de Wahlund igual (0,0658) y que se encuentren muy próximos hace sospechar que se heredan ligados. En el test adultos progenitores frente a no progenitores se han identificado 3 SNP en la población YRI: rs3819292 (localizado en el intrón 5), rs4999718 (localizado en el intrón 2) y rs9266198 (localizado en el intrón 1). Sus varianzas de Wahlund son 0,0324, 0,0314 y 0,0211, y en el primer caso es posible asignarle un patrón continental europeo. También se han identificado 5 SNPs en la población CEU, todos localizados en el exón 2 de HLA-B: rs1131214, rs1140404, rs41541616, rs41543121 y rs1131202. Sus varianzas de Wahlund son 0,0189, 0,0319, 0,0292, 0,0295 y 0,0599; y en ningún caso ha sido posible asignar un patrón de distribución continental. De los 12 SNPs localizados para la población MXL, cabe destacar que se encuentran repartidos por todo el gen. Así, rs1058067 y rs1056429 se encuentran en la región reguladora 3´, rs3819294 se encuentra en el intrón 5, rs1050517 está en el exón 2,...

En cuanto a los test de Chi2, se ha determinado que el SNP rs709055 para la población CEU presenta diferencias significativas entre los genotipos observados y los esperados por azar. Este SNP presenta un valor de varianza de Wahlund de 0,0161, y no ha sido posible asignar su patrón de distribución a ningún continente.

TEST	POBLACIONES			
	YRI	ASW	CEU	MXL
HW p1	30	3	27	16
HW p2	19	4	29	11
HW p3	8	1	2	1
HW p4	0	0	0	0
Fisher p12	0	0	0	0
Fisher p13	0	0	1	2
Fisher p23	1	4	2	2
Fisher p14	2	0	11	0
Fisher p24	3	0	8	0
Fisher p34	2	0	8	0
Chi2	1	0	0	0

Tabla R-12: Número de SNPs que muestran significación para los test estadísticos realizados en el gen HLA-C.

En la tabla R-12 podemos observar la información obtenida para el gen HLA-C. Se han hallado 35 SNPs distintos que presentan desequilibrio Hardy-Weinberg en el conjunto total de las poblaciones de la base de datos HapMap, localizados en la región reguladora 3´ (como rs67157575 o rs1130558), exones (como por ejemplo rs41542414 o rs34794906 en el exón 5) e intrones (como rs2523609 en el intrón 6); y 38 marcadores distintos que presentan desequilibrio Hardy-Weinberg en el conjunto de adultos progenitores, localizados en la región reguladora 3´ (como rs3189472 o rs572896906), exones (como por ejemplo rs1130838 o rs34794906 en el exón 5) e intrones (como rs68037221 en el intrón 5).

En el caso de los adultos no progenitores, se han identificado 11 SNPs en desequilibrio: rs572896906 (región reguladora 3´), rs1130586 (región reguladora 3´), rs3207555 (región reguladora 3´), rs41542414 (exón 5), rs1050105 (exón 5), rs9264626 (intrón 4), rs9264627 (intrón 4) y rs2308628 (exón 4) para la población YRI; rs3189472 (región reguladora 3´) para la población ASW; rs1130554 y rs1130558 para la población CEU, ambos en la región reguladora 3´; y rs3189472 para la población CEU, localizado también en la región reguladora 3´. Sus varianzas de Wahlund son no alcanzan en ningún caso un valor de 0,1. Los marcadores rs1130554 y rs1130558 presentan la misma varianza de Wahlund, por lo que cabe esperar que se hereden ligados. Para el caso de los SNPs rs1050105 se ha establecido que posee un patrón de distribución asiático.

No se han hallado resultados significativos para los test HW p4 ni Fisher 12. En el test conjunto total frente a no progenitores se han identificado 1 SNP en la población CEU, rs17885557 (localizado en el intrón 7). Su varianza de Wahlund es 0,0662, y no es posible asignarle un patrón continental. También se han identificado 2 marcadores en la población MXL: rs373826500 (en el intrón 5) y rs11757919 (en el exón 5). Sus varianzas de Wahlund son 0,0126 en ambos casos, lo que junto a su posición, sugiere que se heredan ligados. En el test adultos progenitores frente a no progenitores se han identificado 1 SNP en la población YRI: rs41542414 (localizado en el exón 5),

siendo su varianza de Wahlund 0,0599, y no siendo posible asignarle un patrón continental. También se han identificado 4 SNPs en la población ASW, todos localizados en el exón 2 de HLA-B: rs1130559 (región reguladora 3´), rs17879195 (intrón 2), rs17880655 (intrón 2) y rs2074491 (región reguladora 5´). Sus varianzas de Wahlund son 0,0357, 0,1308, 0,1304 y 0,0376; y en el caso de los dos marcadores intrónicos presentan un patrón de distribución continental asiático. Se han detectado dos SNPs para la población CEU: rs17885557 (intrón 7) y rs1131118 (exón 3). Sus varianzas de Wahlund son 0,0662 y 0,0138, pero en ningún caso ha sido posible asignar un patrón continental. Para la población MXL se han detectado los marcadores rs373826500 (intrón 5) y rs11757919 (exón 5). Sus varianzas de Wahlund son 0,0126 en ambos casos, lo que hace sospechar que existe un desequilibrio de ligamiento. En ningún caso ha sido posible asignar un patrón continental. Para el test de Fisher entre el conjunto total de la población frente a los descendientes se han identificado 2 SNPs para YRI y 11 marcadores para CEU. En el caso de la población YRI, los SNP son rs9264648 (intrón 3) y rs2074491 (región reguladora 5´), siendo sus varianzas de Wahlund 0,0079 y 0,0376 respectivamente, aunque no se ha hallado ningún patrón continental asociable a ninguno de los dos. De los 11 marcadores de CEU destacar que todos se encuentran en intrones, salvo rs7767581 y rs2074491, que se encuentran en la región reguladora 5´; sin embargo no es posible asociar ninguno de los 11 marcadores a un patrón continental. Para el test de Fisher entre el conjunto de adultos progenitores frente a los descendientes se han identificado 3 SNPs para YRI y 8 marcadores para CEU. En el caso de la población YRI, los SNP son rs2001181 (intrón 7), rs9264648 (intrón 3) y rs2074491 (región reguladora 5´). Sus varianzas de Wahlund son 0,0285, 0,0079 y 0,0376, y no se ha podido caracterizar ningún grupo continental para ninguno de los tres. Los 8 marcadores detectados en CEU son rs9264594 (intrón 7), rs9264601 (intrón 5), rs9264603 (intrón 5), rs9264606 (intrón 5), rs9264636 (intrón 3), rs9264638 (intrón 3 y patrón de distribución asiático), rs9264648 (intrón 3) y rs2074491 (región reguladora 5´). Ninguna de las varianzas de Wahlund de estos 8 marcadores alcanza un valor de 0,1. Para el test de Fisher entre el conjunto de adultos no progenitores frente a los descendientes se han identificado 2 SNPs para YRI y 8 marcadores para CEU. En el caso de la población YRI, los SNP son los mismos que en el caso del test Fisher entre el conjunto de adultos progenitores frente a los descendientes, salvo que rs2001181 (intrón 7) no está presente. Los 8 marcadores detectados en CEU son los mismos que en el caso del test Fisher entre el conjunto de adultos progenitores frente a los descendientes.

En cuanto a los test de Chi2, se ha determinado que el SNP rs3189472 para la población YRI presenta diferencias significativas entre los genotipos observados y los esperados por azar. Este SNP, situado en la región reguladora 3´, presenta un valor de varianza de Wahlund de 0,0301, y no ha sido posible asignar su patrón de distribución a ningún continente.

TEST	POBLACIONES			
	YRI	ASW	CEU	MXL
HW p1	0	1	0	0
HW p2	0	13	0	0
HW p3	0	0	0	0
HW p4	0	0	1	0
Fisher p12	0	0	0	0
Fisher p13	24	0	9	0
Fisher p23	25	0	12	0
Fisher p14	1	1	3	1
Fisher p24	1	1	3	1
Fisher p34	12	1	9	1
Chi2	0	2	0	0

Tabla R-13: Número de SNPs que muestran significación para los test estadísticos realizados en el gen HLA-DRA.

En la tabla R-13 podemos observar la información obtenida para el gen HLA-DRA. Se ha hallado 1 SNP que presentan desequilibrio Hardy-Weinberg en el conjunto total de la población ASW: rs3129877, localizado en el intrón 1, posee una varianza de Wahlund 0,0514 y no ha sido posible asociarlo a un patrón de distribución continental.

En el caso de los adultos progenitores, se han identificado 11 marcadores en desequilibrio en la población ASW: rs14004 (región reguladora 5´), rs3129876 (intrón 1), rs3129879 (intrón 1), rs9268657 (intrón 1), rs3129884 (intrón 1), rs3129885 (intrón 1), rs2239806 (intrón 3), rs532328824 (intrón 4), rs539718048 (intrón 4), rs115317719 (intrón 4), rs1131541 (región reguladora 3´), rs1051336 (región reguladora 3´) y rs1041885 (región reguladora 3´). En ningún caso los valores de la varianza de Wahlund alcanzan 0,1 ni es posible asignar un patrón de distribución continental.

No se han hallado resultados significativos para los test HW p3, por lo que cabe decir que no se ha hallado diferencias significativas entre las frecuencias alélicas observadas y esperadas para el conjunto de adultos no progenitores en el gen HLA-DRA.

En lo referente al conjunto de descendientes se ha hallado un resultado significativo para la población CEU: el SNP rs11544315, localizado en el exón 4, que presenta una varianza de Wahlund de 0,0401 y un patrón continental europeo.

No se han hallado resultados significativos para el test Fisher 12. En el test conjunto total frente a no progenitores se han identificado 24 marcadores en la población YRI, de los que únicamente rs8084 (exón 3), rs7192 (exón 4), y rs3177928, rs7194, rs7195, rs7196 y rs7197 (exón 5), no se hallan en regiones intrónicas. De estos 8 marcadores exónicos, ninguno alcanza una varianza de Wahlund de 0,1, y únicamente se puede asociar un patrón de distribución continental

asiático a rs7197. También se han identificado 9 marcadores en la población CEU, de los que solo rs3177928 (exón 5) se encuentra en una región no intrónica, con una varianza de Wahlund de 0,0423 y sin patrón de distribución continental asociado. En el test adultos progenitores frente a no progenitores se han identificado 25 marcadores en la población YRI. Los marcadores detectados son los mismo que en el test de Fisher conjunto total frente a no progenitores, con la inclusión de rs114369132, localizado en el intrón 1, con una varianza de Wahlund de 0,0074 y un patrón de distribución continental africano. También se han identificado 12 SNPs en la población CEU: los marcadores detectados son los mismo que en el test de Fisher conjunto total frente a no progenitores, con la inclusión de rs17496549, rs16822616 y rs70993830, encontrándose todos en regiones intrónicas, y unos valores de varianza de Wahlund de 0,0556, 0,0046 y 0,0251, sin poder asociarse patrón de distribución continental a ninguno. Para el test de Fisher entre el conjunto total de la población frente a los descendientes se han identificado 1 SNPs para YRI, 1 para ASW, 3 para CEU y 1 para MXL. En todos los casos está presente el SNP rs4935356, localizado en el intrón 4, con una varianza de Wahlund de 0,0048 y sin patrón continental asociado. En el caso de CEU, se han detectado diferencias en los SNPs rs6911419 (intrón 1) y rs7196 (exón 5), siendo sus varianzas de Wahlund 0,0407 y 0,0238, y no teniendo patrón de distribución asociado. Para el test de Fisher entre el conjunto de adultos progenitores frente a los descendientes se han detectado exactamente las mismas diferencias que en el test de Fisher entre el conjunto total de la población frente a los descendientes. Para el test de Fisher entre el conjunto de adultos no progenitores frente a los descendientes se han identificado 12 marcadores para YRI, 1 para ASW, 9 para CEU y 1 para MXL. En todos los casos está presente el SNP rs4935356, localizado en el intrón 4, con una varianza de Wahlund de 0,0048 y sin patrón continental asociado. En el caso de los 12 marcadores de la población YRI, únicamente rs8084 (exón 3), rs7192 (exón 4), y, rs7194, rs7195, y rs7197 (exón 5), no se hallan en regiones intrónicas. De estos 8 marcadores exónicos, ninguno alcanza una varianza de Wahlund de 0,1, y únicamente se puede asociar un patrón de distribución continental asiático a rs7197. De los 9 marcadores con diferencias significativas en la población CEU, solo rs3177928 (exón 5) se encuentra en una región no intrónica, con una varianza de Wahlund de 0,0423 y sin patrón de distribución continental asociado.

En cuanto a los test de Chi2, se ha determinado que los SNPs rs2239803 y rs9281809 para la población ASW presentan diferencias significativas entre los genotipos observados y los esperados por azar. Ambos se sitúan en el intrón 4, presentan valores de varianza de Wahlund de 0,0628 y 0,0056, y no tienen patrón continental asociado.

TEST	POBLACIONES			
	YRI	ASW	CEU	MXL
HW p1	0	0	0	0
HW p2	0	0	0	0
HW p3	0	0	0	0
HW p4	0	0	0	0
Fisher p12	0	0	0	0
Fisher p13	0	0	0	0
Fisher p23	0	0	1	0
Fisher p14	0	0	0	0
Fisher p24	0	0	0	0
Fisher p34	0	0	0	0
Chi2	0	0	0	0

Tabla R-14: Número de SNPs que muestran significación para los test estadísticos realizados en el gen HLA-DMA.

En la tabla R-14 podemos observar la información obtenida para el gen HLA-DMA. Únicamente se ha hallado una diferencia significativa en las frecuencias alélicas entre la población de adultos progenitores frente a la de no progenitores para la población CEU: el SNP rs142827383 se sitúa en el intrón 2 de HLA-DMA, posee una varianza de Wahlund de 0,0134 y no ha sido posible asignarle un patrón de distribución continental. No se han hallado más resultados significativos ni para este test ni para ningún otro en el gen HLA-DMA.

TEST	POBLACIONES			
	YRI	ASW	CEU	MXL
HW p1	0	2	0	0
HW p2	0	1	0	0
HW p3	0	1	0	0
HW p4	1	0	0	0
Fisher p12	0	0	0	0
Fisher p13	2	0	0	0
Fisher p23	3	3	0	0
Fisher p14	0	0	0	0
Fisher p24	0	0	0	0
Fisher p34	0	0	0	0
Chi2	0	0	0	0

Tabla R-15: Número de SNPs que muestran significación para los test estadísticos realizados en el gen HLA-DMB.

En la tabla R-15 podemos observar la información obtenida para el gen HLA-DMB. Se han hallado 2 SNP que presentan desequilibrio Hardy-Weinberg en el conjunto total de la población ASW: rs11540148 (en la región reguladora 3´) y rs58879393 (en el intrón 3). Sus

varianzas de Wahlund son 0,0334 y 0,0806, y en el caso de rs58879393 ha sido posible asignarle un patrón de distribución continental africano. Igualmente se ha hallado para la población ASW un SNP con desequilibrio Hardy-Weinberg para el conjunto de adultos progenitores: rs1042337, localizado en el exón 3, posee una varianza de Wahlund de 0,0285 y no tiene patrón de distribución continental. También para ASW se ha encontrado un SNP con desequilibrio para el conjunto de adultos no progenitores: rs11540148, que ya hemos visto que estaba en desequilibrio para el conjunto total de la población. En el caso de los descendientes, se ha hallado un SNP en desequilibrio para la población YRI: rs1007636, localizado en el intrón 3, posee una varianza de Wahlund de 0,0243 y un patrón de distribución continental europeo.

No se han hallado resultados significativos para los test de Fisher 12, 14, 24 y 34. En el caso del test conjunto total frente a no progenitores se han identificado 2 SNPs para la población YRI: rs73396789 y rs114828043. Ambos se encuentran localizados en el intrón 3, sus varianzas de Wahlund son 0,0367 y 0,0311, y en el caso de rs114828043 se ha visto que posee un patrón de distribución correspondiente al continente africano. Para el test de Fisher adultos progenitores frente a no progenitores, se han hallado 3 SNPs para la población YRI y 3 SNPs para la población ASW. En el caso de YRI, los SNPs son rs10751 (región reguladora 3´), rs73396789 (intrón 3) y rs114828043 (intrón 3), siendo sus varianzas de Wahlund 0,0234, 0,0367 y 0,0311, y como ya hemos visto anteriormente, para rs114828043 se ha encontrado un patrón de distribución africano. En el caso de ASW, los 3 SNPs encontrado en desequilibrio para esta comparación son rs112744519 (intrón 2), rs114032390 (intrón 2) y rs190602907 (intrón 1). Sus varianzas de Wahlund son 0,019, 0,0191 y 0,0155 respectivamente, y para el caso de rs190602907 se ha dictaminado que posee un patrón de distribución continental africano.

No se han hallado resultados significativos para los test de Chi2, por lo que cabe señalar que no hay diferencias significativas entre los genotipos observados y los esperados por azar.

TEST	POBLACIONES			
	YRI	ASW	CEU	MXL
HW p1	0	6	0	6
HW p2	2	1	0	3
HW p3	0	3	0	0
HW p4	0	1	0	1
Fisher p12	0	0	0	0
Fisher p13	1	0	0	0
Fisher p23	2	0	2	0
Fisher p14	0	0	0	0
Fisher p24	0	0	0	0
Fisher p34	1	0	0	0
Chi2	0	0	0	0

Tabla R-16: Número de SNPs que muestran significación para los test estadísticos realizados en el gen HLA-DOA.

En la tabla R-16 podemos observar la información obtenida para el gen HLA-DOA. Se han hallado 6 SNPs para la población ASW y 6 SNPs para la población MXL en el conjunto total de la población. Para ASW, los SNPs son rs3129304 (región reguladora 3´), rs3129303 (región reguladora 3´), rs2267647 (intrón 2), rs364950 (exón 2), rs743771 (intrón 1) y rs404557 (intrón 1). Sus varianzas de Wahlund son 0,0859, 0,0832, 0,0617, 0,148, 0,0595 y 0,1493. Resulta llamativo que tanto rs364950 como rs404557, muestren ambos una varianza de Wahlund mayor de 0,1 y al mismo tiempo un patrón de distribución africano. Para MXL, los SNPs rs3129304 (región reguladora 3´), rs3129303 (región reguladora 3´), rs378352 (exón 4), rs369150 (intrón 2), rs375256 (exón 2) y rs86567 (intrón 1). Sus varianzas de Wahlund son 0,0859, 0,0832, 0,1129, 0,0862, 0,0872 y 0,0531. Ninguno tiene un patrón continental asociado.

Para el conjunto de adultos progenitores se han identificado 2 marcadores en desequilibrio para la población YRI: rs1044429 y rs592625, ambos situados en la región reguladora 3´, poseen varianzas de Wahlund de 0,0851 y 0,0633, pero no poseen un patrón continental reconocible. También se ha detectado un SNP para la población ASW: rs376892, en la región reguladora 3´, posee una varianza de Wahlund de 0,0527 y un patrón continental africano. En el caso de la población MXL se han detectado 3 SNPs: rs369150 (intrón 2), rs375256 (exón 2) y rs86567 (intrón 1); siendo sus varianzas de Wahlund 0,0862, 0,0872 y 0,0531, pero sin patrón continental conocido.

Para el conjunto de adultos no progenitores se han encontrado 3 marcadores en desequilibrio para la población ASW: rs142850513, rs3129304 y rs3129303. Los tres se encuentran en la región reguladora 3´, poseen varianzas de Wahlund de 0,1133, 0,0859 y 0,0832. El caso de rs142850513 resulta curioso porque es un pequeño indel de 20 pares de bases que posee un patrón de distribución continental africano.

Para el conjunto de descendientes se han encontrado un SNP en la población ASW (rs453779, localizado en el intrón 2) y otro en la MXL (rs11575906, localizado en el exón 2). Sus varianzas de Wahlund son 0,0423 y 0,0148, y no se ha identificado un patrón continental para ninguno.

No se han hallado resultados significativos para los test de Fisher 12, 14 y 24. En el caso del test conjunto total frente a no progenitores se ha detectado un SNP con diferencias alélicas significativas en la población YRI: rs2267647, localizado en el intrón 2, posee una varianza de Wahlund de 0,0617 y no tiene patrón continental asociado. En el caso del test de Fisher de progenitores frente a no progenitores se han detectado 2 SNPs para la población YRI y otros 2 para la población CEU. En el caso de YRI los SNPs son rs71565360 (región reguladora 3´) y rs2267647 (intrón 2). Sus varianzas de Wahlund son 0,0347 y 0,0617, y en el caso de rs71565360 presenta un patrón de distribución asiático. En el caso de la población CEU, los SNPs son rs41270506 y rs41270510. Ambos se sitúan en el intrón 1, poseen varianzas de Wahlund de 0,0234 y 0,048, pero no poseen un patrón distributivo conocido. En el caso del test de Fisher de no progenitores frente

a descendientes se ha detectado un SNP en la población YRI: rs2267647, localizado en el intrón 2, posee una varianza de Wahlund de 0,0617, pero no se le conoce patrón de distribución continental.

No se han hallado resultados significativos para los test de Chi2, por lo que cabe señalar que no hay diferencias significativas entre los genotipos observados y los esperados por azar.

TEST	POBLACIONES			
	YRI	ASW	CEU	MXL
HW p1	8	34	25	3
HW p2	8	30	15	2
HW p3	0	3	0	0
HW p4	0	0	1	0
Fisher p12	0	0	0	0
Fisher p13	0	0	34	2
Fisher p23	2	0	44	3
Fisher p14	3	1	3	1
Fisher p24	3	1	3	1
Fisher p34	3	1	4	3
Chi2	0	0	0	0

Tabla R-17: Número de SNPs que muestran significación para los test estadísticos realizados en el gen HLA-DPA1.

En la tabla R-17 podemos observar la información obtenida para el gen HLA-DPA1. Se han hallado 63 SNPs distintos que presentan desequilibrio Hardy-Weinberg en el conjunto total de las poblaciones de la base de datos HapMap, localizados en la región reguladora 3´ (como rs8807 o rs7905), exones (como por ejemplo rs1062481 o rs1042176 en el exón 2), intrones (como rs77403406 en el intrón 4 o rs4422658 en el intrón 1) y en la región reguladora 5´ (como rs1042117); y 54 marcadores distintos que presentan desequilibrio Hardy-Weinberg en el conjunto de adultos progenitores, localizados en la región reguladora 3´ (como rs9277338 o rs7905), exones (como por ejemplo rs2308927 o rs1126544 en el exón 3), intrones (como rs7769592 en el intrón 1) y en la región reguladora 5´ (como rs1042117).

Para el conjunto de adultos no progenitores se han encontrado 3 marcadores en desequilibrio para la población ASW: rs58249824 (intrón 4), rs73741626 (intrón 1) y rs7760936 (intrón 1). Sus varianzas de Wahlund son 0,1662, 0,018 y 0,0155. En el caso de rs58249824 se ha detectado que posee un patrón continental europeo.

Para el conjunto de descendientes se ha encontrado un SNP en desequilibrio Hardy-Weinberg para la población CEU: rs2301225, localizado en el intrón 4, posee una varianza de Wahlund de 0,1195 y patrón de distribución continental europeo.

No se han hallado resultados significativos para los test de Fisher 12. En el caso del test conjunto total frente a no progenitores se han detectado 36 marcadores que presentan diferencias significativas en el conjunto de poblaciones de la base de datos HapMap, localizados en la región reguladora 3´ (como rs17220948 o rs7905), exones (como rs1042177 o rs1042174 en el exón 2), intrones (como rs66951571 en el intrón 4) y la región reguladora 5´ (como rs1126511 o rs1126513). En el caso del test progenitores frente a no progenitores se han detectado 49 marcadores, situados en la región reguladora 3´ (como rs9277338 o rs17220948), exones (como rs2308931 en el exón 3), intrones (como rs66977180 en el intrón 1) y la región reguladora 5´ (como rs1126511 o rs1126513). En el caso del test conjunto total frente a descendientes se han detectado 3 SNPs para la población YRI: rs17509489 (exón 4), rs6457711 (intrón 1) y rs7770418 (intrón 1). Sus varianzas de Wahlund son 0,0405, 0,1162 y 0,0272. En el caso de rs6457711 se ha detectado un patrón de distribución europeo. Para las poblaciones ASW y MXL se ha detectado un SNP: rs1126513, localizado en la región reguladora 5´, posee una varianza de Wahlund de 0,0212, pero no ha sido posible asignarle un patrón de distribución continental concreto. Para la población CEU se han encontrado 3 SNPs: rs17509489, rs1042178 y rs6457711. Como podemos ver son casi los mismo que para la población YRI, aunque en este caso se sustituye el SNP rs7770418 de aquella por el SNP rs1042178, presente en el exón 2, posee una varianza de Wahlund de 0,1575 y posee un patrón continental europeo. En el caso del test progenitores frente a descendientes se han detectado las mismas diferencias que en el caso del test conjunto de población total frente a descendientes para todas las poblaciones. En el caso del test no progenitores frente a descendientes se han detectado las mismas diferencias que en los dos test anteriores para la población YRI y ASW. Para la población CEU, aparte de los detectados en los dos test anteriores, se ha encontrado que el SNP rs7770418 también presenta diferencias significativas. En el caso de MXL, aparte del ya detectado, se ha encontrado que los SNPs rs2301226 (intrón 4) y rs2856830 (intrón 1) presentan diferencias significativas. Sus valores de varianza de Wahlund son 0,0352 y 0,0303, y ninguno presenta un patrón de distribución concreto.

No se han hallado resultados significativos para los test de Chi2, por lo que cabe señalar que no hay diferencias significativas entre los genotipos observados y los esperados por azar.

TEST	POBLACIONES			
	YRI	ASW	CEU	MXL
HW p1	41	6	35	3
HW p2	40	2	20	5
HW p3	0	4	0	0
HW p4	0	0	1	0
Fisher p12	0	0	0	0
Fisher p13	0	0	31	0
Fisher p23	15	0	53	2
Fisher p14	9	4	11	4
Fisher p24	9	4	11	4
Fisher p34	10	4	14	4
Chi2	0	0	0	0

Tabla R-18: Número de SNPs que muestran significación para los test estadísticos realizados en el gen HLA-DPB1.

En la tabla R-18 podemos observar la información obtenida para el gen HLA-DPB1. Se han hallado 72 SNPs distintos que presentan desequilibrio Hardy-Weinberg en el conjunto total de las poblaciones de la base de datos HapMap, localizados en exones (como por ejemplo rs1042117 en el exón 1), intrones (como rs7769592 en el intrón 1 o rs7770370 en el intrón 2) y en la región reguladora 3´ (como rs144401610 o rs9501263); y 56 marcadores distintos que presentan desequilibrio Hardy-Weinberg en el conjunto de adultos progenitores, localizados en exones (como por ejemplo rs1042117 en el exón 2), intrones (como rs6415133 en el intrón 1) y en la región reguladora 5´ (como rs3128970).

En el caso de los adultos no progenitores se han encontrado 4 SNPs en desequilibrio para la población ASW: rs73741626 (intrón 1), rs7760936 (intrón 1), rs112104961 (intrón 2) y rs9277518 (intrón 5). Sus varianzas de Wahlund son 0,018, 0,0155, 0,255 y 0,0985. En los dos últimos casos se ha detectado que poseen patrones de distribución africano y europeo respectivamente.

En el caso de descendientes se ha detectado un SNP en desequilibrio Hardy-Weinberg para la población CEU: rs9501249, localizado en el intrón 2, posee una varianza de Wahlund de 0,0467 y un patrón de distribución continental africano.

No se han hallado resultados significativos para los test de Fisher 12. En el caso del test conjunto total frente a no progenitores se han detectado 31 marcadores para la población CEU, localizados en intrones (como rs2071352 en el intrón 1 o rs9501253 en el intrón 2), exones (como rs1126511 o rs1126513 en el exón 2) y en la región reguladora 3´ (como rs9501263). En el caso del test progenitores frente a no progenitores se han detectado 57 marcadores, situados en intrones (como rs9469347 en el intrón 1), exones (como rs1126511 o rs1126513 en el exón 2) y en la región reguladora 3´ (como rs73743107). Para los test conjunto total de la población frente a descendientes y progenitores para descendientes se han detectado diferencias significativas en los

mismos 16 marcadores, localizados en intrones (como rs6457711 en el intrón 1), exones (como rs1126513 en el exón 2) y en la región reguladora 3´ (como rs9296078). En el caso del test no progenitores frente a descendientes se han detectado 20 marcadores distintos, localizados en intrones (como rs6457711 en el intrón 1), exones (como rs1126513 en el exón 2) y en la región reguladora 3´ (como rs9501263). Muchos de estos marcadores también se habían detectado en los dos test anteriores.

No se han hallado resultados significativos para los test de Chi2, por lo que cabe señalar que no hay diferencias significativas entre los genotipos observados y los esperados por azar.

TEST	POBLACIONES			
	YRI	ASW	CEU	MXL
HW p1	314	245	376	282
HW p2	311	244	386	234
HW p3	53	33	32	50
HW p4	0	0	0	0
Fisher p12	0	0	0	0
Fisher p13	74	16	2	1
Fisher p23	99	94	4	1
Fisher p14	2	0	2	1
Fisher p24	2	0	2	1
Fisher p34	3	0	2	0
Chi2	13	9	0	4

Tabla R-19: Número de SNPs que muestran significación para los test estadísticos realizados en el gen HLA-DQA1.

En la tabla R-19 podemos observar la información obtenida para el gen HLA-DQA1. Se han hallado 495 SNPs distintos que presentan desequilibrio Hardy-Weinberg en el conjunto total de las poblaciones de la base de datos HapMap, localizados en la región reguladora 5´ (como rs9272426 o rs200004305), exones (como por ejemplo rs1042117 en el exón 1 o rs1129740 en el exón 2), intrones (como rs9272444 en el intrón 1 o rs9272713 en el intrón 2) y en la región reguladora 3´ (como rs9272990 o rs9272981). Hay 496 SNPs distintos en desequilibrio en el conjunto de progenitores de las poblaciones de la base de datos HapMap, muchos de los cuales son compartidos con el conjunto total. En lo referente al conjunto de no progenitores, señalar que se han detectado 137 SNPs distintos en desequilibrio, localizados en intrones (como rs28383350 en el intrón 1), exones (como rs707950 en el exón 3) y la región reguladora 3´ (como rs7757583). No se han detectado SNPs en desequilibrio Hardy-Weinberg para el conjunto de los descendientes en ninguna de las poblaciones.

Tampoco se han hallado diferencias alélicas significativas entre el conjunto total de la población y los progenitores. En el caso del test conjunto total frente a no progenitores se han detectado 93 marcadores que presentan diferencias significativas, estando repartidos por todo el

gen. En el caso del test progenitores frente a no progenitores se han detectado 195 SNPs distintos localizados en las diversas regiones génicas que presentan diferencias alélicas significativas. Para las comparaciones del conjunto total de la población frente a los descendientes, y de los progenitores frente a los descendientes se han hallado 2 SNPs con diferencias significativas en las poblaciones YRI y CEU: rs7771972 y rs9272535, ambos localizados en el intrón 1, poseen unas varianzas de Wahlund de 0,0442 y 0,0817, y no tienen patrón continental asociado. En la población MXL también se ha detectado diferencias significativas para el SNP rs9272535. En el caso de la comparación de los adultos no progenitores frente a los descendientes se han identificado 3 SNPs para la población YRI y 2 SNPs para la CEU: En adición a los observados en los dos test anteriores, se ha detectado que rs9272775 presenta diferencias significativas para YRI. Este SNP, localizado en el intrón 3, presenta un valor de varianza de Wahlund de 0,0339 y no tiene patrón de distribución asociado.

En cuanto a los test de Chi2, se ha determinado que 13 marcadores presentan diferencias significativas entre los genotipos observados y los esperados por azar para la población YRI. Los marcadores rs9272502, rs9272525, rs9272553, rs77888377, rs9272603, rs9272614, rs9272656 y rs9272661 se localizan en el intrón 1, mientras que rs1130151, rs9272962, rs9272972, rs9272975 y rs9272981 se sitúan en el exón 5. En ningún caso alcanzan un valor de varianza de Wahlund de 0,1 ni es posible asignarles un patrón de distribución continental concreto. Para la población ASW se ha detectado rs9272435 (intrón 1), rs9272451 (intrón 1), rs2213287 (intrón 1), rs9272472 (intrón 1), rs9272485 (intrón 1), rs1048027 (exón 2), rs397843773 (intrón 3), rs79351547 (intrón 3) y rs116022266 (intrón 3). En ningún caso alcanzan un valor de varianza de Wahlund de 0,1 ni es posible asignarles un patrón de distribución continental concreto. Para la población MXL se han detectado los SNPs rs2187668 (intrón 1), rs9272622 (intrón 1), rs531139227 (intrón 1) y rs9272729 (intrón 2). Sus valores de varianza de Wahlund son 0,0279, 0,0512, 0,0391 y 0,0275. En el caso de rs2187668 y rs9272729 se ha detectado que poseen un patrón continental europeo.

TEST	POBLACIONES			
	YRI	ASW	CEU	MXL
HW p1	422	344	458	414
HW p2	415	318	430	346
HW p3	42	109	107	89
HW p4	0	0	0	0
Fisher p12	0	0	0	0
Fisher p13	114	14	17	0
Fisher p23	140	92	27	1
Fisher p14	1	0	0	0
Fisher p24	1	0	0	0
Fisher p34	1	0	0	0
Chi2	16	0	1	0

Tabla R-20: Número de SNPs que muestran significación para los test estadísticos realizados en el gen HLA-DQB1.

En la tabla R-20 podemos observar la información obtenida para el gen HLA-DQB1. Se han hallado 606 SNPs distintos que presentan desequilibrio Hardy-Weinberg en el conjunto total de las poblaciones de la base de datos HapMap, localizados en la región reguladora 3´ (como rs9273410), exones (como por ejemplo rs1140343 en el exón 4 o rs1049083 en el exón 2), intrones (como rs9273472 en el intrón 4 o rs9273781 en el intrón 3) y en la región reguladora 5´ (como rs1049053). Hay 596 SNPs distintos en desequilibrio en el conjunto de progenitores de las poblaciones de la base de datos HapMap, muchos de los cuales son compartidos con el conjunto total. En lo referente al conjunto de no progenitores, señalar que se han detectado 208 SNPs distintos en desequilibrio, localizados en la región reguladora 3´ (como rs9273418), intrones (como rs9273482 en el intrón 4) y exones (como rs1063318 en el exón 2). No se han detectado SNPs en desequilibrio Hardy-Weinberg para el conjunto de los descendientes en ninguna de las poblaciones.

Tampoco se han hallado diferencias alélicas significativas entre el conjunto total de la población y los progenitores. En el caso del test conjunto total frente a no progenitores se han detectado 144 marcadores que presentan diferencias significativas, estando repartidos por todo el gen. En el caso del test progenitores frente a no progenitores se han detectado 253 SNPs distintos localizados en las diversas regiones génicas que presentan diferencias alélicas significativas. Para las comparaciones del conjunto total de la población frente a los descendientes, y de los progenitores frente a los descendientes se ha hallado 1 SNP con diferencias significativas en la población YRI: rs3830058, localizado en el intrón 1, con una varianza de Wahlund de 0,0618, no posee un patrón continental asociado. Para la comparación de los adultos no progenitores frente a los descendientes se ha detectado diferencias alélicas significativas en el SNP rs12918 en la población YRI. Este SNP, localizado en la región reguladora 3´, presenta un valor de varianza de Wahlund de 0,0372 y no tiene un patrón de distribución continental conocido.

En cuanto a los test de Chi2, se ha determinado que 16 marcadores presentan diferencias significativas entre los genotipos observados y los esperados por azar para la población YRI. Los marcadores rs9273424, rs9273425 y rs1130455 se sitúan en la región reguladora 3´; rs9273514, rs9273515, rs3020628, rs9273556, rs9273559 y rs28724238 están en el intrón 4; y rs376975457, rs281863930, rs281863610, rs281863357, rs281874967, rs281874965 y rs281874963 se localizan en el intrón 4. Únicamente rs281863357 (0,1135) posee un valor de varianza de Wahlund mayor de 0,1, y solo ha sido posible asignar un patrón de distribución continental asiático a rs281863930. También se ha detectado un resultado significativo para la población CEU: rs9274284, localizado en el intrón 2, posee una varianza de Wahlund de 0,0881, pero no tiene un patrón continental concreto.

Análisis de una muestra de SNPs en Gitanos del País Vasco

Se han estudiado 16 SNPs presentes en la población gitana del País Vasco (Tabla R-21). Salvo en el caso del SNP rs116818505 que es monomórfico, todos los demás SNPs estudiados son

polimórficos. Sin contar con el caso del SNP monomórfico, el rango de valores de las frecuencias alélicas es relativamente amplio, con un valor de media de 0,410 y un valor de mediana de 0,313. Desde el mínimo de 0,104 del SNP rs72873921 hasta el máximo de 0,973 del SNP rs2071350, podemos observar valores variados.

SNP	Alelo de referencia	Frecuencia alélica	Error estándar
rs3823324	A	0,671	0,038
rs79244404	C	0,149	0,029
rs2770	C	0,222	0,035
rs9273352	G	0,638	0,039
rs369150	A	0,295	0,041
rs9277332	A	0,629	0,041
rs200789833	G	0,463	0,043
rs72873921	C	0,104	0,025
rs72873922	A	0,097	0,024
rs2071350	C	0,973	0,013
rs9277413	A	0,173	0,030
rs9277418	C	0,818	0,032
rs116818505	G	0,000	0,000
rs9277498	C	0,160	0,029
rs72500564	C	0,838	0,031
rs9374640	A	0,331	0,039

Tabla R-21: Frecuencias alélicas y errores estándar de los SNPs analizados.

Se han estudiado también las frecuencias genotípicas (Tabla R-22). Aparte del caso ya mencionado del SNP monomórfico, llama la atención el caso de los SNPs rs72873921 y rs72873922. En ambos casos, se presentan individuos heterocigotos, es decir, que presentan ambos alelos; pero sin embargo no hay frecuencias genotípicas homocigotas para el alelo de referencia (C y A respectivamente). En el caso de rs2071350 ocurre lo contrario, es el genotipo homocigoto para el alelo de referencia (en este caso el C) y el heterocigoto los que están presentes, mientras que el genotipo homocigoto para el alelo T no está presente. Cabe destacar también el caso de rs200789833 que presenta un valor de frecuencia genotípica de heterocigotos de 0,897, siendo esta la más alta de entre todos los SNPs estudiados. A modo de resumen, podemos señalar que en el caso de rs9273352, rs9277332 y rs200789833 es más frecuente el genotipo heterocigoto; en el caso de rs3823324, rs2071350, rs9277418 y rs72500564 es más frecuente el genotipo homocigoto para el alelo de referencia; y en el caso de rs79244404, rs2770, rs369150, rs72873921, rs72873922, rs9277413, rs116818505, rs9277498 y rs9374640 es más frecuente el genotipo homocigoto para el alelo alternativo.

SNP	Genotipo			Frecuencia		
	1	2	3	1	2	3
rs3823324	A:T	A:A	T:T	0,395	0,474	0,132
rs79244404	C:T	C:C	T:T	0,273	0,013	0,714
rs2770	C:T	C:C	T:T	0,389	0,028	0,583
rs9273352	G:T	G:G	T:T	0,434	0,421	0,145
rs369150	A:G	A:A	G:G	0,361	0,115	0,525
rs9277332	A:G	A:A	G:G	0,657	0,300	0,043
rs200789833	G:T	G:G	T:T	0,897	0,015	0,088
rs72873921	C:T	C:C	T:T	0,208	0,000	0,792
rs72873922	A:G	A:A	G:G	0,195	0,000	0,805
rs2071350	C:T	C:C	T:T	0,053	0,947	0,000
rs9277413	A:G	A:A	G:G	0,269	0,038	0,692
rs9277418	C:T	C:C	T:T	0,284	0,676	0,041
rs116818505	G:T	G:G	T:T	0,000	0,000	1,000
rs9277498	C:T	C:C	T:T	0,295	0,013	0,692
rs72500564	C:T	C:C	T:T	0,296	0,690	0,014
rs9374640	A:G	A:A	G:G	0,419	0,122	0,459

Tabla R-22: Frecuencias genotípicas de los SNPs analizados.

Se han realizado test de desequilibrio Hardy-Weinberg para los SNPs estudiados (Tabla R-23).

SNP	Chi-2	p
rs3823324	0,852	0,356
rs79244404	0,414	0,520
rs2770	1,125	0,289
rs9273352	0,272	0,602
rs369150	1,080	0,299
rs9277332	11,615	0,001
rs200789833	43,942	0,000
rs72873921	1,035	0,309
rs72873922	0,897	0,344
rs2071350	0,056	0,812
rs9277413	0,276	0,600
rs9277418	0,175	0,675
rs116818505	-	-
rs9277498	0,712	0,399
rs72500564	0,569	0,451
rs9374640	0,218	0,641

Tabla R-23: Test de equilibrio Hardy-Weinberg de los SNPs analizados. En rojo se señalan los valores de p en los que es menor que el nivel de significación establecido, y por tanto se rechaza la hipótesis nula de que el SNP se encuentra en equilibrio Hardy-Weinberg.

Dos SNPs han mostrado desequilibrio Hardy-Weinberg. En general, se observan unas tendencias similares a otras poblaciones en las frecuencias genotípicas, con exceso o defecto de heterocigotos según los casos, pero tan sólo dos han mostrado significación.

Se ha realizado un escalamiento multidimensional (MDS, *Multidimensional scaling*) para estudiar las relaciones entre la población gitana y distintas poblaciones tanto de distribución europea por ser el área que habitan en el presente como del sur de Asia, por ser la zona de origen putativa de la población proto-romaní (Figura R-39).

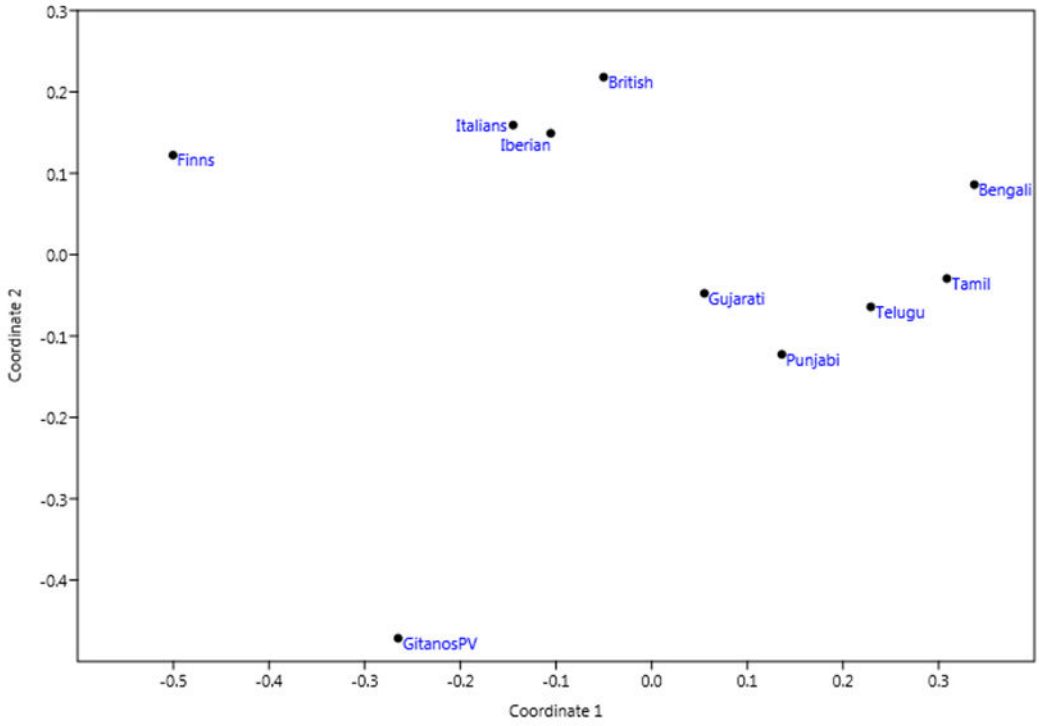


Figura R-39: MDS realizado a partir de una matriz de distancias R de Harpending y Jenkins.

El valor del stress de Kruskal, que aumenta con el tamaño de la muestra y el número de variables, es en realidad una medida de la bondad del ajuste (Guerrero-Casas & Ramírez-Hurtado 2002). Dado que tiene un valor de 0,016, el ajuste es excelente según la escala de Kruskal (1964), y por tanto se trata de una configuración robusta.

Aparece una agrupación de muestras del sur de Asia (*Gujarati, Punjabi, Telugu, Tamil* y *Bengali*) y otra de Europa (*Italians, Iberian* y *British*), si bien Finlandia (*Finns*) aparece un tanto distanciada. Los gitanos se alejan de ambos grupos poblacionales, reflejando una posible acción de la deriva.

Se ha realizado un análisis factorial de correspondencias (AFC) para analizar las relaciones de interdependencia entre variables, en este caso las relaciones interpoblacionales teniendo en cuenta los SNPs estudiados (Figura R-40).

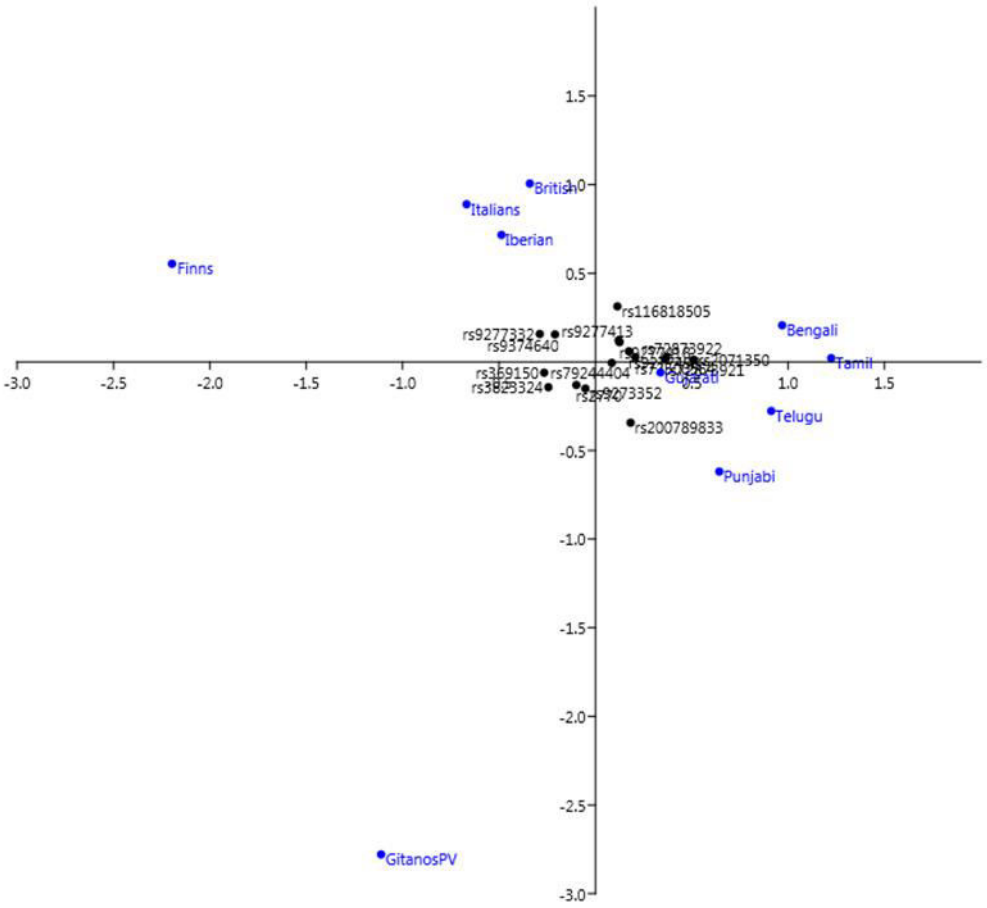


Figura R-40: AFC realizado sobre un grupo de poblaciones de Europa y Sur de Asia a partir de varios SNPs del cromosoma 6. La interpretación del gráfico va en función de la distancia existente entre cada una de las variables (frecuencias de los SNPs) y una población concreta o un grupo de poblaciones. Se considera la influencia de las frecuencias de los SNPs sobre la posición de las poblaciones.

El AFC mostró una configuración muy similar de las poblaciones, caracterizando a la población gitana como claramente diferenciada del resto. Los SNPs que subrayan esta diferenciación son rs369150, rs3823324, rs2770 y rs9273352, para los cuales presentan los valores máximos.

Se ha realizado un análisis del centroide (Figura R-41).

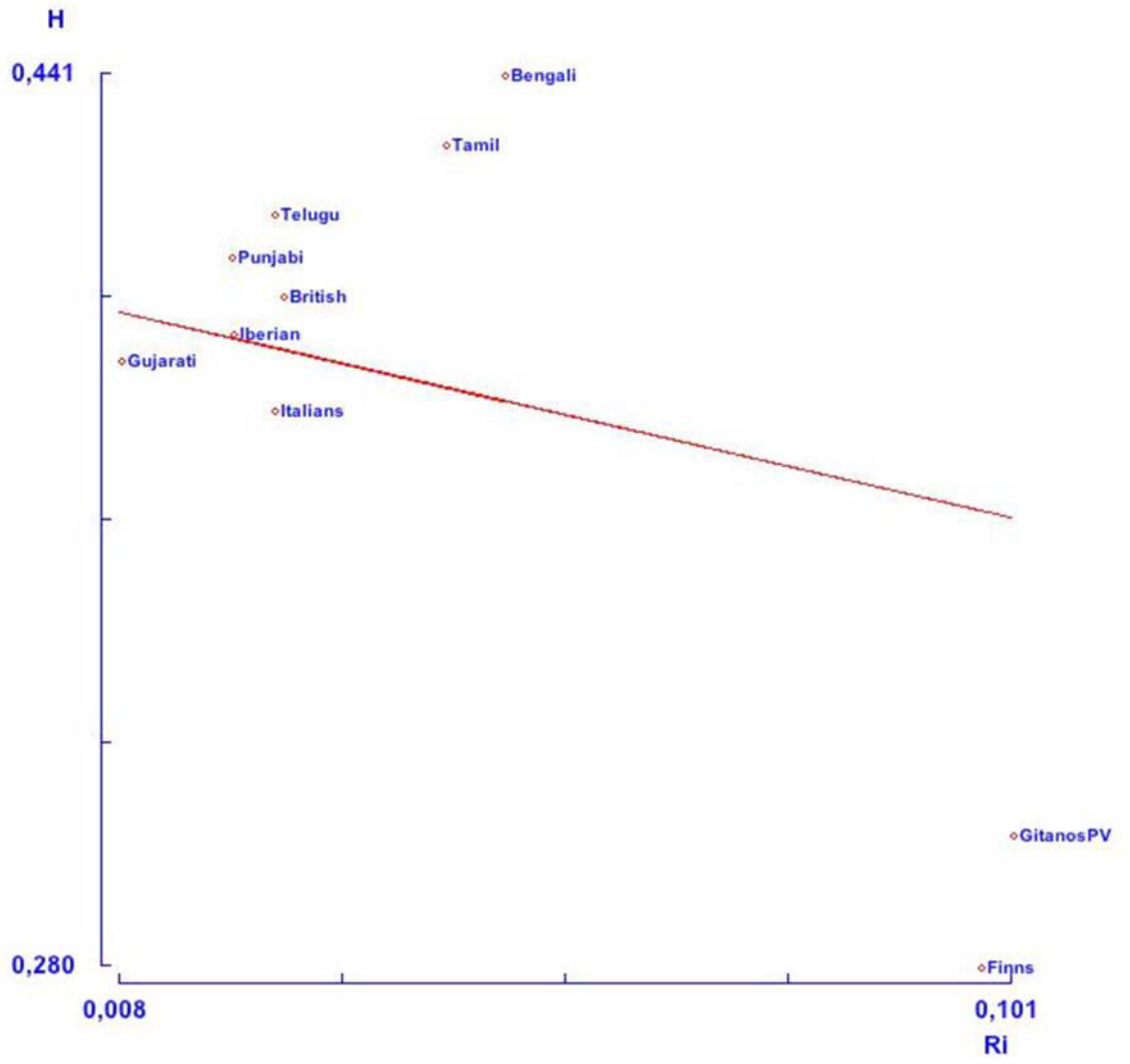


Figura R-41: Análisis del centroide realizado sobre un grupo de poblaciones de Europa y Sur de Asia a partir de varios SNPs del cromosoma 6.

El análisis del centroide revela un notable grado de aislamiento de la población gitana, rasgo que comparte con la población de Finlandia. La línea roja indica la heterocigosidad (eje vertical) esperada según la varianza interpoblacional o varianza de Wahlund observada (eje horizontal). Las poblaciones por debajo de la línea son poblaciones aisladas con valores bajos de heterocigosidad, y las poblaciones por debajo de la línea son poblaciones afectadas por flujo génico. Al mismo tiempo las poblaciones que están por encima de la línea y se sitúan a la izquierda (más heterocigosidad, menos varianza) son poblaciones que reciben flujo génico de poblaciones cercanas, como por ejemplo, poblaciones que se conforman como capital de poblaciones cercanas.

Las poblaciones que se sitúan a la derecha por encima de la línea roja (más heterocigosidad, más varianza) son poblaciones que reciben flujo génico de poblaciones externas a las analizadas.

Podemos observar que en el caso de nuestras poblaciones, tanto los finlandeses como los gitanos del País Vasco son poblaciones muy aisladas, ya que presentan una heterocigosidad muy baja para la varianza tan grande que poseen. Los Gujarati (población de la India) y los italianos, si bien presentan menos heterocigosidad de la esperada, no son casos tan acusados como los dos anteriores. La población ibérica se sitúa casi sobre la línea roja, revelando que tienen la heterocigosidad que cabría esperar para la varianza que presentan. El resto de poblaciones de la Figura R-17 presentan más heterocigosidad de la que cabría esperar para sus valores de varianza de Wahlund, y en general, se sitúan hacia la izquierda del gráfico.

Se analizaron las posibles diferencias en las frecuencias alélicas de los progenitores y sus descendientes en los tríos que se han establecido en la muestra de población gitana. Ningún SNP mostró diferencias estadísticamente significativas (Tabla R-24).

SNP	Fisher exact test
rs3823324	0,7719
rs79244404	0,7181
rs2770	1
rs9273352	0,77469
rs369150	1
rs9277332	0,7817
rs200789833	1
rs72873921	1
rs72873922	1
rs2071350	1
rs9277413	0,3599
rs9277418	0,3599
rs116818505	-
rs9277498	0,3599
rs72500564	0,3599
rs9374640	1

Tabla R-24: Test exactos de Fisher de comparación entre las frecuencias alélicas observadas entre los progenitores y sus descendientes en un grupo de tríos.

También se han comparado las frecuencias observadas y esperadas de los cruzamientos genotípicos en las parejas progenitoras de los distintos tríos caracterizados en la muestra de población gitana estudiada (Tabla R-25). De igual manera que en la Tabla R-24, no se observan diferencias significativas entre los valores estudiados.

SNP	Chi-2	p	Monte Carlo p
rs3823324	0,6	0,98800324	0,8836
rs79244404	0,022675737	0,99999591	0,8977
rs2770	3,782716049	0,58110412	0,1459
rs9273352	0,285039636	0,99791502	0,9795
rs369150	3,928571429	0,55974516	0,1185
rs9277332	3,393313609	0,63958848	0,0678
rs200789833	3,412152778	0,63671969	0,0607
rs72873921	0,12345679	0,9997274	0,7361
rs72873922	0,12345679	0,9997274	0,7303
rs2071350	0,027700831	0,99999327	0,6097
rs9277413	1,715131109	0,88698801	0,4929
rs9277418	1,715131109	0,88698801	0,4865
rs116818505	0	1	1
rs9277498	1,715131109	0,88698801	0,5001
rs72500564	1,715131109	0,88698801	0,4965
rs9374640	2,519723866	0,77352173	0,3335

Tabla R-25: Test Chi-cuadrado de comparación entre las frecuencias observadas y esperadas de cruzamientos genotípicos en las parejas de progenitores de un grupo de tríos.

Relación entre genes HLA de Neandertal y humanos anatómicamente modernos

El estudio de las posibles introgresiones de genes HLA de neandertal ha arrojado resultados dispares en los diferentes alelos analizados, lo que implica necesariamente un entrecruzamiento diferencial con distintas poblaciones humanas en un amplio espacio geográfico y temporal. Se presentan a continuación los árboles de relaciones evolutivas realizados con el programa MEGA para cada uno de los genes estudiados. Dado que la cantidad de alelos es distinta en cada gen, y que en algunos casos el número de alelos ronda los varios cientos, es necesario realizar una preselección de los alelos representados. En cada caso se indicará si ha sido necesario realizar dicha preselección. En todos los casos, el porcentaje de árboles replicados en los que los taxones asociados se encuentran agrupados en la prueba de *bootstrap* se muestran junto a los nodos de las ramas. Los árboles están dibujados a escala, con longitudes de rama en las mismas unidades que las distancias evolutivas utilizadas para inferir el árbol filogenético. Las distancias evolutivas están analizadas en unidades del número de sustituciones por locus.

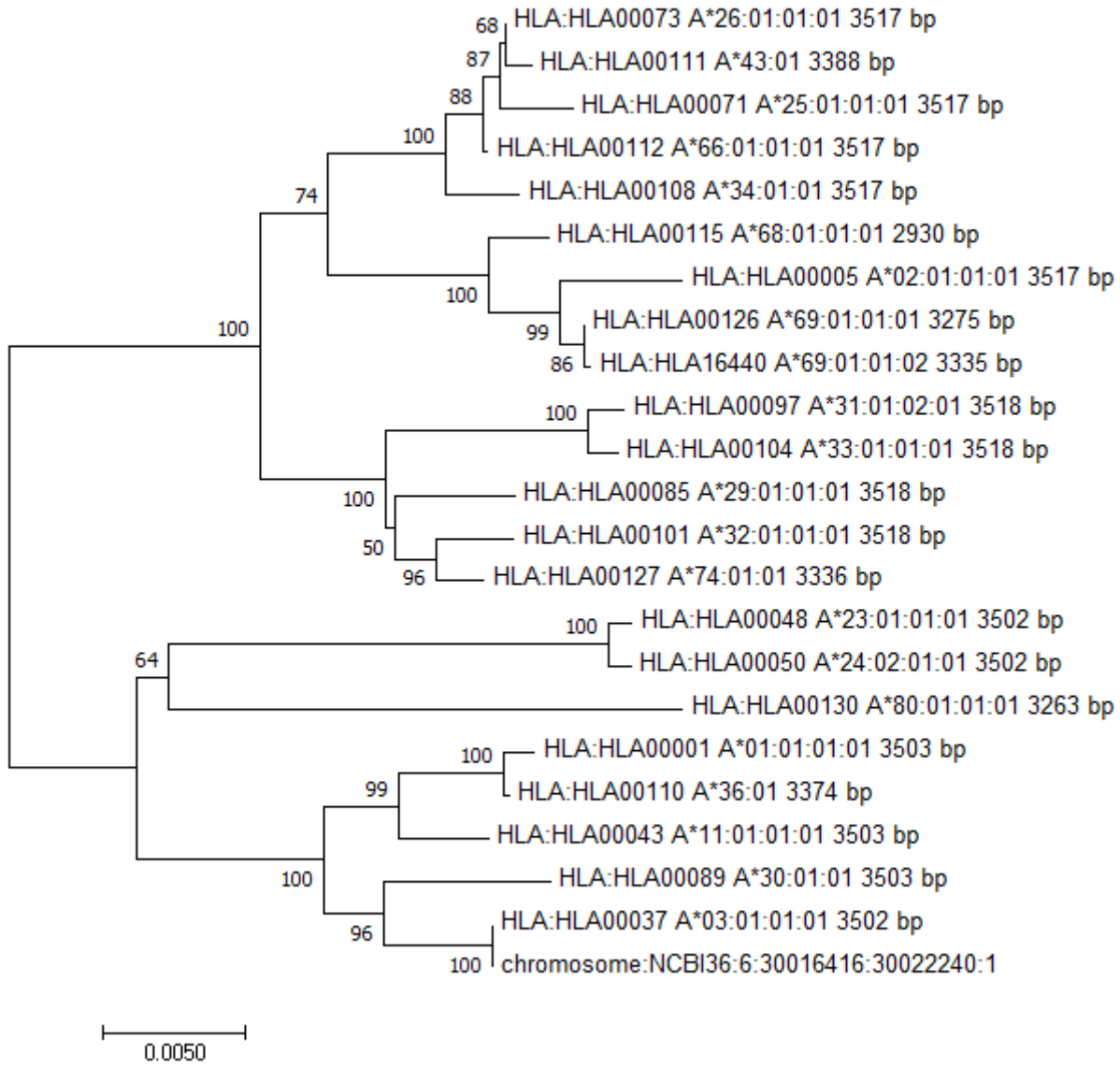


Figura R-42: Árbol de relaciones evolutivas para el gen HLA-A. Se presenta el árbol óptimo con la suma de longitud de rama = 0,13428520. Se han analizado 23 secuencias nucleotídicas, y un total de 2903 loci. El archivo original de HLA-A de humanos modernos contaba con 849 alelos. La preselección ha consistido en elegir un alelo de cada grupo, según el código de alelo, teniendo en cuenta el primer par de dígitos de dicho código.

En general, las relaciones representadas en la Figura R-42 presentan una alta robustez, por ejemplo, el clúster más próximo en el que se agrupa el gen HLA-A Neandertal presenta un valor de presencia en las distintas réplicas del 96%. Podemos observar como el alelo HLA-A de Neandertal, al que se le ha dado el nombre de *chromosome:NCBI36:6:30016416:30022240:1*, aparece asociado a A*03:01:01:01 con un valor de *bootstrap* de 100, por lo que es altamente probable que este alelo sea el que proviene de la introgresión del alelo Neandertal. Este alelo está presente sobre todo en poblaciones de Centroeuropa y relacionadas: tiene una frecuencia de 0,148 en individuos caucásicos de EEUU, una frecuencia de 0,129 en individuos del este de Europa residentes en EEUU, un 0,1281 en la población de Polonia (donde el fenotipo asociado a este alelo

está presente en el 23,9% de la población) y un 0,114 en población de EEUU con ancestría italiana. Está distribución geográfica es concordante con el área en el que se habría dado el entrecruzamiento entre Neandertal y humanos anatómicamente modernos, por lo que hace más probable que este sea el alelo proveniente de la introgresión.

También se relaciona con el alelo A*30:01:01, presente en poblaciones con ancestría ibérica y mestizas caribeñas como la afrocaribeña de Costa Rica (0,108) donde el fenotipo asociado al alelo está presente en un 21,8% de la muestra poblacional, residentes en EEUU con ancestría española (0,061) o poblaciones de origen étnico mixto del norte de África como la de Libia Cirenaica (0,042) o población de Sudáfrica de ancestría mixta (0,04). Este alelo está presente, si bien a bajas frecuencias en diversas poblaciones mestizas del centro y sur de América como nicaragüenses, costarricense y mexicanos. Por último, cabe destacar que este alelo, está presente en la población de Canarias (0,0093), donde el fenotipo asociado está presente en un 1,9% de la población.

Se asocia además a A*11:01:01:01, alelo presente en poblaciones con orígenes históricos en Europa, norte de África y sudoeste y este de Asia; a A*36:01 presente en diversas poblaciones de ancestría africanas como las de Kenia (0,07), de la isla de Sao Tomé (0,063) y afroamericanos de Bethesda en EEUU (0,05) donde el fenotipo asociado está presente en el 7,9% de la población.

Por último, en el mismo clúster está presente el alelo A*:01:01:01:01, que posee una frecuencia de 0,16 en población caucásica de San Francisco (EEUU), un 0,152 en población italoamericana, un 0,1365 en población de Polonia (donde el fenotipo asociado a este alelo está presente en el 25,5% de la población) y un 0,114 en la población de Libia Cirenaica del norte de África.

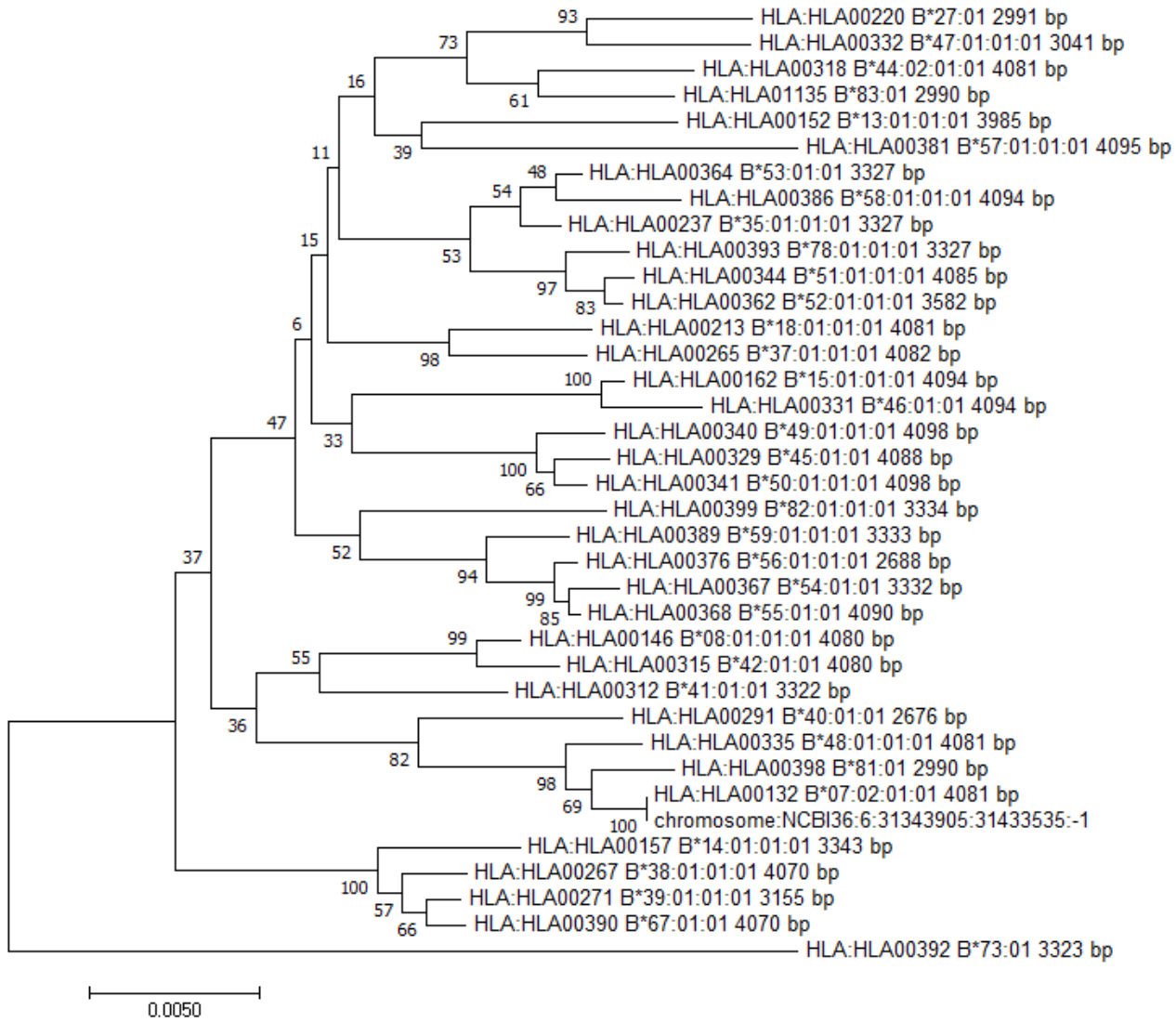


Figura R-43: Árbol de relaciones evolutivas para el gen HLA-B. Se presenta el árbol óptimo con la suma de longitud de rama = 0,21143180. Se han analizado 37 secuencias nucleotídicas, y un total de 2657 loci. El archivo original de HLA-A de humanos modernos contaba con 1144 alelos. La preselección ha consistido en elegir un alelo de cada grupo, según el código de alelo, teniendo en cuenta el primer par de dígitos de dicho código.

Las relaciones representadas en la Figura R-43 presentan una robustez variable, por ejemplo, el clúster en el que se agrupa el gen HLA-B Neandertal presenta un valor de presencia en las distintas réplicas de entre el 100 y el 36% dependiendo el nivel que observemos, y en general, el árbol de HLA-B presenta valores muy dispares a todos los niveles. Podemos observar como el alelo HLA-B de Neandertal, que aparece identificado como *chromosome:NCBI36:6:31343905:31433535:-1*, se encuentra asociado con B*07:02:01:01 con un valor de *bootstrap* de 100. Este alelo está presente en habitantes preeuropeos de Australia y poblaciones con orígenes históricos en el este y sudoeste de Asia, Europa y norte de África. Por ejemplo, en Polonia presenta una frecuencia alélica de 0,1138, con presencia fenotípica en el 21,4%

de la población. Resulta un tanto extraño que se encuentre presente en población aborigen australiana, cuya insularidad, hace pensar que estarían aislados. Sin embargo, los valores de frecuencias aléllicas observados en las poblaciones europeas y asiáticas, es coincidente con el rango de distribución de Neandertal, y por tanto, el alelo B*07:02:01:01 es un buen candidato a ser el alelo introgresado a partir del de Neandertal.

También se relaciona con B*81:01, presente en poblaciones con origen histórico en África subsahariana, como miembros de la tribu Luo de Kenia (0,05) donde el fenotipo asociado aparece en el 10% de la población, población de Yaounde en Camerún (0,044), o población Zulú de Sudáfrica (0,035) donde el fenotipo aparece en el 7% de la población. También llama la atención el caso de los Semang de Pahang (Malasia). En esta población la frecuencia de este alelo es de 0.079 y la presencia fenotípica alcanza el 15,8% de la población. Cabe señalar que los Semang tienen características asociadas a los primeros pobladores de Asia que colonizaron progresivamente la costa sur del continente hace 60000 años provenientes de África, y que presentan afinidad con grupos como los Jarawa de las islas Andamán y los Aeta de Filipinas. De manera marginal, B*81:01 también está presente en poblaciones mestizas como los mulato de Cuba (0,012) donde la frecuencia fenotípica es del 2,4%, población sudafricana de ancestría mixta (0,01) y algunas poblaciones de la península arábiga como la de Omán (0,013) donde el fenotipo está presente en el 1,3% de la población o en Abu Dhabi (Emiratos Árabes Unidos) donde la frecuencia alélica es 0,0096.

En el mismo clúster se sitúa B*48:01:01:01 presente en poblaciones con orígenes geográficos muy diversos: residentes preeuropeos americanos, poblaciones con origen histórico en el este de Asia, y diversas poblaciones mestizas. Llama la atención la bajísima frecuencia alélica (0,0007) de la población de Polonia, donde el fenotipo alcanza únicamente al 0,1% de la población.

Además podemos observar que se relaciona también con el alelo B*40:01:01, presente en poblaciones caucasoides con orígenes en Europa, norte de África o el sudoeste de Asia, incluyendo la India, y orientales con orígenes en el este de Asia. Así, la frecuencia alélica es de 0,155 entre los chinos Han de la provincia de Guangdong, o de 0,081 entre los Han del norte de China. En la población de Sunda y Java la frecuencia alélica es de 0,035, y en los Kannada de Karnataka en el suroeste de la India la frecuencia alélica es de 0,014, alcanzando el fenotipo al 2,9% de la población. Cabe destacar también la presencia de este alelo en población de origen hispánico: en la población con ancestría española residente en EEUU la frecuencia alélica es de 0,12, en la población mestiza mexicana de EEUU es del 0,016, y en los mestizos de Valle Central de Costa Rica es de 0,014 con el fenotipo presente en el 2,8% de la población. Por último señalar la presencia en poblaciones del este de Europa: en residentes de EEUU con este origen étnico la frecuencia alélica es del 0,044 y entre los gitanos de Bulgaria la frecuencia alélica alcanza el 0,045.

Por último puede destacarse el caso de B*73:01 que se sitúa como *outlier* del resto de alelos estudiados. Este alelo está presente en poblaciones de origen judío como los Kavkazi (0,021), los Ashkenazi (0,023) y las poblaciones judías de Iraq y árabe de Israel (0,0094). También está presente entre los Seri del desierto de Sonora en México (0,015) y los Beti-Pahuin de Camerún (0,011).

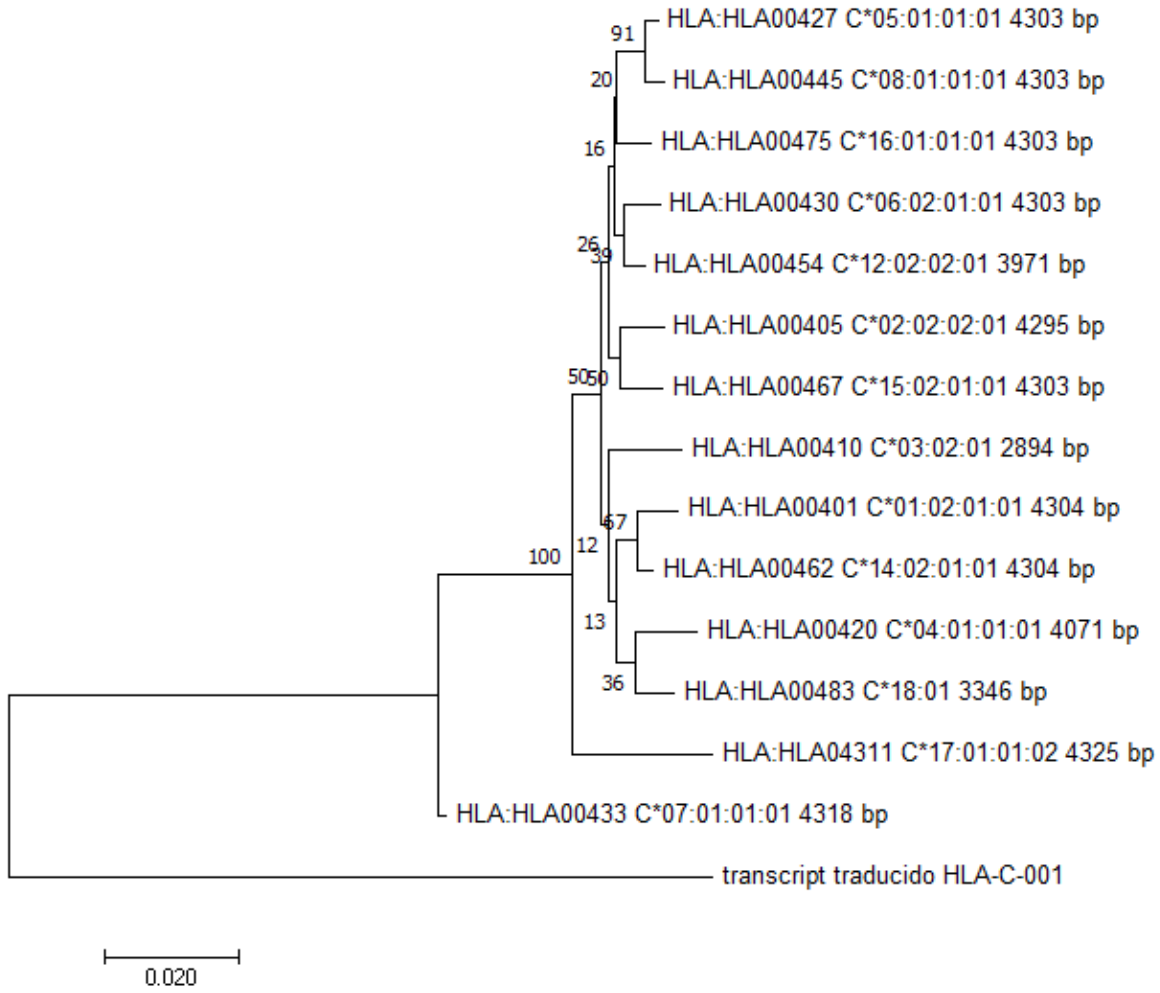


Figura R-44: Árbol de relaciones evolutivas para el gen HLA-C. Se presenta el árbol óptimo con la suma de longitud de rama = 0,30581405. Se han analizado 15 secuencias nucleotídicas, y un total de 1098 loci. El archivo original de HLA-A de humanos modernos contaba con 1141 alelos. La preselección ha consistido en elegir un alelo de cada grupo, según el código de alelo, teniendo en cuenta el primer par de dígitos de dicho código.

Las relaciones representadas en la Figura R-44 presentan una robustez relativamente baja, y además el alelo de HLA-C de Neandertal (“*transcript traducido HLA-C-001”)* se presenta como *outlier*. Este hecho junto a que, en este caso, se hizo imposible conseguir los datos del gen HLA-C Neandertal de manera directa, teniendo que hacerse una traducción inversa a partir de la

secuencia de ARN mensajero de la que se disponía, nos hace pensar que la calidad de la secuencia de HLA-C Neandertal no sea lo suficientemente alta como para poder aceptar el análisis.

Respecto a las relaciones del alelo HLA-C Neandertal, esté alelo está más relacionado con C*07:01:01:01 y C*17:01:01:02, alelos que están presentes en poblaciones de África subsahariana, del este y sudoeste de Asia, del norte de África y de Europa, destacando que en Polonia la frecuencia alélica es de 0,0744 y 0,0015 con una frecuencia fenotípica de 14,1% y 0,3% en la población, respectivamente.

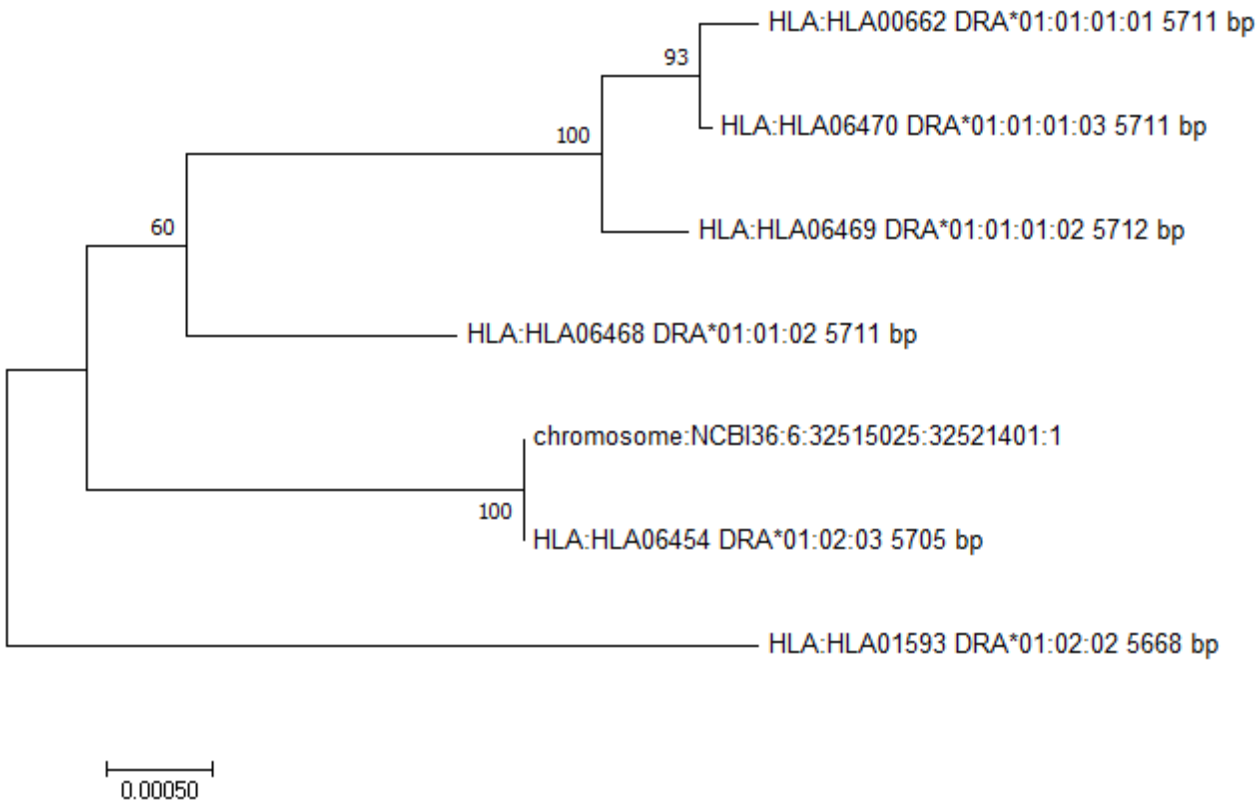


Figura R-45: Árbol de relaciones evolutivas para el gen HLA-DRA. Se presenta el árbol óptimo con la suma de longitud de rama = 0,01100521. Se han analizado 7 secuencias nucleotídicas, y un total de 5668 loci. El archivo original de HLA-A de humanos modernos contaba con 6 alelos. No se ha realizado preselección dado que la baja cantidad de alelos de HLA-DRA de humanos modernos disponibles ha permitido incluirlos todos en el árbol.

Las relaciones representadas en la Figura R-45 presentan una robustez, en general, muy alta, de entre el 100 y el 60% dependiendo el nivel que observemos. Podemos observar como el alelo HLA-DRA de Neandertal, al que se le ha dado el nombre de *chromosome:NCBI36:6:32515025:32521401:1*, se encuentra asociado, con un valor de *bootstrap* de 100, a DRA*01:02:03. Si bien no ha sido posible obtener datos de frecuencias de este alelo en concreto, el alelo DRA*01:02, relacionado íntimamente con DRA*01:02:03. DRA*01:02 está presente poblaciones caucasoides de origen europeo, norteafricano o del sudoeste de Asia,

incluyendo la India, por lo que DRA*01:02:03 podría ser un buen candidato a alelo introgresado a partir del alelo HLA-DRA de Neandertal. En otro grupo aparecen DRA*01:01:02, DRA*01:01:01:02 y DRA*01:01:01:03, presentes también en poblaciones caucasoides de origen europeo, norteafricano o del sudoeste de Asia, incluyendo la India.

Respecto a DRA*01:01:01:01, cabe señalar que si bien comparte la presencia en poblaciones caucasoides de origen europeo, norteafricano o del sudoeste de Asia, incluyendo la India, también se observa en poblaciones cuyo origen histórico se sitúa en África subsahariana. Así entre la población afroamericana la frecuencia alélica alcanza el 0,1053.

Por último, el *outlier* representado por el alelo DRA*01:02:02 comparte la presencia en poblaciones caucasoides de origen europeo, norteafricano o del sudoeste de Asia, incluyendo la India.

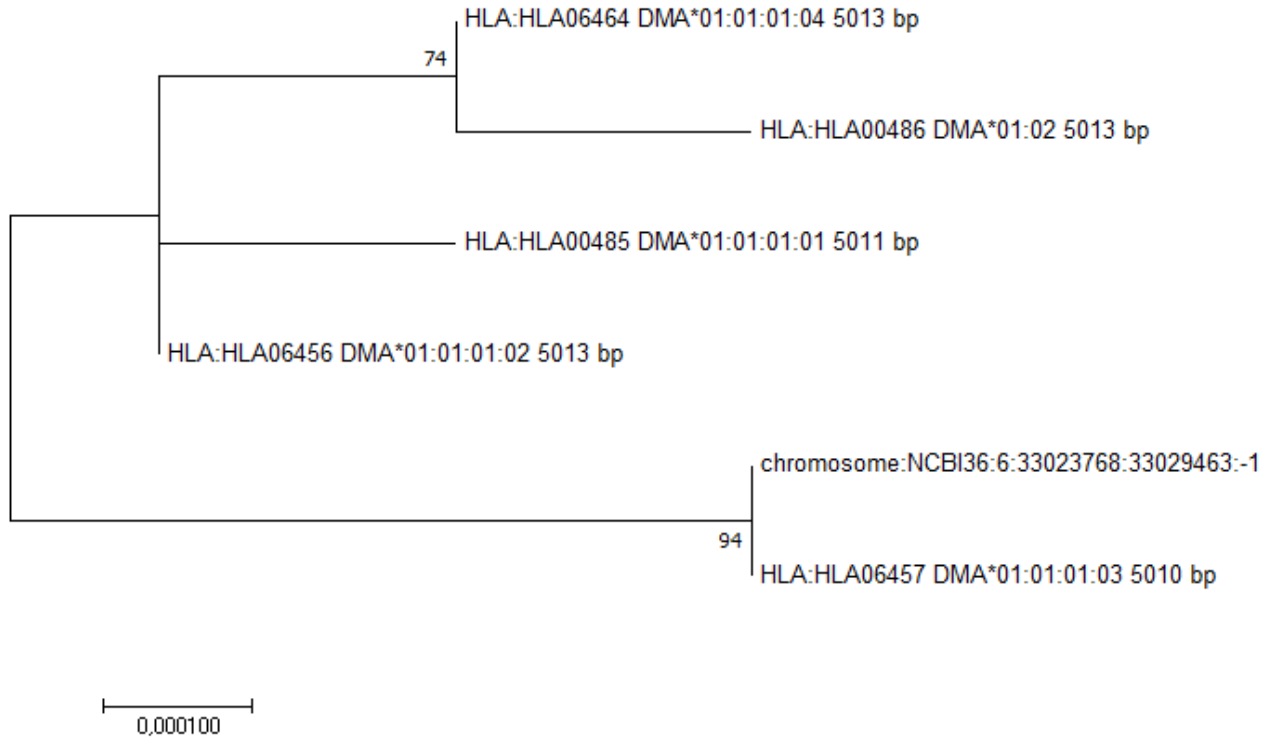


Figura R-46: Árbol de relaciones evolutivas para el gen HLA-DMA. Se presenta el árbol óptimo con la suma de longitud de rama = 0,00119891. Se han analizado 6 secuencias nucleotídicas, y un total de 5008 loci. El archivo original de HLA-A de humanos modernos cuenta con 5 alelos. No se ha realizado preselección dado que la baja cantidad de alelos de HLA-DMA de humanos modernos conseguidos ha permitido incluirlos todos en el árbol.

Las relaciones representadas en la Figura R-46 presentan una robustez, en general, muy alta, de entre el 94 y el 74% dependiendo el nivel que observemos. También se aprecia la presencia de varias agrupaciones polimórficas. Podemos observar como el alelo HLA-DMA de Neandertal, al que

se le ha dado el nombre de *chromosome:NCBI36:6:33023768:33029463:-1*, se encuentra asociado a DMA*01:01:01:03 con un valor de bootstrap de 94. Este alelo está presente en poblaciones caucasoides de origen europeo, norteafricano o del sudoeste de Asia, incluyendo la India, lo cubre el rango geográfico para un posible entrecruzamiento entre Neandertales y humanos anatómicamente modernos, por lo que este alelo es un buen candidato para provenir de la introgresión del alelo Neandertal. DMA*01:01:01:02, DMA*01:01:01:01 y DMA*01:01:01:04 presentan una distribución similar (poblaciones caucasoides de origen europeo, norteafricano o del sudoeste de Asia, incluyendo la India) en humanos modernos.

El alelo DMA*01:01, relacionado con todos los anteriores, está presente en poblaciones de origen europeo, de Oriente Próximo y del este de Asia con muy alta frecuencia. Así, en la población caucasoide de Virginia en EEUU la frecuencia de este alelo es de 0,854, en la población de Turquía alcanza una frecuencia de 0,849, en Alemania es de 0,7902 alcanzando la frecuencia fenotípica al 95,6% de la población. En China, los Han del norte tiene una frecuencia alélica de 0,7283, mientras que los Han del sur tienen una frecuencia de 0,6622.

En cuanto al alelo DMA*01:02, ocurre algo similar al caso anterior, aunque las frecuencias son más bajas en todos los casos. En la población caucasoide de Virginia en EEUU la frecuencia de este alelo es de 0,1050, en la población de Turquía alcanza una frecuencia de 0,094, en Alemania es de 0,1567 alcanzando la frecuencia fenotípica al 28,9% de la población. En China, los Han del norte tiene una frecuencia alélica de 0,2505, mientras que los Han del sur tienen una frecuencia de 0,3074.

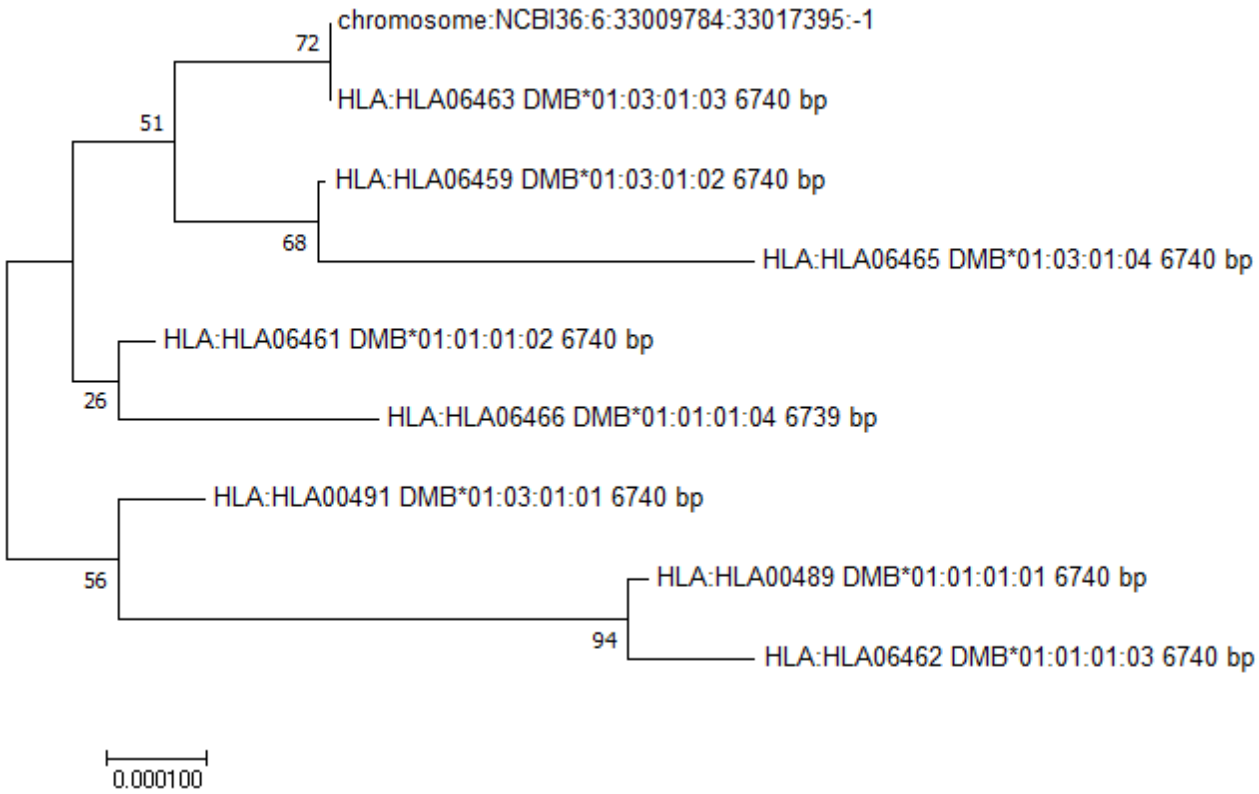


Figura R-47: Árbol de relaciones evolutivas para el gen HLA-DMB. Se presenta el árbol óptimo con la suma de longitud de rama = 0,00210647. Se han analizado 9 secuencias nucleotídicas, y un total de 6739 loci. El archivo original de HLA-A de humanos modernos contaba con 8 alelos. No se ha realizado preselección dado que la baja cantidad de alelos de HLA-DMB de humanos modernos conseguidos ha permitido incluirlos todos en el árbol.

Las relaciones representadas en la Figura R-47 presentan una robustez variable, de entre el 94 y el 26% dependiendo el nivel que observemos. Podemos observar como el alelo HLA-DMB de Neandertal, al que se le ha dado el nombre de *chromosome:NCBI36:6:33009784:33017395:-1*, se asocia con DMB*01:03:01:03 con un valor de *bootstrap* de 72. Si bien este valor no es tan alto como en casos anteriores, la distribución geográfica actual de DMB*01:03:01:03 se superpone con la localización geográfica de los lugares putativos de entrecruzamiento entre Neandertal y humanos modernos, por lo que este alelo es un buen candidato para provenir de la introgresión del alelo Neandertal. Relacionados con ambos, también se encuentran DMB*01:03:01:02 y DMB*01:03:01:04, ambos presentes, al igual que DMB*01:03:01:03, en poblaciones caucasoides de origen europeo, norteafricano o del sudoeste de Asia, incluyendo la India.

El alelo DMB*01:03, relacionado con los anteriores, está presente en poblaciones de origen europeo, de Oriente Próximo y del este de Asia. Así, en la población Han del norte de China tiene una frecuencia alélica de 0,3256, mientras que los Han del sur tienen una frecuencia de 0,3074. En la población caucasoide de Virginia en EEUU la frecuencia de este alelo es de 0,196, en la

población de Turquía alcanza una frecuencia de 0,156, y en Alemania es de 0,0227 alcanzando la frecuencia fenotípica al 4,5% de la población.

Los alelos DMB*01:01:01:02, DMB*01:01:01:01 y DMB*01:01:01:03 están de igual modo presentes en poblaciones de origen europeo, de Oriente Próximo y del este de Asia, con muy alta frecuencia. Así, en la población de Alemania la frecuencia es de 0,8516 alcanzando la frecuencia fenotípica al 97,8% de la población. En la población de Turquía alcanza una frecuencia de 0,809. En la población caucasoide de Virginia en EEUU la frecuencia de este alelo es de 0,749. Por último, en los Han del sur de China tiene una frecuencia alélica de 0,5275, mientras que los Han del norte tienen una frecuencia de 0,5222.

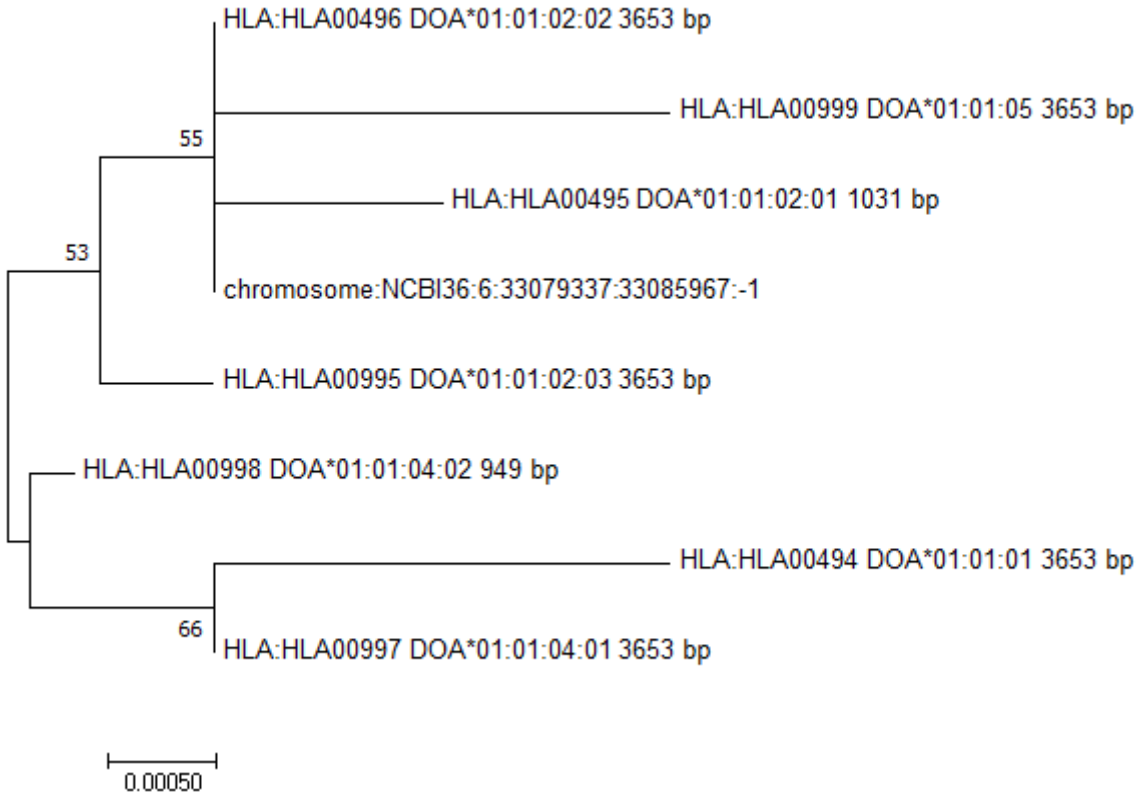


Figura R-48: Árbol de relaciones evolutivas para el gen HLA-DOA. Se presenta el árbol óptimo con la suma de longitud de rama = 0,00792589. Se han analizado 8 secuencias nucleotídicas, y un total de 949 loci. El archivo original de HLA-A de humanos modernos contaba con 7 alelos. No se ha realizado preselección dado que la baja cantidad de alelos de HLA-DOA de humanos modernos conseguidos ha permitido incluirlos todos en el árbol.

Las relaciones representadas en la Figura R-48 presentan una robustez media, de entre el 66 y el 53% dependiendo el nivel que observemos. También se aprecia la presencia de varias agrupaciones polimórficas. Podemos observar como el alelo HLA-DOA de Neandertal, al que se le ha dado el nombre de *chromosome:NCBI36:6:33079337:33085967:-1*, se sitúa en un clúster polimórfico junto con DOA*01:01:02:02, DOA*01:01:05 y DOA*01:01:02:01.

Mientras que DOA*01:01:05 presenta un patrón de distribución caucasoide, hallándose en poblaciones caucasoides de origen europeo, norteafricano o del sudoeste de Asia, incluyendo la India; DOA*01:01:02:02 está presente además en poblaciones del este de Asia; y DOA*01:01:02:01 además del patrón de distribución caucasoide y estar presente en poblaciones del este de Asia, se halla en poblaciones preeuropeas de América. Esta distribución junto al hecho de que el grupo sea polimórfico y su valor de bootstrap sea algo mediocre, no permite afirmar que alguno de los alelos mencionados sea candidato a provenir de una introgresión de Neandertal, si bien la distribución geográfica de estos alelos es coincidente, en algunos casos, con la de los lugares donde se habría dado este entrecruzamiento.

Los alelos DOA*01:01:02:03, DOA*01:01:01 y DOA*01:01:04:01 tienen igualmente un patrón de distribución caucasoide como el anteriormente mencionado.

No ha sido posible atribuir un origen putativo al alelo DOA*01:01:04:02.

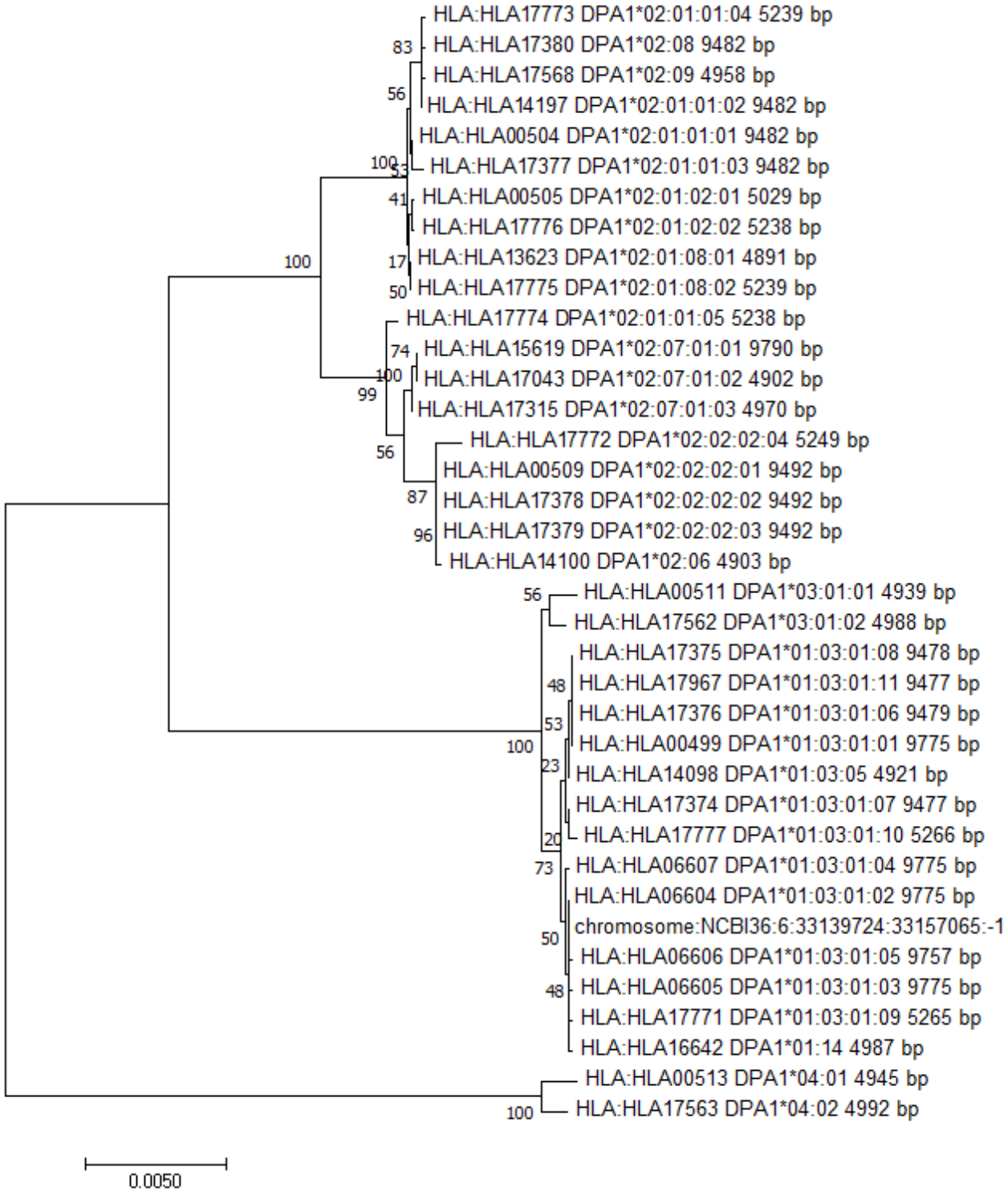


Figura R-49: Árbol de relaciones evolutivas para el gen HLA-DPA1. Se presenta el árbol óptimo con la suma de longitud de rama = 0,06209569. Se han analizado 37 secuencias nucleotídicas, y un total de 4864 loci. El archivo original de HLA-A de humanos modernos contaba con 36 alelos. No se ha realizado preselección dado que la baja cantidad de alelos de HLA-DPA1 de humanos modernos conseguidos ha permitido incluirlos todos en el árbol.

Las relaciones representadas en la Figura R-49 presentan una robustez variable, por ejemplo, el clúster en el que se agrupa el gen HLA-DPA1 Neandertal presenta un valor de presencia en las distintas réplicas de entre el 100 y el 20% dependiendo el nivel que observemos, y en general, el árbol de HLA-DPA1 presenta valores dispares a todos los niveles. Se aprecia cómo los 37 alelos estudiados se dividen en 3 macrogrupos: uno en la parte superior del árbol formado por 19 alelos, un segundo grupo de 16 alelos (incluyendo el alelo de HLA-DPA1 de Neandertal) y un último grupo formado por el *outlier* de DPA1*04:01 y DPA1*04:02.

Podemos observar como el alelo HLA-DPA1 de Neandertal, al que se le ha dado el nombre de *chromosome:NCBI36:6:33139724:33157065:-1*, se sitúa en un clúster polimórfico junto a DPA1*01:03:01:02, DPA1*01:03:01:05, DPA1*01:03:01:03, DPA1*01:03:01:09 y DPA1*01:14. Todos estos alelos están presentes en poblaciones con orígenes históricos europeos, del norte de África y del sudoeste de Asia, incluyendo la India. Además, en el caso de DPA1*01:03:01:02 también se halla en poblaciones con origen en África subsahariana. De nuevo ocurre algo similar a lo visto en HLA-DOA: el bajo valor de bootstrap (48) y el carácter polimórfico del grupo donde se sitúa el alelo HLA Neandertal no permiten asignar a ninguno de los alelos HLA-DPA1 de humanos modernos el posible origen introgresado, aunque de nuevo, la distribución geográfica de estos es coincidente con la zona donde se habría dado el entrecruzamiento.

El alelo DPA1*01:03:01, relacionado con todos los anteriores salvo con DPA1*01:14, está presente en poblaciones amerindias, europeas e isleñas del Pacífico. Así, en los Pima de la zona de rio Gila en Arizona (EEUU) la frecuencia alélica de DPA1*01:03:01 es de 0,9783. En Eslovenia es de 0,8267, llegando la frecuencia fenotípica al 97% de la población. En cuanto a las poblaciones isleñas del Pacífico, este alelo se encuentra en Tokelau (0,73), las islas Cook (0,54), Tonga (0,46) y Samoa (0,42).

En cuanto al *outlier* formado por DPA1*04:01 y DPA1*04:02 cabe destacar que son dos alelos presentes en multitud de poblaciones de orígenes diversos (Papúa-Nueva Guinea, Uganda, Hong Kong, Seri de México, Reino Unido, Japón,...) pero a muy bajas frecuencias: alcanza un máximo de frecuencia alélica de 0,045 en Papúa-Nueva Guinea y un mínimo de 0,0013 en Japón.

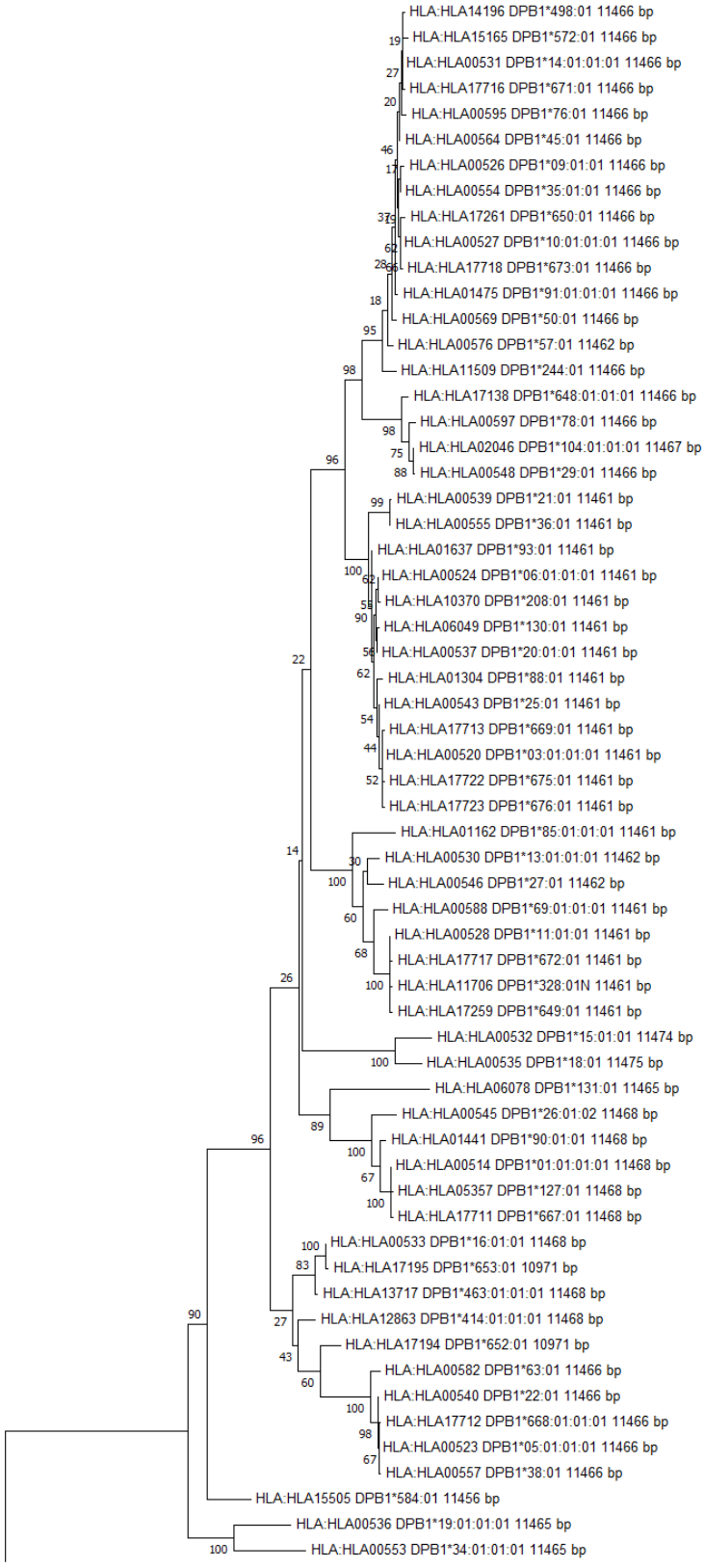


Figura R-50: Árbol de relaciones evolutivas para el gen HLA-DPB1. Se presenta el árbol óptimo con la suma de longitud de rama = 0,009131333. Se han analizado 117 secuencias nucleotídicas, y un total de 10937 loci. El archivo original de HLA-DPB1 de humanos modernos contaba con 290 alelos. La preselección ha consistido en elegir un alelo de cada grupo, según el código de alelo, teniendo en cuenta el primer par de dígitos de dicho código.

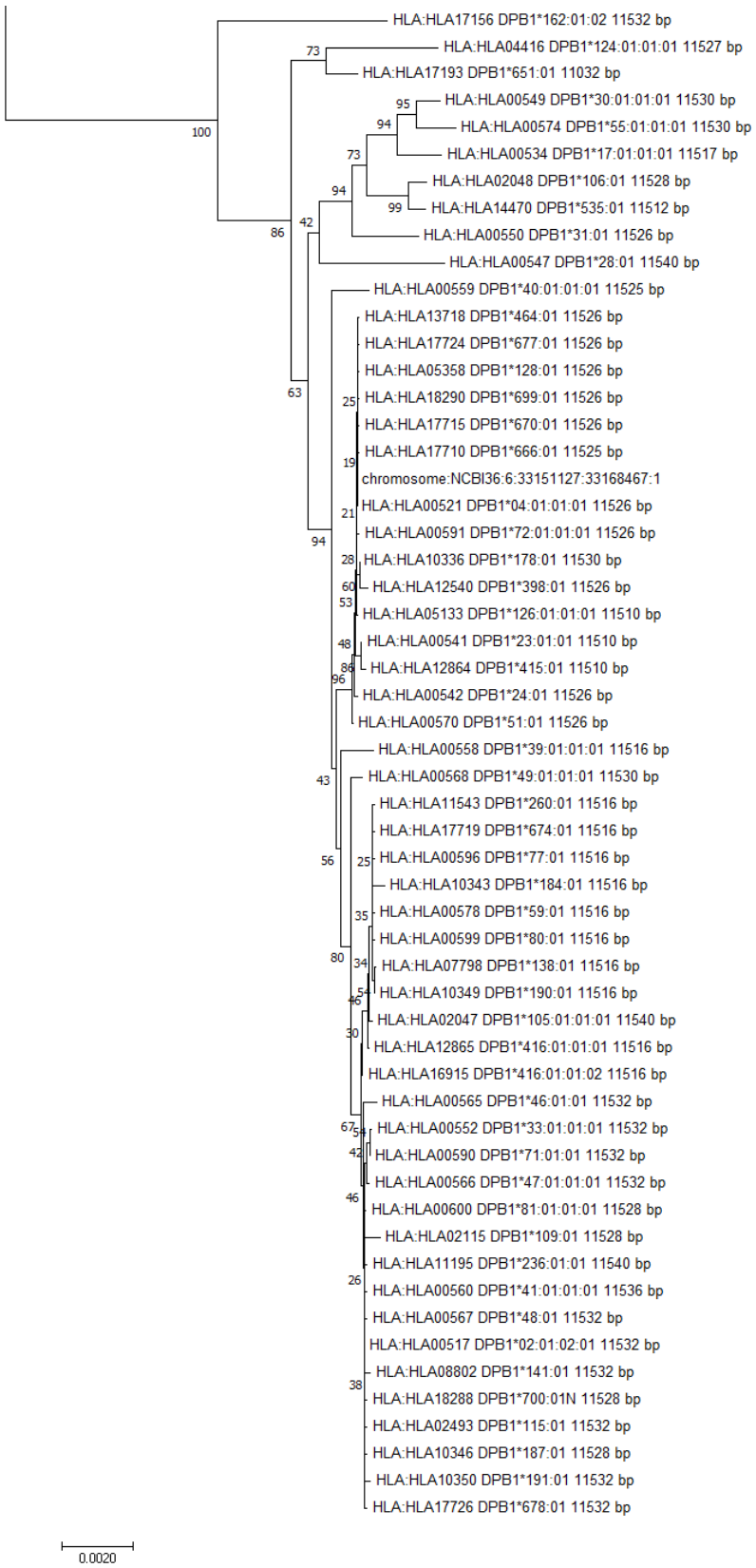


Figura R-51: Continuación de Figura R-50. Árbol de relaciones evolutivas para el gen HLA-DPB1. Se presenta el árbol óptimo con la suma de longitud de rama = 0,009131333. Se han analizado 117 secuencias nucleotídicas, y un total de 10937 loci. El archivo original de HLA-DPB1 de humanos modernos contaba con 290 alelos. La preselección ha consistido en elegir un alelo de cada grupo, según el código de alelo, teniendo en cuenta el primer par de dígitos de dicho código.

Las relaciones representadas en las Figuras R-50 y R-51 presentan una robustez variable, por ejemplo, el clúster en el que se agrupa el gen HLA-DPB1 Neandertal presenta un valor de presencia en las distintas réplicas de entre el 96 y el 19% dependiendo el nivel que observemos, y en general, el árbol de HLA-DPB1 presenta valores dispares a todos los niveles. Se aprecia cómo los 117 alelos estudiados se dividen en 2 macrogrupos: uno en la parte superior del árbol representado en la Figura R-50 y un segundo grupo, en el que se incluye el alelo de HLA-DPB1 de Neandertal, en la parte inferior del árbol, representado en la Figura R-51.

Podemos observar como el alelo HLA-DPB1 de Neandertal, al que se le ha dado el nombre de *chromosome:NCBI36:6:33151127:33168467:1*, se sitúa en un clúster junto a DPB1*04:01:01:01, que está presente en poblaciones con orígenes históricos europeos, del norte de África y del sudoeste de Asia, incluyendo la India; y además, poblaciones caucasoides mediterráneas y amerindias. Si bien la distribución de DPB1*04:01:01:01 es coincidente con la localización putativa de los entrecruzamientos entre Neandertales y humanos, el valor de bootstrap tan bajo (19) y el hecho de que sea un grupo polimórfico, no permiten sugerir con seguridad que sea un candidato a alelo introgresado. No ha sido posible asignar un origen étnico putativo a los alelos DPB1*666:01, DPB1*670:01, DPB1*699:01, DPB1*128:01, DPB1*677:01 y DPB1*464:01, que se encuentran próximos en el árbol al alelo de HLA-DPB1 de Neandertal.

Comentar que otros alelos como DPB1*72:01:01:01, cercano al clúster mencionado anteriormente, tiene orígenes caucasoides. El alelo DPB1*126:01:01:01 está presente en poblaciones de la región de África subsahariana. Y el alelo DPB1*23:01:01 está presente en poblaciones con orígenes históricos en Europa, norte de África y sudoeste de Asia (incluyendo la India), e incluso, en población de la región de Mongolia Interior (China) con una frecuencia alélica de 0,001 y una presencia fenotípica de 0,2%.

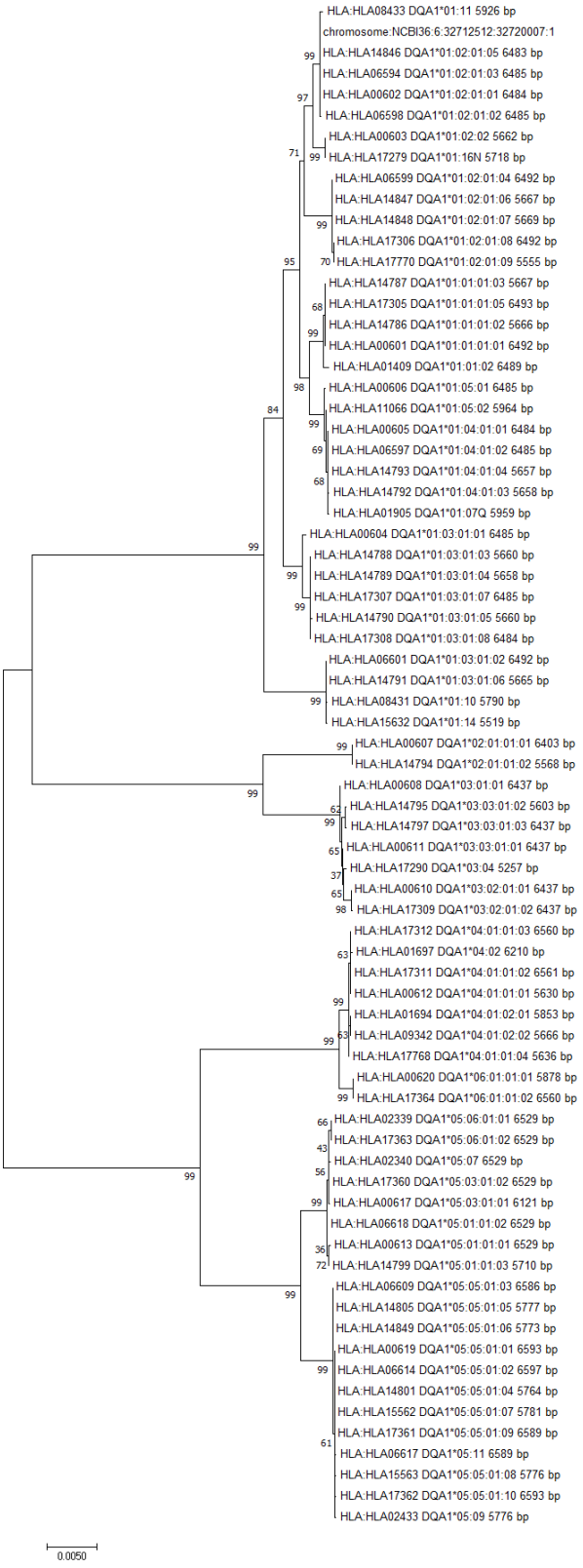


Figura R-52: Árbol de relaciones evolutivas para el gen HLA-DQA1. Se presenta el árbol óptimo con la suma de longitud de rama = 0,15029865. Se han analizado 73 secuencias nucleotídicas, y un total de 5123 loci. El archivo original de HLA-DPB1 de humanos modernos contaba con 290 alelos. No se ha realizado preselección dado que la baja cantidad de alelos de HLA-DQA1 de humanos modernos conseguidos ha permitido incluirlos todos en el árbol.

Las relaciones representadas en las Figura R-52 presentan una robustez variable, por ejemplo, el clúster en el que se agrupa el gen HLA-DQA1 Neandertal presenta un valor de presencia en las distintas réplicas de entre el 99 y el 70% dependiendo el nivel que observemos, aunque en general, el árbol de HLA-DQA1 presenta valores altos de robustez a casi todos los niveles. Se aprecia cómo los 73 alelos estudiados se dividen en 2 macrogrupos: uno en la parte superior del árbol, en el que se incluye el alelo de HLA-DQA1 de Neandertal, y un segundo macrogrupo en la parte inferior del árbol. Se observan algunos grupos polimórficos en ambos macrogrupos.

Podemos observar como el alelo HLA-DQA1 de Neandertal, al que se le ha dado el nombre de *chromosome:NCBI36:6:32712512:32720007:1*, se sitúa en un clúster junto a DQA1*01:11, DQA1*01:02:01:05 y DQA1*01:02:01:01 que está presente en poblaciones con orígenes históricos en Europa, norte de África y sudoeste de Asia (incluyendo la India). También observamos que se relaciona con DQA1*01:02:01:02 y DQA1*01:02:01:03, estando ambas presentes en poblaciones de origen subsahariano y DQA1*01:02:01:02, además, en poblaciones con orígenes históricos en el este de Asia. Si bien este grupo es polimórfico e incluye algunos alelos con distribuciones no coincidentes con la distribución geográfica de los entrecruzamientos entre Neandertales y humanos, el hecho de que tenga un valor de *bootstrap* muy alto (99) y que incluya alelos con distribuciones coincidentes con la de los entrecruzamientos, hace sospechar que alguno de esos alelos es el candidato a alelo introgresado.

El alelo DQA1*01, asociado a los anteriores, tiene una distribución cosmopolita: posee una frecuencia alélica de 0,59 en población surasiática de Trinidad, un 0,51 en población de Turquía, un 0,474 en población mestiza brasileña de Sao Paulo (donde se observa el fenotipo en el 86,1% de la población), un 0,441 en Bélgica donde el fenotipo asociado a este alelo está presente en el 68,1% de la población, un 0,433 en los bretones del oeste de Francia, un 0,408 en los Tuva de Rusia, un 0,389 en la población de Marruecos, o un 0,218 en la población China general, donde el fenotipo asociado a este alelo se observa en un 38,8% de la población.

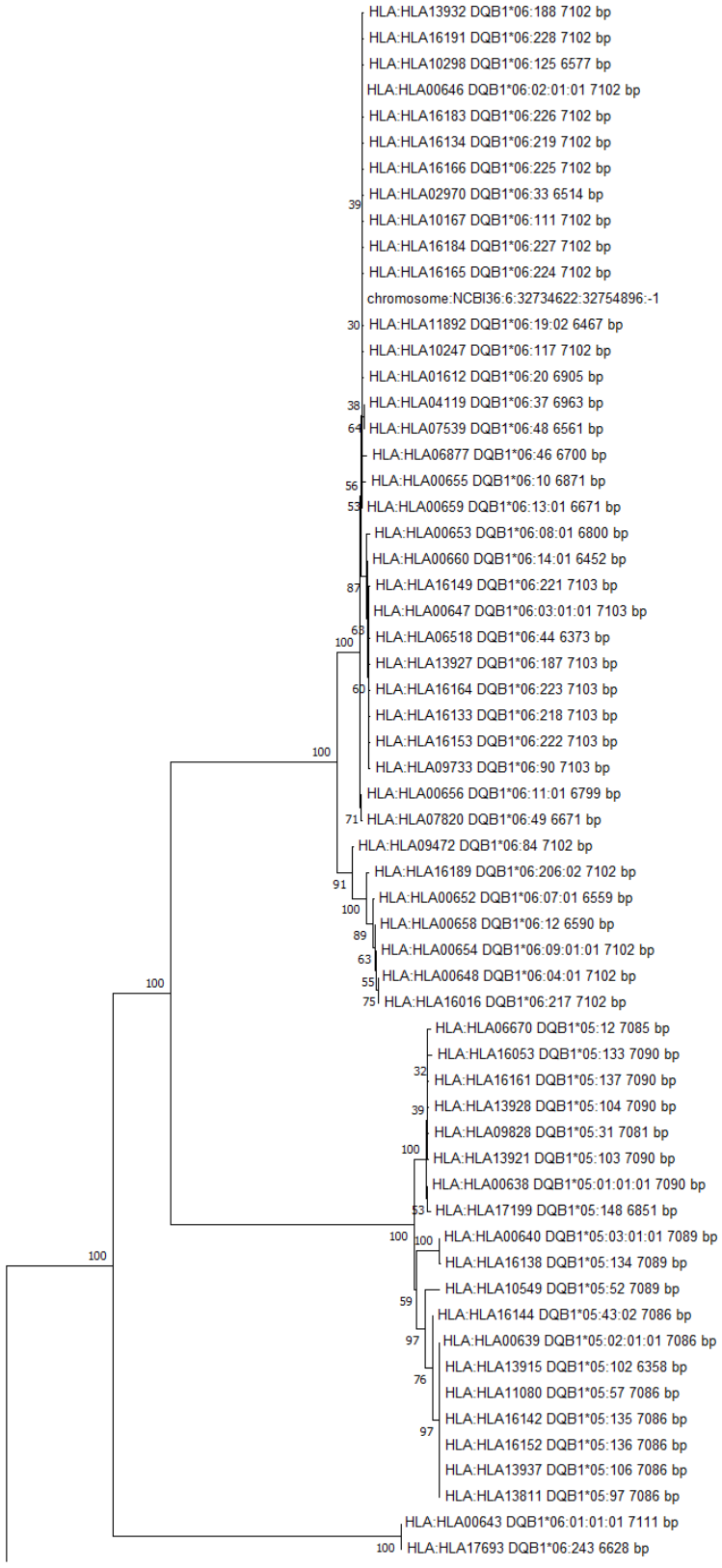


Figura R-53: Árbol de relaciones evolutivas para el gen HLA-DQB1. Se presenta el árbol óptimo con la suma de longitud de rama = 0,23178144. Se han analizado 105 secuencias nucleotídicas, y un total de 5346 loci. El archivo original de HLA-DPB1 de humanos modernos contaba con 196 alelos. La preselección ha consistido en elegir un alelo de cada grupo, según el código de alelo, teniendo en cuenta los dos primeros pares de dígitos de dicho código.

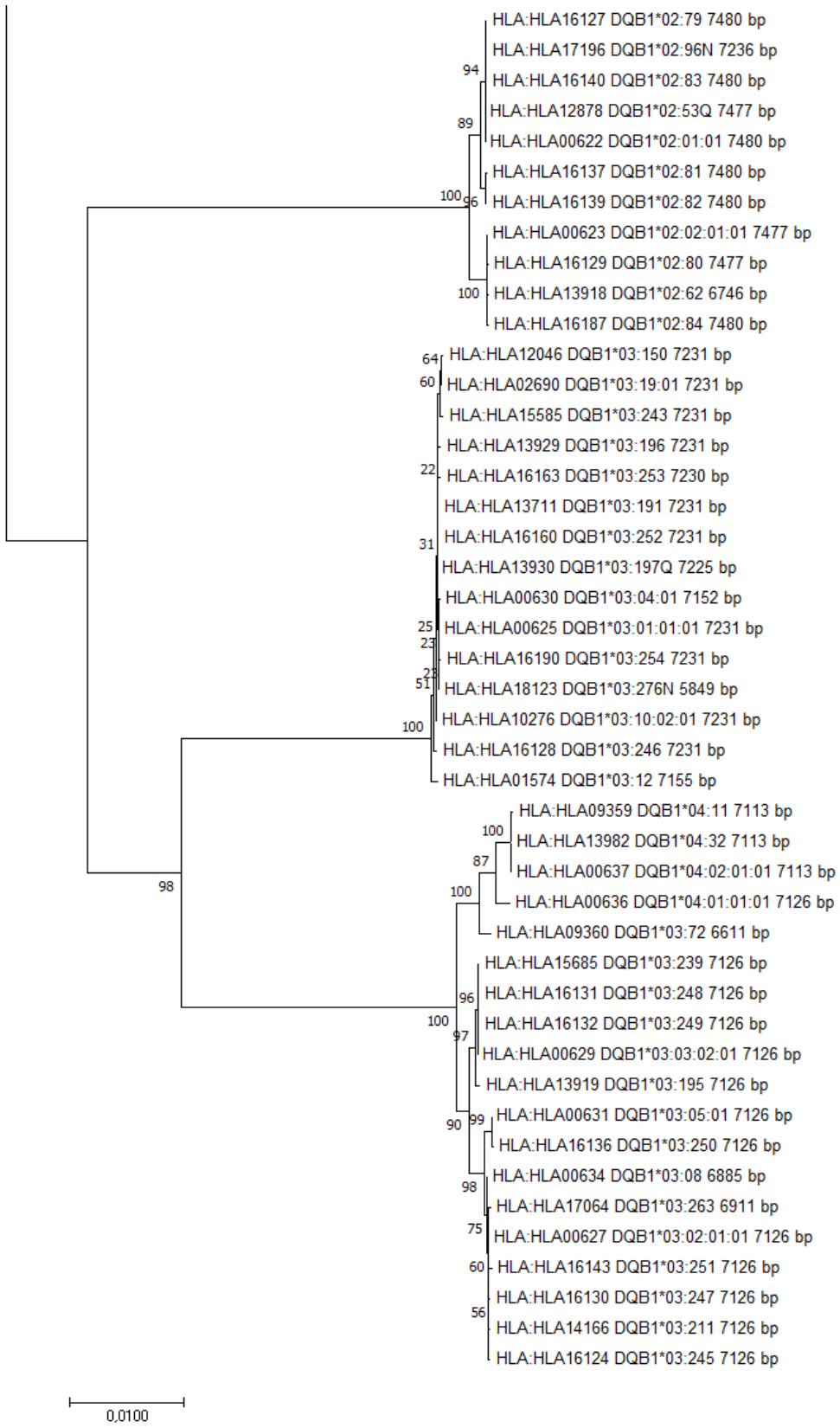


Figura R-54: Continuación de Figura R-53. Árbol de relaciones evolutivas para el gen HLA-DQB1. Se presenta el árbol óptimo con la suma de longitud de rama = 0,23178144. Se han analizado 105 secuencias nucleotídicas, y un total de 5346 loci. El archivo original de HLA-DPB1 de humanos modernos contaba con 196 alelos. La preselección ha consistido en elegir un alelo de cada grupo, según el código de alelo, teniendo en cuenta los dos primeros pares de dígitos de dicho código.

Las relaciones representadas en las Figuras R-53 y R-54 presentan una robustez variable, por ejemplo, el clúster en el que se agrupa el gen HLA-DQB1 Neandertal presenta un valor de presencia en las distintas réplicas de entre el 100 y el 39% dependiendo el nivel que observemos, y en general, el árbol de HLA-DPB1 presenta valores dispares a todos los niveles. Se aprecia cómo los 105 alelos estudiados se dividen en 2 macrogrupos: uno en la parte superior del árbol, en el que se incluye el alelo de HLA-DQB1 de Neandertal, y un segundo macrogrupo en la parte inferior del árbol. Se observan algunos grupos polimórficos en ambos macrogrupos.

Podemos observar como el alelo HLA-DQB1 de Neandertal, al que se le ha dado el nombre de *chromosome:NCBI36:6:32734622:32754896:-1*, se sitúa en un gran clúster polimórfico junto a DQB1*06:188, DQB1*06:228, DQB1*06:125, DQB1*06:02:01:01, DQB1*06:226, DQB1*06:219, DQB1*06:225, DQB1*06:33, DQB1*06:111, DQB1*06:227 y DQB1*06:224. De este numeroso grupo únicamente es posible asignar un origen étnico a 3 alelos.

DQB1*06:02:01:01 está presente en poblaciones caucasoides con orígenes históricos en Europa, norte de África y sudoeste de Asia (incluyendo la India), y además en poblaciones subsaharianas. Destacar que en la población de Polonia presenta una frecuencia alélica de 0,1195 y el fenotipo asociado se observa en el 20,9% de la población.

DQB1*06:33 está presente en poblaciones caucasoides con orígenes históricos en Europa, norte de África y sudoeste de Asia (incluyendo la India), habitantes americanos preeuropeos, e irlandeses y británicos.

DQB1*06:111 está presente en poblaciones con orígenes históricos en el este de Asia.

El hecho de que este grupo sea polimórfico, posea un valor de bootstrap relativamente pobre (39), que sea un grupo tan numeroso y que falte información geográfica de tantos alelos hace altamente complicado asignar a alguno de los alelos el posible origen introgresado. Dicho esto, en los alelos en los que se conoce la distribución geográfica, esta coincide, al menos parcialmente, con la localización de los entrecruzamientos entre Neandertales y humanos.

El alelo DQB1*06, relacionado con todos los anteriores, está ampliamente distribuido por todo el mundo, salvo en poblaciones isleñas del Pacífico. Así por ejemplo, en la población de la región de Venda en la provincia de Limpopo (Sudáfrica) este alelo tiene una frecuencia de 0,437, entre la población sin presencia Raika de Bikaner (India) la frecuencia es 0,352 (con una prevalencia del fenotipo del 58% de la población), entre la población rusa de Osetia del Norte la frecuencia alélica alcanza el 0,3425 con el fenotipo asociado observándose en el 54,3% de la población, en la población mestiza de Sudán la frecuencia es de 0,318 con un fenotipo observable en el 60,5% de la población, entre los chinos de Linqu en la provincia de Shangdong la frecuencia es de 0,3158 con una presencia fenotípica del 53,2%, entre la población de Noruega la frecuencia es del 0,3034 y el fenotipo asociado alcanza el 51,2% de la población y entre la población de

Valencia (España) la frecuencia es de 0,2012, observándose el fenotipo asociado a este alelo en el 36,2% de la población.

DISCUSIÓN

DISCUSIÓN

Si bien en trabajos anteriores como el realizado por Graffelman et al (2017) ya se ha trabajado con datos en bruto de 1000Genomes en cuanto a desequilibrio Hardy-Weinberg a escala genómica, o se han usado datos de HLA para estudiar el desequilibrio de ligamiento (Gourraud et al. 2014), es en esta tesis doctoral cuando se han utilizado por primera vez datos en bruto de todas las poblaciones de la base de datos de 1000Genomes al mismo tiempo para analizar en profundidad los genes HLA clase I y II del CMH, a fin de establecer una caracterización poblacional de los desequilibrios como potenciales evidencias de selección.

Indels

La existencia de indels polimórficos cubriendo por completo las regiones de los genes HLA-C y HLA-DQA1 sugiere una serie de interesantes expectativas acerca de los posibles efectos de la hemicigosis en estos genes.

Se han encontrado algunos síndromes asociados a hemicigosis en alelos concretos de otros genes. Así, por ejemplo, según Yeung et al. en un estudio de 2013 examinan la prevalencia de eventos de pérdida de heterocigosidad (una de las características más importantes de la hemicigosis) y sus relaciones con la supervivencia general en glioblastoma multiforme (GBM). Este tipo de tumor posee una expresión menor de antígenos leucocitarios humanos clase I, lo que hace que las células T citotóxicas no lo detecten correctamente y limitando la eficacia de la inmunoterapia. Su estudio les permite afirmar que la pérdida de heterocigosidad en la región HLA clase I es frecuente en pacientes con glioblastoma multiforme adultos. Determinan que la asociación de una menor supervivencia en individuos con pérdida de heterocigosidad en esta región sugiere un papel crucial para estos genes en la inmunovigilancia.

En un trabajo de 2018, Berteau et al. hacen una revisión de la literatura existente entre la enfermedad de Behcet familiar autosómica dominante y haploinsuficiencia A20. La enfermedad de Behcet es un tipo de vasculitis sistémica que involucra vasos sanguíneos de cualquier tamaño con diversas características clínicas. La mayoría de los casos de BD son multifactoriales y están asociados con el antígeno HLA-B*51. La haploinsuficiencia es la situación en la que una sola copia del alelo estándar (*wild-type*) en un locus en combinación heterocigótica con un alelo variante es insuficiente para producir el fenotipo estándar, es decir, el individuo haploinsuficiente es incapaz de producir proteína en cantidad o calidad suficiente para asegurar la función normal.

En un artículo reciente (Hori et al. 2019), se menciona el caso concreto de una familia de tres generaciones en el que la haploinsuficiencia del gen que codifica la proteína 3 alfa-inducida para factor de necrosis tumoral (TNFAIP3, también conocido como A20) se relaciona con la presencia de la tiroiditis de Hashimoto, una enfermedad autoinmune, que se caracteriza por la destrucción de la glándula tiroides, mediada por autoanticuerpos. A20 es un regulador negativo de múltiples rutas señalizadoras intracelulares del sistema inmune, incluyendo señales de necrosis tumoral. Determinan que la tiroiditis de Hashimoto se relaciona con alelos HLA concretos como A2, A*02:07, DR3, DRB1*03:01, DR4 (DR53, DRB4), DR5, y DRB1*11:04; y, debido a que tanto los genes HLA como el gen TNFAIP3 están en el cromosoma 6, especulan que existe alguna asociación entre ambos.

Sin embargo, no hemos encontrado ninguna literatura sobre efectos de la hemicigosis en HLA-C y HLA-DQA1, a pesar de las altas frecuencias de ambas delecciones. En efecto, las frecuencias de cada deleción oscilan entre 0 y 0,402 en el caso de la que afecta a HLA-C y entre 0,289 y 0,492, con valores máximos en la población de indios Telugu (ITU) y la población de ancestría mexicana de Los Ángeles en EEUU (MXL), respectivamente.

Por otra parte, dada la elevada frecuencia de los indels mencionados, particularmente en algunas poblaciones, cabe esperar una cierta frecuencia de homocigotos para la deleción. Esto implicaría la no existencia del gen HLA-C o HLA-DQA1 en algunos individuos. Se han discutido en algunos trabajos las causas y efectos de las variaciones en la expresión de estos genes. Por ejemplo, Kaur et al. (2017) muestran que la variación en los exones 2 y 3, que codifican los dominios α1 / α2, impulsa la expresión diferencial de los alomorfos de HLA-C en la superficie celular al influir en la estructura de la unión a péptidos y la diversidad de péptidos unidos por el HLA-C moléculas.

Incluso se ha discutido la existencia de no expresión de genes HLA como una característica fundamental en algunos linfomas. En un estudio realizado por Riemersma et al. (2000) se establece que en los linfomas de células B, la pérdida de las moléculas de antígeno leucocitario humano (HLA) clase I y II podría contribuir al escape inmune de las células T citotóxicas CD8 (+) y CD4 (+), especialmente porque las células B pueden presentar su propio idiotipo. Investigaron la pérdida de la expresión de HLA y las posibles alteraciones genómicas subyacentes en 28 linfomas testiculares, 11 en el sistema nervioso central y 21 linfomas difusos nodales de células B grandes (DLCL). La pérdida total de la expresión de HLA-A se encontró en el 60% de los casos extranodales y en el 10% de los casos nodales, mientras que la pérdida de la expresión de genes HLA-DR se encontró en el 56% y el 5%, respectivamente. Esto fue acompañado por una pérdida extensa de heterocigosidad dentro de la región HLA en

los DLCL extranodales. En 3 casos, la retención de heterocigosidad para el microsatélite D6S1666 en la región de clase II sugirió una deleción homocigótica. Este hallazgo fue confirmado por un análisis FISH en interfase que mostró deleciones homocigóticas en los genes de clase II en 11 de los 18 linfomas extranodales pero en ninguno de los 7 DLCL nodales. Además se detectaron deleciones variables que siempre incluían genes HLA-DQ y HLA-DR. Se encontraron deleciones hemicigotas y recombinaciones mitóticas que a menudo involucran a todos los genes HLA en 13 de 18 linfomas extranodales y 2 de 7 nodales. Concluyen que una pérdida estructural de la expresión de HLA de clase I y II podría ayudar a las células de linfoma de células B a escapar del ataque inmune.

Todo esto abre la posibilidad de que en el caso de HLA-C y HLA-DQA1 exista una variabilidad fenotípica, tanto por una no expresión como por expresión diferencial, debida a la variación estructural producida por las grandes deleciones asociadas a estos genes. Esto supone una gran oportunidad de investigación en futuros estudios.

Hardy- Weinberg y selección

Hay diferentes trabajos en los que se ha especulado acerca de la posible relación entre desequilibrios HW y presiones selectivas.

Nielsen, Ehm & Weir (1998) discutieron sobre la idea de que la localización de un locus de susceptible a una enfermedad compleja puede realizarse de acuerdo a las desviaciones del equilibrio de Hardy-Weinberg entre los individuos afectados en comparación con los del resto de la. Es decir, un hallazgo de desequilibrio Hardy-Weinberg para un marcador implica la heterogeneidad de un marcador asociado a la enfermedad en cuestión, y por tanto, el desequilibrio de ligamiento para el par marcador-enfermedad. Aunque la falta de desviación del desequilibrio de Hardy-Weinberg en los loci marcadores implica que los desequilibrios de ligamiento ponderados por susceptibilidad a la enfermedad son cero, dada la heterogeneidad de la enfermedad, no se deduce que las medidas habituales de desequilibrio de ligamiento sean cero. Por lo tanto, para los loci de susceptibilidad a la enfermedad con más de dos alelos, se necesita cuidado en la extracción de inferencias a partir de desequilibrios Hardy-Weinberg. Se espera que el desequilibrio sea mayor en el locus susceptible a la enfermedad, ya que este es el factor que determina el criterio de selección en ese estudio. Los loci que son fenotípicamente neutros pero que de alguna manera están asociados con el locus susceptible a la enfermedad, como los marcadores genéticos en el desequilibrio de ligamiento con el locus de susceptible a la enfermedad, también experimentan una selección desproporcionada del genotipo. A medida que disminuye el grado de asociación entre la susceptibilidad a la enfermedad

y los distintos marcadores, también se espera que disminuya el desequilibrio de HW en el locus marcador.

En un trabajo de Ohashi et al. (2004) se establece que desequilibrio de ligamiento y la reducida diversidad haplotípica observada para la variante E de la hemoglobina (HbE) en la población tailandesa es debida a la presión selectiva positiva derivada de la endemicidad de la malaria en dicha población. Si bien reconocen que hay evidencia de múltiples orígenes para HbE en el sudeste asiático, afirman que los resultados obtenidos proporcionan evidencia sólida de un origen único de la variante HbE en la población tailandesa de referencia. Su análisis de simulación mostró que el intervalo de credibilidad del 95% de la edad estimada de la variante de HbE estaba entre 1.240 y 4.440 años. Sin una selección positiva contra la infección por malaria, la variante de HbE nunca se habría extendido tan rápidamente en esa población.

Como hemos visto en el apartado de Resultados, en total se han hallado 80647 puntos de presión evolutiva potencial (entendidos como marcadores en desequilibrio Hardy-Weinberg) distribuidos a lo largo de los 11 genes y 26 poblaciones estudiadas, por lo que podemos afirmar que, en efecto, los genes clase I y II de la región del Complejo Mayor de Histocompatibilidad son regiones hipervariables posiblemente sometidas a selección diversificadora, siendo la primera vez que se mapea el alcance completo del desequilibrio Hardy-Weinberg para estos genes en un conjunto de poblaciones de diferentes continentes. Se han hallado diferencias significativas en las distribuciones alélicas de multitud de marcadores, tanto entre poblaciones de un mismo grupo continental como entre grandes grupos continentales. Igualmente, estas diferencias son muy variables entre cada gen HLA estudiado. Asimismo, la presencia de distribuciones alélicas coincidentes entre ciertas poblaciones está en concordancia con las explicaciones aceptadas actualmente sobre el proceso *Out of Africa* de *Homo sapiens* (Macaulay et al. 2005, Posth et al. 2016, Rito et al. 2019, Haber et al. 2019).

Igualmente, hemos observado diferencias intra e intergénicas en cuanto al número de poblaciones que se encuentran en desequilibrio Hardy-Weinberg. Si bien solo se ha usado Haploview para visualizar un llamativo bloque de ligamiento en el gen HLA-DPB1 (Figura R-10), se han hallado numerosos bloques más pequeños distribuidos por todos los genes, indicando que hay marcadores que se heredan juntos en todos los genes estudiados. También se ha observado una tendencia al acumulamiento de SNPs en desequilibrio Hardy-Weinberg en ciertas regiones de ciertos genes, como por ejemplo algunas pequeñas agrupaciones en HLA-A, entre 5 y 6 regiones de varios SNPs con distintos valores de desequilibrio en la segunda mitad de HLA-B, 2 regiones distintas de varios SNPs en el extremo 3´de HLA-C, una región en DPA1, una región en DPB1, y lo que resulta más llamativo, una parte muy importante del gen DQA1 y de DQB1. Del mismo modo la presencia de diferencias significativas entre los genotipos

observados y esperados (test de Chi2 en la sección "Diferencias en los procesos de selección" del apartado de Resultados) sugiere que sobre dichos marcadores ha incurrido algún tipo de fuerza evolutiva, que bien pudiera deberse a procesos selectivos, aunque no es posible afirmarlo con total seguridad.

En un estudio para la población Han del sur de China (Trachtenberg et al. 2007), se encontró que la distribución de frecuencias alélicas para los genes HLA clase I, expresada en términos del estadístico F de Watterson para la homocigosidad, se encontraban dentro de lo esperado bajo la asunción de neutralidad. En contraste, en los genes HLA clase II DRB1, DQA1 y DQB1, las distribuciones de frecuencia de alelos son más uniformes de lo esperado bajo neutralidad, lo que sugiere la presencia de una selección diversificadora en estos loci.

Varianza de Wahlund

Tradicionalmente, se ha considerado que la variación entre los grandes grupos continentales es pequeña en nuestra especie. Supondría poco más del 10% de la variación total del patrimonio genético humano (Relethford 2002), y la variación genética total es más bien pequeña en nuestra especie (Barbujani, Ghirotto & Tassi 2013), por debajo de la de chimpancés o gorilas, por mencionar a nuestros parientes más próximos. En humanos, el nivel de diferenciación intraespecífico es menos de un tercio de los observados en chimpancés (Fst = 0,32) o gorilas (Fst = 0,38).

Así, las estimaciones de la varianza de Wahlund calculadas a partir de diferentes tipos de marcadores genéticos para nuestra especie han dado siempre valores bajos, de alrededor de 0,12 en promedio (Barbujani & Colonna 2010). Estos valores se han visto confirmados en trabajos recientes sobre análisis genéticos a gran escala del genoma humano. En el trabajo de Akey et al. (2002) el valor promedio es también de 0,12. A partir del análisis de la distribución de los valores de FST entre marcadores, que sigue una función exponencial negativa, se ha propuesto que los SNPs más interesantes para la detección de presiones selectivas serían aquellos que mostrasen valores extremos (Elhaik 2012).

Nuestros resultados para los diferentes genes HLA arrojan valores promedio de entre 0,015 (HLA-DMA) y 0,062 (HLA-DPB1), claramente muy por debajo de los valores promedio. Los valores medianos son similares a estos. Estas cifras claramente apuntan a una tendencia homogeneizadora, lo que podría entenderse como una fuerte presión selectiva generalizada. Lo cual es fácil de entender, por cuanto que estos genes son fundamentales para el sistema inmune. Por otra parte, también cabe una presión selectiva diversificadora, al menos en puntos concretos de las cadenas, lo que podría

facilitar alcanzar una gran diversidad en las configuraciones de polipéptidos generados por estos genes. Ello explicaría los elevados valores máximos encontrados, que en el caso de HLA-DPA1 y HLA-DPB1, son de 0,371. Estos, junto con HLA-DRB1 precisamente son los genes aparentemente sometidos a una mayor presión selectiva según diferentes autores (Trachtenberg et al. 2007). El valor máximo de Fst en HLA-DRA (0,850) podría ser accidental, fruto de algún proceso de deriva, o bien reflejar alguna presión selectiva puntual. En el caso de los marcadores con valores tan altos, se observa que no es algo puntual de una población que desvirtúe el valor de la media, si no que todas la poblaciones en general presentan valores altos. Por tanto podemos señalar que se aprecia una presión selectiva diferencial entre los distintos genes, si bien la varianza de Wahlund también se ve afectada por la deriva y el flujo génico.

Marcadores de ancestralidad

No son raros los trabajos en los que se han detectado marcadores de ancestralidad entre los SNPs de la región HLA (Nakaoka et al. 2013) y también abundan los trabajos que versan sobre los marcadores de ancestralidad asociados a diferentes enfermedades genéticas de tipo autoinmune (Herráez et al. 2013). En general, los genes estudiados en la presente tesis doctoral presentan un porcentaje muy bajo de marcadores de ancestralidad entre el conjunto de SNPs de cada uno. Sin embargo, HLA-DPA1 y DPB1 destacan precisamente por el hecho de que una gran parte de sus SNPs (21,7% y 42,7%, respectivamente) son marcadores informativos de ancestralidad, y además son los únicos genes que presentan AIMs asociables a patrones de distribución continental africano, europeo y asiático. El hecho de que estos genes sean precisamente los que presentan valores medios de la varianza de Wahlund más altos, nos sugiere la existencia de una posible relación entre ambos fenómenos. Dado que los propios AIMs nos sirven para trazar la historia evolutiva de las poblaciones, un gen con gran cantidad de AIMs resultará muy informativo para conocer el proceso evolutivo de dicha población y sus relaciones con las demás.

Análisis de una muestra de SNPs en Gitanos del País Vasco

En primer lugar, cabe señalar que hubo dificultades para el análisis de los SNPs. El método elegido comprendía una prueba previa *in silico* de los SNPs y gran parte de ellos no eran analizables, por estar sometidos a la zona de influencia de indels, estar caracterizados como bialélicos cuando eran polialélicos o indels, etc. De ahí la importancia del apartado de análisis de control de datos presente en Resultados: se analizaron SNPs sobre los que no cabía duda de su caracterización como marcadores

bialélicos no afectados por variaciones estructurales. Esto da una idea de la complejidad de esta región.

Se ha observado desequilibrio Hardy-Weinberg en rs9277332. Casualmente se encuentra junto al extremo 5' de HLA-DPB1 y el 3' de HLA-DPA1. El otro SNP en desequilibrio, rs200789833, se encuentra en la misma región.

Desde un punto de vista de genética de poblaciones, la población de los gitanos parece haber experimentado una deriva genética intensa, ya que se diferencia claramente de las poblaciones europeas y del sur de Asia (Gresham et al. 2001, Martínez-Cruz et al. 2016). A ello parece haber contribuido un intenso aislamiento, seguramente asociado a un comportamiento endogámico (Martínez-Frías & Bermejo 1992).

Relación entre genes HLA de Neandertal y humanos modernos

En cuanto al estudio de las posibles introgresiones de genes HLA de Neandertal ha arrojado resultados dispares en los diferentes alelos analizados, lo que implica necesariamente un entrecruzamiento diferencial con distintas poblaciones humanas en un amplio espacio geográfico y temporal. Se han encontrado en todos los genes alelos íntimamente ligados al alelo Neandertal. Si bien no en todos los casos las asociaciones presentan valores bootstrap de 100, y en algunos casos son francamente bajas (19 en el caso de HLA-DPB1), estos alelos se distribuyen, entre otras regiones, siempre por Europa y Próximo/Medio Oriente y regiones asociadas, lo que parece ser un indicio claro de su origen. Aunque, obviamente, a partir de los datos utilizados no pueden obtenerse unos resultados taxativos, por lo que posteriores estudios deberán investigar los datos usados a mayor resolución a fin de establecer si los alelos introgresados son los asignados, y conocer su distribución actual en la población humana.

CONCLUSIONES

CONCLUSIONES

1. Los resultados obtenidos sobre la existencia de grandes indels polimórficos afectando a genes completos (HLA-C y HLA-DQA1) con frecuencias diferentes en las poblaciones estudiadas, en conjunción con trabajos anteriores de otros autores, sugieren que la existencia de una enorme variabilidad en la expresión génica y el fenotipo, desde un funcionamiento normal hasta la completa falta de producción proteica por la ausencia de ambas copias del gen en el caso de homocigotos para la deleción, pasando por todo el arco de expresión diferencial, debida a la variación estructural producida por las deleciones parciales. Si bien se sospecha que esto afectaría a factores de reconocimiento autoinmune y de alorreconocimiento, lo cual es de gran importancia para la respuesta inmune, no es posible afirmarlo con rotundidad, por lo que futuros estudios deberán ahondar en los efectos de estas grandes deleciones que afectan a HLA-C y a HLA-DQA1.

2. El número de marcadores en desequilibrio Hardy-Weinberg distribuidos a lo largo de los 11 genes y 26 poblaciones estudiadas permiten confirmar que los genes clase I y II de la región del Complejo Mayor de Histocompatibilidad son regiones hipervariables, posiblemente sometidas a selección diversificadora, lo cual es de gran importancia en la respuesta inmune.

3. Los valores de varianza de Wahlund posibilitan realizar un doble análisis sobre los genes HLA estudiados. Por un lado, la menor varianza media general observada en el conjunto de genes al completo, parece informar de la existencia de presiones selectivas conservadoras, lo que está en concordancia con el hecho de que estos genes son fundamentales para el sistema inmune. Sin embargo los valores puntuales observados en regiones concretas apuntan a la existencia de presiones selectivas diversificadoras que añaden variabilidad a la región, y por tanto, a la evolución del sistema inmunitario de nuestra especie.

4. Los resultados obtenidos sugieren que, en general, los genes HLA estudiados no son buenos candidatos a regiones génicas informativas de ancestría con el fin de conocer los procesos colonizadores de nuestra especie, dado el bajo número de AIMs asociables a patrones de distribución continental. Sin embargo, en el caso de HLA-DPA1 y HLA-DPB1, se ha descubierto que son tremendamente informativos en este aspecto dado que presentan gran número de AIMs de varios grupos continentales distintos.

5. En lo referente a la población de gitanos del País Vasco, se caracteriza a este grupo como una población notablemente aislada y relativamente independiente del resto de poblaciones a las que está asociada, tanto geográficamente como históricamente. También se observa que han sido fruto de una fuerte deriva génica, en parte debida a la casi ausencia de flujo génico del exterior de la población.

6. Los datos usados para realizar los análisis de introgresión de ADN Neandertal en humanos modernos no permiten caracterizar al completo y en profundidad la presencia de introgresiones en todos los genes estudiados. Sin embargo, la distribución geográfica coincidente observada en todos los genes para las relaciones estudiadas de las secuencias estudiadas de ADN Neandertal y humano modernos hacen sospechar que se dio un entrecruzamiento entre la población Neandertal y humana moderna en la zona del este de Europa, oeste de Asia y Oriente Próximo, y que, al menos bajo la definición biológica, neandertales y humanos modernos podríamos ser la misma especie.

BIBLIOGRAFÍA

BIBLIOGRAFÍA

Abdellaoui A., Verweij K.J.H. & Zietsch B.P. 2014. *No evidence for genetic assortative mating beyond that due to population stratification*. Proceedings of the National Academy of Sciences 111 (40): E4137-E4137.

Abi-Rached L., Jobin M. J., Kulkarni S. et al. 2011. *The Shaping of Modern Human Immune Systems by Multiregional Admixture with Archaic Humans*. Science. 334 (6052): 89-94.

Ackermann R.R. & Cheverud J.M. 2004. *Detecting genetic drift versus selection in human evolution*. Proceedings of the National Academy of Sciences of the USA 101: 17946-17951.

Akey J.M. 2009. *Constructing genomic maps of positive selection in humans: where do we go from here?.* Genome Res. 19:711.

Akey J.M., Zhang G., Zhang K. et al. 2002. *Interrogating a high-density SNP map for signatures of natural selection*. Genome research 12(12): 1805-1814.

Akkuratov E.E., Gelfand M.S., & Khrameeva E.E. 2018. *Neanderthal and Denisovan ancestry in Papuans: A functional study*. Journal of Bioinformatics and Computational Biology. 16 (2): 1840011.

Altman R.B. 2004. *Building successful biological databases*. Brief. Bioinformatics. 5 (1): 4-5.

Altshuler D. et al. 2005. *A haplotype map of the human genome*. Nature 437, 1299-1320.

Ambrose S.H. 1998. *Late Pleistocene human population bottlenecks, volcanic winter, and differentiation of modern humans*. Journal of Human Evolution. 34 (6): 623-651.

Andersson M.B. 1994. *Sexual selection*. Princeton University Press.

Apanius V. et al. 1997. *The nature of selection on the major histocompatibility complex*. Crit Rev Immunol. 17: 179-224.

Appenzeller T. 2012. *Human migrations: Eastern odyssey. Humans had spread across Asia by 50,000 years ago. Everything else about our original exodus from Africa is up for debate*. Nature 485 (7396).

Armitage S.J. et al. 2011. *The southern route "out of Africa": evidence for an early expansion of modern humans into Arabia*. Science. 331 (6016): 453-456.

Auton A., Abecasis G.R., Altshuler D.M., et al. 2015. *A global reference for human genetic variation*. Nature 526 (7571): 68-74.

Balter M. 2011. *Was North Africa the launch pad for modern human migrations?*. Science 331 (6013): 20-23.

Barbujani G. & Colonna V. 2010. *Human genome diversity: frequently asked questions*. Trends in Genetics 26(7): 285-295.

Barbujani G., Ghirotto S. & Tassi F. 2013. *Nine things to remember about human genome diversity*. Tissue Antigens 82(3): 155-164.

Barnes C. et al. 2008. *A robust statistical method for case-control association testing with copy number variation*. Nature Genet. 40: 1245-1252.

Barrett J.C. & Cardon L. R. 2006. *Evaluating coverage of genome-wide association studies*. Nature Genet. 38: 659-662.

Begon M., Harper J. L. & Townsend C.R. 1996. *Ecology: Individuals, populations and communities* Blackwell Science.

Berger S.L. et al. 2009. *An operational definition of epigenetics*. Genes & Development 23 (7): 781-783.

Bernstein H. & Bernstein C. 2010. *Evolutionary origin of recombination during meiosis*. BioScience 60 (7): 498-505.

Bernstein H., Bernstein C. & Michod R.E. 2011. *Meiosis as an Evolutionary Adaptation for DNA Repair*. Capítulo 19 en *DNA Repair*. ed. Inna Kruman. InTechOpen.

Bernstein H., Payne C.M., Bernstein C., Garewal H. & Dvorak K. 2008. *Cancer and aging as consequences of un-repaired DNA damage*. En: *New Research on DNA Damages* (Editors: Honoka Kimura and Aoi Suzuki) Nova Science Publishers, Inc., New York, Chapter 1, pp. 1-47.

Berteau F., Rouviere B., Delluc A. et al. 2018. *Autosomic dominant familial Behçet disease and haploinsufficiency A20: A review of the literature*. Autoimmun Rev. 17(8): 809-815.

Beyin A. 2011. *Upper Pleistocene Human Dispersals out of Africa: A Review of the Current State of the Debate*. International Journal of Evolutionary Biology (615094): 1-17.

Bird A. 2007. *Perceptions of epigenetics*. Nature 447 (7143): 396-398.

Bourne P. 2005. *Will a biological database be different from a biological journal?*. PLoS Comput. Biol. 1 (3): 179-181.

Brahic C. 2012. *Our True Dawn*. New Scientist. Reed Business Information (2892): 34-37.

Bruder C.E.G. et al. 2008. *Phenotypically Concordant and Discordant Monozygotic Twins Display Different DNA Copy-Number-Variation Profiles*. The American Journal of Human Genetics 82 (3): 763-771.

Bullini L. 1994. *Origin and evolution of animal hybrid species*. Trends in Ecology and Evolution 9 (11): 422-426.

Cabrera V.M., Marrero J.P., Abu-Amero K.K. & Larruga J.M. 2018. *Carriers of mitochondrial DNA macrohaplogroup L3 basic lineages migrated back to Africa from Asia around 70,000 years ago*. BioRxiv: 233502.

Callaway E. 2012. *Hunter-gatherer genomes a trove of genetic diversity.* Nature News.

Campbell N.A. 1996. *Biology*. Benjamin/Cummings Series in the *Life Sciences* (4th ed.). Menlo Park, CA: Benjamin/Cummings Pub. Co.

Capittini C., Martinetti M. & Cuccia M. 2008. *MHC variation, mate choice and natural selection: the scent of evolution*. Riv Biol. 101(3): 463-480.

Cariaso M. & Lennon G. 2012. S*NPedia: a wiki supporting personal genome anotation, interpretation and analysis*. Nucleic Acid Research (40) Database issue: 1-5.

Cavalli-Sforza L.L., Menozzi P. & Piazza A. 1996. *The History and Geography of Human Genes* (Abridged paperback ed.). Princeton, N.J.: Princeton University Press.

Chaix R., Cao C., & Donnelly P. 2008. *Is mate choice in humans MHC-dependent?*. PLoS genetics, 4(9): e1000184.

Chandler V.L. 2007. *Paramutation: from maize to mice*. Cell 128 (4): 641-645.

Charlesworth B, Morgan M.T. & Charlesworth D. 1993. *The effect of deleterious mutations on neutral molecular variation*. Genetics. Genetics Society of America 134 (4): 1289-1303.

Charlesworth B. 2009. *Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation*. Nature Reviews. Genetics. Nature Publishing Group 10 (3): 195-205.

Choi C. Q. 2013. *Toba Supervolcano Not to Blame for Humanity's Near-Extinction*. Livescience.com.

Church K. & Wimber D.E. 1969. *Meiosis in the grasshopper: chiasma frequency after elevated temperature and x-rays*. Can. J. Genet. Cytol. 11 (1): 209-216.

Clark A. G. & Li J. 2007. *Conjuring SNPs to detect associations*. Nature Genet. 39: 815-816.

Clarkson C. et al. 2017. *Human occupation of northern Australia by 65,000 years ago*. Nature 547: 306-310.

Condemi S., Mounier A., Giunti P., Lari M., Caramelli D. & Longo L. 2013. *Possible interbreeding in late Italian Neanderthals? New data from the Mezzena jaw (Monti Lessini, Verona, Italy)*. PLoS One 8(3): e59781.

Conrad D.F. et al. 2006. *A high-resolution survey of deletion polymorphism in the human genome*. Nature Genet. 38: 75-81.

Cooper A. & Stringer C.B. 2013. *Did the Denisovans Cross Wallace's Line?*. Science. 342 (6156): 321-323.

Cooper G.M. et al. 2008. *Systematic assessment of copy number variant detection via genome-wide SNP genotyping*. Nature Genet. 40: 1199-1203.

Cooper G.M., Nickerson D.A. & Eichler E.E. 2007. *Mutational and selective effects on copy-number variants in the human genome*. Nature Genet. 39: S22-S29.

Cornuet J.M. & Luikart G. 1996. *Description and Power Analysis of Two Tests for Detecting Recent Population Bottlenecks from Allele Frequency Data*. Genetics. Bethesda, MD: Genetics Society of America. 144 (4): 2001-2014.

Creighton H. & McClintock B. 1931. *A Correlation of Cytological and Genetical Crossing-Over in Zea Mays*. Proc Natl Acad Sci USA. 17 (8): 492-497.

Cruciani F. et al. 2011. *A revised root for the human Y chromosomal phylogenetic tree: the origin of patrilineal diversity in Africa*. AJHG. 88 (6): 814-818.

Cuozzo C., Porcellini A., Angrisano T., Morano A., Lee B., Di Pardo A., Messina S., Iuliano R., Fusco A., Santillo M.R., Muller M.T., Chiariotti L., Gottesman M.E. & Avvedimento E.V. 2007. *DNA damage, homology-directed repair, and DNA methylation*. PLoS Genet. 3 (7): e110.

Dangel N.J., Knoll A., & Puchta H. 2014. *MHF1 plays Fanconi anaemia complementation group M protein (FANCM)-dependent and FANCM-independent roles in DNA repair and homologous recombination in plants*. Plant J. 78 (5): 822-833.

Darwin C. 1859. *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life* London: John Murray.

Delson E. et al. 2000. *Encyclopedia of human evolution and prehistory*. Taylor & Francis. 677- 680.

Der R., Epstein C.L. & Plotkin J.B. 2011. *Generalized population models and the nature of genetic drift*. Theoretical Population Biology. Elsevier 80 (2): 80-99.

Ding Q., Hu Y., Xu S., Wang J. & Jin L. 2014. *Neanderthal Introgression at Chromosome 3p21.31 was Under Positive Natural Selection in East Asians*. Molecular Biology and Evolution 31 (3): 683-695.

Dowling T.E. & Secor C.L. 1997. *The role of hybridization and introgression in the diversification of animals*. Annual Review of Ecology and Systematics 28: 593-619.

Duarte C., Maurício J., Pettitt P.B. et al. 1999. *The early Upper Paleolithic human skeleton from the Abrigo do Lagar Velho (Portugal) and modern-human emergence in Iberia*. Proceedings of the National Academy of Sciences 96 (13): 7604-7609.

Dupont C. et al. 2009. *Epigenetics: definition, mechanisms and clinical perspective.* Seminars in Reproductive Medicine 27 (5): 351-357.

Durbin R.M., Abecasis G.R., Altshuler R.M. et al. 2010. *A map of human genome variation from population-scale sequencing*. Nature 467 (7319): 1061-1073.

Durvasula A. & Sankararaman S. 2018. *Recovering signals of ghost archaic admixture in the genomes of present-day Africans*. BioRxiv.

Eberle M.A. et al. 2007. *Power to detect risk alleles using genome-wide tag SNP panels*. PLoS Genet. 3: e170.

Eichler E.E. et al. 2007. *Completing the map of human genetic variation*. Nature 447: 161-165.

Ejsmond M.J., Radwan J. & Wilson A.B. 2014. *Sexual selection and the evolutionary dynamics of the major histocompatibility complex*. Proceedings of the Royal Society of London B: Biological Sciences 281(1796): 20141662.

Elhaik E. 2012. *Empirical distributions of FST from large-scale human polymorphism data*. PloS one 7(11): e49837.

Esposito M. 1978. *Evidence that Spontaneous Mitotic Recombination Occurs at the Two-Strand Stage*. Proceedings of the National Academy of Sciences of the USA. 75 (9): 4436-4440.

Evans P.D. et al. 2006. *Evidence that the adaptive allele of the brain size gene microcephalin introgressed into Homo sapiens from an archaic Homo lineage*. Proceedings of the National Academy of Sciences of the United States of America. 103 (48): 18178-18183.

Evans P.D., Mekel-Bobrov N., Vallender E.J., Hudson R.R. & Lahn B.T. 2006. *Evidence that the adaptive allele of the brain size gene microcephalin introgressed into Homo sapiens from an archaic Homo lineage*. Proceedings of the National Academy of Sciences. 103 (48): 18178-18183.

Fernandes et. al. 2006. *Absence of post-Miocene Red Sea land bridges: biogeographic implications*. Journal of Biogeography 33 (6): 961-966.

Finlayson C. 2009. *The humans who went extinct: why Neanderthals died out and we survived*. Oxford University Press US. p. 68.

Fisher R.A. 1922. *On the dominance ratio*. Proceedings of the Royal Society of Edinburgh 42 (1922): 321-341.

Flicek P., Aken B.L., Ballester B., et al. 2010. *Ensembl's 10th year*. Nucleic Acids Res. 38 (Database issue): 557-562.

Flicek P., Amode M.R., Barrell D., et al. 2010. *Ensembl 2011*. Nucleic Acids Res. 39 (Database issue): 800-806.

Frazer K.A. et al. 2007. *A second generation human haplotype map of over 3.1 million SNPs*. Nature 449: 851-861.

Freitas A.A. & de Magalhães J.P. 2011. *A review and appraisal of the DNA damage theory of ageing*. Mutat. Res. 728 (1-2): 12-22.

Friedl A.A., Mazurek B. & Seiler D.M. 2012. *Radiation-induced alterations in histone modification patterns and their potential impact on short-term radiation effects*. Front Oncol. 2: 117.

Fu Q., Meyer M., Gao X. et al. 2013. *DNA analysis of an early modern human from Tianyuan Cave, China*. Proceedings of the National Academy of Sciences. 110 (6): 2223-2227.

Futuyma D. 1998. *Evolutionary Biology* (3rd ed.). Sunderland, MA: Sinauer Associates.

Futuyma D. J. 2013. *Evolution*. Third Edition. Sinauer Associates, Inc: Sunderland, MA.

Gao J.X., Park J.M., Huang S. et al. 2013. *MSH3 Mismatch Repair Protein Regulates Sensitivity to Cytotoxic Drugs and a Histone Deacetylase Inhibitor in Human Colon Carcinoma Cells*. PLoS ONE 8 (5): e65369.

Garamszegi L.Z. & Nunn C.L. 2011. P*arasite-mediated evolution of the functional part of the MHC in primates*. Journal of evolutionary biology 24(1): 184-195.

Garcia-Dorado A. 2015. *On the consequences of ignoring purging on genetic recommendations of MVP rules*. Heredity 115: 185-187.

Génin E., Ober C., Weitkamp L. & Thomson G. 2000. *A robust test for assortative mating*. Eur J Hum Genet 8(2):119-124.

Gherman A. et al. 2007. *Population bottlenecks as a potential major shaping force of human genome architecture*. PLoS Genetics 3: e119.

Gibbons A. 1993. *Pleistocene Population Explosions*. Science 262 (5130): 27-28.

Gilbert S.C. et al. 1998. *Association of malaria parasite population structure, HLA, and immunological antagonism*. Science 279: 1173-1177.

Gillespie J.H. 1991. *The causes of molecular evolution*. Oxford University Press, New York.

Gillespie J.H. 2000. *Genetic Drift in an Infinite Population: The Pseudohitchhiking Model*. Genetics. 155 (2): 909-919.

Gillespie J.H. 2001. *Is the population size of a species relevant to its evolution?*. Evolution 55 (11): 2161-2169.

Gorer P.A. 1936. *The detection of a hereditary antigenic difference in the blood of mice by means of human group A serum*. J. Genet. 32: 17-31.

Gottschalk A.J., Timinszky G., Kong S.E., Jin J., Cai Y., Swanson S.K., Washburn M.P., Florens L., Ladurner A.G., Conaway J.W. & Conaway R.C. 2009. *Poly(ADP-ribosyl)ation directs recruitment and activation of an ATP-dependent chromatin remodeler*. Proc. Natl. Acad. Sci. U.S.A. 106 (33): 13770-13774.

Gourraud P.A., Khankhanian P., Cereb N., Yang S.Y., Feolo M., Maiers M., et al. 2014. *HLA Diversity in the 1000 Genomes Dataset*. PLoS ONE 9(7): e97282.

Graffelman J., Jain D., & Weir B. 2017. *A genome-wide study of Hardy-Weinberg equilibrium with next generation sequence data*. Human genetics 136(6): 727-741.

Grant P.R., Grant B.R. & Petren K. 2005. *Hybridization in the Recent Past*. The American Naturalist. 166: 56-67.

Green R.E., Krause J., Briggs A.W. et al. 2010. *A Draft Sequence of the Neandertal Genome*. Science 328 (5979): 710-722.

Greenfield M.D., Alem S., Limousin D., Bailey N.W. 2014. *The dilemma of Fisherian sexual selection: Mate choice for indirect benefits despite rarity and overall weakness of trait-preference genetic correlation*. Evolution 68: 3524-3536.

Gregory M.D., Kippenhan J.S., Eisenberg D.P. et al. 2017. *Neanderthal-Derived Genetic Variation Shapes Modern Human Cranium and Brain*. Scientific Reports 7 (1): 6308.

Gresham D., Morar B., Underhill P.A. et al. 2001. *Origins and Divergence of the Roma (Gypsies)*. American Journal of Human Genetics 69 (6): 1314-1331.

Guerrero-Casas F.M. & Ramírez-Hurtado J.M. 2002. *El Análisis de Escalamiento Multidimensional: Una alternativa y un complemento a otras técnicas multivariantes*. *X Jornadas ASEPUMA* (Asociación Española de Profesores Universitarios de Matemáticas para la Economía y la Empresa). Madrid.

Guo G. et al. 2014. *Genomic Assortative Mating in Marriages in the United States*. PLoS ONE 9 (11).

Haber M., Jones A.L., Connell B.A., Asan, Arciero E., Yang H. & Tyler-Smith C. 2019. *A Rare Deep-Rooting D0 African Y-Chromosomal Haplogroup and Its Implications for the Expansion of Modern Humans Out of Africa*. Genetics 212(4), 1421-1428.

Hacquard-Bouder C., Ittah M., Breban M. 2005. *Animal models of HLA-B27 associated diseases: new outcoumes*. Joint Bone Spine 73: 132-138.

Hammer M.F., Woerner A.E., Mendez F.L. et al. 2011. *Genetic evidence for archaic admixture in Africa*. Proceedings of the National Academy of Sciences 108 (37): 15123-15128.

Hedrick P.W. & Black F.L. 1997. *HLA and mate selection: no evidence in South Amerindians*. Am J Hum Genet. 61(3): 505-511.

Hedrick P.W. & Loeschcke V. 1996. *MHC and mate selection in humans?* Trends Ecol Evol 11: 24.

Hedrick P.W. 1992. *Female choice and variation in the major histocompatibility complex*. Genetics 132 (2): 575-581.

Helgason A., Nicholson G., Stefansson K., and Donnelly P. 2003. *A reassessment of genetic diversity in Icelanders: strong evidence from multiple loci for relative homogeneity caused by genetic drift*. Annals of Human Genetics 67: 281-297.

Hermisson J. & Pennings P. S. 2005. *Soft sweeps: molecular population genetics of adaptation from standing genetic variation*. Genetics 169: 2335-2352.

Herráez D.L., Martínez-Bueno M., Riba L. et al. 2013. *Rheumatoid arthritis in Latin Americans enriched for Amerindian ancestry is associated with loci in chromosomes 1, 12, and 13, and the HLA class II region*. Arthritis & Rheumatism 65(6): 1457-1467.

Hershkovitz I., Marder O., Ayalon A. et al. 2015. *Levantine cranium from Manot Cave (Israel) foreshadows the first European modern humans*. Nature 520 (7546): 216-219.

Hey J. 2005. *On the number of New World founders: a population genetic portrait of the peopling of the Americas*. PLoS Biology 3: e193.

Hey J., Fitch W. M., Ayala F. J. 2005. *Systematics and the Origin of Species: On Ernst Mayr's 100th Anniversary*. Washington, D.C.: National Academies Press.

Hinds D.A. et al. 2005. *Whole-genome patterns of common DNA variation in three human populations*. Science 307: 1072-1079.

Hoeijmakers J.H. 2009. *DNA damage, aging, and cancer*. N Engl J Med. 361 (15): 1475-1485.

Hoffecker J. 2009. *The spread of modern humans in Europe*. PNAS 106 (38): 16040-16045.

Holliday T.W. 2003. *Species concepts, reticulations, and human evolution*. Current Anthropology 44 (5): 653-673.

Hong E.L., Sloan C.A., Chan E.T., et al. 2016. *Principles of metadata organization at the ENCODE data coordination center (2016 update)*. Database.

Hori T., Ohnishi H., Kadowaki T. et al. 2019. A*utosomal dominant Hashimoto's thyroiditis with a mutation in TNFAIP3*. Clinical pediatric endocrinology: case reports and clinical investigations: official journal of the Japanese Society for Pediatric Endocrinology 28(3): 91-96.

Hosken D.J. & House C.M. 2011. *Sexual Selection*. Current Biology 21: 62-65.

Howard D.J. & Berlocher S.H. 1998. *Endless Forms: Species and Speciation.* New York: Oxford University Press.

Hublin J.J. et al. 2017. *New fossils from Jebel Irhoud, Morocco and the pan-African origin of Homo sapiens*. Nature 546: 289-292.

Huerta-Sánchez E., Jin X., Asan B. et al. 2014. *Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA*. Nature 512 (7513): 194-197.

Hunter P. 2008. *What genes remember*. Prospect Magazine. Web.archive.org.

Hussels I.E. & Morton N.E. 1972. *Pingelap and Mokil Atolls: achromatopsia*. American Journal of Human Genetics 24: 304-309.

Ilnytskyy Y. & Kovalchuk O. 2011. *Non-targeted radiation effects-an epigenetic connection*. Mutat. Res. 714 (1-2): 113-125.

Janeway C.A. Jr., Travers P., Walport M., et al. 2001. *The Major Histocompatibility Complex and Its Functions* in *Immunobiology: The Immune System in Health and Disease*. 5th edition. New York: Garland Science.

Jin K., Speed T.P. & Thomson G. 1995. *Test of random mating for a highly polymorphic locus: application to HLA data*. Biometrics 51:1064-1076.

Jurmain R., Kilgore L. & Trevathan W. 2008. *Essentials of Physical Anthropology*. Cengage Learning. 266-268.

Kamiya T. et al. 2014. *A quantitative review of MHC-based mating preference: the role of diversity and dissimilarity*. Molecular ecology 23(21): 5151-5163.

Karlsson E.K., Kwiatkowski D.P. & Sabeti, P.C. 2014. *Natural selection and infectious disease in human populations*. Nature Reviews Genetics 15(6): 379-393.

Kataria R.K. & Brent L.H. 2004. *Spondyloarthropathies*. American Family Physician 69(12): 2853-2860.

Kaur G., Gras S., Mobbs J.I. et al. 2017. *Structural and regulatory diversity shape HLA-C protein expression levels*. Nature communications 8: 15924.

Keeney S., Giroux C.N. & Kleckner N. 1997. *Meiosis-Specific DNA Double-Strand Breaks Are Catalyzed by Spo11, a Member of a Widely Conserved Protein Family*. Cell 88 (3): 375-384.

Khaja, R. et al. 2006. *Genome assembly comparison identifies structural variants in the human genome*. Nature Genet. 38: 1413-1418.

Kidd J.M. et al. 2008. *Mapping and sequencing of structural variation from eight human genomes*. Nature 453: 56-64.

Kim B.Y. & Lohmueller K.E. 2015. S*election and Reduced Population Size Cannot Explain Higher Amounts of Neandertal Ancestry in East Asian than in European Human Populations*. The American Journal of Human Genetics 96 (3): 454-461.

Kimura M. 1968. *Evolutionary Rate at the Molecular Level*. Nature 217:624-626.

King J.L. & Jukes T.H. 1969. *Non-Darwinian Evolution*. Science 164: 788-797.

Kocsor F. et al. 2011. *Preference for Facial Self-Resemblance and Attractiveness in Human Mate Choice* Arch Sex Behav 40: 1263.

Köhler K., Ferreira P., Pfander B. & Boos D. 2016. *The Initiation of DNA Replication in Eukaryotes*. Springer, Cham. pp. 443-460.

Korbel J.O. et al. 2007. *Paired-end mapping reveals extensive structural variation in the human genome*. Science 318: 420-426.

Korn J.M. et al. 2008. *Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs*. Nature Genet. 40: 1253-1260.

Kovalchuk O. & Baulch J.E. 2008. *Epigenetic changes and nontargeted radiation effects—is there a link?*. Environ. Mol. Mutagen. 49(1): 16-25.

Krings M., Stone A., Schmitz R.W. el at. 1997. *Neandertal DNA Sequences and the Origin of Modern Humans*. Cell. 90(1): 19-30.

Kruglyak L. & Nickerson D.A. 2001. *Variation is the spice of life*. Nature Genet. 27: 234-236.

Kruskal J.B. 1964. *Nonmetric multidimensional scaling: a numerical method*. Psychometrika 29: 115-129.

Kuhlwilm M., Gronau I., Hubisz M.J. et al. 2016. *Ancient gene flow from early modern humans into Eastern Neanderthals*. Nature. 530 (7591): 429-433.

Kumar A., Bassi F. & Paux E. 2012. *DNA repair and crossing over favor similar chromosome regions as discovered in radiation hybrid of Triticum*. BMC Genomics. 13: 339.

Kunkel T.A. & Erie D.A. 2005. *Dna Mismatch Repair*. Annual Review of Biochemistry 74 (1): 681-710.

LaBar T. & Adami C. 2017. *Evolution of drift robustness in small populations*. Nature Communications 8(1): 1012.

Lachance J., Vernot B., Elbers C.C. et al. 2012. *Evolutionary History and Adaptation from High-Coverage Whole-Genome Sequences of Diverse African Hunter-Gatherers*. Cell 150 (3): 457-469.

Lande R. 1976. *Natural selection and random genetic drift in phenotypic evolution*. Evolution 30: 314-334.

Lari M., Rizzi E., Milani L. et al. 2010. *The Microcephalin Ancestral Allele in a Neanderthal Individual*. PLoS ONE 5 (5): e10648.

Ledford H. 2008. *Disputed definitions*. Nature. 455 (7216): 1023-1028.

Levy S. et al. 2007. *The Diploid Genome Sequence of an Individual Human* PLoS Biol 5(10): e254.

Lin J.C., Jeong S., Liang G., Takai D., Fatemi M., Tsai Y.C., Egger G., Gal-Yam E.N. & Jones P.A. 2007. *Role of nucleosomal occupancy in the epigenetic silencing of the MLH1 CpG island*. Cancer Cell 12 (5): 432-444.

Little C.C. 1941. *The genetics of tumor transplantation* in *Biology of the Laboratory Mouse*, pp 279-309, ed by Snell GD, New York: Dover.

Liu H., Prugnolle F., Manica A., Balloux F. 2006. *A geographically explicit genetic model of worldwide human-settlement history*. AJHG. 79(2): 230-237.

Liu L. et al. 2012. *Global, regional, and national causes of child mortality: an updated systematic analysis for 2010 with time trends since 2000*. The Lancet 379(9832): 2151-2161.

Liu W. et al. 2015. *The earliest unequivocally modern humans in southern China*. Nature 526: 696-699.

Lohse K. & Frantz L.A.F. 2014. *Neandertal Admixture in Eurasia Confirmed by Maximum-Likelihood Analysis of Three Genomes*. Genetics 196(4): 1241-1251.

Macaulay V. et al. 2005. *Single, Rapid Coastal Settlement of Asia Revealed by Analysis of Complete Mitochondrial Genomes*. Science 308 (5724): 1034-1036.

Malanga M. & Althaus F.R. 2005. *The role of poly(ADP-ribose) in the DNA damage signaling network*. Biochem Cell Biol. 83 (3): 354-364.

Marcos Sarobe I. 2016. I*nserciones Alu y heterogeneidad genética de la población gitana del País Vasco*. Online: https://addi.ehu.es/handle/10810/18055

Martínez-Cruz B., Mendizabal I., Harmant C. et al. 2016. *Origins, admixture and founder lineages in European Roma*. European journal of human genetics: EJHG 24(6): 937-943.

Martinez-Frías M.L. & Bermejo E. 1992. *Prevalence of congenital anomaly syndromes in a Spanish gypsy population*. J. Med. Genet. 29: 483-486.

Masel J. 2011. *Genetic drift*. Current Biology. Cambridge, MA: Cell Press. 21 (20): 837-838.

Mason P.H. & Short R.V. 2011. *Neanderthal-human Hybrids*. Hypothesis 9 (1): e1.

Mattei J. F. 2002. *El genoma humano (Ethical eye: the human genome)*. Sáez García M.A., Chao Crecente M., Vázquez D.A. & Rodríguez-Roda Stuart J. trad. Colección La Mirada de la Ciencia. Madrid: Council of Europe/Editorial Complutense. Glosario p. 201.

McCarroll S.A. & Altshuler D.M. 2007. *Copy-number variation and association studies of human disease*. Nature Genet. 39: S37-S42.

McClintock B. 1984. *The significance of response of the genome to challenge*. Science 226: 792-801.

McCoy R.C., Wakefield J. & Akey J.M. 2017. *Impacts of Neanderthal-Introgressed Sequences on the Landscape of Human Gene Expression*. Cell 168 (5): 916-927.

McVean G.A., Abecasis D.M., Auton A. et al. 2012. *An integrated map of genetic variation from 1,092 human genomes*. Nature 491 (7422): 56-65.

Mendel J.G. 1866. *Versuche über Pflanzenhybriden*, Verhandlungen des naturforschenden Vereines in Brünn, Bd. Abhandlungen: 3-47. A través de la traducción en inglés: Druery C.T. & Bateson W. 1901. *Experiments in plant hybridization.* Journal of the Royal Horticultural Society. 26: 1-32.

Metspalu M. et al. 2004. *Most of the extant mtDNA boundaries in south and southwest Asia were likely shaped during the initial settlement of Eurasia by anatomically modern humans*. BMC Genet. 5: 26.

Metzenberg A.B. et al. 1991. *Homology Requirements for Unequal Crossing over in Humans*. Genetics 128(1): 143-161.

Meyer M., Kircher M., Gansauge M.T. et al. 2012. *A High-Coverage Genome Sequence from an Archaic Denisovan Individual*. Science 338 (6104): 222-226.

Mirazón Lahr M. et al. 2012. S*earching for traces of the Southern Dispersal*. Wayback Machine.

Moran P.A.P. 1958. *Random processes in genetics*. Mathematical Proceedings of the Cambridge Philosophical Society 54 (1): 60-71.

Nakaoka H., Mitsunaga S., Hosomichi K. et al. 2013. *Detection of ancestry informative HLA alleles confirms the admixed origins of Japanese population*. PLoS One 8(4): e60793.

NCBI Resource Coordinators. 2012. *Database resources of the National Center for Biotechnology Information*. Nucleic Acids Research. 41 (Database issue): D8-D20.

Nedelcu M., Marcu O. & Michod R.E. 2004. *Sex as a response to oxidative stress: a twofold increase in cellular reactive oxygen species activates sex genes*. Proc. R. Soc. B. 271: 1591-1596.

Neher R.A. & Shraiman B.I. 2011. *Genetic draft and quasi-neutrality in large facultatively sexual populations*. Genetics. Genetics Society of America. 188 (4): 975-996.

Nei M. & Roychoudhury A. K. 1974. *Genic variation within and between the three major races of man, Caucasoids, Negroids, and Mongoloids*. The American Journal of Human Genetics. 26: 421-443.

Nei M. 1987. *Molecular Evolutionary Genetics* (Chapter 9). New York: Columbia University Press.

Neves A. & Serva M. 2012. E*xtremely Rare Interbreeding Events Can Explain Neanderthal DNA in Living Humans*. PLoS ONE. 7 (10): e47076.

Nielsen D.M., Ehm M.G. & Weir B.S. 1998. *Detecting marker-disease association by testing for Hardy-Weinberg disequilibrium at a marker locus*. The American Journal of Human Genetics 63(5): 1531-1540.

Nielsen R. 2005. *Molecular signatures of natural selection*. Annu. Rev. Genet. 39: 197-218.

Nielsen R., Akey J.M., Jakobsson M. et al. 2017. *Tracing the peopling of the world through genomics*. Nature. 541 (7637): 302-10.

Nordborg M., Hu T.T., Ishino Y. et al. 2005. *The pattern of polymorphism in Arabidopsis thaliana*. PLoS Biology. Public Library of Science 3 (7): e196.

Ober C. et al. 1997. *HLA and mate choice in humans*. Am J Hum Genet 61: 497-504.

O'Corry-Crowe G. 2008. *Climate change and the molecular ecology of arctic marine mammals*. Ecological Applications. Washington, D.C.: Ecological Society of America. 18 (2, Supplement: Arctic Marine Mammals): 56-76.

O'Hagan H.M., Mohammad H.P. & Baylin S.B. 2008. *Double strand breaks can initiate gene silencing and SIRT1-dependent onset of DNA methylation in an exogenous promoter CpG island*. PLoS Genet. 4 (8): e1000155.

Ohashi J., Naka I., Patarapotikul J. et al. 2004. *Extended linkage disequilibrium surrounding the hemoglobin E variant due to malarial selection*. Am. J. Hum. Genet. 74: 1198-1208.

Oppenheimer C. 2002. *Limited global change due to largest known Quaternary eruption, Toba ≈74 kyr BP?.* Quaternary Science Reviews 21: 1593-1609.

Orr H.A. 1998. *Testing natural selection vs. genetic drift in phenotypic evolution using quantitative trait locus data*. Genetics 196: 2099-2104.

Orr H.A. 2009. *Fitness and its role in evolutionary genetics*. Nat Rev Genet. 10 (8): 531-539.

Ostrer H. 2001. *A genetic profile of contemporary Jewish populations*. Nature Reviews. Genetics 2: 891-898.

Pe'er I. et al. 2006. *Evaluating and improving power in whole-genome association studies using fixed marker sets*. Nature Genet. 38: 663-667.

Pearson et al. 1903. *Assortative Mating in Man: A Cooperative Study*. Biometrika. 2 (4): 481-498.

Piertney S.B. & Oliver M.K. 2006. *The evolutionary ecology of the major histocompatibility complex*. Heredity 96(1): 7-21.

Pope K.O. & Terrell J. E. 2007. *Environmental setting of human migrations in the circum-Pacific region.* Journal of Biogeography 35 (1): 1-21.

Posth C. et al. 2016. *Pleistocene Mitochondrial Genomes Suggest a Single Major Dispersal of Non-Africans and a Late Glacial Population Turnover in Europ*e. Current Biology 26: 827-833.

Posth C. et al. 2017. *Deeply divergent archaic mitochondrial genome provides lower time boundary for African gene flow into Neanderthals*. Nature Communications 8: 16046.

Potts W.K. & Wakeland E.K. 1993. *Evolution of MHC genetic diversity: a tale of incest, pestilence and sexual preference*. Trends Genet. 9: 408-412.

Presgraves D.C. 2005. *Recombination enhances protein adaptation in Drosophila melanogaster*. Current Biology. Cell Press 15 (18): 1651-1656.

Prüfer K., de Filippo C., Grote S. et al. 2017. *A high-coverage Neandertal genome from Vindija Cave in Croatia*. Science 358 (6363): 655-58.

Prüfer K., Racimo F., Patterson N. et al. 2014. *The complete genome sequence of a Neanderthal from the Altai Mountains*. Nature 505 (7481): 43-49.

Przeworski M. et al. 2005. *The signature of positive selection on standing genetic variation*. Evolution 59: 2312-2323.

Puchta H. 2005. *The repair of double-strand breaks in plants: mechanisms and consequences for genome evolution*. Journal of Experimental Botany 56 (409): 1-14.

Racowsky C. 2002. *High rates of embryonic loss, yet high incidence of multiple births in human ART: is this paradoxical?.* Theriogenology 57: 87-96.

Rampino M.R. & Self S. 1993. *Bottleneck in the Human Evolution and the Toba Eruption*. Science 262 (5142): 1955.

Raney B.J. et al. 2010. *ENCODE whole-genome data in the UCSC genome browser (2011 update)*. Nucleic Acids Res. Nucleic Acids Research. 39: D871-D875.

Rasmussen M., Guo X., Wang Y. et al. 2011. *An Aboriginal Australian Genome Reveals Separate Human Dispersals into Asia*. Science 334 (6052): 94-98.

Redon R. et al. 2006. *Global variation in copy number inthe human genome*. Nature 444: 444-454.

Reich D., Green R.E., Kircher M. et al. 2010. *Genetic history of an archaic hominin group from Denisova Cave in Siberia*. Nature 468 (7327): 1053-1060.

Reich D., Patterson N., Kircher M. et al. 2011. *Denisova Admixture and the First Modern Human Dispersals into Southeast Asia and Oceania*. The American Journal of Human Genetics 89 (4): 516-528.

Reik W. 2007. *Stability and flexibility of epigenetic gene regulation in mammalian development*. Nature 447 (7143): 425-432.

Relethford J.H. 2002. *Apportionment of global human genetic diversity based on craniometrics and skin color*. American Journal of Physical Anthropology: The Official Publication of the American Association of Physical Anthropologists 118(4): 393-398.

Riemersma S.A., Jordanova E.S., Schop R.F. et al. 2000. *Extensive genetic alterations of the HLA region, including homozygous deletions of HLA class II genes in B-cell lymphomas arising in immune-privileged sites*. Blood 96 (10): 3569-3577.

Rigden D.J. & Fernández X.M. *The 2018 Nucleic Acids Research database issue and the online molecular biology database collection*. Nucleic Acids Research 46 (D1): D1-D7.

Riggs A.D., Russo V.E. & Martienssen R.A. 1996. *Epigenetic mechanisms of gene regulation*. Plainview, N.Y.: Cold Spring Harbor Laboratory Press.

Rito T., Vieira D., Silva M., Conde-Sousa E., Pereira L., Mellars P. & Soares P. 2019. *A dispersal of Homo sapiens from southern to eastern Africa immediately preceded the out-of-Africa migration*. Scientific reports 9(1): 4728.

Robberecht C. et al. 2012. *Non-allelic homologous recombination between retrotransposable elements is a driver of de novo unbalanced translocations*. Genome Research. 23: gr.145631.112.

Robinson R. 2003. *Population Bottleneck*. Genetics. 3. New York: Macmillan Reference USA.

Ruffier M., Kähäri A., Komorowska M., et al. 2017. *Ensembl core software resources: storage and programmatic access for DNA sequence and genome annotation*. Database. 2017 (1).

Sahney S., Benton M.J. & Ferry P.A. 2010. *Links between global taxonomic diversity, ecological diversity and the expansion of vertebrates on land*. Biology letters 6(4): 544-547.

Sánchez-Quinto F., Botigué L.R., Civit S. et al. 2012. *North African Populations Carry the Signature of Admixture with Neandertals*. PLoS ONE 7 (10): e47765.

Sankararaman S., Mallick S., Dannemann M. et al. 2014. *The genomic landscape of Neanderthal ancestry in present-day humans*. Nature 507 (7492): 354-357.

Sankararaman S., Patterson N., Li H., Pääbo S., Reich D. & Akey J.M. 2012. *The Date of Interbreeding between Neandertals and Modern Humans*. PLoS Genetics 8 (10): e1002947.

Saponaro M., Callahan D., Zheng X. & Liberi G. 2010. *Cdk1 Targets Srs2 to Complete Synthesis-Dependent Strand Annealing and to Promote Recombinational Repair*. PLoS Genet. 6 (2): e1000858.

Sauvageau S., Stasiak A.Z., Banville I. et al. 2005. *Fission Yeast Rad51 and Dmc1, Two Efficient DNA Recombinases Forming Helical Nucleoprotein Filaments*. Molecular and Cellular Biology 25 (11): 4377-4387.

Scerri E.M.L., Drake N.A., Jennings R., Groucutt H.S. 2014. *Earliest evidence for the structure of Homo sapiens populations in Africa*. Quaternary Science Reviews 101: 207-216.

Sebat J. 2007. *Major changes in our DNA lead to major changes in our thinking*. Nature Genet. 39: S3-S5.

Serre D., Langaney A., Chech M. et al. 2004. *No Evidence of Neandertal mtDNA Contribution to Early Modern Humans*. PLoS Biology 2 (3): 313-317.

Shackelton L.A. et al. 2006. *JC virus evolution and its association with human populations*. Journal of Virology 80: 9928-9933.

Shen G. et al. 2002. *U-Series dating of Liujiang hominid site in Guangxi, Southern China*. J. Hum. Evol. 43 (6): 817-829.

Shinohara A., Ogawa H., Matsuda Y. et al. 1993. *Cloning of human, mouse and fission yeast recombination genes homologous to RAD51 and recA*. Nature Genetics 4 (3): 239-243.

Singh P.B. 2001. *Chemosensation and genetic individuality*. Reproduction 121: 529-539.

Skoglund P. & Jakobsson M. 2011. *Archaic human ancestry in East Asia*. Proceedings of the National Academy of Sciences 108 (45): 18301-18306.

Skoglund P., Thompson J.C., Prendergast M.E. et al. 2017. *Reconstructing Prehistoric African Population Structure*. Cell 171 (1): 59-71.

Slatkin M. 2008. *Linkage disequilibrium — understanding the evolutionary past and mapping the medical future*. Nature Rev. Genet. 9: 477-485.

Smith G.P. 1976. *Evolution of Repeated DNA Sequences by Unequal Crossover*. Science 191 (4227): 528-535.

Smith J.M. & Haigh J. 1974. *The hitch-hiking effect of a favourable gene*. Genetical Research 23 (1): 23-35.

Smith T.M. 2007. *Earliest evidence of modern human life history in North African early Homo sapiens*. PNAS 104 (15): 6128-6133.

Snell G.D. & Higgins G.F. 1951. *Alleles at the histocompatibility-2 locus in the mouse as determined by tumor transplantation*. Genetics 36: 306-310.

SNPedia. 2018. *How many SNPs are in SNPedia?*. En SNPedia: FAQs General. Online.

Soficaru A., Dobos A. & Trinkaus E. 2006. *Early modern humans from the Peştera Muierii, Baia de Fier, Romania*. Proceedings of the National Academy of Sciences 103 (46): 17196-17201.

Star B. & Spencer H.G. 2013. *Effects of genetic drift and gene flow on the selective maintenance of genetic variation*. Genetics 194 (1): 235-244.

Steinboeck F. 2010. *The relevance of oxidative stress and cytotoxic DNA lesions for spontaneous mutagenesis in non-replicating yeast cells*. Mutat Res. 688 (1-2): 47-52.

Stix G. 2008. *The Migration History of Humans: DNA Study Traces Human Origins Across the Continents*. Scientific American.

Stringer C. 2003. *Human evolution: Out of Ethiopia*. Nature 423 (6941): 692-693, 695.

Surtees J.A., Argueso J.L. & Alani E. 2004. *Mismatch Repair Proteins: Key Regulators of Genetic Recombination*. Cytogenetic and Genome Research 107: 146-159.

Tabish A.M., Poels K., Hoet P. & Godderis L. 2012. *Epigenetic factors in cancer risk: effect of chemical carcinogens on global DNA methylation pattern in human TK6 cells*. PLoS ONE 7 (4): e34674.

Tattersall I. 2009. *Human origins: Out of Africa*. PNAS 106 (38): 16018-16021.

Templeton A.R. 2005. *Haplotype trees and modern human origins*. Am. J. Phys. Anthropol. 128 (41): 33-59.

Trachtenberg E., Vinson M., Hayes E. et al. 2007. *HLA class I (A, B, C) and class II (DRB1, DQA1, DQB1, DPB1) alleles and haplotypes in the Han from southern China*. Tissue antigens 70 (6): 455-463.

Trinkaus E., Moldovan O., Milota S. et al. 2003. *An early modern human from the Peştera cu Oase, Romania*. Proceedings of the National Academy of Sciences 100 (20): 11231-36.

Trinkaus E. 2007. *European early modern humans and the fate of the Neandertals*. Proceedings of the National Academy of Sciences 104 (18): 7367-7372.

Tuzun E. et al. 2005. *Fine-scale structural variation of the human genome*. Nature Genet. 37: 727-732.

Vernot B. & Akey J.M. 2014. *Resurrecting Surviving Neandertal Lineages from Modern Human Genomes*. Science 343 (6174): 1017-1021.

Vernot B. & Akey J.M. 2015. *Complex History of Admixture between Modern Humans and Neandertals*. The American Journal of Human Genetics 96 (3): 448-453.

Wahl L.M. 2011. *Fixation when N and s vary: classic approaches give elegant new results*. Genetics. Genetics Society of America 188 (4): 783-785.

Wall J.D. & Hammer M.F. 2006. *Archaic admixture in the human genome*. Current Opinion in Genetics & Development 16 (6): 606-10.

Wall J.D., Yang M.A., Jay F. et al. 2013. *Higher Levels of Neanderthal Ancestry in East Asians than in Europeans*. Genetics 194 (1): 199-209.

Walter R.C. et al. 2000. *Early human occupation of the Red Sea coast of Eritrea during the last interglacial*. Nature 405 (6782): 65-69.

Wang C.C., Farina S.E. & Li H. 2013. *Neanderthal DNA and modern human origins*. Quaternary International 295: 126-129.

Wang E.T. et al. 2006. *Global landscape of recent inferred Darwinian selection for Homo sapiens*. Proc. Natl. Acad. Sci. 103: 135-140.

Weaver T.D. et al. 2007. *Were Neandertal and modern human cranial differences produced by natural selection or genetic drift?*. Journal of Human Evolution 53: 135-145.

Wedekind C. & Furi S. 1997. *Body odor preferences in men and women: do they aim for specific MHC combinations or simply heterozygosity?* Proc R Soc Lond B 264: 1471-1479.

Wedekind C., Seebeck T., Bettens F. & Paepke A.J. 1995. *MHC-dependent mate preferences in humans*. Proc R Soc Lond B 260: 245-249.

Wells S. 2003. T*he Journey of Man: A Genetic Odyssey*. New York: Random House Trade Paperbacks.

Westerman M. 1971. *The effect of x-irradiation on chiasma frequency in Chorthippus brunneus*. Heredity 27 (1): 83-91.

Wills C. 2011. *Genetic and Phenotypic Consequences of Introgression Between Humans and Neanderthals*. Advances in Genetics 76: 27-54.

Winking J. et al. 2014. *A population-level test of human negative assortative mating along HLA class I and class II loci.* The University of New Mexico.

Winternitz J.C. et al. 2013. *Sexual selection explains more functional variation in the mammalian major histocompatibility complex than parasitism*. Proceedings of the Royal Society of London B: Biological Sciences 280 (1769).

Winternitz J.C. & Abbate J.L. 2015. *Examining the evidence for major histocompatibility complex-dependent mate selection in humans and nonhuman primates*. Research and Reports in Biology 6: 73-88.

Winternitz J.C et al. 2017. *Patterns of MHC-dependent mate selection in humans and nonhuman primates: a meta-analysis*. Molecular Ecology 26 (2): 668-688.

Wolf J.B., Brodie E.D. III & Wade M.J. 2000. *Epistasis and the Evolutionary Process*. Oxford, UK; New York: Oxford University Press.

Wolpoff M.H., Hawks J. & Caspari R. 2000. *Multiregional, not multiple origins*. Am. J. Phys. Anthropol. 112 (1): 129-136.

Xu D., Pavlidis P., Taskent O.R. et al. 2017. *Archaic Hominin Introgression in Africa Contributes to Functional Salivary MUC7 Genetic Variation*. Molecular Biology and Evolution 34 (10): 2704-2715.

Yamazaki K. et al. 1983. *Recognition of H-2 types in relation to the blocking of pregnancy in mice*. Science 221: 186-188.

Yang M.A., Gao X., Theunert C. et al. 2017. *40,000-Year-Old Individual from Asia Provides Insight into Early Population Structure in Eurasia*. Current Biology 27 (20): 3202-3208.

Yang M.A., Malaspinas A.S., Durand E.Y. & Slatkin M. 2012. *Ancient Structure in Africa Unlikely to Explain Neanderthal and Non-African Genetic Similarity*. Molecular Biology and Evolution 29 (10): 2987-2995.

Yeung J.T., Hamilton R.L., Ohnishi K. et al. 2013. *LOH in the HLA class I region at 6p21 is associated with shorter survival in newly diagnosed adult glioblastoma*. Clinical cancer research: an official journal of the American Association for Cancer Research 19(7): 1816-1826.

Yotova V. et al. 2011. *An X-linked haplotype of Neandertal origin is present among all non-African populations*. Molecular Biology and Evolution 28 (7): 1957-1962.

Zhivotovsky L.A. et al. 2003. *Features of Evolution and Expansion of Modern Humans, Inferred from Genomewide Microsatellite Markers*. American Journal of Human Genetics 72 (5): 1171-1186.