UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE CIENCIAS

EVALUACIÓN DE MÉTODOS DE FASEO EN DATOS GENÓMICOS DE ADN ANTIGUO

T E S I S

QUE PARA OBTENER EL TÍTULO DE:

LICENCIADO EN CIENCIAS DE LA COMPUTACIÓN

P R E S E N T A :

JAZEPS MEDINA TRETMANIS

DIRECTORA DE TESIS:

DR. MARÍA DEL CARMEN ÁVILA ARCOS

CIUDAD DE MÉXICO, 2021

Índice

Agradecimientos

Estoy profundamente agradecido con mis asesoras, las doctoras María Ávila Arcos y Emilia Huerta Sánchez. Su paciencia mientras aprendía de cero tanto conocimiento de biología que ha sido necesario para este trabajo fue increíble. Gracias a ustedes he podido conocer a tanta gente y lugares nuevos, nunca hubiera imaginado que esta tesis pudiera resultar en tantas oportunidades.

De igual manera, agradezco a todos los integrantes de los laboratorios Ávila-Arcos y Huerta-Sánchez por su ayuda constante siempre que fue posible. Muchísimas gracias al doctor Diego Ortega del Vecchyo, con quien siempre pudimos consultar sobre cualquier etapa del proyecto.

Este trabajo también recibió apoyo de Luis Aguilar, Alejandro de León, Carlos S. Flores, y Jair García del Laboratorio Nacional de Visualización Científica Avanzada (LAVIS). Se le agradece de sobremanera a todo el personal de los laboratorios LIIGH y LAVIS, por su ayuda técnica, ayuda académica, y hospitalidad.

Este trabajo fue sustentado por el programa PAPIIT de la DGAPA UNAM con claves IA206817 e IA201219, y por la subvención CN 17-12 del programa UC MEXUS-Conacyt.

1. Resumen

En este trabajo, se explora la relación entre la calidad de datos de secuenciación a partir de ADN antiguo, y los resultados que se pueden obtener con diferentes estrategias de estimación de haplotipos (faseo) sobre estos datos. Por calidad nos referimos a la cantidad de ADN recuperable de una muestra, la contaminación del ADN antiguo con ADN moderno, y el daño por deaminación que ocurre en ADN antiguo. Los datos recuperables a partir de ADN antiguo son susceptibles a estos cambios en la calidad, principalmente por el efecto del paso del tiempo y las condiciones de preservación del ADN.

El uso de métodos analíticos de inferencia demográfica basados en haplotipos puede ofrecer ventajas sobre métodos basados en frecuencias alélicas, sin embargo, la relación entre el daño y baja calidad de datos genómicos antiguos (paleogenómicos) y los haplotipos inferidos a partir de ellos no ha sido explorada a profundidad.

En este trabajo se simularon datos paleogenómicos con distintos parámetros de daño, calidad, e historias demográficas representativas de estudios recientes que trabajan sobre

ADN antiguo, para cuantificar la relación entre estas variables y la confiabilidad de los métodos existentes de faseo. Esta relación fue explorada en contexto de dos diferentes métodos de faseo, el faseo con panel de referencia, y el faseo poblacional.

La combinación de nuevas herramientas de extracción, secuenciación y análisis de ADN antiguo, junto al continuo descubrimiento de material biológico antiguo, pueden aclarar cada vez más preguntas sobre las historias demográficas y evolutivas de las poblaciones humanas actuales, lo cual se puede ver considerablemente beneficiado mediante el uso de métodos analíticos basados en haplotipos. En consecuencia, es sumamente importante desarrollar un marco teórico que describa la confiabilidad de las inferencias de haplotipos a partir de datos genómicos antiguos, sobre todo si consideramos los avances del campo de la paleogenómica y la gran cantidad de datos de este tipo que en estos momentos se generan por varios grupos de investigación.

Si bien este trabajo no contempla todas las posibles combinaciones de estrategias de faseo y calidad de datos paleogenómicos, los resultados permiten apreciar la relación general entre la calidad de los datos simulados y la exactitud y precisión del faseo sobre estos, lo cual es suficiente para orientar su aplicación en futuros estudios con datos reales.

2. Introducción

2.1. ADN antiguo, limitantes y aplicaciones

El campo emergente de la paleogenómica se enfoca en el procesamiento y análisis del ADN antiguo (ADNa) a escala genómica.

La investigación en el campo de la paleogenómica debe considerar varios factores que dificultan la recuperación y análisis del ADNa, el cual es sujeto a una inevitable degradación a través del tiempo a comparación de muestras de ADN moderno obtenidas en condiciones controladas. Una de las principales formas en las que el ADNa es dañado por el paso del tiempo es la deaminación *post-mortem* de las bases nitrogenadas que lo conforman. Específicamente la deaminación ocurre en citosinas (C), dando lugar a uracilos (U), los cuales son leídos por las tecnologías de secuenciación como timinas (T). Esto ocurre principalmente en los extremos terminales de los fragmentos de ADN, los cuales tienden a estar en cadena sencilla[1]. Estas tasas de deaminación han sido medidas de manera experimental[2]. Otro aspecto del daño al ADNa es la alta fragmentación de

las secuencias que pueden ser recuperadas, lo cual puede dificultar encontrar de manera definitiva la ubicación de estos fragmentos cortos dentro de otras secuencias de ADN[3]. Otros ejemplos de complicaciones al trabajar con ADNa son la contaminación por ADN ambiental de microorganismos que colonizan las muestras antiguas o de los individuos modernos que han manipulado la muestra sin los cuidados necesarios, y la poca cantidad de material genético endógeno (muchas veces menor al 1 %) que es recuperable de estas muestras[4]. Las tecnologías de secuenciación de siguiente generación (NGS, por sus siglas en inglés) han permitido aumentar la cantidad de ADN recuperable de muestras antiguas[5]. Asimismo, algunos estudios han determinado cuáles son los huesos que mejor preservan el ADN[6]. Gracias a estos avances, la paleogenómica ha aportado al desarrollo de la biología evolutiva[7][8], al conocimiento sobre la historia demográfica de poblaciones humanas y de otras especies[9][10], y a la investigación de patógenos[11], entre otras contribuciones.

Dentro de la genética de poblaciones, el ADNa ha permitido la reconstrucción de los cambios poblacionales de diferentes especies a diferentes escalas de tiempo[9]. Ejemplos específicos incluyen: inferencias sobre el origen y expansión de los primeros humanos[12], evidencia de mezcla entre humanos antiguos y homínidos arcaicos[13], efectos de la conquista sobre las poblaciones americanas actuales[10], y mecanismos por los cuales las islas polinesias fueron originalmente pobladas[14], por mencionar algunos.

Un ejemplo de interés es el caso del gen *MCM6*, en el cual pueden existir variantes genéticas que permiten la digestión de la lactosa en humanos adultos. El estudio de varios genomas antiguos europeos, permitió revelar que esta adaptación existía desde hace $8,000$ años, pero que era todavía poco común hasta hace $3,000$ años[9]. Esto nos dice que la adaptación no fue común hasta mucho tiempo después de que se comenzara la práctica del consumo de lácteos, lo cual lleva a preguntarse cómo se manejaban los efectos secundarios de la intolerancia a la lactosa en los individuos que carecían de esa variante.

En el campo de la medicina, el análisis de ADNa ha contribuido dando un contexto evolutivo a ciertas adaptaciones fisiológicas de interés clínico y médico[15]. Por ejemplo, estudios recientes han revelado que la mezcla genética entre *Homo Sapiens* y los homínidos arcaicos denisovanos (hace aproximadamente $30,000$ años[16]), favoreció la adaptación de poblaciones Tibetanas a ambientes hipóxicos en las grandes elevaciones del Tibet [17]. Por otro lado, la recuperación y análisis de patógenos recuperados de individuos antiguos

nos ha permitido trazar el origen, evolución, y expansión de patógenos modernos[18]. Por ejemplo, se han utilizado genomas antiguos de la bacteria *Y. pestis*, agente patógeno responsable de la peste negra, para reconstruir las diferentes adaptaciones de esta bacteria que incrementaron su virulencia, junto con los movimientos demográficos que la esparcieron a lo largo de Europa y el resto del mundo[11].

La reconstrucción de varios eventos demográficos del pasado ha sido posible gracias a la disponibilidad de muestras antiguas que corresponden al tiempo de estos eventos; a su vez, estas reconstrucciones nos permiten contextualizar la estructura genética de las poblaciones modernas. Algunos ejemplos son la caracterización de eventos migratorios en Gran Bretaña antes de las migraciones anglosajonas[19], el efecto de migraciones zoroastras en poblaciones actuales de Irán e India[20], cambios genómicos en poblaciones europeas siguiendo las transiciones entre edades de piedra, bronce, y hierro[21], y evidencia de migraciones barbáricas hacia el norte de Italia entre los siglos IV y VI A.D. [22]. Estos estudios son de especial interés por las metodologías empleadas. En todos ellos, se hizo uso de herramientas de faseo estadístico (el cual se definirá más adelante), sin ahondar en los posibles efectos de la calidad de los datos paleogenómicos sobre la precisión de los haplotipos obtenidos.

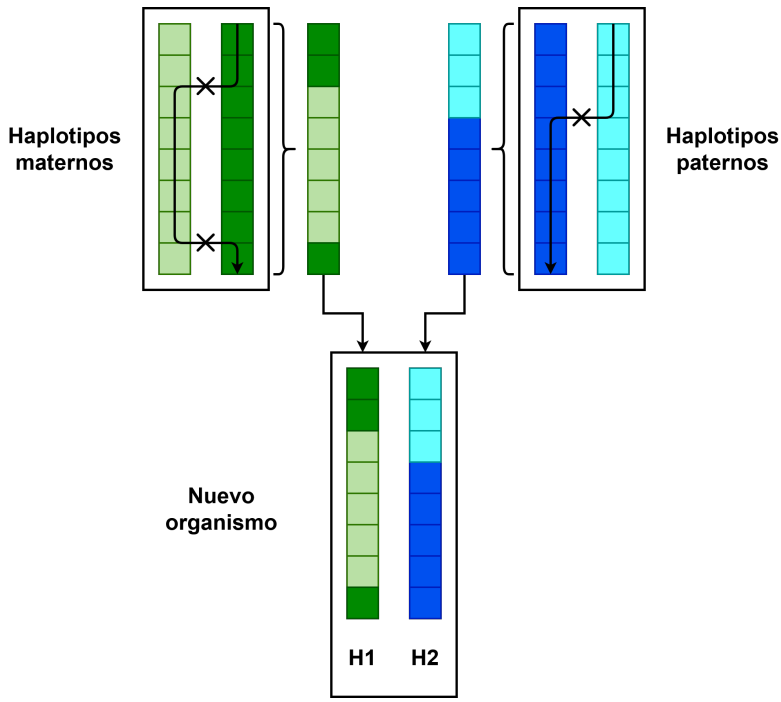
Algunos de los estudios paleogenómicos que hacen inferencias sobre demografía humana del pasado, incluyendo los mencionados arriba[19][20][21][22], hacen uso de la estimación de haplotipos para la implementación de aproximaciones analíticas basadas en este tipo de información. Sin embargo, la estimación de haplotipos es un procedimiento que no se ha evaluado sistemáticamente en ADNa, lo cual es una de las principales motivaciones de la presente tesis.

2.2. Inferencia de haplotipos: estrategias y métodos de faseo

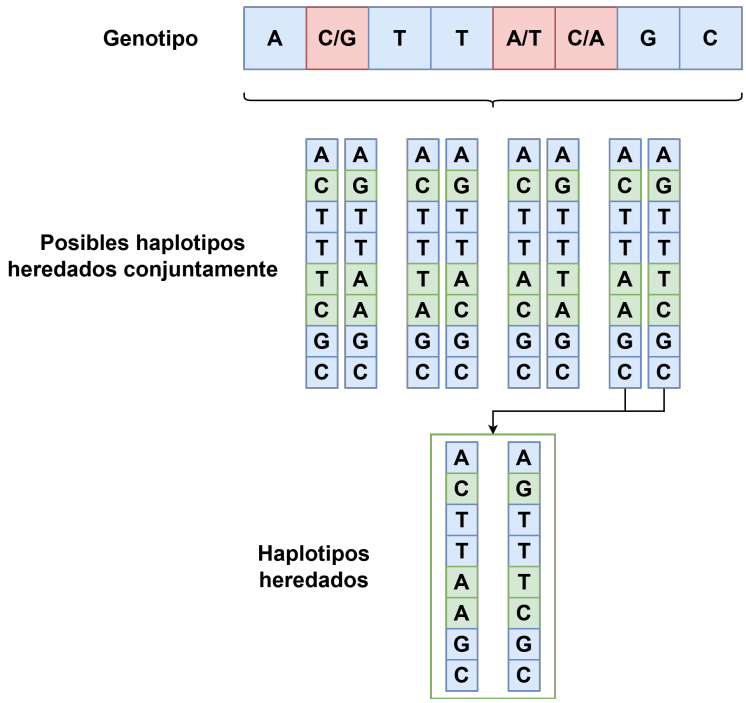
Por haplotipos entendemos el grupo de alelos que son heredados conjuntamente en un mismo cromosoma de su ancestro inmediato anterior. Hablando específicamente de humanos, los dos haplotipos de un organismo corresponden al material genético heredado de la madre y del padre biológicos (Figura 1a). En otras palabras, el genotipo de un individuo implica varias posibles «fases» o haplotipos. Esto es porque, aunque la información genotípica nos permite identificar las posiciones en las que los haplotipos difieren, es ambiguo respecto a las relaciones entre estas posiciones (Figura 1b). El «faseo» («phasing»

en inglés) del genotipo de un individuo nos permite reconstruir e identificar sus haplotipos, es decir, distinguir los alelos del genotipo que fueron heredados conjuntamente, lo cual aporta un nivel adicional de información que se pierde al estudiar los genotipos de forma independiente. En consecuencia, conocer los haplotipos de los individuos de una población permite utilizar métodos análiticos que incorporan esta información de «fase» para la inferencia demográfica y detección de selección positiva[23][24].

Por otro lado, los métodos de secuenciación comúnmente empleados en años recientes (*short read sequencing*) complican la reconstrucción precisa de haplotipos dado que las secuencias recuperadas son cortas (en el mejor de los casos 300 bp[25]). Por ello, es necesaria la implementación de métodos estadísticos y computacionales para la estimación de haplotipos; a estos métodos se les conoce como métodos de «faseo» (*phasing* en inglés). Si bien la implementación de tecnologías de secuenciación que producen lecturas mucho más largas (por ejemplo *PacBio* u *Oxford Nanopore*) ha permitido perfeccionar la estimación de haplotipos en muestras de ADN contemporáneo[26], este tipo de tecnologías no es aplicable para el ADNa, dado que éste sufre de una alta fragmentación (con un rango promedio de 60 bp a 150 bp)[3]. En consecuencia, el faseo de ADNa tiene complicaciones adicionales a las ya existentes para datos modernos, por lo que es indispensable evaluar el desempeño de los métodos de faseo en el contexto del ADNa. Esto no se ha realizado de manera sistemática a pesar de que varios estudios[19][20][21][22] han empleado haplotipos faseados de ADNa para realizar inferencias sobre la historia demográfica de poblaciones humanas.



(a)



(b)

Figura 1: Estimación de haplotipos. El panel **(a)** muestra cómo la recombinación de los haplotipos de los padres biológicos (tonos de verde para los haplotipos de la madre, y azul para el padre) de un individuo crean haplotipos nuevos, que pueden ser representados como un mosaico de los haplotipos originales. Los eventos de recombinación (indicados con cruces) muestran cómo un individuo diploide puede heredar material genético de ambos conjuntos de cromosomas. El panel **(b)** muestra cómo la estimación de haplotipos puede eliminar la ambigüedad generada al considerar sólo datos genotípicos. Los sitios rojos en el genotipo indican sitios heterocigotos para los que la fase es ambigua, mientras que los sitios en verde indican posibles fases.

Para el análisis de datos genéticos, existen tres estrategias de faseo: faseo por pedigree, faseo estadístico, y faseo basado en lecturas (también conocido como ensamblado de haplotipos o *haplotype assembly*). Para cada una de estas estrategias (que se describirán más a detalle en la sección 3.1.1), existen diferentes métodos que varían respecto a los algoritmos con los que son implementados[27] (Figura 2). Considerando el alto nivel de fragmentación del ADNa[3] y que en pocas ocasiones podemos identificar con precisión las relaciones de parentesco de los individuos bajo estudio, únicamente resulta viable aplicar métodos de faseo estadístico en datos paleogenómicos. En cuanto al faseo estadístico, podemos hablar de dos diferentes métodos: faseo con panel de referencia, y faseo poblacional (del inglés *population phasing*)[28].

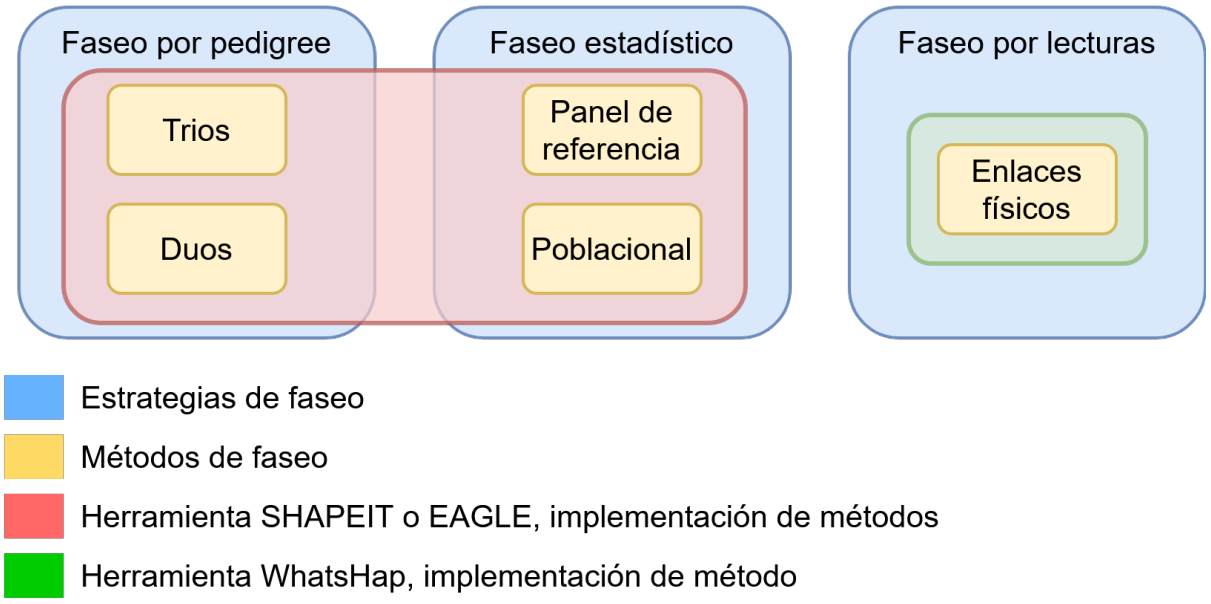


Figura 2: Esquema de las relaciones entre estrategias, métodos, y herramientas.

El faseo con panel de referencia hace uso de un panel de haplotipos conocidos, es decir, un conjunto confiable de haplotipos que describe a una o más poblaciones, a partir del cual se pueden estimar los haplotipos de un nuevo individuo calculando la probabilidad de que ciertos alelos sean heredados en conjunto. Tres de los paneles de referencia de haplotipos más utilizados son los generados por el proyecto de los *1,000 Genomas*[29], el proyecto del *Consorcio para Referencia de Haplotipos*[30], y el proyecto *TOPMed*[31]. Por otro lado, el faseo poblacional utiliza solamente los datos genotípicos de un grupo de individuos para hacer esta inferencia, sin tener conocimiento sobre la estructura haplotípica subyacente. Ambos métodos generalmente son implementados con el uso de modelos

ocultos de Markov[28][32][33] y serán descritos más a profundidad en la siguiente sección.

El proyecto de los *1,000 Genomas* estudió un total de 2,504 individuos modernos pertenecientes a 26 poblaciones humanas, esto incluye la recopilación de toda la información haplotípica para cada individuo. Las poblaciones estudiadas pueden ser divididas en cinco súper categorías: africanas, asiáticas del este, europeas, sudasiáticas, y poblaciones mezcladas de América[29]. Este panel de referencia ha sido utilizado en numerosos estudios[34]. Como se verá más adelante, algunas de estas publicaciones[19][20][21][22] utilizan este panel de referencia para la inferencia de haplotipos de muestras antiguas.

2.3. Efectos de cobertura y profundidad genómica en calidad de faseo

Dado que estos métodos de faseo dependen de la cantidad de información genética disponible para los individuos que se pretende analizar, es importante definir los términos *cobertura* y *profundidad*. La información genética sobre una secuencia de ADN se obtiene a través de lecturas. Una lectura es una secuencia de pares de bases que corresponden a alguna parte de un fragmento de ADN. Los métodos de secuenciación comúnmente empleados generan en el orden de millones a miles de millones de lecturas por muestra[35]. Estas lecturas son «mapeadas» contra un genoma de referencia para reconstruir la variación genética del individuo o individuos bajo estudio. De estos «mapeos» se desprenden los valores de cobertura y profundidad, los cuales son indicativos de la calidad del ADN de la muestra y de los datos. La cobertura nos indica la fracción del genoma (o secuencia) de referencia para la que existe al menos una lectura que la cubre, por lo general la cobertura se expresa en porcentajes. Por su parte, la profundidad en un sitio específico de la secuencia de referencia nos dice cuántas lecturas tienen información de ese sitio (Figura 3). Asimismo es posible calcular la profundidad promedio de un genoma de referencia, lo cual nos indica el promedio de veces que se observa cada sitio de la referencia, lo que es equivalente a decir cuántas lecturas están cubriendo cada sitio en promedio[36].

Es importante tener varias lecturas cubriendo cada sitio; los genomas diploides pueden estar formados por dos alelos diferentes en un mismo sitio, y las máquinas de secuenciación pueden cometer errores. Tener múltiples lecturas que observan la misma base incrementa la confianza en la secuencia obtenida.

La confiabilidad de diferentes implementaciones de métodos para la estimación de

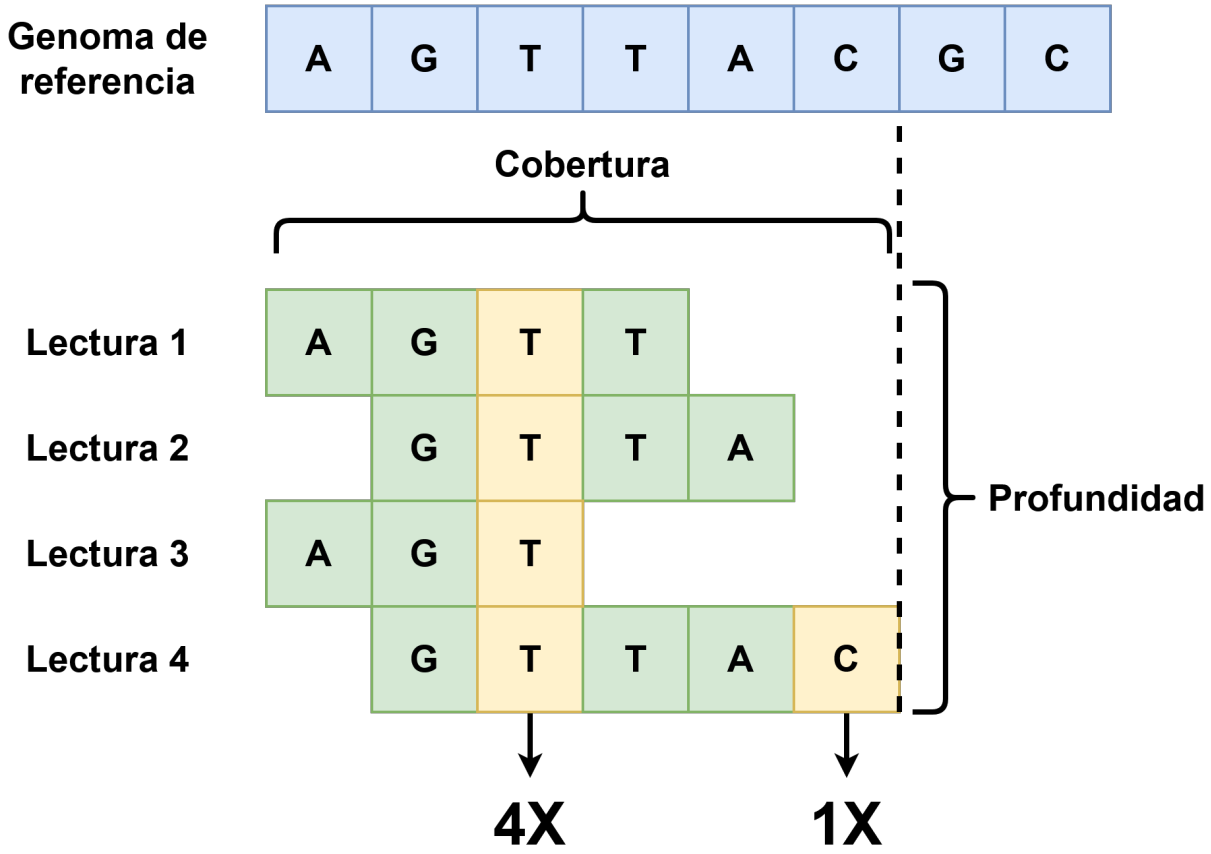


Figura 3: Cobertura y profundidad. La cobertura del genoma de referencia es del 75 %, ya que las últimas dos bases de la secuencia no fueron cubiertas por ninguna lectura. Las profundidades de la tercera y sexta base de la secuencia son de 4× y 1× respectivamente. La profundidad promedio del genoma de referencia es $\frac{2+4+4+3+3+2+1}{8} = 2{,}4\times$. Una mayor profundidad de cobertura nos da más certeza sobre el genotipo recuperado, ya que el genotipo de un individuo diploide está definido por dos cromosomas no necesariamente idénticos (haplotipos).

haplotipos se ha reportado para datos genéticos de poblaciones actuales[32][33][37]. En estos estudios se ha evaluado la precisión de diferentes métodos de faseo sobre datos simulados de secuenciación a diferentes profundidades, y han dejado claro que a menor profundidad ($< 5\times$) hay una importante disminución en la calidad del faseo. Hablando en general respecto a la estrategia de faseo estadístico, se ha mostrado que una mayor cantidad de individuos al utilizar el método de faseo poblacional resulta en una mayor precisión[33]. Este resultado se obtuvo como parte parte de la publicación de la herramienta *BEAGLE*, a partir de observar que el faseo en grupos de 500 individuos diploides mostró resultados notablemente mejores que en grupos de 50 individuos. Incrementar el tamaño del grupo considerado a $2,500$ resultó en una calidad aún mayor, aunque esta mejoría no fue tan marcada como la anterior[33]. En cuanto al faseo basado en lecturas,

si se cuenta con lecturas alineadas con una longitud promedio de $\sim 8,000$ bp (utilizando *long-read sequencing*) y una profundidad de al menos $5\times$, se obtienen mejores resultados que al utilizar datos de una menor profundidad[37].

El desempeño de estos programas no ha sido propiamente evaluado para datos generados a partir de ADNa. Consideremos también el hecho de que los haplotipos de una población cambian a través del tiempo debido a la recombinación [24][38], y a la potencial falta de haplotipos ancestrales en el panel de referencia de los *1,000 Genomas*[29]. Esta combinación hace difícil evaluar la confiabilidad de las estimaciones de haplotipos a partir de ADNa, y por consecuencia, la confiabilidad de las inferencias sobre eventos demográficos y estructura poblacional.

Este trabajo propone que a través de la simulación de datos genómicos antiguos se puede establecer una relación fidedigna entre los niveles de daño, contaminación, e historia demográfica de los individuos y la confiabilidad de los métodos existentes de faseo. Esta relación fue explorada tanto en el contexto del faseo con panel de referencia, como el faseo poblacional. Con este fin, se simularon haplotipos modernos y antiguos, a partir de los cuales se generaron secuencias de ADNa con diferentes combinaciones de daño, contaminación moderna, profundidad de cobertura, y estructuras demográficas similares a los de las muestras de ADNa en los estudios mencionados. Posteriormente, se emplearon los métodos de faseo estándar (poblacional y referencia) para poder evaluar la precisión y exactitud de las estimaciones de haplotipos comparando los haplotipos originales simulados contra los inferidos por el programa.

Se espera que estos resultados puedan ser un punto de referencia para cualquier estudio futuro que planee utilizar métodos de faseo estadístico sobre datos paleogenómicos. Dada la importancia de los resultados que se pueden obtener a partir de estos datos, y el aumento contínuo de la disponibilidad de datos de ADNa, es vital contar con bases firmes para sustentar cualquier análisis realizado. Los resultados presentados en este trabajo se limitan solamente a la implementación de la herramienta *SHAPEIT*[32] para dos métodos de faseo, y es importante recordar que se basan en datos simulados que no contemplan todas las posibles combinaciones de parámetros observables en datos paleogenómicos. Aún así, se considera que los resultados son ilustrativos de la relación general entre diversas características de datos poblacionales paleogenómicos y la exactitud y precisión del faseo sobre ellas.

3. Antecedentes

3.1. Faseo

3.1.1. Estrategias de faseo

En general, podemos hablar de tres diferentes estrategias de faseo computacional: *faseo por pedigree*, *faseo estadístico*, y *faseo basado en lecturas* (Figuras 2); las cuales se describen a continuación.

Faseo por pedigree: dentro de esta estrategia, los métodos se basan en relaciones de madres, padre, e hijos conocidas entre los individuos analizados. Conociendo los genotipos de los linajes paterno y materno, es posible inferir de cuál linaje provienen los alelos en ciertos sitios heterocigotos.

Por ejemplo, supongamos que tenemos un individuo con genotipo G/C en el sitio x, si sabemos que los linajes materno y paterno en el mismo sitio son G/C y G/G respectivamente, el alelo C necesariamente debe provenir de la madre biológica.

Hablando de ADNa, este método ha sido usado cuando se cuenta con documentos históricos o estructuras sociales reflejadas en los entierros de donde provienen las muestras[39]. Sin embargo, no es aplicable en casos donde no se cuenta con individuos directamente emparentados, lo cual es el caso en la mayoría de los estudios de ADN antiguo.

faseo basado en lecturas: La estrategia de faseo basado en lecturas, consiste en identificar lecturas de ADN que contienen más de un sitio heterocigoto para reconstruir los haplotipos de un individuo[28]. Si bien el tamaño de las lecturas es por lo general corto (en el mejor de los casos 300 bp[25]), y es poco frecuente encontrar más de un sitio heterocigoto en lecturas de este tamaño, las tecnologías de secuenciación más utilizas (e.g. *Illumina*) permiten realizar secuenciación de lecturas «pareadas»[40] (*paired-end sequencing*) en las que se secuencían los extremos de fragmentos largos. Esto permite identificar la fase de dos sitios heterocigotos en los extremos de una una lectura pareada. Se ha mostrado que la precisión de los haplotipos reconstruidos con esta estrategia aumenta al utilizar insertos largos[41]. Estos insertos largos tienen longitudes en el rango de $2,000$ bp a $10,000$ bp[42].

Reconstruir la fase de un individuo con esta estrategia basada en lecturas es más preciso entre más largas sean las lecturas[37]. Dado que una de las características del ADNa es la longitud corta de los fragmentos [6],el uso de estos métodos no es recomendable para

este tipo de datos.

Faseo estadístico: la estrategia de faseo estadístico se refiere a la inferencia de haplotipos sobre un conjunto de datos de genotipos, ya sea usando un panel de referencia de haplotipos (faseo con panel de referencia), o bien realizando la inferencia a partir de los genotipos de una población (faseo poblacional). Estos métodos se valen de poder reconocer bloques de ADN compartidos por individuos emparentados aún muy lejanamente. Dado que es un proceso estadístico, su exactitud depende de tener un número suficiente de individuos, ya sea a partir de los cuales se inferirán los haplotipos con el métodos de faseo poblacional, o bien en el panel de referencia de haplotipos en el caso del método de faseo con panel de referencia. A continuación se describen con mayor detalle ambos métodos de faseo.

- *Faseo con panel de referencia*: El faseo con panel de referencia hace uso de un panel de haplotipos pre-computados con la suficiente información como para poder fasear a uno o más individuos que idealmente pertenezcan a la misma población. El tener un panel de referencia pre-computado acelera dramáticamente el proceso de faseo, como se mostrará en secciones posteriores. Como se mencionó al describir la inferencia de haplotipos (Sección 2.2), tres de los paneles de referencia más utilizados son los correspondientes al proyecto de los *1,000 Genomas*[29], el proyecto del *Consorcio para Referencia de Haplotipos*[30], y el proyecto *TOPMed*[31]. Aunque estos paneles comprenden una gran cantidad de individuos de diferentes poblaciones, existe un sesgo importante hacia individuos de poblaciones europeas. Todavía hay muchas poblaciones humanas no representadas (ya sean modernas o antiguas). Si no se cuentan con referencias para la población que está siendo faseada, o si existen errores de secuenciación en los datos del panel de referencia, la eficiencia de estos se vería afectada.

- *Faseo poblacional*: El faseo poblacional se refiere al faseo simultáneo de muchos individuos de los que sólo se tiene información de genotipo. Para reconstruir los haplotipos, se debe hacer una comparación entre el genotipo de cada individuo con el resto de los genotipos en el grupo. De igual manera, para obtener buenos resultados es necesario contar con muchos individuos para obtener resultados confiables[33]. La herramienta *SHAPEIT* recomienda un mínimo de 100 individuos bajo condiciones típicas[32], es decir, datos modernos.

En este trabajo, exploramos en específico los dos métodos de faseo estadístico: faseo con panel de referencia y faseo poblacional, ya que son los que se han empleado en estudios de ADNa. Existen varias herramientas para el faseo estadístico[28]. En este trabajo se utilizó la herramienta *SHAPEIT*[32], ya que es la empleada en los estudios de ADNa mencionados previamente[19][20][21][22]. Otras herramientas como *EAGLE*[43] y *Beagle*[33] también han sido utilizadas sobre datos paleogenómicos, sin embargo se ha demostrado que la precisión y exactitud de los resultados son similares entre estas herramientas y *SHAPEIT*[44].

3.1.2. Exactitud del faseo estadístico

Para hablar de la exactitud de los métodos de faseo, es necesario definir la métrica con la cual se evaluarán. La tasa de error de intercambio, o SWE (del inglés *SWitch Error rate*) indica la cantidad de veces que erróneamente se invierten los haplotipos paterno y materno dentro de los haplotipos reconstruidos. Esta métrica generalmente se indica como un porcentaje, calculado como el número de intercambios erróneos dividido por la cantidad de sitios donde un intercambio era posible, o sea, sitios heterocigotos.

Un bloque de intercambio erróneo ocurre cuando los haplotipos reconstruidos para un individuo entre los sitios x y y corresponden al genotipo entre esos sitios, pero se invierte el orden verdadero de los linajes paterno y materno. Todo bloque de intercambio erróneo implica dos errores de intercambio, ya que debe haber un segundo intercambio para regresar a los linajes correctos previos al primer intercambio (Figura 4).

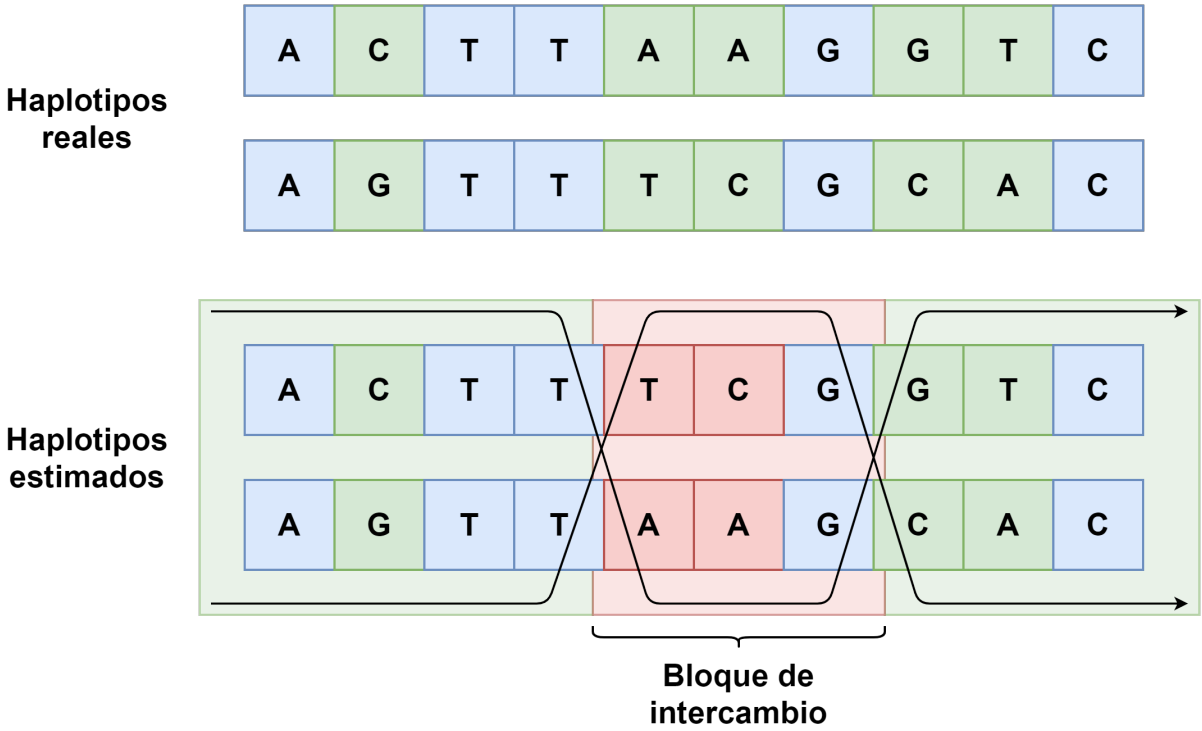


Figura 4: Bloques de intercambio erróneo. Los sitios coloreados con verde corresponden a los sitios heterocigotos, estos son los sitios en los que es posible que ocurra un error de intercambio. El bloque de intercambio erróneo coloreado en rojo muestra cómo los haplotipos estimados pueden tener los alelos correctos, pero en el cromosoma equivocado.

Los autores de las herramientas de faseo estadístico *SHAPEIT*[32] y *BEAGLE*[33] han presentado sus herramientas junto con evaluaciones de la exactitud de sus resultados. A continuación se presenta una comparación de sus resultados, relativos al número de sitios analizados con el que se obtuvieron[32]. Nótese que a menor cantidad de información genética (número de SNPs), mayor tasa de error. La longitud de las regiones consideradas fue la misma en todos los casos.

SNPs considerados	Individuos considerados	SWE SHAPEIT	SWE BEAGLE
$29,227$	740	$\approx 1{,}5\,\%$	$\approx 2\,\%$
$7,985$	$1,229$	$\approx 3\,\%$	$\approx 5\,\%$
$7,871$	925	$\approx 5\,\%$	$\approx 7\,\%$

3.1.3. Implementación del faseo estadístico

El algoritmo de faseo estadístico que se utiliza para este trabajo es el implementado en la herramienta *SHAPEIT*[32].

Este algoritmo es de orden lineal sobre la cantidad de haplotipos a fasear. Específica-

mente, expresando el conjunto de marcadores o variantes considerados en el conjunto de haplotipos como M, y el número de haplotipos a fasear como K, el algoritmo corre en tiempo $O(|M|K)$.

Este algoritmo utiliza un modelo oculto de Markov (HMM por sus siglas en inglés) para estimar el haplotipo desconocido h con base en dos objetos; el genotipo observado G, y el espacio de todos los haplotipos estimados actualmente H.

A grandes rasgos, este algoritmo construye una gráfica de segmentos de haplotipos S en base a G y H, con los segmentos delimitados por los marcadores $m \in M$, de tal manera que todos los haplotipos posibles a partir de G se pueden representar como un camino $x = x_1, x_2, x_3, ..., x_M \mid x_i \in S$ entre los nodos de S.

Esto permite modelar una cadena de Markov cuyos estados y transiciones son representados por los nodos y aristas de S, asumiendo que se cumple la propiedad de Markov para transicionar de un segmento haplotípico a otro, es decir, la probabilidad de transicionar al estado x_{m+1} depende solamente del estado x_m y no de cualquier transición previa. Por lo tanto, tenemos la siguiente ecuación para estimar la probabilidad de un camino x en base al conjunto de haplotipos H[32]:

$$P(x|H) = P(\{x_1, x_2, x_3, ..., x_M\}|H) = P(x_1|H) \prod_{m=2}^{M} P(x_m|x_{m-1}, H)$$

Donde la probabilidad de las transiciones se calcula en base a H al inspeccionar cuáles segmentos haplotípicos son heredados juntos comúnmente. Esto recalca cómo aumentar el tamaño del conjunto H da más certeza a las transiciones calculadas para el HMM, lo cual afecta la confiabilidad de los estimados de $P(x|H)$.

3.1.4. Efectos de la profundidad y contaminación sobre el faseo

Debido a que los algoritmos de faseo estadístico se han desarrollado para ser aplicados en datos de ADN moderno, no se tiene una noción clara sobre los efectos que la baja profundidad y la contaminación pueden tener sobre la calidad de los haplotipos inferidos.

Dado que una baja profundidad en los datos genómicos implica una menor cantidad de sitios (SNPs) a considerar, la exactitud del faseo se vería negativamente afectada si sólo se cuenta con datos a baja profundidad (y por lo tanto baja certidumbre, Sección 2.3) en los individuos a fasear.

Un estudio previo realizado por los desarrolladores de la herramienta *whatshap*[37]

reportó los efectos de diferentes profundidades sobre el SWE. Este estudio estimó que para datos de buena profundidad ($10\times$), el SWE se aproxima al $1\,\%$, mientras que con datos de media o baja profundidad ($< 5\times$), este error se aproxima al $10\,\%$.

Otro aspecto a considerar es la contaminación con ADN de un individuo de la misma especie sobre las muestras de ADNa obtenido a partir de restos arqueológicos. Esta contaminación puede resultar en que se identifique a los haplotipos contaminantes y se asignen como verdaderos. Los efectos conjuntos de la contaminación y los datos de baja profundidad no han sido medidos para herramientas de faseo estadístico.

3.2. Estudios previos con faseo en ADNa

Durante la última década se han publicado algunos artículos que hacen uso de datos paleogenómicos de poblaciones antiguas humanas para reconstruir eventos demográficos a partir de haplotipos. En ellos se han realizado diversas aproximaciones para el faseo de muestras antiguas, sin que reporten la confiabilidad que se tiene sobre este procedimiento. Los métodos de estimación de haplotipos empleados en estos artículos incluyen faseo con panel de referencia y faseo poblacional. En los casos de faseo con panel de referencia[19][21][22], todos los trabajos utilizaron el panel de referencia de haplotipos del proyecto *1,000 Genomas*[29]. A continuación se describen con mayor detalle los estudios a los que hacemos referencia:

En un estudio por Martiniano et al. (2016)[19], utilizaron 9 muestras antiguas, con un rango de antigüedad de entre $1,700$ y $1,000$ años. El genoma de estas muestras se secuenció a una profundidad de aproximadamente $1\times$, y se estimó una contaminación con ADN moderno humano del $2\,\%$. Los genomas fueron faseados con la herramienta *SHAPEIT* utilizando faseo con panel de referencia. Los datos faseados se procesaron con los programas *ChromoPainter*[38] y *fineSTRUCTURE* para identificar estructura poblacional y generar agrupaciones de los individuos antiguos con poblaciones europeas modernas.

En López et al. (2017)[20], se analizaron los genomas antiguos de 8 individuos reportados en publicaciones anteriores junto con datos de genotipos de poblaciones actuales de Irán e India. 7 de los genomas antiguos tenían antigüedades en el rango de $12,000$ a $4,500$ años, y el restante $45,000$ años de antigüedad. El rango de profundidad de cobertura de los genomas antiguos fue de entre $2\times$ y $21\times$. El rango de contaminación en los datos es

del 0,2 % hasta el 1,5 %.

En este estudio, se obtuvieron los haplotipos de los datos antiguos a través de faseo poblacional. Esto se realizó con *SHAPEIT* sobre un conjunto de datos de 2,553 individuos, incluyendo los 8 genomas antiguos, así como datos de genotipado de poblaciones actuales de India, Irán y el resto del mundo. Finalmente, se obtuvieron agrupamientos de los datos modernos y antiguos con *ChromoPainter* y *fineSTRUCTURE*. Estas agrupaciones fueron utilizadas para inferir la época en la cual hubo mezcla entre grupos parsi zoroástricos después de su llegada a India.

En Gamba et al. (2014)[21] se secuenciaron 13 genomas antiguos con edades desde el periodo neolítico hasta la edad de hierro (antigüedades entre 12,000 y 2,500 años). La profundidad de los genomas producidos caen en un rango de entre 0,1× hasta 22×, con estimados de contaminación moderna de $< 0{,}7\,\%$. Estos datos fueron faseados utilizando el panel de referencia de haplotipos del proyecto de los *1,000 Genomas*. Los resultados de este estudio dan evidencia de movimientos poblacionales y mezclas genéticas en el este de Europa concordantes con las transiciones al periodo neolítico, la edad de bronce, y la edad de hierro.

Finalmente, Amorim et al. (2018)[22] utilizaron 63 genomas de cementerios fechados a los siglos VI y VII A.D. (antigüedades entre 1,500 y 1,300 años), con estimados de contaminación de 1 % en promedio y un estimado aislado para una sola muestra de 22 % de contaminación (no se especifica un rango). Para 10 de los individuos se obtuvo una profundidad promedio del genoma de 11,3×, mientras que para las otras 53 el valor fue de aproximadamente 1,5×. El faseo de estos genomas antiguos se realizó usando el panel de referencia de haplotipos del proyecto de los *1,000 Genomas*.

3.3. Simulación de datos genómicos

3.3.1. Teoría de coalescencia

La simulación de secuencias de ADN está basada en la teoría del *n-coalescente*. Un *n-coalescente* es una herramienta que nos permite modelar la ancestría compartida entre n individuos miembros de una población haploide[45][46][47].

Un *n-coalescente* está definido como una cadena de Markov de tiempo continuo, $\{R_t \mid t \geq 0\}$, con las siguientes características:

Se cumple que R_t es una relación de identidad sobre el conjunto $[n] = \{1, 2, 3, ..., n\}$, con $|R_t|$ denotando el número de clases de equivalencia generadas por R_t. También se cumple que R_0 es la relación identidad:

$$R_0 \ = \ \{(i, i) \mid i = 1, 2, 3, ..., n\}$$

y las probabilidades de transición entre estados R_m, R_n están dadas por:

$$q_{m,n} = \begin{cases} 1 & \text{si } m \prec n \\ 0 & e.o.c. \end{cases}$$

Decimos que $m \prec n$ si, siendo P_m y P_n las particiones inducidas por R_m y R_n respectivamente, podemos formar P_n al combinar dos clases de equivalencia de P_m, tal que:

$$P_m \subset P_n$$

$$|P_m| = |P_n| + 1$$

Modelar el *n-coalescente* de esta manera nos permite considerar que la partición inducida por R_t consiste de los pares (i, j) con un ancestro vivo en el tiempo $t - t_0$. A partir de esta herramienta es posible construir árboles genealógicos que representan a los individuos simulados. Por ejemplo, considerando 5 poblaciones en el presente, podemos fácilmente representar la siguiente genealogía como una serie de eventos de coalescencia, combinando en cada nivel dos clases de equivalencia del nivel inferior (Figura 5).

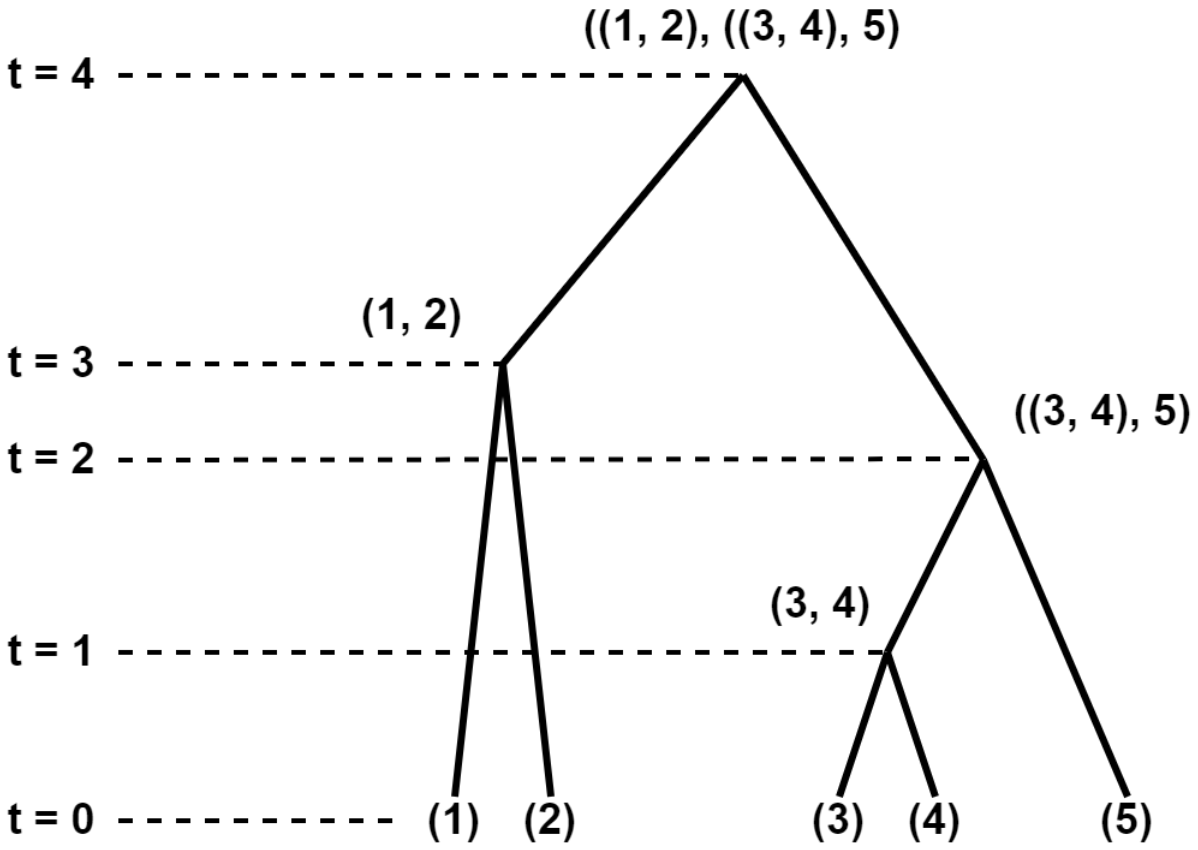


Figura 5: Árbol genealógico con eventos de coalescencia. Esta estructura nos permite representar la genealogía de una cantidad arbitraria de poblaciones o linajes. En este ejemplo, se consideran 5 linajes representados como las hojas del árbol. Los nodos internos del árbol indican eventos de coalescencia (viendo desde el presente hacia el pasado) en los cuales dos o más linajes se unen para formar uno solo.

Todos los datos genómicos simulados para cada individuo en este estudio fueron representados primero como un árbol de coalescencia. Esto permitió tomar muestras (secuencias) de un tiempo pasado que fueran ancestrales a los individuos modernos que se utilizarían tanto en el faseo poblacional como en el faseo con panel de referencia. De igual forma, esta metodología permite fácilmente codificar eventos demográficos como divergencia entre poblaciones.

3.3.2. Simulado de evolución y mutación en secuencias genómicas

A partir de los árboles de coalescencia, podemos simular secuencias de ADN que evolucionan a lo largo de las ramas del árbol. Una manera de simular esta evolución es a través de modelos de Markov[48].

Estos modelos asumen que la evolución de cada sitio en una secuencia de ADN es independiente e idéntica a la de otros sitios. De igual manera, la evolución de una rama del árbol es idéntica e independiente a las otras ramas[48].

Para modelar esta evolución, es necesario tener una matriz de probabilidad de transición. Sabiendo que cada sitio de ADN puede tener uno de cuatro estados dentro de $S = \{A, C, G, T\}$, se utiliza una matriz M con dimensiones $|S| \times |S| = 4 \times 4$ con las probabilidades de que en un sitio arbitrario ocurra una transición o transversión entre estados de S, parametrizadas sobre el tiempo t.

$$M = \begin{pmatrix} P_{A\to A}(t) & P_{A\to C}(t) & P_{A\to G}(t) & P_{A\to T}(t) \\ P_{G\to A}(t) & P_{G\to C}(t) & P_{G\to G}(t) & P_{G\to T}(t) \\ P_{C\to A}(t) & P_{C\to C}(t) & P_{C\to G}(t) & P_{C\to T}(t) \\ P_{T\to A}(t) & P_{T\to C}(t) & P_{T\to G}(t) & P_{T\to T}(t) \end{pmatrix}$$

Donde $P_{i\to j}(t) \mid i \in S$ denota la probabilidad de que una base en el estado i se transforme (o mantenga) al estado j en un tiempo t.

El modelo de Hasegawa, Kishino y Yano[49] permite diferenciar entre las tasas de transición $(A, C) \iff (G, T)$ y transversión $(A, G) \iff (C, T)$ a través de los parámetros α y β respectivamente. Este modelo también toma en cuenta la frecuencia esperada de cierta base, denotada por $\pi_i \mid i \in S$ para las probabilidades de mutación[49]. Los valores de las diagonales son los complementos necesarios para que la suma de cada hilera tenga un valor unitario.

$$HKY = \begin{pmatrix} - & \pi_C\beta & \pi_G\alpha & \pi_T\beta \\ \pi_A\beta & - & \pi_G\beta & \pi_T\alpha \\ \pi_A\alpha & \pi_C\beta & - & \pi_T\beta \\ \pi_A\beta & \pi_C\alpha & \pi_G\beta & - \end{pmatrix}$$

3.3.3. Eventos demográficos

Al hablar de eventos demográficos, nos referimos a cambios en la estructura de una población o un subconjunto de ésta. Para este trabajo, se simularon tres tipos de eventos demográficos: continuidad, cuellos de botella y divergencia entre un par de poblaciones.

Un cuello de botella indica una reducción drástica y súbita en el tamaño efectivo de una población, lo cual tiene como efecto la pérdida de diversidad genética y haplotípica en esta población. Por otro lado, la divergencia entre dos poblaciones indica dos poblaciones que formaban una sola población en algún punto del pasado. Esto corresponde a los eventos de coalescencia descritos anteriormente.

Esta investigación también busca caracterizar los efectos de los cambios en la estructura de las poblaciones estudiadas sobre la inferencia de haplotipos. Para modelar eventos demográficos utilizaremos las siguientes definiciones.

El conjunto:

$$P$$

Que contiene el conjunto de poblaciones de la forma $P_i \mid i \in \mathbb{N}$ consideradas por el modelo demográfico.

La función:

$$Pops(x) \ : \ \mathbb{N} \rightarrow \mathcal{P}(P)$$

Que regresa el conjunto de poblaciones existentes en el tiempo x, medido en generaciones desde el presente hacia el pasado.

Y finalmente, la función:

$$Ne(p, x) \ : \ P \times \mathbb{N} \rightarrow \mathbb{N}$$

Que dada una población p y un tiempo x medido en generaciones desde el presente hacia el pasado, regresa el tamaño de p en el tiempo x.

Con estas definiciones podemos construir tres modelos demográficos de interés. Es importante recordar que mientras los individuos modernos siempre corresponden a un tiempo 0 en relación al presente, los individuos antiguos pueden corresponder a cualquier

tiempo $x \in \mathbb{N}$ del pasado.

El escenario más simple es el caso de la continuidad, en el que los individuos antiguos y modernos pertenecen a la misma población, y el tamaño de la población se mantiene igual a una constante c a través del tiempo. En este caso, lo único que diferencía a un individuo antiguo de un individuo moderno es el daño acumulado en el tiempo y el distanciamiento genético entre la población ancestral y la presente.

Más formalmente, en este caso asignamos:

$$P = \{P_0\}$$

Y observamos que las siguientes condiciones se cumplen:

$$Pop(x) = \{P_0\} \; \forall x \in \mathbb{N}$$

$$Ne(P_0, x) = c \; \forall x \in \mathbb{N}$$

O su representación equivalente en diagrama, donde las poblaciones existentes se indican con líneas gruesas, y el tamaño de la población se muestra como líneas más delgadas rodeando la población correspondiente:

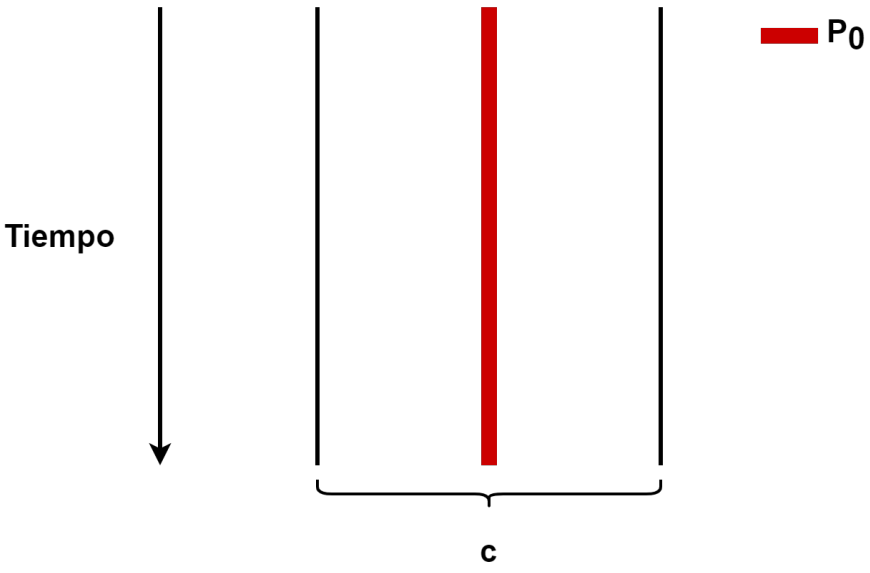


Figura 6: Representación gráfica del modelo de continuidad poblacional. Considerando el pasado y el presente como la parte superior e inferior de la gráfica respectivamente, el conjunto de poblaciones (en este caso, P_0) y su tamaño se mantienen constantes.

El segundo modelo demográfico considera un cuello de botella. Es decir, observando desde el presente hacia el pasado, la población se mantiene constante con un valor c_0 hasta llegar al tiempo del cuello de botella b, donde la población incrementa a un valor $c_1 \mid c_1 > c_0$, y se mantiene constante de nuevo. En otros términos el cuello de botella implica un evento en el pasado en el que la población sufrió una disminución considerable en su tamaño poblacional. El objetivo de simular este escenario es observar cómo cambia la calidad del faseo, dependiendo de si los individuos faseados son previos o posteriores al tiempo del cuello de botella.

Para este modelo igualmente consideramos:

$$P = \{P_0\}$$

Y consideramos el siguiente comportamiento de Pop y Ne:

$$Pop(x) = \{P_0\} \; \forall x \in \mathbb{N}$$

$$Ne(P_0, x) = \begin{cases} c_0 & \text{si } x \leq b \\ c_1 & \text{si } x > b \end{cases}$$

$$c_1 > c_0$$

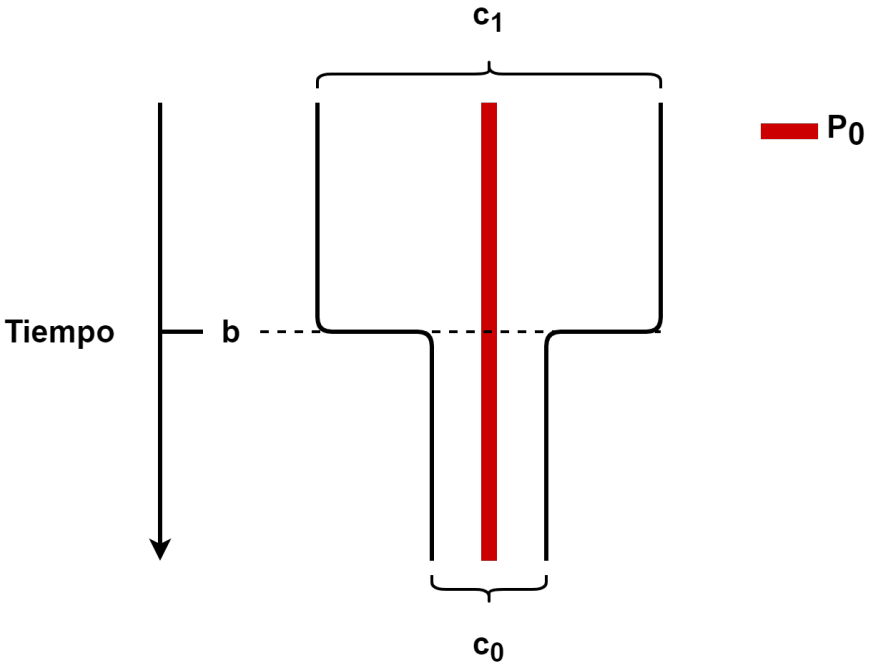


Figura 7: Representación gráfica del modelo con cuello de botella.

El último modelo comprende divergencia hace s generaciones entre una población formada por individuos antiguos, y una población moderna. El objetivo de este modelo es medir cómo cambia la calidad del faseo si la población de referencia es diferente a la población faseada. Esto se modela viendo desde el presente hacia el pasado, a dos poblaciones P_0, P_1 de tamaño constante c. Estas poblaciones se mantienen separadas hasta llegar al tiempo de divergencia s, donde la población P_1 se une a P_0, resultando desde ese momento en solo una población P_0 con tamaño $2c$. En este caso tenemos:

$$P = \{P_0, P_1\}$$

Con las siguientes condiciones en Pop y Ne:

$$Pop(x) = \begin{cases} P_0, P_1 & \text{si } x \leq s \\ P_0 & \text{si } x > s \end{cases}$$

$$Ne(P_1, x) = c \qquad \text{si } x \leq s$$

$$Ne(P_0, x) = \begin{cases} c & \text{si } x \leq s \\ 2c & \text{si } x > s \end{cases}$$

Que corresponde al siguiente diagrama:

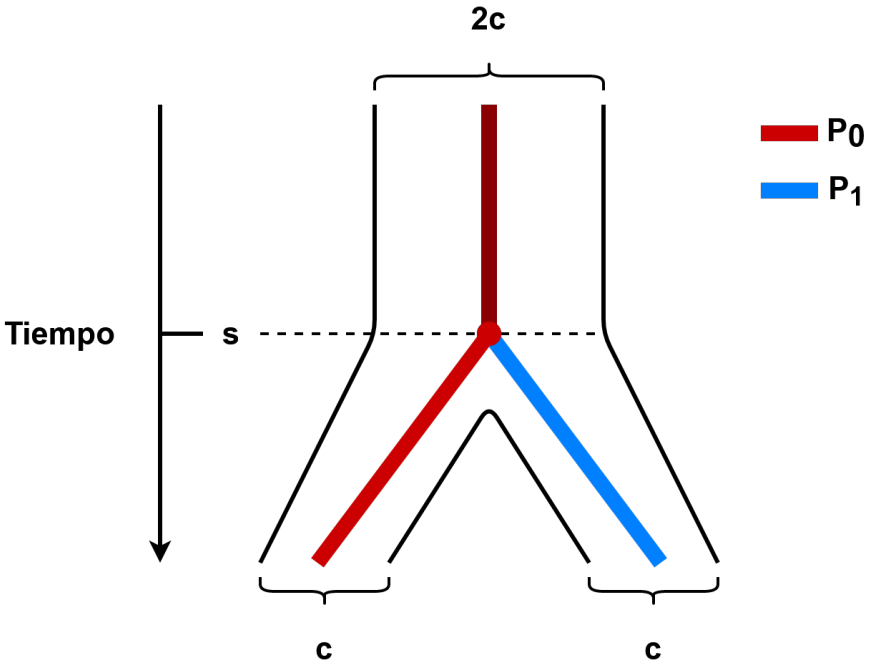


Figura 8: Representación gráfica del modelo con divergencia de poblaciones.

3.4. Identificación de divergencia de poblaciones a partir de Análisis de Componentes Principales (PCA)

Los haplotipos inferidos por las herramientas de faseo pueden ser analizados para, a su vez, inferir eventos demográficos en las poblaciones analizadas. Un ejemplo de cómo se puede hacer uso de información haplotípica para hacer inferencias sobre las poblaciones bajo estudio, es la detección de estructura genética[50][19][20]. Este tipo de análisis de estructura nos permite detectar eventos de mezcla genética, detectar eventos de migración, y medir distancia genética entre poblaciones actuales.

El programa *ChromoPainter*[38], dado un conjunto de N haplotipos donadores y M haplotipos recipientes, reconstruye cada haplotipo recipiente como un mosaico de haplotipos donadores, capturando así cuáles haplotipos donadores son esenciales para reconstruir el haplotipo recipiente. Esto lo hace *pintando* cada cromosoma recipiente como una combinación de cromosomas donadores.

Este algoritmo fue utilizado para medir la similitud genética, mediante haplotipos, entre poblaciones Británicas actuales e individuos antiguos de la edad de hierro[19]. También se empleó para detectar mezcla genética entre zoroastras iraníes con poblaciones Parsi modernas[20]. En ambos casos fue necesario primero inferir los haplotipos a partir de los datos de secuenciación de las muestras antiguas.

Para el presente trabajo, se utilizó este tipo de análisis para la reconstrucción de eventos de divergencia de poblaciones. Esto se hace con la intención de identificar la relación entre la antigüedad y calidad (contaminación y profundidad) de los datos genómicos faseados, y su utilidad para reconstruir estos eventos de divergencia.

También es importante mencionar que este análisis de estructura de poblaciones puede ser realizado sin tener acceso a datos haplotípicos, utilizando solamente los genotipos de los individuos considerados. En un estudio por Nievergelt et al. (2007), se mostró que se puede obtener una mejor diferenciación entre poblaciones humanas al utilizar datos haplotípicos[51], recalcando la importancia de contar con datos haplotípicos confiables en relación al análisis de poblaciones humanas. A continuación se describen ambos métodos, genotípico y haplotípico.

3.4.1. PCA sobre matriz genotípica

Para este tipo de análisis los únicos datos considerados son los genotipos de todos los individuos para cada sitio variante considerado. Esto significa que representamos a N individuos a lo largo de M variantes en una matriz de $N \times M$ donde cada elemento es parte del conjunto $\{0, 1, 2\}$; 0 para representar que ambos alelos de ese individuo son el alelo de referencia, 1 para indicar que sólo uno es el alelo de referencia, y 2 si ningún alelo es el de referencia[52][53].

Para construir la estructura de las poblaciones descritas en esta matriz de correlación, se examinan las variantes que maximizan la varianza dentro de esta matriz[53].

3.4.2. PCA sobre matriz de donadores (datos haplotípicos)

Para estos análisis, la matriz de donadores $M = R \times D$ representa R haplotipos recipientes como una combinación de D haplotipos donadores, esto significa que un valor $x = M_{i,j}$ corresponde a la proporción del cromosoma recipiente i que puede ser descrita utilizando al cromosoma donador j[54].

Como en el caso de PCA sobre datos genotípicos, los individuos que maximizan la varianza de esta matriz son clave para recuperar la estructura poblacional subyacente[38].

4. Objetivos e hipótesis

4.1. Objetivo general

El objetivo general de esta tesis es cuantificar el efecto de la antigüedad, la calidad, y la estructura demográfica de datos genómicos sobre los resultados de dos métodos diferentes de faseo: con panel de referencia y poblacional. La evaluación de dichos resultados comprende dos aspectos: la precisión con la que los haplotipos correctos son inferidos, y la calidad de posibles inferencias hechas a partir de estos haplotipos inferidos.

4.2. Objetivos específicos

- Implementar un *software* capaz de la simulación de datos genómicos. Es de vital importancia que esta herramienta sea flexible respecto a parámetros como cantidad

de individuos, longitud de secuencias, antigüedad y calidad de datos, e historia demográfica.

- Implementar una tubería de análisis para el procesamiento de la gran cantidad de datos genómicos simulados para la inferencia de haplotipos. Esta tubería también debe ser lo más similar posible a las tuberías de análisis utilizadas en los estudios tomados como referencia. Esto incluye también la flexibilidad de elegir el faseo con panel de referencia o el faseo poblacional.

- Analizar la precisión de los haplotipos inferidos en relación a los parámetros con los que fueron simulados los datos genómicos.

- Usar los haplotipos inferidos para la reconstrucción de la historia demográfica con la que se simularon los datos. De igual manera, se busca relacionar la habilidad de reconstruir la historia originalmente simulada con los parámetros de los datos.

4.3. Hipótesis

Disminuir la calidad (i.e. disminuir la profundidad y aumentar la contaminación) y aumentar la antigüedad de los datos simulados tendrá un efecto negativo sobre ambos métodos de inferencia de haplotipos, y cualquier aplicación o inferencia que dependa de estos haplotipos. Crucialmente, existirán datos simulados con tan poca calidad que no se tendrá certidumbre sobre los haplotipos recuperados y por lo tanto de cualquier inferencia hecha a partir de ellos. Esto a raíz de que la herramienta estudiada no fue diseñada con datos paleogenómicos en mente (aunque se haya utilizado con ese fin en estudios previos).

5. Diseño experimental

Las simulaciones producidas para este estudio pueden ser definidas con base en los siguientes componentes:

- Antigüedad y calidad (profundidad y contaminación) de datos genómicos simulados.

- Métodos de faseo.

- Eventos demográficos.

- Longitud de las secuencias simuladas.

A continuación se describen estos componentes, junto con sus propósitos y justificaciones para los valores escogidos.

5.1. Daño y calidad de datos genómicos simulados

Para tener un alcance apropiado, se eligieron múltiples combinaciones de antigüedad y calidad para los datos de ADNa simulados. La simulación considera tres parámetros para controlar esto:

- Edad: en generaciones, comenzando desde el presente hacia el pasado.
- Profundidad: valor de profundidad de cobertura promedio, antes de cualquier tipo de filtros de calidad.
- Contaminación: porcentaje de ADN moderno que contamina la muestra.

La siguiente tabla muestra los valores escogidos para estos parámetros.

Parámetro de calidad	Valores simulados
Antigüedad (generaciones)	0 (presente), 25, 50, 100, 200, 400
Profundidad	1×, 5×, 10×
Contaminación	0 %, 2 %, 5 %, 10 %

Esto resulta en 72 diferentes combinaciones.

Estos valores fueron escogidos para ser ilustrativos de los datos comúnmente observados en estudios de ADNa. Las profundidades entre 1× y 10× pueden representar a la mayoría de los datos utilizados en los estudios descritos previamente[19][20][21][22]. Por otro lado, aunque estos estudios no sobrepasan una contaminación en muestra del 2 %, se agregaron valores de contaminación del 5 % y 10 % para tener más información sobre el efecto de la contaminación en este tipo de análisis.

Finalmente, se eligió simular el rango de antigüedad de entre 0 y 400 generaciones para tener valores representativos desde el periodo neolítico europeo (antigüedad de aproximadamente 10,000 años) hasta el siglo XVI A.D. Estas edades abarcan los rangos de tiempo estudiados en los artículos usados como referencia[19][20][21][22]. Es importante mencionar que datos simulados con una edad de 25 generaciones serán representativas de muestras que corresponden al tiempo de la conquista de América, lo cual es de especial interés cuando se simularon cuellos de botella como eventos demográficos.

5.2. Métodos de faseo

Se realizó el análisis del faseo con panel de referencia, y faseo poblacional.

Para implementar estos métodos de faseo sobre datos antiguos simulados, es necesario también simular datos modernos. En la literatura, normalmente se utiliza el panel de referencia creado por *1,000 Genomas*, el cual consiste completamente de individuos modernos. De igual manera, los ejemplos de faseo poblacional en investigaciones previas agrupan individuos antiguos y modernos que después son faseados en conjunto.

Por lo tanto, en el caso del análisis de faseo con panel de referencia, se construyó un panel de referencia de haplotipos a partir de cromosomas modernos simulados. Para el faseo poblacional, se simularon individuos modernos junto a los individuos antiguos, creando un solo grupo de individuos para el faseo.

En adelante, estos individuos modernos serán referenciados también como individuos de referencia al hablar del faseo de haplotipos.

5.3. Inclusión de eventos demográficos

Los valores elegidos para las constantes que parametrizan estos modelos (definidas junto con los eventos demográficos en la sección 3.3.3) se encuentran en la siguiente tabla:

Divergencia de poblaciones	
c	1×10^4
s	50, 100, 200

Cuello de botella	
c_0	1×10^3
c_1	1×10^4
s	25

Considerando un tiempo de generación de 25 años, estos tiempos de divergencia entre poblaciones coinciden aproximadamente con periodos migratorios del sudeste de Asia y Melanesia (3,000 a 950 B.C.)[55], o el periodo de migraciones germánicas (siglos III a VII A.D.)[56].

El tiempo y magnitud del cuello de botella fueron elegidos para coincidir con el importante decremento en el tamaño poblacional de la población indígena a raíz de la conquista y colonización de América hace 500 años[57].

5.4. Estructura de los datos simulados

Se comparte la siguiente estructura para todas las simulaciones, de todos los modelos:

Tipo de individuos	Número de individuos simulados
Modernos	500
Antiguos	100
Referencia (alineamiento)	1
Contaminación	1

5.5. Longitud de las secuencias simuladas

Para el cálculo de SWE, se simularon secuencias de ADNa con longitud de 2 millones de pares de bases (2 Mbp). Esta longitud fue suficiente para los análisis de exactitud de faseo. Incrementar la longitud de las secuencias para todas las simulaciones no era viable al considerar el tiempo de cómputo resultante.

En cuanto a las simulaciones que fueron utilizadas para el análisis de componentes principales a través de *ChromoPainter* en la etapa de reconstrucción de eventos demográficos (ver Figura 9), se utilizaron secuencias con una longitud de 20 Mbp. Esta longitud es más comparable con la del cromosoma 21 en los humanos (46 Mbp). La razón para incrementar la longitud de las secuencias es el hecho de que *ChromoPainter* espera entradas de cromosomas completos, y esto nos permitirá obtener resultados más cercanos a los que se obtendrían con datos reales[38]. Dado que simular, procesar, y fasear secuencias más largas es computacionalmente más costoso, no se simularon secuencias más grandes a 20 Mbp.

6. Metodología

6.1. Diseño de la tubería de análisis o *pipeline*

En la figura 9 se presenta un esquema de la tubería desarrollada para la simulación de datos genómicos, procesamiento y faseado de éstos, y análisis de los resultados.

Todo el *pipeline* de simulación, procesamiento, faseo, y análisis de datos está publicado en línea como un repositorio de GitHub[58]. Todas las etapas de esta tubería se ejecutaron en un cluster de cómputo de alto rendimiento del Laboratorio Nacional de Visualización Científica Avanzada (*LAVIS*). Aunque este cluster cuenta con más de 1,000 núcleos y 4 TB de memoria, solo se llegaron a utilizar un máximo de 80 núcleos y 500 GB de memoria.

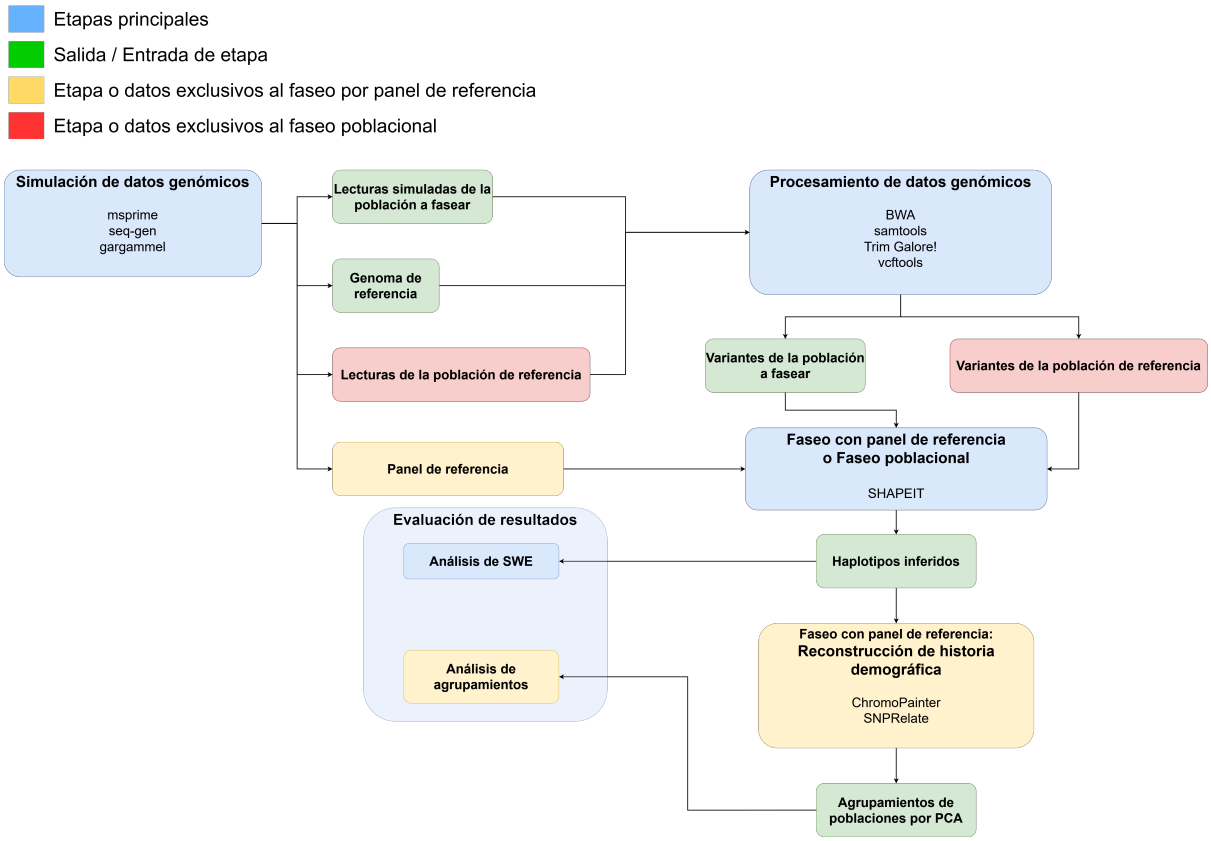


Figura 9: Se consideran cuatro etapas principales (azul): simulación, procesamiento, faseo, y análisis de resultados. Cada una de estas etapas incluye las herramientas utilizadas para su ejecución. La entrada y salida de estas etapas puede variar dependiendo del método de faseo elegido (amarillo y rojo), mientras que algunas de las estructuras de datos de entrada o salida se mantienen constantes para todas las ejecuciones del *pipeline* (verde).

A continuación se presentan descripciones detalladas de la ejecución de todas es-

tas etapas, haciendo referencia a los archivos de implementación en el repositorio de GitHub[58].

6.2. Simulación de datos genómicos

La figura 10 muestra el flujo para la simulación de datos genómicos modernos y antiguos. Esto incluye simular los parámetros de calidad de los datos.

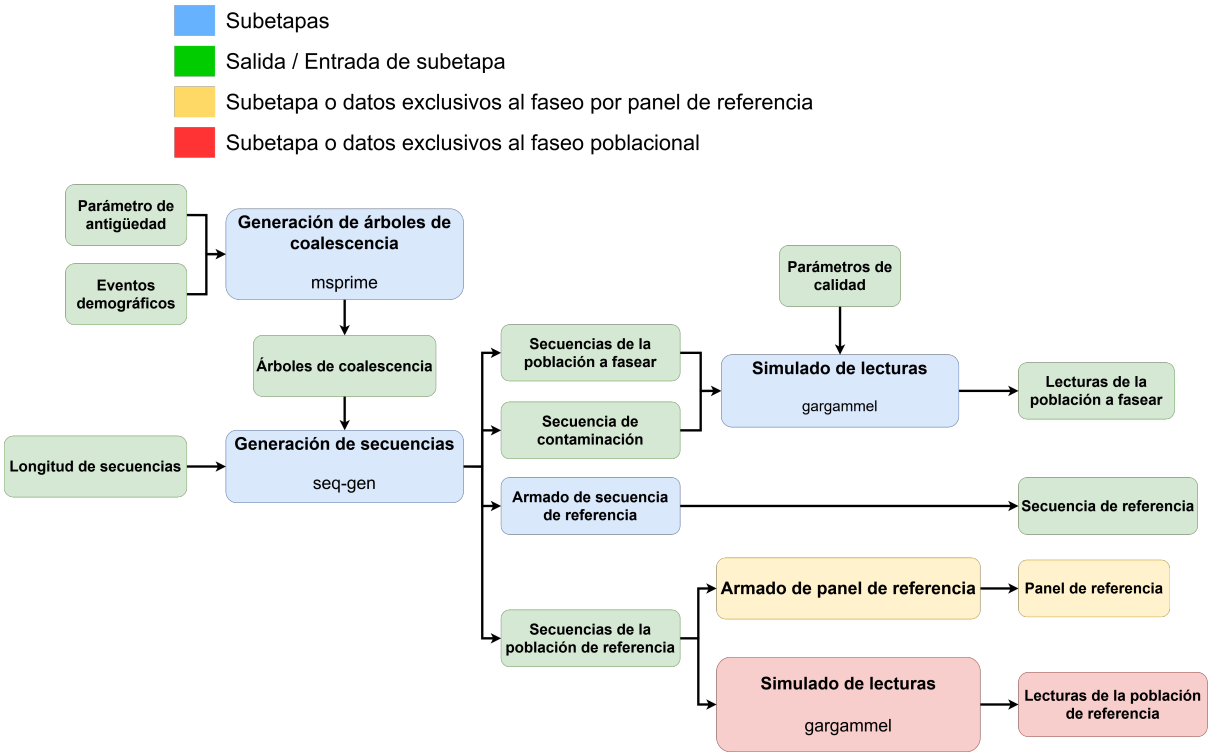


Figura 10: Flujo para la etapa de simulación de datos.

- Generación de árboles de coalescencia (**genomic_datasim/gen_genomes.py**): modelado de los árboles que describen la población o poblaciones de las cuales generamos datos genómicos. Esto incluye especificar la antigüedad de los individuos ancestrales, eventos demográficos, y tasas de mutación y recombinación. En todos los casos, se eligió el valor 2×10^{-8} para las tasas de mutación y recombinación, este valor es parecido a la tasa de mutación por base por generación promedio en los humanos de $\sim 1{,}6 \times 10^{-8}$[59][60].

  Para el modelado de árboles de coalescencia, se utilizó la biblioteca *msprime*[61] a través de su API para *Python*. Esta herramienta es una reimplementación de la

herramienta *ms*[62], pensada para ser mucho más eficiente en cuanto a tiempo de ejecución y espacio necesario para representar los árboles generados.

- Generación de secuencias (**genomic_datasim/gen_genomes.py**): a partir de los árboles de coalescencia de la etapa previa, se generan dos secuencias por individuo para simular organismos diploides. Se simularon 502 individuos modernos, 500 conformaron la población de referencia al fasear. Los dos individuos restantes se utilizaron para el genoma de referencia y para la contaminación moderna respectivamente. En cuanto a individuos antiguos, se simularon 100 individuos para fasear.

  Éstas secuencias se aparean y son tratadas como unidad durante el resto del procesamiento. Es en este paso en el que se especifica la longitud de las secuencias simuladas. Las secuencias fueron simuladas con longitudes de 2 Mbp para los cálculos de SWE, y 20 Mbp para la reconstrucción de eventos demográficos.

  Para esta generación de secuencias se utilizó la herramienta *seq-gen*[48]. La producción de secuencias funciona a partir de los árboles de coalescencia junto con el modelo HKY[49] (especificado a través de la bandera `-mHKY`) explicado previamente para la simulación de mutaciones.

  Se tuvieron que realizar modificaciones al código fuente en *C* de *seq-gen*, ya que al generar miles de secuencias con longitudes de millones de pares base se observaban problemas de corrupción de memoria. Esto ocurría ya que *seq-gen* utiliza una estrategia de manejo de memoria similar a una *arena de memoria*[63]. El tamaño de la arena no fue suficiente para los requisitos de este trabajo así que se realizó una mejora al código de *seq-gen* eliminando esta falla[64].

- Armado de secuencia de referencia (**genomic_datasim/merge_reference.py**): uno de los individuos modernos simulados en el paso anterior es utilizado para crear un genoma de referencia. Este genoma fue utilizado solamente para el alineamiento de nuestras lecturas simuladas. Para armar esta referencia, se camina a lo largo de las dos secuencias, eligiendo un alelo al azar cada vez que encontremos un sitio segregante.

- Armado de panel de referencia (**genomic_datasim/create_panel.py**): sólo en el caso de faseo por panel de referencia. Utilizando el formato de *SHAPEIT*[65] (Figura

11 se consideraron todos los sitios posibles, creando una matriz de $n \times m$. Aquí, n corresponde al número de sitios simulados, y m al número de haplotipos de los individuos modernos. En el caso donde simulamos 500 individuos modernos para nuestro panel de referencia, $m = 1,000$. Cada celda (x, y) indicará si el haplotipo y tiene en el sitio x el alelo de referencia (0), el alelo alterno (1).

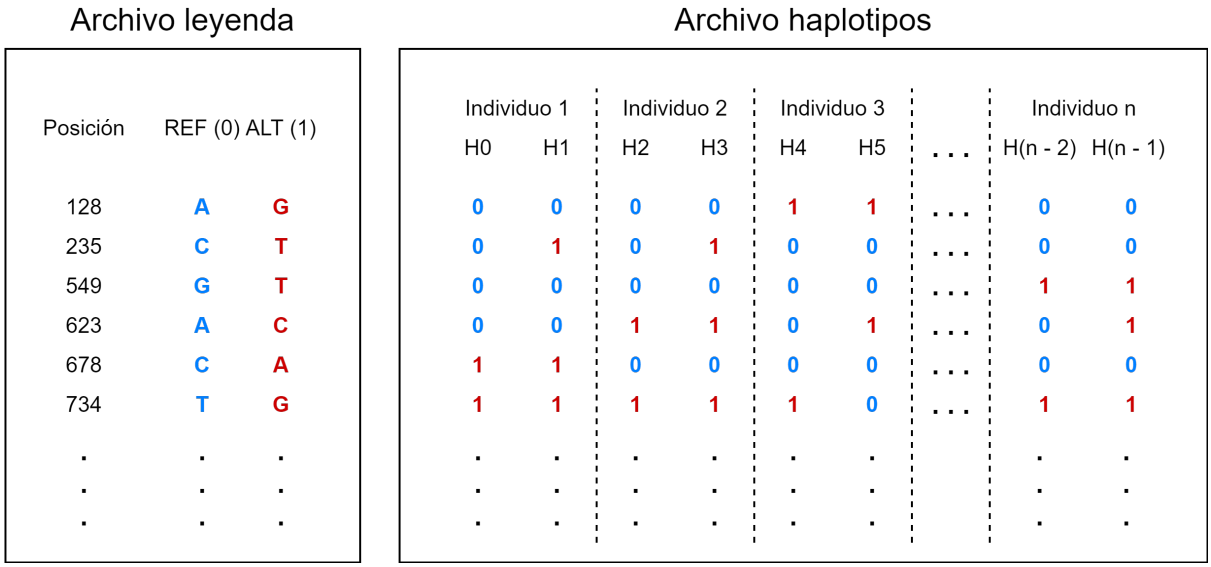


Figura 11: Representación en matriz del panel de referencia. El panel debe considerar tanto alelos de referencia (azul) como alelos alternativos (rojo).

- Simulado de lecturas (**genomic_datasim/gargammel.sh**): a partir de las secuencias ancestrales simuladas, se generaron lecturas que incorporan la baja calidad característica de las muestras antiguas: deaminación, fragmentación, baja profundidad, contaminación, errores de secuenciación, etc. Esto se realizó con el programa *gargammel*[66]. Este programa emplea parámetros estimados a partir de datos reales observados en muestras de ADNa, como la distribución de longitud de lecturas, la probabilidad de eventos de deaminación a lo largo de las lecturas, el promedio de profundidad para cada sitio, y errores de secuenciación[67]. Asimismo *gargammel* también es capaz de introducir contaminación en las lecturas simuladas.

  En el caso de faseo con panel de referencia, sólo se simularon lecturas para la población a fasear (ancestral). Mientras que para el faseo poblacional adicionalmente se generaron lecturas de alta calidad para los individuos modernos.

Dado que *gargammel* fue originalmente concebido para trabajar con *ms* en vez de *msprime*, la funcionalidad implementada en **genomic_datasim/gen_genomes.py** re-

sultó más eficiente y configurable que el *driver* ejemplo oficial de *gargammel*. Una versión modificada del *driver* creado para este trabajo fue aceptado como nuevo ejemplo en la página oficial de *gargammel*[68].

6.3. Procesamiento

Para procesar todos los individuos ancestrales de manera paralela, se utilizaron las opciones de paralelización del software de manejo de clusters *SGE* con el que trabaja el *cluster* de cómputo utilizado. La figura 12 muestra las herramientas utilizadas para transformar las lecturas simuladas en variantes que pueden ser faseadas.

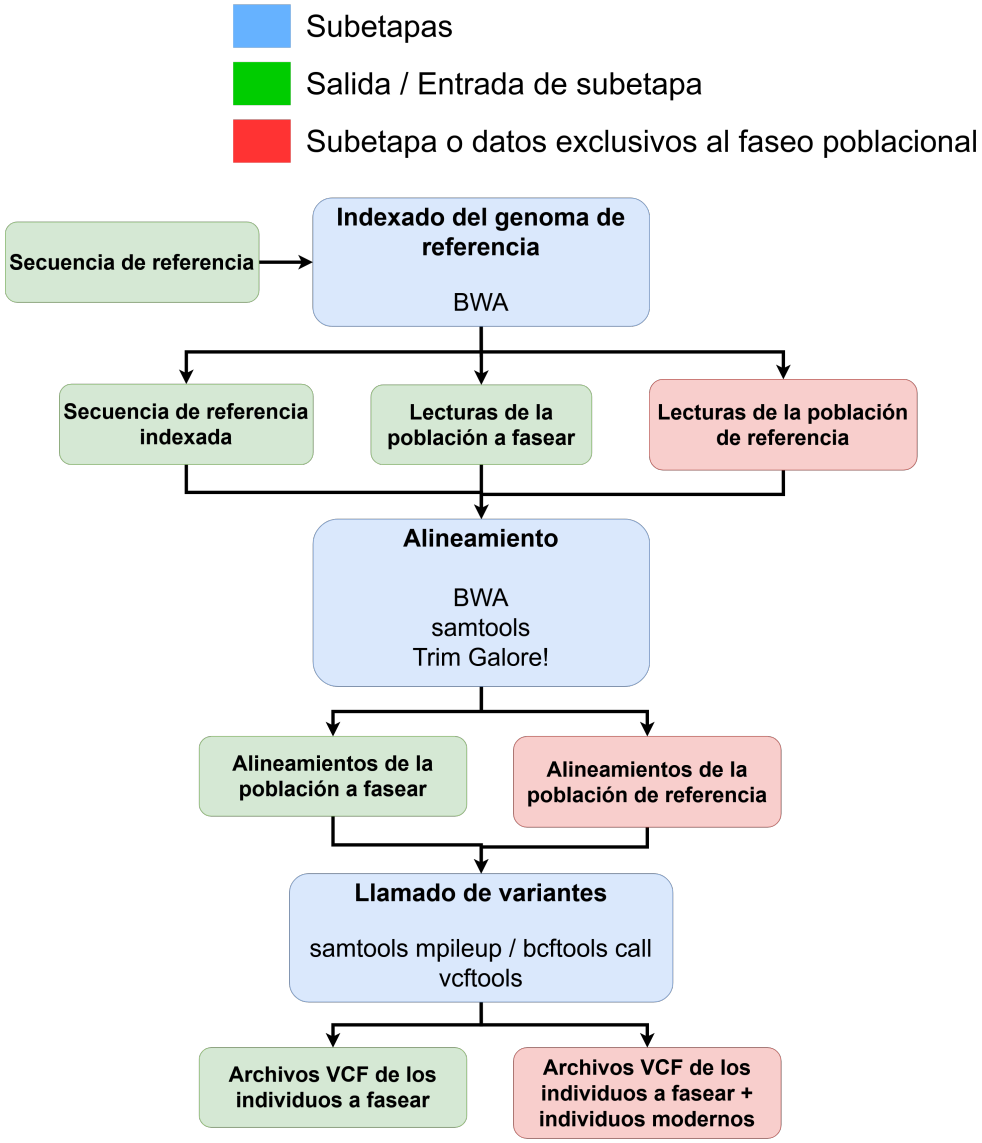


Figura 12: Flujo para la etapa de procesamiento.

- Indexado del genoma de referencia (**genomic_datasim/pre_map.sh**): el indexado de un genoma o secuencia de referencia es realizado para acelerar la etapa de alineamiento a este mismo. El resultado de indexar un genoma es una base de datos que contiene diferentes sufijos de pares de bases vistos como cadenas de texto, junto con su ubicación en la secuencia de referencia[69].

  Para este paso se utilizó la herramienta BWA[69] con la bandera `-a bwtsw`, esta bandera especifica el uso del algoritmo BWT-SW, el cual funciona con regiones más largas que el algoritmo default[69].

- Alineamiento (**genomic_datasim/map.sh**): se alinearon las lecturas contra el genoma de referencia previamente simulado e indexado. Este alineado se puede modelar como una búsqueda de subcadenas (lecturas) dentro de una cadena mucho más grande (genoma de referencia). Contar con el índice de sufijos de la secuencia de referencia permite evitar la solución a fuerza bruta de buscar linealmente cada lectura[69].

  Antes del alineamiento, es preciso eliminar los extremos de las lecturas de ADN simulado usando una herramienta de recortado (*trimming*). Esto es porque *gargammel* también simula la secuencia de los adaptadores utilizados en la construcción de librerías de secuenciación real. Esto se realizó con la herramienta *Trim Galore!*[70] en la modalidad de recortado de lecturas pareadas (bandera `--paired`).

  Después del recortado, las lecturas se alinearon al genoma de referencia utilizando BWA *mem*. Estos alineamientos se ordenaron respecto a coordenadas genómicas y fueron filtrados para eliminar duplicados utilizando la herramienta *samtools*[71] y sus comandos `sort` y `rmdup`.

  En el caso de faseo poblacional, todos estos pasos se realizaron también para las lecturas de la población moderna.

- Llamado de variantes (**genomic_datasim/call_variants.sh**): se generaron archivos con llamados de variantes, o VCF, para todas las lecturas alineadas en el paso anterior. Este formato nos permitió saber en cuáles posiciones los alelos de un genoma difieren de los alelos del genoma de referencia (variantes)[52].

  Este llamado de variantes se realizó con los comandos `samtools mpileup` y `bcftools call`. Ambos pertenecientes a `samtools`[71]. Se utilizó la bandera `-m` para `bcftools`

`call`, esto indica un método de llamado de variantes que supera algunas limitaciones del método por defecto[72].

Este paso incluyó remover sitios trialélicos del análisis y filtrar la calidad de las variantes recuperadas a través de la herramienta *vcftools*[52]. La calidad de una variante se indica a través de un *puntaje de calidad Phred*. Este puntaje Q se define como $Q = -10log_{10}E$, donde E es la probabilidad de que la base variante haya sido incorrectamente identificada[73]. Todas las variantes utilizadas en este trabajo fueron filtradas para un puntaje de calidad Phred mínimo de 90.

Este paso se realiza en paralelo para todos los individuos para los que se llamaron variantes.

6.4. Faseo

La etapa de faseo descrita en la figura 13 infiere los haplotipos originalmente simulados para cada individuo.

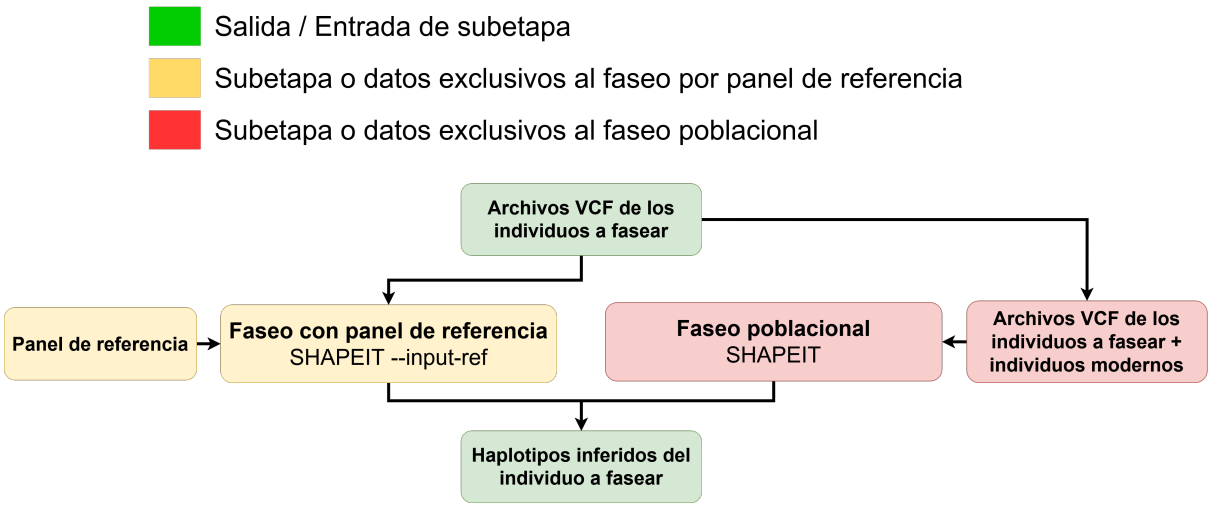


Figura 13: Flujo para la etapa de faseo. Esta etapa se repitió 500 veces en paralelo para cada simulación, una vez por cada individuo a fasear.

- Faseo con panel de referencia (**genomic_datasim/phase_ref.sh**): se realizó el faseo para cada individuo de manera paralela utilizando *SHAPEIT*[32]. Se construyó un archivo VCF para cada individuo a fasear. Se especificaron los archivos del panel de referencia generados en la etapa de simulación de datos del *pipeline* (Sección 6.2) con la bandera `--input-ref`.

- Faseo poblacional (**genomic_datasim/phase_pop.sh**): para el faseo poblacional, se generó un archivo VCF para todo individuo de la población a fasear. Este archivo VCF incluyó al individuo a fasear y a todos los individuos de la población moderna. Contrastando con los casos de faseo con panel de referencia, donde cada archivo VCF contenía solamente la información del individuo a fasear. Esto nos permitió fasear a cada individuo de la población ancestral de manera independiente, conservando siempre los mismos datos de la población moderna. Aunque los individuos modernos también fueron faseados, no se tomaron en cuenta los resultados de la inferencia de haplotipos de estos individuos.

  De igual manera, se realizó el faseo para cada individuo de manera paralela con la herramienta *SHAPEIT*[32].

Para ambos métodos de faseo, se especificó también la tasa de recombinación (2×10^{-8}), que es la misma con la que se simularon los árboles de coalescencia.

6.5. Reconstrucción de eventos demográficos

La figura 14 muestra cómo se utilizaron los haplotipos inferidos y datos genotípicos para reconstruir la historia demográfica de los datos.

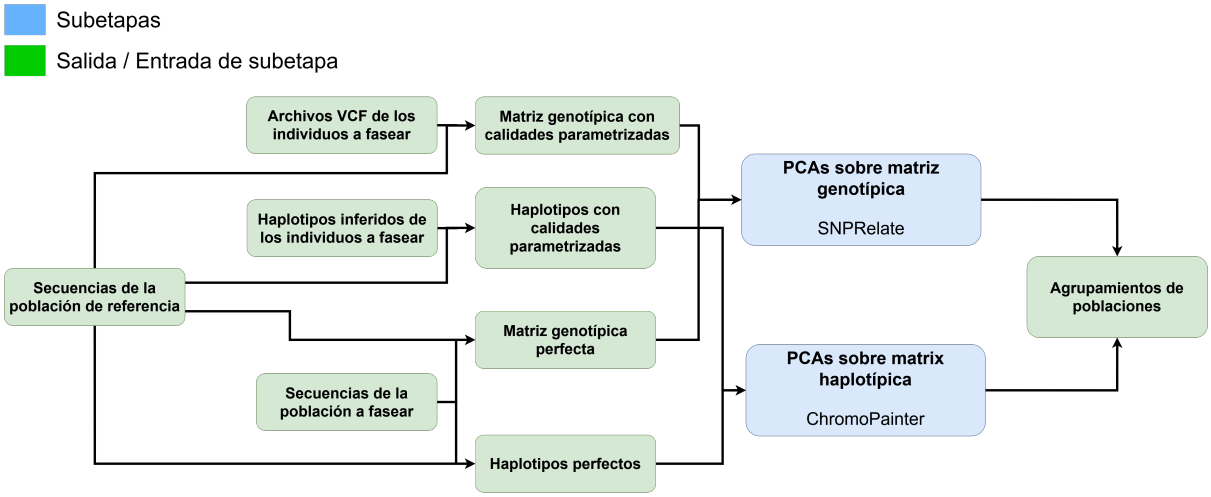


Figura 14: Flujo para la etapa de reconstrucción de eventos demográficos. Se utilizaron datos con calidades parametrizadas y datos perfectos para medir los efectos de la calidad sobre la reconstrucción.

Esta etapa sólo se ejecutó sobre los datos obtenidos con el faseo con panel de referencia (por razones de tiempo de ejecución), y el evento demográfico de divergencia de

poblaciones hace 200 generaciones. Asimismo, solamente se consideraron los datos simulados con coberturas de 10× y 5×, dado que los datos con 1× resultaban en tan pocos sitios recuperados que las herramientas no podían operar adecuadamente sobre ellos.

Para evaluar el efecto de los diferentes parámetros sobre el Análisis de Componentes Principales (PCA) generados con *ChromoPainter*, el tamaño de las regiones simuladas aumentó a 20 Mbp. Este aumento en el tamaño fue para es para acercarse más al tamaño de un cromosoma humano pequeño, ya que *ChromoPainter* está diseñado para este tipo de longitudes[38].

Finalmente, se hicieron PCAs sobre cuatro tipos de datos diferentes:

- Datos genotípicos «perfectos» (**pca_analysis/run_perfect_vcf.sh**).

  Sin ningún daño por antigüedad, sin contaminación, y sin ninguna incertidumbre respecto a las variantes existentes. En este caso, el único factor que determina la diferencia entre individuos antiguos y modernos es la divergencia genética esperada dado el tiempo desde su separación (200 generaciones). Esta divergencia puede ser utilizada para diferenciar las poblaciones[74]. Estos datos genotípicos se obtuvieron de las secuencias generadas por *seq-gen*, en la etapa de simulación de datos genómicos (Sección 6.2).

  El tener datos genotípicos «perfectos» permite contar con una referencia para interpretar cómo el decremento de la calidad al usar datos de secuenciación antiguos con diferentes parámetros (antigüedad, contaminación y profundidad) puede afectar los resultados del PCA. Asimismo sirve como punto de comparación para evaluar diferencias entre el uso de genotipos y haplotipos «perfectos» en este tipo de análisis.

- Datos haplotípicos «perfectos» (**pca_analysis/run_perfect_chromo.sh**).

  Similar a los datos genotípicos perfectos, estos haplotipos perfectos se calcularon directamente a partir de las secuencias generadas por *seq-gen*. Estos datos se usaron como referencia para comparar los resultados de análisis de estructura de poblaciones a partir de datos haplotípicos con peores calidades.

- Datos genotípicos con calidades parametrizadas (**pca_analysis/run_vcf.sh**).

  Estos datos se obtuvieron después del paso de llamado de variantes en la etapa de procesamiento (Sección 6.3). Contar con estos datos nos permitió comparar con los

resultados de datos genotípicos perfectos, y datos haplotípicos con calidades parametrizadas. Estos archivos se generaron a partir de los archivos de variantes para la población antigua junto con las secuencias originales de los individuos modernos.

Al comparar con los datos genotípicos perfectos, se pudieron aislar los efectos de diferentes niveles de calidad sobre los PCA a base de datos genotípicos. Por otro lado, al comparar con los datos haplotípicos con calidades parametrizadas, observamos si contar con datos haplotípicos mejora el desempeño de un análisis de estructura de poblaciones que podría haber sido negativamente afectado por la baja calidad.

- Datos haplotípicos con calidades parametrizadas (**pca_analysis/run_chromo.sh**).

  Estos datos son los resultados de introducir los parámetros de calidad (contaminación, profundidad, y daño por el tiempo) a los datos antes de fasearlos. Suponiendo que la calidad de los datos influye sobre la calidad de los haplotipos recuperados, el propósito es analizar qué tanto la calidad de los haplotipos inferidos afecta la reconstrucción de la historia demográfica simulada.

Para los análisis con datos genotípicos, se creó un archivo VCF que contiene todas las variantes de todos los individuos en ambas poblaciones. Este archivo VCF es creado a partir de las secuencias de *seq-gen* directamente (**pca_analysis/perf_vcf.py**) en el caso de los datos genotípicos perfectos, o juntando todos los archivos VCF generados en el paso de llamado de variantes para los datos con calidades parametrizadas. La matriz genotípica en ambos casos se construyó utilizando el paquete de *R*, *SNPRelate*[53], el cual opera directamente sobre archivos VCF.

Cuando se utilizaron datos haplotípicos, las matrices de donadores para PCA (Sección 3.4.2) fueron construidas utilizando *ChromoPainter*. Estos datos haplotípicos fueron generados a partir de las secuencias simuladas con *seq-gen* para los datos perfectos, o después de la etapa de faseo para los datos haplotípicos con calidades parametrizadas.

Se utilizaron los parámetros recomendados por los desarrolladores de *ChromoPainter*[54], de tal manera que todo haplotipo simulado (tanto moderno como ancestral) fuera *pintado* con todos los otros haplotipos. Finalmente, el PCA fue realizado sobre la matriz de donadores, de acuerdo a las recomendaciones disponibles en el sitio de *ChromoPainter*.

6.6. Cálculo de SWE y análisis de agrupamientos

Los haplotipos inferidos en la etapa de faseo fueron comparados con los haplotipos verdaderos de las secuencias que generadas por *seq-gen* en la etapa de simulación (**genomic_datasim/get_switch_err.py**). Se obtuvo el SWE para cada individuo, y después estos valores se procesaron juntos para obtener una distribución del SWE (Figura 15).

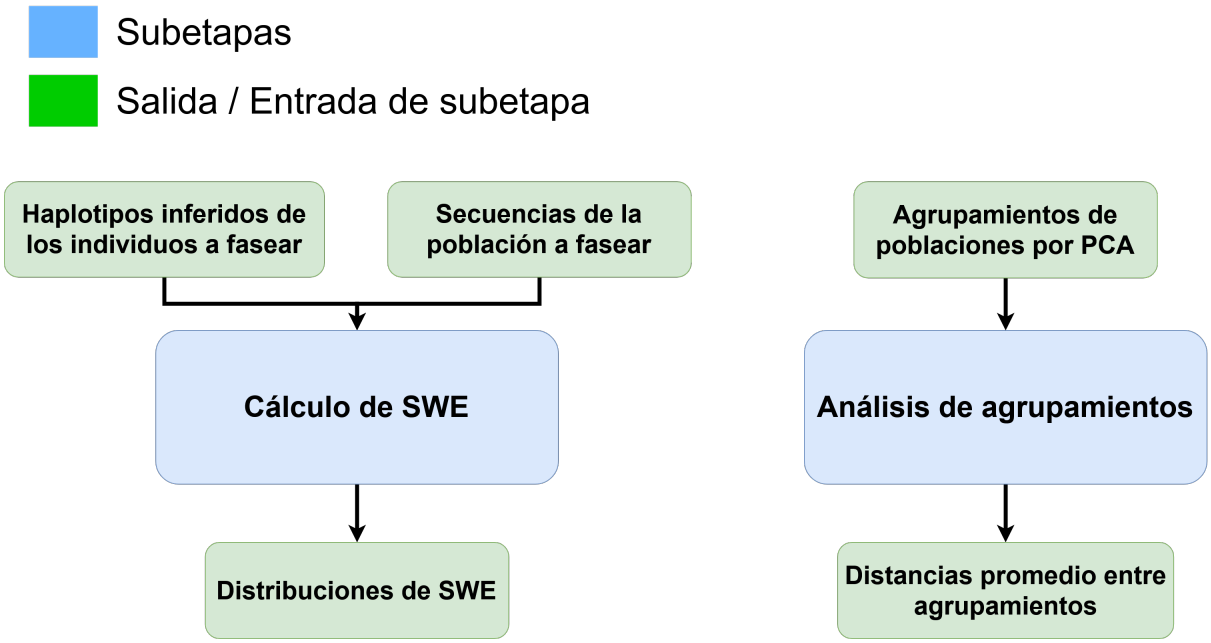


Figura 15: Flujo para la etapa de análisis de resultados. Todas las figuras resultantes de estas etapas fueron generadas en *R* utilizando el paquete *ggplot2*[75].

Estas distribuciones se importaron a un script de *R*, donde se generaron las visualizaciones finales. Este script hizo uso de varias librerías para manipulación de matrices y gráficas*ply*[76][77][78].

Una vez teniendo los resultados de los PCA para la identificación de estructura de poblaciones, éstos se graficaron para medir qué tan efectivamente se separaron las poblaciones ancestrales y modernas al utilizar diferentes datos. Estas separaciones de agrupamientos se midieron como las distancias euclidianas entre los centroides del agrupamiento moderno y el agrupamiento ancestral.

7. Resultados

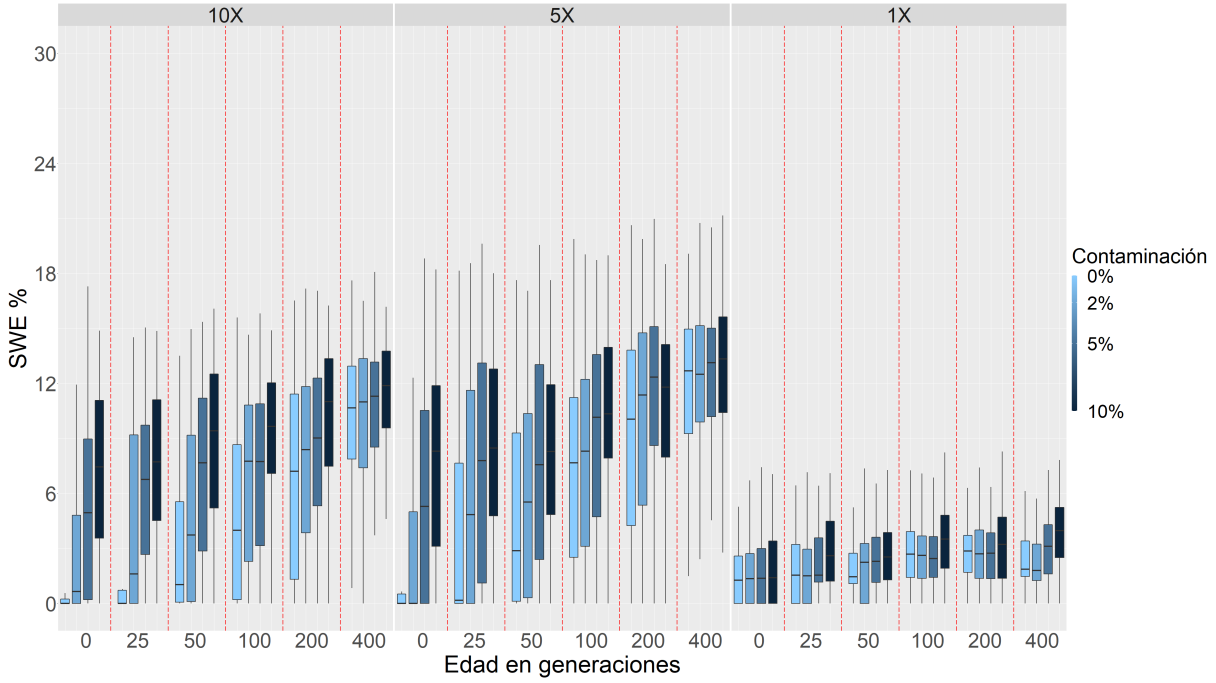
7.1. Calidad de faseo con panel de referencia

Para el faseo con panel de referencia se consideraron tres categorías de simulaciones: Simulaciones de continuidad poblacional, simulaciones con divergencia entre población moderna y antigua, y simulaciones con cuello de botella.

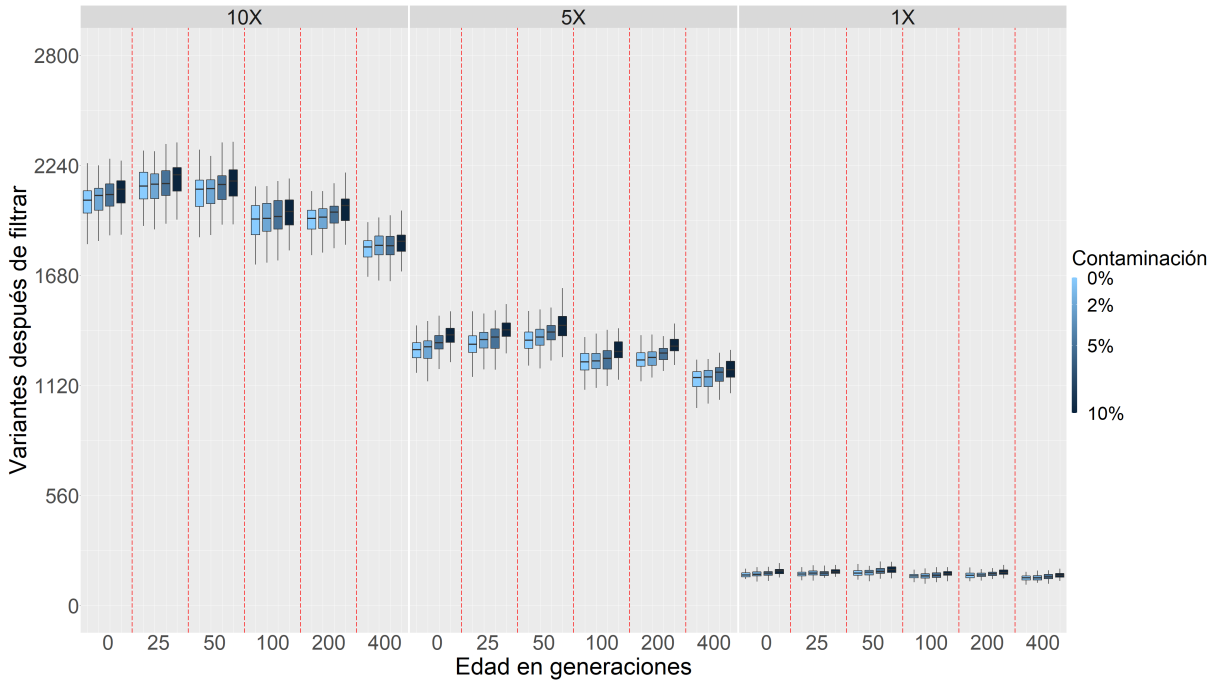
7.1.1. Simulaciones de continuidad poblacional

Los primeros resultados (Figura 16) no consideran ningún evento demográfico en la población, es decir, el tamaño poblacional permanece constante y la población actual que se usa como panel de referencia es descendiente directa de la población antigua (Figura 6). Esto significa que los únicos parámetros que cambian son la antigüedad de los individuos, la profundidad, y el nivel de contaminación.

El simular y evaluar el faseo de individuos con una edad de 0 generaciones (modernas) nos permite calcular la precisión esperada para datos modernos, es decir el tipo de datos para los que se desarrolló la herramienta *SHAPEIT*. De igual manera, simular datos paleogenómicos de continuidad poblacional nos permite evaluar cómo se ve afectado el desempeño de la herramienta al existir diversos eventos demográficos (divergencia entre poblaciones y cuellos de botella). Así bien, tener estos dos puntos de referencia permite aislar los efectos de diferentes calidades e historias demográficas.



(a)



(b)

Figura 16: Puntuaciones de SWE y variantes recuperadas para simulaciones de continuidad poblacional. Faseo con panel de referencia. **(a)** Las puntuaciones de SWE se presentan divididas en tres facetas correspondientes a profundidades promedio de 10×, 5×, y 1×. Dentro de estas facetas existe una distribución de valores de SWE para los 100 individuos simulados y faseados. Para cada combinación de antigüedad y contaminación. Las divisiones verticales rojas dividen las distribuciones correspondientes a diferentes valores de edad en generaciones. Los diferentes valores de contaminación (0 %, 2 %, 5 % y 10 %) se representan con diferentes tonos de azul. **(b)** Cantidad de variantes recuperadas después de filtrar para cada conjunto de individuos faseados. Nótese que los datos con profundidad promedio de 1× muestran muy pocas variantes recuperadas, y que generalmente recuperamos menos variantes para individuos más antiguos. Agregar contaminación siempre incrementa las variantes recuperadas.

En este sentido, se observó que incrementar la antigüedad o la contaminación de los datos resulta en una tasa de error más elevada, junto con una distribución del error más amplia. Podemos ver un incremento en el SWE de 1 % hasta una media de 9,9 % cuando comparamos los datos con antigüedad de 0 generaciones a los datos con antigüedad de 400 generaciones, ambos con alta profundidad (10×). La contaminación también tiene efectos importantes sobre el SWE, cuando consideramos datos faseados con antigüedad de 0 generaciones y alta profundidad, solamente la contaminación ocasiona un salto de 1 % hasta una media de 7 %. Por otro lado, incrementar la antigüedad puede causar un aumento en el número de sitios que son identificados como una base (T) pero en realidad podrían ser otra (C) a causa de la deaminación de citosinas[1], mientras que la contaminación agrega variantes (o haplotipos) que no pertenecen a los haplotipos antiguos que se intentan reconstruir.

En cuanto a la profundidad, podemos ver que el salto de 10× a una peor profundidad de 5× también resulta en una mayor tasa de error y desviación, incrementando el SWE un promedio de 3 %. Paradójicamente, parecería que las simulaciones con profundidad de 1× fueron las más precisas. Esto se debe a que los casos en que no había información recuperable para el faseo, es decir, no había sitios variantes para la inferencia de haplotipos, fueron contados con una tasa de error del 0 %. La figura 16b muestra la distribución de sitios variantes recuperados para cada individuo faseado. Como era esperado, la cantidad de información faseable es proporcional a la profundidad de secuenciación de los datos.

Es importante notar que al incrementar la contaminación, también incrementa la cantidad de sitios variantes recuperados. Esta observación también coincide con lo esperado, dado que al introducir contaminación de ADN moderno, se incorporan variantes (SNPs) modernas que no se encontraban en las poblaciones antiguas.

7.1.2. Simulaciones con divergencia entre poblaciones

En el siguiente conjunto de simulaciones se modelaron divergencias entre los individuos antiguos y los modernos utilizados para crear el panel de referencia, o el conjunto moderno en el caso del faseo poblacional. Una divergencia de poblaciones implica modelar dos poblaciones que, vistas desde el presente hacia el pasado, se vuelven más y más similares genéticamente hasta convertirse en la misma población (Figura 8).

Se consideraron tres diferentes momentos en el pasado para estas divergencias: 50, 100,

y 200 generaciones (Figuras 17, 18, y 19 respectivamente). En todos los casos, se puede observar un patrón común: la tasa de error incrementa a medida que los individuos se alejan temporalmente del momento de la divergencia. Es decir, individuos que son tanto menos como más antiguos al punto de divergencia. Estos valores mínimos de SWE se obtienen siempre en los casos de alta profundidad, con valores mínimos iguales al 1,9 %, 3,4 %, y 3,3 % respectivamente.

Este aumento en el error para individuos más antiguos que la divergencia es esperado, ya que equivale a las simulaciones donde no hubo una divergencia entre poblaciones (individuos con edades de 200 y 400 generaciones en la figura 16a). Por otro lado, el aumento del error conforme los individuos son más recientes se explica por una mayor distancia genética entre la población faseada y la población de referencia. Esto significa que incluso al fasear individuos modernos, si éstos pertenecen a una población que ha divergido de la población de referencia, aunque esta divergencia sea de solo 25 generaciones, se obtendrán resultados menos confiables que si utilizáramos haplotipos de la misma población como referencia.

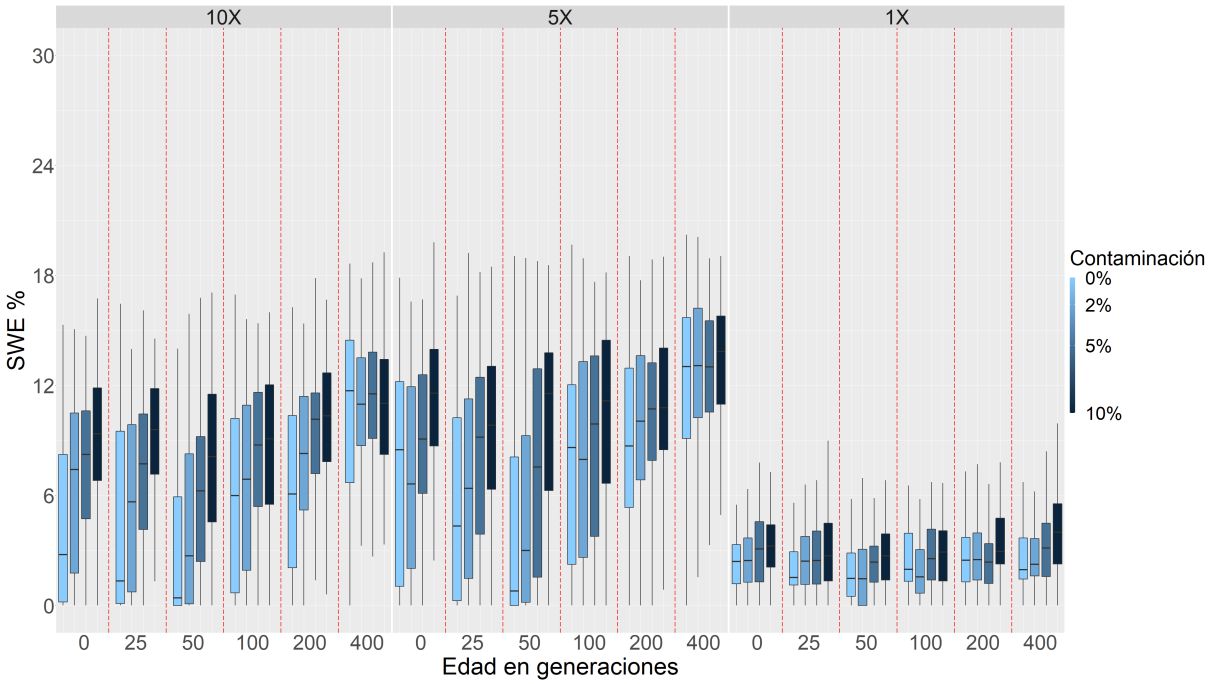


Figura 17: Puntuaciones de SWE para simulaciones con divergencia de poblaciones hace 50 generaciones. Faseo con panel de referencia. La tasa de error mínima corresponde a los individuos con edad de 50 generaciones, exactamente igual al tiempo de divergencia.

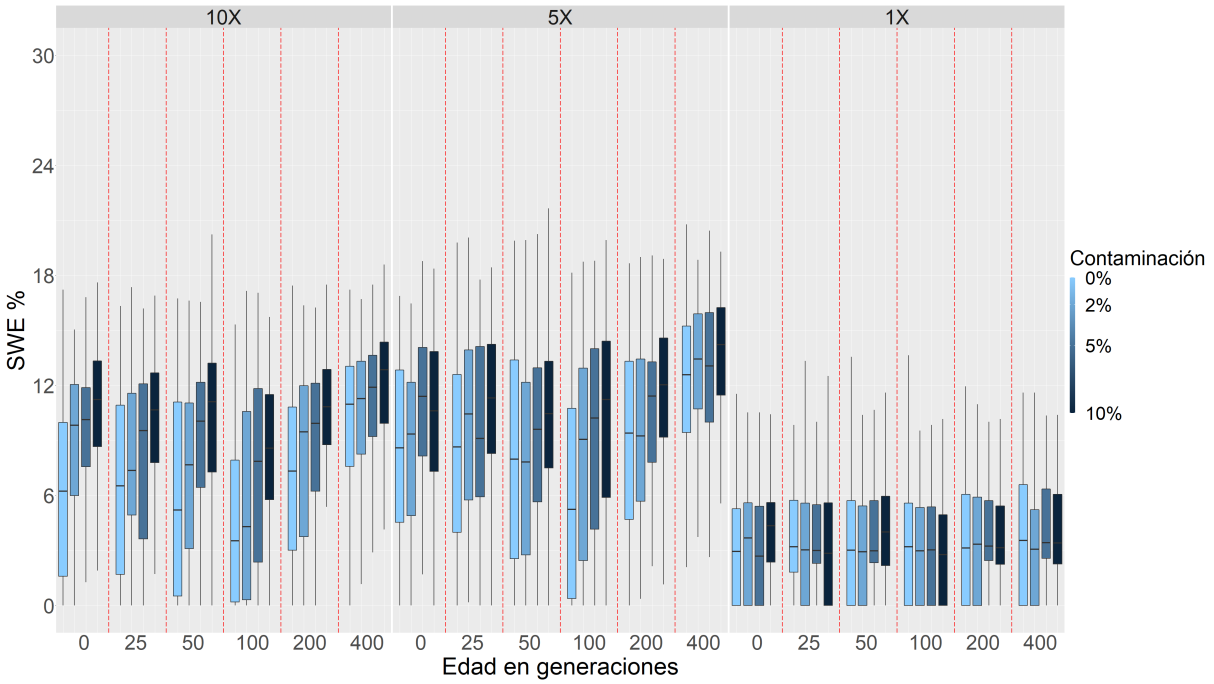


Figura 18: Puntuaciones de SWE para simulaciones con divergencia de poblaciones hace 100 generaciones. Faseo con panel de referencia. La tasa de error mínima corresponde a los individuos con una edad de 100 generaciones.

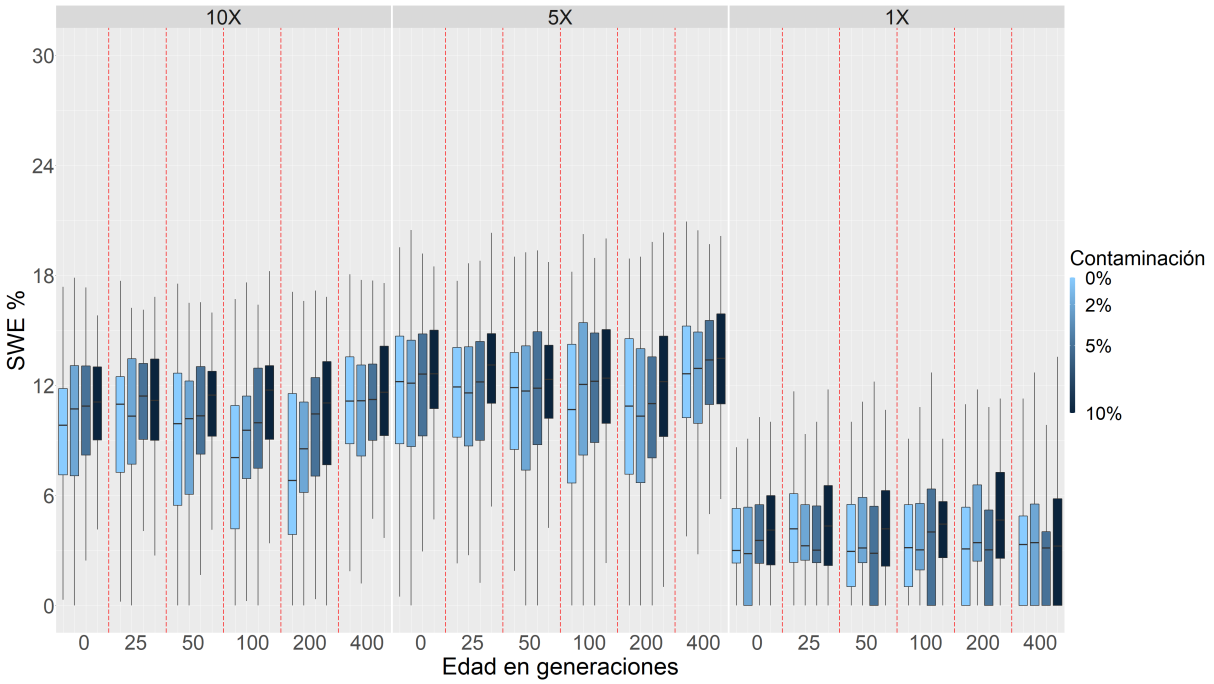


Figura 19: Puntuaciones de SWE para simulaciones con divergencia de poblaciones hace 200 generaciones. Faseo con panel de referencia. El error mínimo se alcanza en los individuos con edad de 200 generaciones. Aunque los individuos de 0 generaciones y 400 generaciones de edad tienen la misma distancia temporal al punto de divergencia (200 generaciones), el error para los individuos con 0 generaciones de edad decrementó 1,0 %.

7.1.3. Simulaciones con cuellos de botella

Se realizaron simulaciones bajo el escenario de un cuello de botella hace 25 generaciones (Figura 7). Un cuello de botella se simula como una población que sufre una reducción drástica y repentina en un solo momento en el pasado (Figura 7). El cuello de botella simulado corresponde a una reducción del 90 % al tamaño efectivo de la población ($N_e = 10,000 \rightarrow 1,000$). Recordemos que el tiempo y magnitud del cuello de botella coinciden con el importante decremento de la población indígena a raíz de la conquista y colonización de América hace 500 años (25 generaciones)[57].

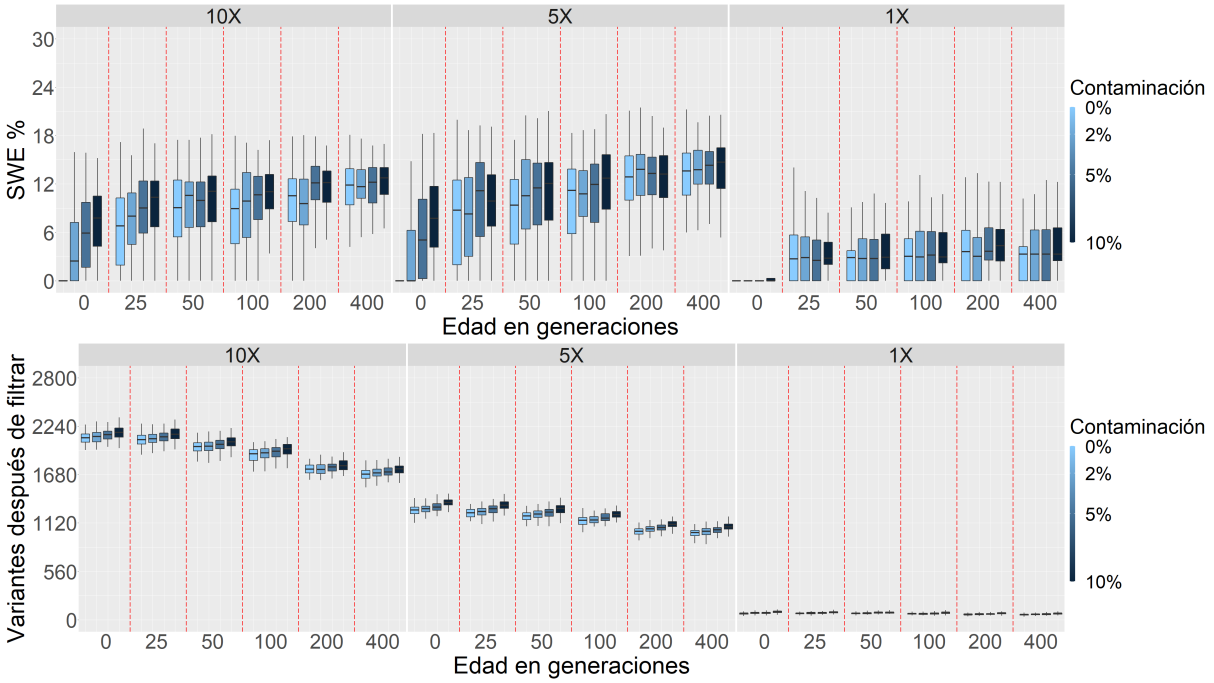


Figura 20: Puntuaciones de SWE para simulaciones con cuello de botella hace 25 generaciones. Faseo con panel de referencia. **(a)** La tasa de error de individuos posteriores al cuello de botella (individuos con 0 generaciones de antigüedad) obtienen tasas de error bajas. La calidad de faseo para individuos posteriores al cuello de botella muestra errores elevados al comparar con la figura 16a. **(b)** El número de variantes recuperadas en individuos previos al cuello de botella se ve reducido al comparar con la figura 16b.

Las distribuciones del error de faseo en estas simulaciones también difieren de las observadas para simulaciones de continuidad poblacional. Mientras que en las simulaciones de continuidad poblacional se observa un aumento gradual en la tasa de error conforme aumenta la antigüedad de los individuos faseados, bajo el modelo de cuello de botella se aprecia un incremento en SWE de 0,4 % a 10 % para los individuos faseados con una edad de 25 generaciones, es decir, previos al cuello de botella.

Este comportamiento se puede explicar por la reducción en la diversidad genética causada por el cuello de botella. El panel de referencia moderno no cuenta con información para sitios variantes o haplotipos enteros que fueron perdidos con la reducción en la población.

7.2. Faseo poblacional

Algunas de las simulaciones (Sección 5.3) fueron repetidas sin el uso de un panel de referencia, en su lugar, los individuos simulados fueron faseados como parte de un grupo compuesto por 1 individuo antiguo y 500 individuos modernos para los cuales sólo se contaba con información genotípica (Figuras 21, 22, y 23). El número de variantes no se muestra en estas figuras, ya que la cantidad de variantes esperadas es la misma que en la sección anterior.

Los resultados de las simulaciones de continuidad poblacional (Figura 21) son contraintuitivos, el SWE promedio al fasear individuos con antigüedad de 0 generaciones incrementó del 1 % al 6,4 % al comparar con la figura 6a. Al contrario, el SWE promedio al fasear individuos con antigüedad de 400 generaciones decreció del 9,9 % al 5,8 % comparando con la figura 6a. De forma inesperada, mientras que la contaminación tiene el mismo efecto sobre el SWE que al fasear con panel de referencia (aumentar la tasa de error), el salto de alta a mediana cobertura resultó una reducción promedio de 0,8 % al SWE.

En el caso de la divergencia de poblaciones hace 200 generaciones, se observa el mismo efecto que en los resultados del faseo con panel de referencia (Figura 22). Es decir, el valor mínimo de error corresponde a los individuos con la misma antigüedad al tiempo de divergencia, con un valor mínimo del 6 %. Hubo un incremento del 2,7 % al comparar los individuos faseados con panel de referencia bajo las mismas condiciones de simulación.

Para la simulación con un cuello de botella hace 25 generaciones (Figura 23), observamos de nuevo que el SWE mínimo de 11,9 % se alcanza al fasear individuos con 400 generaciones de antigüedad. Aún así, el desempeño del faseado fue mejor en casi todos los casos bajo el mismo escenario demográfico al utilizar el faseo con panel de referencia (Figura 20).

Otro cambio importante es la reducción en las varianzas de las distribuciones de SWE. Esto concuerda con lo esperado al utilizar el algoritmo de *SHAPEIT* (Sección 3.1.3). Al

utilizar el faseo poblacional, se incrementa la cantidad de haplotipos a fasear. Esto permite a *SHAPEIT* usar los 500 individuos modernos y 1 individuo ancestral para actualizar la segmentación de los haplotipos utilizada en la construcción de la gráfica S. Esto resulta en estimados más consistentes (aunque con un SWE generalmente elevado) a comparación de fasear solamente un individuo contra un panel de haplotipos conocidos[32].

Finalmente, otra desventaja importante, comparado con el faseo con panel de referencia, es el tiempo de ejecución. En el *cluster* de cómputo utilizado, fasear una secuencia de 2 Mbp toma alrededor de 3 horas con un panel de referencia conformado por $1,000$ haplotipos, y alrededor de 25 horas para el faseo poblacional. Esto se debe a que al fasear un individuo usando un panel de referencia, sólo se infiere un haplotipo contra cientos de haplotipos ya conocidos, por otro lado, el faso poblacional implica estimar haplotipos para 501 individuos (1 individuo antiguo y 500 de referencia).

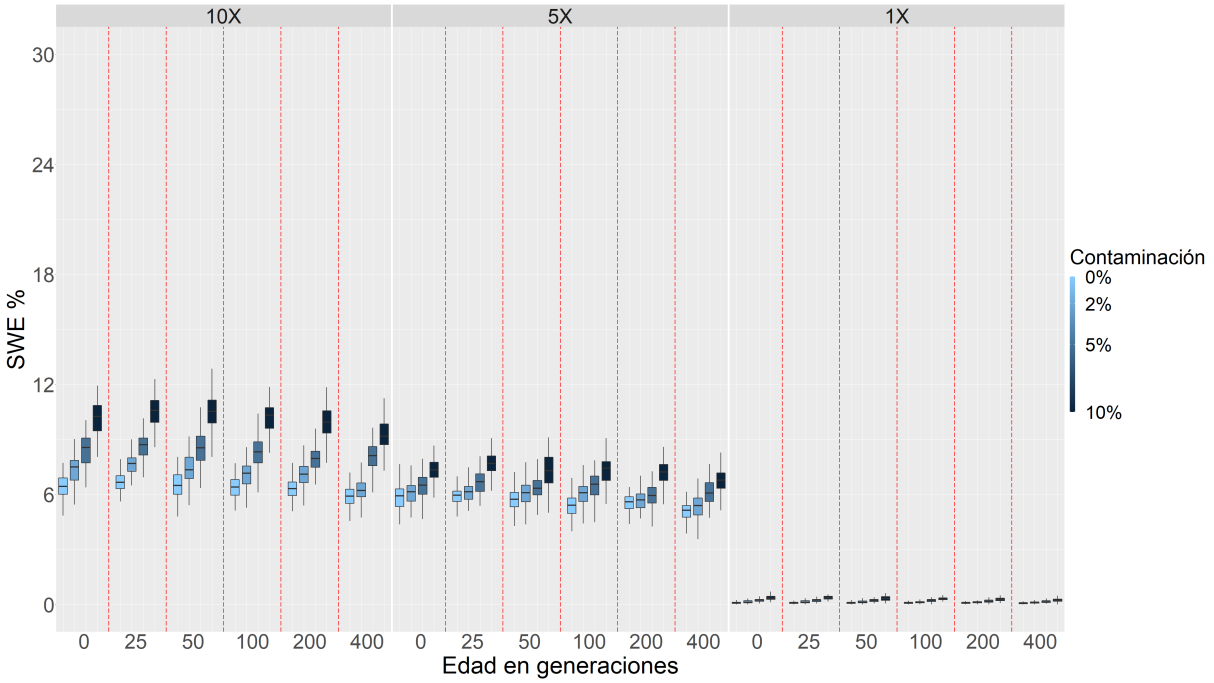


Figura 21: Puntuaciones de SWE junto con variantes recuperadas para simulaciones de continuidad poblacional. Se utilizó el método de faseo poblacional. En este caso las tasas de error mínimo corresponden a los individuos con mayor edad en generaciones (400 generaciones). Los resultados con cobertura de 5× muestran mejor calidad de faseo que los resultados con cobertura al 10×. La contaminación incrementa el error como es esperado.

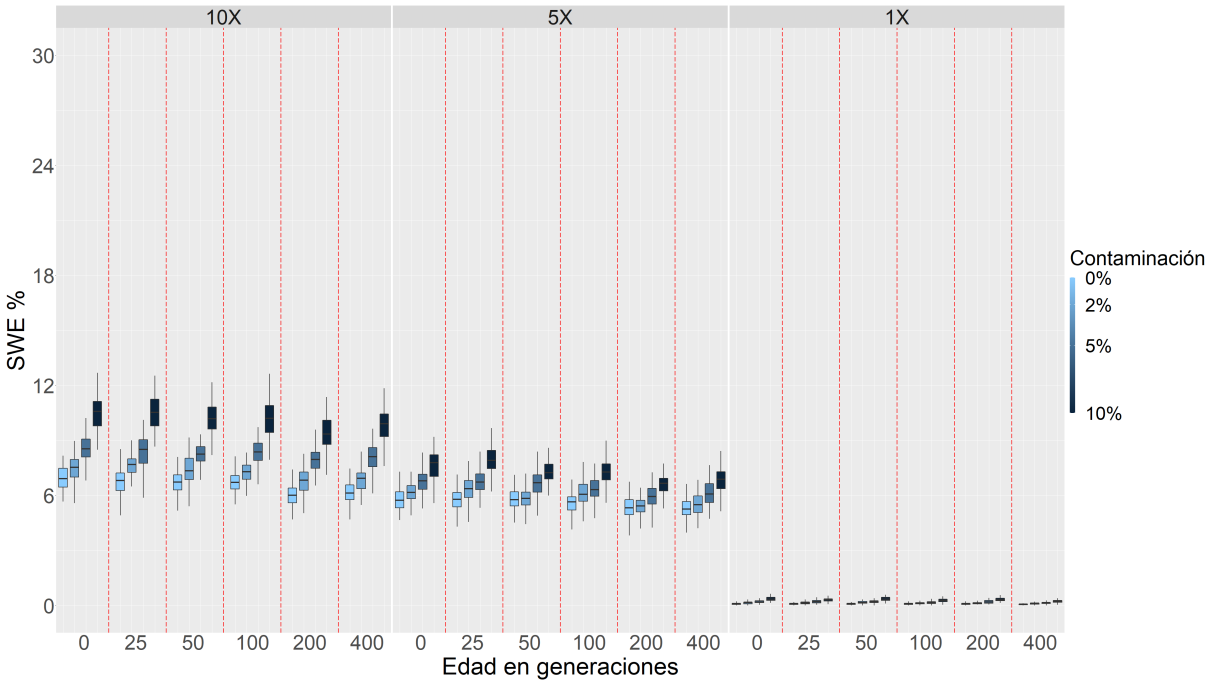


Figura 22: Puntuaciones de SWE para simulaciones con divergencia de poblaciones hace 200 generaciones. Se utilizó el método de faseo poblacional. Los individuos con edad de 400 generaciones presentan menores tasas de error que los individuos con 0 generaciones de edad. Esta vez, el error mínimo corresponde a los individuos con 200 generaciones de edad, estos individuos coinciden con el tiempo de la divergencia.

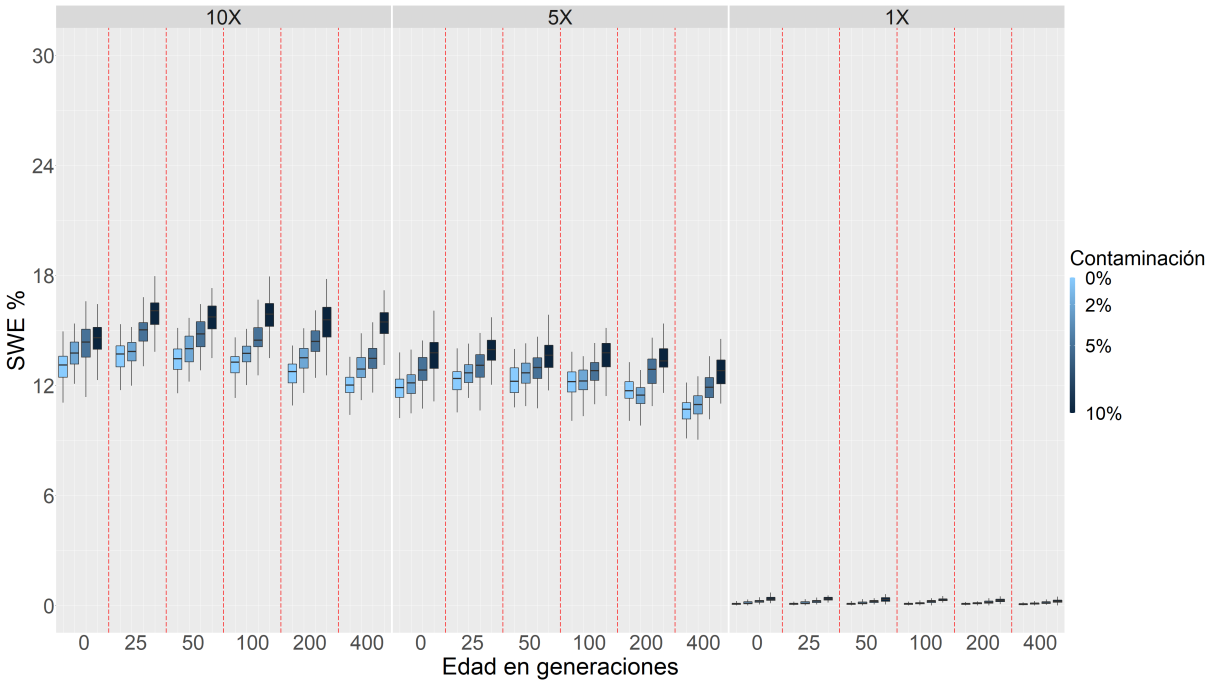


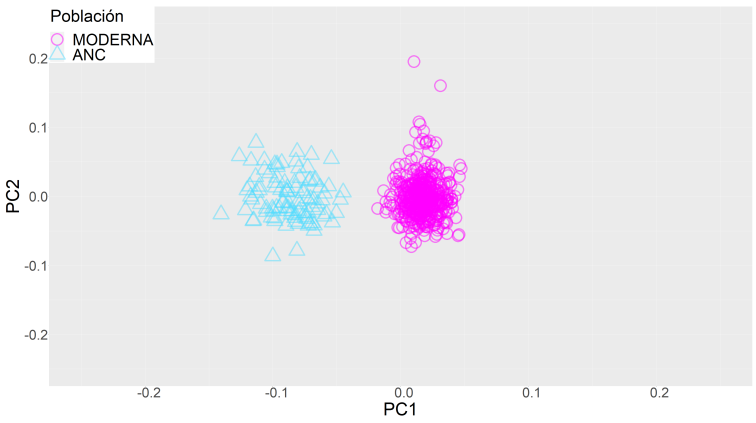
Figura 23: Puntuaciones de SWE para simulaciones con cuello de botella hace 25 generaciones. Se utilizó el método de faseo poblacional. Nuevamente los individuos más antiguos muestran mejor calidad de faseo que los individuos más recientes. Todas las tasas de error se elevan al comparar con la figura 20.

7.3. Análisis de Componentes Principales

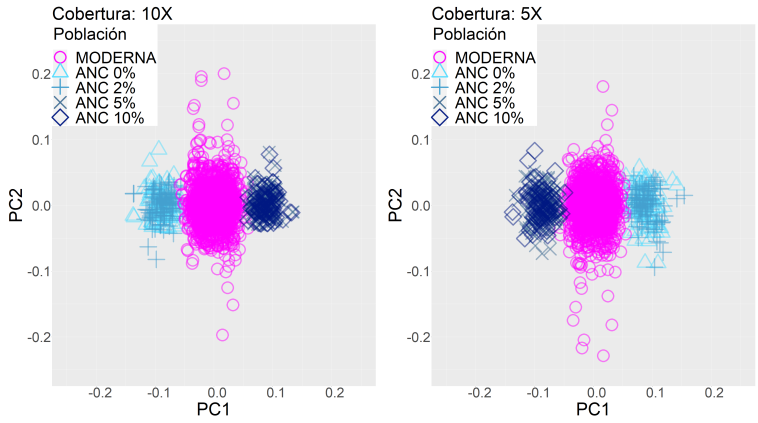
Aunque los resultados sobre la precisión y exactitud del faseo son ilustrativos de los efectos del tiempo, profundidad, contaminación, y eventos demográficos relevantes, es importante evaluar el efecto sobre los análisis estándar basados en haplotipos. Para ello se realizaron simulaciones con divergencia entre poblaciones hace 200 años, y análisis de componentes principales basados en cuatro diferentes tipos de datos resultantes de estas simulaciones (Sección 6.5).

7.3.1. Datos genotípicos

Las figuras 24 y 25 muestran los agrupamientos obtenidos con antigüedad en las poblaciones ancestrales de 0 (presente) y 25 generaciones. Al contrastar los paneles **(a)** y **(b)** de estas figuras, el agrupamiento en los datos «perfectos» se aprecia claramente, mientras que la resolución de los agrupamientos es menor en el caso de los datos con parámetros de calidad. Un efecto interesante de la contaminación en los paneles **(b)** de estas figuras es la aparición de un tercer agrupamiento para valores altos de contaminación ($\geq 5\,\%$).

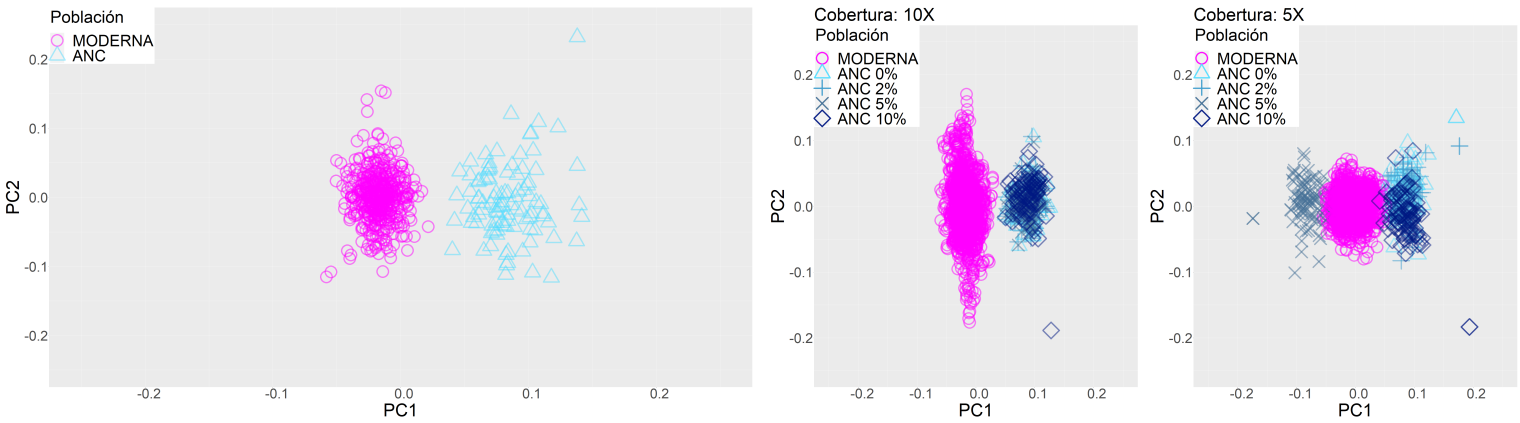


(a)



(b)

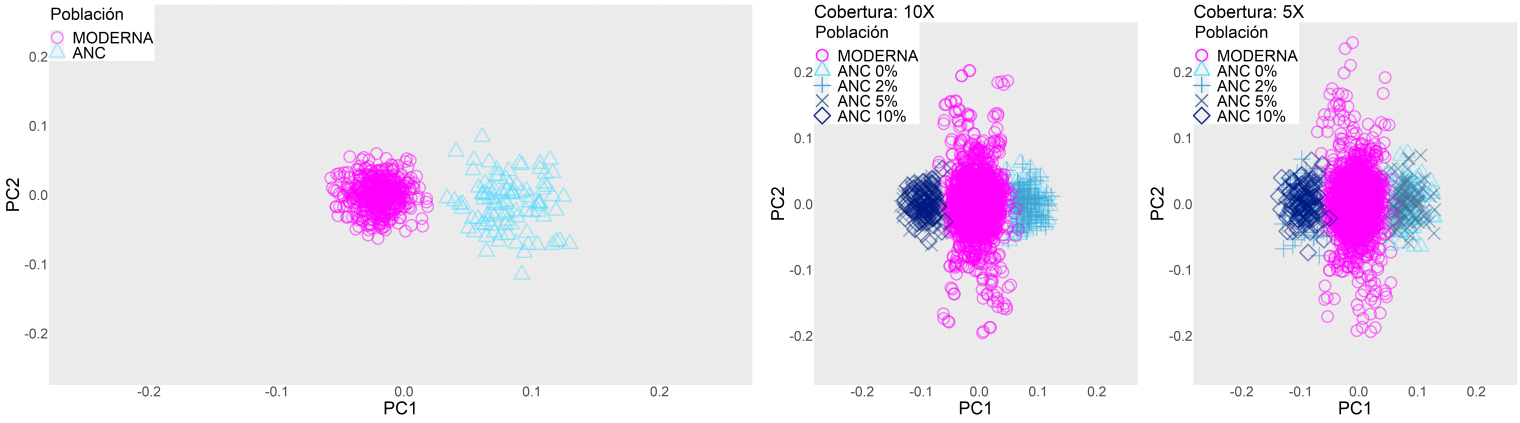
Figura 24: Agrupamientos con antigüedad de 0 generaciones para la población ancestral, divergencia de poblaciones hace 200 generaciones. **(a)** matriz genotípica perfecta y **(b)** genotipos con diferentes parámetros de calidad. Nótese que en el panel **(b)**, los individuos con alta contaminación forman un tercer agrupamiento.



(a) (b)

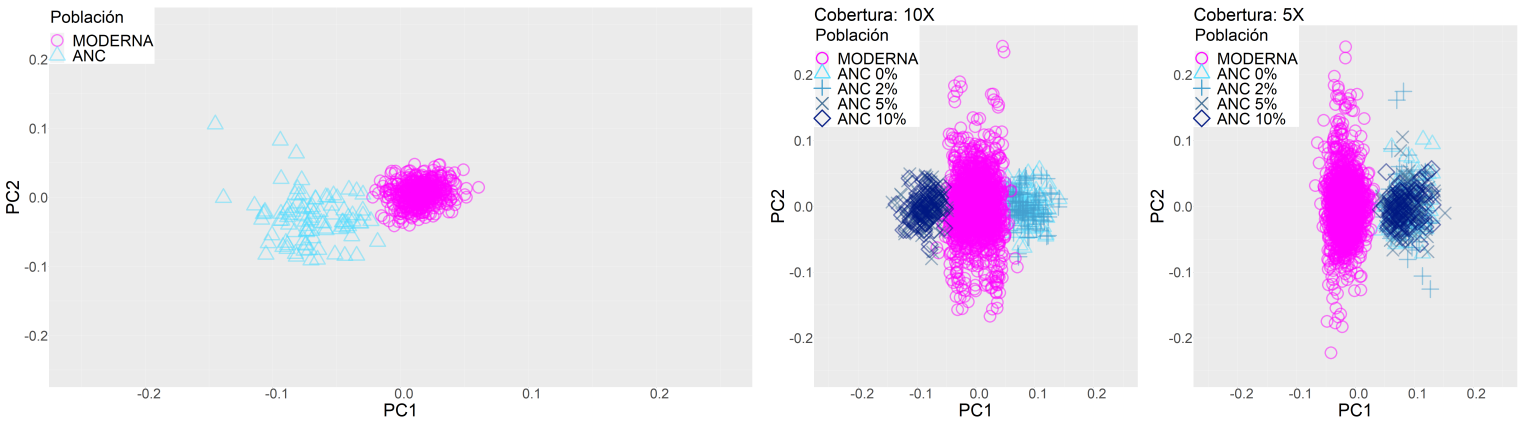
Figura 25: Agrupamientos con antigüedad de 25 generaciones para la población ancestral, divergencia de poblaciones hace 200 generaciones. **(a)** matriz genotípica perfecta y **(b)** genotipos con parámetros de calidad.

Las figuras 26 y 27 muestran los agrupamientos cuando la población ancestral tiene una antigüedad de 50 y 100 generaciones. En estos casos, tanto los agrupamientos perfectos como con parámetros de calidad muestran un menor distanciamiento al compararlos con las figuras 24 y 25. Los agrupamientos con datos genotípicos perfectos no muestran superposición, pero los agrupamientos con parámetros de calidad sí se sobreponen.



(a) (b)

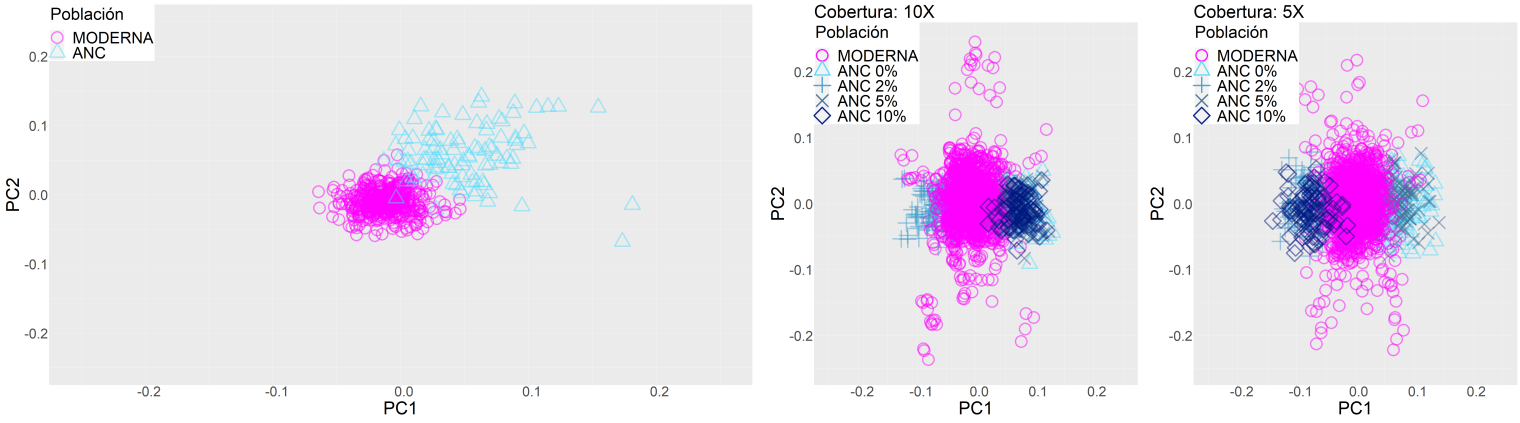
Figura 26: Agrupamientos con antigüedad de 50 generaciones para la población ancestral, divergencia de poblaciones hace 200 generaciones. **(a)** matriz genotípica perfecta y **(b)** genotipos con parámetros de calidad.



(a) (b)

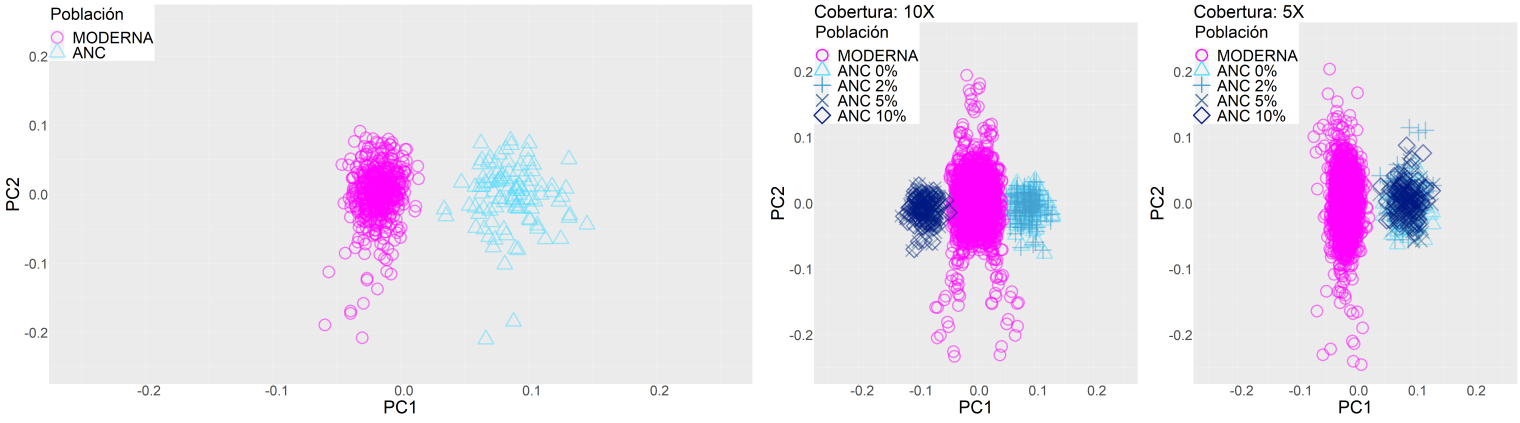
Figura 27: Agrupamientos con antigüedad de 100 generaciones para la población ancestral, divergencia de poblaciones hace 200 generaciones. **(a)** matriz genotípica perfecta y **(b)** genotipos con parámetros de calidad.

La antigüedad de los datos analizados parece ser el factor más importante en la estructura que obtenemos en el PCA. En la figura 28 se puede ver que los individuos que se encuentran más cercanos al tiempo de la divergencia (antigüedad de 200 generaciones) se agrupan más cercanamente con los individuos de referencia. Aunque los individuos antiguos con edad simulada de 400 generaciones (Figura 29) todavía no son afectados por esta divergencia, se agrupan como una población diferente, esto se puede explicar por la deriva génica entre los individuos contemporáneos y los individuos de hace 400 generaciones.



(a) (b)

Figura 28: Agrupamientos con antigüedad de 200 generaciones para la población ancestral, divergencia de poblaciones hace 200 generaciones. **(a)** matriz genotípica perfecta y **(b)** genotipos con parámetros de calidad.



(a) (b)

Figura 29: Agrupamientos con antigüedad de 400 generaciones para la población ancestral, divergencia de poblaciones hace 200 generaciones. **(a)** matriz genotípica perfecta y **(b)** genotipos con parámetros de calidad.

La introducción de contaminación moderna del 5 % o más resulta en agrupamientos que sugieren tres distintas poblaciones, es decir, se pueden apreciar 3 agrupamientos distintos en la gráfica: modernos, ancestrales de baja contaminación ($< 5\,\%$) y ancestrales de alta contaminación ($\geq 5\,\%$) (Figuras 24 a 29). Esto sugiere que la mezcla de haplotipos contaminantes modernos junto con los haplotipos ancestrales tiene un efecto lo suficientemente fuerte como para que se diferencien en una tercera población.

En las figuras 24 a 29, la diferencia en profundidades del 10× a 5× no tienen un gran efecto sobre la estructura inferida, es decir, la distancia promedio entre las agrupaciones no cambia considerablemente. Esto lo podemos medir con la distancia promedio entre los centroides de los agrupamientos de las poblaciones antiguas (ANC) y los individuos modernos. En la figura 30 podemos apreciar que la distancia promedio de los datos con profundidad al 10× y 5× es muy similar, y que ambas siguen la misma curva que la de los agrupamientos perfectos.

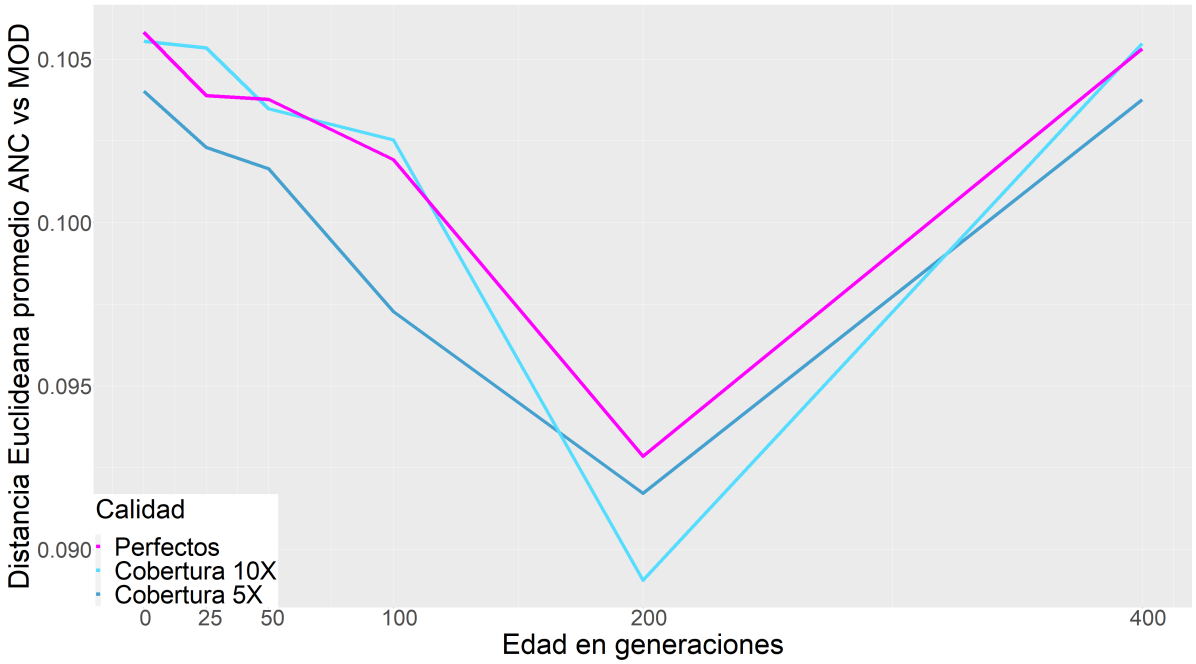
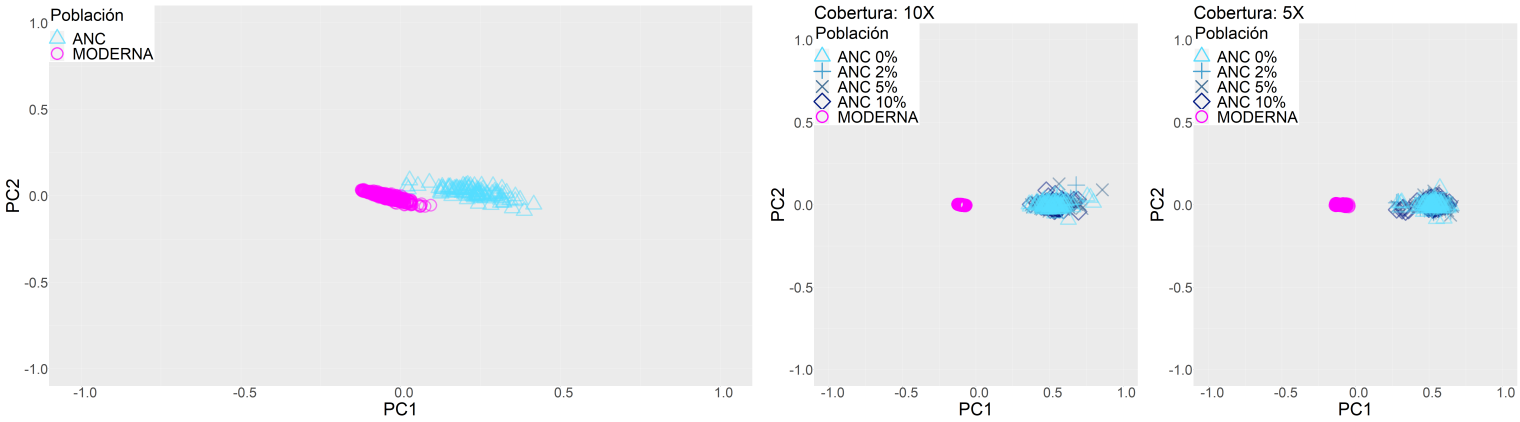


Figura 30: Distancia promedio entre agrupamientos ancestrales y modernos al utilizar datos genotípicos perfectos y con calidades parametrizadas. Tiempo de divergencia simulado hace 200 generaciones. Nótese cómo los agrupamientos se vuelven más cercanos a medida que la antigüedad de los individuos se acerca al tiempo de divergencia.

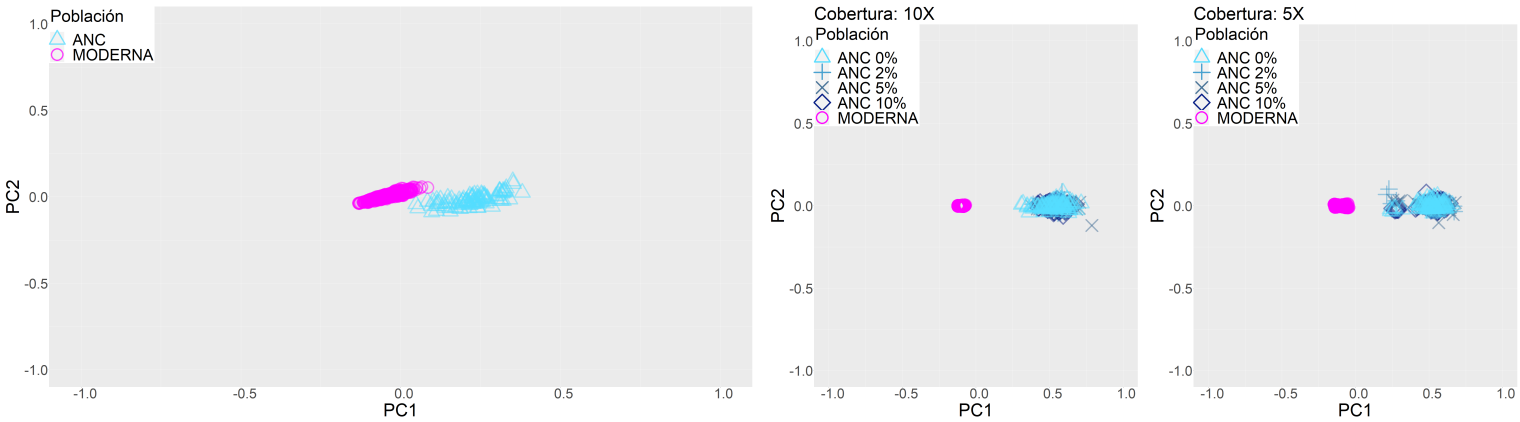
7.3.2. Datos haplotípicos

Las figuras 31 y 32 muestran los agrupamientos al utilizar datos haplotípicos con antigüedad de la población ancestral igual a 0 (presente) y 25 generaciones. En estas figuras, la estructura recuperada se aproxima mucho más a la estructura simulada que en los PCAs donde se utilizan sólo datos genotípicos (Figuras 24 a 29). Esto se puede apreciar en mayor medida comparando las figuras 30 y 37.



(a) (b)

Figura 31: Agrupamientos con antigüedad de 0 generaciones para la población ancestral, divergencia de poblaciones hace 200 generaciones. **(a)** haplotipos perfectos y **(b)** haplotipos inferidos con parámetros de calidad.

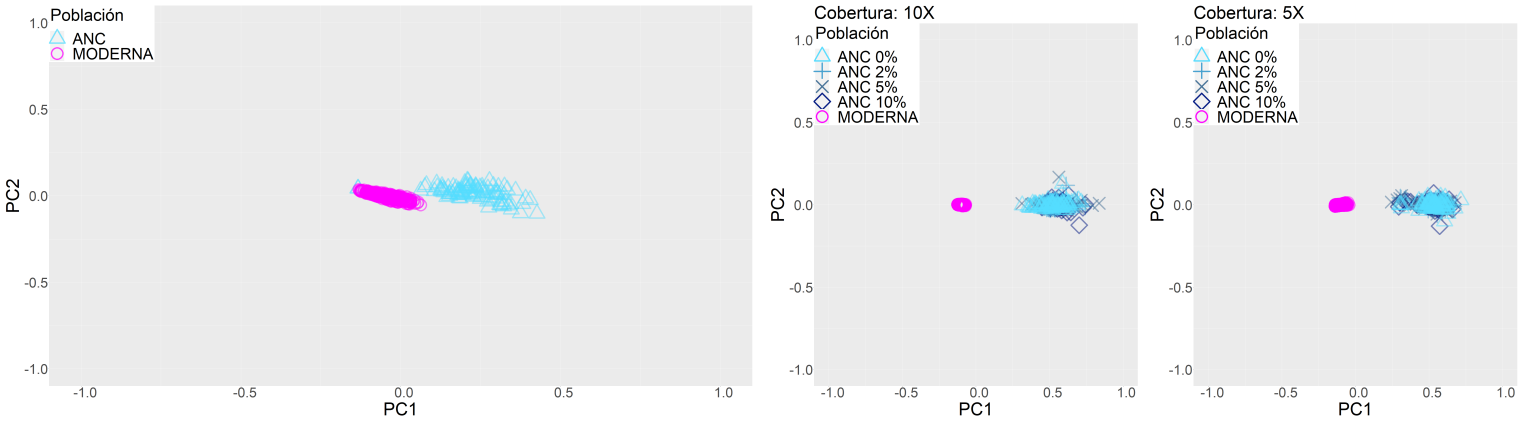


(a) (b)

Figura 32: Agrupamientos con antigüedad de 25 generaciones para la población ancestral, divergencia de poblaciones hace 200 generaciones. **(a)** haplotipos perfectos y **(b)** haplotipos inferidos con parámetros de calidad.

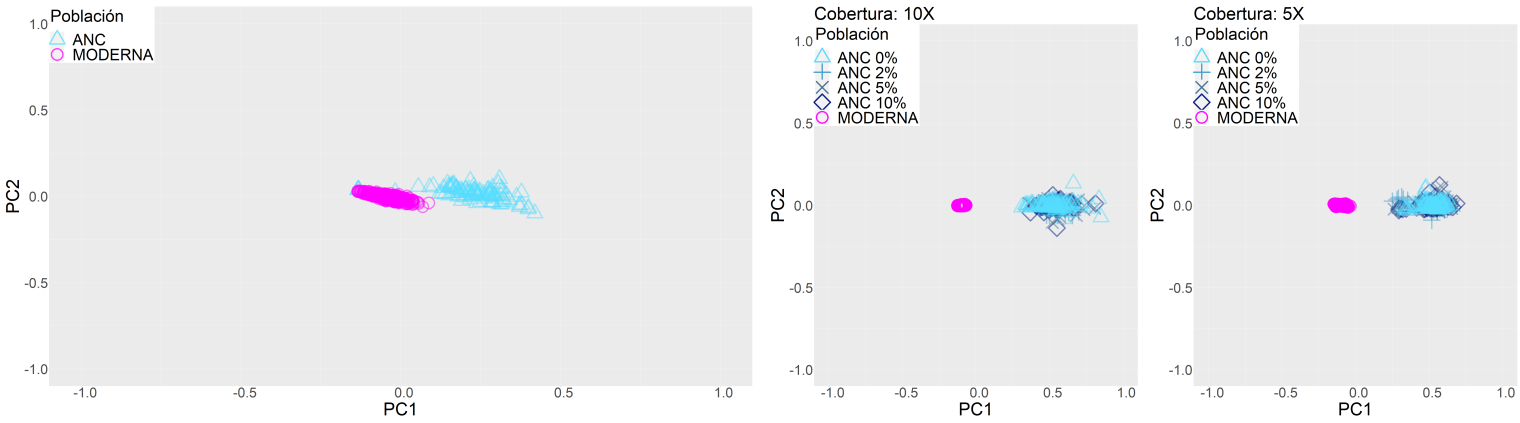
Las figuras 33 y 34 muestran un incremento a la antigüedad de las poblaciones ances-

trales: 50 y 100 generaciones. De nuevo, los agrupamientos se diferencian menos a medida que se acercan al tiempo de divergencia entre las poblaciones ancestrales y modernas.



(a) (b)

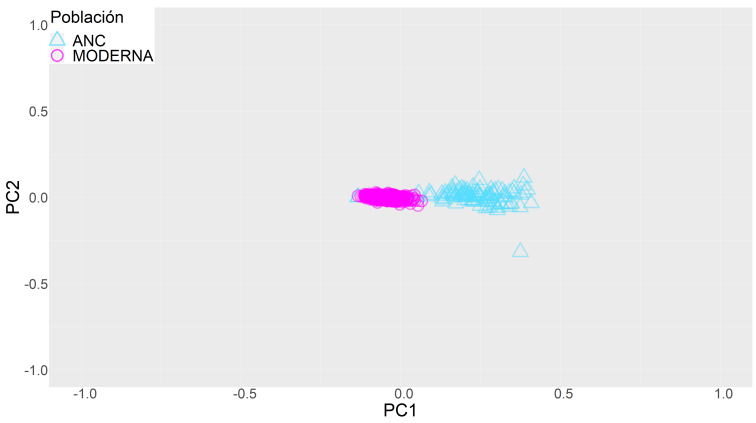
Figura 33: Agrupamientos con antigüedad de 50 generaciones para la población ancestral, divergencia de poblaciones hace 200 generaciones. **(a)** haplotipos perfectos y **(b)** haplotipos inferidos con parámetros de calidad.



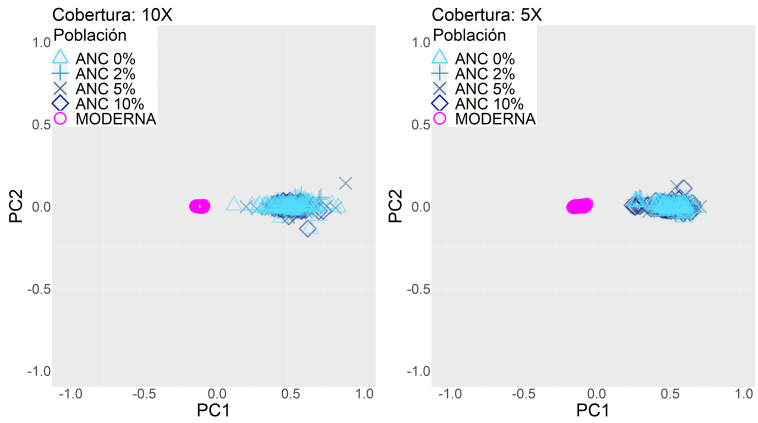
(a) (b)

Figura 34: Agrupamientos con antigüedad de 100 generaciones para la población ancestral, divergencia de poblaciones hace 200 generaciones. **(a)** haplotipos perfectos y **(b)** haplotipos inferidos con parámetros de calidad.

Finalmente, las figuras 35 y 36 muestran los agrupamientos para antigüedades de la población ancestral de 200 y 400 respectivamente. Nótese que en la figura 28, utilizando datos genotípicos y antigüedad de 200 generaciones, existía superposición de los agrupamientos incluso con datos perfectos. En el caso de la figura 35, esta superposición no existe aún utilizando los haplotipos inferidos después de aplicar los parámetros de calidad.

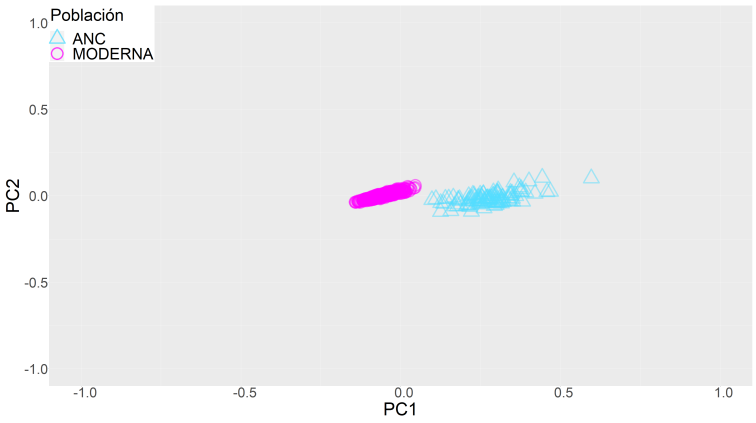


(a)

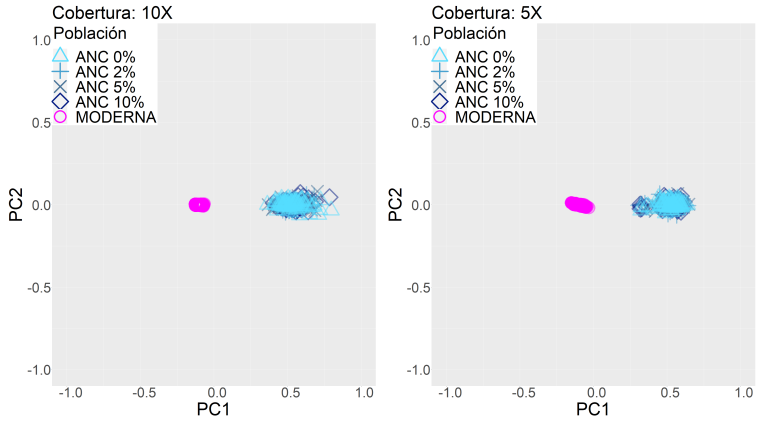


(b)

Figura 35: Agrupamientos con antigüedad de 200 generaciones para la población ancestral, divergencia de poblaciones hace 200 generaciones. **(a)** haplotipos perfectos y **(b)** haplotipos inferidos con parámetros de calidad.



(a)



(b)

Figura 36: Agrupamientos con antigüedad de 400 generaciones para la población ancestral, divergencia de poblaciones hace 200 generaciones. **(a)** haplotipos perfectos y **(b)** haplotipos inferidos con parámetros de calidad.

A diferencia de lo observado en los análisis de datos de genotipo (Figuras 24 a 29) al realizar PCA con datos haplotípicos la contaminación parece no tener un efecto tan drástico. Incluso con niveles altos ($> 5\,\%$) de contaminación, se reconoce que las muestras ancestrales contaminadas y no contaminadas pertenecen a la misma población, es decir, no se separan en dos agrupamientos basados en contaminación. Esto sugiere que la inferencia de estructura poblacional usando haplotipos es menos sensible a la contaminación moderna.

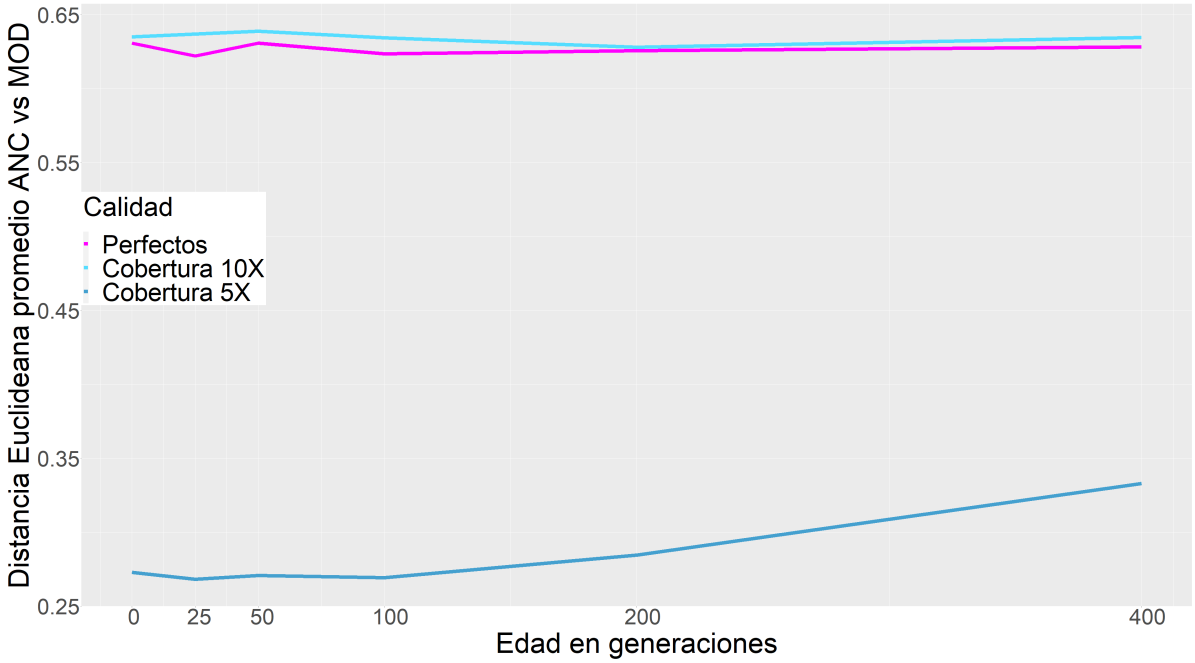


Figura 37: Distancia promedio entre agrupamientos ancestrales y modernos al utilizar datos haplotípicos. El efecto de la cobertura es mucho más marcado cuando comparamos con los agrupamientos de datos genotípicos (Figura 30).

Dado que *ChromoPainter* depende de la calidad de los haplotipos inferidos, la diferencia en resolución de poblaciones se beneficia de la mejor inferencia al utilizar datos con alta (10×) a comparación de media (5×) profundidad. Mientras que la distancia entre agrupamientos de profundidad alta se aproxima muy cercanamente a la distancia promedio de los agrupamientos perfectos, con un valor promedio de 0,62 unidades en una escala del −1,0 al 1,0, esta distancia promedio es mucho menor (es decir, las poblaciones se distinguen menos) cuando la profundidad promedio es del 5×, llegando sólo a 0,29 unidades en promedio.

8. Discusión y conclusiones

Los análisis realizados como parte de este trabajo han permitido revelar varios comportamientos interesantes del faseo aplicado a datos genómicos antiguos simulados. La calidad del faseo y posibles inferencias hechas a partir de los haplotipos recuperados pudieron ser medidas de manera concreta en relación a la calidad de los datos, historias demográficas, y métodos de faseo utilizados.

Las distribuciones de SWE obtenidas a partir del faseo de los individuos simulados nos permiten cuantificar la relación entre la calidad del faseo de individuos antiguos simulados, y diferentes parámetros de calidad como la cobertura, la contaminación, y el daño por antigüedad de los individuos. De igual manera, nos permiten comparar dos métodos de faseo dentro de la estrategia de faseo estadístico: el faseo con panel de referencia (Sección 7.1, y el faseo poblacional (Sección 7.2).

En cuanto al efecto de la contaminación sobre la calidad de faseo, todas las simulaciones (Figuras 16 a 23) revelaron que ésta tiene un efecto negativo. En todos los casos, fijando los demás parámetros (edad y cobertura), la tasa de error incrementó junto con la cantidad de contaminación moderna. Este comportamiento está presente tanto en los resultados de faseo con panel de referencia, como los resultados de faseo poblacional.

Esto nos hace pensar que el nivel de contaminación presente al fasear un individuo es una de las medidas más útiles para poner en cuestión la calidad de los haplotipos inferidos. Aún cuando los individuos faseados fueron simulados para tener poca antigüedad (25 generaciones o 500 años) respecto al individuo de contaminación, y que dicho individuo pertenecía a la misma población, la contaminación tiene un fuerte efecto sobre la tasa de SWE. Basta referirse a la figura 16a, donde consideramos datos faseados con antigüedad de 0 generaciones, alta profundidad, y contaminación de un individuo de la misma población que el individuo faseado. La contaminación por sí sola ocasiona un salto en la media del SWE de 1 % hasta una media de 7 %.Excluyendo la contaminación, los resultados de las dos diferentes estrategias de faseo muestran tendencias muy diferentes, como se discutirá más adelante.

En relación al parámetro de antigüedad de los individuos, la calidad del faseo con panel de referencia se ve negativamente afectada con el incremento en la antigüedad de los individuos faseados. Esto ocurre cuando no hay eventos demográficos o cuando se simula un cuello de botella (Figuras 16a y 20), pero no se observa de la misma manera

en las simulaciones con divergencia entre poblaciones. Inesperadamente, la tendencia es la opuesta cuando se utilizó el faseo poblacional; en este caso (Figuras 21 y 23), la tasa de error mínima se obtiene al considerar los individuos con mayor antigüedad. Esto va en contra de la intuición inicial, y debería ser explorado más a detalle considerando los efectos de otros parámetros, como el número de individuos de referencia o la longitud de las secuencias simuladas.

Cuando se simula una divergencia entre las poblaciones a fasear y de referencia (Figuras 17, 18, 19 y 22), el faseo poblacional y con panel de referencia comparten el mismo comportamiento respecto a la antigüedad de los individuos a fasear. La tasa de error promedio mínima corresponde siempre a los individuos con antigüedad igual al tiempo de la divergencia. Este comportamiento es lo esperado por dos razones. Los individuos con antigüedad menor al tiempo de divergencia pertenecen a otra población que se vuelve cada vez más diferente a la que conforma la población de referencia. En el caso opuesto, los individuos con antigüedad mayor al tiempo de divergencia pertenecen a la misma población que la población de referencia, esto es equivalente a las simulaciones de continuidad poblacional, por lo que es de esperar que mayor antigüedad resulta en mayor tasa de error. Esto tiene la importante implicación de que para obtener el mejor faseo posible de individuos antiguos, es necesario considerar también a los individuos más cercanos genéticamente de los que se puedan extraer datos genómicos de alta calidad para usar como referencia.

Otra implicación importante que tienen nuestras observaciones en relación a las características del panel de referencia es que cuando dicho panel contiene poblaciones lejanas, el faseo puede presentar tasas de error altas. Esto es de considerar pues, aunque el panel de haplotipos del proyecto de los *1,000 Genomas* incluye decenas de poblaciones, su uso puede resultar en tasas de error inesperadamente altas al ser usado para fasear individuos antiguos o modernos que no son cercanos genéticamente a las poblaciones que contiene. De igual manera, si se planea usar el faseo poblacional, debe considerarse la distancia genética entre los individuos modernos utilizados.

En cuanto a la cobertura, observamos un comportamiento inesperado. Mientras que los mejores resultados al fasear con panel de referencia se obtuvieron con la más alta cobertura (10×), los resultados con faseo poblacional fueron mejores en las simulaciones con cobertura media (5×). Esto también debe ser investigado más a detalle.

La simulación de diferentes modelos demográficos como un cuello de botella y una divergencia poblacional, permitió evaluar el efecto de éstos bajo diferentes parámetros.

Las simulaciones con de cuellos de botella (Figuras 20 y 23) tuvieron tasas de error elevadas tanto para el faseo con panel de referencia como en el faseo poblacional. Para el faseo poblacional, todas las medias de tasa de error sobrepasaron el 12 %, la calidad de faseo fue mejor en todos los casos para el faseo con panel de referencia. Sin embargo, incluso dentro de la categoría de simulaciones con panel de referencia, las tasas de error fueron elevadas a comparación de la simulación de continuidad poblacional cuando los individuos tenían una antigüedad igual o mayor al tiempo del cuello de botella (Figura 16a). Esto es consistente con lo esperado, dada la pérdida de variación genética y haplotípica que implica un cuello de botella. Podemos decir que la inferencia de haplotipos para individuos previos a un cuello de botella siempre presenta problemas de calidad, independientemente de los parámetros de las simulaciones.

Cuando las simulaciones incluyen divergencia entre poblaciones (Figuras 17, 18, 19 y 22), el faseo poblacional obtuvo tasas de error menores a las del faseo por panel de referencia. En promedio, el SWE mejoró un 5 % al usar el faseo poblacional. Por otro lado, la calidad del faseo por panel de referencia fue mejor en los casos de continuidad poblacional y con cuellos de botella. Esto sugiere que la distancia genética entre poblaciones es un factor a considerar al decidir qué método de faseo se empleará.

Respecto al tiempo de ejecución, existe una ventaja definitiva del faseo con panel de referencia sobre el faseo poblacional. El faseo poblacional aumenta tanto el tiempo de ejecución, que no fue factible utilizarlo para secuencias más largas (20 Mbp). Una ventaja aparente del faseo poblacional es una menor varianza en las distribuciones de SWE, aunque en general este sigue siendo más alto que en el faseo por panel de referencia, excepto en el escenario demográfico de divergencia poblacional.

En cuanto a los resultados de análisis de componentes principales (Sección 7.3), contar con datos haplotípicos (Figuras 31 a 36) siempre resultó en una mejor separación de poblaciones que al considerar sólo datos genotípicos (Figuras 24 a 29). Incluso las agrupaciones de datos haplotípicos con contaminación y baja cobertura, fueron mejores que aquéllas obtenidas a partir de datos genotípicos perfectos (Figuras 35b y 28a).

La cantidad de contaminación simulada tuvo diferentes efectos en los agrupamientos haplotípicos y genotípicos. Mientras que los datos genotípicos con alta contaminación

llegaron a sugerir una tercera población diferente a las poblaciones ancestrales y modernas (Figuras 24b a 29b), esto no ocurrió cuando se utilizaron datos haplotípicos con los mismos parámetros de calidad (Figuras 31b a 36b). Esto sugiere que una ventaja al usar datos haplotípicos en vez de genotípicos es una mayor «tolerancia» a altos niveles de contaminación.

En cuanto a cobertura, tanto los datos haplotípicos como genotípicos mostraron agrupamientos menos distintos al simular cobertura media (5×) comparados con los agrupamientos observados con cobertura alta (10×). Los datos genotípicos llegaron a mostrar superposición entre agrupamientos (Figuras 26 a 28), pero esto no ocurrió en las gráficas de datos haplotípicos.

Aún si el faseo tiene una tasa de error relativamente alta, por ejemplo del 13 % para los individuos con alta contaminación y cobertura media (Figura 19), la información haplotípica es ventajosa para los análisis de agrupamiento de las poblaciones simuladas; incluso superando los agrupamientos que se obtienen al utilizar datos genotípicos de calidad perfecta.

En resumen, en esta tesis se estudiaron varios aspectos del faseo de ADNa que previamente no habían sido considerados formalmente o a detalle. Aún así, siempre quedan más parámetros de calidad, estructuras demográficas, o propiedades de los datos simulados a analizar en un futuro. Se espera que los resultados presentados sean un buen punto de partida para cualquier estudio que necesite inferir haplotipos a partir de datos paleogenómicos, y que los resultados de estas inferencias sean más predecibles cuando se cuente con información sobre la historia y calidad de los datos faseados. De igual manera, se considera que este trabajo es un buen punto de partida para otros trabajos que midan la precisión de la inferencia de haplotipos sobre datos de ADNa simulados, incorporando diferentes herramientas, modelos poblacionales, longitudes de regiones simuladas, etc.

Referencias

[1] Jesse Dabney, Matthias Meyer y Svante Pääbo. «Ancient DNA Damage». En: *Cold Spring Harbor Perspectives in Biology* 5.7 (2013).

[2] Hákon Jónsson y col. «mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters». En: *Bioinformatics* 29.13 (2013), págs. 1682-1684.

[3] Kay Prüfer y col. «Computational challenges in the analysis of ancient DNA». En: *Genome Biology* 11.5 (2010).

[4] Joachim Burger y col. «DNA preservation: A microsatellite-DNA study on ancient skeletal remains». En: *ELECTROPHORESIS* 20.8 (1999), págs. 1722-1728.

[5] Michael Knapp y Michael Hofreiter. «Next Generation Sequencing of Ancient DNA: Requirements, Strategies and Perspectives». En: *Genes* 1.2 (2010), págs. 227-243.

[6] Ron Pinhasi y col. «Optimal Ancient DNA Yields from the Inner Ear Part of the Human Petrous Bone». En: *PloS one* 10.6 (2015).

[7] Caroline Pont, Stefanie Wagner, Antoine Kremer y col. «Paleogenomics: reconstruction of plant evolutionary trajectories from modern and ancient DNA». En: *Genome Biology* 20.19 (2019).

[8] Beth Shapiro y Michael Hofreiter. «A Paleogenomic Perspective on Evolution and Gene Function: New Insights from Ancient DNA». En: *Science* 343.6169 (2014).

[9] Pontus Skoglund y Iain Mathieson. «Ancient Genomics of Modern Humans: The First Decade». En: *Annual Review of Genomics and Human Genetics* 19.1 (2018), págs. 381-404.

[10] Bastien Llamas y col. «Ancient mitochondrial DNA provides high-resolution time scale of the peopling of the Americas». En: *Science Advances* 2.4 (2016).

[11] Maria A. Spyrou, Kirsten I. Bos, Alexander Herbig y col. «Ancient pathogen genomics as an emerging tool for infectious disease research». En: *Nature Reviews Genetics* 20 (2019), págs. 323-340.

[12] Rasmus Nielsen y col. «Tracing the peopling of the world through genomics». En: *Nature* 541.7637 (2017), págs. 302-310.

[13] Fernando Racimo y col. «Evidence for archaic adaptive introgression in humans». En: *Nature Reviews Genetics* 16.6 (2015).

[14] Pontus Skoglund y col. «Genomic insights into the peopling of the Southwest Pacific». En: *Nature* 538.7624 (2016), págs. 510-513.

[15] Stephanie Marciniak y George H. Perry. «Harnessing ancient genomes to study the history of human adaptation». En: *Nature Reviews Genetics* 18 (2017), págs. 659-674.

[16] Peggy Hsieh y col. «Model-based analyses of whole-genome data reveal a complex evolutionary history involving archaic introgression in central African pygmies». En: *Genome Research* 26 (2016), págs. 291-300.

[17] Emilia Huerta-Sánchez, Xin Jin y col. «Neolithic and medieval virus genomes reveal complex evolution of hepatitis B». En: *Nature* 512 (2014), págs. 194-197.

[18] Ben Krause-Kyora y col. «Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA». En: *eLife* 7 (2018).

[19] Rui Martiniano y col. «Genomic signals of migration and continuity in Britain before the Anglo-Saxons». En: *Nature Communications* (2016).

[20] Saioa López, Mark G. Thomas y col. «The Genetic Legacy of Zoroastrianism in Iran and India: Insights into Population Structure, Gene Flow, and Selection». En: *The American Journal of Human Genetics* (2017).

[21] Cristina Gamba y col. «Genome flux and stasis in a five millennium transect of European prehistory». En: *Nature Communications* (2014).

[22] Carlos Eduardo G. Amorim y col. «Understanding 6th-century barbarian social organization and migration through paleogenomics». En: *Nature Communications* (2018).

[23] Stephan Schiffels y Richard Durbin. «Inferring human population size and separation history from multiple genome sequences». En: *Nature Genetics* 46 (2014).

[24] Pardis C. Sabeti, David E. Reich, John M. Higgins y col. «Detecting recent positive selection in the human genome from haplotype structure». En: *Nature* 419 (2002).

[25] Sara Goodwin, John D. McPherson y W. Richard McCombie. «Coming of age: ten years of next-generation sequencing technologies». En: *Nature Reviews Genetics* 17 (2016), págs. 333-351.

[26] Sina Majidian, Mohammad Hossein Kahaei y Dick de Ridder. «Minimum error correction-based haplotype assembly: Considerations for long read data». En: *PloS one* (2020).

[27] Yongwook Choi y col. «Comparison of phasing strategies for whole human genomes». En: *PloS genetics* 14.4 (2018).

[28] Sharon R. Browning y Brian L. Browning. «Haplotype phasing: Existing methods and new developments». En: *Nature Reviews Genetics* 12 (2011), págs. 703-714.

[29] The 1000 Genomes Project Consortium. «A global reference for human genetic variation». En: *Nature* 526 (2015).

[30] the Haplotype Reference Consortium. «A reference panel of 64,976 haplotypes for genotype imputation». En: *Nature Genetics* 48 (2016).

[31] Sayantan Das y col. «Next-generation genotype imputation service and methods». En: *Nature Genetics* 48 (2016), págs. 1284-1287.

[32] Olivier Delaneau, Jonathan Marchini y Jean-François Zagury. «A linear complexity phasing method for thousands of genomes». En: *Nature Methods* 9 (2012), págs. 179-181.

[33] Sharon R. Browning y Brian L. Browning. «Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies By Use of Localized Haplotype Clustering». En: *American journal of human genetics* 81 (2007).

[34] Ewan Birney y Nicole Soranzo. «The end of the start for population sequencing». En: *Nature* 526 (2015).

[35] *Considerations for RNA-Seq read length and coverage*. URL: https://support.illumina.com/bulletins/2017/04/considerations-for-rna-seq-read-length-and-coverage-.html.

[36] David Sims y col. «Sequencing depth and coverage: key considerations in genomic analyses». En: *Nature Reviews Genetics* 15 (2014), págs. 121-132.

[37] Marcel Martin y col. «WhatsHap: fast and accurate read-based phasing». En: *bioRxiv* (2016).

[38] Daniel John Lawson y col. «Inference of Population Structure using Dense Haplotype Data». En: *PLoS genetics* 8 (1 2012).

[39] Mateusz Baca y col. «Ancient DNA reveals kinship burial patterns of a pre-Columbian Andean community». En: *BMC Genetics* 13 (2012).

[40] *Paired-End vs. Single-Read Sequencing Technology*. URL: https://www.illumina.com/science/technology/next-generation-sequencing/plan-experiments/paired-end-vs-single-read.html.

[41] Peter Edge, Vineet Bafna y Vikas Bansal. «HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies». En: *Genome Research* 27.5 (2017).

[42] Ryan Tewhey y col. «The importance of phase information for human genomics». En: *Nature Reviews Genetics* 12.3 (2011).

[43] Po-Ru Loh y col. «Reference-based phasing using the Haplotype Reference Consortium panel». En: *Nature Genetics* 48 (2016).

[44] Olivier Delaneau y col. «Accurate, scalable and integrative haplotype estimation». En: *Nature Communications* 10 (2019).

[45] John Frank Charles Kingman. «The coalescent». En: *Stochastic Processes and their Applications* 13 (3 1982), págs. 235-248.

[46] Fumio Tajima. «Evolutionary relationship of DNA sequences in finite populations». En: *Genetics* 105 (1983).

[47] Richard R. Hudson. «Properties of a neutral allele model with intragenic recombination». En: *Theoretical Population Biology* 23 (1983).

[48] Andrew Rambaut y Nicholas C Grassly. «Seq-Gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees». En: *Computer Applications in the Biosciences* 13 (1993), págs. 235-238.

[49] Masami Hasegawa, Hirohisa Kishino y Taka-aki Yano. «Dating of the Human-Ape Splitting by a Molecular Clock of Mitochondrial DNA». En: *Journal of Molecular Evolution* 22 (oct. de 1985), págs. 160-174. DOI: 10.1007/BF02101694.

[50] John Novembre y Benjamin M. Peter. «Recent advances in the study of fine-scale population structure in humans». En: *Current opinion in genetics & development* 41 (2016), págs. 98-105.

[51] Caroline M. Nievergelt y col. «Generalized Analysis of Molecular Variance». En: *PLoS genetics* 3 (4 2007).

[52] Petr Danecek y col. «The variant call format and VCFtools». En: *Bioinformatics* 27 (15 2011), págs. 2156-2158.

[53] Xiuwen Zheng y col. «A high-performance computing toolset for relatedness and principal component analysis of SNP data». En: *Bioinformatics* 28 (24 2012), págs. 3326-3328.

[54] Daniel Lawson. *Summary of tools for data preparation*. 2012. URL: https://people.maths.bris.ac.uk/~madjl/finestructure/toolsummary.html.

[55] Per Hage y Jeff Mark. «Matrilineality and the Melanesian Origin of Polynesian Y Chromosomes». En: *Current Anthropology* 44 (2003).

[56] Guy Halsall. *Barbarian Migrations and the Roman West, 376–568*. 2007.

[57] Rodolfo Acuna-Soto y col. «Megadrought and Megadeath in 16th Century Mexico». En: *Emerging infectious diseases* 8 (4 2002), págs. 360-362.

[58] *Implementación del* pipeline *descrito*. URL: https://github.com/Jazpy/Paleogenomic-Datasim.

[59] Mark Lipson y col. «Calibrating the Human Mutation Rate via Ancestral Recombination Density in Diploid Genomes». En: *PloS Genetics* 11 (11 2015).

[60] Laure Ségurel y col. «Determinants of Mutation Rate Variation in the Human Germline». En: *Annual Review of Genomics and Human Genetics* 15 (2014).

[61] Jerome Kelleher, Alison M Etheridge y Gilean McVean. «Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes». En: *PLoS Computational Biology* 12.5 (mayo de 2016), págs. 1-22.

[62] Richard R. Hudson. «Generating samples under a Wright-Fisher neutral model of genetic variation». En: *Bioinformatics* 18 (2002).

[63] David R. Hanson. «Fast allocation and deallocation of memory based on object lifetimes». En: *Software: Practice and Experience* 20.1 (1989).

[64] *Pull request para Seq-Gen*. URL: https://github.com/rambaut/Seq-Gen/pull/13.

[65] *HAP / LEGEND / SAMPLE file format for haplotype reference panels*. URL: https://mathgen.stats.ox.ac.uk/genetics_software/shapeit/shapeit.html#haplegsample.

[66] Gabriel Renaud y col. «gargammel: a sequence simulator for ancient DNA». En: *Bioinformatics* 33.4 (nov. de 2016), págs. 577-579. ISSN: 1367-4803.

[67] Huang Weichun y col. «ART: a next-generation sequencing read simulator». En: *Bioinformatics* 28.4 (dic. de 2011), págs. 593-594.

[68] *Pull request para Gargammel*. URL: https://github.com/grenaud/gargammel/pull/5.

[69] Heng Li y Richard Durbin. «Fast and accurate short read alignment with Burrows-Wheeler Transform». En: *Bioinformatics* 25 (2009), págs. 1754-1760.

[70] Felix Krueger. *Trim Galore!* URL: http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/.

[71] Heng Li y col. «The Sequence Alignment/Map format and SAMtools». En: *Bioinformatics* 25.16 (2009), págs. 2078-2079.

[72] Genome Research Ltd. *bcftools call*. URL: http://samtools.github.io/bcftools/bcftools.html#call.

[73] GATK Team. *Phred-scaled quality scores*. URL: https://gatk.broadinstitute.org/hc/en-us/articles/360035531872-Phred-scaled-quality-scores.

[74] Gil McVean. «A Genealogical Interpretation of Principal Components Analysis». En: *PLoS genetics* 5 (2009).

[75] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN: 978-3-319-24277-4. URL: https://ggplot2.tidyverse.org.

[76] Hadley Wickham. «The Split-Apply-Combine Strategy for Data Analysis». En: *Journal of Statistical Software* 40.1 (2011), págs. 1-29. URL: http://www.jstatsoft.org/v40/i01/.

[77] Baptiste Auguie. *gridExtra: Miscellaneous Functions for "Grid" Graphics*. R package version 2.3. 2017. URL: https://CRAN.R-project.org/package=gridExtra.

[78] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2019. URL: https://www.R-project.org/.