BENEMÉRITA UNIVERSIDAD AUTÓNOMA DE PUEBLA

FACULTAD DE CIENCIAS BIOLOGÍA

"DETECCIÓN DE SEÑALES DE SELECCIÓN EN GENOMAS COMPLETOS DE NATIVOS AMERICANOS"

T E S I S

PRESENTADA PARA OBTENER EL TÍTULO DE:

LICENCIADO (A) EN BIOTECNOLOGÍA

PRESENTA:
MARÍA FERNANDA MIRÓN TORUÑO

DIRECTOR: DR. ISRAEL AGUILAR ORDOÑEZ
CO-DIRECTORA: DRA WENDY ARGELIA GARCÍA SUASTEGUI

JUNIO 2022

BIO TECNOLOGÍA

A todas las comunidades Nativas Mexicanas,
por permitirme entender la vida
a través de su genoma.

AGRADECIMIENTOS

A mi mamá María Guadalupe Toruño, por su amor y apoyo incondicional a lo largo de toda mi vida. A mis hermanos Ingrid Mirón, Marilupe Mirón y Luis Mirón por motivarme día con día, inspirarme a ser mejor persona e interesarse por los temas que me apasionan. A mi abuela Guadalupe Montaño, por preocuparse por mí en cada día de desvelo e infundirme valor en cada momento. A mi abuelo Gonzalo Toruño, por jugar el rol de papá en mi vida. A mi familia, por estar ahí.

Al M.C Israel Aguilar Ordóñez, por transmitir su pasión por la bioinformática, por brindarme oportunidades que soñaba imposibles, por su tiempo cada reunión de cada semana y por creer en el potencial de cada miembro del laboratorio.

A la Dra. Wendy García Suastegui, por ser una increíble docente, por asesorarme en cada paso de este proceso y retroalimentar este trabajo. Al Dr. Salvador Galicia Isasmendi, por nutrir mi conocimiento y amor por la evolución y aceptar ser revisor de este proyecto. Al Dr. Luis Ramiro Caso, por fundamentar mis bases de conocimiento en genética y aceptar ser revisor de este trabajo. Al Dr. Austin Reynolds, por mejorar mi entendimiento de la genómica de poblaciones y orientarme en el desarrollo teórico de este proyecto. Al Dr. Juan Enrique Morett Sánchez, por ser un ejemplo de humildad y confiar ampliamente en mis capacidades.

Al Instituto Nacional de Medicina Genómica (INMEGEN), la Benemérita Universidad Autónoma de Puebla (BUAP) y al Consejo Nacional de Ciencia y Tecnología (CONACYT) por brindarme herramientas para el aprendizaje y autorrealización académica.

A José Eduardo García, por ser el mejor compañero bioinformático y un increíble amigo, por todos los aprendizajes juntos, por todas las fiestas y por su ayuda cada madrugada para resolver problemas de código. A Paulina Pérez y Josué Guzmán por ser los mejores compañeros de laboratorio, por las sesiones de música, por los juegos de mesa y por su apoyo incondicional en cada proyecto.

A Juan Pablo Vernet, por ser un cuarto hermano, por darme su opinión honesta en el desarrollo de cada gráfico, por las veces que nos quedamos trabajando en casa y por su apoyo absoluto en cada decisión de mi vida. A mis amigos Chiara Blanno, César Juárez, Christopher Rivera, Alan Santiago, Nayma García por hacer mi vida mucho más feliz.

A cada uno de los 95 individuos Nativos Mexicanos pertenecientes al proyecto 100G-MX. Sin ellos, nada de esto sería posible.

ÍNDICE GENERAL

1. INTRODUCCIÓN 8
2. ANTECEDENTES 10
2.1 TIPOS DE SELECCIÓN NATURAL. 10
2.2 IDENTIFICANDO POR PRIMERA VEZ EL IMPACTO DE LA SELECCIÓN EN HUMANOS. 12
2.4 MÉTODOS MICROEVOLUTIVOS: FUNDAMENTO 13
2.4.1 MÉTODOS MICROEVOLUTIVOS: BASADOS EN ESPECTROS DE FRECUENCIAS 15
2.4.2 MÉTODOS MICROEVOLUTIVOS: BASADOS EN DIFERENCIACIÓN POBLACIONAL 16
2.4.3 MÉTODOS MICROEVOLUTIVOS: BASADOS EN DESEQUILIBRIO DE LIGAMIENTO. 17
2.5 HIPÓTESIS DE SELECCIÓN EN DIVERSAS DISTRIBUCIONES GEOGRÁFICAS 18
2.5.1 HIPÓTESIS DE SELECCIÓN EN POBLACIONES EUROPEAS 19
2.5.2 HIPÓTESIS DE SELECCIÓN EN POBLACIONES AFRICANAS. 21
2.5.4 HIPÓTESIS DE SELECCIÓN EN POBLACIONES AMERICANAS. 23
3. MARCO CONCEPTUAL 24
3.1 DESCRIPCIÓN DE POPULATION BRANCH STATISTIC (PBS). 24
3.1.1 DESCRIPCIÓN DE PBS: MODELADO DEMOGRÁFICO CON FASTSIMCOAL2 26
3.2 DESCRIPCIÓN DE INTEGRATED HAPLOTYPE SCORE (iHS) 27
3.2.1 DESCRIPCIÓN DE iHS: FASEADO DE HAPLOTIPOS. 29
3.2.2 DESCRIPCIÓN DE iHS: HERRAMIENTAS BIOINFORMÁTICAS 29
3.3 DESCRIPCIÓN DE iHS VS PBS 30
4. PLANTEAMIENTO DEL PROBLEMA 31
5. HIPÓTESIS 32
6. OBJETIVOS 32
7. MATERIAL Y MÉTODOS 33
7.1 WORKFLOW BIOINFORMÁTICO 33
7.2 OPERACIÓN DEL PIPELINE 33
7.2.1 OPERACIÓN DEL PIPELINE: iHS 35
7.2.2 OPERACIÓN DEL PIPELINE: PBS 37
7.2.3 OPERACIÓN DEL PIPELINE: IHS VS PBS 38
7.3 DETECCIÓN DE SEÑALES DE SELECCIÓN EN GENOMAS COMPLETOS DE NATIVOS AMERICANOS. 39
7.3.1 MUESTRAS. 39
7.3.2 DISEÑO EXPERIMENTAL. 39
7.3.3 PRETRATAMIENTO DE LOS DATOS. 40
8. RESULTADOS. 41
8.1 RESULTADOS: PBS. 41
8.2 RESULTADOS: iHS 43
8.3 RESULTADOS: PBS VERSUS iHS 45

9. DISCUSIÓN 49
10. CONCLUSIONES 52
11. REFERENCIAS 53
12. SOPORTE GRÁFICO 67

ÍNDICE DE FIGURAS

Figura 1. Tipos de selección y su impacto en la variabilidad genética 10

Figura 2. ¿Qué es un barrido selectivo? 14

Figura 3. Fundamento de los métodos basados en espectros de frecuencias. 15

Figura 4. Fundamento de los métodos basados en diferenciación poblacional 17

Figura 5. Ejemplos de genes bajo presión selectiva local 20

Figura 6. Descripción gráfica del estadístico EHH 27

Figura 7. Decremento de EHH en datos simulados 28

Figura 8. Workflow esquemático del pipeline desarrollado para detectar señales de selección. Branch iHS. 36

Figura 9. Workflow esquemático del pipeline desarrollado para detectar señales de selección. Branch PBS por variante genética. 37

Figura 10. Workflow esquemático del pipeline desarrollado para detectar señales de selección. Branch iHS vs PBS por variante genética. 38

Figura 11. Manhattan plot PBS 41

Figura 12. Histograma y spiderplot PBS 42

Figura 13. Manhattan plot iHS 43

Figura 14. Histograma de valores de iHS por variante core. 44

Figura 15. Circus plot 45

Figura 16. Clasificación ontológica de genes bajo selección 46

Figura 17. Clasificación ontológica de genes bajo selección. 15 categorías 47

ÍNDICE DE TABLAS

Tabla 1. Métodos utilizados para la identificación de señales de selección. 14

Tabla 2. Softwares utilizados en pipeline de Nextflow. 34

CSV 1. Input iHS 33

CSV 2. Input PBS 34

ABSTRACT

Over the past few years, the considerable increase in computational power and the development of next-generation sequencing technologies have given way to the opportunity to identify genomic regions that appear to be shaped by natural selection. Although detection of signatures of selection shed light on the understanding of evolution and genetic structure, the required use of statistical and bioinformatics methods represents a barrier for scientists unfamiliar with whole-genome data manipulation. Here we provide a brief overview of a simple Nextflow pipeline for detecting recent evidence of selection in whole-genome data through Population Branch Statistic (PBS) and Integrated Haplotype Score (iHS). To verify the effectiveness of the developed tool, a selection analysis was carried out with data belonging to 76 Native American individuals. Of interest, we find evidence of adaptation to pathogenic environments in populations of the central region of Mexico. Code and brief manual (from installation to testing run) of the developed pipeline are publicly available from the following Github repository:
https://github.com/fernanda-miron/nf-selection

RESUMEN

A través de los últimos años, el aumento considerable del poder computacional y el desarrollo de tecnologías de secuenciación de nueva generación, han permitido identificar regiones del genoma que exhiben evidencia selectiva. La detección de señales de selección contribuye en la generación de nuevas asociaciones genotipo-fenotipo y mejora el entendimiento de la estructura genética poblacional. Sin embargo, el uso requerido de métodos estadísticos y bioinformáticos representa una barrera para los científicos poco familiarizados con la manipulación de datos genómicos masivos. Este trabajo presenta el desarrollo de un pipeline bioinformático para la detección de señales de selección con Population Branch Statistic (PBS) e Integrated Haplotype Score (iHS). Para verificar la efectividad de la herramienta, se llevó a cabo un análisis de selección con datos pertenecientes a 76 individuos Nativos Americanos. Notablemente, se identificó evidencia de adaptación a ambientes patogénicos en poblaciones de la región central de México. El código para instalar y la documentación para usar la herramienta desarrollada se encuentran públicamente disponibles en el siguiente repositorio GitHub:
https://github.com/fernanda-miron/nf-selection

1. INTRODUCCIÓN

En 1858 Charles Darwin y Alfred Russel Wallace describieron por primera vez y de forma independiente el principio de la selección natural [1]. Esta teoría revolucionaria, publicada en el libro "On the Origin of Species by Means of Natural Selection", articulaba la idea que aquellos rasgos benéficos que aumentaban las oportunidades de supervivencia y reproducción de un organismo, tendían a volverse más frecuentes a lo largo del tiempo [2,3]. Algunos años después, el concepto darwiniano de selección natural fue combinado con el redescubrimiento de las leyes Mendelianas para definir a la selección como cualquier variación no aleatoria en la propagación de un alelo, como consecuencia de su efecto fenotípico [4,5]. La redefinición del concepto de selección desde una perspectiva genética dió paso a la posibilidad de estudiar e identificar el impacto de la misma al nivel más fundamental: el genoma.

A través de los últimos años, el aumento considerable del poder computacional y el desarrollo constante de las tecnologías de secuenciación de nueva generación han permitido identificar patrones característicos de selección natural en el material genético de diversas especies y organismos (también llamado detección de señales de selección) [6]. Con ello, la comunidad científica ha comenzado a comprender no sólo la evolución y adaptación de diversas poblaciones, sino la distribución geográfica de rasgos y enfermedades [7,8,9].

Si bien la detección de señales de selección ha mejorado el entendimiento de la evolución y la estructura genética de las especies, el uso obligado de métodos estadísticos y bioinformáticos representa una barrera para los científicos no familiarizados con la manipulación de datos genómicos masivos. Para superar este problema, se han desarrollado diversos pipelines[1] bioinformáticos que permiten simplificar la obtención de hipótesis de selección a partir de datos de genotipificación. Muestra de esto es la herramienta “Selection Browser 1.0” que tiene por objeto determinar el impacto de la selección natural en datos pertenecientes al proyecto 1000 Genomas [10]. Asimismo, Murray Cadzow y

[1] Un pipeline es una práctica informática en la que uno o varios conjuntos de datos se modifican a través de una serie de procesos. Los procesos suelen ser secuenciales y la salida de un proceso es la entrada de otro.

colaboradores desarrollaron "Selection Tools", un pipeline diseñado para identificar señales de selección mediante la ejecución de diversos análisis [8]. A pesar de que los recursos previamente citados permiten la detección simplificada de señales de selección, ambos se encuentran basados en los mismos métodos estadísticos - Tajima's D, CLR, Fay and Wu's H, XPEHH, iHH, iHS, F_{ST}, DAF, and XPCLR - dejando así el panorama global incompleto.

El Population Branch Statistic (PBS) es un método basado en diferenciación poblacional con gran capacidad para detectar selección natural reciente [11]. En los últimos años, ha sido ampliamente utilizado para detectar regiones del genoma bajo selección en diversas poblaciones [12-15]. De igual forma, el Integrated Haplotype Score (iHS) es un método basado en desequilibrio de ligamiento que permite la detección de señales de selección en una escala temporal cercana [16]. Ambos métodos han sido utilizados en complementariedad a lo largo de diversas investigaciones [12,13]. A pesar de esto, no existen herramientas bioinformáticas públicas que permitan calcular PBS e iHS en conjunto y de forma automatizada.

Esta tesis presenta el desarrollo de un pipeline bioinformático destinado a la detección completamente reproducible de señales de selección con PBS e iHS. La herramienta desarrollada lleva a cabo el cómputo de los estadísticos previamente citados en datos de genotipificación de genoma completo pertenecientes a cualquier especie diploide — con enfoque particular en la especie *Homo Sapiens* —. Asimismo, para validar el desarrollo bioinformático con datos reales, se llevó a cabo la identificación de regiones bajo selección con 76 genomas completos pertenecientes a 27 comunidades Nativas Mexicanas. Notablemente, se identificaron señales de selección en *CD164*, un gen relacionado con la respuesta antiretroviral innata con un papel específico en las infecciones provocadas por el virus VIH. Asimismo, se identificaron señales de selección nuevas —no reportadas previamente en la bibliografía— en los genes *CXCL9, ARMC6, SOX6, MSH3, ACVR2A, ESCO2 y RAB3C*, los cuales juegan un papel fundamental en el funcionamiento del sistema inmune. Lo anteriormente citado evidencia procesos adaptativos a ambientes patogénicos, un fenotipo previamente descrito en poblaciones Nativas Mexicanas, y también en otras regiones del mundo.

2. ANTECEDENTES

2.1 TIPOS DE SELECCIÓN NATURAL.

La selección natural opera en una amplia variedad de patrones que impactan de forma diferente en la estructura genética de las poblaciones. Dentro de su categorización, es común encontrar a la selección direccional, la selección estabilizadora y la selección disruptiva [17,2]. La selección direccional ocurre cuando el fenotipo promedio de una población "cambia" en una dirección específica; cuando un alelo (variaciones genéticas) se encuentra favorecido se habla de selección direccional positiva, mientras que un alelo no favorecido hace referencia a la selección direccional negativa. La selección direccional tiende a reducir la diversidad genética en las poblaciones al promover la fijación (frecuencia alélica igual a 1.0) de los alelos favorables y la pérdida (frecuencia alélica igual a 0.0) de los alelos desfavorables [17,2] (**Figura 1A**). Un ejemplo clásico de selección direccional lo constituye el caso de los picos de los pinzones en las islas Galápagos estudiados por Charles Darwin, Peter y Rosemary Grant [18].

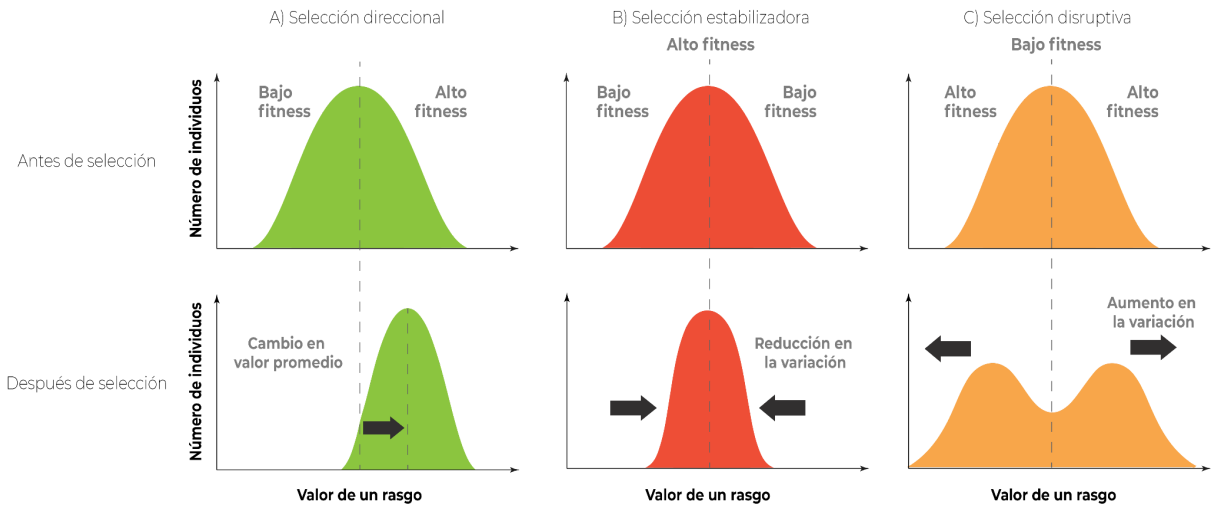


Fig. 1 Tipos de selección y su impacto en la variabilidad genética. **A)** Selección direccional. **B)** Selección estabilizadora. **C)** Selección disruptiva. Adaptado de Freeman, S. (2008)

Por otro lado, la selección estabilizadora ocurre cuando se ven favorecidos valores fenotípicos intermedios y por consecuencia se desfavorecen los valores que se encuentran en los extremos de la distribución (**Figura 1B**). La selección

estabilizadora tiende a reducir la variación genética poblacional sin cambiar el valor promedio de un rasgo a lo largo del tiempo. Asimismo, sus mecanismos pueden abarcar la presencia de alelos codominantes o alelos que implican fenotipos intermedios bajo presión selectiva positiva [17,2]. Un ejemplo característico de selección estabilizadora fue identificado en los hospitales de Reino Unido al reportar que aquellos bebés con un peso promedio tenían mayor probabilidad de sobrevivir. Caso contrario, los bebés con pesos pequeños y pesos muy elevados poseían una mayor tasa de mortalidad [19].

En contraste con la selección estabilizadora, la selección disruptiva favorece aquellos fenotipos que se encuentran en los extremos de la distribución y que tienden a ser opuestos (**Figura 1C**). Este tipo de selección es también conocida como “selección diversificadora” porque provoca el aumento en la variación genética de una población [17,2]. Un ejemplo distintivo de selección estabilizadora se ha observado en los pinzones cascanueces de vientre negro originarios de Camerún. En este caso, los pinzones con picos extremadamente cortos o largos poseen mejores tasas de supervivencia que aquellos con fenotipos intermedios. El efecto causal de la selección se ha atribuido a la presencia exclusiva de semillas muy pequeñas o muy grandes que pueden ser consumidas de mejor forma con picos muy cortos o muy largos de forma respectiva. En algunos casos, la selección disruptiva puede desencadenar eventos de especiación [20].

A pesar de la gran diversidad de patrones de selección reportados en la literatura, un gran porcentaje de la investigación evolutiva se ha enfocado en el desarrollo de métodos genómicos y estadísticos dedicados exclusivamente a identificar selección direccional positiva. Esta última afirmación se ha sustentando en la observación de huellas genómicas más características cuando los individuos son sujetos a eventos de selección positiva [2, 9].

2.2 IDENTIFICANDO POR PRIMERA VEZ EL IMPACTO DE LA SELECCIÓN EN HUMANOS.

Durante años, entender el impacto de la selección en las poblaciones humanas expuestas a diversos ambientes ha capturado el interés de la comunidad científica. El biólogo J.B.S Haldane fue el primer científico en exponer la selección reciente en

la especie *Homo Sapiens* al describir la posibilidad de que los individuos africanos heterocigotos para talasemia — trastorno sanguíneo hereditario que ocurre cuando el cuerpo no produce la cantidad necesaria de hemoglobina [21] — podían ser más resistentes a la infección por malaria [9]. La teoría de Haldane fue confirmada años después por A.C Allison al demostrar que las mutaciones en el gen que codifica la *Hemoglobina-B* eran el objetivo molecular de la selección y provocaban resistencia en las poblaciones con malaria endémica [22].

La capacidad de identificar el impacto de la selección a nivel genético (sustentado por las investigaciones de Allison), supuso un parteaguas para el desarrollo de metodologías y estadísticos [23,24,25] enfocados en dilucidar el papel de los genes en fenotipos con hipótesis previas de adaptación [2,26]. A pesar del gran potencial que ofrecía la metodología fenotipo adaptativo-genotipo bajo selección, la identificación de loci bajo esta estrategia enfrentaba un gran desafío; las hipótesis de selección dependían directamente de las observaciones empíricas. Lo anterior limitaba las observaciones a los fenotipos más evidentes.

El advenimiento de las tecnologías de secuenciación de genoma completo, el desarrollo de herramientas estadísticas y el aumento considerable del poder computacional, representaron un punto de inflexión en el estudio de la selección en humanos contemporáneos. El enfoque de los estudios evolutivos pasó de una perspectiva "*hyphotesis-testing*" a una perspectiva "*hyphotesis-generating*", permitiendo la identificación de nuevas regiones del genoma bajo posible selección y reevaluando candidatos previamente propuestos [2, 27]. Ahora es posible cuestionar millones de loci en busca de señales de selección, aún sin evidencia empírica que permita sospechar.

2.3 MÉTODOS PARA LA DETECCIÓN DE SEÑALES DE SELECCIÓN.

La estructura genética de los organismos es influida por eventos demográficos que impactan directamente en la variabilidad genómica. Particularmente, la selección natural puede ser identificada por "huellas" o patrones distintivos de variación genética que son identificados a través de estadísticas resumidas. Los métodos para el estudio de la selección pueden dividirse en dos grandes ramas: métodos macroevolutivos y microevolutivos. Los métodos macroevolutivos se encuentran

basados en comparaciones interespecies con el objeto de identificar eventos selectivos que tuvieron lugar en el pasado profundo[2]. Por otro lado, los métodos microevolutivos son utilizados para identificar eventos selectivos intraespecie con una profundidad temporal reducida. El enfoque de este capítulo se centrará en su totalidad en los métodos microevolutivos.

2.4 MÉTODOS MICROEVOLUTIVOS: FUNDAMENTO

Las variantes o mutaciones bajo la influencia de la selección positiva aumentan su frecuencia en la población a través del tiempo. A medida que esto sucede, por efecto del desequilibrio de ligamiento [2], aquellos alelos cercanos a la mutación bajo selección también aumentarán su frecuencia (Genetic hitchhiking[3]).

Este fenómeno — conocido como barrido selectivo — genera una disminución en la diversidad genética que rodea al alelo causal, hasta que la recombinación o mutaciones aleatorias restauran la diversidad en el locus (**Figura 2**) [28,29].

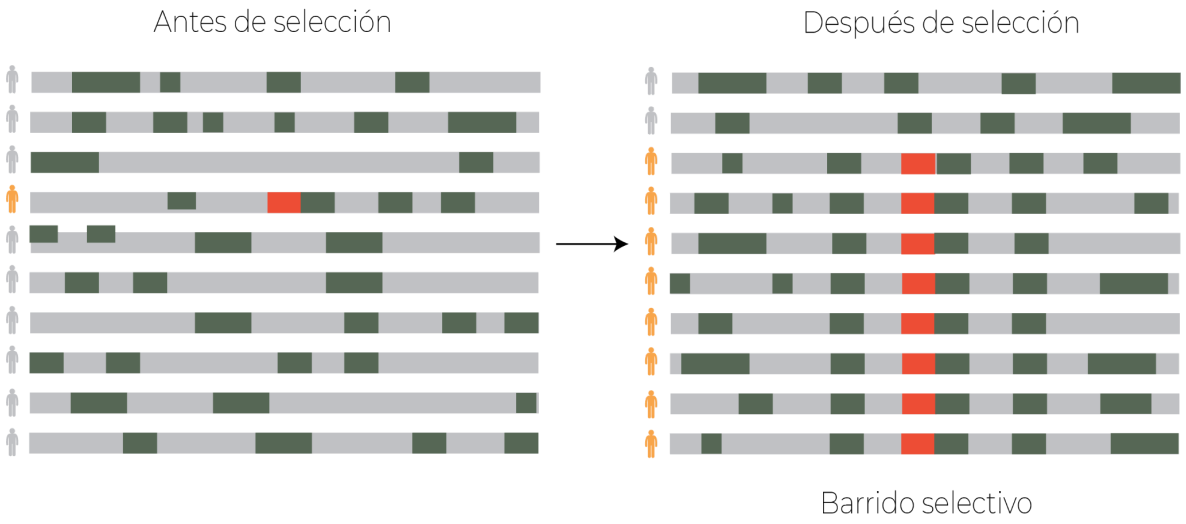


Fig. 2 ¿Qué es un barrido selectivo?. Bajo un modelo de selección natural positiva, una variante benéfica aumenta su prevalencia en la población. La figura esquematiza los polimorfismos a lo largo de un cromosoma antes y después de la selección. Conforme un nuevo alelo bajo selección (rojo)

[2] Desequilibrio de ligamiento: Asociación no aleatoria de alelos. Ocurren juntos con más frecuencia de lo que puede explicarse por casualidad debido a su proximidad física en el cromosoma.Tendencia de ciertas variantes en el mismo cromosoma a heredarse en conjunto. Recuperado de: Montgomery, 2008 [31] .

[3] Propuesto por Maynard Smith y Haigh en 1974, asume que la selección positiva opera en un único locus que se encuentra parcialmente ligado a polimorfismos neutrales. Por tanto, describe la reducción de la variación en las posiciones nucleotídicas neutrales por la fijación de un nuevo alelo benéfico. Recuperado de: Stephan 2010 [32]

alcanza frecuencias alélicas elevadas, los alelos cercanos aumentan su frecuencia en conjunto por efecto del desequilibrio de ligamiento. Adaptado de Learn Science at Scitable. “A Selective Sweep.”

Los métodos microevolutivos identifican regiones del genoma bajo selección a través de la identificación de patrones característicos de un barrido selectivo; regiones con variación reducida, cambios en el espectro de frecuencias, entre otros [30]. Los métodos microevolutivos se han clasificado en métodos basados en espectros de frecuencias, métodos basados en desequilibrio de ligamiento, y métodos basados en diferenciación poblacional [2] (**Tabla 1**).

TABLA 1 | Métodos utilizados para la identificación de señales de selección. Adaptado de: Vitti, J. J., Grossman, S. R., & Sabeti, P. C. (2013)

Acercamiento	Tipo de selección	Fundamento	Método Representativo
Métodos basados en frecuencias	Positiva, balanceadora	En un barrido selectivo, la variante bajo selección junto con otras variantes cercanas alcanzan una alta prevalencia. De este fondo homogéneo, surgen nuevos alelos que inicialmente son muy raros. Estos métodos se basan en detectar zonas con excedentes de alelos raros.	Tajima’s D and derivatives
			Fay & Wu’s H
Métodos basados en desequilibrio de ligamiento	Positiva	Los barridos selectivos hacen que una región genética tenga una alta prevalencia en una población, incluida la variante causal y sus “vecinas”. La asociación entre estas variantes define un haplotipo que persiste en la población y que reduce la diversidad genética.	EHH
			iHS
			XP-EHH
			IBD
Métodos basados en diferenciación poblacional	Positiva y negativa	Distintas poblaciones están sujetas a diferentes presiones selectivas. Si la selección actúa en un locus dentro de una población pero no en otra, las frecuencias alélicas en ese locus pueden variar significativamente.	PBS
			Fst
			LSBL

2.4.1 MÉTODOS MICROEVOLUTIVOS: BASADOS EN ESPECTROS DE FRECUENCIAS

Como se ha discutido con anterioridad, un barrido selectivo genera fondos genéticos homogéneos caracterizados por la poca diversidad genética. En este contexto, la “aparición” aleatoria de nuevas mutaciones a través del tiempo, degenera en

genotipos conformados por variantes inicialmente raras. A pesar de que el espectro de frecuencias alélicas es normalizado a través del tiempo, la presencia de variantes nuevas poco frecuentes persiste por generaciones [2] (**Figura 3**).

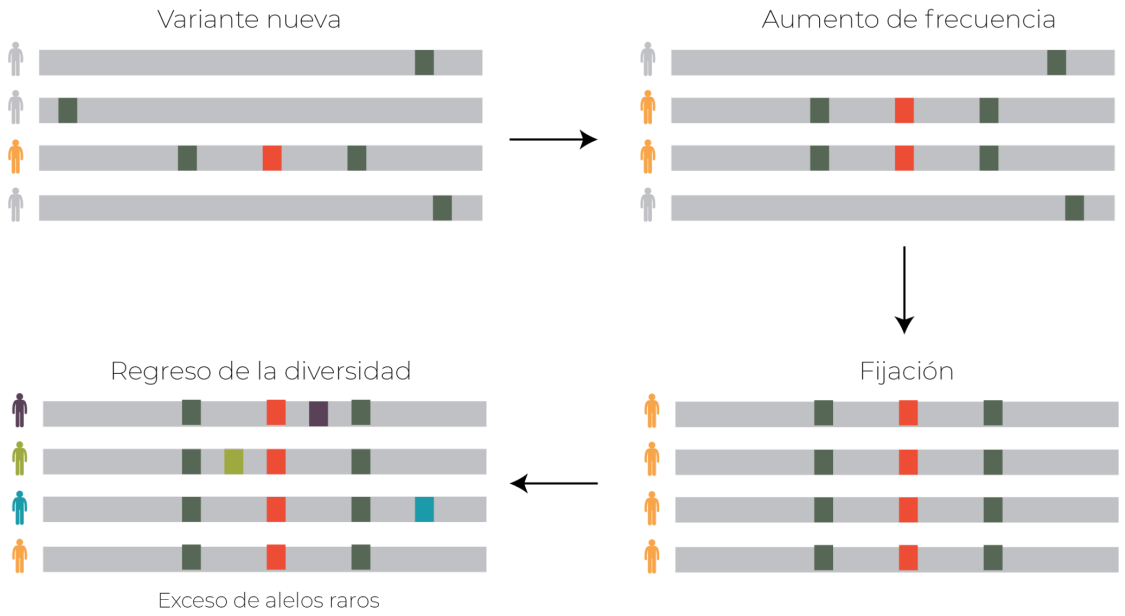


Fig. 3 Fundamento de los métodos basados en espectros de frecuencias. Un barrido selectivo genera fondos genéticos caracterizados por la poca diversidad genética (Fijación). En este contexto, la "aparición" aleatoria de nuevas mutaciones a través del tiempo, degenera en genotipos conformados por variantes inicialmente raras (Regreso de la diversidad).

El test de Tajima's D es un estadístico clásico que aprovecha el patrón genético previamente descrito. Planteado por primera vez en 1989 por Tajima [33], este estadístico se encarga de comparar las diferencias entre pares de individuos y el número de sitios de segregación[4] en datos de genotipificación [34]. En un escenario bajo selección positiva, los alelos raros contribuyen en menor medida a las diferencias entre pares de individuos y provocan un valor estadístico negativo [2]. En algunos casos, los valores positivos del test de Tajima's D pueden ser indicativos de selección balanceadora [35]. Bajo el mismo fundamento que el test de Tajima's D se pueden encontrar los estadísticos Fu and Li's F y Fu and Li's D [34].

De forma complementaria, los barridos selectivos pueden distorsionar el espectro de frecuencias alélicas al aumentar la frecuencia de los alelos derivados[5] como

[4] Los sitios de segregación son definidos como posiciones a lo largo del genoma que pueden variar entre un individuo y otro. Recuperado de: Fu, YX 1995 [38]

[5] Alelo derivado: Alelo que surge como resultado de una nueva mutación y no logra fijarse en una población.

producto del efecto hitchhiking. El estadístico Fay & Wu's H [36] compara el número de diferencias entre pares de individuos con el número de individuos que son homocigotos para el alelo derivado. Valores reducidos de H indican un exceso de alelos derivados con alta frecuencia, un fenómeno sugestivo de selección positiva [2].

2.4.2 MÉTODOS MICROEVOLUTIVOS: BASADOS EN DIFERENCIACIÓN POBLACIONAL

Los métodos basados en diferenciación poblacional se fundamentan en el hecho de que distintas poblaciones han experimentado diversas presiones selectivas. Por ende, los rasgos adaptativos de una población pueden ser distintos de aquellos rasgos adaptativos presentes en otra.

De esta forma, si la selección actúa en un locus específico de una población, pero no en el mismo locus en una población distinta, se podrán observar frecuencias alélicas significativamente diferentes (**Figura 4**). En contraste con otros enfoques, los métodos basados en diferenciación poblacional poseen la habilidad de detectar un mayor rango de patrones selectivos [2, 8, 39].

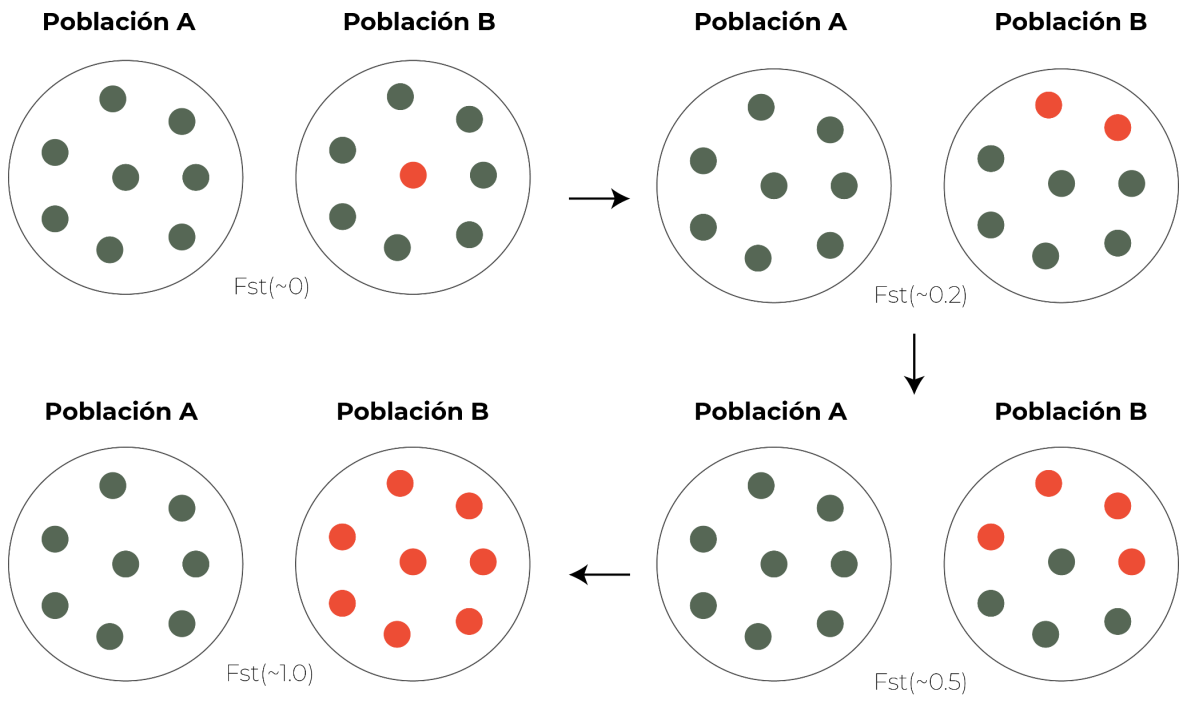


Fig. 4 Fundamento de los métodos basados en diferenciación poblacional. Las diferencias en las frecuencias alélicas pueden reflejar el impacto de la selección en una población específica. Esto provoca que el índice de fijación de Wright (Fst) entre dos poblaciones aumente.

Tradicionalmente, el Wright's Fixation Index (F_{ST}) es el estadístico basado en diferenciación poblacional más utilizado para la detección de señales de selección [39]. Esta métrica se encarga de comparar la variación de las frecuencias alélicas dentro de una población y entre poblaciones distintas. Si existe una diferenciación marcada entre dos poblaciones en relación con un locus específico, los valores de Fst serán relativamente altos. Por otro lado, si dos poblaciones son homogéneas en un locus, los valores de Fst serán relativamente pequeños [2].

A través de los años, una serie de estadísticos han derivado del Fst, entre ellos, se puede encontrar Lewontin-Krakauer Test (LKT), Locus-specific Length Branch Metric (LSBL), Cross-Population Composite Likelihood Ratio (XP-CLR), y Population Branch Statistic (PBS) [2].

2.4.3 MÉTODOS MICROEVOLUTIVOS: BASADOS EN DESEQUILIBRIO DE LIGAMIENTO.

Conforme un alelo bajo presión selectiva aumenta su frecuencia poblacional, este se mantiene en desequilibrio de ligamiento con las variantes cercanas. La combinación de la variante causal y sus alelos “vecinos” definen un haplotipo [41,42]. Los métodos basados en desequilibrio de ligamiento detectan señales de selección mediante la búsqueda de regiones con haplotipos largos. Este tipo de estadísticos son particularmente útiles para detectar variantes que han experimentado barridos selectivos parciales o incompletos, es decir, variantes que no se han fijado en la población [2] .

Un gran porcentaje de las pruebas basadas en LD se centra en el estadístico Extended Haplotype Homozygosity Statistic (EHH) [5]. La aplicación del test EHH define una región “core” (por ejemplo, un alelo con hipótesis de selección) y específica una distancia específica río arriba y río abajo de la variante. Posteriormente, calcula la probabilidad de obtener una región core idéntica al tomar aleatoriamente dos cromosomas (que incluyen la región core) de la población [5].

Partiendo del estadístico EHH, Voight y colaboradores desarrollaron en el año 2006 el análisis Integrated Haplotype Score (iHS), el cual compara el área bajo la curva definida por el test EHH [16]. Otra variación del EHH es el estadístico Cross

Population Extended Haplotype Homozygosity (XP-EHH) encargado de comparar el largo de los haplotipos entre poblaciones[39]. Otros estadísticos basados en desequilibrio de ligamiento incluyen los test long-range haplotype (LRH) [43], LD decay (LDD) [44], identity-by-descent (IBD) [45,46], entre otros. El enfoque de este trabajo estará totalmente dirigido a PBS e iHS.

2.5 HIPÓTESIS DE SELECCIÓN EN DIVERSAS DISTRIBUCIONES GEOGRÁFICAS

El desarrollo constante de la investigación evolutiva ha permitido ampliar el conocimiento acerca del origen y expansión de nuestra propia especie. La evidencia genómica [47,48] y diversos restos fósiles [49,50,51] sugieren que los humanos anatómicamente modernos surgieron hace aproximadamente 200,000 años en África y se expandieron a lo largo del mundo. En este proceso, nuestra especie se ha enfrentado a diversas presiones selectivas que han dado forma a patrones distintivos de variación genética [52]. En los siguientes capítulos se discutirán brevemente los estudios más relevantes de selección en función de la ubicación geográfica poblacional (**Figura 5**).

2.5.1 HIPÓTESIS DE SELECCIÓN EN POBLACIONES EUROPEAS

En la mayoría de los seres humanos, la habilidad de digerir la lactosa decrementa rápidamente después de la etapa de lactancia debido a una disminución de la enzima lactasa-florizina hidrolasa (LPH) [53,54,55]. Particularmente, la enzima LPH es la encargada de degradar la lactosa en carbohidratos menos complejos que pueden ser fácilmente metabolizados por el sistema [55]. En el año 1973, Cavalli-Sforza et al. propusieron que la tolerancia a la lactosa en adultos era un rasgo común en poblaciones que practicaban la domesticación de ganado y poseían una dieta basada en productos lácteos [56]. Así bien, plantearon una hipótesis de selección positiva en un gran porcentaje de las poblaciones europeas del norte.

A través de los últimos años el estudio de la selección en el gen *LCT* (codificante de la enzima LPH) se ha basado en el análisis de datos de genotipificación con herramientas estadísticas; las variantes cercanas al locus LCT muestran algunas de las señales de selección más evidentes en el genoma humano [39]. Asimismo,

estudios in vitro de las variantes identificadas demostraron un efecto en la sobreexpresión del gen encargado de la degradación de la lactosa [52].

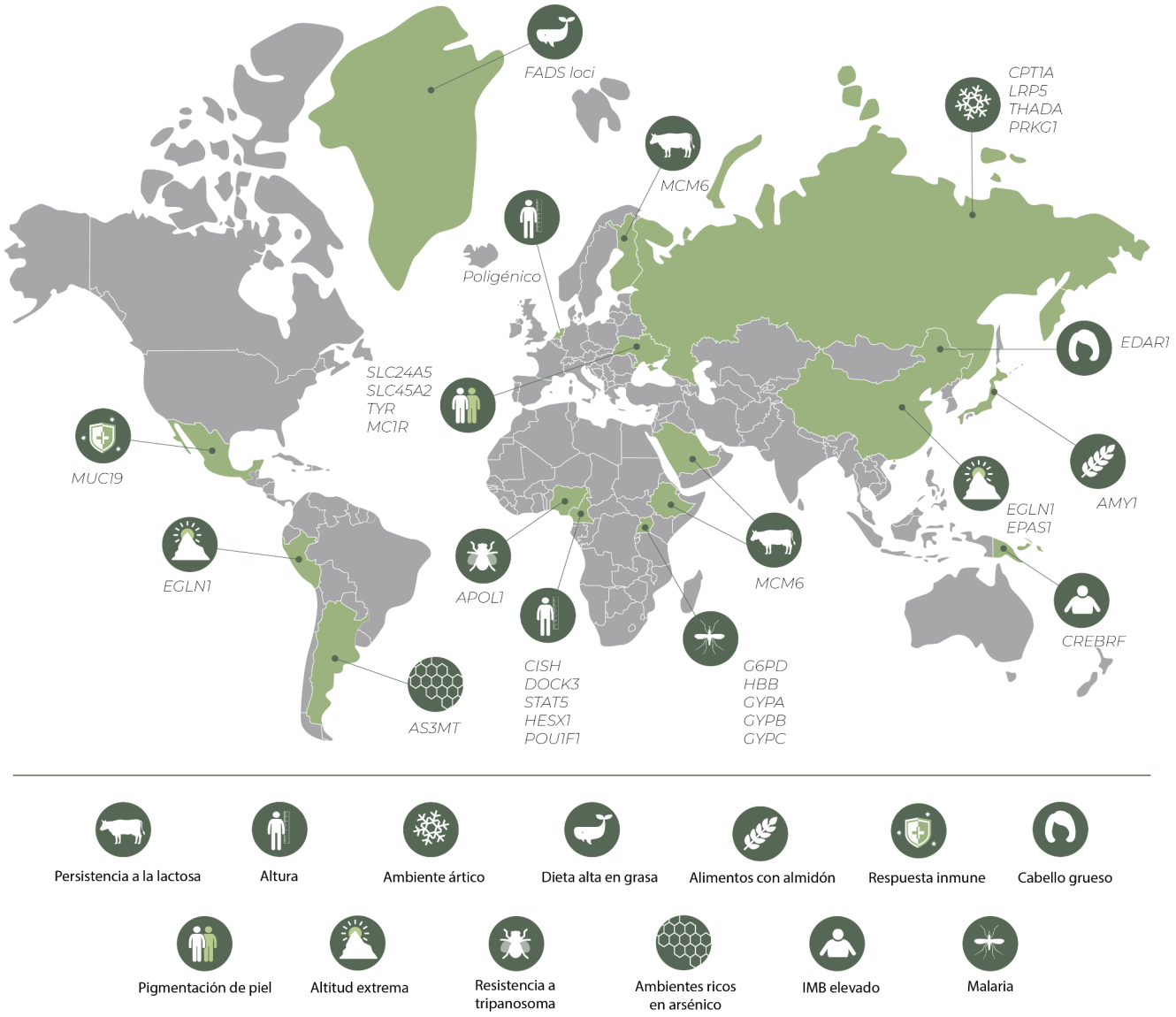


Fig. 5 Ejemplos de genes bajo presión selectiva local; cada uno marcado por el fenotipo o presión selectiva y el loci genético bajo selección. Adaptado de Fan, S et al [52]

Por otro lado, la variación en el color de piel es uno de los rasgos fenotípicos más característicos de las poblaciones. John Mitchell y Samuel Stanhope Smith fueron los primeros en reconocer una correlación directa entre la pigmentación de la piel y la ubicación geográfica de las poblaciones [57,58,59]. Actualmente sabemos que la radiación ultravioleta (UV) es una presión selectiva que ha influido directamente en

la pigmentación evolutiva de la piel; las pieles más oscuras se asocian con bajas altitudes (protección de rayos UV) y las pieles más claras con altitudes elevadas (síntesis de vitamina D). Los primeros estudios de selección relacionados con la pigmentación de piel en poblaciones humanas se fundamentaron en el estudio de genes candidatos identificados previamente. Años después, el avance en las tecnologías de secuenciación permitió detectar conjuntos de genes (*OCA2, TYRP1, TYR, SLC24A5* y *SLC45A2*) bajo posible selección relacionados con el color de piel claro en poblaciones europeas [60,52].

Finalmente, Michael Turchin y colaboradores demostraron la influencia de la selección en uno de los rasgos poligénicos más comunes, la altura. En este caso, las frecuencias de los alelos asociados con mayor altura se encontraban sistemáticamente elevadas en las poblaciones pertenecientes al norte de Europa en comparación con las poblaciones del Sur de Europa [61].

2.5.2 HIPÓTESIS DE SELECCIÓN EN POBLACIONES AFRICANAS.

Como se ha mencionado anteriormente, los ambientes patogénicos son uno de los principales agentes de selección en las poblaciones. En particular, la malaria, una infección parasitaria transmitida por mosquitos, es una de las mayores causas de mortalidad en África [62]. Las observaciones propuestas por la teoría de la malaria de Haldane demostraron en un futuro la existencia de variantes que confieren resistencia a la enfermedad. Particularmente, los alelos causantes de talasemia y anemia falciforme, así como variantes en los genes *ABO, GYPA, GYPB, GYPE y G6PD* han demostrado proporcionar resistencia a la misma [63]. Sin embargo, es importante considerar que las mutaciones con acción ventajosa, a su vez conllevan enfermedades Mendelianas, por esto, se han identificado señales de selección balanceadora [52,63].

Diversos autores han reportado señales fuertes de selección natural en las regiones genéticas cercanas a los genes *MYH9* y *APOLI* [64,65,66]. Las variantes en los genes anteriormente citados se han asociado con nefropatías crónicas, sin embargo, confieren resistencia a la tripanosomiasis africana.

Por otra parte, la expansión de diversos grupos poblacionales involucró la adaptación a regiones de gran altitud con baja disponibilidad de oxígeno [67]. En altitudes extremas, las diferencias en la presión barométrica resultan en hipoxia, hipertensión pulmonar y un riesgo aumentado de preeclampsia [54,68] . Sin embargo, diversos estudios fisiológicos sugieren un fenotipo adaptativo en poblaciones pertenecientes a Etiopía. Los residentes nativos de este país africano, viven a 3,560 metros por arriba del nivel del mar sin desarrollar hemoglobinopatías, inflamaciones crónicas o alteraciones de la presión arterial [69] .

Como es lógico, las señales de selección más fuertes en esta población fueron identificadas en una gran variedad de genes relacionados con el suministro de sangre y en la vía de factor inducible por hipoxia. Especialmente, mutaciones en los genes *THRB* y *ARNT2* presentaron correlaciones con los niveles de hemoglobina en sangre [54]. Finalmente, así como fue descrito con las poblaciones europeas, la habilidad de digerir lactosa en la etapa adulta es un rasgo característico de las poblaciones africanas. Específicamente, la variante bajo selección C-14010 incrementa la transcripción del gen *LCT* y posee una alta prevalencia en Kenia y Tanzania. De la misma forma, dos variantes diferentes - G-13915 y G-13907 - a aquellas identificadas con alta frecuencia en Tanzania, son ampliamente comunes en el norte de Sudán y Kenia [54]. La genotipificación específica de las poblaciones africanas, ha demostrado la presencia de mutaciones en el gen LCT que provienen de fondos genéticos diferentes a las reportadas en poblaciones europeas [55]. De esta forma, es importante resaltar que distintos alelos bajo presión selectiva pueden conllevar al mismo fenotipo bajo selección.

2.5.4 HIPÓTESIS DE SELECCIÓN EN POBLACIONES ASIÁTICAS.

Una señal de selección particularmente fuerte en las poblaciones asiáticas, implica al gen receptor de ectodisplasina *EDAR*. En el año 2008, Fujimoto y colaboradores demostraron que la presencia de un polimorfismo de un solo nucleótido (SNP) en el gen *EDAR* se asocia fuertemente con el grosor del cabello, particularmente de los grupos nativos de Japón, Tailandia e Indonesia [70]. Asimismo, se ha demostrado una correlación directa de las variantes en el gen *EDAR* con la morfología dental de las poblaciones asiáticas [71]. Desafortunadamente, el entendimiento a nivel evolutivo

de estás variantes permanece desconocido. Esto muestra que identificar las señales de selección en el genoma es solo la mitad de la historia.

Las variantes estructurales (SVs) - incluyendo duplicaciones, variantes de número de copia, deleciones, entre otras - pueden contribuir ampliamente a la adaptación de las poblaciones humanas. En concreto, Perry et al. describieron la presencia de un mayor número de copias del gen codificante de la enzima amilasa salival (*AMY1*) en poblaciones con dietas basadas en el consumo de almidón [72]. Entre estas, la población nativa asiática Hadza, con dieta basada en raíces y tubérculos ricos en almidón, presentó un valor más alto de copias del gen *AMY1* en comparación con otras poblaciones.

De la misma forma que Etiopía, las comunidades pertenecientes a la meseta tibetana presentan evidencia fuerte de adaptación a las altitudes extremas. En contraste con las poblaciones africanas, el genoma tibetano revela señales de selección en *EGLN1* [73] , un gen vinculado con los niveles de hemoglobina en sangre, y frecuencias alélicas diferentes en *EPAS1*, un gen encargado de codificar un factor de transcripción relacionado con la respuesta a hipoxia [52,67].

2.5.4 HIPÓTESIS DE SELECCIÓN EN POBLACIONES AMERICANAS.

Las investigaciones basadas en la detección de señales de selección en poblaciones indígenas del Norte de América se han enfocado parcialmente en la detección de genes relacionados con el sistema inmune y la adaptación a ambientes fríos.

Con respecto al primer punto, en el año 2016 Lindo [74] y colaboradores reportaron señales de selección fuertes en comunidades canadienses. Particularmente, el gen *HLA-DQA1* encargado de la codificación del complejo principal de histocompatibilidad clase II, presentó evidencias de adaptación. Por otro lado, diversos genes bajo selección involucrados directamente con la respuesta y trastornos inmunes, han sido identificados en comunidades nativas mexicanas [12, 13, 14,]. Con respecto a la adaptación climática estudios basados en poblaciones nativas de Groenlandia revelaron una señal fuerte de adaptación en los genes codificantes de las enzimas ácido graso desaturasas, al igual que en el gen TBX15 asociado

con la diferenciación de adipocitos. Ambas señales de selección son atribuibles a la adaptación a ambientes árticos [12].

Por otra parte, las investigaciones basadas en la detección de señales de selección en América del Sur se han enfocado parcialmente en la detección de genes relacionados con la adaptación a altitudes extremas y ambientes tóxicos. De la misma forma que Etiopía y el Tibet, las poblaciones pertenecientes al altiplano andino han evidenciado fenotipos adaptados a la altitud. En el año 2001, Moore et al. buscaron señales adaptativas en genomas andinos e identificaron cuatro genes bajo selección involucrados en la vía de factor inducible por hipoxia [75]. Más recientemente, Bigham et al.[76], Simonson et al. [77], Beall et al [78], y otros grupos de investigación, han hecho uso de de datos de genotipificación y secuenciación de exomas [67] para caracterizar nuevos genes y variantes bajo selección en poblaciones nativas andinas y tibetanas.

También se ha demostrado que los habitantes del norte de los Andes argentinos (una región árida con elevadas concentraciones de arsénico en el agua) poseen frecuencias altas de variantes “protectoras” en el gen AS3MT, lo que sugiere una adaptación humana al arsénico [79]. Mientras que el análisis de genoma completo de más de 1,500 brasileños nororientales mostró SNP´s bajo selección dentro o cerca de genes asociados con la función inmunológica, rasgos metabólicos y desarrollo embrionario [80].

3. MARCO CONCEPTUAL

3.1 DESCRIPCIÓN DE POPULATION BRANCH STATISTIC (PBS).

Tradicionalmente, el F_{ST} es el estadístico basado en diferenciación poblacional más utilizado para la detección de señales de selección [39]. Esta métrica se encarga de comparar la variación de las frecuencias alélicas dentro de una población y entre dos poblaciones distintas. A pesar de que el estadístico F_{ST} puede ser utilizado para detectar posibles targets de selección, la obtención de valores por este medio no revela cuál de las dos poblaciones estudiadas fue directamente afectada por la selección natural. Para superar este problema, Yi y colaboradores describieron por primera vez el PBS [67].

Este estadístico incluye a dos poblaciones filogenéticamente relacionadas - pop1 y pop2 - y una tercera población filogenéticamente más lejana - popout -. Posteriormente, al comparar los valores por pares de F_{ST} entre estas tres muestras - pop1vspop2, pop1vspopout, y pop2vspopout - el PBS puede estimar el cambio en las frecuencias alélicas de la población de interés (pop1) desde su divergencia de la población filogenéticamente cercana (pop2). De esta forma, el valor de PBS cuantifica la magnitud de cambio en las frecuencias alélicas de una población en un locus específico desde su divergencia de las otras dos poblaciones [67].

Este enfoque es altamente similar al utilizado por la métrica LSBL, en la cual se calculan valores por pares de F_{ST} para tres o más poblaciones con el fin de lograr la identificación de cambios específicos en las frecuencias alélicas de una población. A pesar de que ambos estadísticos comparten el mismo fundamento, el PBS utiliza la clásica transformación logarítmica de Cavalli-Sforza [67, 81],

$$T = -log(1 - F_{st})$$

Y es calculado de la siguiente manera,

$$\text{PBS} = \frac{T^{pop1pop2} + T^{pop1popout} - T^{pop2popout}}{2}$$

Donde,

$$T^{pop1pop2} = -\ log(1\ -\ F_{st}^{pop1,pop2})$$

$$T^{pop1popout} = -\ log(1\ -\ F_{st}^{pop1,popout})$$

$$T^{pop2popout} = -\ log(1\ -\ F_{st}^{pop2,popout})$$

El uso de PBS para la detección de señales de selección ha sido descrito a nivel de genes y a nivel de variantes. En el primer caso, se computa un valor ponderado de F_{ST} para una región genética completa, así, un valor de PBS calculado bajo estas características permite conocer si la región genética en su totalidad se encuentra bajo selección o no. Por otro lado, la determinación de PBS por variante requiere el cálculo de F_{ST} para cada variante, en este caso el valor de PBS bajo estas características permite conocer si la variante en si misma se encuentra o no bajo selección.

3.1.1 DESCRIPCIÓN DE PBS: MODELADO DEMOGRÁFICO CON FASTSIMCOAL2

A pesar de que los estadísticos basados en diferenciación poblacional presentan ventajas considerables en comparación con otros enfoques, existen un número de limitaciones que deben ser consideradas al interpretar sus resultados. Particularmente, los eventos demográficos como deriva génica[6], migraciones, expansiones, cuellos de botella, entre otros, pueden mimetizar el efecto de la selección en el genoma [2]. El reconocimiento de una correlación directa entre el impacto de la selección y determinados eventos demográficos, ha conllevado al desarrollo de diversos métodos que pretenden modelar la estructura poblacional bajo diversos parámetros [2]. fastsimcoal2 es un software de modelado demográfico desarrollado en el año 2013 por Laurent Excoffier y colaboradores, que permite modelar escenarios evolutivos neutrales[7] a partir de datos de genotipificación

[6] Deriva génica: Refiere a fluctuaciones aleatorias en las frecuencias alélicas debido a eventos fortuitos. Recuperado de: NHGRI "Genetic drift" [106]

[7] Teoría neutralista de la evolución molecular: Propuesta por Kamura en el año 1985, fundamenta que un gran porcentaje de los cambios genéticos observados son atribuibles a la deriva génica en lugar de la selección darwiniana. Recuperado de: X

previamente observados [82]. En los tests basados en diferenciación poblacional es recomendable comparar los datos empíricos con datos generados bajo simulaciones neutrales en fastsimcoal2 para poder obtener una hipótesis nula y por ende, datos estadísticamente significativos.

3.2 DESCRIPCIÓN DE INTEGRATED HAPLOTYPE SCORE (iHS)

El estadístico integrated Haplotype Score (iHS) se encuentra basado en el test Extended Haplotype Homozygosity (EHH) propuesto por primera vez en el año 2002 por Sabeti et al [5]. La aplicación del test EHH define una “región core” (por ejemplo, un alelo con hipótesis de selección) y específica una “distancia x” río arriba y río abajo de la variante. El EHH a una “distancia x” de la “región core” es definido como la probabilidad de que dos cromosomas (portadores de la región core) seleccionados aleatoriamente, sean homocigotos para todos los SNP´s que se encuentran en la “distancia x” . El valor de EHH es computado en una escala de 0 (no homocigosidad, todos los haplotipos son diferentes) a 1 (homocigacidad completa, todos los haplotipos son iguales) [5] (**Figura 6**)

Fig. 6 Descripción gráfica del estadístico Extended Haplotype Homozygosity (EHH). Izquierda: Cuando la región core es neutral (Sin selección) no existe homocigocidad entre los haplotipos comparadas; todos los haplotipos son diferentes. Derecha: Cuando la región core se encuentra bajo selección, existe homocigocidad completa entre los haplotipos comparados; todos los haplotipos son iguales.

Cuando un alelo bajo selección aumenta rápidamente su frecuencia por efecto de un barrido selectivo, se observarán altos niveles de homocigocidad haplotípica en contraste con lo esperado bajo un modelo de evolución neutral. Por lo tanto, en los gráficos de EHH contra Distancia (**Figura 7**) el área bajo la curva de EHH será

usualmente más grande para un alelo bajo selección en comparación de lo observado para un alelo neutral [16].

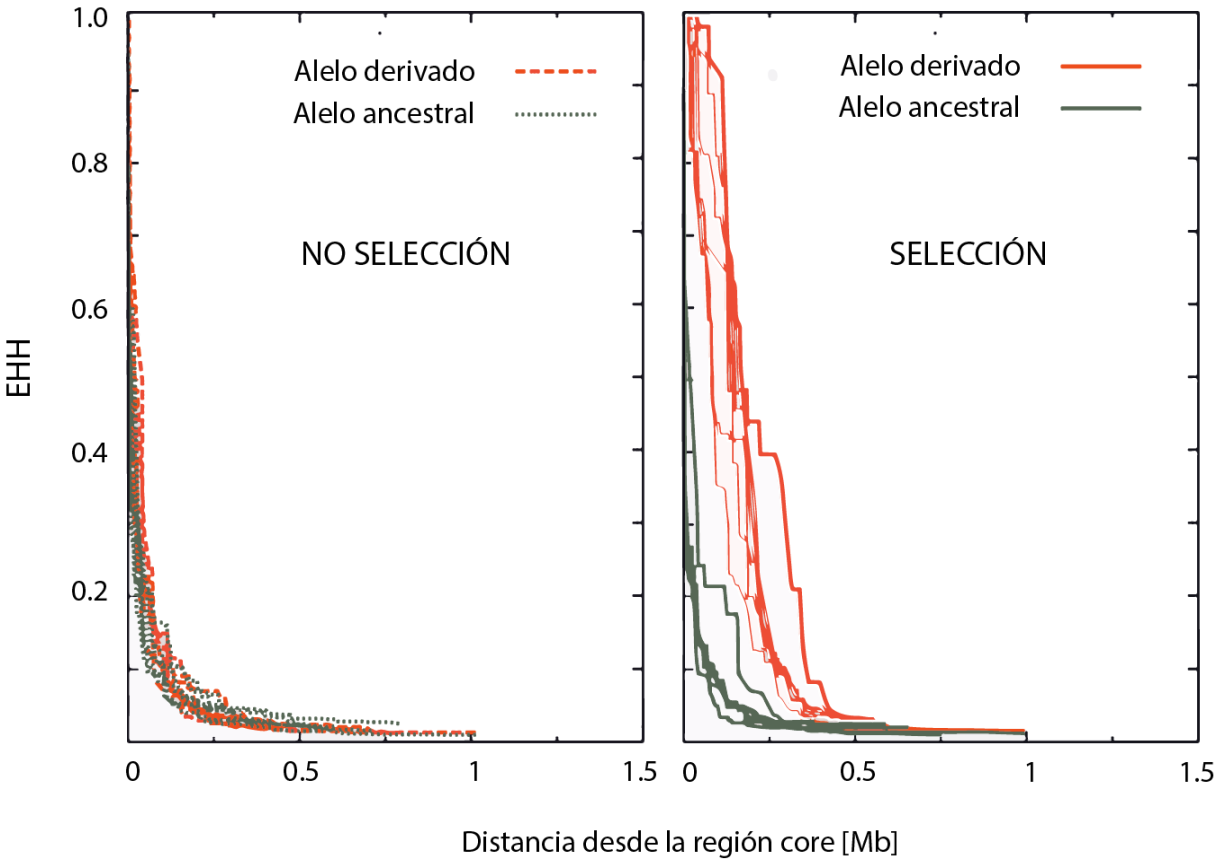


Fig. 7 Decremento de EHH en datos simulados para un alelo con frecuencia de 0.5. Cuando la región core es neutral (No selección), la homocigosidad de los haplotipos disminuye de forma similar en los alelos derivados y ancestrales. Cuando la región core se encuentra bajo selección, la homocigocidad entre haplotipos disminuye en menor medida para los alelos derivados en comparación con los alelos ancestrales. La discrepancia entre las áreas de ambas curvas representa el fundamento del estadístico iHS. Adaptado de: Voight et al. 2006[16]

Voight y colaboradores capturaron el efecto anteriormente descrito, al calcular la integral de la disminución de homocigocidad observada desde una región core específica hasta alcanzar un valor de EHH igual a 0.05 [16]. Este valor de EHH integrado (iHH) es denotado como iHH_A en caso de ser computado con respecto al alelo ancestral y como iHH_D si es computado con respecto al alelo derivado. Así bien, el estadístico iHS es obtenido de la siguiente forma:

$$iHS\ no\ estandarizado\ =\ In\ (\frac{IHH_A}{IHH_D})$$

Cuando el valor de decremento del estadístico EHH es similar en los alelos ancestrales y derivados, la división de iHH_A entre IHH_D es igual a 1.0 y por ende el valor no estandarizado de iHS es igual a 0 (ln(1) = 0). Por el contrario, los valores negativos grandes indican haplotipos inusualmente largos con el alelo derivado y los valores positivos grandes indican haplotipos largos con el alelo ancestral [16]. Finalmente, la fórmula de iHS normalizada en bins de frecuencias alélicas se describe a continuación:

$$iHS = \frac{In\left(\frac{IHH_A}{IHH_D}\right) - E_p\left[In\left(\frac{IHH_A}{IHH_D}\right)\right]}{SD_P\left[In\left(\frac{IHH_A}{IHH_D}\right)\right]}$$

Donde, $E_p\left[In\left(\frac{IHH_A}{IHH_D}\right)\right]$ y $SD_P\left[In\left(\frac{IHH_A}{IHH_D}\right)\right]$ son la expectativa y la desviación estándar en el bin de frecuencia p.

3.2.1 DESCRIPCIÓN DE iHS: FASEADO DE HAPLOTIPOS.

Los datos genómicos obtenidos a través de tecnologías de secuenciación de nueva generación, poseen generalmente un formato no faseado. La inferencia de haplotipos, también llamada “faseado”, es una estimación sobre qué alelos se encuentran co-localizados en un mismo cromosoma [83]. Este preprocesamiento de los datos genómicos es fundamental para todos los métodos de detección de señales de selección basados en desequilibrio de ligamiento. Afortunadamente, existe una gran variedad de softwares que permiten alcanzar este objetivo. Encontrar una revisión detallada en Material Suplementario: material_suplementario

3.2.2 DESCRIPCIÓN DE iHS: HERRAMIENTAS BIOINFORMÁTICAS

En contraste con el PBS, el estadístico iHS ha sido implementado en una variedad de softwares o programas bioinformáticos. *rehh* es un paquete desarrollado en el software estadístico R que permite la identificación de señales de selección en datos de genoma completo usando EHH e iHS [84]. En el año 2014, Szpiech and Hernandez realizaron una implementación de *rehh* a través del desarrollo de Selscan, un programa que demostró ser dos veces más rápido para el cálculo de métodos basados en desequilibrio de ligamiento[85]. A pesar de la efectividad de

ambas herramientas, su uso implica el acceso obligado a clústers computacionales con gran capacidad de procesamiento [86]. Así bien, Colin A. Maclean y colaboradores desarrollaron hapbin, un programa que permite calcular EHH, iHS y XP-EHH 3400 veces más rápido que selscan en una computadora de escritorio estándar [86].

3.3 DESCRIPCIÓN DE iHS VS PBS

Cada uno de los estadísticos utilizados para la detección de señales de selección se fundamenta en diversos patrones genéticos y permite la detección a diversas escalas temporales. Diferentes estudios han utilizado multiples métodos en conjunto para beneficiarse de la complementariedad entre estadísticos y disminuir la probabilidad de obtener resultados falsos positivos [87]. Particularmente, el iHS detecta señales de selección positivas con barridos selectivos incompletos o en curso. De la misma forma, el PBS exhibe un gran poder para detectar selección natural reciente [67]. Ambos test se han usado en conjunto para aumentar el poder estadístico de detección.

4. PLANTEAMIENTO DEL PROBLEMA

A través de los últimos años el aumento considerable del poder computacional y el desarrollo de tecnologías de secuenciación de nueva generación, han permitido detectar e identificar regiones del genoma bajo selección natural. A pesar de que la detección de señales de selección logra identificar nuevas asociaciones genotipo-fenotipo y fomenta el entendimiento de la estructura genética poblacional, el uso requerido de métodos estadísticos y bioinformáticos representa una barrera para los científicos poco familiarizados con la manipulación de datos genómicos masivos. Esto último ha creado un área de oportunidad creciente para el desarrollo de herramientas bioinformáticas que faciliten el preprocesamiento de datos y la obtención de hipótesis de selección en datos genómicos. Asimismo, el Population Branch Statistic (PBS) — un test basado en diferenciación poblacional para la detección de señales de selección — carece de softwares que permitan su cálculo reproducible y automatizado.

Por lo anterior, este trabajo presenta el desarrollo de un pipeline bioinformático que computa PBS e Integrated Haplotype Score (iHS) para lograr la detección de señales de selección positivas en una escala temporal reciente. Esta herramienta se fundamenta en un desarrollo ampliamente documentado, escalable y reproducible. De igual forma, se desarrolló un análisis piloto con datos pertenecientes a poblaciones Nativas Mexicanas con el fin de contribuir al estudio de señales en genomas de poblaciones NatAm, las cuales se encuentran subrepresentadas en los estudios evolutivos. Las relaciones entre variantes y fenotipos particulares reportadas, así como la herramienta provista para su estudio, contribuye al desarrollo de la genómica poblacional nacional.

5. HIPÓTESIS

El desarrollo de una herramienta bioinformática facilita la detección automatizada de señales de selección en datos de secuenciación de genoma completo.

6. OBJETIVOS

6.1 GENERAL

Desarrollar un pipeline bioinformático para la detección automatizada de señales de selección a través del uso de PBS e iHS.

6.2 ESPECÍFICOS

- Desarrollar un script en el ambiente R para el cálculo de PBS.
- Desarrollar un pipeline para la obtención de datos bajo modelado neutral con Fastsimcoal2.
- Detectar señales de selección en poblaciones nativas mexicanas (Proyecto 100G-MX) con PBS e iHS.
- Describir el número de variantes bajo selección utilizando datos de genoma completo de poblaciones nativas mexicanas (Proyecto 100G-MX)
- Describir el contexto biológico de las variantes bajo selección en poblaciones nativas mexicanas.

7. MATERIAL Y MÉTODOS

7.1 WORKFLOW BIOINFORMÁTICO

Para automatizar el proceso de detección de señales de selección con PBS e iHS, se desarrolló un pipeline de Nextflow (**nf-selection**) que logra la implemetación de los análisis requeridos para obtener resultados.

El pipeline, scripts desarrollados para la operación y un test data se encuentran publicamente disponibles vía GitHub. La documentación de la herramienta incluye instrucciones para instalación y uso. Los siguientes apartados describen los análisis implementados y el diagrama de flujo del pipeline.

7.2 OPERACIÓN DEL PIPELINE

nf-selection corre dentro de una distribución estándar de Linux (análisis probado exitosamente en Ubuntu 18.04 LTS y Ubuntu 20.04 LTS) y requiere la instalación del software descrito en la **Tabla 2**.

La herramienta recibe dos inputs principales; un archivo con valores separados por comas (comma-separated values, csv) con los paths de los archivos requeridos para el cálculo de iHS por cromosoma (**CSV 1**) y un archivo csv con los paths de los archivos necesarios para calcular PBS (**CSV 2**).

CSV 1 | Input iHS. path_vcf; path hacia el archivo VCF[8] que contiene la información de genotipificación de la población de interés. path_genetic_map; path hacia el mapa genético de cada cromosoma. path_reference_vcf; path hacia el archivo VCF de referencia. path_index_reference; path hacia el index de la referencia.

chromosome	path_vcf	path_genetic_map	path_reference_vcf	path_index_reference
21	file.vcf	file.map	file.vcf	file.vcf.tbi

CSV 2 | Input PBS. path_vcf; path hacia el archivo VCF que contiene la información de genotipificación de la población de interés. path_pop1; path hacia el archivo txt con los identificadores individuales de la población de interés. path_pop2; path hacia el archivo con los identificadores de la población filogenéticamente emparentada. path_pop3; path hacia el archivo con los identificadores de la población filogenéticamente lejana.

[8] Los archivos VCF son textos planos delimitados con tabs que contienen información acerca del cromosoma, coordenadas, alelos de referencia y alternos, así como ID´s de genotipificación por muestra. Recuperado de: X.

path_vcf	path_pop1	path_pop2	path_popout
file.vcf	pop_1	pop_2	pop_out

Cada proceso del pipeline es realizado en un entorno computacional paralelo de alto nivel, permitiendo la ejecución de múltiples tareas a través de diversos nodos de cómputo.

TABLE 2 | Software utilizados en pipeline de Nextflow.

Aplicación y versión	Uso en el workflow	Sitio web o repositorio
Nextflow ≥ 21.04.2	Orquestación del Workflow, funcionamiento del pipeline	https://www.nextflow.io/
SHAPEIT4	Faseado de haplotipos	https://odelaneau.github.io/shapeit4/
aa_annotate.py	Anotación de alelo ancestral	https://github.com/MerrimanLab/selectionTools/blob/master/selection_pipeline/aa_annotate.py
hapbin	Cálculo de iHS	https://github.com/evotools/hapbin
VCFtools = 0.1.15	Weir and Cockerham Fst calculation, Allele frequency calculation	http://vcftools.sourceforge.net/
bedtools	Anotación de resultados	https://bedtools.readthedocs.io/en/latest/
R ≥ 4.0.5	stringr, dplyr, ggplot2, ggrepel, tidyr, cowplot, vroom	https://www.r-project.org/
stringr	preprocesamiento; manipulación de expresiones regulares	https://cran.r-project.org/web/packages/stringr/index.html

dplyr	preprocesamiento y cálculo de PBS; manipulación de dataframes; manipulación de resultados de iHS	https://cran.r-project.org/web/packages/dplyr/index.html
ggplot2	Cálculo de PBS; desarrollo de plots	https://cran.r-project.org/web/packages/ggplot2/index.html
ggrepel	Cálculo de PBS; desarrollo de plots	https://cran.r-project.org/web/packages/ggrepel/index.html
tidyr	preprocesamiento y cálculo de PBS; manipulación de dataframes	https://cran.r-project.org/web/packages/tidyr/index.html
cowplot	Calculo de PBS; desarrollo de plots	https://cran.r-project.org/web/packages/cowplot/index.html
vroom	Calculo de PBS; importación de data frames	https://cran.r-project.org/web/packages/vroom/index.html
purr	Manipulación automatizada de vectores	https://cran.r-project.org/web/packages/purrr/index.html
BioCircos	Desarrollo Circus plot	https://cran.r-project.org/web/packages/BioCircos/index.html

7.2.1 OPERACIÓN DEL PIPELINE: iHS

El flujo de trabajo de la herramienta se divide en una etapa de preprocesamiento y una etapa de cómputo del estadístico (**Figura 8**). Dado que el test iHS requiere genotipos faseados, el primer paso del preprocesamiento involucra el uso de SHAPEIT4 para este propósito. Considerando que el faseado a través de métodos computacionales puede alcanzar una alta precisión a través del uso de paneles genómicos externos[89], **nf-selection** fue programado para recibir dentro del input principal una referencia genómica en formato VCF. El usuario puede encontrar

referencias del proyecto 1000 Genomas en formato VCF y sus index tbi en los siguientes enlaces:

Referencias GRCh37:
https://ftp-trace.ncbi.nih.gov/1000genomes/ftp/release/20130502/

Referencias GRCh38:
20201028_3202_phased.

Asimismo, la inferencia de haplotipos requiere un mapa genético[9] que debe ser proporcionado dentro del input principal. Mapas genéticos para el genoma humano en sus versiones GRCh37 y 38 pueden ser encontradas dentro del repositorio principal. El segundo paso del preprocesamiento involucra la anotación del alelo ancestral en los datos de genotipificación con el script annotate.py. A pesar de ser un paso ampliamente recomendado, el usuario puede omitirlo y proceder con el análisis. De igual forma, si los datos han sido previamente faseados, el uso de SHAPEIT4 puede ser omitido.

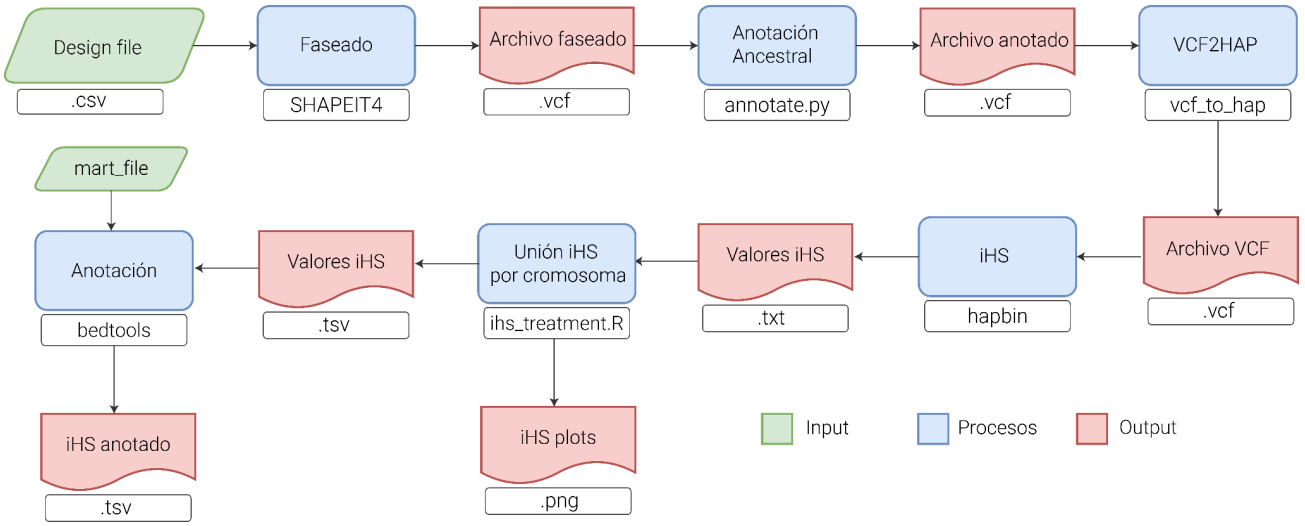


FIGURA 8 | Workflow esquemático del pipeline desarrollado para detectar señales de selección. Branch iHS. Diagrams created with draw.io.

La etapa de cómputo del estadístico implica el uso de hapbin para determinar valores de iHS por posición genómica. Por default, el iHS no es reportado en sitios

[9] Mapa genético: Archivos .map que contiene la posición genómica (cM) de los genes. Recuperado de: Brown, 2017 [89]

con frecuecia del alelo menor (MAF) < 0.05, sin embargo, el valor de MAF es un parámetro flexible y puede ser definido a través del argumento **--maff**. Considerando que hapbin procesa el valor de iHS por cromosoma, los pasos finales de la etapa de cómputo consisten en integrar los resultados de cada proceso, anotarlos con base en un archivo mart y realizar gráficos que fomentan la interpretación visual de los resultados.

7.2.2 OPERACIÓN DEL PIPELINE: PBS

De la misma forma que el estadístico iHS, el workflow de PBS se divide en una etapa de preprocesamiento de datos y una etapa de computación de estadístico (**Figura 9**). Tomando en cuenta que el PBS requiere como input el cálculo de valores de F_{ST} entre poblaciones, el preprocesamiento de datos incluye el uso de VCFTools para calcular el F_{ST} de Weir and Cockerham. En paralelo a este proceso, el pipeline calcula las frecuencias alélicas para cada población.

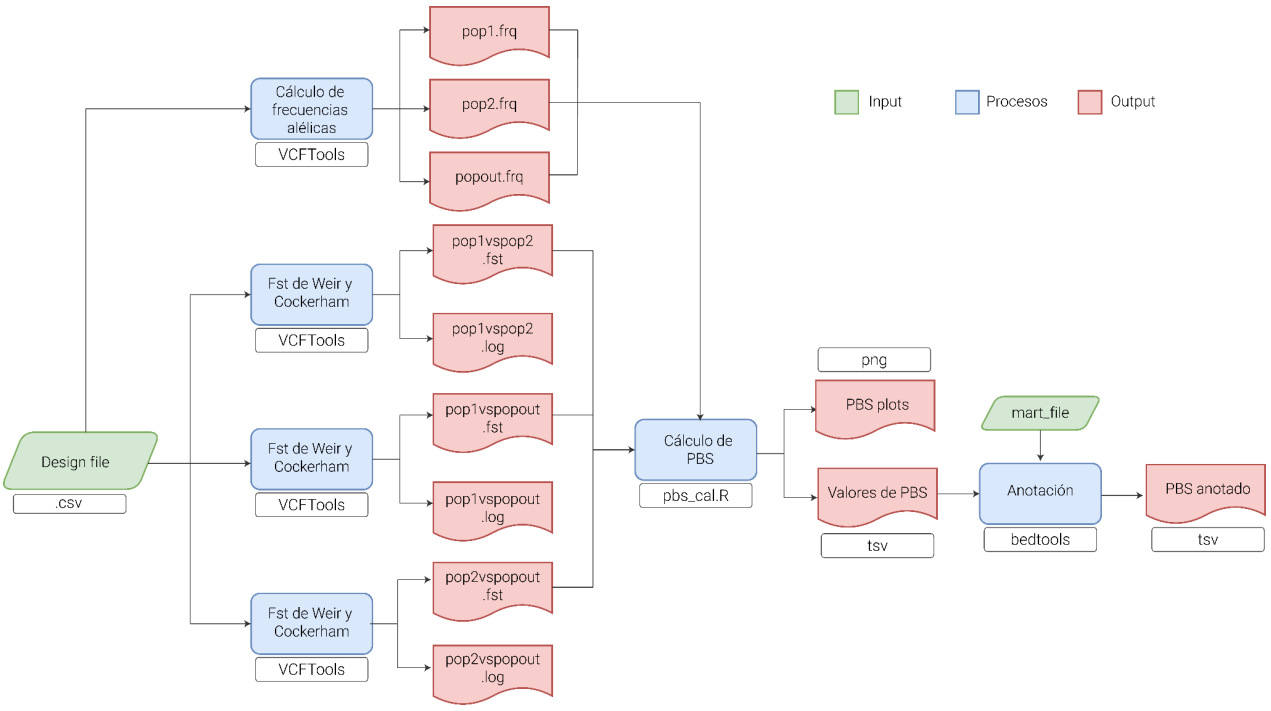


FIGURA 9 | Workflow esquemático del pipeline desarrollado para detectar señales de selección. Branch PBS por variante genética. Diagramas creados con draw.io.

La etapa de cómputo del estadístico involucra el cálculo de PBS por variante genética. Considerando la ausencia de una herramienta para calcular PBS, se desarrolló un script en un ambiente R que permite calcular un valor final de PBS por cada una de las variantes previstas en el archivo VCF. El resultado final de la etapa de cómputo es un archivo separado por tabuladores con los valores de PBS y diversos gráficos que permiten la visualización de los resultados. Asimismo, si el usuario provee un archivo mart el pipeline lleva a cabo una anotación genómica final.

Por la dificultad que implica la automatización del modelado demográfico, se optó por el desarrollo exclusivo de una herramienta que permite la obtención de datos simulados bajo un modelo neutral. El pipeline y scripts desarrollados para la operación se encuentran publicamente disponibles vía el siguiente repositorio GitHub:

https://github.com/fernanda-miron/nf-demographic-modelling

7.2.3 OPERACIÓN DEL PIPELINE: IHS VS PBS

Finalmente, **nf-selection** realiza una interesección entre el 1% de los valores más elevados de cada estadístico para identificar las variantes con mayor probabilidad de estar bajo selección (**Figura 9**).

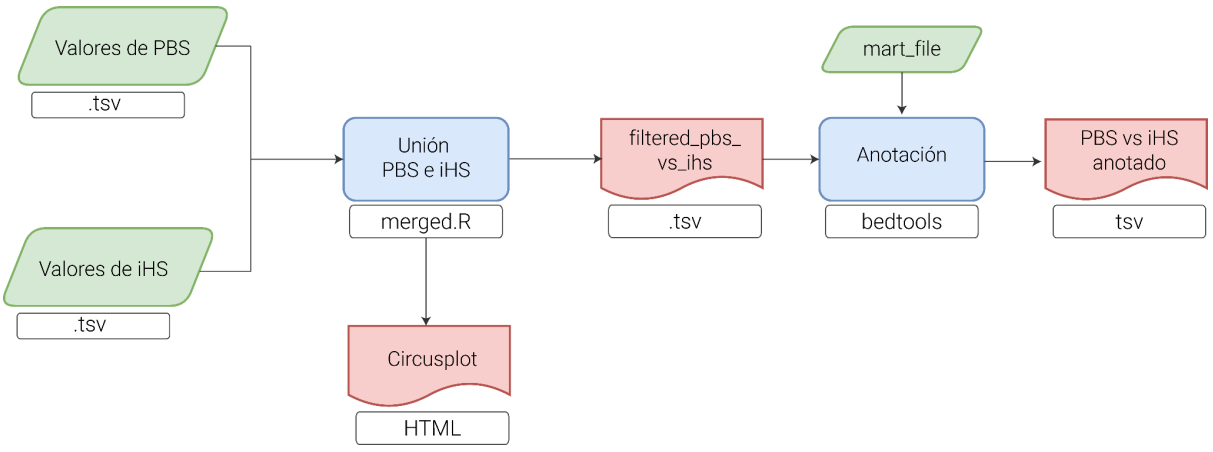


FIGURA 10 | Workflow esquemático del pipeline desarrollado para detectar señales de selección. Branch iHS vs PBS por variante genética. Diagramas creados con draw.io.

Los resultados de este proceso son visualizados gráficamente a través de un circusplot desarrollado con la paquetería “BioCircus” del ambiente R. De la misma forma, si el usuario provee un archivo mart en el procesamiento de iHS o PBS, el pipeline retoma el input y lleva cabo una anotación genómica final.

7.3 DETECCIÓN DE SEÑALES DE SELECCIÓN EN GENOMAS COMPLETOS DE NATIVOS AMERICANOS.

Para verificar la efectividad de la herramienta desarrollada, se llevó a cabo un análisis de detección de selección con datos pertenecientes a 76 individuos Nativos Mexicanos. Los siguiente apartados describen brevemente la metodología utilizada para este propósito.

7.3.1 MUESTRAS.

El Instituto Nacional de Medicina Genómica (INMEGEN) compartió un conjunto de variantes poblacionales que incluye 76 individuos no emparentados pertenecientes a 27 grupos indígenas. Los individuos incluidos son parte del proyecto 100G-MX del mismo instituto [13]. El estudio se hizo en conformidad con la declaración de Helsinki, y fue aprobado por el comité de ética e investigación humana del INMEGEN. Los individuos presentan un promedio de 97.22% de genoma nativo americano.

El conjunto de datos está reportado en la versión GRCh38 del genoma humano y fue previamente sometido a un proceso de llamado de variantes. Los datos para este proyecto fueron proporcionados en formato VCF.

7.3.2 DISEÑO EXPERIMENTAL.

La detección de señales de selección se llevó a cabo en las poblaciones clusterizadas en la región central de México [13]. Con base en la teoría del poblamiento de América fundamentada en el puente de Beringia, se determinó a los individuos pertenecientes a la región del norte de México como outgroup y a los individuos de la región del sur como ingroup (pop2, filogenéticamente cercana). Los cohortes finales del experimento fueron 51 individuos de 16 grupos indígenas del

centro diferentes, 18 individuos de 8 grupos indígenas del sur diferentes y 7 individuos de 3 grupos indígenas del norte diferentes.

7.3.3 PRETRATAMIENTO DE LOS DATOS.

El VCF con datos de genoma completo para los 76 individuos Nativos Mexicanos fue prefiltrado con bcftools [90] para obtener únicamente variantes en los cromosomas 1-22 y X. El resultado de este primer proceso fue posteriormente filtrado para conservar variantes exclusivamente bialélicas y con 0% de missing data. Después de los procesos de curación, el archivo VCF preservo 8982521 variantes.

Para obtener un archivo VCF exclusivo del análisis iHS, se utilizó VCFTools para conservar exclusivamente a los individuos pertenecientes a la región central de México. Posteriormente, con el script split_vcf.py se obtuvo un archivo VCF por cada uno de los cromosomas 1-22 y X. Las referencias genómicas para el faseado de datos en su versión GChr38 fueron descargadas del repositorio oficial del proyecto 1000 Genomas. Se utilizaron los mapas genómicos proporcionados por SHAPEIT4 para realizar una interpolación de distancias genéticas para cada variante presente en los archivos de referencia con predictGMAP (https://github.com/szpiech/predictGMAP). Finalmente, se descargó un archivo mart en Ensembl para realizar la anotación genómica de las variantes

Con el path de cada uno de los archivos previamente citados se escribieron los dos inputs principales del pipeline (Archivos CSV) y se realizó el análisis en un servidor externo.

8. RESULTADOS.

La implementación de este proyecto ha permitido el desarrollo de una herramienta bioinformática que logra la detección de señales de selección con los estadísticos iHS y PBS. La automatización del código a través de la generación de programas, softwares o pipelines es un proceso fundamental en el área bioinformática. A continuación, se enlistan los resultados obtenidos para el conjunto de datos del proyecto 100G-MX.

8.1 RESULTADOS: PBS.

El cómputo de PBS bajo el modelo experimental previamente descrito, permitió la obtención de >28,000 variantes que exceden el corte mínimo de 0.2 en el estadístico (**Figura 11 y 12A**). Dentro de los 23 cromosomas incluídos en el análisis, el cromosoma 2 presentó la mayor cantidad de variantes bajo selección putativa.

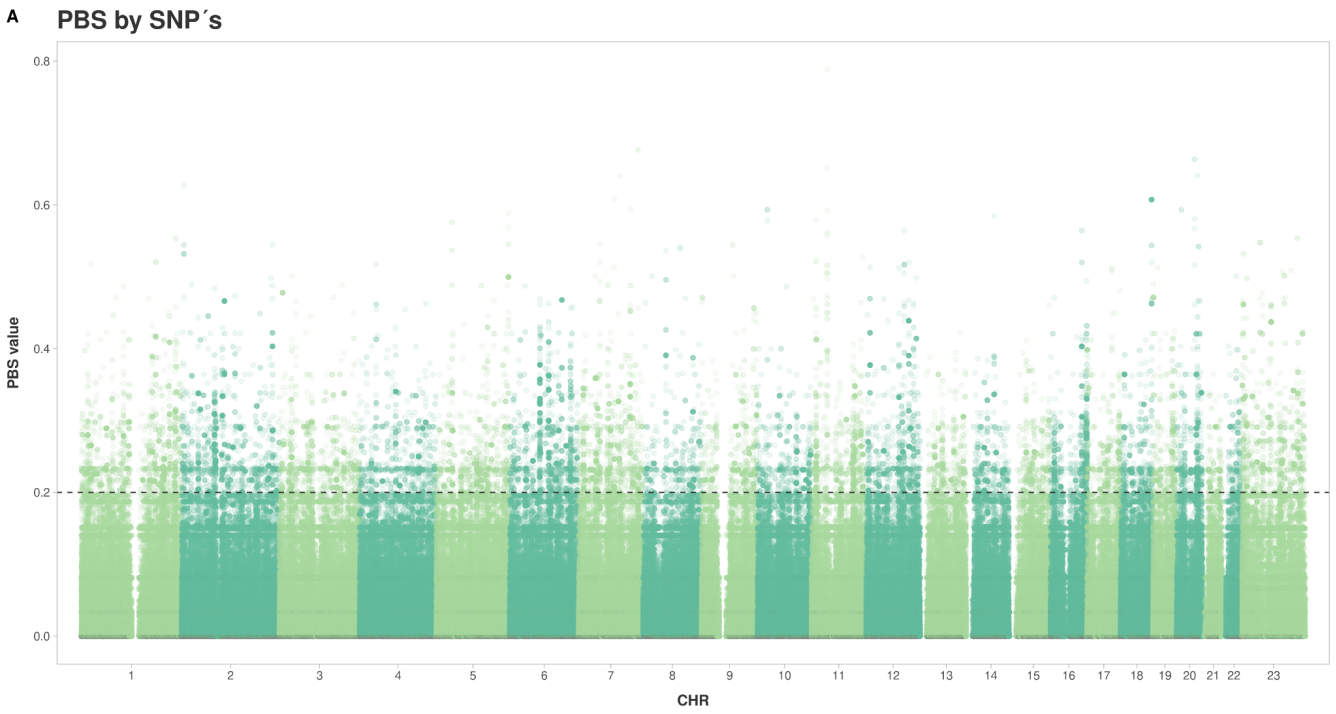


FIGURA 11 | Manhattan plot PBS. Distribución genómica de los valores de PBS calculados por nf-selection. Línea negra punteada representa un corte de 0.2. Variantes arriba de este valor son consideradas candidatas bajo selección.

El histograma de frecuencias de PBS (**Figura 12A**) demuestra que la mayor proporción de variantes se distribuyen en los bines de menor valor. Mientras que,

existen pocas variantes que alcanzan valores de PBS en el rango de 0.5 a 0.8. Por otro lado, la variante con el valor de PBS más elevado (**Figura 12B)** presenta una frecuencia alélica equivalente a 0.7142 para la población outgroup, 0.1388 para la población ingroup y 0.0098 para la población de interés. La disminución de la frecuencia alélica asociada con la variante alterna puede ser indicativa de un evento de selección negativa.

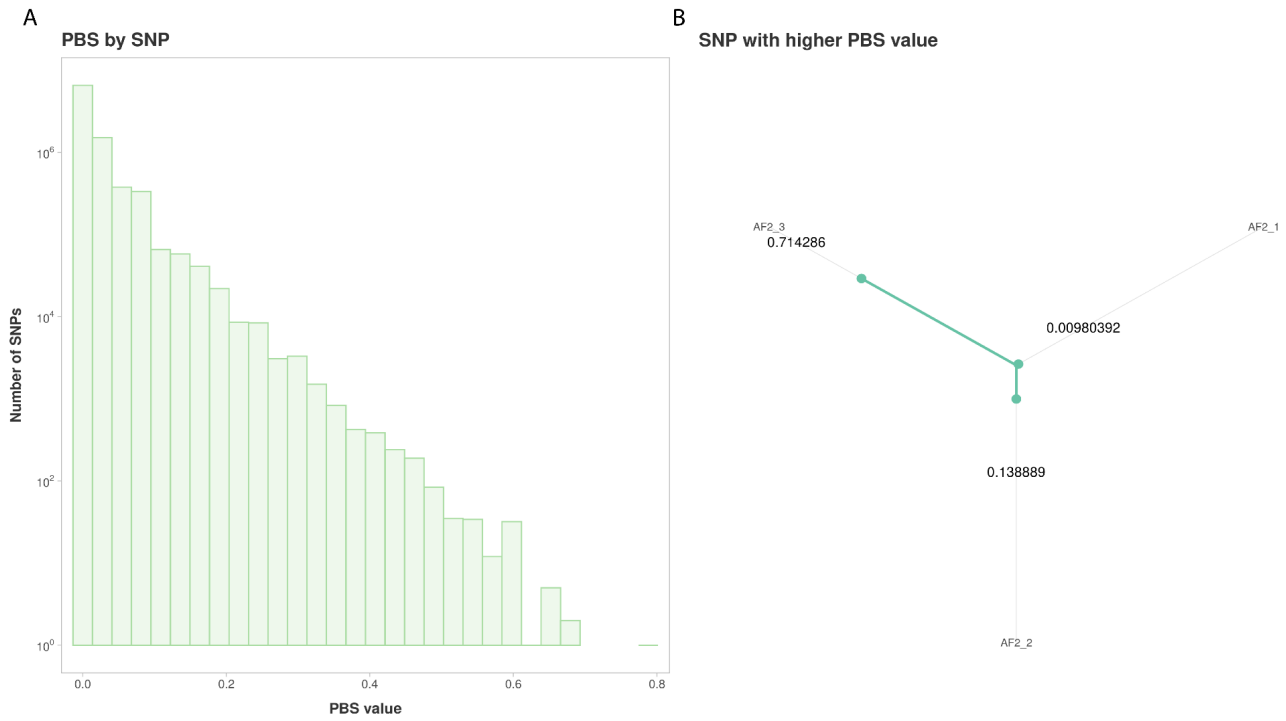


FIGURA 12 | A) Histograma de valores de PBS por variante. **B)** Spider plot con las frecuencias alélicas de la variante con mayor valor de PBS

Los genes *CNTNAP2*, *LINC01620, COL26A1, LINC01879 y ARL2BPP1* pobraron tener hipótesis de selección al exhibir variantes con valores de PBS > 0.6. *CNTNAP2* es un gen encargado de codificar a Caspr2, una proteína involucrada con el funcionamiento del sistema nervioso y ampliamente relacionada con desórdenes del neurodesarrollo [91]. Por otro lado, *LINC01620* y *LINC01879* codifican para dos RNA largos no codificantes para proteínas [99]. *COL26A1* es un gen encargado de codificar la cadena alfa del colágeno tipo XXVI y se ha asociado con reacciones alérgicas exhacerbadas a fármacos anti-inflamatorios no esteroideos [92]. Finalmente, *ARL2BPP1* es un pseudogen codificante de una proteína de unión a GTPasa [99].

8.2 RESULTADOS: iHS

El cómputo de iHS en las poblaciones nativas pertenecientes a la región central de México, arrojó aproximadamente 245649 variantes que exceden el corte mínimo de 0.2 en el estadístico (**Figura 13 y 14**). De la misma forma que el PBS, el cromosoma 2 presentó la mayor cantidad de variantes bajo selección putativa. El valor de MAF establecido en el pipeline fue de 0.01, por lo cual únicamente se obtuvieron valores de iHS para el 58% de las variantes. El histograma de frecuencias de iHS (**Figura 14**) demuestra que la mayor proporción de variantes se distribuyen en los bines de menor valor. Mientras que, existen pocas variantes que alcanzan valores de iHS en el rango de 5 a 10.

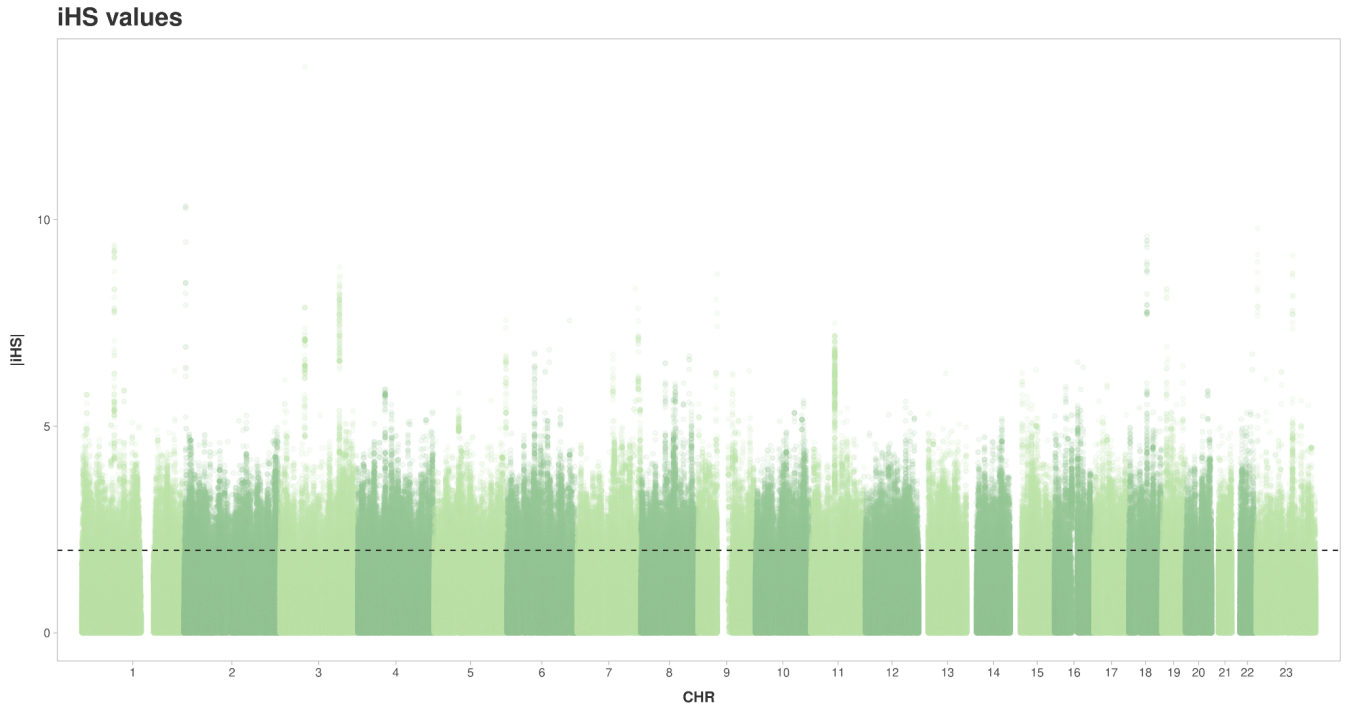


FIGURA 13 | Manhattan plot iHS. Distribución genómica de los valores de iHS calculados por nf-selection. Línea negra punteada representa un corte de 2. Variantes arriba de este valor son consideradas candidatas bajo selección.

Notablemente, la variante con iHS más elevado (iHS = 13.69) fue mapeada en el gen *DOCK3,* el cual se ha relacionado con fenotipos de altura adaptativos en poblaciones africanas [93,94]. Adicionalmente, existen una amplia variedad de genes que pobraron tener hipótesis de selección al exhibir variantes con valor de iHS elevados. En este apartado se describen los cinco genes con valores más

destacados. *KIAA1328* es el gen codificante de la hinderina, una proteína relacionada con el mantenimiento de la estabilidad cromosómica [95, 99]. De forma interesante, múltiples variantes con valores de iHS superiores a 2.0 fueron identificados en *KIAA1328*. Por otro lado, *MIER1* codifica a la proteína 1 de respuesta temprana a la inducción del mesodermo (MIER1) [99]. MIER1 funciona como un regulador transcripcional que recluta a HDAC1, una histona relacionada con el silenciamiento de la cromantina [96]. Al igual que *KIAA1328*, existen múltiples variantes con valores elevados de iHS en este gen. *STXBP5L* es un gen codificador de la proteína de unión a sintaxina 5L, la cual se ha relacionado con el tráfico vesicular y la secreción de neurotransmisores [99]. Notoriamente, las señales de selección en genes neurológicos son observadas de forma regular en poblaciones que viven en altitudes con niveles de oxígeno variables [97]. Adicionalmente, *RNASEH1* codifica la proteína Ribonucleasa H1 encargada de degradar específicamente el ARN de los híbridos ARN-ADN; mutaciones en este gen se han relacionado con oftalmoplejía externa progresiva [98]. Finalmente, *ZNF812P* es un pseudogen [99].

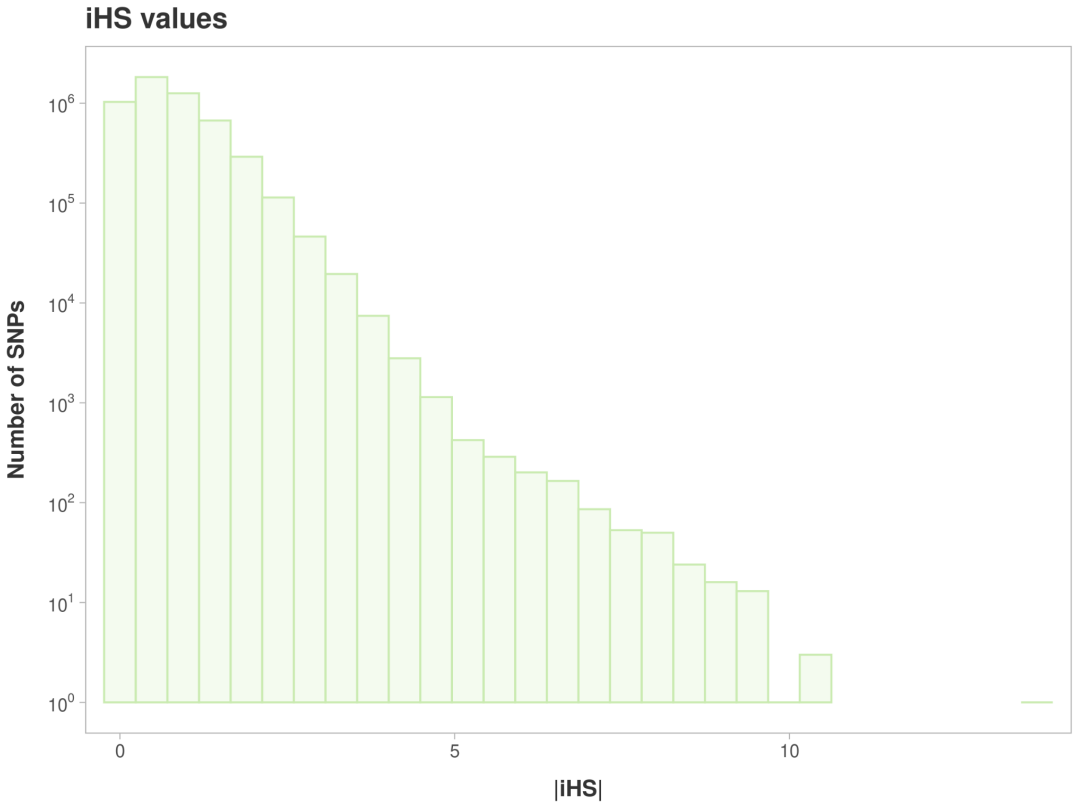


FIGURA 14 | Histograma de valores de iHS por variante core.

8.3 RESULTADOS: PBS VERSUS iHS

El resultado de la interesección entre el 1% de los valores más elevados de cada estadístico fueron 463 variantes distribuidas a través de los autosomas y el cromosomas sexual X de las poblaciones nativas mexicanas de la región centro. El circus plot de la **Figura 15** resume los resultados obtenidos para cada estadístico y destaca en rojo las posiciones genómicas producto de la referencia cruzada entre PBS e iHS. La anotación genómica de las variantes permitió identificar 159 genes distintos.

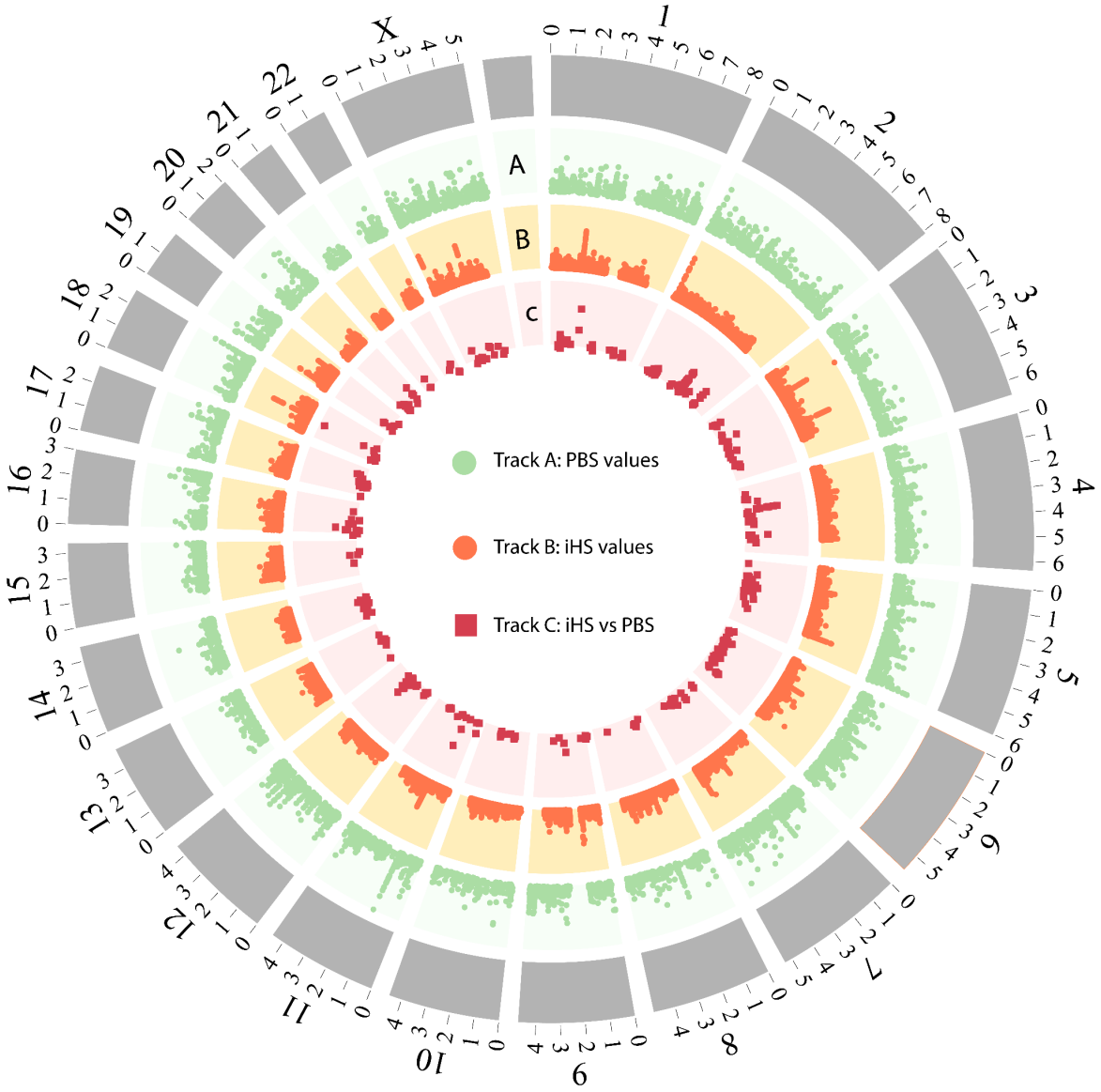


FIGURA 15 | Circus plot. Track A. Distribución de variantes con valores de PBS mayores a 0.2. Track B. Distribución de variantes con valores de iHS. Track C. Intersección entre el 1% de los valores más elevados de PBS e iHS.

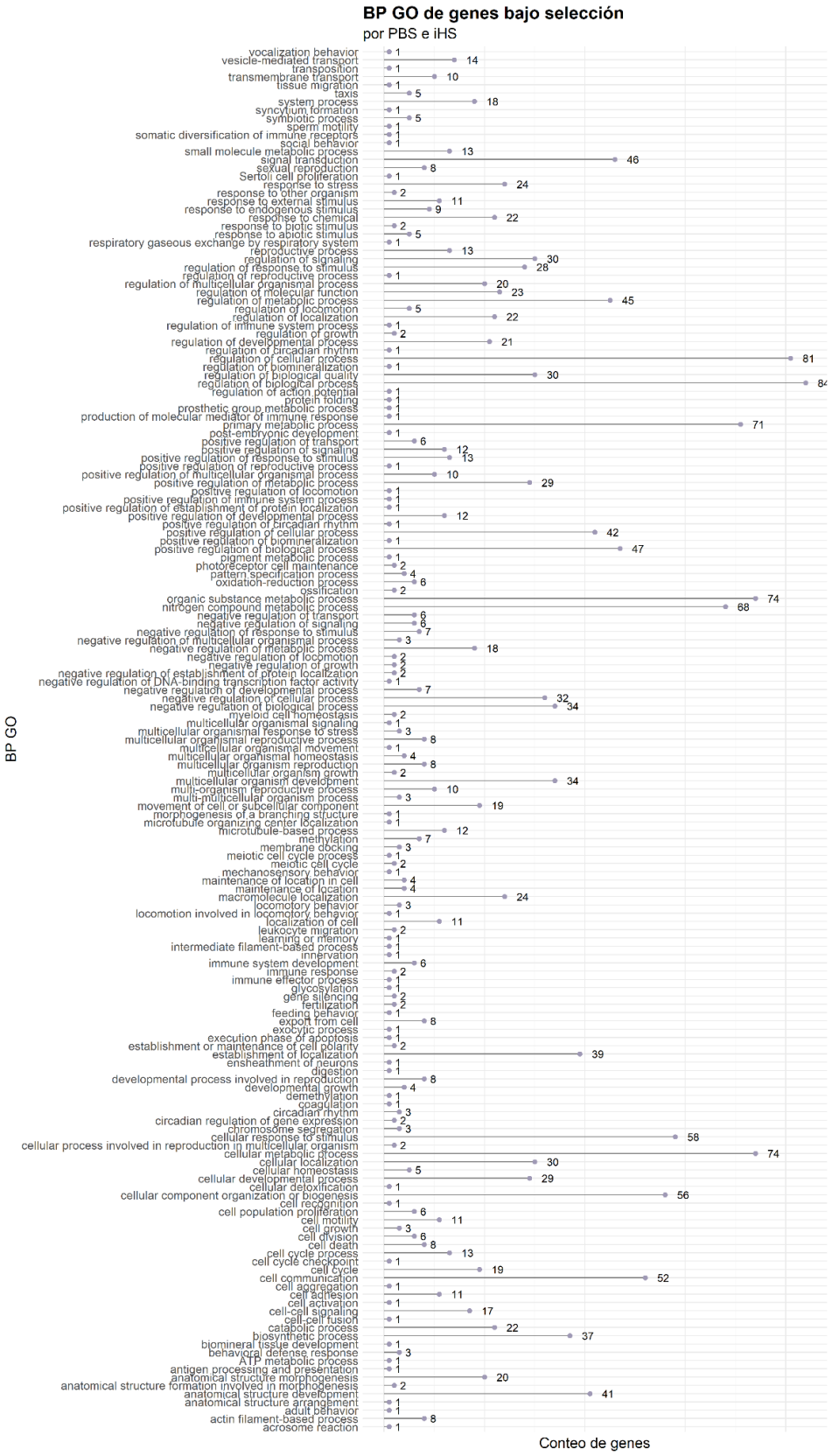


FIGURA 16 | Clasificación ontológica de genes bajo selección. Categoría utilizada: Biological Processes.

Utilizando el paquete clusterProfiler [100] se realizó una clasificación ontológica de genes para identificar procesos biológicos (BP GO) relacionados con los genes bajo hipótesis de selección (**Figura 16 y Figura 17**). Señales de selección involucradas con las categorías inmunológicas fueron encontradas en los siguientes genes: *CD164* codifica una sialomucina transmembranal involucrada con la proliferación, adhesión y migración de células hematopoyéticas [99]. Notablemente, *CD164* se ha relacionado con adaptación positiva a virus patogénicos [101]. *CXCL9*, es un gen antimicrobiano codificante de una quimiocina inducida por interferón gamma involucrada en procesos inflamatorios e inmunoregulatorios [102]; *ARMC6*, *SOX6*, *MSH3*, *ACVR2A*, *ESCO2* son genes involucrados en el desarrollo y regulación del sistema inmune [103]; *RAB3C* codifica una proteína GTPasa pequeña encargada del tráfico de vesículas, y el procesamiento y presentación de antígenos en el MHC de clase 1 [99].

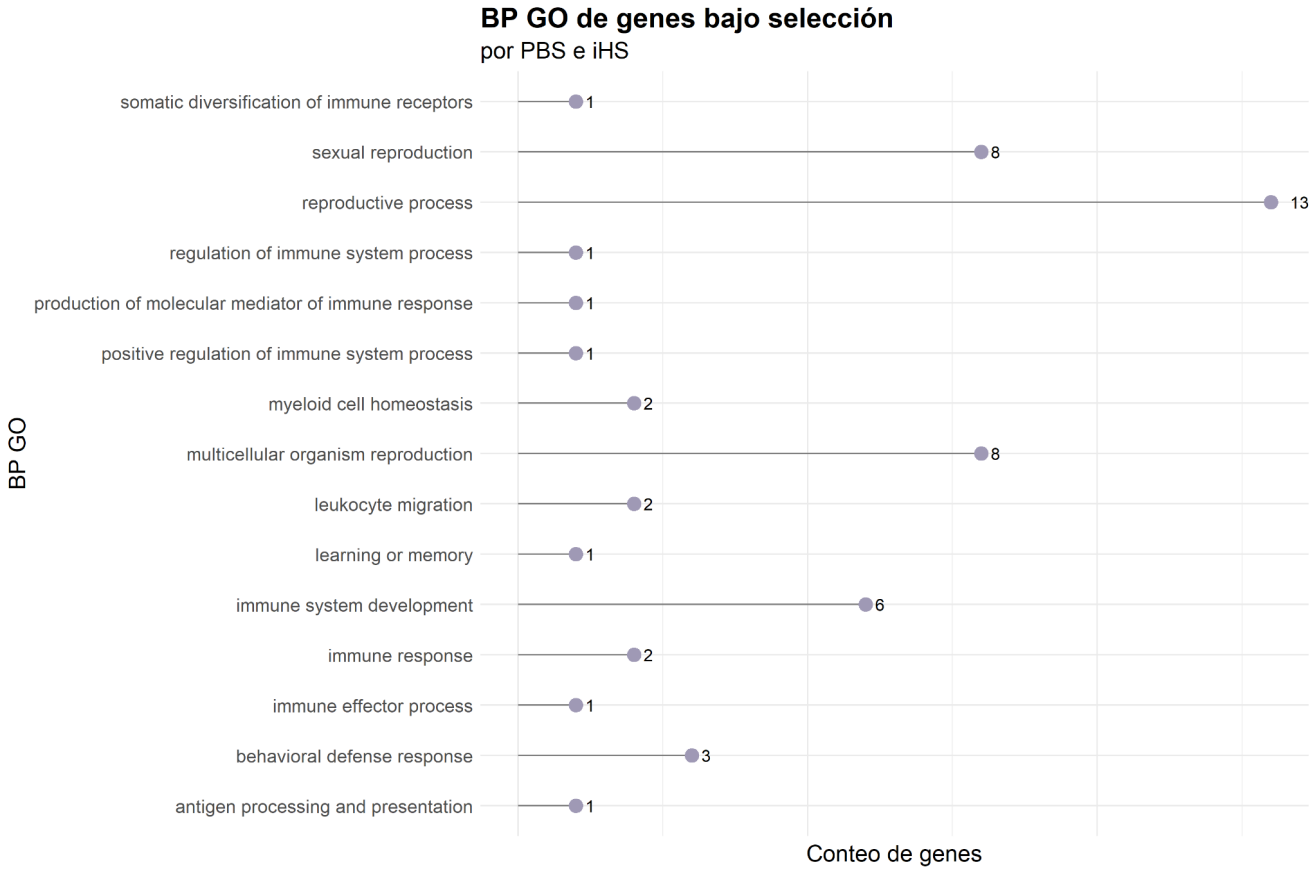


FIGURA 17 | Clasificación ontológica de genes bajo selección. Categoría utilizada: Biological Processes. Se muestran las 15 categorías más significativas.

Adicionalmente, *NRG1* — gen codificante de la neuregulina — mostró hipótesis de selección en las poblaciones de interés. *NRG1* posee evidencia significativa de selección positiva en el cohorte del proyecto HapMap y ha sido de amplio interés por su rol fundamental en procesos neuronales [105]. Por otro lado, los genes *THRAP3, MYCBP2* y *PRMT5* reportan un papel en la regulación del ciclo circadiano, un proceso previamente descrito bajo selección en poblaciones Nativas Mexicanas [13]. Los genes candidatos restantes están involucrados en fertilidad y reproducción, transducción de señales, comunicación celular, procesos catabolicos, regulación de procesos metabólicos, homeostasis, aprendizaje y memoria, entre otros.

9. DISCUSIÓN

La habilidad de detectar e identificar señales de selección ha adquirido un interés creciente en el ámbito de la genómica poblacional. A lo largo de este trabajo se ha presentado una herramienta bioinformática que logra identificar señales de selección en organismos diploides. El flujo de trabajo permite al usuario partir de datos de genotipificación codificados en un archivo VCF y obtener, como resultado final, regiones del genoma que exhiben hipótesis de selección confirmadas por PBS e iHS. Particularmente, el uso de Nextflow como un orquestador de flujo de trabajo ofrece la posibilidad de paralelizar y agilizar los procesos bioinformáticos requeridos; el análisis realizado con 23 cromosomas, 76 individuos y ≈9,000,000 de variantes — en un servidor Linux con 12 cores — requirió 4h 38m y 34s de tiempo computacional. Un beneficio adicional de **nf-selection** es el desarrollo de un script en R para el cálculo automatizado de PBS, una implementación que, a conocimiento de los autores, no existía con anterioridad. Además, el pipeline fue planteado para utilizar herramientas o softwares de libre acceso.

Existen una serie de mejoras que podrían aplicarse directamente a la herramienta desarrollada. Para empezar, el modelado demográfico — planteado actualmente como una herramienta ajena al pipeline principal — debería implementarse como una rama de desarrollo que evite los pasos intermedios y de manipulación manual que debe llevar a cabo el usuario. Asimismo, el estadístico iHS computado por el software hapbin no permite la obtención de resultados estadísticamente significativos (en contraste con el paquete rehh del ambiente R que arroja p-values para cada región core). Adicionalemente, la intersección entre PBS e iHS posibilita la detección de señales de selección bajo la misma escala temporal y el mismo modelo de selección. Para obtener un panorama global más completo, los autores sugieren añadir estadísticos que detecten barridos selectivos completos, entre ellos, el EHH. Finalmente, es necesario reconocer que el desarrollo del pipeline fue restringido a datos de genoma, transcriptoma o exoma completo, por lo que datos de genotipificación obtenidos a través de arreglos deben ser sometidos a pretratamientos bioinformáticos. A pesar de lo anteriormente citado, se cree que el pipeline representa una herramienta ampliamente útil para detectar el impacto de la

selección en individuos de diversas especies, por lo que se espera resulte útil para la comunidad científica.

Para someter la herramienta desarrollada a un proceso de evaluación, se optó por realizar la detección del impacto de la selección natural en 76 genomas Nativos Mexicanos. Históricamente, los estudios dedicados a la detección de señales de selección se han enfocado en poblaciones pertenecientes a Europa, África y Asia, subrepresentando a poblaciones Nativas Americanas (NatAm) y dejando el panorama global incompleto. Sabiendo que la detección de variantes o genes bajo presión selectiva es evidencia de procesos evolutivos y permite la identificación de posibles variantes clínicamente funcionales, se considera imperativo el estudio de señales de selección en datos de genoma completo pertenecientes a poblaciones NatAm.

Los estudios que han volcado su atención en el tema, han identificado en mayor proporción variantes bajo selección en genes relacionados con el sistema inmune. Esto último puede reflejar las respuestas a patógenos endémicos del continente Americano y la adaptación ante las enfermedades infecciosas que conllevó la colonización Europea [106]. Asimismo, diversas investigaciones han reportado genes involucrados en procesos metabólicos bajo presión selectiva fuerte; consistente con las nuevas prácticas dietéticas a las que se expusieron las comunidades al poblar el continente Americano [106]. De forma congruente, el análisis de regiones bajo selección en las poblaciones Nativas Mexicanas de la región centro, arrojó genes involucrados directamente con la respuesta inmunológica.

De interés, *CD164* es un gen bajo selección relacionado con la respuesta antiretroviral innata y se ha demostrado su papel específico durante infecciones provocadas por el virus VIH [101]. Por otro lado, el ligando de la quimiocina 9 codificado por el gen *CXCL9,* juega un papel fundamental en el reclutamiento de células efectoras (Linfocitos B y linfocitos T) en los sitios de inflamación [109]. CXCL9 también regula la acumulación de eosinófilos en el asma, la prognosis en enfermedades infecciosas[108] y posee una actividad angio estática que inhibe la expansión de vasos sanguíneos asociados a tumores [109]. Hipótesis de selección en genes codificantes de quimiocinas han sido planteadas con anterioridad en

poblaciones europeas del norte, representando una posible adaptación a patógenos ambientales [110]. Adicionalemente se identificaron señales de selección nuevas — no reportadas previamente en la bibliografía — en los genes *ARMC6*, *SOX6*, *MSH3*, *ACVR2A*, *ESCO2* y *RAB3C*, los cuales juegan un papel fundamental en el funcionamiento del sistema inmune. Particularmente, RAB3C codifica una GTPasa pequeña encargada del procesamiento y presentación de antígenos en el MHC de clase 1, este gen ha sido clasificado como un miembro de la familia de oncogenes RAS [99]. Curiosamente, RAB27A, otro gen perteneciente a la familia RAS, ha sido previamente identificado como un gen asociado a la pigmentación con hipótesis de selección en poblaciones con ancestría Nativa Americana [104]. Otra variante de interés considerable fue encontrada en *NRG1,* un gen de susceptibilidad para diversas alteraciones neurobiológicas. En contraste, no se detectó alguna variante que tuviera una relación directa con el metabolismo o dieta de las poblaciones Nativas Mexicanas.

Es necesario considerar que, por efectos prácticos, la detección de señales de selección en el cohorte de interés se realizó sin modelado demográfico. Asimismo, la elección de poblaciones para el PBS difirió de los paramétros usualmente considerados en la bibliografía.

10. CONCLUSIONES

Lo expuesto a lo largo de este trabajo permite llegar a las siguientes conclusiones:

1. La identificación de señales de selección en datos de secuenciación de genoma completo, permite dilucidar aspectos fundamentales de la evolución y comprender asociaciones directas entre genotipo y fenotipo adaptativo. Para este propósito, la bioinformática aborda las necesidades requeridas para el tratamiento de datos, aplicación de estadísticos y obtención de resultados.
2. El desarrollo de un pipeline bioinformático que estandariza la detección de señales de selección con PBS e iHS de forma escalable y reproducible, facilita la transición de datos de genotipificación a hipótesis de selección.
3. La aplicación de PBS e iHS en datos de secuenciación de genoma completo pertenecientes a poblaciones Nativas Mexicanas arrojó señales putativas en 463 variantes distribuidas a través de los autosomas y el cromosoma sexual X. La anotación genómica permitió identificar 159 genes distintos.
4. De interés, los genes *CD164, CXCL9, ARMC6*, *SOX6*, *MSH3*, *ACVR2A*, *ESCO2* y *RAB3C* presentaron variantes bajo hipótesis de selección. Estos últimos han demostrado un papel en el funcionamiento o regulación del sistema inmune; un fenotipo bajo selección característico de las poblaciones Nativas Mexicanas.

11. REFERENCIAS

1. Margulis, Lynn. “Ernst Mayr: What Evolution Is.” International Microbiology 5, no. 2 (May 29, 2002): 103–4. https://doi.org/10.1007/s10123-002-0072-1.
2. Vitti, Joseph J., Sharon R. Grossman, and Pardis C. Sabeti. “Detecting Natural Selection in Genomic Data.” *Annual Review of Genetics* 47, no. 1 (November 23, 2013): 97–120. https://doi.org/10.1146/annurev-genet-111212-133526.
3. Wood, Bernard. *Human Evolution: A Very Short Introduction*. Oxford University Press, 2019. http://dx.doi.org/10.1093/actrade/9780198831747.001.0001.
4. Hughes, A L. “Looking for Darwin in All the Wrong Places: The Misguided Quest for Positive Selection at the Nucleotide Sequence Level.” *Heredity* 99, no. 4 (July 11, 2007): 364–73. https://doi.org/10.1038/sj.hdy.6801031.
5. Sabeti, Pardis C., David E. Reich, John M. Higgins, Haninah Z. P. Levine, Daniel J. Richter, Stephen F. Schaffner, Stacey B. Gabriel, et al. “Detecting Recent Positive Selection in the Human Genome from Haplotype Structure.” Nature 419, no. 6909 (October 2002): 832–37. https://doi.org/10.1038/nature01140.
6. Horscroft, Clare, Sarah Ennis, Reuben J Pengelly, Timothy J Sluckin, and Andrew Collins. “Sequencing Era Methods for Identifying Signatures of Selection in the Genome.” *Briefings in Bioinformatics* 20, no. 6 (July 24, 2018): 1997–2008. https://doi.org/10.1093/bib/bby064.
7. Hejase, Hussein A., Noah Dukler, and Adam Siepel. “From Summary Statistics to Gene Trees: Methods for Inferring Positive Selection.” *Trends in Genetics* 36, no. 4 (April 2020): 243–58. https://doi.org/10.1016/j.tig.2019.12.008.
8. Cadzow, Murray, James Boocock, Hoang T. Nguyen, Phillip Wilcox, Tony R. Merriman, and Michael A. Black. “A Bioinformatics Workflow for Detecting Signatures of Selection in Genomic Data.” *Frontiers in Genetics* 5 (August 26, 2014). https://doi.org/10.3389/fgene.2014.00293.
9. Sabeti, P. C., S. F. Schaffner, B. Fry, J. Lohmueller, P. Varilly, O. Shamovsky, A. Palma, T. S. Mikkelsen, D. Altshuler, and E. S. Lander. “Positive Natural

Selection in the Human Lineage." *Science* 312, no. 5780 (June 16, 2006): 1614–20. https://doi.org/10.1126/science.1124309.

10. Pybus, Marc, Giovanni M. Dall'Olio, Pierre Luisi, Manu Uzkudun, Angel Carreño-Torres, Pavlos Pavlidis, Hafid Laayouni, Jaume Bertranpetit, and Johannes Engelken. "1000 Genomes Selection Browser 1.0: A Genome Browser Dedicated to Signatures of Natural Selection in Modern Humans." *Nucleic Acids Research* 42, no. D1 (November 25, 2013): D903–9. https://doi.org/10.1093/nar/gkt1188.
11. Yi, Xin, Yu Liang, Emilia Huerta-Sanchez, Xin Jin, Zha Xi Ping Cuo, John E. Pool, Xun Xu, et al. "Sequencing of 50 Human Exomes Reveals Adaptation to High Altitude." *Science* 329, no. 5987 (July 2, 2010): 75–78. https://doi.org/10.1126/science.1190371.
12. Reynolds, A. W., Mata-Míguez, J., Miró-Herrans, A., Briggs-Cloud, M., Sylestine, A., Barajas-Olmos, F., ... & Bolnick, D. A. (2019). Comparing signals of natural selection between three Indigenous North American populations. Proceedings of the National Academy of Sciences, 116(19), 9312-9317.
13. Aguilar-Ordoñez, I., Pérez-Villatoro, F., García-Ortiz, H., Barajas-Olmos, F., Ballesteros-Villascán, J., González-Buenfil, R., ... & Morett, E. (2021). Whole-genome variation in 27 Mexican indigenous populations, demographic and biomedical insights. PloS one, 16(4), e0249773.
14. Ávila-Arcos, M. C., McManus, K. F., Sandoval, K., Rodríguez-Rodríguez, J. E., Villa-Islas, V., Martin, A. R., ... & Moreno-Estrada, A. (2020). Population history and gene divergence in Native Mexicans inferred from 76 human exomes. Molecular biology and evolution, 37(4), 994-1006.
15. 1000 Genomes Project Consortium. (2015). A global reference for human genetic variation. Nature, 526(7571), 68.
16. Voight, Benjamin F, Sridhar Kudaravalli, Xiaoquan Wen, and Jonathan K Pritchard. "A Map of Recent Positive Selection in the Human Genome." PLOS Biology 4, no. 3 (March 7, 2006). https://doi.org/10.1371/journal.pbio.0040072.
17. Freeman, Scott. *Biological Science*. Benjamin Cummings, 2010.

18. Brown, Charles R., and Mary Bomberger Brown. “INTENSE NATURAL SELECTION ON BODY SIZE AND WING AND TAIL ASYMMETRY IN CLIFF SWALLOWS DURING SEVERE WEATHER.” *Evolution* 52, no. 5 (October 1998): 1461–75. https://doi.org/10.1111/j.1558-5646.1998.tb02027.x.
19. Karn, Mary N., Helen Lang-Brown, Helen MacKENZIE, and L. S. Penrose. “BIRTH WEIGHT, GESTATION TIME AND SURVIVAL IN SIBS.” *Annals of Eugenics* 15, no. 1 (January 1949): 306–22. https://doi.org/10.1111/j.1469-1809.1949.tb02450.x.
20. Smith, Thomas Bates. “Bill Size Polymorphism and Intraspecific Niche Utilization in an African Finch.” *Nature* 329, no. 6141 (October 1987): 717–19. https://doi.org/10.1038/329717a0.
21. “Redirect Notice.” Accessed April 26, 2022. https://www.google.com/url?q=https://www.cdc.gov/ncbddd/spanish/thalassemia/facts.html%23:~:text%3DLa%2520talasemia%2520es%2520un%2520trastorno,importante%2520de%2520los%2520gl%25C3%25B3bulos%2520rojos&sa=D&source=docs&ust=1650948050414609&usg=AOvVaw2U4znD67TpJm4ttz1daIY4.
22. Allison, A. C. “Protection Afforded by Sickle-Cell Trait Against Subtertian Malarial Infection.” *BMJ* 1, no. 4857 (February 6, 1954): 290–94. https://doi.org/10.1136/bmj.1.4857.290.
23. Price, George R. “Selection and Covariance.” *Nature* 227, no. 5257 (August 1970): 520–21. https://doi.org/10.1038/227520a0.
24. Robertson, Alan. “A Mathematical Model of the Culling Process in Dairy Cattle.” *Animal Science* 8, no. 1 (February 1966): 95–108. https://doi.org/10.1017/s0003356100037752.
25. Lande, Russell, and Stevan J. Arnold. “THE MEASUREMENT OF SELECTION ON CORRELATED CHARACTERS.” *Evolution* 37, no. 6 (November 1983): 1210–26. https://doi.org/10.1111/j.1558-5646.1983.tb00236.x.
26. Gould, S. J. (1978). Sociobiology: the art of storytelling. New Scientist, 80(1129), 530-33.

27. Kelley, Joanna. “Detecting Natural Selection in the Genome.” *Oxford Bibliographies Online Datasets*, March 30, 2017. https://doi.org/10.1093/obo/9780199941728-0088.
28. Learn Science at Scitable. “A Selective Sweep.” Accessed April 26, 2022. https://www.nature.com/scitable/content/a-selective-sweep-24827/.
29. Hermisson, Joachim, and Pleuni S. Pennings. “Soft Sweeps and beyond: Understanding the Patterns and Probabilities of Selection Footprints under Rapid Adaptation.” *Methods in Ecology and Evolution* 8, no. 6 (June 2017): 700–716. https://doi.org/10.1111/2041-210x.12808.
30. Pavlidis, Pavlos, and Nikolaos Alachiotis. “A Survey of Methods and Tools to Detect Recent and Strong Positive Selection.” *Journal of Biological Research-Thessaloniki* 24, no. 1 (April 8, 2017). https://doi.org/10.1186/s40709-017-0064-0.
31. Slatkin, Montgomery. “Linkage Disequilibrium — Understanding the Evolutionary Past and Mapping the Medical Future.” *Nature Reviews. Genetics* 9, no. 6 (June 1, 2008). https://doi.org/10.1038/nrg2361.
32. Stephan, Wolfgang. “Genetic Hitchhiking versus Background Selection: The Controversy and Its Implications.” *Philosophical Transactions of the Royal Society B: Biological Sciences* 365, no. 1544 (April 27, 2010). https://doi.org/10.1098/rstb.2009.0278.
33. Tajima, F. “Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism.” *Genetics* 123, no. 3 (November 1, 1989): 585–95. https://doi.org/10.1093/genetics/123.3.585.
34. Ma, Y, X Ding, S Qanbari, S Weigend, Q Zhang, and H Simianer. “Properties of Different Selection Signature Statistics and a New Strategy for Combining Them.” *Heredity* 115, no. 5 (May 20, 2015): 426–36. https://doi.org/10.1038/hdy.2015.42.
35. Carlson, Christopher S., Daryl J. Thomas, Michael A. Eberle, Johanna E. Swanson, Robert J. Livingston, Mark J. Rieder, and Deborah A. Nickerson. “Genomic Regions Exhibiting Positive Selection Identified from Dense Genotype Data.” *Genome Research* 15, no. 11 (October 26, 2005): 1553–65. https://doi.org/10.1101/gr.4326505.

36. Fay, Justin C, and Chung-I Wu. “Hitchhiking Under Positive Darwinian Selection.” *Genetics* 155, no. 3 (July 1, 2000): 1405–13. https://doi.org/10.1093/genetics/155.3.1405.
37. Fu, Y.X. “Statistical Properties of Segregating Sites.” *Theoretical Population Biology* 48, no. 2 (October 1995): 172–97. https://doi.org/10.1006/tpbi.1995.1025.
38. Fu, Y.X. “Statistical Properties of Segregating Sites.” Theoretical Population Biology 48, no. 2 (October 1995): 172–97. https://doi.org/10.1006/tpbi.1995.1025.
39. Sabeti, Pardis C., Patrick Varilly, Ben Fry, Jason Lohmueller, Elizabeth Hostetter, Chris Cotsapas, Xiaohui Xie, et al. “Genome-Wide Detection and Characterization of Positive Selection in Human Populations.” *Nature* 449, no. 7164 (October 2007): 913–18. https://doi.org/10.1038/nature06250.
40. Pollard, Katherine S., Sofie R. Salama, Nelle Lambert, Marie-Alexandra Lambot, Sandra Coppens, Jakob S. Pedersen, Sol Katzman, et al. “An RNA Gene Expressed during Cortical Development Evolved Rapidly in Humans.” *Nature* 443, no. 7108 (August 16, 2006): 167–72. https://doi.org/10.1038/nature05113.
41. Smith, John Maynard, and John Haigh. “The Hitch-Hiking Effect of a Favourable Gene.” *Genetical Research* 23, no. 1 (February 1974): 23–35. https://doi.org/10.1017/s0016672300014634.
42. Oleksyk, Taras K., Michael W. Smith, and Stephen J. O’Brien. “Genome-Wide Scans for Footprints of Natural Selection.” *Philosophical Transactions of the Royal Society B: Biological Sciences* 365, no. 1537 (January 12, 2010): 185–205. https://doi.org/10.1098/rstb.2009.0219.
43. Zhang, C., D. K. Bailey, T. Awad, G. Liu, G. Xing, M. Cao, V. Valmeekam, et al. “A Whole-Genome Long-Range Haplotype (WGLRH) Test for Detecting Imprints of Positive Selection in Human Populations.” *Bioinformatics* 22, no. 17 (July 15, 2006): 2122–28. https://doi.org/10.1093/bioinformatics/btl365.
44. Wang, Eric T., Greg Kodama, Pierre Baldi, and Robert K. Moyzis. “Global Landscape of Recent Inferred Darwinian Selection for Homo Sapiens.” *Proceedings of the National Academy of Sciences* 103, no. 1 (December 21, 2005): 135–40. https://doi.org/10.1073/pnas.0509691102.

45. Cai, Zheng, Nicola J Camp, Lisa Cannon-Albright, and Alun Thomas. “Identification of Regions of Positive Selection Using Shared Genomic Segment Analysis.” *European Journal of Human Genetics* 19, no. 6 (February 9, 2011): 667–71. https://doi.org/10.1038/ejhg.2010.257.
46. Han, Lide, and Mark Abney. “Using Identity by Descent Estimation with Dense Genotype Data to Detect Positive Selection.” *European Journal of Human Genetics* 21, no. 2 (July 11, 2012): 205–11. https://doi.org/10.1038/ejhg.2012.148.
47. Tishkoff, Sarah A., and Scott M. Williams. “Genetic Analysis of African Populations: Human Evolution and Complex Disease.” *Nature Reviews Genetics* 3, no. 8 (August 2002): 611–21. https://doi.org/10.1038/nrg865.
48. Mellars, Paul. “Going East: New Genetic and Archaeological Perspectives on the Modern Human Colonization of Eurasia.” *Science* 313, no. 5788 (August 11, 2006): 796–800. https://doi.org/10.1126/science.1128402.
49. Stringer, Chris. “Modern Human Origins: Progress and Prospects.” *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 357, no. 1420 (April 29, 2002): 563–79. https://doi.org/10.1098/rstb.2001.1057.
50. White, Tim D., Berhane Asfaw, David DeGusta, Henry Gilbert, Gary D. Richards, Gen Suwa, and F. Clark Howell. “Pleistocene Homo Sapiens from Middle Awash, Ethiopia.” *Nature* 423, no. 6941 (June 2003): 742–47. https://doi.org/10.1038/nature01669.
51. Rockman, MatthewV. “Human Evolutionary Genetics: Origins, Peoples, and Disease (2004).” *Human Genetics* 115, no. 2 (May 13, 2004). https://doi.org/10.1007/s00439-004-1138-2.
52. Fan, Shaohua, Matthew E. B. Hansen, Yancy Lo, and Sarah A. Tishkoff. “Going Global by Adapting Local: A Review of Recent Human Adaptation.” *Science* 354, no. 6308 (October 7, 2016): 54–59. https://doi.org/10.1126/science.aaf5098.
53. Bersaglieri, Todd, Pardis C. Sabeti, Nick Patterson, Trisha Vanderploeg, Steve F. Schaffner, Jared A. Drake, Matthew Rhodes, David E. Reich, and Joel N. Hirschhorn. “Genetic Signatures of Strong Recent Positive Selection at the

Lactase Gene." *The American Journal of Human Genetics* 74, no. 6 (June 2004): 1111–20. https://doi.org/10.1086/421051.

54. Lachance, Joseph, and Sarah A. Tishkoff. "Population Genomics of Human Adaptation." *Annual Review of Ecology, Evolution, and Systematics* 44, no. 1 (November 23, 2013): 123–43. https://doi.org/10.1146/annurev-ecolsys-110512-135833.
55. Tishkoff, Sarah A, Floyd A Reed, Alessia Ranciaro, Benjamin F Voight, Courtney C Babbitt, Jesse S Silverman, Kweli Powell, et al. "Convergent Adaptation of Human Lactase Persistence in Africa and Europe." *Nature Genetics* 39, no. 1 (December 10, 2006): 31–40. https://doi.org/10.1038/ng1946.
56. Cavalli-Sforza, L. L. (1973). Analytic review: some current problems of human population genetics. *American journal of human genetics*, *25*(1), 82.
57. Jablonski, Nina G., and George Chaplin. "Human Skin Pigmentation as an Adaptation to UV Radiation." *Proceedings of the National Academy of Sciences* 107, no. supplement_2 (May 5, 2010): 8962–68. https://doi.org/10.1073/pnas.0914628107.
58. Smith, Samuel StanhopeHG. *An Essay on the Causes of the Variety of Complexion and Figure in the Human Species*. Harvard University Press, 1965. http://dx.doi.org/10.4159/harvard.9780674866331.
59. "IV. An Essay upon the Causes of the Different Colours of People in Different Climates; by John Mitchell, M. D. Communicated to the Royal Society by Mr. Peter Collinson, F. R. S." *Philosophical Transactions of the Royal Society of London* 43, no. 474 (December 31, 1744): 102–50. https://doi.org/10.1098/rstl.1744.0033.
60. Han, J., G. A. Colditz, and D. J. Hunter. "Polymorphisms in the MTHFR and VDR Genes and Skin Cancer Risk." *Carcinogenesis* 28, no. 2 (July 8, 2006): 390–97. https://doi.org/10.1093/carcin/bgl156.
61. Turchin, Michael C, Charleston WK Chiang, Cameron D Palmer, Sriram Sankararaman, David Reich, and Joel N Hirschhorn. "Evidence of Widespread Selection on Standing Variation in Europe at Height-Associated SNPs." *Nature Genetics* 44, no. 9 (August 19, 2012): 1015–19. https://doi.org/10.1038/ng.2368.

62. Who, World Health Organization: “Malaria.” *World Health Organization: WHO*, April 6, 2022. https://www.who.int/news-room/fact-sheets/detail/malaria.
63. Kwiatkowski, Dominic P. “How Malaria Has Affected the Human Genome and What Human Genetics Can Teach Us about Malaria.” *The American Journal of Human Genetics* 77, no. 2 (August 2005): 171–92. https://doi.org/10.1086/432519.
64. Genovese, Giulio, David J. Friedman, Michael D. Ross, Laurence Lecordier, Pierrick Uzureau, Barry I. Freedman, Donald W. Bowden, et al. “Association of Trypanolytic ApoL1 Variants with Kidney Disease in African Americans.” *Science* 329, no. 5993 (August 13, 2010): 841–45. https://doi.org/10.1126/science.1193032.
65. Barreiro, Luis B, Guillaume Laval, Hélène Quach, Etienne Patin, and Lluís Quintana-Murci. “Natural Selection Has Driven Population Differentiation in Modern Humans.” *Nature Genetics* 40, no. 3 (February 3, 2008): 340–45. https://doi.org/10.1038/ng.78.
66. “A Second Generation Human Haplotype Map of over 3.1 Million SNPs.” *Nature* 449, no. 7164 (October 2007): 851–61. https://doi.org/10.1038/nature06258.
67. Yi, Xin, Yu Liang, Emilia Huerta-Sanchez, Xin Jin, Zha Xi Ping Cuo, John E. Pool, Xun Xu, et al. “Sequencing of 50 Human Exomes Reveals Adaptation to High Altitude.” *Science* 329, no. 5987 (July 2, 2010): 75–78. https://doi.org/10.1126/science.1190371.
68. Scheinfeldt, Laura B, and Sarah A Tishkoff. “Living the High Life: High-Altitude Adaptation.” *Genome Biology* 11, no. 9 (2010): 133. https://doi.org/10.1186/gb-2010-11-9-133.
69. Beall, Cynthia M., Michael J. Decker, Gary M. Brittenham, Irving Kushner, Amha Gebremedhin, and Kingman P. Strohl. “An Ethiopian Pattern of Human Adaptation to High-Altitude Hypoxia.” *Proceedings of the National Academy of Sciences* 99, no. 26 (December 5, 2002): 17215–18. https://doi.org/10.1073/pnas.252649199.
70. Fujimoto, Akihiro, Jun Ohashi, Nao Nishida, Taku Miyagawa, Yasuyuki Morishita, Tatsuhiko Tsunoda, Ryosuke Kimura, and Katsushi Tokunaga. “A Replication Study Confirmed the EDAR Gene to Be a Major Contributor to

Population Differentiation Regarding Head Hair Thickness in Asia." *Human Genetics* 124, no. 2 (August 13, 2008): 179–85. https://doi.org/10.1007/s00439-008-0537-1.

71. Park, Jeong-Heuy, Tetsutaro Yamaguchi, Chiaki Watanabe, Akira Kawaguchi, Kuniaki Haneji, Mayako Takeda, Yong-Il Kim, et al. "Effects of an Asian-Specific Nonsynonymous EDAR Variant on Multiple Dental Traits." *Journal of Human Genetics* 57, no. 8 (May 31, 2012): 508–14. https://doi.org/10.1038/jhg.2012.60.
72. Perry, George H, Nathaniel J Dominy, Katrina G Claw, Arthur S Lee, Heike Fiegler, Richard Redon, John Werner, et al. "Diet and the Evolution of Human Amylase Gene Copy Number Variation." *Nature Genetics* 39, no. 10 (September 9, 2007): 1256–60. https://doi.org/10.1038/ng2123.
73. Simonson, Tatum S., Yingzhong Yang, Chad D. Huff, Haixia Yun, Ga Qin, David J. Witherspoon, Zhenzhong Bai, et al. "Genetic Evidence for High-Altitude Adaptation in Tibet." *Science* 329, no. 5987 (July 2, 2010): 72–75. https://doi.org/10.1126/science.1189406.
74. Lindo, John, Emilia Huerta-Sánchez, Shigeki Nakagome, Morten Rasmussen, Barbara Petzelt, Joycelynn Mitchell, Jerome S. Cybulski, Eske Willerslev, Michael DeGiorgio, and Ripan S. Malhi. "A Time Transect of Exomes from a Native American Population before and after European Contact." *Nature Communications* 7, no. 1 (November 15, 2016). https://doi.org/10.1038/ncomms13175.
75. Moore, Lorna G., Stacy Zamudio, Jianguo Zhuang, Shinfu Sun, and Tarshi Droma. "Oxygen Transport in Tibetan Women during Pregnancy at 3,658 m." *American Journal of Physical Anthropology* 114, no. 1 (January 2001): 42–53. https://doi.org/10.1002/1096-8644(200101)114:1<42::aid-ajpa1004>3.0.co;2-b .
76. Bigham, Abigail W., Xianyun Mao, Rui Mei, Tom Brutsaert, Megan J. Wilson, Colleen Glyde Julian, Esteban J. Parra, Joshua M. Akey, Lorna G. Moore, and Mark D. Shriver. "Identifying Positive Selection Candidate Loci for High-Altitude Adaptation in Andean Populations." *Human Genomics* 4, no. 2 (December 2009). https://doi.org/10.1186/1479-7364-4-2-79.

77. Simonson TS, Yang Y, Huff CD, Yun H, Qin G, Witherspoon DJ, Bai Z, Lorenzo, FR, Xing J, Jorde LB, Prchal, JT, Ge R“Genetic Evidence for High-Altitude Adaptation in Tibet.” *Science* 329, no. 5987 (July 2, 2010): 72–75. https://doi.org/10.1126/science.1189406.
78. Beall, Cynthia M., Gianpiero L. Cavalleri, Libin Deng, Robert C. Elston, Yang Gao, Jo Knight, Chaohua Li, et al. “Natural Selection on EPAS1 (HIF2α) Associated with Low Hemoglobin Concentration in Tibetan Highlanders.” *Proceedings of the National Academy of Sciences* 107, no. 25 (June 7, 2010): 11459–64. https://doi.org/10.1073/pnas.1002443107.
79. Schlebusch, Carina M., Lucie M. Gattepaille, Karin Engström, Marie Vahter, Mattias Jakobsson, and Karin Broberg. “Human Adaptation to Arsenic-Rich Environments.” *Molecular Biology and Evolution* 32, no. 6 (March 3, 2015): 1544–55. https://doi.org/10.1093/molbev/msv046.
80. Mychaleckyj, Josyf C., Alexandre Havt, Uma Nayak, Relana Pinkerton, Emily Farber, Patrick Concannon, Aldo A. Lima, and Richard L. Guerrant. “Genome-Wide Analysis in Brazilians Reveals Highly Differentiated Native American Genome Regions.” *Molecular Biology and Evolution*, January 18, 2017, msw249. https://doi.org/10.1093/molbev/msw249.
81. Nielsen, Rasmus, Joanna L. Mountain, John P. Huelsenbeck, and Montgomery Slatkin. “MAXIMUM-LIKELIHOOD ESTIMATION OF POPULATION DIVERGENCE TIMES AND POPULATION PHYLOGENY IN MODELS WITHOUT MUTATION.” *Evolution* 52, no. 3 (June 1998): 669–77. https://doi.org/10.1111/j.1558-5646.1998.tb03692.x.
82. Excoffier, Laurent, Nina Marchi, David Alexander Marques, Remi Matthey-Doret, Alexandre Gouy, and Vitor C Sousa. “Fastsimcoal2: Demographic Inference under Complex Evolutionary Scenarios.” *Bioinformatics* 37, no. 24 (June 23, 2021): 4882–85. https://doi.org/10.1093/bioinformatics/btab468.
83. Browning, Sharon R., and Brian L. Browning. “Haplotype Phasing: Existing Methods and New Developments.” *Nature Reviews Genetics* 12, no. 10 (September 16, 2011): 703–14. https://doi.org/10.1038/nrg3054.
84. Gautier, Mathieu, and Renaud Vitalis. “Rehh: An R Package to Detect Footprints of Selection in Genome-Wide SNP Data from Haplotype Structure.”

Bioinformatics 28, no. 8 (March 7, 2012): 1176–77. https://doi.org/10.1093/bioinformatics/bts115.

85. Szpiech, Z. A., and R. D. Hernandez. “Selscan: An Efficient Multithreaded Program to Perform EHH-Based Scans for Positive Selection.” *Molecular Biology and Evolution* 31, no. 10 (July 10, 2014): 2824–27. https://doi.org/10.1093/molbev/msu211.
86. Maclean, Colin A., Neil P. Chue Hong, and James G.D. Prendergast. “Hapbin: An Efficient Program for Performing Haplotype-Based Scans for Positive Selection in Large Genomic Datasets: Fig. 1.” *Molecular Biology and Evolution* 32, no. 11 (August 6, 2015): 3027–29. https://doi.org/10.1093/molbev/msv172.
87. Ma, Y., Ding, X., Qanbari, S., Weigend, S., Zhang, Q., & Simianer, H.“Properties of Different Selection Signature Statistics and a New Strategy for Combining Them.” *Heredity* 115, no. 5 (May 20, 2015): 426–36. https://doi.org/10.1038/hdy.2015.42.
88. Loh, Po-Ru, Petr Danecek, Pier Francesco Palamara, Christian Fuchsberger, Yakir A Reshef, Hilary K Finucane, Sebastian Schoenherr, et al. “Reference-Based Phasing Using the Haplotype Reference Consortium Panel.” Cold Spring Harbor Laboratory, May 10, 2016. http://dx.doi.org/10.1101/052308.
89. Brown, T. A. “Mapping Genomes.” In *Genomes 4*, 55–86. Other titles: Genomes | Genomes four Description: 4th. | New York, NY : Garland Science, [2017] | Preceded by: Garland Science, 2018. http://dx.doi.org/10.1201/9781315226828-3.
90. Danecek, Petr, James K Bonfield, Jennifer Liddle, John Marshall, Valeriu Ohan, Martin O Pollard, Andrew Whitwham, et al. “Twelve Years of SAMtools and BCFtools.” *GigaScience* 10, no. 2 (January 29, 2021). https://doi.org/10.1093/gigascience/giab008.
91. Poot, Martin. “Connecting the CNTNAP2 Networks with Neurodevelopmental Disorders.” *Molecular Syndromology* 6, no. 1 (February 3, 2015): 7–22. https://doi.org/10.1159/000371594.

92. NCBI. “COL26A1 Collagen Type XXVI Alpha 1 Chain [Homo Sapiens (Human)] - Gene.” Accessed April 27, 2022. https://www.ncbi.nlm.nih.gov/gene/136227.
93. Rees, Jasmin S., Sergi Castellano, and Aida M. Andrés. “The Genomics of Human Local Adaptation.” *Trends in Genetics* 36, no. 6 (June 2020): 415–28. https://doi.org/10.1016/j.tig.2020.03.006.
94. Sugden, Lauren Alpert, Elizabeth G. Atkinson, Annie P. Fischer, Stephen Rong, Brenna M. Henn, and Sohini Ramachandran. “Localization of Adaptive Variants in Human Genomes Using Averaged One-Dependence Estimation.” *Nature Communications* 9, no. 1 (February 19, 2018). https://doi.org/10.1038/s41467-018-03100-7.
95. Patel, C. A., & Ghiselli, G. (2005). Hinderin, a five-domains protein including coiled-coil motifs that binds to SMC3. BMC cell biology, 6(1), 1-10.
96. Thorne, Leanne B., Aaron L. Grant, Gary D. Paterno, and Laura L. Gillespie. “Cloning and Characterization of the Mouse Ortholog Ofmi-Er1.” *DNA Sequence* 16, no. 3 (June 2005): 237–40. https://doi.org/10.1080/10425170500069783.
97. Szpiech, Zachary A., Taylor E. Novak, Nick P. Bailey, and Laurie S. Stevison. “Application of a Novel Haplotype-based Scan for Local Adaptation to Study High-altitude Adaptation in Rhesus Macaques.” *Evolution Letters* 5, no. 4 (May 22, 2021): 408–21. https://doi.org/10.1002/evl3.232.
98. NCBI. “Progressive External Ophthalmoplegia with Mitochondrial DNA Deletions, Autosomal Recessive 1 (Concept Id: C4225153) - MedGen.” Accessed April 27, 2022. https://www.ncbi.nlm.nih.gov/medgen/897191.
99. “GeneCards: Integrating Information about Genes, Proteins and Diseases.” *Trends in Genetics* 13, no. 4 (April 1997): 163. https://doi.org/10.1016/s0168-9525(97)01103-7.
100. Wu, Tianzhi, Erqiang Hu, Shuangbin Xu, Meijun Chen, Pingfan Guo, Zehan Dai, Tingze Feng, et al. “ClusterProfiler 4.0: A Universal Enrichment Tool for Interpreting Omics Data.” *The Innovation* 2, no. 3 (August 2021): 100141. https://doi.org/10.1016/j.xinn.2021.100141.
101. McLaren, Paul J, Ali Gawanbacht, Nitisha Pyndiah, Christian Krapp, Dominik Hotter, Silvia F Kluge, Nicola Götz, et al. “Identification of Potential

HIV Restriction Factors by Combining Evolutionary Genomic Signatures with Functional Analyses." *Retrovirology* 12, no. 1 (May 16, 2015). https://doi.org/10.1186/s12977-015-0165-5.

102. NCBI. "CXCL9 C-X-C Motif Chemokine Ligand 9 [Homo Sapiens (Human)] - Gene," n.d. https://www.ncbi.nlm.nih.gov/gene/4283.
103. Ashburner, Michael, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, et al. "Gene Ontology: Tool for the Unification of Biology." *Nature Genetics* 25, no. 1 (May 2000): 25–29. https://doi.org/10.1038/75556.
104. Williamson, Scott, Melissa J Hubisz, Andrew G. Clark, Bret A Payseur, Carlos D Bustamante, and Rasmus Nielsen. "Localizing Recent Adaptive Evolution in the Human Genome." *PLoS Genetics* preprint, no. 2007 (2005): e90. https://doi.org/10.1371/journal.pgen.0030090.eor.
105. Crespi, Bernard, Kyle Summers, and Steve Dorus. "Adaptive Evolution of Genes Underlying Schizophrenia." *Proceedings of the Royal Society B: Biological Sciences* 274, no. 1627 (September 4, 2007): 2801–10. https://doi.org/10.1098/rspb.2007.0876.
106. Mendoza-Revilla, Javier, Juan Camilo Chacón-Duque, Macarena Fuentes-Guajardo, Louise Ormond, Ke Wang, Malena Hurtado, Valeria Villegas, et al. "Disentangling Signatures of Selection before and after European Colonization in Latin Americans." Cold Spring Harbor Laboratory, November 19, 2021. http://dx.doi.org/10.1101/2021.11.15.467418.
107. NHGRI. "Genetic Drift." Genome.gov. Accessed April 27, 2022. https://www.genome.gov/genetics-glossary/Genetic-Drift.
108. Pineda-Tenor, Daniel, Juan Berenguer, María A. Jiménez-Sousa, María Guzmán-Fulgencio, Teresa Aldámiz-Echevarria, Ana Carrero, Mónica García-Álvarez, et al. "CXCL9, CXCL10 and CXCL11 Polymorphisms Are Associated with Sustained Virologic Response in HIV/HCV-Coinfected Patients." *Journal of Clinical Virology* 61, no. 3 (November 2014): 423–29. https://doi.org/10.1016/j.jcv.2014.08.020.
109. Laurent, G. J., & Shapiro, S. D. (Eds.). (2006). Encyclopedia of respiratory medicine (Vol. 3). Academic Press Elsevier.

110. Novembre, John, Alison P Galvani, and Montgomery Slatkin. “The Geographic Spread of the CCR5 Δ32 HIV-Resistance Allele.” PLoS Biology 3, no. 11 (October 18, 2005): e339. https://doi.org/10.1371/journal.pbio.0030339.

12. SOPORTE GRÁFICO

SEÑALES DE SELECCIÓN: TIPOS DE SELECCIÓN

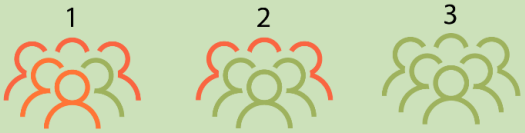
"La selección natural se basa en la observación de que aquellos rasgos que favorecen la supervivencia y reproducción se vuelven más frecuentes a lo largo del tiempo"

En 1858, los científicos Charles Darwin y Alfred Wallace sentaron las bases de la evolución al describir por primera vez el principio de la selección natural

Años después, el redescubrimiento de las Leyes de Mendel permitió definir a la selección como un cambio en las frecuencias alélicas de una población como consecuencia de su ambiente.

Tipos de selección

Selección positiva



Conservación y aumento de frecuencia de variantes ventajosas

Mutación ventajosa

Selección negativa

Pérdida selectiva de variantes desfavorables.

Mutación desfavorable

Selección neutral

Favorece la conservación de dos o más variantes en una población