UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

POSGRADO EN CIENCIAS BIOLÓGICAS

INSTITUTO DE ECOLOGÍA

BIOLOGÍA EVOLUTIVA

INFIRIENDO EL PASADO MEDIANTE EL ANÁLISIS DE DATOS GENÓMICOS MODERNOS Y ANTIGUOS EN CONJUNCIÓN CON DATOS ESPACIALES PARA MEDIR EL IMPACTO DE LA SELECCIÓN NATURAL

TESIS

QUE PARA OPTAR POR EL GRADO DE

MAESTRO EN CIENCIAS BIOLÓGICAS

PRESENTA:

HÉCTOR ALESSANDRO LÓPEZ HERNÁNDEZ

TUTOR(A) PRINCIPAL DE LA TESIS: DR. VICENTE DIEGO ORTEGA DEL VECCHYO
INSTITUTO DE ECOLOGÍA
COMITÉ TUTOR: DRA. ANGÉLICA GONZALEZ OLIVER
FACULTAD DE CIENCIAS, UNAM.
DRA. LUCIA GUADALUPE MORALES REYES
LABORATORIO INTERNACIONAL DE INVESTIGACIÓN SOBRE EL GENOMA HUMANO

Ciudad Universitaria, CD. MX. **MAYO, 2023**

COORDINACIÓN DEL POSGRADO EN CIENCIAS BIOLÓGICAS

ENTIDAD INSTITUTO DE ECOLOGÍA

OFICIO CPCB/322/2023

ASUNTO: Oficio de Jurado

M. en C Ivonne Ramírez Wence
Directora General de Administración Escolar, UNAM

P r e s e n t e

Me permito informar a usted que, que el Comité Académico, del Posgrado en Ciencias Biológicas, en su reunión ordinaria del día **23 de enero de 2023**, aprobó el siguiente jurado para el examen de grado de **MAESTRO EN CIENCIAS BIOLÓGICAS** en el campo de conocimiento de **Biología Evolutiva** del alumno **LÓPEZ HERNÁNDEZ HÉCTOR ALESSANDRO,** con número de cuenta: **311005523** con la tesis titulada: **"INFIRIENDO EL PASADO MEDIANTE EL ANÁLISIS DE DATOS GENÓMICOS MODERNOS Y ANTIGUOS EN CONJUNCIÓN CON DATOS ESPACIALES PARA MEDIR EL IMPACTO DE LA SELECCIÓN NATURAL",** bajo la dirección del **DR. VICENTE DIEGO ORTEGA DEL VECCHYO,** Tutor Principal, quedando integrado de la siguiente manera**:**

Presidente: DR. DANIEL IGNACIO PIÑERO DALMAU
Vocal: DRA. MASHAAL SOHAIL
Vocal: DRA. MARÍA DEL CARMEN ÁVILA ARCOS
Vocal: DR. LUIS MEDRANO GONZÁLEZ
Secretario: DRA. ANGÉLICA GONZÁLEZ OLIVER

Sin otro particular, me es grato enviarle un cordial saludo.

A T E N T A M E N T E
"POR MI RAZA HABLARÁ EL ESPÍRITU"
Ciudad Universitaria, Cd. Mx., a 30 de marzo de 2023

COORDINADOR DEL PROGRAMA

DR. ADOLFO GERARDO NAVARRO SIGÜENZA

POSGRADO EN CIENCIAS BIOLOGICAS
COORDINACIÓN

Agradecimientos Institucionales

- Al Posgrado en Ciencias Biológicas de la Universidad Nacional Autónoma de México por haberme dado la oportunidad de desarrollarme profesionalmente y académicamente al abrirme las puertas a su programa.

- Al Consejo Nacional de Ciencia y Tecnología por el apoyo económico brindado durante la realización de este proyecto.

- Al *PAPIIT* por los apoyos económicos recibidos para la realización de este proyecto: IA200620 «Análisis temporal de los efectos de la selección natural en nuevos alelos» y IA206222 «Nuevos métodos para estudiar la evolución de alelos bajo selección natural mediante el uso de árboles genealógicos».

- Al Doctor Vicente Diego Ortega Del Vecchyo, tutor, por su apoyo académico y emocional, además de darme la oportunidad de integrarme en su campo de investigación.

- A la Doctora Angélica González Oliver, secretaria sinodal y miembro de mi comité tutoral, por sus comentarios y su tutoría en la realización de esta tesis además de su apoyo durante este trayecto académico.

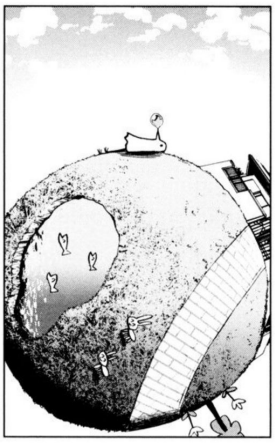
- A la Doctora Lucía Guadalupe Morales Reyes, miembro del comité tutoral, por sus comentarios, su tutoría y su invaluable apoyo para la realización de esta tesis.

Agradecimientos Personales

- Al Doctor Daniel Ignacio Piñero Dalmau, presidente del Jurado para Examen de Grado por sus comentarios para la realización de esta tesis y sus observaciones sobre métodos de detección de selección natural.
- A la doctora María C. Ávila Árcos, vocal del Jurado para Examen de Grado, por sus comentarios generales sobre este trabajo.
- Al doctor Luis Medrano González, vocal del Jurado para Examen de Grado, por sus comentarios generales sobre este trabajo y sus observaciones puntuales sobre la selección natural.
- A la doctora Mashaal Sohail, vocal del Jurado para Examen de Grado, por sus comentarios sobre la discusión de este proyecto así como su disponibilidad.
- Al Laboratorio Nacional de Visualización LAVIS, integrado por Luis Aguilar, Alejandro De León y Jair García, por el apoyo técnico para la realización de este trabajo.
- A la Universidad Nacional Autónoma de México, la Escuela Nacional Preparatoria No. 6 «Antonio Caso», la Facultad de Ciencias y al Laboratorio Internacional de Investigación sobre el Genoma Humano.
- Alexandra Elbakyan, quien quizás nunca llegue a leer estas palabras pero que gracias a su proyecto de investigación yo y muchos otros hemos tenido aún ese contacto con la ciencia; destruyendo un muro que muchos no podríamos superar.

A mi madre, Estela.

Tu voz es un refugio
en el caos de este mundo loco,
en tu abrazo siento el alivio
y la ternura que siempre busco.



Índice general

Agradecimientos Institucionales II

Agradecimientos Personales III

V

1 Resumen 1

2 Abstract 2

3 Introducción 3

3.1 Selección natural 3

3.2 Métodos de detección de selección natural 8

3.2.1 Datos genéticos como series de tiempo 12

3.2.2 Métodos markovianos 14

3.3 Datos genéticos de poblaciones humanas europeas 17

3.3.1 Datos genómicos 17

3.3.2 Poblaciones europeas 18

4 Objetivos 20

5 Antecedentes 21

5.1 Inferencia de selección natural a través de cadenas de Markov 21
5.2 Clasificación de alelos deletéreos, neutros y ventajosos 23
5.3 Análisis de la distribución espacio temporal de frecuencias alélicas 24

6 Métodos **25**
6.1 Datos genómicos 25
6.2 Análisis de los datos 25
6.3 Detección de alelos tentativamente deletéreos, neutros y ventajosos 27
6.4 Modelo de inferencia de selección y geografía 28

7 Resultados **31**
7.1 Los datos genómicos de humanos de Europa comprenden una amplia distribución geográfica y temporal. 31
7.2 La mayor parte de los alelos investigados tienden a ser neutros o ligeramente deletéreos. 31
7.3 Los alelos neutros se dispersan a una gran velocidad a lo largo de Europa. 33
7.4 La inferencia de los valores de selección natural sugieren cambios temporales en las presiones selectivas. 37

8 Discusión **48**
8.1 El coeficiente de selección probablemente no es continuo en el tiempo en alelos bajo selección positiva. 48
8.2 Los alelos con puntaje CADD mayor a 10 probablemente aparecieron en los últimos 30000 años. 50
8.3 Los alelos bajo selección tienen velocidades de dispersión diferentes a los alelos neutros. 51

9 Conclusiones **53**

10 Glosario **55**

Bibliografía **57**

1

Resumen

El análisis de genomas antiguos humanos ha permitido estimar los cambios en las frecuencias alélicas que han ocurrido en nuestra especie a lo largo del tiempo. Estos cambios permiten inferir diversos procesos evolutivos que abarcan migraciones pasadas hasta el impacto de la selección natural. Sin embargo, poco se conoce sobre cómo estos procesos moldean la distribución espacial de los alelos. En esta tesis se desarrolló una metodología estadística para estudiar la distribución espacio temporal de la frecuencia alélica para inferir la dispersión de alelos y el impacto de la selección natural a partir de datos genómicos antiguos y modernos de poblaciones de seres humanos ubicadas en distintas zonas de Europa y cuyos registros de datación abarcan desde los 30000 años en el pasado hasta el presente. Nuestra metodología agrupa a poblaciones con una distribución geográfica y temporal cercana mediante un algoritmo de *K-medias*. Posteriormente inferimos el impacto de la selección natural y de la dispersión de alelos en estas poblaciones utilizando una cadena oculta de Markov. Se clasificaron los alelos basados en su puntaje *CADD*, el cual predice qué tan deletéreo es un alelo y, basado en este puntaje, inferir el coeficiente de selección de alelos tentativamente deletéreos. Los resultados muestran que los alelos tentativamente deletéreos tienen un coeficiente de selección igual a -0.2, y también inferimos que los alelos se mueven a 10 celdas (de 91x44 km a 28x44 km por celda) por generación. Adicionalmente, los resultados muestran que tomar en cuenta el tiempo de inicio de la presión selectiva es clave para inferir correctamente el impacto de la selección natural en las poblaciones humanas estudiadas. Nuestra metodología muestra que es posible conjuntar datos genómicos espacio temporales para inferir el impacto de alelos bajo selección natural tanto para alelos ventajosos como para alelos deletéreos.

2

Abstract

Ancient human genome analyses have allowed us to estimate the change in allele frequency that have occurred in our species over time. These changes allow us to infer evolutionary processes, such as past migrations or the impact of natural selection. However, little is known about how natural selection has influenced the spatial distribution of alleles. Here, we developed a statistical method to study the spatiotemporal distribution of allele frequencies in order to infer the dispersion of alleles and the impact of natural selection from ancient and modern genomic data of human populations located in different areas of Europe, with dating records ranging from 30 000 years ago to the present. Our methodology groups populations with a similar spatiotemporal distribution using a K-means algorithm. Then, we infer the impact of natural selection and the dispersion of alleles in these populations using a Hidden Markov Model. We classified these alleles based on their CADD scores, which allow us to classify how deleterious a particular allele is and infer the coefficient selection of putatively deleterious alleles. Our results show that putatively deleterious alleles have a selective coefficient equal to -0.2, and we also infer that the alleles move 10 grid cells per generation (from 91x44 km to 28x44 km on average per grid cell). Additionally, our results show that it is essential to estimate the starting time of a selective pressure in order to correctly infer the impact of natural selection. Our methodology demonstrates that it is possible to group spatiotemporal genomic data to infer the impact of alleles under natural selection for both advantageous and deleterious alleles.

3

Introducción

3.1. Selección natural

La selección natural es un mecanismo evolutivo que fue descrito por Darwin (1856) para explicar cómo los seres vivos logran adaptarse a su entorno a través de la variación. La selección actúa sobre la diversidad de rasgos presentes en una población favoreciendo aquellos que son más ventajosos en un ambiente determinado. Los individuos que poseen estos rasgos ventajosos experimentan un aumento en su capacidad de sobrevivir y reproducirse en su ambiente, lo que se conoce como *adecuación biológica* o *aptitud*. Aquellos individuos que logran adaptarse mejor a su ambiente tienen una mayor adecuación y, por lo tanto, aumentan sus posibilidades de dejar descendencia. Si estos rasgos ventajosos son heredables, se transmiten a las generaciones futuras, aumentando su frecuencia en la población hasta que se fijan en la población. A lo largo del tiempo, estos cambios graduales pueden llevar a la formación de una nueva especie, que será lo suficientemente diferente de la población original.

El origen de la variación y cómo es que estos rasgos se heredan fue un problema dentro de la teoría de la evolución por selección natural. De hecho, en la primera edición del Origen (1859), capítulo 1, Darwin escribiría

> *"The laws governing inheritance are quite unknown; no one can say why the same peculiarity in different individuals of the same species, and in individuals of different species, is sometimes inherited and sometimes not so...*

En 1868, Darwin propondría a las *gémulas* como las partículas portadoras de la herencia de rasgos hacia la progenie, los cuáles a su vez eran sensibles al ambiente. Sin embargo, esta teoría sería rápidamente refutada por medio de los experimentos de su propio primo, Francis Galton. Estos experimentos consistieron en la donación de sangre proveniente de conejos mutilados hacia conejos sanos con el propósito de observar descendencia mutilada, lo cuál no sucedió (Liu,

2018). Curiosamente, de manera casi paralela, Gregor Mendel comenzaría sus experimentos sobre la herencia de caracteres con los chícharos, permitiendo publicar sus resultados en 1865 en un artículo llamado *Experiments in Plant Hybridization*. Sin embargo estos trabajos estarían olvidados durante los próximos 35 años (Kimura, 2020).

La integración de ambas teorías parecía natural, no obstante el surgimiento de los grupos *mendelianos*, *neo-lamarckianos*, *mutacionistas* y *biometricionistas* durante principios del siglo XX segregaron a los grupos *darwinistas*. Personajes como Thomas Hunt Morgan, quien se encontraba investigando las mutaciones inducidas en *Drosophila* (Morgan, 1911) mantuvieron durante mucho tiempo una postura *antidarwiniana* [1], y al ser personajes muy influyentes en su campo, mantuvieron a la selección natural al margen de los laboratorios y algunos centros académicos. Esto se debía principalmente a que existía un espíritu positivista en la época, el cual no aceptaba evidencias empíricas sino sólo evidencia que surgiera por experimentación. La necesidad de los biólogos de la época, de colocar a la biología en un contexto experimental más que de *narrativa*, pretendía alejar el carácter *metafísico* de la biología, esperando encontrarse con una *teoría unificada de la biología* (Smocovitis, 1992).

Los primeros esfuerzos para integrar la teoría evolutiva de Darwin serían mediante un contexto biométrico bajo las *Leyes de Mendel*. G. H. Hardy y W. Weinberg lo lograrían de manera independiente, y durante los años 1918-1931 R. A. Fisher, J. B. S. Haldane y S. Wright sentarían las bases para definir el marco teórico que explicaría más adelante cómo la selección natural modifica las frecuencias alélicas de las poblaciones, convirtiéndolos en los precursores de un nuevo campo de estudio en el contexto de la Síntesis Moderna Evolutiva: la *Genética de Poblaciones*. Para poder entender los modelos tradicionales de detección de selección en genética de poblaciones, hay que explicar que la gran mayoría están hechos para trabajar con dos **alelos**, A y a debido a que es raro encontrarse sitios en el genoma con 3 o 4 alelos. Por ejemplo, en humanos, Nelson *et al.* (2012) encontró que los sitios con 3 alelos representan 2 % de los SNPs y mientras que los sitios con 4 alelos representan 1.6 % de los SNPs.

La inferencia de selección en poblaciones diploides se produce cuando la adecuación, representada por el término w y definida como la cantidad de descendientes viables engendrados

[1] Se entiende como *antidarwiniano* en el sentido de proponer una fuerza alternativa que origina a las especies. Entre los 1900 y 1915, la selección natural era una propuesta evolutiva rechazada por la mayoría de los círculos de la academia estadounidense al considerarla contraria a las Leyes de Mendel. Paralelamente a la selección natural, Hugo de Vries, uno de los re-descubridores del trabajo de Mendel propuso que las mutaciones eran un factor a tener en cuenta por el que las especies se originaban y no así el ambiente actuando a favor de ciertos rasgos: "*... Species have arisen from one another by a discontinuous, as opposed to a continuous process [...]. The new species appears all at once; it originates from the parent species without any visible preparation and without any obvious series of transitional forms...*". T. H. Morgan escribiría: "*It appears that new species are born; they are not made by Darwinian methods, and the theory of natural selection has nothing to do with the origin of species, but with the survival of already formed species*". Más tarde por influencia de su alumno más destacado, T. Dobzhansky, se adheriría al darwinismo (Gould, 2002)".

por individuos con un genotipo particular, no es la misma para los tres posibles genotipos. Por ejemplo, si en un determinado caso la adecuación del genotipo homocigoto AA es superior a la de los otros dos genotipos, se produce selección direccional favoreciendo al grupo portador de ese genotipo:

$$w_{\mathrm{AA}} > w_{\mathrm{Aa}} > w_{\mathrm{aa}} \tag{3.1}$$

La selección direccional es uno de los muchos términos con los que se ha descrito los tipos de selección dentro de la terminología de la evolución molecular y que se describen a continuación:

- Selección purificadora: también conocida como selección negativa, ocurre cuando una nueva mutación influye de manera negativa en el desarrollo de los portadores y lleva a un éxito reproductivo inferior comparado a otros organismos no portadores. Los alelos que se encuentren en la población suelen ser removidos. Dependiendo de la fuerza de selección, tipo de herencia o tipo de dominancia, la frecuencia de estos alelos suele ser baja.

- Selección direccional: también conocida como selección positiva, ocurre cuando una nueva mutación influye de manera positiva y se puede observar en un mayor éxito reproductivo además de un aumento en la frecuencia de portadores de dicha mutación en generaciones posteriores, comenzando en $1/2\mathrm{N}$ y tendiendo al $100\,\%$ de frecuencia.

- Selección balanceadora: se da mediante tres mecanismos distintos. *Ventaja del heterocigoto* que ocurre cuando los individuos que son heterocigotos tienen una adecuación mayor a la de sus formas homocigotas, por ejemplo, porque las formas homocigotas son deletéreas[2]. *Selección según la frecuencia* (Takahata *et al.*, 1975), la cuál consiste cuando en una población está determinada la eficacia relativa en diferentes situaciones o contextos. Mientras que la *la selección direccional en ambientes heterogéneos en el tiempo y/o en el espacio* ocurre cuando la selección natural favorece un cambio gradual en una característica de una población hacia una forma particular en función de las condiciones cambiantes del ambiente (Johri *et al.*, 2022).

No obstante, en algunos casos, una mutación puede no afectar la adecuación (w) de los individuos portadores, lo que la convierte en una mutación neutral. Esto se debe a que la mutación no se encuentra sometida a ninguna forma de selección y, por lo tanto, se mantiene por medio de lo que se conoce como *evolución neutral.*. La evolución neutral fue ampliamente estudiada por Kimura *et al.* (1968; véase Figura 3.1) el cuál sostiene que, bajo un modelo de selección natural como única fuerza moldeadora de secuencias, la **carga genética** de la secuencia de un gen debería ser alta, pero sus observaciones en los aminoácidos en α y β

[2]De hecho, el mismo Hugo de Vries descubrió este fenómeno en su planta modelo *Oenothera lamarckiana* para sus trabajos en mutaciones (Gould, 2002)

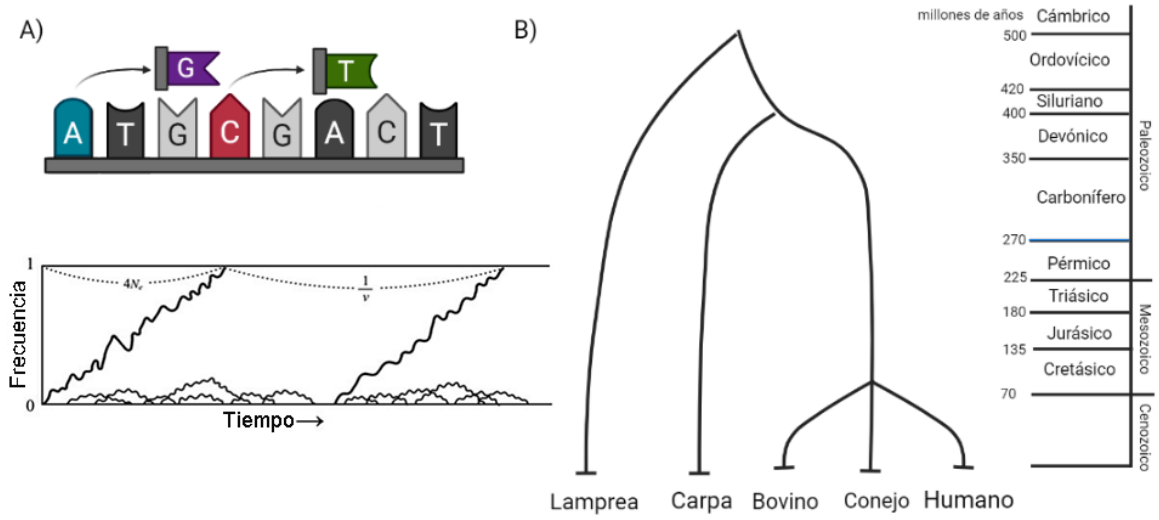


Figura 3.1: Pauling y Zuckerland (1962) habían sugerido que la tasa de sustitución es constante en todo el genoma al contrario de lo propuesto por Freese y Yoshida (1965)sugiriendo mutaciones que no están bajo selección natural. Por otro lado, Kimura (1969) describió que la proporción de mutaciones era mayor en las zonas no funcionales de la hemoglobina que en las partes funcionales. *Figura A*: Representación gráfica de sustitución de nucleótidos que dan lugar a nuevos alelos (A y C, respectivamente) y frecuencia alélica de los alelos derivados de una secuencia a lo largo del tiempo. La tasa de sustitución es mayor en la tercera base de cada triplete debido a la degeneración del código, según Kimura (2020), la tasa de sustitución en proteínas como H4 de las histonas o hemoglobinas es de $3x10^{-9}$ a $4x10^{-9}$ al año. Las mutaciones generadas pueden fijarse en los sitios neutros debido a la deriva génica. *Figura B*: A partir de las secuencias homólogas de aminoácidos de hemoglobina de distintas especies, Kimura estimó la edad aproximada de divergencia de cada una de las proteína a partir de la tasa de sustitución de sitios neutros. Imagen modificada y extraída de Kimura (1969) y Kimura (2020).

globulinas le indicaron que, bajo un modelo neutro, esta carga era casi nula. Al analizar la variación que existe entre las secuencias de diferentes especies, señaló como responsables de la variación en las secuencias a los fenómenos de deriva génica que actúan concretamente en los sitios neutros. La existencia de estas mutaciones serían relevantes más adelante para la construcción de modelos nulos para la detección de selección en el genoma (Nielsen, 2001).

En Genética de Poblaciones, se puede saber el tipo de selección actuando en una mutación e inferir su fuerza por medio de las frecuencias alélicas. Al analizar los cambios de las **frecuencias alélicas**, podemos inferir a la selección matemáticamente y comprender cómo la selección natural incide en los patrones de variación genética[3]. Se considera *s* o *coeficiente de selección* a una medida de las diferencias en la adecuación entre portadores de distintos alelos en una

[3]Estos análisis toman en cuenta el efecto de otras fuerzas evolutivas, como la deriva génica, y pueden inferir el impacto de la selección natural al contrastar los atributos de los cambios en las frecuencias alélicas contra las expectativas de los cambios en frecuencias alélicas en variantes neutrales

población. Esta medida surge de una relación proporcional entre el número de portadores de un alelo N_A y un segundo alelo N_a de una población[4] con N cantidad de individuos. En secuencias haploides, la frecuencia de individuos con el alelo A se obtendría mediante la relación de A:$f_A = N_A/(N_A + N_a)$. Para secuencias diplodes se considera solamente la cantidad de alelos en la población (A en los homocigotos AA y los heterocigotos Aa), por lo que se asume que solamente existen dos alelos. Suponiendo que en la siguiente generación (cuyo valor es considerado como tiempo discreto, es decir, poblaciones sin sobrelapamiento generacional) es $t=1$, entonces, cada uno de los hijos será portador del alelo A o a respectivamente, por lo que cada padre tendrá la posibilidad de dejar w cantidad de hijos portadores de un alelo (w_A y w_a respectivamente). Nótese que w_A y W_a se refiere al número de portadores promedio del alelo A y a, respectivamente, que se producen exitosamente de sus padres. Matemáticamente, para una generación t, podemos obtener la cantidad de portadores A como $f_A(t) = w_A N_A$, mientras que para obtener la frecuencia de portadores del alelo A en una población de tamaño N, esta se obtendría como:

$$f_\mathrm{A}(1) = \frac{w_\mathrm{A} N_\mathrm{A}}{w_\mathrm{A} N_\mathrm{A} + w_\mathrm{a} N_\mathrm{a}} \tag{3.2}$$

Donde el resultado de $w_A N_A$ y $w_a N_a$ respectivamente será el número de portadores de A y a en la generación $t=1$. Si sólo estuviéramos interesados en la frecuencia del alelo A entonces cambiaremos el número de portadores N_A a la frecuencia f_A y f_a en la población.

Sin embargo, vale la pena señalar que esta expresión puede simplificarse como:

$$f_\mathrm{A}(t+1) = \frac{f_\mathrm{A(t)}}{f_\mathrm{A}(t) + (w_\mathrm{a}/w_\mathrm{A}) f_\mathrm{a}(t)} \tag{3.3}$$

Por lo que la frecuencia de A dependerá entonces de la relación w_a/w_A. Es justamente esta relación la que nos proporcionará el valor de coeficiente de selección s del alelo A en la población en términos de una proporción:

$$\frac{w_\mathrm{a}}{w_\mathrm{A}} = 1 - s \tag{3.4}$$

Por lo tanto s es:

$$s = 1 - W \tag{3.5}$$

Donde W es la *adecuación relativa* (relativo en este caso al alelo A) o *fitness* de un genotipo o *la probabilidad relativa de que el genotipo se produzca.* Es decir, si en una población de tamaño constante de $N=100$, el alelo A y el alelo a se encuentran en igual frecuencia y en la generación $t+1$ el alelo A se encuentra en el 65 % de la generación, teniendo un valor $s(A)$=0.46 (Nielsen

[4]En la mayoría de los modelos de genética de poblaciones se trabaja únicamente con dos alelos representados como A y a en este trabajo. Uno representado al alelo de referencia o de mayor frecuencia y otro al alternativo o el segundo de mayor frecuencia.

y Slatkin, 2013). Los valores de s van desde 0 hasta 1, donde 0 es prácticamente neutro y 1 un valor muy alto de selección direccional para crecer hasta prácticamente fijarse en pocas generaciones. Este concepto también nos da una idea de qué tan rápido se pueden propagar los alelos con alto s en una población, sin embargo, la dinámica de los alelos a lo largo del tiempo depende de N. Si este valor crece a lo largo del tiempo, la selección actuará de forma cada vez más efectiva. Es decir, si se encuentran alelos con un coeficiente de selección muy alto en una población en expansión a lo largo del tiempo, la selección actuará de forma más efectiva en dichos alelos (Enard, 2021; Ohta, 1972).

También, se han desarrollado distintos métodos para inferir s en poblaciones panmícticas (Van Valen, 1965); observar cambios en s en función de N_E (Ohta, 1972), e inclusive estimar s a partir de distintas frecuencias a través del tiempo (Bollback *et al.*, 2008; Takahata *et al.*, 1975). La importancia de s es tal que su uso es bien conocido para la industria alimenticia y farmacéutica. Como ejemplo, el s ha sido utilizado ampliamente en plantas (Primack y Kang, 1989; Scheiner y Callahan, 1999) para el mejoramiento de algunos caracteres con potencial en la agricultura o en la búsqueda de mutaciones que podrían mejorar el *fitness* de algunos microorganismos contra algunos medicamentos (Gordo *et al.*, 2011).

La inferencia de selección en variantes alélicas en genes de seres humanos es uno de los campos que también se ha estudiado en los últimos años desde que se ha secuenciado el genoma humano. Los estudios, enfocados principalmente en el sector salud, nos han permitido conocer cómo los seres humanos nos hemos adaptado a distintos ambientes (Byars *et al.*, 2010; Mathieson *et al.*, 2015; Moorad y Walling, 2017), pero la gran mayoría de estas inferencias solo representan un tiempo muy limitado al estudiarse en genomas modernos, lo que implica una visión limitada de la historia. Afortunadamente, ahora contamos con tecnologías que mejoran la extracción y secuenciación de genomas a partir de restos biológicos que datan de hace cientos o miles de años, lo que abre la puerta a la posible exploración de los cambios en la dinámica de las poblaciones humanas antiguas y cómo las condiciones del pasado moldearon los genomas del presente (Dehasque *et al.*, 2020; Pickrell y Reich, 2014).

3.2. Métodos de detección de selección natural

Diferenciar qué alelos se encuentran bajo selección de aquellos que han cambiado sus frecuencias exclusivamente por eventos demográficos (expansiones, vicarianza, deriva génica y migración) representa un gran trabajo metodológico (Sabeti *et al.*, 2006). Sin embargo, mediante métodos estadísticos, se han podido detectar alelos bajo selección natural a partir de la frecuencia alélica de dos o más **demes** en contacto (Vitalis *et al.*, 2014) de un solo gen de

interés comparando las frecuencias alélicas contra la frecuencia de alelos tentativamente neutros (Ameur *et al.*, 2012; Mathieson y Mathieson, 2018) o buscando desequilibrio de ligamiento en múltiples alelos (Weir y Cockerham, 1978). Sin embargo, cada metodología infiere alelos bajo selección dentro de una ventana específica de tiempo (Sabeti *et al.*, 2006):

- de millones de años, infiriendo qué alelos resultaron ventajosos entre especies basados en las mutaciones funcionales que existen entre ambos organismos, teniendo como punto de comparación las secuencias de una proteína homóloga entre ambas especies (i.e. McDonald y Kreitman, 1991).

- de menos de 250 mil de años, a través de la detección de zonas neutras que sufrieron *arrastre por selección* o *selective sweep* (i.e. Hudson *et al.*, 1987; Tajima, 1989).

- de menos de 80 mil años, reconociendo qué alelos son ancestrales y derivados y calculando la frecuencia de los alelos derivados en una población y buscando alelos derivados con una frecuencia alta (i.e. Fay y Wu, 2000).

- de entre 50 a 70 mil años, detectando diferencias entre poblaciones, ya sea por medio de *FST* (Lewontin y Krakauer, 1973) o por medio del uso de haplotipos largos analizando señales de *arrastre* en sitios neutros (Toomajian *et al.*, 2003).

En general, la gran mayoría de los métodos se basan en la comparación de un modelo nulo a partir de las frecuencias/distribución de alelos o sitios tentativamente neutros (concepto heredado de Kimura *et al.*, 1968) comparando los alelos/sitios de interés y a partir de estas comparaciones se puede conocer si existe selección (Kreitman, 2000; Nielsen, 2001; Vitalis *et al.*, 2014). Los modelos que parten de un **equilibrio neutral**[5] de las frecuencias permiten diferenciar aquellos alelos que no se encuentran bajo este *equilibrio* y, si es el caso, determinar qué tipo de fuerza de selección actúa en ellos. Los sitios tentativamente neutrales permiten el desarrollo de un modelo nulo sencillo y, posteriormente, un buen análisis estadístico puede determinar la neutralidad de un sitio (Kreitman, 2000).

Una de las primeras propuestas hechas para detectar eventos de selección fue realizada por Tajima (1983,1989), utilizando como base el modelo neutral de Kimura como hipótesis nula, demostrando que las diferencias que existen entre el número de sitios segregantes observados (S) y esperados dada la variación genética observada (π) puede indicar que la selección natural está actuando en un sitio particular (Tajima, 1989). Debe notarse que el método de Tajima no trata con frecuencias de alelos específicamente. El método de Tajima utiliza un par de parámetros conocidos como Theta y pi que dependen de frecuencias alélicas para determinar la

[5]Equilibrio neutral se refiere a los patrones de variación genética de los alelos bajo el supuesto de que los alelos sólo evolucionan debido al impacto de mutaciones y de la deriva génica (Achaz, 2009).

neutralidad de un loci. El método de Tajima puede indicar si algún loci en particular puede estar bajo selección positiva o negativa. Sin embargo, este estadístico es susceptible al impacto de la historia demográfica (Nielsen, 2001). Esta particularidad es compartida con otros métodos y, por lo tanto, debe considerarse la historia demográfica para detectar alelos bajo selección natural.

Los métodos de detección de selección se pueden clasificar en dos tipos de métodos: 1) aquellos que utilizan la distribución de las frecuencias alélicas (*Site Frequency Spectrum, SFS*) y realizan análisis estadísticos con respecto a la distribución de alelos sin selección y 2) aquellos que comparan la variabilidad o divergencia de diferentes clases de mutaciones tales como las mutaciones *no-sinónimas* y las *sinónimas* (Nielsen, 2001). En todos los métodos para identificar regiones nucleotídicas bajo selección natural se han realizado esfuerzos importantes para incorporar el impacto de la demografía histórica para obtener modelos cada vez más realistas y detectar de mejor manera alelos bajo selección natural (Mathieson y McVean, 2013; Muktupavela *et al.*, 2021; Vitalis *et al.*, 2014).

Los métodos para detectar alelos bajo selección a partir de datos de frecuencias alélicas se dividen en dos tipos: de un solo locus (a) y de múltiples loci (b) y utilizan estadísticos que resumen los valores obtenidos de un Espectro de Frecuencias por Sitio o *Site Frequency Spectrum SFS*. Entre los más populares del tipo *1a* está el test Ewens-Waterson (Watterson, 1977) que utiliza frecuencias alélicas y detecta alelos bajo selección con base en la diferencia estadística entre los heterocigotos observados y los esperados, tal como el antes mencionado D de Tajima (Tajima, 1989). Vale la pena recordar que π es un valor que se ve afectado por la pérdida de diversidad producto del barrido selectivo, mientras que S no se ve fuertemente afectada, por lo tanto, en cuanto *D de Tajima* tome valores muy grandes o muy pequeños, la hipótesis nula de neutralidad es rechazada (Enard, 2021; Nielsen, 2005). En los modelos tradicionales de inferencia de selección descritos anteriormente, los cambios demográficos (migración, cuellos de botella, expansión) en las poblaciones pueden variar el valor de π, por lo que trabajan bajo supuestos como *no migración* o *tamaño poblacional constante*, aunque en los últimos años se han realizado análisis que consideran estas variables Freedman *et al.* (2016). Por el otro lado, el estadístico *1b* más popular, aquellos que trabajan con múltiples loci, es el test de Lewontin-Krakauer (Lewontin y Krakauer, 1973). Mientras que para secuencias, el test HKA (Hudson *et al.*, 1987) compara secuencias dentro de especies y entre especies. En este test, en ausencia de selección, el número de sitios segregantes (sitios polimórficos) dentro de la especie y el número de sitios fijados (divergencia) entre especie son proporcionales a la tasa de mutación, y por lo tanto, la proporción entre estos dos valores debería ser el mismo para todos los loci. Entonces, cuando se analiza cada loci por separado, si la tasa de divergencia es muy alta, existe evidencia de selección natural positiva del loci estudiado (Nielsen, 2001).

Finalmente, del tipo 2, los test que usan dos diferentes tipos de mutaciones: las mutaciones sinónimas y no sinónimas. Básicamente, trabajan bajo el supuesto de que la selección favorecerá la formación de variantes raras y compara estas variantes entre especies, por lo que es importante señalar que se utilizan para conocer selección en periodos muy anteriores al presente al buscar mutaciones adaptativas con un importante impacto en la funcionalidad (aminoácidos, secciones reguladoras, secciones codificantes, etc.). El test McDonald-Kreitman (McDonald y Kreitman, 1991) es de los más populares en su tipo. Compara la tasa de mutaciones no sinónimas y sinónimas dentro de la misma especie y después compara con la tasa de mutaciones no sinónimas y sinónimas entre especies cercanas. Si la selección sólo afecta a las mutaciones *NS*, la selección negativa actuará sobre estas variantes reduciendo el número de mutaciones *NS*, caso contrario al de la selección positiva la cual aumentará el número de mutaciones *NS* de ser el caso; ambas en comparación al número de mutaciones *S*. Es importante señalar que una de sus debilidades es que detectar selección negativa contra mutaciones deletéreas *NS* puede ser engañoso, ya que en muchas ocasiones, cuando existen cuellos de botella largos, la selección purificadora no es tan eficiente, lo que permite a este tipo de mutaciones alcanzar frecuencias relativamente altas (Enard, 2021).

Los estadísticos que infieren eventos de selección natural son dependientes del conocimiento de la historia demográfica. Para muchos de estos eventos no se tienen datos por falta de registros históricos, pero esta historia se puede inferir a través del análisis de datos genómicos. Recientemente se han construido metodologías que integran la *historia demográfica* para la inferencia de genes bajo selección. Un ejemplo de ello es un estudio reciente que identificó genes seleccionados durante la domesticación de los perros (Freedman *et al.*, 2016). Muy recientemente, desde la década de 1980 y gracias a la disponibilidad de series de datos genómicos de diferentes periodos, se han desarrollado métodos estadísticos que permiten la inferencia de la selección en periodos recientes de tiempo. Este tipo de datos y estadísticos tienen una mayor ventaja sobre los estadísticos de un solo periodo tales como conocer el cambio de frecuencias de un alelo. Este dato a su vez nos puede dar información sobre la fuerza de la selección sobre ese alelo (Mathieson y McVean, 2013); estimar el tamaño efectivo de las poblaciones (Foll *et al.*, 2014, 2015); estimar la edad de una mutación alélica (Malaspinas *et al.*, 2012), etc. Todos estos métodos utilizan como modelo nulo los cambios de las frecuencias de un alelo neutro desde su emergencia hasta el presente y también analizan los cambios en el aumento o la disminución de las frecuencias alélicas. Como existen periodos en los que el registro de datos de frecuencias alélicas es nulo, se utilizan procesos Ocultos de Markov (*Hidden Markov Chains*, *HMC* o cadenas ocultas de Markov), para *rellenar* los huecos de información e inferir la frecuencia alélica y otros parámetros como la selección natural y el tamaño poblacional. Este tipo de inferencia de selección ha sido conocido popularmente como *Detección de Selección a partir de Datos*

Genéticos en Series de Tiempos (Bank *et al.*, 2014). Hay que hacer notar también se pueden contrastar rasgos de variación contra marcadores neutrales como la región control mitocondrial (Cohen, 2002) y algunos intrones nucleares (Parsch *et al.*, 2010).

3.2.1. Datos genéticos como series de tiempo

Llamamos *Series de Tiempo* (**ST**) a un conjunto de datos ordenados cronológicamente. Estos datos pueden ser encontrados en un tiempo llamado T (por ejemplo T=1950-Presente), mientras que las observaciones hechas en un tiempo específico t (arbitrariamente, el año 1960) se denomina como *X(t)* (por ejemplo, todos los datos disponibles de 1960; Eshleman *et al.*, 2003). El interés por este campo se debe principalmente por su uso en las ciencias económicas, meteorológicas y físicas; campos que han usado a las ST para poder analizar, agrupar y hasta predecir los fenómenos de su interés ya que la secuencia de datos puede explicar la dinámica del sistema que genera. Las herramientas para el minado[6] de datos (nombre con el que se refiere al proceso de descubrir patrones y relaciones útiles en grandes conjuntos de datos en el campo de *Ciencia de Datos*), habían sido desarrolladas desde la década de los 2000 (Keogh y Kasetty, 2002) popularizándose desde entonces entre la investigación y la industria. No obstante, bajo el concepto empleado por el campo antes mencionado, los primeros minados de datos ocurrieron en la biología a través del estudio de las poblaciones a nivel molecular durante la década de 1960 a través del análisis de la diversidad de aminoácidos en proteínas[7] (análisis de *fenotipos electroforéticos* o *electromorfos*; Dayhoff, 1969) debido a la información que pudiera arrojar la relación entre ancestro-descendencia analizando los cambios mutacionales entre las secuencias de aminoácidos. Cabe destacar que serían estos mismos datos que Kimura usaría para proponer la teoría neutral de la evolución molecular (Lewontin, 1991), y no fue sino hasta el 2008 que las herramientas del minado de datos se integraron en genética de poblaciones analizando las frecuencias de los alelos de CCR5-Δ32 en poblaciones humanas de Europa en distintas temporalidades (Bollback *et al.*, 2008). Aunque bien Bollback y sus colaboradores no fueron los primeros en integrar los datos biológicos en series de tiempo (Anteriormente, Fisher y Ford (1947) y Wright (1948) habían estudiado los distintos fenotipos de alas expuestos por la mariposa nocturna, *Panaxia dominula*, mientras que Wright y Chambers (1920) realizarían las

[6]Aunque el término *minado* puede traducirse perfectamente del inglés *mining*, estamos ante un caso de un fenómeno lingüístico conocido como *falsos amigos* o *false cognates*. En esta tesis solo estamos empleando el término traducido literalmente al español y su significado original empleado en la Ciencia de Datos para honrrar al idioma.

[7]Cabe destacar que Margaret Dayhoff fue la primera persona que aplicó métodos matemáticos y computacionales a la bioquímica tal como la matriz de substituciones, la cual nos permiten puntuar los alineamientos de secuencias; a ella se le atribuye el código de una sola letra para aminoácidos y el *Atlas of Protein Sequence and Structure,* Sin su contribución es probable que éste y muchos otros trabajos relacionados con la bioinformática serían impensables.

mismas observaciones sobre el ganado), sí fue el primero en proponer un modelo de verosimilitud para inferir selección suponiendo un modelo Wright-Fisher a partir de un modelo oculto de Markov. Es importante recordar que la máxima verosimilitud es un método que permite estimar los parámetros desconocidos de un modelo basándose en la probabilidad de observar un conjunto de datos, es decir, hallar los valores desconocidos del parámetro de la selección que hacen que los datos observados (frecuencias alélicas) sean más probables. Estos modelos se utilizan para poder inferir, a partir de un estado observable, un conjunto de parámetros. En los modelos ocultos de Markov, las frecuencias alélicas fueron utilizadas como el estado observable mientras que el número de alelos en la población es el estado oculto. Bajo los modelos ocultos de Markov es posible trazar la trayectoria completa del número de alelos en la población a través del tiempo. Debido a la ausencia de modelos que puedan calcular la probabilidad de transición entre diferentes estados, Bollback utilizó *Ecuaciones hacia atrás de Kolgomorov*, las cuales permiten conocer la distribución de probabilidad del estado *X(t)* en un momento t, a través del análisis de los estados X_i en otro tiempo t_i. Malaspinas *et al.* (2012) por otro lado, agregaron análisis de máxima verosimilitud para poder inferir s, el tamaño de la población y la edad del alelo (entendiéndose como el tiempo en generaciones en la que la mutación apareció). Steinrücken *et al.* (2014) logró, por medio del análisis espectral de la probabilidad de distribución, la misma que permite conocer la trayectoria más probable del alelo, inferir selección en secuencias diploides. Sin embargo, la metodología de Steinrücken *et al.* (2014) no permite inferir la edad del alelo. Por otra parte, Mathieson y McVean (2013) analizaron las frecuencias alélicas del mismo modo que Bollback, obteniendo los valores de máxima verosimilitud de migración m en un espacio de dos dimensiones a través del análisis de varias *demes* de polillas siendo el único trabajo hasta ahora que ha podido inferir migración y selección natural en un espacio de dos dimensiones. Schraiber *et al.* (2016) desarrollaron otro método utilizando *Cadenas de Markov-Montecarlo* (*Markov Chain Monte-Carlo, MCMC*) la cual permite analizar los cambios de frecuencias más probables del alelo, lo que permite inferir la fuerza de la selección natural en historias demográficas con tamaños de población variables. Sin embargo este último método es el menos recomendado por su costo computacional. Finalmente, He *et al.* (2020) desarrollarían un método que extendería el trabajo de Bollback, el cual consiste en tomar valores de *Ne* no constantes, analizando las trayectorias más probables de los alelos y finalmente infiriendo selección.

El análisis de las series de tiempo con datos de DNA pueden ser más confiables que los datos de un sólo punto en el tiempo ya que estos proporcionan más información sobre la selección debido a que se puede conocer la trayectoria de las frecuencias alélicas a lo largo del tiempo. Dichos cambios pueden permitir calcular con mayor exactitud la fuerza de selección y s (Bollback *et al.*, 2008) y poder tomar en cuenta distintos modelos demográficos (Mathieson

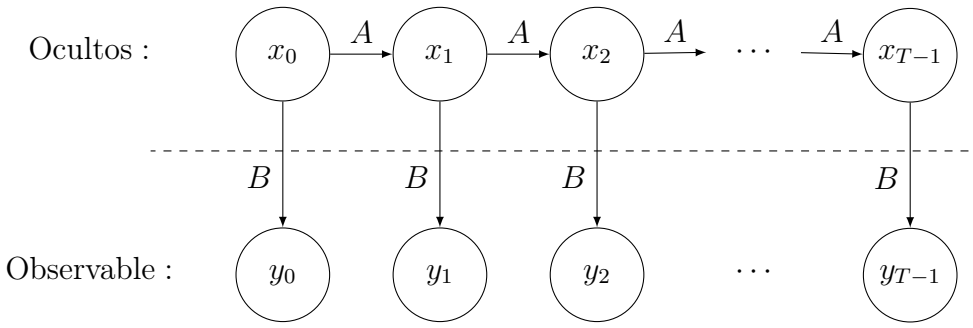


Figura 3.2: Estructura básica de un Modelo Oculto de Markov. Se compone principalmente del conjunto de estados ocultos se representan como x_{T-1} y los estados observables y_{T-1}. Las probabilidades de transición se representan como un conjunto de valores $A = a_1, a_2, a_3...a_n$ mientras que las probabilidades de emisión como el conjunto $B = b_1, b_2, b_3...b_n$. Las probabilidades varían de acuerdo a la cantidad de estados posibles que tengamos.

y Mathieson, 2018). En seres humanos, debido a la creciente disponibilidad y el aumento en la calidad del DNA antiguo (*aDNA*), se han podido entender mejor los procesos evolutivos que moldearon nuestro genoma tales como cambios en la dieta y adaptación a distintos patógenos (Dehasque *et al.*, 2020; Marciniak y Perry, 2017), por lo que su disponibilidad permitió conocer qué alelos y qué regiones se encontraban bajo selección (Mathieson *et al.*, 2015).

3.2.2. Métodos markovianos

Las Cadenas Ocultas de Markov (Modelos Ocultos) son un modelo probabilístico que busca determinar parámetros desconocidos (o el *estado oculto* x_t) a partir de parámetros conocidos (y_t) en un tiempo t, por lo que en $t + 1$ el estado dependerá de t, es decir, que la probabilidad de un estado y_{t+1} depende entonces de que haya ocurrido X_{t+1}; este a su vez dependerá de que haya ocurrido X_t, que a su vez dependió del estado X_{t-1}, por lo que al final se obtiene una cadena de eventos compuestos por X y Y de tamaño t_M. Por lo anterior, se observa que una secuencia Y es de:

$$P(Y) = \sum_X P(Y|X)P(X) \tag{3.6}$$

Es decir, la probabilidad de una secuencia Y es igual a la suma de las probabilidades de Y dado X por la probabilidad de X, donde X son los nodos ocultos y Y son los estados observables. En la figura 3.2, se puede observar la estructura general de un modelo oculto de Markov.

Estos cambios de estados están compuestos a su vez por un conjunto de estados ocultos $Q = x_0, x_1, x_2...x_{T-1}$, un conjunto de valores $V = y_0, y_1, y_2...y_{T-1}$ observables; sus probabilidades iniciales π_i, el cual es la probabilidad de que el estado Q_o (estado inicial) siga siendo Q_o en la

siguiente serie de tiempo. Se define el conjunto de probabilidades $A = a_i j$ para el cambio entre estados:

$$a_{ij} = P(q_t = j | q_{t-1} = i) \tag{3.7}$$

el cual se lee como la probabilidad de estar en el estado j dado que en la serie anterior $t-1$ se estaba en un estado i. También se define el término $B = b_j v_k$ como:

$$b_i(v_k) = P(y_t = v_k | q_t = j) \tag{3.8}$$

el cual se lee como la probabilidad de observar el valor v_k cuando se está en un estado j en el instante q_t; por lo que al final se obtendrá una matriz de probabilidades dado los estados observados y los estados no observados. El clima es uno de los ejemplos clásicos que nos permitirá entender las HMC:

- Suponiendo que durante la cuarentena de COVID-19 se desea saber si tendrá que usar una chamarra o no durante los días lluviosos. Usted se encuentra encerrado y no puede ver en su totalidad el cielo, pero si puede ver que algunos vecinos suelen salir con su paraguas (P) o sin él (NP) y por lo tanto aprendió a relacionar el estado del clima (oculto) a partir de la cantidad de vecinos que ve con paraguas (observable). Suponiendo que de 10 vecinos que observamos, 8 de ellos llevan un paraguas y 2 no en los **días con lluvia** y por el contrario, 4 llevan paraguas y 6 no en los **días sin lluvia**; para el clima, sabemos que si un día es soleado, la probabilidad de que al día siguiente sea soleado (S) es del 0.8, por el contrario la probabilidad de que llueva (R) al día siguiente es del 0.2. Si el día fue lluvioso, la probabilidad de que se mantenga así es del 0.6, siendo su complemento (que el día fuera lluvioso y al siguiente sea soleado) es del 0.4. Todas estas probabilidades son llamadas formalmente como *Probabilidades de transición*, las cuales describen las probabilidades de cambios de estado ocultos. Por otro lado, las *probabilidades de emisión* describen la probabilidad de observar individuos con o sin paraguas dado el clima. Gráficamente se observaría de la siguiente forma:

Si llegamos a obtener una secuencia de datos (por ejemplo, 15 días), tendríamos una tabla de la siguiente forma:

S	S	S	S	R	R	R	S	S	S	R	S	S	S	R
P	NP	NP	NP	P	P	P	NP	NP	P	NP	NP	NP	P	P

Cuadro 3.1: Tabla de registro de 15 días del clima y la presencia o ausencia de personas usando Paraguas. S y R son los estados *soleados* y *lluviosos*; P y NP se refiere a *paraguas* y *sin paraguas* respectivamente.

Calculando la probabilidad en una secuencia de 15 días (probabilidad a priori) de cada estado

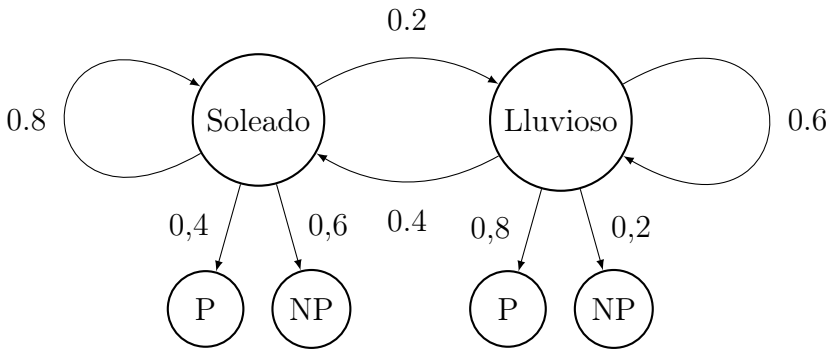


Figura 3.3: Diagrama de Markov para el clima.

oculto sería $P(S) = Dias_{soleados}/Dias_{totales} = 0{,}67$ y $P(R) = Dias_{lluviosos}/Dias_{totales} = 0{,}33$.

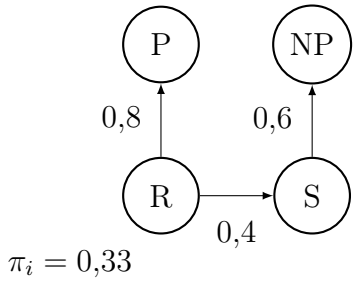


Figura 3.4: Diagrama de Markov para los primeros dos días de observación.

Entonces, suponiendo que en nuestras siguientes tres mañanas hemos observado la secuencia *Con Paraguas, Sin Paraguas, Sin Paraguas*, por lo que al final tendremos hasta ocho posibles combinaciones de estados ($2 * 2 * 2 = 8$). Basados en los estados anteriores, calcularemos la probabilidad de cada una de las combinaciones para saber cuál fue la secuencia del clima más probable (Estimación de la probabilidad a *posteriori*). Para el caso de este ejemplo, usaremos los primeros dos días (Figura 3.4). Como probabilidad inicial (a *priori*) tenemos que $\pi_i = P(R) = 0{,}33$; la probabilidad de transición de R a S es de 0,4; la probabilidad de usar paraguas dado que es un día lluvioso $P(P|R) = 0{,}8$; la probabilidad de no usar paraguas dado que es un día soleado es de $P(P|S) = 0{,}6$; nótese que la secuencia es $Paraguas, SinParaguas$ para los primeros dos días. De esta forma, dado una de las propiedades de Markov que habla de la independencia de los estados dado t también serán independientes en el resto de T, por lo tanto podemos empezar calculando la probabilidad de una de estas rutas:

$$P(R \to S|P \to NP) = P(Priori(R_t)) * P(S_{t+1}|R_t) * P(P|R_t) * P(NP|S_{t+1}) \tag{3.9}$$

Sustituyendo

$$P(R \rightarrow S|P \rightarrow NP) = 0{,}33 * 0{,}4 * 0{,}8 * 0{,}6 = 0{,}06336 \tag{3.10}$$

Al obtener esta nueva probabilidad, podemos comenzar con el siguiente ciclo, sin embargo para efectos prácticos no se hará. Lo que obtendremos será las P de cada secuencia, siendo la secuencia *Soleado* $\rightarrow$ *Soleado* de los primeros dos días del ejemplo anterior ($P = 0{,}67 * 0{,}8 * 0{,}4 * 0{,}6 = 0{,}12864$) la secuencia con la P más alta. De esta forma es como conocemos la secuencia de los estados ocultos en las HMC.

3.3. Datos genéticos de poblaciones humanas europeas

3.3.1. Datos genómicos

Analizar la historia demográfica de poblaciones a partir de datos genéticos ha consistido en su mayor parte de información proveniente de polimorfismos de un solo nucleótido (*SNP* en Inglés). Desde que en el 2001 se comenzaron a identificar los sitios polimórficos, se ha comenzado la identificación de variantes de DNA (con importancia médica, en la mayoría de los casos) como principal objetivo en la genética humana (LaFramboise, 2009). A pesar del esfuerzo por representar la diversidad genética en distintas poblaciones humanas, es prácticamente casi imposible para muchas poblaciones en el globo ya que la mayoría de los recursos económicos, tecnológicos y la mano de obra calificada para realizar la secuenciación del genoma humano se encuentran en su mayoría en los países desarrollados (Miga y Wang, 2021). Estos estudios pueden partir de secuencias de genoma completo (WGS, *Whole-Genome Sequence* por sus siglas en inglés) y se realizan analizando regiones ya sea secuencias cortas ($\approx$ 50 pb., por medio de tecnologías *Illumina*) o secuencias largas ($\approx$ 10000 pb. por *Pacific Biosciences* o *Oxford Nanopore Technologies*). Sin embargo, estas últimas pueden tener un error de aproximadamente $\approx$ 1-2 % (Delahaye y Nicolas, 2021), además de que representan un costo mucho mayor en comparación al análisis de secuencias cortas (Lappalainen *et al.*, 2019) y no se usan para secuenciar genomas antiguos. Para la identificación de variantes alélicas, se secuencia varias veces una misma región (a la cantidad de veces que una misma región se secuenció le llamaremos profundidad), dando como resultado un determinado n de secuencias. Posteriormente, estas lecturas tienen que ser unidas por medio de programas de computadora para posteriormente ser comparadas entre sí contra una secuencia de referencia. En general, el análisis de secuencias de variantes raras se recomienda hacer con una profundidad >20x (Dehasque *et al.*, 2020), lo cual es poco frecuente en datos de DNA antiguo donde en ocasiones los datos llegan a tener una cobertura menor de 1x (Marciniak y Perry, 2017). En el pasado, la obtención de genomas antiguos era suma-

mente complicado. Las metodologías para la obtención de *aDNA* estaban limitadas a factores abióticos tales como la reacción de β-eliminación, la cuál rompe la hebra de DNA así como la desaminación de las citosinas, lo que provoca *artefactos* o errores en la secuenciación, observando timinas *artificiales* (al igual que sucede de G a A); otro factor que influye en la calidad baja de las secuencias antiguas es la poca disponibilidad de genoma endógeno (siendo del 1 % y rara vez del 5 %, Carpenter *et al.* 2013). Los primeros genomas que se obtuvieron iban del 0.5x al 2x de cobertura[8] (Green *et al.*, 2006, 2010; Reich *et al.*, 2010) hasta llegar a los >20x en la calidad promedio de algunos genotipos (Fu *et al.*, 2013; Mafessoni *et al.*, 2020; Meyer *et al.*, 2012; Prüfer *et al.*, 2014, 2017; Rasmussen *et al.*, 2010), lo que no sólo se traduce en una mejora significativa, sino también en una calidad mayor en las investigaciones de genética de poblaciones con datos antiguos.

3.3.2. Poblaciones europeas

Uno de los grupos de los que más datos genómicos se han obtenido son las poblaciones humanas europeas. Éste grupo, a comparación de otros grupos humanos, se considera poco diverso y con una estructura genética que corresponde con la geografía del continente (Novembre *et al.*, 2008). Los primeros datos genómicos de *Homo sapiens* en Europa provienen de individuos de hace 30000 años aproximadamente (Fu *et al.*, 2016) y se extienden hasta el presente (Fairley *et al.*, 2020).

Las poblaciones de *Homo sapiens* modernos comenzaron a ingresar en este continente hace aproximadamente 45-43000 años antes del presente (A.P.[9]), coincidiendo con la abrupta caída del registro fósil de *H. neanderthalensis* en este continente (Benazzi *et al.*, 2011). Aún se discute la posibilidad de una extinción de los neandertales por causas humanas o climáticas (Timmermann, 2020). Posteriormente, se desarrollarían en este continente dos grupos culturales humanos: el *Gravetiense* (35000 A.P.) y el *Solutrense* (24000 A.P.), los cuáles a su vez pertenecieron al grupo de los *cazadores recolectores*. Estos grupos se caracterizaban por el uso de herramientas tales como las flechas triangulares con bordes suaves y el uso de materiales como huesos y maderas. Lazaridis *et al.* (2014) encontró que este grupo se separó hace 24000 A.P. de los *Antiguos Euroasiáticos del Norte* (*Ancient North Eurasian*, *ANE*), dando origen a los *Cazadores Recolectores Occidentales* (*Western Hunter-Gatherers*, *WHG*) que representa a uno de los componentes de ancestría mayoritarios en europeos modernos. Hace 20000 años A.P. los grupos culturales permanecieron inalterables, sin embargo, hace 15-10000 A.P., comenzó un cambio en el desarrollo tecnológico de las culturas establecidas en Europa, lo que dio

[8] cada x a lado de un valor numérico representa la cantidad de veces que se secuenció un genoma

[9] Antes del Presente, según la traducción literal de *Before Present* de acuerdo a la Comisión Internacional de Estratigrafía y que comienza a partir del año 1950 hacia atrás

como resultado el surgimiento de la agricultura y por lo tanto el inicio del Mesolítico. Posteriormente, hace 10000-5000 años, grupos de la estepa Póntica, ubicado en Rusia, Crimea y Ucrania, conocido como los Yamnaya ingresaron a Europa, contribuyendo en gran parte al remplazamiento genético de los *WHG*, y convirtiéndose así en el segundo componente europeo más común (Mathieson *et al.*, 2015). Durante el Neolítico Europeo (en algunas partes de Europa abarca de los 7000 a los 3000 en el sur, mientras que en el norte abarca sólo de los 6000 a los 3000 A.P.) las poblaciones agricultoras comenzaron un remplazamiento poblacional sobre los *WHG*. Este grupo comenzaría a extenderse desde lo que hoy es el mar Egeo, Grecia, hacia el Occidente Europeo, siendo conocidos como los *Primeros Europeos Agricultores* (*Early European Farmers*, *EEF*). A su vez, se tiene identificado un grupo Indo-Europeo que comenzó su ingreso a este continenten hace aproximadamente 7000 A.P. (Lazaridis *et al.*, 2014). Para el último periodo de la prehistoria europea, la Edad de los Metales, un grupo proveniente de Siberia se introdujo como el tercer mayor componente de las poblaciones europeas del presente. Este grupo, *ANE* está relacionado fuertemente con la cultura *BellBeaker* (4000 A.P.), otro grupo de las estepas, el cuál tiene yacimientos principales en Francia, Alemania y España (Lazaridis *et al.*, 2014; Olalde *et al.*, 2018). Los movimientos migratorios de la edad de Bronce (4250-3900 A.P.), Hierro (2800-2400-A.P.)y en los tiempos modernos ayudaron a distribuir aún más la ancestría de las culturas de la estepa en las poblaciones europeas modernas de occidente. En la actualidad hay una estructura poblacional muy bien diferenciada entre las poblaciones del norte y sur del continente, en las costas del mar Mediterráneo, siendo estos últimos los cuentan con un mayor componente de poblaciones Neolíticas de Anatolia (Skoglund y Mathieson, 2018), pero encontrándose a nivel continental una estructura genética correspondiente con la geografía (Novembre *et al.*, 2008).

Por todo lo anterior, es importante recalcar que a pesar de contar con las herramientas suficientes para obtener datos genómicos a través de distintos periodos y detectar alelos bajo selección, no se ha intentado antes el análisis de la distribución de los alelos neutros, deletéreos y ventajosos en un espacio geográfico. Si bien podemos conocer cómo se comportan los alelos a lo largo del tiempo, un análisis de distribución espacial nos puede ayudar a entender cómo la selección natural moldea la diversidad de las poblaciones y cómo estas cambian su composición genética a lo largo del tiempo. En esta tesis desarrollamos un nuevo modelo estadístico para estudiar la evolución espacio-temporal de alelos neutrales, deletéreos y ventajosos permitiéndonos obtener un primer acercamiento acerca de la distribución de los alelos a lo largo de un espacio en dos dimensiones en función del tipo de selección que actúa sobre el alelo.

4

Objetivos

- *Objetivo general*:

  Conocer el impacto de la selección natural en la diversidad genética de las poblaciones humanas de Europa a lo largo de 30000 años hasta el presente.

- *Objetivos particulares:*

  - Identificar las frecuencias de los alelos tentativamente neutrales, deletéreos y ventajosos en poblaciones europeas de las cuales existan datos genómicos.
  - Establecer un modelo demográfico que explique la distribución de alelos neutrales y su coeficiente de dispersión en un espacio.
  - Determinar la fuerza de selección que actúa en las variantes deletéreas y ventajosas.

5

Antecedentes

5.1. Inferencia de selección natural a través de cadenas de Markov

Existen distintas metodologías para inferir el impacto de la selección natural mediante distintos estadísticos que toman en cuenta los cambios en frecuencias alélicas a lo largo del tiempo Bollback *et al.* (2008), Malaspinas *et al.* (2012), Mathieson y McVean (2013), Foll *et al.* (2014), Mathieson *et al.* (2015), Marciniak y Perry (2017) y Muktupavela *et al.* (2021). Los métodos estadísticos que infieren la fuerza de la selección natural han sido utilizados para analizar variantes genéticas de distintas especies como caballos, virus y humanos. Podemos analizar las frecuencias alélicas mediante un Modelo Oculto de Markov (Ewens, 2004) en donde la frecuencia alélica es el estado observado y las frecuencias poblacionales son los estados ocultos. Adicionalmente, debemos contar con las probabilidades de cambios de estado como se hizo mención en el apartado *Métodos Markovianos*. Es posible usar las *Ecuaciones de Kolmogorov*, también conocidas como *Ecuaciones hacia adelante* y *Ecuaciones hacia atrás* las cuáles calculan la probabilidad de estar en un estado particular en un tiempo continuo. Dicho en otras palabras, contamos con el estado x_t (es decir, con la distribución de probabilidad $p_t(x)$) y del que queremos conocer su distribución de probabilidad en un tiempo $t+1$ ($p_{t+1}(x)$) por lo que al final tendríamos un sistema de ecuaciones en derivadas parciales para resolver este problema. Para explicar este concepto, utilizaremos la siguiente fórmula:

$$-\frac{\partial}{\partial t}p(x,t) = \mu(x,t)\frac{\partial}{\partial x}p(x,t) + \frac{1}{2}\sigma^2(x,t)\frac{\partial^2}{\partial x^2}p(x,t) \tag{5.1}$$

En donde Bollback *et al.* (2008) sustituyeron los valores de la siguiente forma:

$$\frac{\partial}{\partial t}f(x;p,t) = a(p)\frac{\partial}{\partial p}f(x;p,t) + \frac{1}{2}b(p)\frac{\partial}{\partial p^2}f(x;p,t) \tag{5.2}$$

Aquí, $f(x;p,t) = p(X(t) = x|X(0) = p)$, que se lee como la probabilidad de que la frecuencia del alelo en el tiempo t sea igual a x dado que la frecuencia del alelo es igual a p en el tiempo $t = 0$. En otras palabras se refiere a la frecuencia t veces (unidades de tiempo) después de obtener la frecuencia p en un tiempo inicial; por otro lado, se define a $a(p) = sN_ep(1-p)$ siendo s el coeficiente de selección, N_e el tamaño efectivo de la población y $(1-p)$ la frecuencia del alelo ancestral q, finalmente, el término $b(p)$ se define como $b(p) = p(1-p)$. Por otro lado, es común en las Cadenas de Markov encontrarse con estados de los que es prácticamente imposible salir de ahí. En términos de cálculo de frecuencias, por ejemplo, la única forma de que un estado (el valor de una frecuencia) ya no cambie es cuando un alelo se fije, por lo tanto, decimos que se encuentra en un estado *absorbido*. La probabilidad de que esto ocurra en el alelo p es de:

$$P_1(p) = \lim_{x\to\infty} Pr(X(t) = 1|X(0) = p) = \frac{1 - e^{-1N_esp}}{1 - e^{-2N_es}} \tag{5.3}$$

Que se puede entender como la probabilidad de que un estado en un momento $X(t)$ sea igual a 1 dado que en el estado inicial se tiene una frecuencia p es igual a la frecuencia de p entre todos los alelos[1], por lo que de ser igual a 1, el estado absorbente se alcanza. Para finalizar, el cálculo de las probabilidades de emisión (las frecuencias observadas) pueden resolverse como como un problema de distribución binomial:

$$Pr(Y(t) = y(t)|X(t) = x(t)) = \binom{n}{y(t)} x(t)^{y(t)}(1 - x(t))^{n-y(t)} \tag{5.4}$$

Dado que estamos intentando calcular la probabilidad de que tengamos una frecuencia observada *y(t)* en un tiempo t en un conjunto de n de alelos. Por supuesto, esto tendría que calcularse para todas las posibles rutas del alelo, puesto que para un mismo tiempo t estamos calculando distintos valores de *y(t)*, como se vio en el ejemplo anterior (ver sección 1.2.2; Bollback *et al.*, 2008; Muktupavela *et al.*, 2021; las ecuaciones 3.1-3.4 forman parte del modelo propuesto por Bollback para inferir el impacto de la selección natural mediante un algoritmo de máxima verosimilitud).

[1]Nótese que el denominador y el numerador hacen referencia a la tasa de crecimiento, mismo que hemos utilizado para el cálculo de tasa de crecimiento en bacterias.

5.2. Clasificación de alelos deletéreos, neutros y ventajosos

Para explorar los alelos en función de su impacto en la adecuación, se ha propuesto el uso de *CADD* (del inglés *Combined Annotation Dependent Depletion*, Anotación Combinada Dependiente de Agotamiento, ver sección 6.3; Rentzsch *et al.*, 2021), cuya metodología es capaz de detectar los alelos que son neutros y deletéros otorgándoles un puntaje de 0 a 100, siendo 0-5 tentativamente neutro y 6-100 tentativamente deletéreos. La inferencia de la fuerza de la selección natural de los alelos en el genoma humano ha sido utilizado en su mayoría por estudios de asociación médicos para priorizar el análisis de variantes que pudieran ser patógenas dado su puntaje de CADD (Chen *et al.*, 2021), en especial y de muy alta relevancia aquellos estudios relacionados con una sintomatología grave en pacientes con SARS-CoV-2 (Huffman *et al.*, 2022). Esta metodología no es la mejor para predecir efectos patogénicos en ciertos sitios del genoma (Rowlands *et al.*, 2021). Por ejemplo, hay herramientas que predicen mejor el efecto patogénico de una variante en mutaciones que afectan el empalme[2] alternativo como TrAP(Gelfman *et al.*, 2017). No obstante, TrAP es una herramienta capaz de evaluar la patogenicidad de todos los alelos del genoma y es hasta ahora la única herramienta que predice alelos neutros (Kircher *et al.*, 2014). Además, se ha comprobado la correlación positiva que guarda el puntaje de *CADD* con el coeficiente de selección (Racimo y Schraiber, 2014). Por otra parte, las metodologías para la detección de alelos ventajosos emplean otro tipo de metodologías que detectan alelos que han aumentado muy rápido en frecuencia (Mathieson, 2020). Mediante el uso de datos de DNA antiguo, Mathieson *et al.* (2015) compiló la lista más actual de alelos bajo selección positiva en Europa usando datos de DNA.

ID	Cromosoma	Posición	Genes afectados	Función/ fenotipo relacionado
rs4988235	2	136,608,646	MCM6,LCT	Tolerancia a la lactosa
rs16891982	5	33,951,693	SLC45A2	Pigmentación
rs2269424	6	32,132,233	Región MHC	Inmunidad
rs174546	11	61,569,830	FADS1, FADS2	Metabolismo de ácidos grasos
rs4833103	4	38,815,502	TLR1, TLR6, TLR10	Inmunidad
rs653178	12	112,007,756	ATXN2, SH2B3	Desconocido
rs7944926	11	71,165,625	DHCR7, NADSYN1	Metabolismo de Vitamina D
rs7119749	11	88,515,022	GRM5	Pigmentación
rs272872	5	131,675,864	SLC22A4	Transporte de ergotionina
rs6903823	6	28,322,296	ZKSCAN3, ZSCAN31	Autofagia, función pulmonar
rs1979866	13	38,825,900	-	Desconocido
rs12913832	11	28,365,618	GEC2, 0CA3	Color de Ojos

Cuadro 5.1: Lista de SNP's con alelos tentativamente ventajosos, obtenido de Mathieson *et al.* (2015).

[2]Splicing

5.3. Análisis de la distribución espacio temporal de frecuencias alélicas

El análisis de alelos provenientes de distintos periodos y su relación geográfica es un tópico de alto interés reciente (Bradburd y Ralph, 2019). Por ejemplo, Racimo *et al.* (2020) utilizaron datos genómicos, detectaron la ancestría de algunos alelos y graficaron la frecuencia de estos alelos en distintos puntos de Europa para después obtener su distribución espacial a lo largo del continente, lo que permitió observar cuánto ha cambiado a lo largo del tiempo la composición genética de los grupos europeos. Anteriormente, Mathieson y McVean (2013) habían propuesto un modelo para simular la dispersión de un caracter en un espacio de 2D para posteriormente llevarlo a cabo en un espacio geográfico y en función de éste calcular el coeficiente de selección. Por otra parte, Muktupavela *et al.* (2021) obtuvo los valores de máxima verosimilitud de alelos ventajosos y calculó su coeficiente de verosimilitud, pudiendo obtener la distribución gráfica desde el punto más antiguo en el que surgió un alelo ventajoso y su distribución en poblaciones europeas de un periodo de 10000 años. Sin embargo, aún no ha sido estudiado a detalle la distribución espacio temporal de alelos deletéreos a pesar de que se sabe que la selección negativa tiene un impacto en la distribución geográfica de un alelo (Locke *et al.*, 2019).

6

Métodos

6.1. Datos genómicos

Los datos genómicos de poblaciones de Europa provienen de la base de datos *Allen Ancient DNA Resource* (Mallick *et al.*, 2023). Estos datos compilan aproximadamente 10379 individuos de todo el mundo, de los cuales se analizaron únicamente 3672 individuos de Europa de 40 países distintos y otros 8 de Euroasia. Los datos genéticos se componen del estudio de 1233013 sitios del genoma nuclear por persona[1]. Estos *chips* fueron previamente diseñados para evaluar los sitios más variables compartidos en todas las poblaciones humanas (Lu *et al.*, S.F.), permitiendo observar las mutaciones compartidas entre todas las poblaciones europeas independientemente de su ancestría. La edad de los individuos varía desde los 30000 años hasta el presente.

6.2. Análisis de los datos

Para obtener las frecuencias alélicas de los sitios, se utilizó el programa PLINK (v. 1.9), que puede calcular la frecuencia de cada alelo en cada posición y población. Por lo tanto, tendremos una estructura similar a la siguiente en un archivo *.frq*:

	CHR	SNP	A1	A2	MAF	NCHROBS
POB 1	1	rs13618	A	T	0.50	2
POB 2	1	rs13618	A	T	0.25	4

Donde la primera columna corresponde a la nomenclatura de una población, *CHR* es el cromosoma al que pertenece el sitio, el *SNP* es el ID único en el que se puede encontrar en la base de *NCBI*, *A1* es el alelo menor o el segundo alelo más común, *A2* es el alelo

[1]Este es el número de sitios máximos analizados. No siempre se cubren todos los sitios, por lo que se considera similar al sitio del genoma de referencia.

mayor o el más común, *MAF* es la frecuencia del alelo menor y *NCHROBS* es el número de cromosomas en el que se observa este SNP en una población. En el ejemplo anterior, podemos observar cuál es la frecuencia del SNP $rs13618$ en las poblaciones 1 y 2 además del número de cromosomas observados en cada población. Como los seres humanos somos diploides, estaríamos observando por lo tanto un individuo que representa a la población 1 y de dos individuos para el caso de la población 2. Estos datos están asociados a un archivo *.anno*[2] el cuál posee los datos de los individuos asociados a este análisis tales como la edad (de ser posible obtener), la ubicación exacta en coordenadas, datación, caso asociado, artículo publicado, autores, etc. Las frecuencias alélicas se calcularon para cada subgrupo cultural definido por un método de subagrupación. Para ello comenzamos usando la agrupación cultural definida en la base de datos del *Allen Resource*, posteriormente se realizaron una serie de subdivisiones dentro del mismo grupo cultural según algunos criterios que permiten excluir o incluir a los individuos. Esto debido a que a pesar de que varios individuos formaban parte de una misma cultura, en algunos casos la diferencia de edad media era de hasta por varios miles de años o sus restos fueron hallados a cientos de kilómetros de distancia. Esta subagrupación fue realizada mediante *k-medias* en *Python 3.8.* Este análisis nos permitió agrupar a los individuos dentro de un grupo cultural en otro subgrupo en donde cada individuo no tiene una distancia mayor a 500 km y 500 años. Esto nos ayudó a tener una definición más cercana en tiempo y espacio de los individuos analizados en cada uno de los subgrupos.

Subdivisión poblacional por medio del algoritmo de *k-medias*

Un análisis de *k-medias* se basa en el principio de agrupación por distancias, el cual comienza 1) proyectando uno o más puntos (dependiendo del número de k o grupos que se quieran formar) entre los datos previamente graficados. Posteriormente, 2) se comienza a asignar el primer *cluster* hacia los puntos más cercanos, 3) se calculará la media de distribución para colocar un nuevo punto y a partir de ese nuevo punto se volverá a asignar de nuevo un *cluster* tomando otra vez los mismos o nuevos puntos con respecto a la distancia de la nueva media calculada (véase figura 6.1.). El paso 2 y 3 se repetirán varias veces (iteraciones) hasta que el proceso se detenga o según un cierto número de repeticiones. En nuestro caso, utilizamos de 2 a 10 subgrupos dentro de cada grupo cultural, permitiendo que cada individuo dentro de su propio subgrupo no estuviera a una distancia mayor a 500 km y a 500 años. El mismo fue realizado en *Python* por medio del paquete *sklearn.cluster* (Pedregosa *et al.*, 2011) utilizando un máximo de 100 iteraciones.

[2]Este archivo está disponible para descargar en la página del laboratorio de David Reich

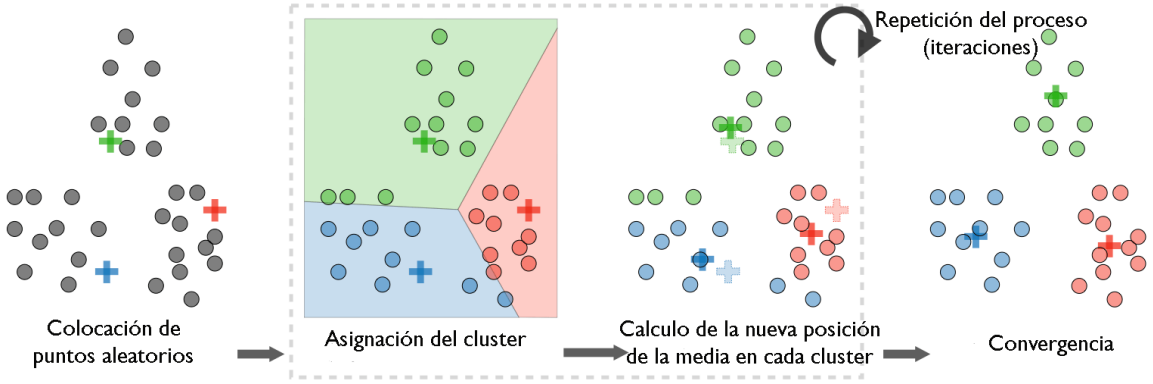
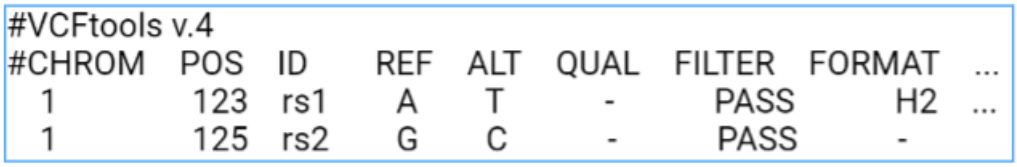


Figura 6.1: Proceso de formación de *cluster*s en *k-medias*. Entre lineas punteadas se puede observar los pasos que involucra la iteración del proceso para construir los *clusters*. Modificado y obtenido de *Jusliu.com*

6.3. Detección de alelos tentativamente deletéreos, neutros y ventajosos

Para la detección de alelos ventajosos se utilizaron los datos de 10 sitios de los 12 originales que Mathieson *et al.* (2015) previamente identificaron en poblaciones de Europa. Se descartaron 2 debido a que estos no se encuentran estudiados en la base de datos consultada (rs4833103 relacionado al sistema inmune y rs7119749 relacionado a la pigmentación). Mientras que para la identificación de alelos tentativamente deletéreos y neutros, se optó por utilizar el programa *CADD* (siglas de *Combined Annotation Dependent Depletion*). *CADD* funciona otorgando a cada mutación un puntaje el cual indica qué tan deletéreo o neutro es, yendo desde 0 a 100 donde un valor de 0 a 5 es prácticamente neutro y >6 es deletéreo. Para ello se transformaron los 3672 datos de los individuos a formato VCF por medio del programa *VCFtools* (Danecek *et al.*, 2011). Para entender el formato *VCF* (de *Variant Call Format*) hay que recordar que las primeras cinco columnas de cada formato arrojan información del alelo tal como la ubicación en el cromosoma, la posición de la base nucleotídica, el ID de ese alelo (en *reference sequence* o *rs*, Figura 6.2.) así como el nucleótido de referencia y el nucleótido alternativo. El resto de las columnas contienen la información de los individuos que presentan este alelo. Al final, se obtuvieron 23 archivos VCF, uno por cada cromosoma, que posteriormente fueron subidos al sitio web de *CADD* para obtener su puntaje. La versión del genoma de referencia utilizado fue la GRCh38-v.1.6. Posteriormente, se obtuvieron un total de 22 archivos excluyendo al cromosoma del par sexual 23 ya que *CADD* aún no contiene información sobre estos cromosomas.



#VCFtools v.4

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	FORMAT	...
1	123	rs1	A	T	-	PASS	H2	...
1	125	rs2	G	C	-	PASS	-	

Figura 6.2: Estructura clásica de un archivo VCF. Los archivos de este tipo tienen entre comentarios (#) descrito el tipo de archivo con el que se está trabajando.

6.4. Modelo de inferencia de selección y geografía

Consideramos a $T = t_1, t_2...t_n$ al conjunto de edades en la que cada alelo fue datado en una población. A $\Gamma = \gamma_1, \gamma_2...\gamma_n$ las latitudes promedio del alelo y a $H = \eta_1, \eta_2...\eta_n$ las longitudes promedio. Para los loci consideramos que hay S sitios muestreados (1233013) en todas las poblaciones. El número de alelos derivados en cada sitio en una población sería:

$$D = (d_{1,1}, d_{1,2}, d_{1,3}...d_{1,s}), (d_{2,1}, d_{2,2}, d_{2,3}...d_{2,s}), (d_{3,1}, d_{3,2}, d_{3,3}...d_{3,s})...(d_{n,1}, d_{n,2}, d_{n,3}...d_{n,s}) \quad (6.1)$$

en donde $d_{i,j}$ indica el número de alelos derivados en la población i de la posición j, y n es el número de poblaciones analizado, mientras que el número de copias de cromosomas observados en cada población es igual a:

$$C = (c_{1,1}, c_{1,2}, c_{1,3}...c_{1,s}), (c_{2,1}, c_{2,2}, c_{2,3}...c_{2,s}), (c_{3,1}, c_{3,2}, c_{3,3}...c_{3,s})...(c_{n,1}, c_{n,2}, c_{n,3}...c_{n,s}) \quad (6.2)$$

Dado que anteriormente pudimos estratificar nuestros alelos por puntaje *CADD*, es posible estratificar los loci de acuerdo a su puntaje. Recordando que estos valores van de 0 a 100, donde si $0<CADD<5$ es neutral (ι), mientras que si $CADD>5$ (ρ) se considera tentativamente deletéreo. Por lo tanto, podemos definir el número de alelos neutros de una población D_ι de la siguiente forma:

$$D_\iota = (d_{\iota,1,1}, d_{\iota,1,2}, d_{\iota,1,3}...d_{\iota,1,s}), (d_{\iota,2,1}, d_{\iota,2,2}, d_{\iota,2,3}...d_{\iota,2,s}), (d_{\iota,3,1}, d_{\iota,3,2}, d_{\iota,3,3}...d_{\iota,3,s})... \\ (d_{\iota,n,1}, d_{\iota,n,2}, d_{\iota,n,3}...d_{\iota,n,s}) \quad (6.3)$$

Lo cual nos permite realizar la misma forma para C_ι, obteniendo el número de copias de cromosomas en una población con este valor:

$$C_\iota = (c_{\iota,1,1}, c_{\iota,1,2}, c_{\iota,1,3}...c_{\iota,1,s}), (c_{\iota,2,1}, c_{\iota,2,2}, c_{\iota,2,3}...c_{\iota,2,s}), (c_{\iota,3,1}, c_{\iota,3,2}, c_{\iota,3,3}...c_{\iota,3,s})... \\ (c_{\iota,n,1}, c_{\iota,n,2}, c_{\iota,n,3}...c_{\iota,n,s}) \quad (6.4)$$

Mientras tanto, dado que tenemos distintos valores de ρ, es posible estratificarlos, como

$$D_{6-10}, D_{11-15}, D_{16-20}, D_{21-25}...D_{i-j} \tag{6.5}$$

De igual manera, el número de copias de cromosomas con diferentes valores de *CADD* es:

$$C_{6-10}, C_{11-15}, C_{16-20}, C_{21-25}...C_{i-j}. \tag{6.6}$$

Para finalizar, nuestro modelo nulo debe explicar la dispersión de los alelos a través de un espacio. Por lo tanto, nuestro modelo calculará las frecuencias de p_ι en tiempo y espacio a través de un modelo de ecuaciones parciales diferenciales de dos dimensiones (EPD) (Bollback *et al.*, 2008; Muktupavela *et al.*, 2021). Este modelo nos permite representar el cambio en frecuencia x en un modelo de dos dimensiones (x, y) en un periodo t[3]:

$$\frac{\partial p}{\partial t} f(x; p, t) = \frac{1}{2}\sigma^2 \frac{\partial^2 p}{\partial x^2} + \frac{1}{2}\sigma^2 \frac{\partial^2 p}{\partial y^2} + \gamma(p, s, d) \tag{6.7}$$

donde utilizaremos la fórmula del cambio de frecuencias alélicas posterior a la selección (Ewens, 2004; Hamilton, 2009)[4].

$$\gamma(p, s, d) = p(1 - p)(pd + s(1 - 2p)). \tag{6.8}$$

Siendo σ el coeficiente de dispersión, s el coeficiente de selección y $d = 2s$, siendo este último valor parte de un modelo aditivo y fijo que nos permitirá conocer qué tanto se dispersan los alelos bajo selección (Henn *et al.*, 2016). En este trabajo propongo una metodología que consiste en: 1) estimar el coeficiente de dispersión de los alelos tentativamente neutrales (aquellos que poseen un puntaje *CADD* menor a 5) y 2) estimar el coeficiente de selección en alelos tentativamente deletéreos/ventajosos. Para calcular el valor de dispersión, se estimó el valor de máxima verosimilitud del parámetro de dispersión. Calcularemos el valor de la función de verosimilitud para los parámetros $\theta_1 = s$ donde los valores evaluados son $s = 0, -0{,}1, -0{,}2, -0{,}3, -0{,}4, -0{,}5$ para alelos deletéreos (tomando el coeficiente de dispersión estimado en el paso 1) y $\theta_2 = \Phi$ donde $\Phi = 2, 4, 6, 8, 10$ siendo Φ el coeficiente de dispersión:

$$LL(\theta|G) = \prod_{i=1}^{g} P(d_i, c_i|\theta, p(x_i, y_i, t_i)) \tag{6.9}$$

Donde G representa la información genética de todos los sitios analizados, d_i es el número de alelos derivados en una población i dado un número c_i de alelos muestreados en una población.

[3]Por tiempo nos referimos a generación. En este trabajo, cada generación equivale a 25 años.

[4]Consultar el capítulo 1 de Ewens (2004) *"The Deterministic Theory"* para el origen de esta fórmula y *Further models of natural selection* de Hamilton (2009) para observar el desarrollo de ella a partir de *W*.

Además, consideraremos que $p(x_i, y_i, t_i)$ es la frecuencia poblacional de ese alelo en una ubicación x_i, y_i y tiempo dados t_i. Los valores de máxima verosimilitud pueden expresarse en términos de una ecuación de probabilidad binomial (Hamilton, 2009; Muktupavela *et al.*, 2021), tendremos entonces que:

$$L(\sigma, \Phi; d_i, c_i) = P(d_i, c_i | p(x, y, t)) = \binom{c_i}{d_i} p(x_i, y_i, t_i)^{d_i} (1 - p(x_i, y_i, t_i))^{c_i - d_i} \tag{6.10}$$

En donde $p(x_i, y_i, t_i)$, se obtiene de la resolución de la ecuación diferencial de dos dimensiones con parámetros σ y Φ. Posteriormente podemos analizar los valores de máxima verosimilitud de la siguiente forma:

$$LL(\sigma, \Phi) = \sum_{i=1}^{g} L(\sigma, \Phi; d_i, c_i) \tag{6.11}$$

La resolución de la ecuación diferencial de dos dimensiones se realizó con los paquetes de R(v 4.2.1) *deSolve* para el sistema de ecuaciones diferenciales parciales y *rootSolve* para la resolución de ecuaciones no lineales. La resolución de la ecuación diferencial nos permite calcular el valor de $p(x_i, y_i, t_i)$ dado un coeficiente de dispersión y un estado inicial. En ambos casos (coeficiente de selección y dispersión) el estado inicial del sistema en donde se resuelve la ecuación diferencial asume que el alelo está presente únicamente en la población i más antigua en donde la frecuencia alélica es mayor a cero y asumimos que la frecuencia alélica en dicha población es igual a d_i/c_i.

Para el cálculo del valor de máxima verosimilitud del coeficiente de dispersión se asume que las poblaciones están dentro de una gradilla de un tamaño de $100x100$, espacio en el que se analizó el $\Phi = 2, 4, 6, 8, 10$ y se estimó la posible distribución de cada alelo según el coeficiente de distribución. Estas gradillas se calcularon normalizando la posición geográfica de cada población, lo que hacía que cada gradilla fuera el equivalente a un grado de latitud y 0.4 grados longitud[5]. El espacio geográfico utilizado abarca desde latitud -25° a 75° y longitud 35° a 75°. Adicionalmente, estos cálculos se realizaron tomando en cuenta alelos que sólo aparecieron en los datos en poblaciones que surgieron hace menos de 10000 años. Para calcular el coeficiente de selección se utilizaron a todas las poblaciones independientemente de su datación.

Cuanto analizamos la distribución de los alelos en ventanas de tiempo, dividimos a las subpoblaciones en 6 grandes grupos por edad tomando en cuenta la media de datación de cada subagrupación: A=>10000, B=10000-7500, C=7500-5000, D=5000-2500, E=2500-150 Y F=poblaciones del presente. Esta distribución nos permitió analizar por ventanas de tiempo lo suficientemente grandes cómo los alelos han cambiado sus frecuencias alélicas a lo largo del tiempo (Aris-Brosou, 2019) y se representaron en un gráfico de cajas y bigotes de acuerdo a si son alelos neutros, deletéreos y ventajosos.

[5]Una celda mide a 35° longitud aproximadamente 91x44 km y a 75°latitud aproximadamente 28x44 km

7

Resultados

7.1. Los datos genómicos de humanos de Europa comprenden una amplia distribución geográfica y temporal.

La cantidad total de individuos total es de $n_{Tot} = 3672$ (véase figura 7.1). Los individuos de las poblaciones humanas europeas provienen aproximadamente 40 países de Europa y 8 Euroasiáticos (véase 7.2). . La cantidad de individuos antiguos es de $n_{Ant} = 3458$, mientras que la de individuos modernos es de $n_{Mod} = 214$. La cantidad de poblaciones modernas viene de 19 países mientras que las poblaciones antiguas que se tienen cubiertas vienen de 46 países. Destacan España con $n_{Esp,Ant} = 376$ y Alemania con $n_{Ale,Ant} = 356$; Francia con $n_{Fra,Mod} = 57$ y Rusia $n_{Rus,Mod} = 54$ individuos (figura 7.2). Tras la subagrupación cultural hecha por *k-medias* de cada uno de los sitios, en total se contabilizaron 697 subgrupos. Los grupos culturales que se formaron por país pueden ser vistos en la figura 7.4, además de la visualización de la ubicación promedio de cada subagrupación cultural en la figura 7.5. Destacan Rusia con $n_j = 71$ subgrupos, Italia con $n_j = 65$ y España con $n_j = 53$ subgrupos. Por otro lado, la cantidad de subgrupos por periodos promedio son $t_A = 51$, $t_B = 63$, $t_C = 171$, $t_D = 265$, $t_E = 111$ y $t_F = 36$ (véase figura 7.6)

7.2. La mayor parte de los alelos investigados tienden a ser neutros o ligeramente deletéreos.

En este trabajo identificamos el puntaje *CADD* de $n_i = 1150638$ alelos del 1233013 obtenidos originalmente. Los puntajes obtenidos abarcaron el rango de 0 a 50, identificando un total de $n_{CADD<5} = 848513$ alelos tentativamente neutros. Además obtuvimos un total de $n_{CADD>5} =$

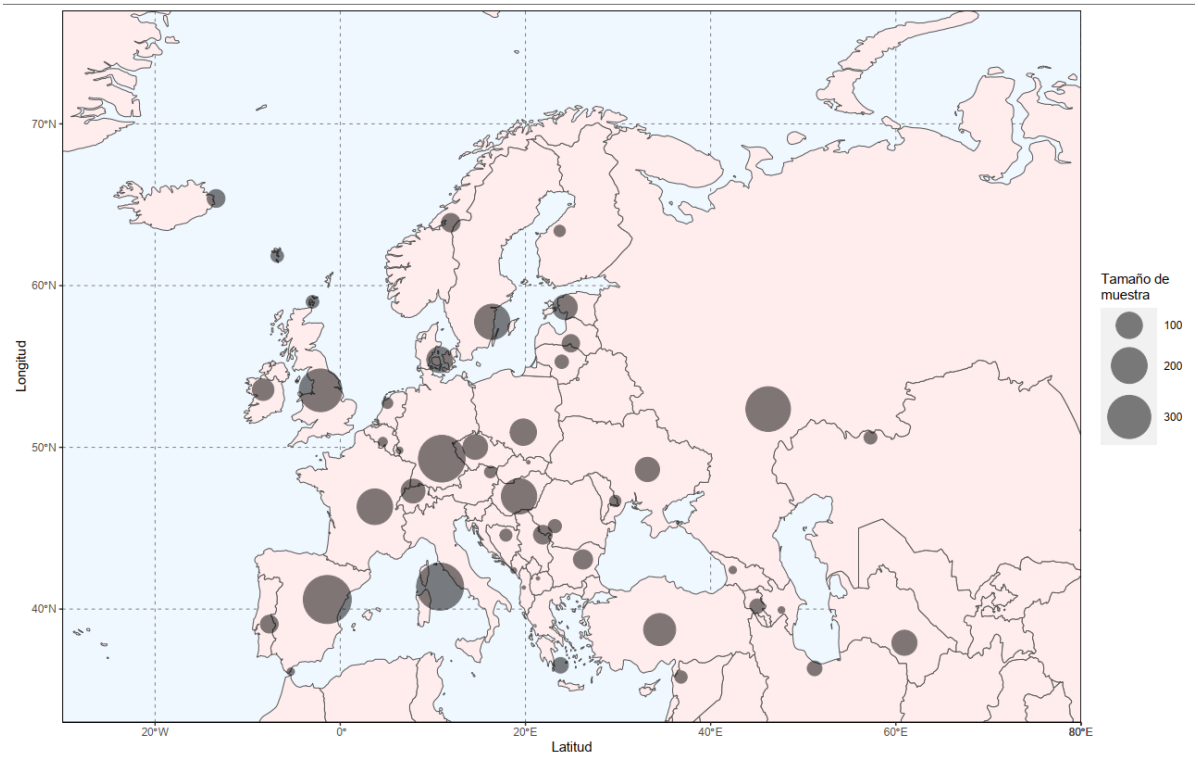


Figura 7.1: Número de individuos por país trabajados en este estudio.

302125 de alelos tentativamente deletéreos. En este trabajo descartamos 82375 alelos para evaluar por *CADD* ya que este no evalúa el grado de patogenicidad de alelos localizados en el cromosoma sexual X, por lo que este par quedó descartado de nuestros análisis (Figura 7.7a, 7.7b y tabla 7.1). Otro de los aspectos a mencionar es que la cantidad de alelos deletéreos $C_{CADD>5}$ iban disminuyendo conforme aumentaba su puntaje, muy acorde a lo que se observa en las poblaciones en donde los alelos segregantes bajo selección (ventajosos o desventajosos) son muy escasos. Además, es posible que observemos menos alelos con valores del puntaje *CADD* más altos ya que hay una correlación positiva entre el impacto de la selección natural y el puntaje de *CADD* (Racimo y Schraiber, 2014), y esta selección natural actúa contra los alelos para que no alcancen frecuencias muy altas en la población que les permitan ser parte de las variantes usadas en los datos analizados. Por otra parte, se identificaron los alelos ancestrales de 366408 sitios. Esto es, la base nitrogenada hipotética del ancestro en común de todos los individuos muestreados. De los anteriores, el número de alelos ancestrales identificados tentativamente neutros es de $i_{CADD<5} = 298373$. En el mismo sentido, $i_{CADD>5} = 81696$. Como se observará, la suma de ambos da un número mayor a la cantidad de alelos ancestrales ya que se identificaron $13,661$ posiciones con más de un alelo alternativo.

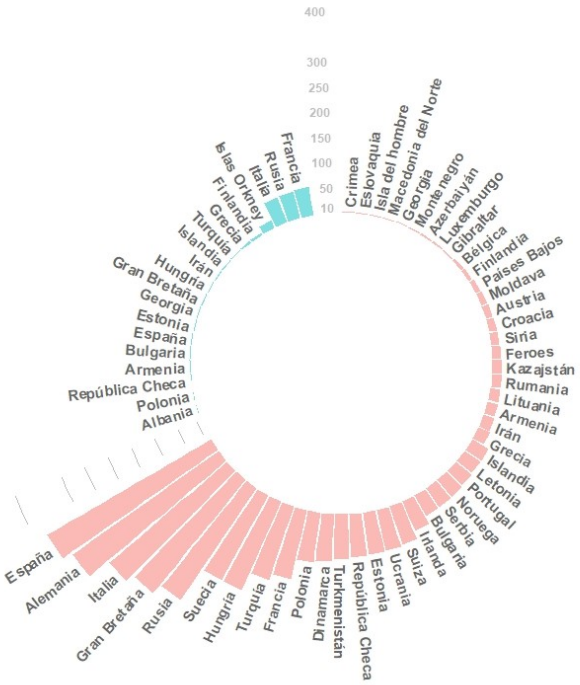


Figura 7.2: Número de individuos modernos (verde) y antiguos (rosa), usados en este proyecto

7.3. Los alelos neutros se dispersan a una gran velocidad a lo largo de Europa.

Los valores de dispersión de un alelo suponen la distribución hipotética de un alelo en la siguiente generación a una distancia de una celda, siendo una celda un grado por latitud y 0.40 grados de longitud. Estos valores nos permiten comparar que tan rápido o lento se dispersa un alelo. En este caso, los alelos evaluados son los alelos neutros (valores de CADD<5), los cuáles fungirán como el *modelo nulo*. Los valores de distribución evaluados ($\sigma = 2, 4, 6, 8, 10$) sugieren que es probable que los alelos neutros se hayan expandido a lo largo del continente rápidamente. La figura 7.8 nos muestra que el valor de máxima verosimilitud de dispersión evaluado corresponde a 10, que indica que los alelos neutros se mueven a una tasa promedio de 10 celdas por unidad de tiempo. Usaremos este valor de dispersión como referencia para futuros análisis para evaluar el movimiento de cambio en frecuencia de alelos deletéreos y ventajosos. Como se observará, no se graficaron aún más valores para alcanzar un valor máximo debido principalmente a limitaciones en nuestra capacidad computacional, sin embargo, no se descarta continuar con más pruebas con más valores. Además, este análisis no considera barreras geográficas (como montañas o ríos), esperando contar en un futuro con las herramientas para

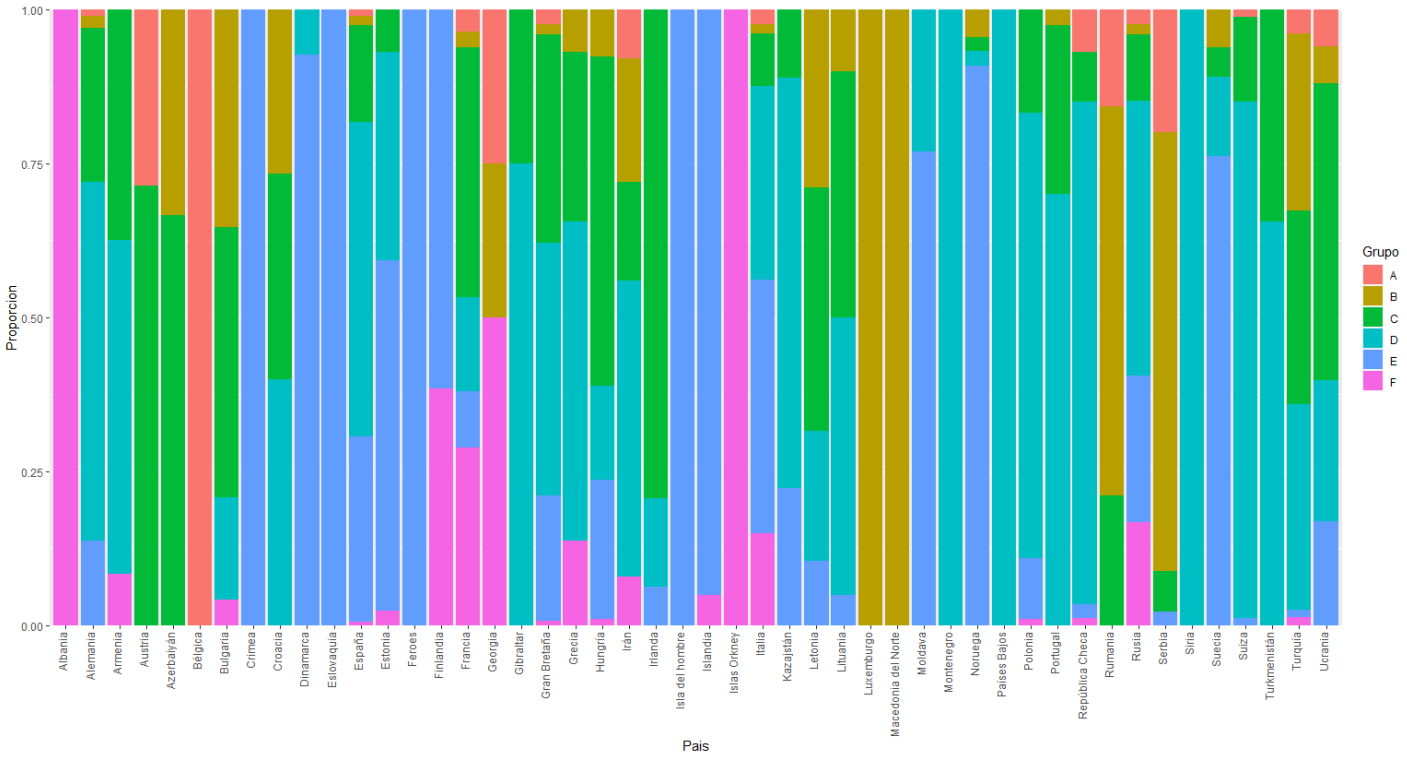


Figura 7.3: Proporción de individuos por país según la datación media estimada. A=>10000, B=10000-7500, C=7500-5000, D=5000-2500, E=2500-150 Y F=poblaciones del presente

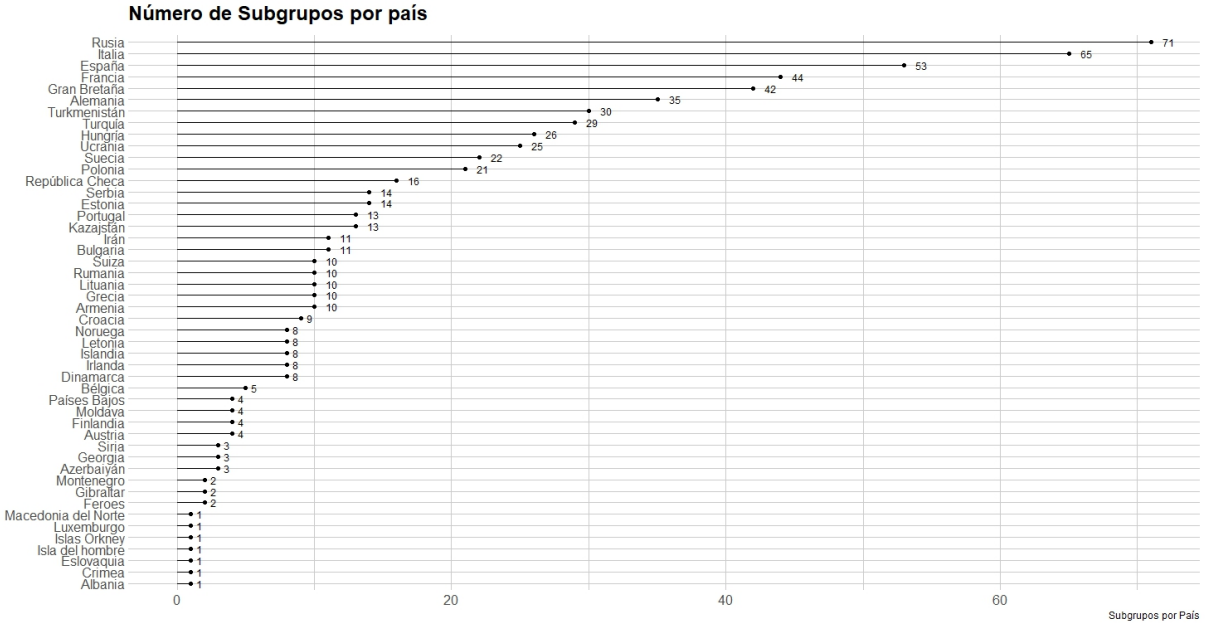


Figura 7.4: Gráfico de barras de la cantidad de subgrupos que se formaron por medio de *k-medias* por país. En este gráfico se incluyen individuos antiguos y modernos.

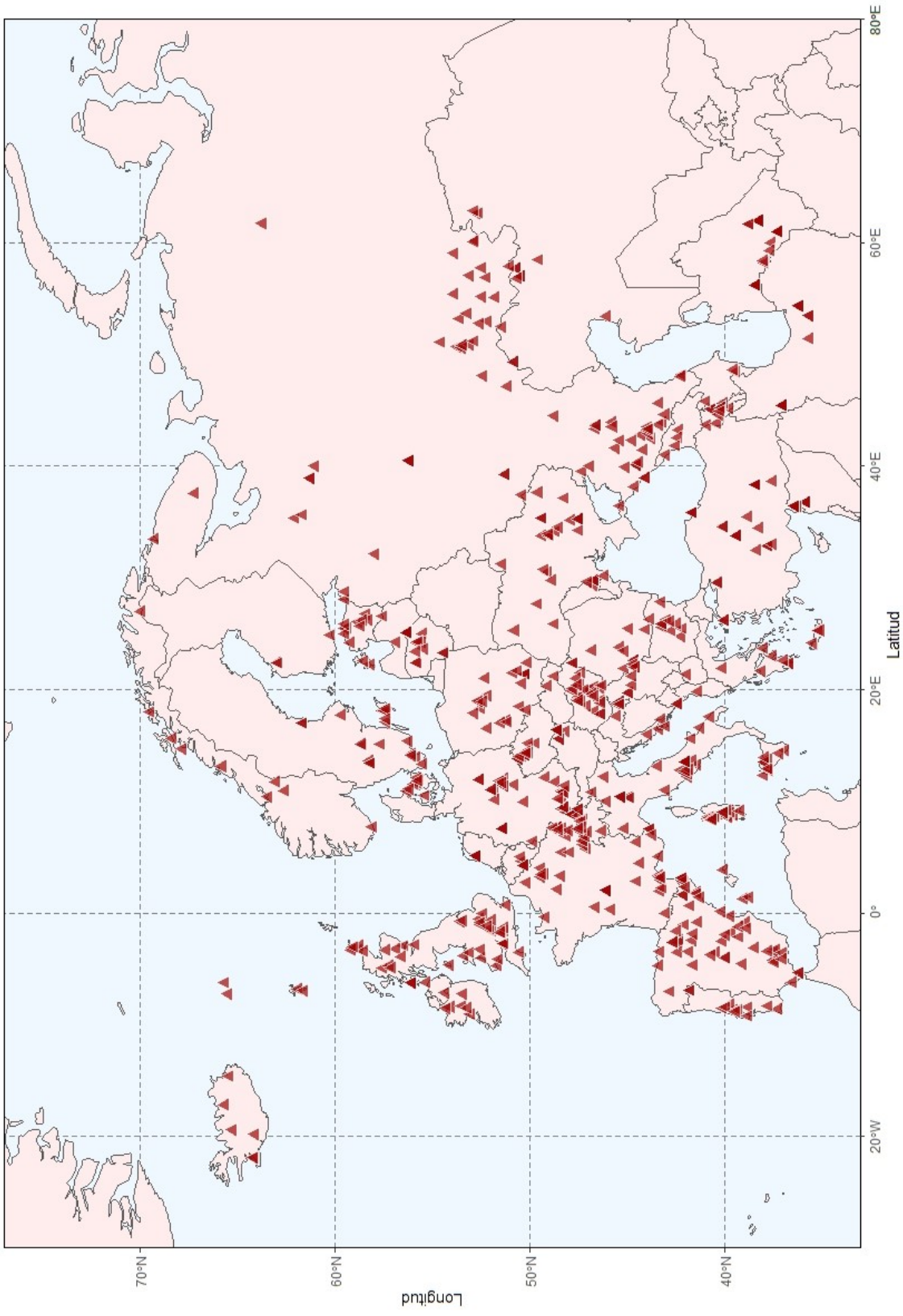


Figura 7.5: Mapa de la ubicación de las 697 subagrupaciones obtenidas realizado por medio de *k-medias*. El espacio geográfico observado es también el utilizado para el análisis de dispersión de los alelos.

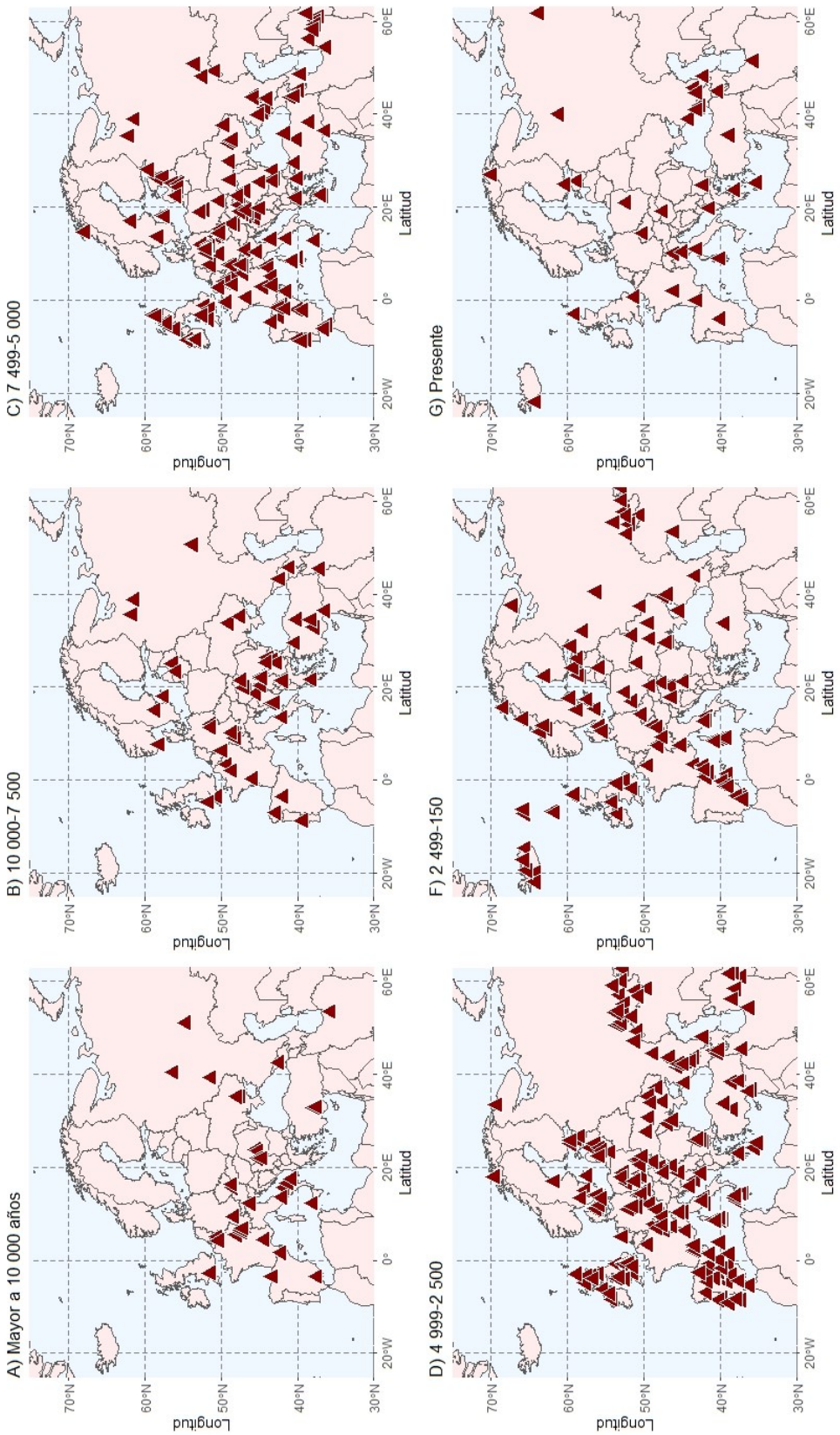


Figura 7.6: Mapa de la ubicación de las subagrupaciones obtenidas por *k-medias* por periodo. Las ubicaciones representan el promedio de ubicación de cada subgrupo.

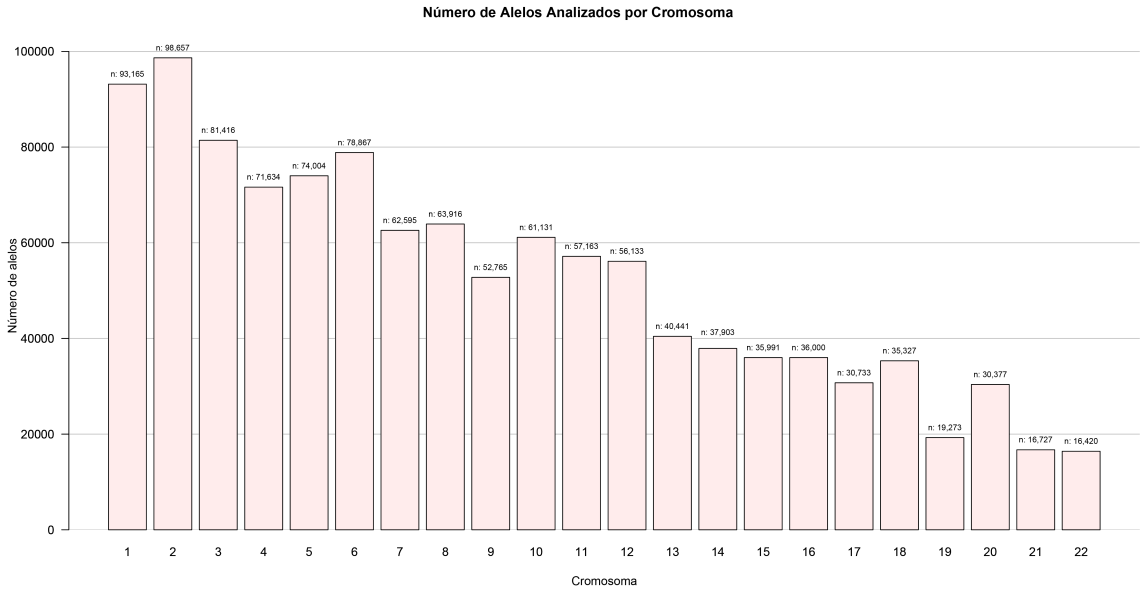
Rango	Número de Alelos
0-4.99	848,513
5-9.99	215,232
10-14.99	59,518
15-19.99	22,101
20-24.99	4,670
25-29.99	435
30-34.99	123
35-39.99	30
40-44.99	10
45-50	6
Total	1,150,638

Cuadro 7.1: Número de Alelos Identificados por Puntaje *CADD*.

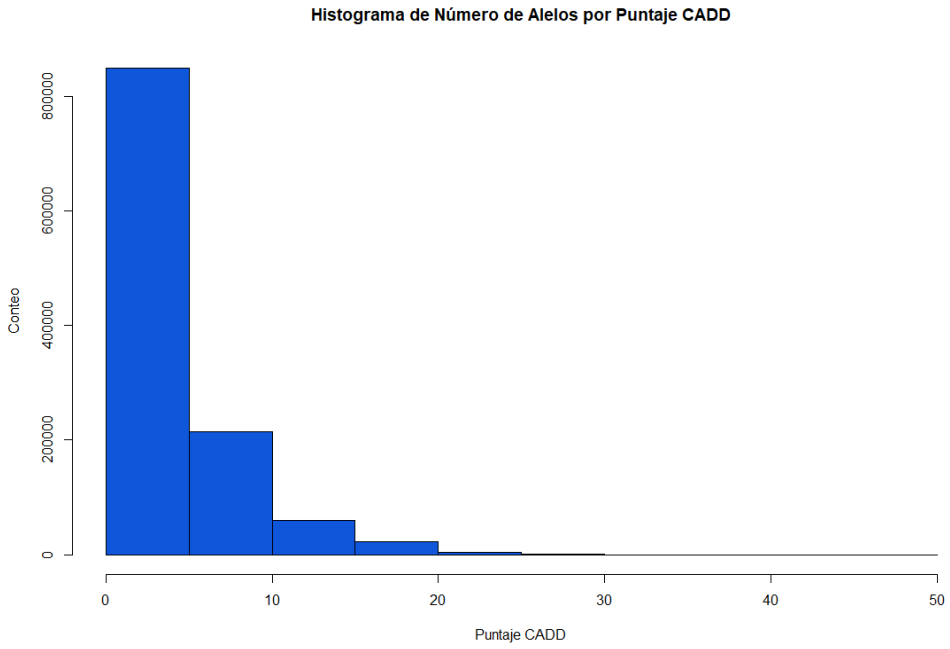
simular migraciones. Por desgracia, se necesitan estudios más profundos para establecer todas las barreras geográficas a lo largo del tiempo en Europa y simular las migraciones de larga distancia (Peter *et al.*, 2020).

7.4. La inferencia de los valores de selección natural sugieren cambios temporales en las presiones selectivas.

En cuanto a los valores de selección, se evaluaron hasta 6 valores, de -1 a 0, siendo -1 totalmente deletéreo y 0 neutro para cada uno de los grupos de *CADD* (grupos de 5 en 5). No se encontraron frecuencias alélicas de loci con $CADD > 10$ posiblemente porque los individuos que portaran alelo con valores altos de CADD no hayan sobrevivido a la etapa adulta. Sólo fue posible calcular los valores del grupo $CADD_{>5}$. En la figura 7.9 se puede observar que el estimado de máxima verosimilitud es negativo $s = -0{,}2$ (recordando que $s = d/2$ y que la figura 7.9 muestra los valores de verosimilitud para la variable d). Por otra parte, se analizó la presencia de alelos neutros, deletéreos (figura 7.10) y ventajosos (figuras 7.11 y 7.12) a lo largo del tiempo para analizar los cambios de frecuencias de estos tres tipos de alelos. Estos fueron agrupados en distintas temporalidades: A=>10000, B=10000-7500, C=7500-5000, D=5000-2500, E=2500-150 y F=Presente. Es de destacar que mientras que algunos alelos tentativamente ventajosos aumentan gradualmente la media de las frecuencias alélicas a medida que vamos acercándonos a las poblaciones del presente, no ocurre así en los alelos tentativamente deletéreos y neutros, siendo considerable este cambio entre la época *E* y *F*. Por otra parte, se analizó la frecuencia de los alelos ventajosos obtenidos por Mathie-



(a)



(b)

Figura 7.7: (a) Número de alelos identificados con puntaje *CADD* por cromosoma. (b) Histograma del número de alelos por puntaje *CADD* en un rango de 5.

son *et al.* (2015), algunos de los cuáles muestran una disminución en la media de su frecuencia: *rs174546, rs2269424, rs272872* y *rs653178*. Cabe destacar que estos alelos recibieron una puntuación *CADD* de rs653178=0.294, rs6903823=0.448, rs272872=1.533, rs2269424=4.119, rs7944926 =5.482, rs174546=10.84, rs4988235=13.88, rs12913832=16.46, rs16891982=24.8. Al tratar de obtener el valor de verosimilitud de los parámetros del coeficiente de selección, arrojaron valores máximos que corresponden a parámetros negativos, es decir, un coeficiente de selección negativo (figura 7.13), contrario a lo descrito en la literatura, en particular llamó la atención un alelo ligado a la tolerancia a la lactosa, (*rs4988235*; McEvoy *et al.* (2009); Ye y Gu (2011)). Al no encontrar concordancia entre los estimados de verosimilitud de los valores de selección con algunos alelos clásicos de ejemplos de selección positiva, se decidió eliminar en el alelo de la lactosa a aquellas poblaciones con una edad promedio mayor a 5000 años, siendo éste el tiempo estimado del inicio de la selección de alelos ligados a la tolerancia a la lactosa (Burger *et al.*, 2020; Mathieson *et al.*, 2015). El resultado fue de un cambio en el valor estimado de selección a 0.05 (figura 7.17), que indica un valor de selección positiva que es cercano a lo descrito por los autores Burger *et al.* (2020). Esto sugirió que es importante considerar criterios de inclusión en las poblaciones que se analizan, por ejemplo la antigüedad, para observar un cambio positivo en la frecuencia y encontrar señales de selección positiva.

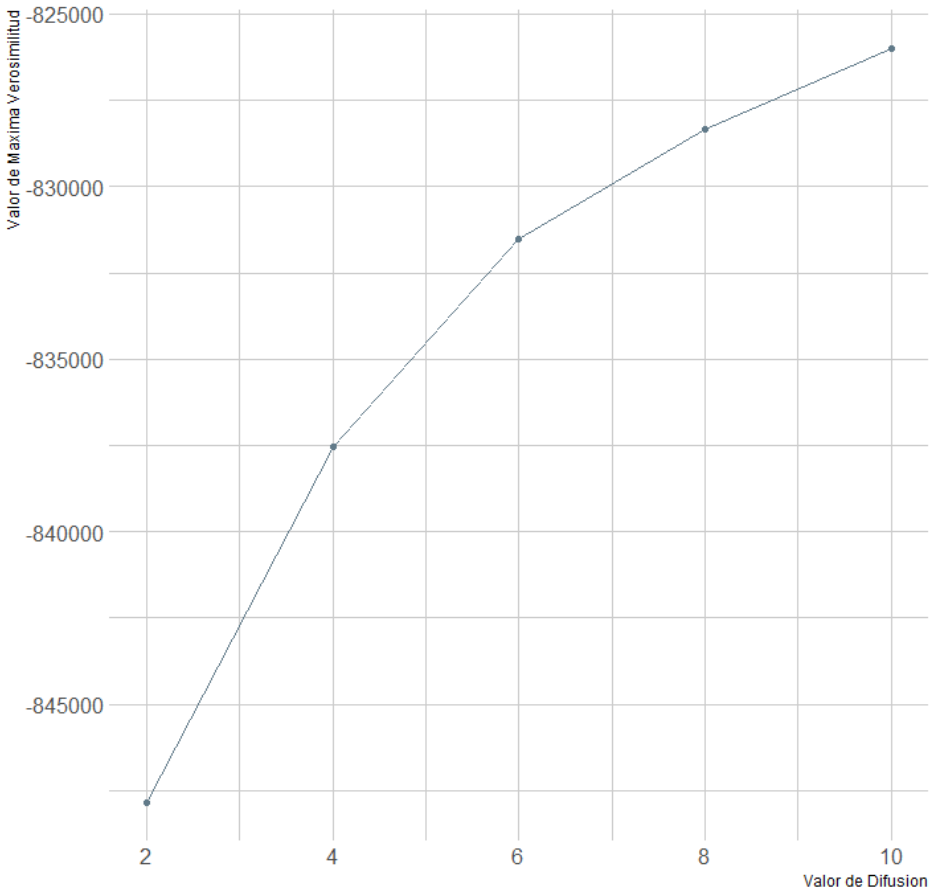


Figura 7.8: Valores de Verosimilitud para el parámetro de dispersión .

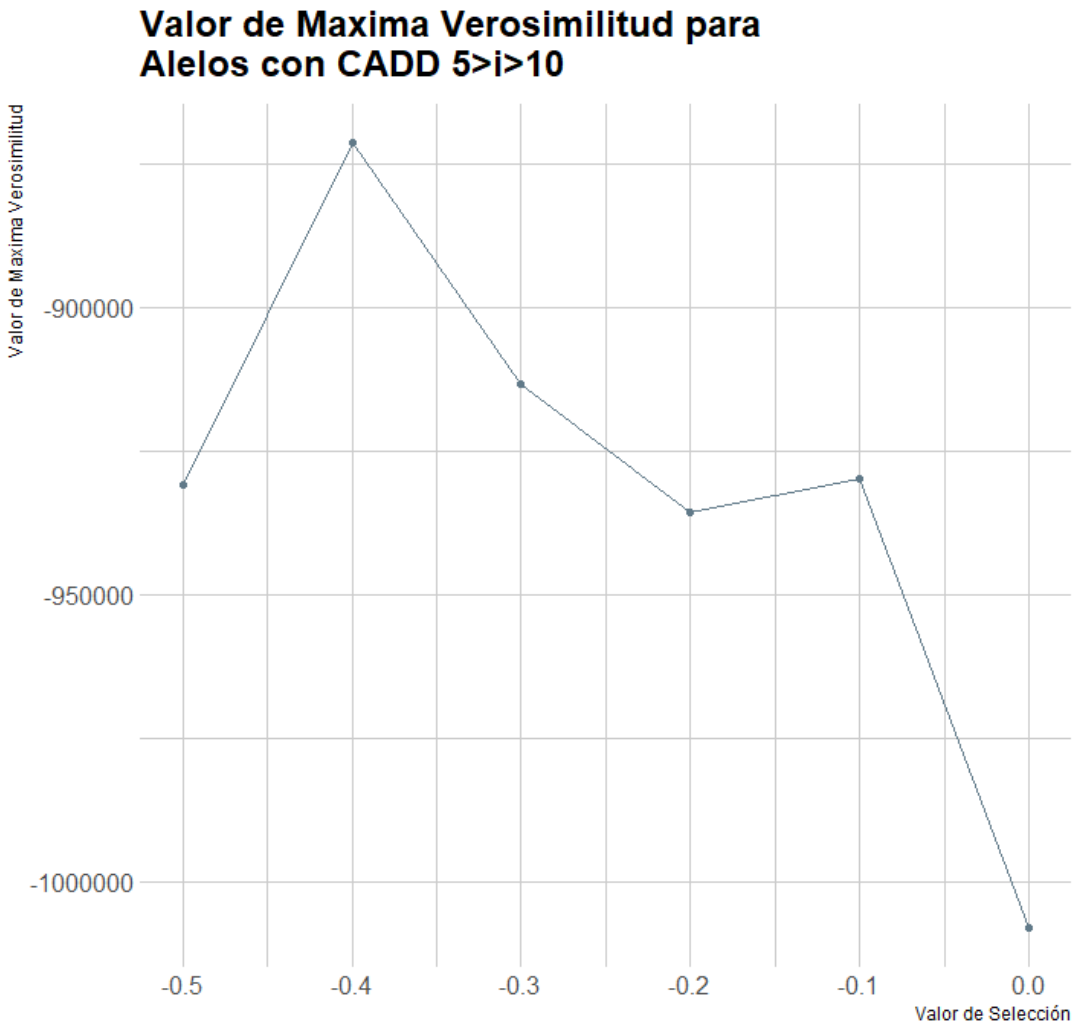


Figura 7.9: Valores de Verosimilitud para el Coeficiente de Selección d de Alelos con puntaje $CADD$=5>i>10.

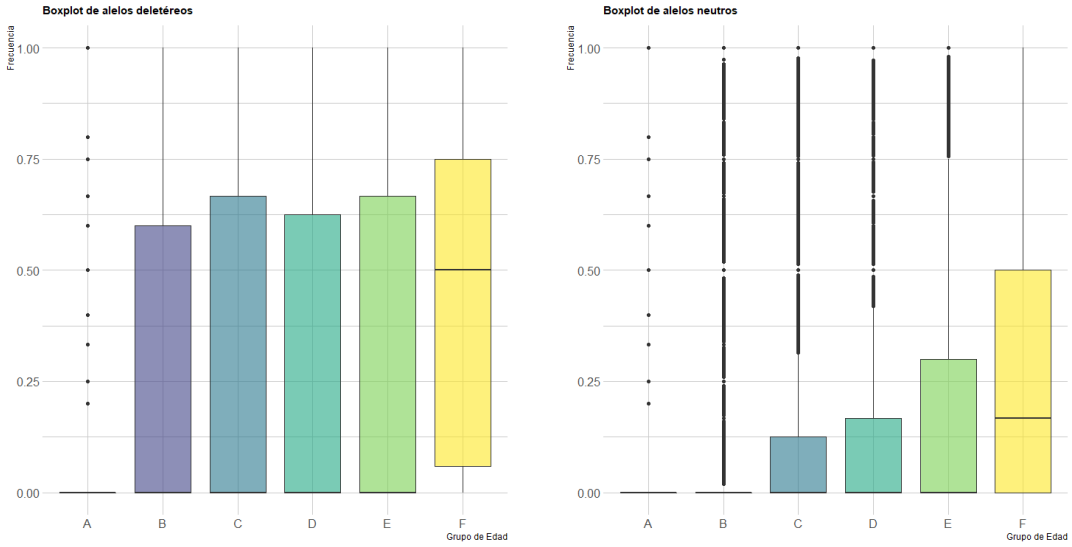


Figura 7.10: Gráfico de *cajas y bigotes* donde se muestra la distribución de las frecuencias alélicas de los subgrupos poblacionales de los primeros 10,000 alelos neutros y deletéreos por puntaje *CADD* por periodos: A=>10000, B=10000-7500, C=7500-5000, D=5000-2500, E=2500-150 y F son subgrupos del presente.

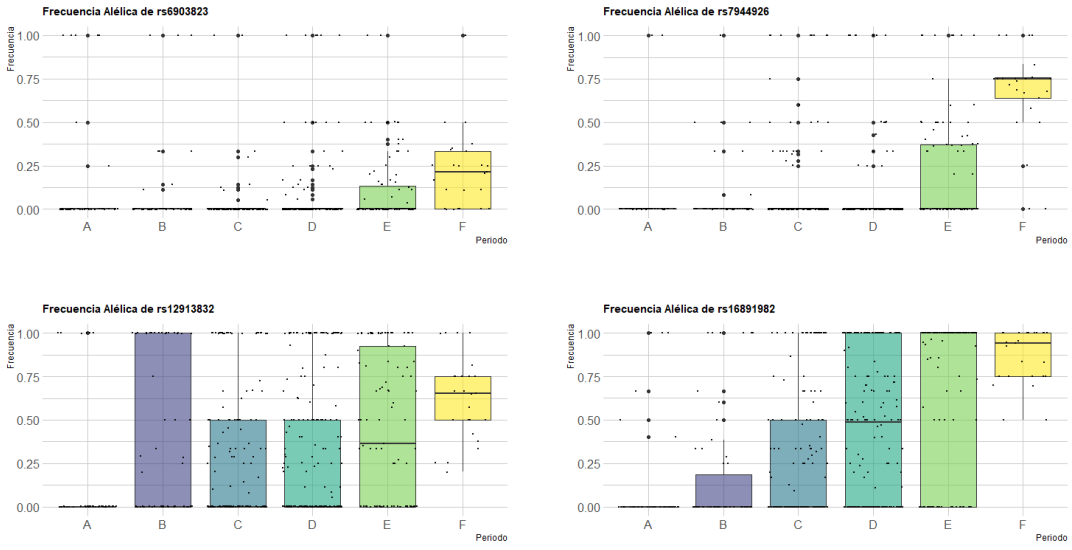


Figura 7.11: Parte 1: Gráfico de *cajas y bigotes* donde muestra la distribución de las frecuencias de los 10 alelos tentativamente ventajosos identificados por Mathieson *et al.* (2015) de los subgrupos poblacionales por periodos: A=>10000, B=10000-7500, C=7500-5000, D=5000-2500, E=2500-150 y F son subgrupos del presente.

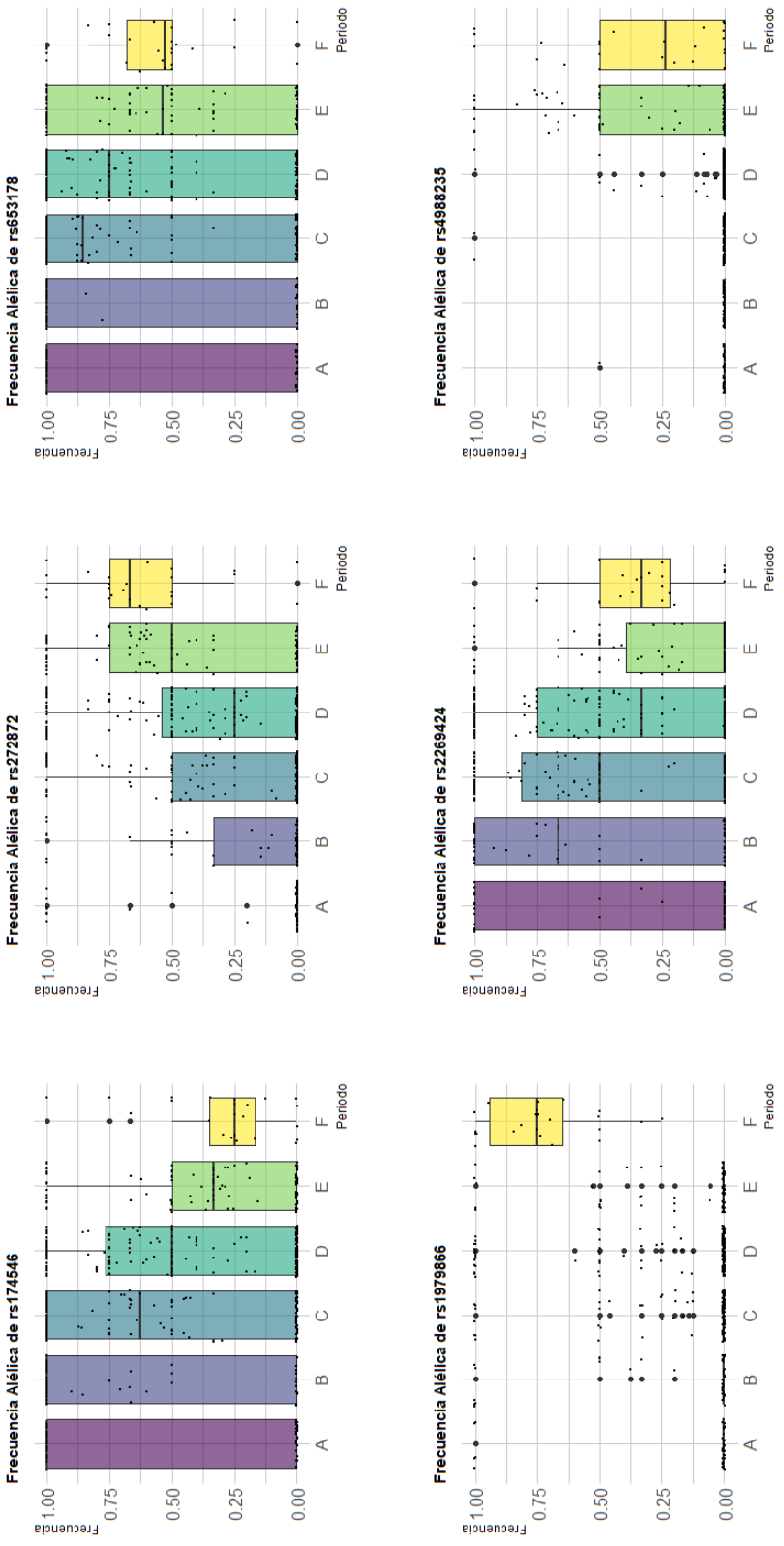


Figura 7.12: Parte 2: Gráfico de *cajas y bigotes* donde muestra la distribución de las frecuencias de alelos tentativamente ventajosos (Mathieson *et al.*, 2015) de los subgrupos poblacionales por periodos: A=>10000, B=10000-7500, C=7500-5000, D=5000-25000, E=2500-150 y F son subgrupos del presente.

Figura 7.13: Valores de verosimilitud de los parámetros de selección de los alelos tentativamente ventajosos según Mathieson *et al.* (2015). En rojo los valores de máxima verosimilitud de cada uno de los alelos tentativamente ventajosos, es decir, el coeficiente de selección más probable calculado. Los valores al infinito negativo fueron normalizados hasta -3500.

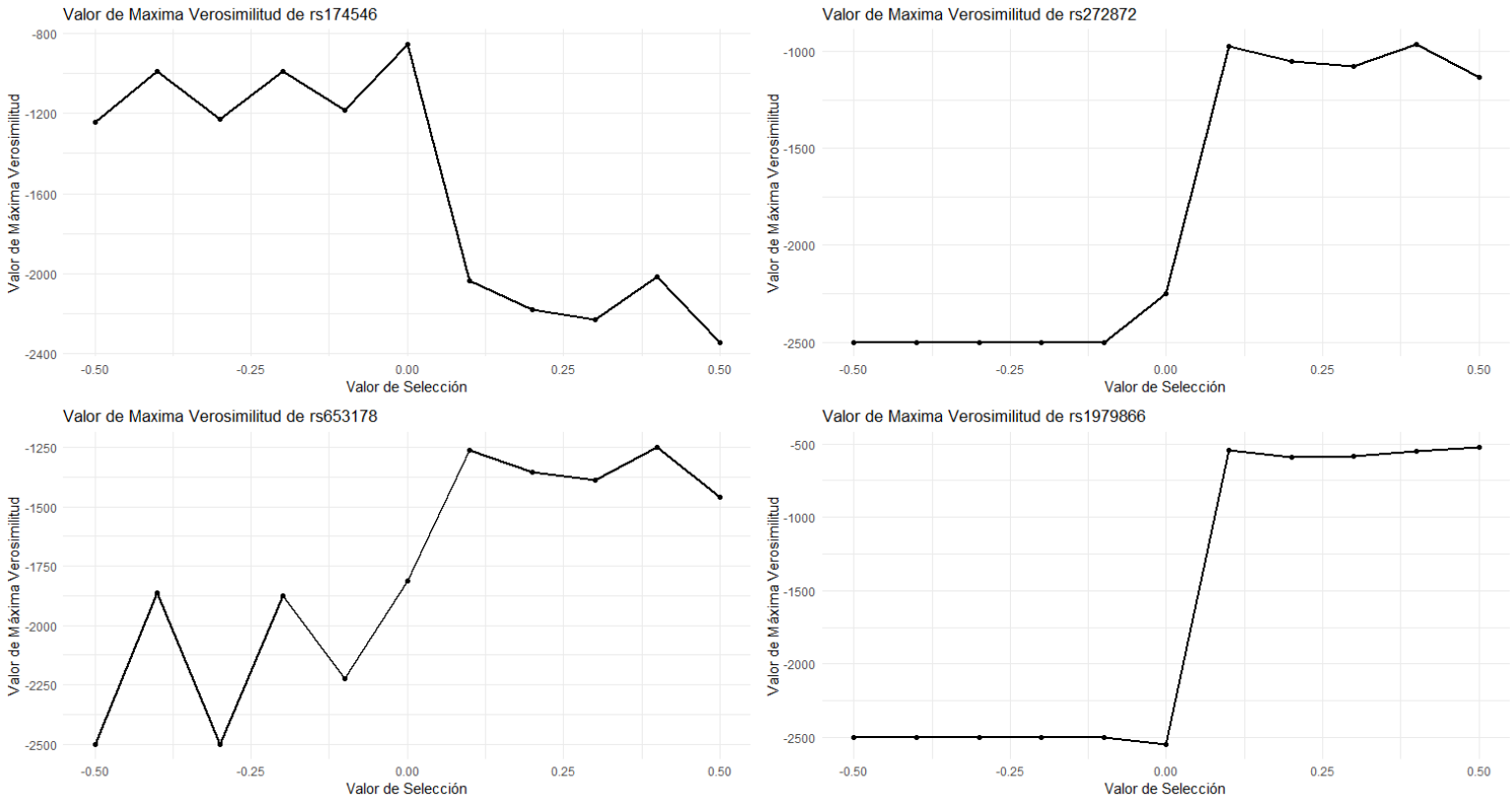


Figura 7.14: Valores de verosimilitud de los parámetros de selección de los alelos tentativamente ventajosos según Mathieson *et al.* (2015). En orden de izquierda a derecha, *rs174546* asociado al metabolismo de ácidos grasos, *rs27872* transporte de ergotionina, *rs6533178* no se conoce su asociación y al igual que *rs1979866*. *Continua en la siguiente figura.*

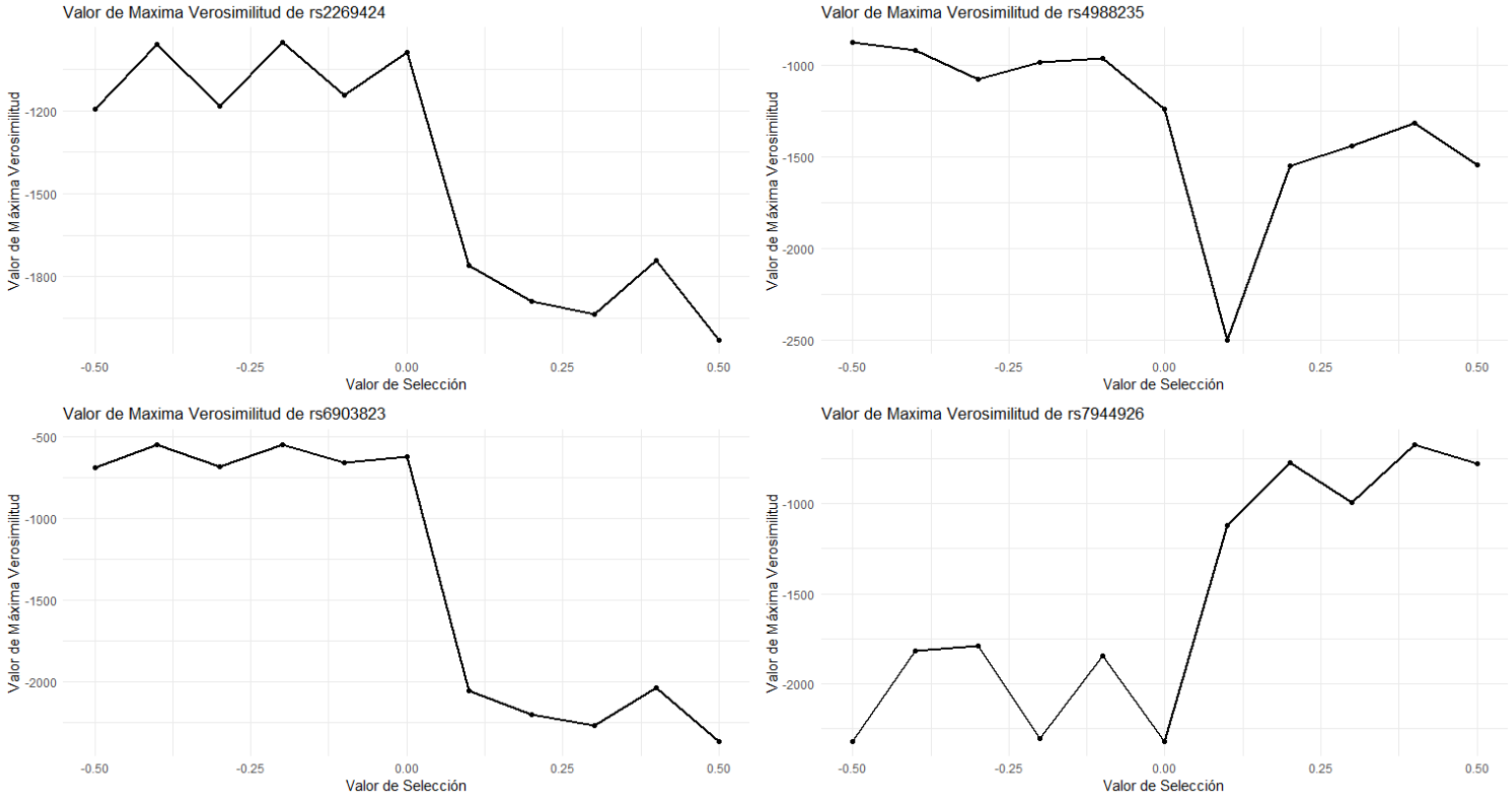


Figura 7.15: *Continuación.* Valores de verosimilitud de los parámetros de selección de los alelos tentativamente ventajosos según Mathieson *et al.* (2015). En orden de izquierda a derecha, *rs2269424* asociado al sistema inmune, *rs4988235* tolerancia a la lactosa, *rs6903823* autofagia y función pulmonas y *rs7944926* metabolismo de vitamina D.

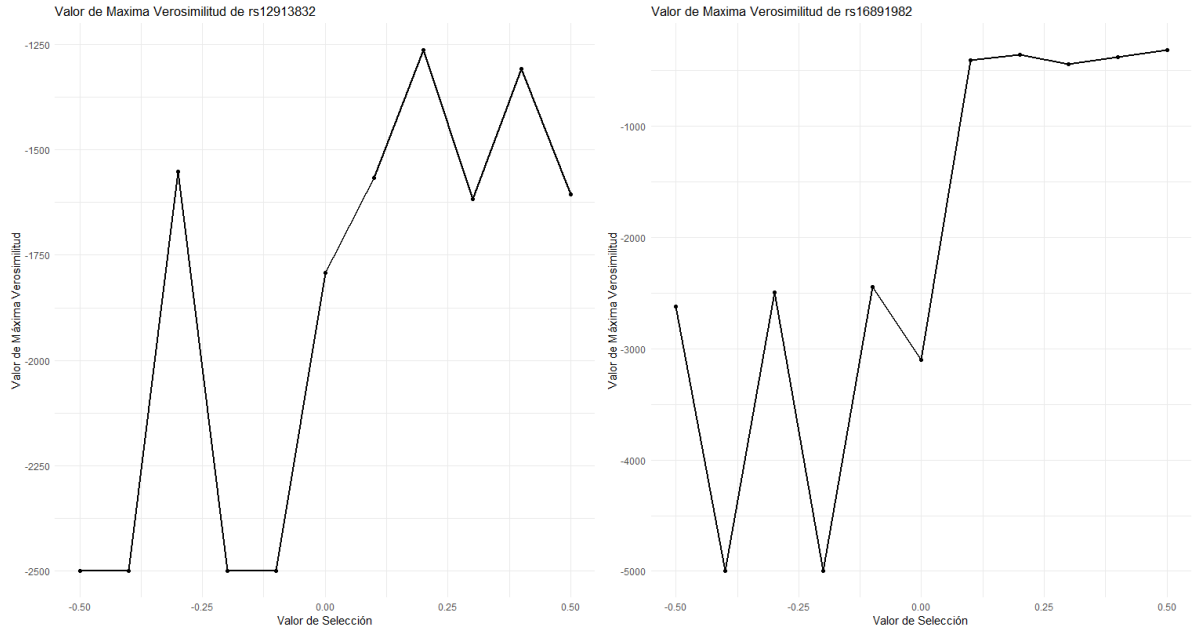


Figura 7.16: *Continuación.* Valores de verosimilitud de los parámetros de selección de los alelos tentativamente ventajosos según Mathieson *et al.* (2015). En orden de izquierda a derecha, *rs12913832* asociado al color de ojos y *rs16891982* asociado con la pigmentación.

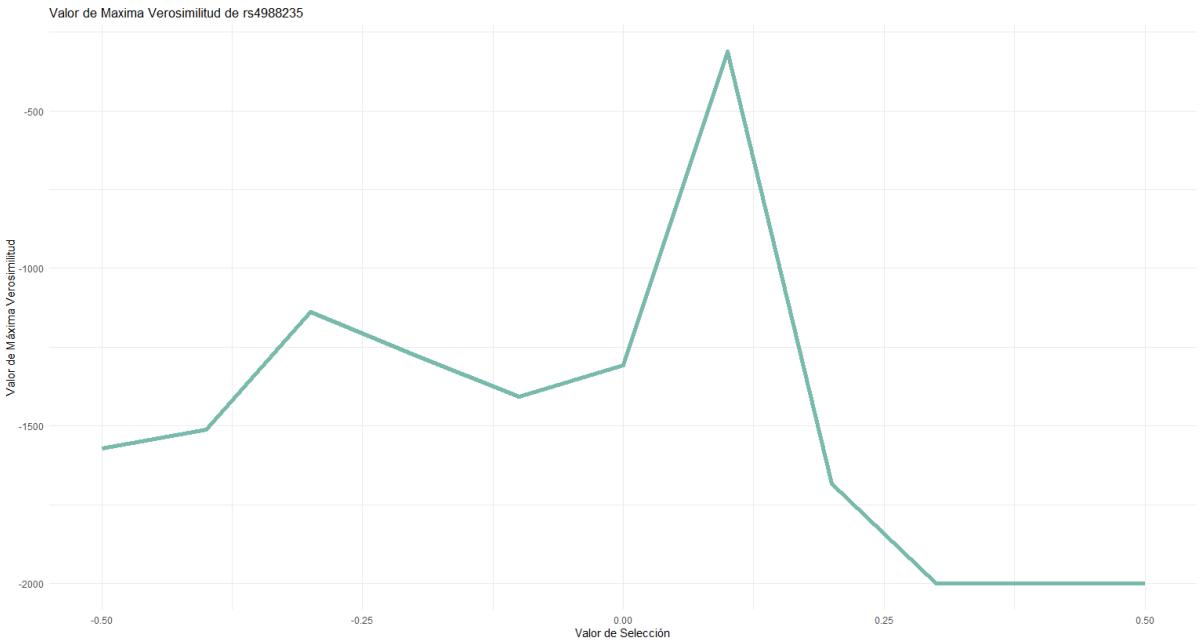


Figura 7.17: Valores de verosimilitud de los parámetros de selección de rs4988235 asociado a tolerancia a la lactosa. En este gráfico no se incluyeron poblaciones mayores a 5000 años de antigüedad.

8

Discusión

8.1. El coeficiente de selección probablemente no es continuo en el tiempo en alelos bajo selección positiva.

La correlación entre los valores de selección negativos y neutros y el puntaje *CADD* ha sido previamente estudiada, encontrando que están estrechamente relacionados entre sí (Racimo y Schraiber, 2014). En esta tesis, se han obtenido estimados de máxima verosimilitud a favor de valores de selección negativos (s=-0.2 en todos los casos) en alelos con un puntaje *CADD* entre 5 y 10. Esto indica la presencia de alelos deletéreos en el *chip* de captura evaluados en esas posiciones. Por otra parte, los valores de selección de los alelos tentativamente ventajosos variaban desde negativos hasta positivos. Los valores negativos provenían de alelos relacionados con la función pulmonar (*rs6903823*, genes ZSCAN31:ZKSCAN3), inmunidad (*rs2269424*; PPT2-EGFL8:EGFL8:2KB) y tolerancia a la lactosa (*rs4988235*; MCM6). Es interesante observar que tanto el primero como el segundo han sido reportados en muchas ocasiones como alelos que tienen una ventaja adaptativa en las poblaciones humanas (Ingram *et al.*, 2009; Liebert *et al.*, 2017; Mathieson *et al.*, 2015), siendo la tolerancia a la lactosa uno de los fenómenos más estudiados. Existen más alelos involucrados en la tolerancia a la lactosa (por ejemplo, *rs4954490, rs4988235*) y se ha encontrado evidencia de selección positiva actuando en esos alelos (Burger *et al.*, 2020). La tolerancia a la lactosa es un proceso que ha sido analizado principalmente a través de análisis de *Desequilibrio de Ligamiento* y *Estudios de Asociación de Genoma Completo*. Aquí concluimos que la razón por la que inferimos coeficiente de selección negativo para el alelo de la lactasa se debe a que el ambiente ha seleccionado a la variante adaptativa durante los últimos 5000-3000 años (Burger *et al.*, 2020). Tomar datos de poblaciones más antiguas al momento de inicio de presión selectiva lleva a un cálculo erróneo en la fuerza de la selección actuando sobre esta variante. Esto nos sugiere que es importante tomar en cuenta los tiempos

a partir de los cuales se empieza observar este cambio positivo en las frecuencias para observar señales de selección positiva. Otra posibilidad para encontrar señales de selección positiva es adaptar el parámetro del coeficiente de selección para que cambie a lo largo del tiempo. Esto nos permitiría detectar el punto en el tiempo en que una presión selectiva empieza a actuar y nos permitiría cuantificar la fuerza de la selección natural actuando sobre un alelo. En un futuro adecuaremos nuestro modelo para permitirle incluir parámetros en el tiempo que indiquen cuándo inicia cierta presión selectiva. Adicionalmente, destacar que el método de detección de alelos bajo selección positiva utilizado por Mathieson *et al.* (2015) fue a partir de un estudio comparativo entre tres poblaciones antiguas de Europa (cazadores-recolectores occidentales[1], europeos agrícolas del neolítico[2] y *Yamnaya*[3]) con frecuencias simuladas bajo un modelo neutral nulo con la idea de obtener los alelos con las frecuencias alélicas más significativas que muestran cambios en frecuencia que no pueden ser explicadas por el modelo neutro.

El alelo *rs6903823*, asociado a la autofagia y la función pulmonar, tiene un valor *CADD* neutro (0.448) con valor de selección negativo s=-0.2 y *rs4988235*, asociado a la tolerancia la lactosa, mostró un valor deletéreo (13.88) con valor de selección s=-0.05, sin embargo ambos mostraron un aumento en las gráficas de cajas y bigotes entre los periodos E (2500-150) al F (presente). Una posible explicación de este fenómeno, así como de la presencia de un posible alelo con fuerte selección positiva según el estudio de Mathieson *et al.* (2015), radica en la existencia de un período en el que las frecuencias alélicas eran insignificantes, anterior a los 10000 A.P. Aunque disponemos de datos que respaldan esta teoría, menos de 10000 años A.P. surgieron en el continente grupos de seres humanos, entre ellos los tres grupos estudiados por Mathieson y colaboradores, y fue en estas poblaciones donde se produjo un proceso de selección a favor de variantes alélicas relacionadas con la tolerancia a la lactosa y la función pulmonar. Para poder observar la evolución de las frecuencias alélicas de estos y otros alelos presumiblemente ventajosos, se requiere una revisión más exhaustiva de genomas antiguos. (Lazaridis *et al.*, 2014).

El alelo *rs174546*, el cuál codifica para una variante alélica del UTR 3 prima del gen *acido graso desaturasa 1* o *FADS1*[4], y obtuvo un puntaje *CADD* deletéreo y un valor de selección igual a 0, lo que lo hace tentativamente neutro. La media de la frecuencia de este alelo ha ido disminuyendo gradualmente hasta llegar a nuestros días. Este alelo junto con otros tres forman parte del haplotipo B del gen *FADS1*, el cuál se encarga del metabolismo de ácidos grasos, y que se encuentra disperso en los seres humanos actuales y neandertales. Tiene una relación

[1] 40000 a 8000 A.P.

[2] 8000 a 4200 A.P.

[3] Cultura ubicada en la actual Ucrania entre los años 5300-4600 A.P.

[4] Perteneciente a un grupo de genes FADS, esta variante está asociada al metabolismo de ácidos grasos. La variante estudiada en éste trabajo está asociada con enfermedades metabólicas.

fuerte de desequilibrio de ligamiento con el alelo *rs174594*, del cuál se hipotetiza se encuentra bajo selección positiva (Buckley *et al.*, 2017). Sin embargo, es posible que la fuerza de selección natural haya cambiado a lo largo del tiempo debido a cambios en el ambiente. Por ejemplo, *rs174546*, el alelo relacionado al metabolismo de ácidos grasos, es un alelo con posibles afectaciones a la salud, relacionado en estudios de asociación con la acumulación de ácidos grasos no saturados en la sangre (Zietemann *et al.*, 2010). Sin embargo, debido a que los recursos para obtener ácidos grasos eran escasos en el pasado, la presencia de genes que permitieran almacenar ácidos grasos en el cuerpo conferían una ventaja ante la posibilidad de no encontrar otros recursos en el futuro. Por otra parte, una persona portadora de estos alelos derivados en el presente puede presentar problemas de salud relacionados con el corazón debido al cambio en la dieta moderna que es alta en ácidos grasos (Amorim *et al.*, 2017), lo que explica la existencia o presencia de alelos con coeficientes de selección positiva en la literatura pero con un puntaje *CADD* poco ventajoso. En el futuro esperamos contar con datos concretos para realizar cortes temporales para aquellos alelos tentativamente ventajosos cuyos coeficientes hayan sido negativos y observar cómo este valor ha cambiado a lo largo del tiempo.

8.2. Los alelos con puntaje CADD mayor a 10 probablemente aparecieron en los últimos 30000 años.

El puntaje del Apunte Combinado Dependiente de Agotamiento *CADD* es una herramienta de aprendizaje máquina o *machine learning* que permite predecir el efecto que tendría una variante en el genoma humano. Aunque su uso es mayoritariamente con fines médicos (van der Velde *et al.*, 2015, 2017), también se ha utilizado para el análisis de estas variantes a través del tiempo (Aris-Brosou, 2019). En esta tesis se encontró que la gran mayoría de las variantes alélicas utilizadas caen en el umbral de *neutros*, siendo un 74.73 % de los sitios analizados identificados como neutros. Además, al momento de inferir el valor de máxima verosimilitud del coeficiente de selección, se encontró que de 50000 alelos, sólo 7030 (14.06 %) han tenido un tiempo de aparición de menos de 10000 años hacia el pasado, dejando al resto de alelos con una antigüedad mayor a 10000 años. Curiosamente, todos los alelos analizados tienen puntajes CADD$<$10, siendo aquellos con puntaje CADD$>$10 descartados y por lo tanto no vistos en nuestras poblaciones ya que que este análisis está limitado en su capacidad computacional para hacer los cálculos con alelos de más de 10000 años. En el futuro esperamos poder analizar algunos alelos con asociación médica importante sin importar su tiempo de aparición. En este análisis asumimos que el puntaje *CADD* es un indicador de qué tan deletéreos es un alelos, tal como Racimo y Schraiber (2014) concluyeron, por lo que podemos inferir que los alelos más deletéreos (CADD$>$10) han estado presentes al menos los últimos 30000 años. Por último, es

importante destacar que los alelos tentativamente ventajosos analizados no tienen un patrón de valores CADD o con valores CADD similares debido a que muchos de ellos fueron analizados bajo la visión de afectaciones de salud médica (Rentzsch *et al.*, 2021), lo que explica la existencia o presencia de alelos con coeficientes de selección positiva en la literatura pero con un puntaje *CADD* poco ventajoso.

En cuanto a los alelos deletéreos, se ha observado que de manera general estos tienden a aumentar con el paso del tiempo. Si bien esto podría indicarnos que existe una tendencia a la acumulación de alelos deletéreos a través del tiempo, esto también ocurre con los alelos neutros y ventajosos. Aris-Brosou (2019) encontró esta relación en estas mismas poblaciones europeas para los alelos deletéreos por puntaje *CADD*, sin embargo, no consideró otros alelos ventajosos en su momento. De igual manera, se encontró con un patrón similar de aumento de la carga genética en estas poblaciones y una tendencia a un aumento en la frecuencia en los tres grupos de alelos bajo selección, sin embargo, no ocurre así en todos los periodos. Al observar las gráficas de cajas y bigotes, podemos observar un salto entre el periodo E (5000-2500 años) y F (2500-150 años), para los alelos neutros, y considerando que estamos partiendo de un modelo basado en un solo grupo, es probable que este aumento de alelos pueda deberse a la entrada de alelos derivados cuyo origen sería diferente a este grupo poblacional estudiado. Por otra parte, Albers y McVean (2020) sugieren que los alelos deletéreos encontrados en la mayoría de las poblaciones modernas han aparecido no hace más de 1000 generaciones. En el futuro planeamos hacer análisis utilizando datos de genoma completo debido al sesgo que podría representa utilizar datos de un conjunto de sitios segregantes a lo largo del genoma.

8.3. Los alelos bajo selección tienen velocidades de dispersión diferentes a los alelos neutros.

Los modelos de dispersión de alelos ventajosos nos han permitido explorar qué tan restringido se podría encontrar un alelo con respecto a un espacio geográfico. En este trabajo utilizamos únicamente un modelo variante de Novembre *et al.* (2005) y de Muktupavela *et al.* (2021) para explorar el coeficiente de dispersión σ. En este caso, la cantidad de celdas que un alelo neutro se logra dispersar por generación es de $\sigma^2 = 10$ celdas. Despejando el exponente y transformándolo a kilómetros equivaldría a $\sigma => 20{,}9$ km longitudinalmente y $\sigma = 16 - 30$ km latitudinalmente (considerando que cada celda está posicionado con respecto al sistema de coordenadas, por lo que un grado puede medir diferente entre más nos acercamos al ecuador). Este valor es mucho menor al reportado por Novembre *et al.* (2005) para alelos ventajosos en seres humanos ($\sigma => 100km$) y con un valor similar a los reportados por Muktupavela *et al.* (2021) (modelo con diferentes σ para la latitud x y longitud y) yendo desde $\sigma_x = 10 - 40$ hasta

$\sigma_y = 10 - 40$, lo que nos indica que los alelos bajo selección difieren su velocidad de dispersión (km/generación) con respecto a los alelos neutros, sin embargo no tenemos aún datos sobre la velocidad de dispersión de alelos tentativamente deletéreos en seres humanos. Una perspectiva a futuro de este trabajo es analizar el parámetro de dispersión con diferentes alelos bajo selección, debido a que la capacidad computacional fue limitada y no se pudo continuar con el análisis con valores de dispersión mayores a 10. Además, analizar el comportamiento de la distribución de los alelos con barreras físicas que impidan su avance. Cabe mencionar que este trabajo no considera eventos demográficos (migraciones de larga distancia, expansiones, deriva génica) o históricos (guerras, pandemias), puesto que son eventos de una ventana temporal muy pequeña, esperando en el futuro considerar estos eventos en un modelo más integrativo.

9

Conclusiones

- Desarrollamos un modelo estadístico para inferir por primera vez la tasa de dispersión de alelos neutrales además de inferir el coeficiente de selección natural de alelos tentativamente ventajosos y deletéreos a partir de datos temporales y espaciales en poblaciones humanas de Europa.

- La fuerza de la selección natural no es una constante que se mueve a lo largo del tiempo, sino que esta cambia según las condiciones del ambiente como lo muestra el alelo *rs4988235*, relacionado con la tolerancia a la lactosa que ha sido seleccionado a favor en los últimos 5,000 años. Es necesario expandir las metodologías que infieren la fuerza de la selección para incluir cambios temporales además de contar con información sobre eventos demográficos y barreras geográficas y su impacto de la selección natural.

- Los hábitos culturales también han tenido un impacto la fuerza de la selección. Estudios previos han encontrado que el alelo *174546* asociado al metabolismo de ácidos grasos, ha sido selecionado en el pasado, a pesar de que obtuvo un puntaje CADD>5, deletéreo. Esto probablemente se deba a un cambio en la dieta actual, la cuál es alta en grasas.

- Calculamos que la velocidad de dispersión de los alelos neutros es de $\sigma^2 = 10$ celdas o 16-30 km por generación. Estos valores de σ similar a los valores reportados por autores como Muktupavela *et al.* (2021) y menor a los reportados para alelos ventajosos por Novembre *et al.* (2005).

- Inferimos que alelos tentativamente deletéreos, tienen un coeficiente de selección deletéreo igual a s=-0.2. Esto indica que nuestro modelo identifica de forma correcta a alelos deletéreos. Además, esto demuestra que el puntaje *CADD* es el instrumento más adecuado para predecir el comportamiento de alelos con importancia médica, como fue reportado previamente por Racimo y Schraiber (2014).

- El uso de herramientas de aprendizaje máquina (agrupación por textit(k-medias, *CADD*, *HMC*) ayudan a afinar aún más los métodos de inferencia de la fuerza de la selección natural actuando en poblaciones humanas.

10

Glosario

- **Alelo**: es una variante de un gen que se encuentra en un locus específico de un cromosoma. Los alelos pueden diferir en la secuencia de ADN en una o varias bases. La distribución de alelos en una población es un factor importante en la genética de poblaciones, ya que influye en la variabilidad genética y en la evolución de las especies.

- **Carga genética**: se refiere a la carga de mutaciones deletéreas o perjudiciales que una población lleva en su material genético. Estas mutaciones pueden ser causantes de enfermedades, defectos de desarrollo o de disminución de la viabilidad reproductiva. La carga genética puede ser evaluada a través del análisis de la frecuencia de mutaciones perjudiciales en una población, así como la tasa de mutación de nuevas mutaciones perjudiciales. También puede calcularse mediante la diferencia en la adecuación que poseen en promedio todos los individuos de la población en comparación con la adecuación que poseería la población si todos los individuos tuvieran genotipos óptimos para ese ambiente particular.

- **Frecuencia alélica**: Se define como el número de copias de un alelo en una población dividido por el número total de copias de todos los alelos en ese locus en la población.

- **FST**: es una medida de la estructuración genética de una población, que describe la distribución de la variación genética entre y dentro de las poblaciones. FST se basa en la varianza de los alelos y su distribución en las poblaciones. El valor de FST varía entre 0 y 1, donde 0 indica que no hay diferencias genéticas entre las poblaciones, y 1 indica que las poblaciones están completamente separadas genéticamente.

- **Sitios Segregantes**: son posiciones específicas en la secuencia de ADN que varían entre individuos de una población debido a las mutaciones que han ocurrido en esos sitios a lo largo del tiempo. Estas variaciones pueden ser una sola base (un cambio de nucleótido) o pueden implicar la inserción o eliminación de segmentos más grandes del ADN.

- **Demes**: se refiere a un grupo de individuos dentro de una población que tienen una estructura genética distintiva y que se reproducen principalmente entre ellos.

- **Cadenas de Markov-Montecarlo**: son una clase de métodos computacionales utilizados para simular la distribución de probabilidad de un sistema complejo. Este enfoque combina dos técnicas: las cadenas de Markov, que son un conjunto de reglas para generar secuencias de estados aleatorios, y el método de Monte Carlo, que utiliza muestras aleatorias para estimar la probabilidad de un evento dado los datos observados.

- **Equilibrio neutral**. es un concepto en la genética de poblaciones que se refiere a un estado en el que la variación genética de una población se mantiene estable a lo largo del tiempo debido a la ausencia de selección natural o la presencia de una selección puramente neutral por lo que los alelos sólo evolucionan debido al impacto de mutaciones y de la deriva génica. En este estado, la frecuencia de los diferentes alelos en una población puede cambiar debido a la deriva genética, es decir, la fluctuación aleatoria en las frecuencias alélicas causada por la reproducción aleatoria. Sin embargo, en ausencia de la selección natural, la tasa a la que los alelos se fijan o desaparecen de la población debido a la deriva genética es equivalente para todos los alelos y, por lo tanto, no hay un cambio neto en la variación genética de la población a lo largo del tiempo. El equilibrio neutral de frecuencias es importante porque proporciona una línea de base para medir los efectos de la selección natural y otras fuerzas evolutivas sobre la variación genética de las poblaciones. Además, la teoría del equilibrio neutral ha sido útil para explicar la variabilidad genética observada en las poblaciones y para identificar las propiedades estadísticas de la evolución molecular.

Bibliografía

Achaz, G. Frequency Spectrum Neutrality Tests: One for All and All for One. *Genetics* **183**(1):249–258 (2009)

Albers, P.K. y McVean, G. Dating genomic variants and shared ancestry in population-scale sequencing data. *PLOS Biology* **18**(1):e3000586 (2020)

Ameur, A., Enroth, S., Johansson, Å., Zaboli, G., Igl, W., Johansson, A.C., Rivas, M.A., Daly, M.J., Schmitz, G., Hicks, A.A. *et al.* Genetic adaptation of fatty-acid metabolism: a human-specific haplotype increasing the biosynthesis of long-chain omega-3 and omega-6 fatty acids. *The American Journal of Human Genetics* **90**(5):809–820 (2012)

Amorim, C.E.G., Nunes, K., Meyer, D., Comas, D., Bortolini, M.C., Salzano, F.M., y Hünemeier, T. Genetic signature of natural selection in first Americans. *Proceedings of the National Academy of Sciences of the United States of America* **114**(9):2195–2199 (2017)

Aris-Brosou, S. Direct Evidence of an Increasing Mutational Load in Humans. *Molecular Biology and Evolution* **36**(12):2823–2829 (2019)

Bank, C., Ewing, G.B., Ferrer-Admettla, A., Foll, M., y Jensen, J.D. Thinking too positive? Revisiting current methods of population genetic selection inference. *Trends in Genetics* **30**(12):540–546 (2014)

Benazzi, S., Douka, K., Fornai, C., Bauer, C.C., Kullmer, O., Svoboda, J., Pap, I., Mallegni, F., Bayle, P., Coquerelle, M., Condemi, S., Ronchitelli, A., Harvati, K., y Weber, G.W. Early dispersal of modern humans in Europe and implications for Neanderthal behaviour. *Nature* **479**(7374):525–528 (2011)

Bollback, J.P., York, T.L., y Nielsen, R. Estimation of 2Nes from temporal allele frequency data. *Genetics* **179**(1):497–502 (2008)

Bradburd, G.S. y Ralph, P.L. Spatial Population Genetics: It's About Time. *Annual Review of Ecology, Evolution, and Systematics* **50**(1):427–449 (2019)

Buckley, M.T., Racimo, F., Allentoft, M.E., Jensen, M.K., Jonsson, A., Huang, H., Hormozdiari, F., Sikora, M., Marnetto, D., Eskin, E. *et al.* Selection in Europeans on fatty acid desaturases associated with dietary changes. *Molecular biology and evolution* **34**(6):1307–1318 (2017)

Burger, J., Link, V., Blöcher, J., Schulz, A., Sell, C., Pochon, Z., Diekmann, Y., Žegarac, A., Hofmanová, Z., Winkelbach, L., Reyna-Blanco, C.S., Bieker, V., Orschiedt, J., Brinker, U., Scheu, A., Leuenberger, C., Bertino, T.S., Bollongino, R., Lidke, G., Stefanović, S., Jantzen, D., Kaiser, E., Terberger, T., Thomas, M.G., Veeramah, K.R., y Wegmann, D. Low Prevalence of Lactase Persistence in Bronze Age Europe Indicates Ongoing Strong Selection over the Last 3,000 Years. *Current Biology* **30**(21):4307–4315.e13 (2020)

Byars, S.G., Ewbank, D., Govindaraju, D.R., y Stearns, S.C. Natural selection in a contemporary human population. *Proceedings of the National Academy of Sciences* **107**(suppl_1):1787–1792 (2010)

Carpenter, M.L., Buenrostro, J.D., Valdiosera, C., Schroeder, H., Allentoft, M.E., Sikora, M., Rasmussen, M., Gravel, S., Guillén, S., Nekhrizov, G., Leshtakov, K., Dimitrova, D., Theodossiev, N., Pettener, D., Luiselli, D., Sandoval, K., Moreno-Estrada, A., Li, Y., Wang, J., Gilbert, M.T.P., Willerslev, E., Greenleaf, W.J., y Bustamante, C.D. Pulling out the 1 %: whole-genome capture for the targeted enrichment of ancient DNA sequencing libraries. *American journal of human genetics* **93**(5):852–864 (2013)

Chen, Y., Graf, L., Chen, T., Liao, Q., Bai, T., Petric, P.P., Zhu, W., Yang, L., Dong, J., Lu, J., Chen, Y., Shen, J., Haller, O., Staeheli, P., Kochs, G., Wang, D., Schwemmle, M., y Shu, Y. Rare variant MX1 alleles increase human susceptibility to zoonotic H7N9 influenza virus. *Science* **373**(6557):918–922 (2021)

Cohen, S. Strong Positive Selection and Habitat-Specific Amino Acid Substitution Patterns in Mhc from an Estuarine Fish Under Intense Pollution Stress. *Molecular Biology and Evolution* **19**(11):1870–1880 (2002)

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., McVean, G., y Durbin, R. The variant call format and VCFtools. *Bioinformatics (Oxford, England)* **27**(15):2156–2158 (2011)

Darwin, C. *The origin of species.* Everyman's library. Dent (1856)

Dayhoff, M.O. *Atlas of protein sequence and structure*, tomo 4. National Biomedical Research Foundation. (1969)

Dehasque, M., Ávila-Arcos, M.C., Díez-del-Molino, D., Fumagalli, M., Guschanski, K., Lorenzen, E.D., Malaspinas, A., Marques-Bonet, T., Martin, M.D., Murray, G.G.R., Papadopulos, A.S.T., Therkildsen, N.O., Wegmann, D., Dalén, L., y Foote, A.D. Inference of natural selection from ancient DNA. *Evolution Letters* **4**(2):94–108 (2020)

Delahaye, C. y Nicolas, J. Sequencing DNA with nanopores: Troubles and biases. *PLOS ONE* **16**(10):e0257521 (2021)

Enard, D. Types of Natural Selection and Tests of Selection BT. En K.E. Lohmueller y R. Nielsen (editores), *Human Population Genomics: Introduction to Essential Concepts and Applications*, págs. 69–86. Springer International Publishing, Cham (2021)

Eshleman, J.A., Malhi, R.S., y Smith, D.G. Mitochondrial DNA Studies of Native Americans: Conceptions and Misconceptions of the Population Prehistory of the Americas. *Evolutionary Anthropology* **12**:7–18 (2003)

Ewens, W.J. Molecular Population Genetics: Introduction BT. En W.J. Ewens (editor), *Mathematical Population Genetics: I. Theoretical Introduction*, págs. 288–327. Springer New York, New York, NY (2004)

Fairley, S., Lowy-Gallego, E., Perry, E., y Flicek, P. The International Genome Sample Resource (IGSR) collection of open human genomic variation resources. *Nucleic Acids Research* **48**(D1):D941–D947 (2020)

Fay, J. y Wu, C. Fay JC, Wu CI. Hitchhiking under positive Darwinian selection. Genetics 155: 1405-1413. *Genetics* **155**:1405–1413 (2000)

Fisher, R.A. y Ford, E.B. The spread of a gene in natural conditions in a colony of the moth Panaxia dominula L. *Heredity* **1**(2):143–174 (1947)

Foll, M., Poh, Y.P., Renzette, N., Ferrer-Admetlla, A., Bank, C., Shim, H., Malaspinas, A.S., Ewing, G., Liu, P., Wegmann, D., Caffrey, D.R., Zeldovich, K.B., Bolon, D.N., Wang, J.P., Kowalik, T.F., Schiffer, C.A., Finberg, R.W., y Jensen, J.D. Influenza Virus Drug Resistance: A Time-Sampled Population Genetics Perspective. *PLOS Genetics* **10**(2):e1004185 (2014)

Foll, M., Shim, H., y Jensen, J.D. WFABC: a Wright-Fisher ABC-based approach for inferring effective population sizes and selection coefficients from time-sampled data. *Molecular ecology resources* **15**(1):87–98 (2015)

Freedman, A.H., Schweizer, R.M., Ortega-Del Vecchyo, D., Han, E., Davis, B.W., Gronau, I., Silva, P.M., Galaverni, M., Fan, Z., Marx, P., Lorente-Galdos, B., Ramirez, O., Hormozdiari, F., Alkan, C., Vilà, C., Squire, K., Geffen, E., Kusak, J., Boyko, A.R., Parker, H.G., Lee, C., Tadigotla, V., Siepel, A., Bustamante, C.D., Harkins, T.T., Nelson, S.F., Marques-Bonet, T., Ostrander, E.A., Wayne, R.K., y Novembre, J. Demographically-Based Evaluation of Genomic Regions under Selection in Domestic Dogs. *PLOS Genetics* **12**(3):e1005851 (2016)

Freese, E. y Yoshida, A. The Role of Mutations in Evolution. En V. Bryson, H.J.B.T.E.G. Vogel, y Proteins (editores), *Evolving Genes and Proteins*, págs. 341–355. Academic Press (1965)

Fu, Q., Meyer, M., Gao, X., Stenzel, U., Burbano, H.A., Kelso, J., y Pääbo, S. DNA analysis of an early modern human from Tianyuan Cave, China. *Proceedings of the National Academy of Sciences* **110**(6):2223–2227 (2013)

Fu, Q., Posth, C., Hajdinjak, M., Petr, M., Mallick, S., Fernandes, D., Furtwängler, A., Haak, W., Meyer, M., Mittnik, A., Nickel, B., Peltzer, A., Rohland, N., Slon, V., Talamo, S., Lazaridis, I., Lipson, M., Mathieson, I., Schiffels, S., Skoglund, P., Derevianko, A.P., Drozdov, N., Slavinsky, V., Tsybankov, A., Cremonesi, R.G., Mallegni, F., Gély, B., Vacca, E., Morales, M.R.G., Straus, L.G., Neugebauer-Maresch, C., Teschler-Nicola, M., Constantin, S., Moldovan, O.T., Benazzi, S., Peresani, M., Coppola, D., Lari, M., Ricci, S., Ronchitelli, A., Valentin, F., Thevenet, C., Wehrberger, K., Grigorescu, D., Rougier, H., Crevecoeur, I., Flas, D., Semal, P., Mannino, M.A., Cupillard, C., Bocherens, H., Conard, N.J., Harvati, K., Moiseyev, V., Drucker, D.G., Svoboda, J., Richards, M.P., Caramelli, D., Pinhasi, R., Kelso, J., Patterson, N., Krause, J., Pääbo, S., y Reich, D. The genetic history of Ice Age Europe. *Nature* **534**(7606):200–205 (2016)

Gelfman, S., Wang, Q., McSweeney, K.M., Ren, Z., La Carpia, F., Halvorsen, M., Schoch, K., Ratzon, F., Heinzen, E.L., Boland, M.J., Petrovski, S., y Goldstein, D.B. Annotating pathogenic non-coding variants in genic regions. *Nature Communications* **8**(1):236 (2017)

Gordo, I., Perfeito, L., y Sousa, A. Fitness Effects of Mutations in Bacteria. *Microbial Physiology* **21**(1-2):20–35 (2011)

Gould, S.J. The Frauitful Facets of Galton's Polyhedron. En *The Structure of evolutionary theory*, 1ª edición, capítulo The Frauit, pág. 427. Belknap-Harvard, Boston (2002)

Green, R.E., Krause, J., Ptak, S.E., Briggs, A.W., Ronan, M.T., Simons, J.F., Du, L., Egholm, M., Rothberg, J.M., Paunovic, M., y Pääbo, S. Analysis of one million base pairs of Neanderthal DNA. *Nature* **444**(7117):330–336 (2006)

Green, R.E., Krause, J., Briggs, A.W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M.H.Y., Hansen, N.F., Durand, E.Y., Malaspinas, A.S., Jensen, J.D., Marques-Bonet, T., Alkan, C., Prüfer, K., Meyer, M., Burbano, H.A., Good, J.M., Schultz, R., Aximu-Petri, A., Butthof, A., Höber, B., Höffner, B., Siegemund, M., Weihmann, A., Nusbaum, C., Lander, E.S., Russ, C., Novod, N., Affourtit, J., Egholm, M., Verna, C., Rudan, P., Brajkovic, D., Kucan, Ž., Gušic, I., Doronichev, V.B., Golovanova, L.V., Lalueza-Fox, C., de la Rasilla, M., Fortea, J., Rosas, A., Schmitz, R.W., Johnson, P.L.F., Eichler, E.E., Falush, D., Birney, E., Mullikin, J.C., Slatkin, M., Nielsen, R., Kelso, J., Lachmann, M., Reich, D., y Pääbo, S. A Draft Sequence of the Neandertal Genome. *Science* **328**(5979):710–722 (2010)

Hamilton, M. *Population Genetics*. 1ª edición. John Wiley & Sons Ltd, West Sussex (2009)

He, Z., Dai, X., Beaumont, M., y Yu, F. Estimation of Natural Selection and Allele Age from Time Series Allele Frequency Data Using a Novel Likelihood-Based Approach. *Genetics* **216**(2):463–480 (2020)

Henn, B.M., Botigué, L.R., Peischl, S., Dupanloup, I., Lipatov, M., Maples, B.K., Martin, A.R., Musharoff, S., Cann, H., Snyder, M.P., Excoffier, L., Kidd, J.M., y Bustamante, C.D. Distance from sub-Saharan Africa predicts mutational load in diverse human genomes. *Proceedings of the National Academy of Sciences* **113**(4):E440–E449 (2016). _eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.1510805112

Hudson, R.R., Kreitman, M., y Aguadé, M. A Test of Neutral Molecular Evolution Based on Nucleotide Data. *Genetics* **116**(1):153–159 (1987)

Huffman, J.E., Butler-Laporte, G., Khan, A., Pairo-Castineira, E., Drivas, T.G., Peloso, G.M., Nakanishi, T., Ganna, A., Verma, A., Baillie, J.K., Kiryluk, K., Richards, J.B., Zeberg, H., y Initiative, C..H.G. Multi-ancestry fine mapping implicates OAS1 splicing in risk of severe COVID-19. *Nature Genetics* **54**(2):125–127 (2022)

Ingram, C.J.E., Mulcare, C.A., Itan, Y., Thomas, M.G., y Swallow, D.M. Lactose digestion and the evolutionary genetics of lactase persistence. *Human genetics* **124**(6):579–591 (2009)

Johri, P., Eyre-Walker, A., Gutenkunst, R.N., Lohmueller, K.E., y Jensen, J.D. On the prospect of achieving accurate joint estimation of selection with population history. *Genome Biology and Evolution* **14**(7):evac088 (2022)

Keogh, E. y Kasetty, S. On the need for time series data mining benchmarks: A survey and empirical demonstration. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* págs. 102–111 (2002)

Kimura, M. The rate of molecular evolution considered from the standpoint of population genetics. *Proceedings of the National Academy of Sciences of the United States of America* **63**(4):1181–1188 (1969)

Kimura, M. Contribution of Mendel. En *My Thoughts on Biological Evolution*, 1ª edición, capítulo Diversity, págs. 1–10. Springer US (2020)

Kimura, M. *et al.* Evolutionary rate at the molecular level. *Nature* **217**(5129):624–626 (1968)

Kircher, M., Witten, D.M., Jain, P., O'Roak, B.J., Cooper, G.M., y Shendure, J. A general framework for estimating the relative pathogenicity of human genetic variants. *Nature genetics* **46**(3):310–315 (2014)

Kreitman, M. Methods to detect selection in populations with applications to the human. *Annual review of genomics and human genetics* **1**:539–559 (2000)

LaFramboise, T. Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. *Nucleic acids research* **37**(13):4181–4193 (2009)

Lappalainen, T., Scott, A.J., Brandt, M., y Hall, I.M. Genomic Analysis in the Age of Human Genome Sequencing. *Cell* **177**(1):70–84 (2019)

Lazaridis, I., Patterson, N., Mittnik, A., Renaud, G., Mallick, S., Kirsanow, K., Sudmant, P.H., Schraiber, J.G., Castellano, S., Lipson, M., Berger, B., Economou, C., Bollongino, R., Fu, Q., Bos, K.I., Nordenfelt, S., Li, H., de Filippo, C., Prüfer, K., Sawyer, S., Posth, C., Haak, W., Hallgren, F., Fornander, E., Rohland, N., Delsate, D., Francken, M., Guinet, J.M., Wahl, J., Ayodo, G., Babiker, H.A., Bailliet, G., Balanovska, E., Balanovsky, O., Barrantes, R., Bedoya, G., Ben-Ami, H., Bene, J., Berrada, F., Bravi, C.M., Brisighelli, F.,

Busby, G.B.J., Cali, F., Churnosov, M., Cole, D.E.C., Corach, D., Damba, L., van Driem, G., Dryomov, S., Dugoujon, J.M., Fedorova, S.A., Gallego Romero, I., Gubina, M., Hammer, M., Henn, B.M., Hervig, T., Hodoglugil, U., Jha, A.R., Karachanak-Yankova, S., Khusainova, R., Khusnutdinova, E., Kittles, R., Kivisild, T., Klitz, W., Kučinskas, V., Kushniarevich, A., Laredj, L., Litvinov, S., Loukidis, T., Mahley, R.W., Melegh, B., Metspalu, E., Molina, J., Mountain, J., Näkkäläjärvi, K., Nesheva, D., Nyambo, T., Osipova, L., Parik, J., Platonov, F., Posukh, O., Romano, V., Rothhammer, F., Rudan, I., Ruizbakiev, R., Sahakyan, H., Sajantila, A., Salas, A., Starikovskaya, E.B., Tarekegn, A., Toncheva, D., Turdikulova, S., Uktveryte, I., Utevska, O., Vasquez, R., Villena, M., Voevoda, M., Winkler, C.A., Yepiskoposyan, L., Zalloua, P., Zemunik, T., Cooper, A., Capelli, C., Thomas, M.G., Ruiz-Linares, A., Tishkoff, S.A., Singh, L., Thangaraj, K., Villems, R., Comas, D., Sukernik, R., Metspalu, M., Meyer, M., Eichler, E.E., Burger, J., Slatkin, M., Pääbo, S., Kelso, J., Reich, D., y Krause, J. Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* **513**(7518):409–413 (2014)

Lewontin, R.C. Twenty-five years ago in Genetics: electrophoresis in the development of evolutionary genetics: milestone or millstone? *Genetics* **128**(4):657–662 (1991)

Lewontin, R.C. y Krakauer, J. Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics* **74**(1):175–195 (1973)

Liebert, A., López, S., Jones, B.L., Montalva, N., Gerbault, P., Lau, W., Thomas, M.G., Bradman, N., Maniatis, N., y Swallow, D.M. World-wide distributions of lactase persistence alleles and the complex effects of recombination and selection. *Human genetics* **136**(11-12):1445–1453 (2017)

Liu, Y. Darwin's Pangenesis: A Theory of Everything? *Advances in genetics* **101**:1–30 (2018)

Locke, A.E., Steinberg, K.M., Chiang, C.W.K., Service, S.K., Havulinna, A.S., Stell, L., Pirinen, M., Abel, H.J., Chiang, C.C., Fulton, R.S., Jackson, A.U., Kang, C.J., Kanchi, K.L., Koboldt, D.C., Larson, D.E., Nelson, J., Nicholas, T.J., Pietilä, A., Ramensky, V., Ray, D., Scott, L.J., Stringham, H.M., Vangipurapu, J., Welch, R., Yajnik, P., Yin, X., Eriksson, J.G., Ala-Korpela, M., Järvelin, M.R., Männikkö, M., Laivuori, H., Dutcher, S.K., Stitziel, N.O., Wilson, R.K., Hall, I.M., Sabatti, C., Palotie, A., Salomaa, V., Laakso, M., Ripatti, S., Boehnke, M., Freimer, N.B., y Project, F. Exome sequencing of Finnish isolates enhances rare-variant association power. *Nature* **572**(7769):323–328 (2019)

Lu, Y., Genschoreck, T., Mallick, S., Ollmann, A., Patterson, N., Zhan, Y., Webster, T., y Reich, D. Application Brief A SNP array for human population genetics studies. Informe Técnico Table 1, Affymetrix (S.F.)

Mafessoni, F., Grote, S., de Filippo, C., Slon, V., Kolobova, K.A., Viola, B., Markin, S.V., Chintalapati, M., Peyrégne, S., Skov, L., Skoglund, P., Krivoshapkin, A.I., Derevianko, A.P., Meyer, M., Kelso, J., Peter, B., Prüfer, K., y Pääbo, S. A high-coverage Neandertal genome from Chagyrskaya Cave. *Proceedings of the National Academy of Sciences* **117**(26):15132–15136 (2020)

Malaspinas, A.S., Malaspinas, O., Evans, S.N., y Slatkin, M. Estimating Allele Age and Selection Coefficient from Time-Serial Data. *Genetics* **192**(2):599–607 (2012)

Mallick, S., Micco, A., Mah, M., Ringbauer, H., Lazaridis, I., Olalde, I., Patterson, N., y Reich, D. The Allen Ancient DNA Resource (AADR): A curated compendium of ancient human genomes. *bioRxiv* (2023)

Marciniak, S. y Perry, G.H. Harnessing ancient genomes to study the history of human adaptation. *Nature Reviews Genetics* **18**(11):659–674 (2017)

Mathieson, I. Human adaptation over the past 40,000 years. *Current opinion in genetics and development* **62**:97–104 (2020)

Mathieson, I. y McVean, G. Estimating selection coefficients in spatially structured populations from time series data of allele frequencies. *Genetics* **193**(3):973–984 (2013)

Mathieson, I., Lazaridis, I., Rohland, N., Mallick, S., Patterson, N., Roodenberg, S.A., Harney, E., Stewardson, K., Fernandes, D., Novak, M., Sirak, K., Gamba, C., Jones, E.R., Llamas, B., Dryomov, S., Pickrell, J., Arsuaga, J.L., de Castro, J.M.B., Carbonell, E., Gerritsen, F., Khokhlov, A., Kuznetsov, P., Lozano, M., Meller, H., Mochalov, O., Moiseyev, V., Guerra, M.A.R., Roodenberg, J., Vergès, J.M., Krause, J., Cooper, A., Alt, K.W., Brown, D., Anthony, D., Lalueza-Fox, C., Haak, W., Pinhasi, R., y Reich, D. Genome-wide patterns of selection in 230 ancient Eurasians. *Nature* **528**(7583):499–503 (2015)

Mathieson, S. y Mathieson, I. FADS1 and the timing of human adaptation to agriculture. *Molecular biology and evolution* **35**(12):2957–2970 (2018)

McDonald, J.H. y Kreitman, M. Adaptive protein evolution at the Adh locus in Drosophila. *Nature* **351**(6328):652–654 (1991)

McEvoy, B.P., Montgomery, G.W., McRae, A.F., Ripatti, S., Perola, M., Spector, T.D., Cherkas, L., Ahmadi, K.R., Boomsma, D., Willemsen, G., Hottenga, J.J., Pedersen, N.L., Magnusson, P.K.E., Kyvik, K.O., Christensen, K., Kaprio, J., Heikkilä, K., Palotie, A., Widen, E., Muilu, J., Syvänen, A.C., Liljedahl, U., Hardiman, O., Cronin, S., Peltonen, L., Martin, N.G., y Visscher, P.M. Geographical structure and differential natural selection among North European populations. *Genome research* **19**(5):804–814 (2009)

Meyer, M., Kircher, M., Gansauge, M.T., Li, H., Racimo, F., Mallick, S., Schraiber, J.G., Jay, F., Prüfer, K., de Filippo, C., Sudmant, P.H., Alkan, C., Fu, Q., Do, R., Rohland, N., Tandon, A., Siebauer, M., Green, R.E., Bryc, K., Briggs, A.W., Stenzel, U., Dabney, J., Shendure, J., Kitzman, J., Hammer, M.F., Shunkov, M.V., Derevianko, A.P., Patterson, N., Andrés, A.M., Eichler, E.E., Slatkin, M., Reich, D., Kelso, J., y Pääbo, S. A High-Coverage Genome Sequence from an Archaic Denisovan Individual. *Science* **338**(6104):222–226 (2012)

Miga, K.H. y Wang, T. The Need for a Human Pangenome Reference Sequence. *Annual review of genomics and human genetics* **22**:81–102 (2021)

Moorad, J.A. y Walling, C.A. Measuring selection for genes that promote long life in a historical human population. *Nature Ecology & Evolution* **1**(11):1773–1781 (2017)

Morgan, T.H. The origin of five mutations in eye color in drosophila and their modes of inheritance. *Science (New York, N.Y.)* **33**(849):534–537 (1911)

Muktupavela, R., Petr, M., Ségurel, L., Korneliussen, T., Novembre, J., y Racimo, F. Modelling the spatiotemporal spread of beneficial alleles using ancient genomes. *bioRxiv* pág. 2021.07.21.453231 (2021)

Nelson, M.R., Wegmann, D., Ehm, M.G., Kessner, D., St Jean, P., Verzilli, C., Shen, J., Tang, Z., Bacanu, S.A., Fraser, D., Warren, L., Aponte, J., Zawistowski, M., Liu, X., Zhang, H., Zhang, Y., Li, J., Li, Y., Li, L., Woollard, P., Topp, S., Hall, M.D., Nangle, K., Wang, J., Abecasis, G., Cardon, L.R., Zöllner, S., Whittaker, J.C., Chissoe, S.L., Novembre, J., y Mooser, V. An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science (New York, N.Y.)* **337**(6090):100–104 (2012)

Nielsen, R. Statistical tests of selective neutrality in the age of genomics. *Heredity* **86**(6):641–647 (2001)

Nielsen, R. Molecular Signatures of Natural Selection. *Annual Review of Genetics* **39**(1):197–218 (2005)

Nielsen, R. y Slatkin, M. *An Introduction to Population Genetics; Theory and Aplications.* 2ª edición. Sinauer Associates, Inc., Sunderland (2013)

Novembre, J., Galvani, A.P., y Slatkin, M. The Geographic Spread of the CCR5 Δ32 HIV-Resistance Allele. *PLOS Biology* **3**(11):e339 (2005)

Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A.R., Auton, A., Indap, A., King, K.S., Bergmann, S., Nelson, M.R., Stephens, M., y Bustamante, C.D. Genes mirror geography within Europe. *Nature* **456**(7218):98–101 (2008)

Ohta, T. Population size and rate of evolution. *Journal of Molecular Evolution* **1**(4):305–314 (1972)

Olalde, I., Brace, S., Allentoft, M.E., Armit, I., Kristiansen, K., Booth, T., Rohland, N., Mallick, S., Szécsényi-Nagy, A., Mittnik, A., Altena, E., Lipson, M., Lazaridis, I., Harper, T.K., Patterson, N., Broomandkhoshbacht, N., Diekmann, Y., Faltyskova, Z., Fernandes, D., Ferry, M., Harney, E., de Knijff, P., Michel, M., Oppenheimer, J., Stewardson, K., Barclay, A., Alt, K.W., Liesau, C., Ríos, P., Blasco, C., Miguel, J.V., García, R.M., Fernández, A.A., Bánffy, E., Bernabò-Brea, M., Billoin, D., Bonsall, C., Bonsall, L., Allen, T., Büster, L., Carver, S., Navarro, L.C., Craig, O.E., Cook, G.T., Cunliffe, B., Denaire, A., Dinwiddy, K.E., Dodwell, N., Ernée, M., Evans, C., Kuchařík, M., Farré, J.F., Fowler, C., Gazenbeek, M., Pena, R.G., Haber-Uriarte, M., Haduch, E., Hey, G., Jowett, N., Knowles, T., Massy, K., Pfrengle, S., Lefranc, P., Lemercier, O., Lefebvre, A., Martínez, C.H., Olmo, V.G., Ramírez, A.B., Maurandi, J.L., Majó, T., McKinley, J.I., McSweeney, K., Mende, B.G., Modi, A., Kulcsár, G., Kiss, V., Czene, A., Patay, R., Endrődi, A., Köhler, K., Hajdu, T., Szeniczey, T., Dani, J., Bernert, Z., Hoole, M., Cheronet, O., Keating, D., Velemínský, P., Dobeš, M., Candilio, F., Brown, F., Fernández, R.F., Herrero-Corral, A.M., Tusa, S., Carnieri, E., Lentini, L., Valenti, A., Zanini, A., Waddington, C., Delibes, G., Guerra-Doce, E., Neil, B., Brittain, M., Luke, M., Mortimer, R., Desideri, J., Besse, M., Brücken, G., Furmanek, M., Hałuszko, A., Mackiewicz, M., Rapiński, A., Leach, S., Soriano, I., Lillios, K.T., Cardoso, J.L., Pearson, M.P., Włodarczak, P., Price, T.D., Prieto, P., Rey, P.J., Risch, R., Rojo Guerra, M.A., Schmitt, A., Serralongue, J., Silva, A.M., Smrčka, V., Vergnaud,

L., Zilhão, J., Caramelli, D., Higham, T., Thomas, M.G., Kennett, D.J., Fokkens, H., Heyd, V., Sheridan, A., Sjögren, K.G., Stockhammer, P.W., Krause, J., Pinhasi, R., Haak, W., Barnes, I., Lalueza-Fox, C., y Reich, D. The Beaker phenomenon and the genomic transformation of northwest Europe. *Nature* **555**(7695):190–196 (2018)

Parsch, J., Novozhilov, S., Saminadin-Peter, S.S., Wong, K.M., y Andolfatto, P. On the utility of short intron sequences as a reference for the detection of positive and negative selection in Drosophila. *Molecular biology and evolution* **27**(6):1226–1234 (2010)

Pauling, L. y Zuckerkandl, E. Molecular disease, evolution and genetic heterogeneity. En M. Kasha y B. Pullman (editores), *Horizons in Biochemistry*, págs. 189–225. Academic Press (1962)

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., y Duchesnay, E. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**:2825–2830 (2011)

Peter, B.M., Petkova, D., y Novembre, J. Genetic Landscapes Reveal How Human Genetic Diversity Aligns with Geography. *Molecular Biology and Evolution* **37**(4):943–951 (2020)

Pickrell, J.K. y Reich, D. Toward a new history and geography of human genes informed by ancient DNA. *Trends in Genetics* **30**(9):377–389 (2014)

Primack, R.B. y Kang, H. Measuring Fitness and Natural Selection in Wild Plant Populations. *Annual Review of Ecology and Systematics* **20**(1):367–396 (1989)

Prüfer, K., Racimo, F., Patterson, N., Jay, F., Sankararaman, S., Sawyer, S., Heinze, A., Renaud, G., Sudmant, P.H., de Filippo, C., Li, H., Mallick, S., Dannemann, M., Fu, Q., Kircher, M., Kuhlwilm, M., Lachmann, M., Meyer, M., Ongyerth, M., Siebauer, M., Theunert, C., Tandon, A., Moorjani, P., Pickrell, J., Mullikin, J.C., Vohr, S.H., Green, R.E., Hellmann, I., Johnson, P.L.F., Blanche, H., Cann, H., Kitzman, J.O., Shendure, J., Eichler, E.E., Lein, E.S., Bakken, T.E., Golovanova, L.V., Doronichev, V.B., Shunkov, M.V., Derevianko, A.P., Viola, B., Slatkin, M., Reich, D., Kelso, J., y Pääbo, S. The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505**(7481):43–49 (2014)

Prüfer, K., de Filippo, C., Grote, S., Mafessoni, F., Korlević, P., Hajdinjak, M., Vernot, B., Skov, L., Hsieh, P., Peyrégne, S., Reher, D., Hopfe, C., Nagel, S., Maricic, T., Fu, Q., Theunert, C., Rogers, R., Skoglund, P., Chintalapati, M., Dannemann, M., Nelson, B.J., Key, F.M., Rudan, P., Kućan, Ž., Gušić, I., Golovanova, L.V., Doronichev, V.B., Patterson, N., Reich, D., Eichler, E.E., Slatkin, M., Schierup, M.H., Andrés, A.M., Kelso, J., Meyer, M., y Pääbo, S. A high-coverage Neandertal genome from Vindija Cave in Croatia. *Science* **358**(6363):655–658 (2017)

Racimo, F. y Schraiber, J.G. Approximation to the Distribution of Fitness Effects across Functional Categories in Human Segregating Polymorphisms. *PLOS Genetics* **10**(11):e1004697 (2014)

Racimo, F., Woodbridge, J., Fyfe, R.M., Sikora, M., Sjögren, K.G., Kristiansen, K., y Linden, M.V. The spatiotemporal spread of human migrations during the European Holocene. *Proceedings of the National Academy of Sciences of the United States of America* **117**(16):8989–9000 (2020)

Rasmussen, M., Li, Y., Lindgreen, S., Pedersen, J.S., Albrechtsen, A., Moltke, I., Metspalu, M., Metspalu, E., Kivisild, T., Gupta, R., Bertalan, M., Nielsen, K., Gilbert, M.T.P., Wang, Y., Raghavan, M., Campos, P.F., Kamp, H.M., Wilson, A.S., Gledhill, A., Tridico, S., Bunce, M., Lorenzen, E.D., Binladen, J., Guo, X., Zhao, J., Zhang, X., Zhang, H., Li, Z., Chen, M., Orlando, L., Kristiansen, K., Bak, M., Tommerup, N., Bendixen, C., Pierre, T.L., Grønnow, B., Meldgaard, M., Andreasen, C., Fedorova, S.A., Osipova, L.P., Higham, T.F.G., Ramsey, C.B., Hansen, T.v.O., Nielsen, F.C., Crawford, M.H., Brunak, S., Sicheritz-Pontén, T., Villems, R., Nielsen, R., Krogh, A., Wang, J., y Willerslev, E. Ancient human genome sequence of an extinct Palaeo-Eskimo. *Nature* **463**(7282):757–762 (2010)

Reich, D., Green, R.E., Kircher, M., Krause, J., Patterson, N., Durand, E.Y., Viola, B., Briggs, A.W., Stenzel, U., Johnson, P.L.F., Maricic, T., Good, J.M., Marques-Bonet, T., Alkan, C., Fu, Q., Mallick, S., Li, H., Meyer, M., Eichler, E.E., Stoneking, M., Richards, M., Talamo, S., Shunkov, M.V., Derevianko, A.P., Hublin, J.J., Kelso, J., Slatkin, M., y Pääbo, S. Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* **468**(7327):1053–1060 (2010)

Rentzsch, P., Schubach, M., Shendure, J., y Kircher, M. CADD-Splice—improving genome-wide variant effect prediction using deep learning-derived splice scores. *Genome Medicine* **13**(1):31 (2021)

Rowlands, C., Thomas, H.B., Lord, J., Wai, H.A., Arno, G., Beaman, G., Sergouniotis, P., Gomes-Silva, B., Campbell, C., Gossan, N., Hardcastle, C., Webb, K., O'Callaghan, C., Hirst, R.A., Ramsden, S., Jones, E., Clayton-Smith, J., Webster, A.R., Ambrose, J.C., Arumugam, P., Bevers, R., Bleda, M., Boardman-Pretty, F., Boustred, C.R., Brittain, H., Caulfield, M.J., Chan, G.C., Fowler, T., Giess, A., Hamblin, A., Henderson, S., Hubbard, T.J.P., Jackson, R., Jones, L.J., Kasperaviciute, D., Kayikci, M., Kousathanas, A., Lahnstein, L., Leigh, S.E.A., Leong, I.U.S., Lopez, F.J., Maleady-Crowe, F., McEntagart, M., Minneci, F., Moutsianas, L., Mueller, M., Murugaesu, N., Need, A.C., O'Donovan, P., Odhams, C.A., Patch, C., Perez-Gil, D., Pereira, M.B., Pullinger, J., Rahim, T., Rendon, A., Rogers, T., Savage, K., Sawant, K., Scott, R.H., Siddiq, A., Sieghart, A., Smith, S.C., Sosinsky, A., Stuckey, A., Tanguy, M., Taylor Tavares, A.L., Thomas, E.R.A., Thompson, S.R., Tucci, A., Welland, M.J., Williams, E., Witkowsa, K., Wood, S.M., Douglas, A.G.L., O'Keefe, R.T., Newman, W.G., Baralle, D., Black, G.C.M., Ellingford, J.M., y Consortium, G.E.R. Comparison of in silico strategies to prioritize rare genomic variants impacting RNA splicing for the diagnosis of genomic disorders. *Scientific Reports* **11**(1):20607 (2021)

Sabeti, P.C., Schaffner, S.F., Fry, B., Lohmueller, J., Varilly, P., Shamovsky, O., Palma, A., Mikkelsen, T.S., Altshuler, D., y Lander, E.S. Positive natural selection in the human lineage. *Science (New York, N.Y.)* **312**(5780):1614–1620 (2006)

Scheiner, S.M. y Callahan, H.S. Measuring natural selection on phenotypic plasticity. *Evolution* **53**(6):1704–1713 (1999)

Schraiber, J.G., Evans, S.N., y Slatkin, M. Bayesian Inference of Natural Selection from Allele Frequency Time Series. *Genetics* **203**(1):493–511 (2016)

Skoglund, P. y Mathieson, I. Ancient Genomics of Modern Humans: The First Decade. *Annual Review of Genomics and Human Genetics* **19**(1):381–404 (2018)

Smocovitis, V.B. Unifying biology: The evolutionary synthesis and evolutionary biology. *Journal of the History of Biology* **25**(1):1–65 (1992)

Steinrücken, M., Bhaskar, A., y Song, Y.S. A novel spectral method for inferring general diploid selection from time series genetic data. *The annals of applied statistics* **8**(4):2203–2222 (2014)

Tajima, F. Evolutionary relationship of DNA sequences in finite populations. *Genetics* **105**(2):437–460 (1983)

Tajima, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**(3):585–595 (1989)

Takahata, N., Ishii, K., y Matsuda, H. Effect of temporal fluctuation of selection coefficient on gene frequency in a population. *Proceedings of the National Academy of Sciences* **72**(11):4541–4545 (1975)

Timmermann, A. Quantifying the potential causes of Neanderthal extinction: Abrupt climate change versus competition and interbreeding. *Quaternary Science Reviews* **238**:106331 (2020)

Toomajian, C., Ajioka, R.S., Jorde, L.B., Kushner, J.P., y Kreitman, M. A method for detecting recent selection in the human genome from allele age estimates. *Genetics* **165**(1):287–297 (2003)

van der Velde, K.J., Kuiper, J., Thompson, B.A., Plazzer, J.P., van Valkenhoef, G., de Haan, M., Jongbloed, J.D., Wijmenga, C., de Koning, T.J., Abbott, K.M. *et al.* Evaluation of CADD scores in curated mismatch repair gene variants yields a model for clinical validation and prioritization. *Human mutation* **36**(7):712–719 (2015)

van der Velde, K.J., de Boer, E.N., van Diemen, C.C., Sikkema-Raddatz, B., Abbott, K.M., Knopperts, A., Franke, L., Sijmons, R.H., de Koning, T.J., Wijmenga, C. *et al.* GAVIN: Gene-Aware variant interpretation for medical sequencing. *Genome Biology* **18**(1):1–10 (2017)

Van Valen, L. Selection in Natural Populations. III. Measurement and Estimation. *Evolution* **19**(4):514–528 (1965)

Vitalis, R., Gautier, M., Dawson, K.J., y Beaumont, M.A. Detecting and Measuring Selection from Gene Frequency Data. *Genetics* **196**(3):799–817 (2014)

Watterson, G.A. Heterosis or neutrality? *Genetics* **85**(4):789–814 (1977)

Weir, B.S. y Cockerham, C.C. Testing hypotheses about linkage disequilibrium with multiple alleles. *Genetics* **88**(3):633–642 (1978)

Wright, S. y Chambers, J. *Principles of Livestock Breeding.* CreateSpace Independent Publishing Platform (1920)

Wright, S. On the Roles of Directed and Random Changes in Gene Frequency in the Genetics of Populations. *Evolution* **2**(4):279–294 (1948)

Ye, K. y Gu, Z. Recent advances in understanding the role of nutrition in human genome evolution. *Advances in nutrition (Bethesda, Md.)* **2**(6):486–496 (2011)

Zietemann, V., Kröger, J., Enzenbach, C., Jansen, E., Fritsche, A., Weikert, C., Boeing, H., y Schulze, M.B. Genetic variation of the FADS1 FADS2 gene cluster and n-6 PUFA composition in erythrocyte membranes in the European Prospective Investigation into Cancer and Nutrition-Potsdam study. *British Journal of Nutrition* **104**(12):1748–1759 (2010)