Detección de selección en poblaciones mezcladas recientemente

Gaston Rijo

Tesis Maestría en Bioinformática PEDECIBA Universidad de la República
Tutora: María Inés Fariello Cotutor: Bertrand Servin Cotutor: Hugo Naya

Junio 2021

Índice general

Prefacio

No se me va a hacer tan difícil como a Hegel hacer un prefacio. Pero bueno, yo no soy Hegel y esto no es filosofía. Usualmente los prefacios tienen que ver directamente con el texto siguiente. Este va a tener que ver en un sentido más indirecto.

Mi experiencia de cursar una Maestría fue una experiencia formativa académica, pero también espiritual. Digo espiritual no en el sentido esotérico, si no en algo mas materialista; crecimiento personal, conocerse a uno mismo, etc. A través de duros aprendizajes intelectuales pude encontrar un área de trabajo que me apasiona, y es la evolución. Incidentalmente, manifestada en genética de poblaciones con el uso de herramientas computacionales.

Esta experiencia, muchas veces dada por sentada por cientificxs más avanzados en su carrera, me marcó profundamente, y es por eso que me tomo el tiempo de escribir un prefacio de este estilo. Acotaciones del tipo "nadie se acuerda que vas a hacer en tu maestría," o "te vas a olvidar eventualmente, la carrera es algo que se construye," tienen una esencia, que por más cínica que sea, es verdadera en el sentido más objetivo de la realidad material. Pero fallan al momento de describir como es vivida este tipo de experiencia desde lo subjetivo, como científico y ser humano. Como hermano, como hijo, como amigo. Como integrante de la comunidad, como parte del mundo.

En una instancia como esta, se forjan conocimientos especializados, y eso es lo que cuantificamos con los créditos de la escolaridad, con la escritura de este mismo texto. Y está perfecto que sea así. Los conocimientos especializados son lo que dictan la ciencia de hoy en día, el progreso cuantificable en horas dedicadas es una de las maneras que tenemos de medir la magnitud de lo adquirido. Pero hay otro tipo de conocimiento que se forja que es personal; cómo relacionarte con tus pares académicos, cómo preguntar, cuándo preguntar (siempre), en fin, todas las particularidades sociales de las cuales la ciencia no se escapa como actividad humana que es.

Pero dentro de lo personal también está cómo uno vive como científico. Cómo se relaciona uno con la comunidad, con sus seres queridos. Cómo uno mantiene el cuidado de su salud mental en una actividad muy intensiva intelectualmente y muy demandante emocionalmente. Claramente este tipo de aprendizaje no es exclusivo de la academia, si no que forma parte del proceso de crecimiento personal en el cual todas las personas se ven involucradas de un modo u otro, pero al transitar este proceso sumergido completamente en la academia, el aprendizaje personal se vuelve académico, y el académico se vuelve personal.

Y a pesar de las dificultades, lo disfruté profundamente. Y como soy un millenial, soy hedonista por imperativa generacional. Así que seguiré haciendo esto porque *me gusta*.

Todo este disfrute no hubiera sido posible sin un pequeño ejército de personas que estuvieron y están apoyandome día a día. Esta tesis de maestría no es un fruto únicamente mío, sino de esas personas también, ya que sin ellas, yo no podría haber hecho nada de esto. Es lo que tiene ser una especie comunitaria bajo un paradigma individualista. A veces atribuímos éxitos personales a esfuerzos colectivos, o bueno, al azar mismo.

En fin.

Agradezco de todo mi corazón a estas personas que me apoyaron siempre. Agradezco a mi Madre y a mi Padre por siempre estar ahí, y darme un apoyo incondicional, y una sensación visceral de felicidad al saber que ellos están orgullosos de mí. Agradezco a mis hermanas porque siempre me dan alegría. No hay día que pase sin que no piense en ellas y que las extrañe, a pesar de verlas cada tanto, pero nunca lo suficiente. Agradezco a mis amigos del alma, que me escuchan y apoyan en todo momento. A Mauri, a Agus, a Bruno, a Fede. A toda mi familia, mi abuela, mi tía, mis primos, que no nos vemos tan seguido, pero cuando lo necesité estuvieron ahí, y espero poder siempre reciprocar. Y agradezco a mis otres amigues; no son menos importantes por ponerles en la categoria "otres" , en momentos clave me han dado apoyo y cariño que atesoro profundamente.

Agradezco un montón a Luisa, que fué mi mentora de grado, pero para mí será mi mentora toda mi vida. Entre mis modelos a seguir, está ella. Su calidad y forma tan humana de hacer ciencia cambiaron mi perspectiva de cómo disfrutar esta actividad más de lo que me doy cuenta. Agradezco a Nata, que mientras trabajaba en cosas que no me gustaban, siempre me dio para adelante, y me dio consejos y un lugar de escape a mis frustraciones. Agradezco a todos mis compañeros y compañeras de la UBi, que no solo son muy divertidos, si no que son excepcionalmente solidarios y cariñosos.

Agradezco a Maine, que me tomó bajo su ala después de salir corriendo de temas que me di cuenta, no me apasionaban. Me dió la posibilidad de tener una segunda oportunidad, a pesar de ella misma estar ocupadísima con un montón de cosas, y siempre desde un lugar muy humano. Agradezco a Hugo, por su paciencia, que siempre fué comprensivo y creyó en mí a pesar de que titubee moviéndome de un campo a otro.

Realmente, no hubiera sido posible sin ustedes. No vale la pena sin ustedes.

Agradezco también a la Universidad de la República, al PEDECIBA y al Institut Pasteur de Montevideo que me dieron el marco institucional para trabajar en esto. Agradezco al Departamento de Genética de Facultad de Medicina por los años formativos como docente. Y por último, agradezco a la Comisión Académica de Posgrado, a la Agencia Nacional de Investigación e Innovación y al Centro Interdisciplinario en Ciencia de Datos y Aprendizaje Automático por los fondos dedicados para mi formación como científico. Agradezco a Alexandra Elbakyan por sus aportes invaluables para una construir una ciencia libre.

La defensa de esta tesis fue grabada, y su link puede encontrarse aquí.

Capítulo 1

Introducción

Uno de los factores más importantes en la evolución de las poblaciones de organismos es el aislamiento. Se refiere a aislamiento (genético) cuando una población que en un momento se encontró en panmixia (reproducción al azar) sufre una división en dos (o más) poblaciones con suspensión de flujo génico (o migración). Ésta suspensión del flujo génico implica que las nuevas poblaciones formadas no intercambian material genético entre sí.

Un tipo particular de aislamiento es el de las poblaciones estructuradas de forma jerárquica. Estas son poblaciones que comparten un ancestro común, y donde el ancestro común fue sufriendo aislamientos sucesivos formando una estructura de árbol.

A pesar de compartir una historia ramificada, las poblaciones pueden, a lo largo de las generaciones, intercambiar material genético a través de migraciones sostenidas en el tiempo, o en forma de pulsos. Nos interesa el caso en el cual las poblaciones han intercambiado material genético recientemente, en forma de pulso de migración o *admixture*. El hecho de que las poblaciones hayan evolucionado de forma jerárquica, y que se den pulsos migratorios entre ellas, induce correlaciones en las frecuencias alélicas de las distintas poblaciones.

En este trabajo estudiaremos algunas estrategias para estimar dichas correlaciones en forma de matrices de covarianza de poblaciones y las evaluaremos en datos simulados. Estas serán: la matriz de covarianza *teórica*, obtenida analíticamente a partir de los parámetros de deriva y migración de las simulaciones, la matriz de covarianza *kinship* obtenida a través de un árbol filogenético estimado con frecuencias alélicas, la matriz de covarianza *empírica*, obtenida a partir de las frecuencias alélicas estimadas del ancestro común a todas las poblaciones, y por último, la matriz de covarianza *treemix*, obtenida a partir del modelado explícito de un pulso de migración bajo un marco de máxima verosimilitud (Pickrell and Pritchard 2012).

Otro factor importante en los procesos evolutivos es la selección natural. Ésta deja huellas en los genomas de las poblaciones que pueden ser aprovechadas por métodos estadísticos para detectar regiones del genoma que estuvieron bajo selección natural, y que pueden indicar la relevancia funcional de dichas regiones. Uno de éstos métodos es hapFLK, un estadístico que detecta selección a partir de información haplotípica y diferenciación poblacional (que puede ser especificada como una matriz de covarianza de poblaciones)(Fariello et al. 2013).

En la presente tesis de Maestría en Bioinformática (PEDECIBA) bajo el título *"Detección de selección en poblaciones mezcladas recientemente"* se expone el trabajo realizado para extender la capacidad de detección del test estadístico hapFLK usando matrices de covarianza alélicas que modelen migraciones recientes en poblaciones que se formaron a través de ramificaciones históricas.

En primer lugar, se exponen las bases teóricas y metodológicas necesarias para comprender la motivación detrás del problema. Primero, se hará una revisión del modelo básico de genética de poblaciones (principio Hardy Weinberg, modelo Wright-Fisher), para luego explorar las desviaciones de los modelos que conciernen los casos de migración y selección. Segundo, se revisarán algunos métodos en particular que permiten detectar y estimar migración (TreeMix) y selección (LK, FLK y hapFLK).

Luego se expondrán los métodos utilizados y con ellos el código desarrollado para obtener los resultados. Se explicitarán los escenarios simulados para obtener datos genotípicos, los métodos de obtención de matrices de

covarianza, y el software para la eficiente simulación y cálculo de estadísticos.

Finalmente, se expondrán los resultados separados en dos secciones. En la primera parte se concluye que todas las formas de estimación de matrices de covarianza resultan en estadísticos hapFLK con distribuciones χ^2 bajo deriva genética, y la mejor forma de estimar sus grados de libertad es a partir de regresión por cuantiles, para el posterior cómputo de p-valores. En la segunda parte se concluye que, en el caso de un *hard sweep*, el poder de hapFLK es relativamente alto, siendo la estimación empírica de la matriz de covarianza la que se desempeña mejor, seguida de la estimación a través del algoritmo TreeMix.

1.1. Modelo básico de genética de poblaciones

1.1.1. Principio de Hardy-Weinberg

Para que sea posible una discusión productiva sobre poblaciones mezcladas (de ahora en más, *admixture*) y selección, es necesario empezar por los pilares de la teoría genética de poblaciones. El primero es conocido comúnmente en cursos introductorios de genética y evolución como el principio de Hardy-Weinberg (HW). Éste explica cómo se comportan las frecuencias alélicas de un único locus bajo ciertas condiciones ideales:

1. Población diploide
2. Población hermafrodita
3. Tamaño de población infinito
4. Apareamientos al azar (panmixia)
5. Ausencia de inmigración
6. Ausencia de mutación
7. Ausencia de selección
8. Generaciones no solapadas

Bajo estas condiciones, en un locus con alelos A_1 y A_2, el principio de HW propone que las frecuencias genotípicas en la generación $t+1$ van a estar dadas por las frecuencias alélicas en la generación t. Siendo las respectivas frecuencias genotípicas $f(A_1A_1)$, $f(A_1A_2)$ y $f(A_2A_2)$, las frecuencias alélicas en la generación t serán:

$$\begin{aligned} p_t = f(A_1) &= f(A_1A_1) + \frac{1}{2}f(A_1A_2) \\ (1-p_t) = f(A_2) &= f(A_2A_2) + \frac{1}{2}f(A_1A_2) \end{aligned}$$

El principio de Hardy Weinberg implica que los organismos en la generación $t+1$ van a obtener sus alelos al azar con reposición, es decir, obtendrán un alelo A_1 con probabilidad p_t y un alelo A_2 con probabilidad $1-p_t$. Entonces, la proporción de genotipos en $t+1$ va a estar dada por:

$$\begin{aligned} f(A_1A_1) &= p_t^2 \\ f(A_1A_2) &= 2p_t(1-p_t) \\ f(A_2A_2) &= (1-p_t)^2 \end{aligned} \tag{1.1}$$

y las frecuencias alélicas se mantienen iguales:

$$\begin{aligned} p_{t+1} = f(A_1A_1) + \frac{1}{2}f(A_1A_2) &= p_t^2 + \frac{1}{2}p_t(1-p_t) = p_t \\ (1-p_{t+1}) = f(A_2A_2) + \frac{1}{2}f(A_1A_2) &= (1-p_t)^2 + \frac{1}{2}p_t(1-p_t) = (1-p_t) \end{aligned}$$

Este resultado implica que las frecuencias alélicas (y por consecuencia, las genotípicas) se mantienen constantes a lo largo del paso de las generaciones, y por lo tanto, no hay evolución.

Claramente, en poblaciones reales, este conjunto de condiciones presenta diversos tipos de desviaciones. Pero como toda aproximación matemática a un problema aplicado, es necesario asumir tales condiciones para llegar a resultados simplificados y comprensibles. Para obtener resultados que se acerquen más a lo observado en poblaciones reales, es necesario ser menos restrictivo y violar alguna(s) condicion(es) del modelo.

1.1.2. Ligamiento entre loci

Nos referimos como haplotipo a una combinación particular de alelos que se encuentran a lo largo de un cromosoma. Tomando como ejemplo dos loci A y B con alelos A_1, A_2 y B_1, B_2, respectivamente, y con frecuencias $f(A_1) = p_A$, $f(B_1) = p_B$, se pueden obtener cuatro haplotipos a partir de las distintas combinaciones alélicas.

Tomaremos como ejemplo uno de los cuatro haplotipos posibles: A_1B_1. El hecho de que los loci se encuentren en el mismo cromosoma va a resultar que estén asociados probabilísticamente de acuerdo a la frecuencia de los haplotipos presentes en la población, de manera que $P(A_1B_1) \neq p_Ap_B$.

De hecho, la probabilidad de encontrar un haplotipo A_1B_1 va a estar relacionada con la probabilidad de que se dé un evento de recombinación entre el locus A y el locus B. Llamaremos a esta probabilidad r, y la probabilidad de obtener a un haplotipo A_1B_1 a partir de eventos de recombinación va a estar dada por rp_Ap_B

Se le llama ligamiento a la propiedad que tienen los loci de estar probabilísticamente asociados a través de las probabilidades de recombinación. Si no existe recombinación entre loci, las frecuencias haplotípicas se comportan como un locus en HW. Un concepto útil para describir que tan asociados se encuentran dos loci en términos probabilísticos es el desequilibrio de ligamiento D. Este concepto lo podemos definir matemáticamente como

$$D^{(t)} = P(A_1B_1)^{(t)} - p_Ap_B \tag{1.2}$$

Conceptualmente, mide la diferencia entre la probabilidad de encontrar el haplotipo A_1B_1 en una generación t y la probabilidad de encontrar dicho haplotipo si no existiera asociación entre los loci, dada por el producto de sus frecuencias alélicas correspondientes.

Si consideramos que la probabilidad de obtener un haplotipo A_1B_1 en una generacion $t+1$ va a estar dada por:

$$P(A_1B_1)^{(t+1)} = (1-r)P(A_1B_1)^{(t)} + rp_Ap_B$$

(la probabilidad de encontrar un haplotipo A_1B_2 sin que haya sido fruto de una recombinación más probabilidad de encontrar el haplotipo fruto de una recombinación).

Siguiendo la relación de recurrencia en t, podemos seguir la dinámica de $P(A_1B_1)$ a lo largo de las generaciones:

$$\begin{aligned} P(A_1B_1)^{(t+1)} - p_Ap_B &= (1-r)P(A_1B_1)^{(t)} + rp_Ap_B - p_Ap_B \\ P(A_1B_1)^{(t+1)} - p_Ap_B &= (1-r)(P(A_1B_1)^{(t)} - p_Ap_B) \end{aligned}$$

Recordamos que $D^{(t)} = P(A_1B_1)^{(t)} - p_Ap_B$, entonces

$$P(A_1B_1)^{(t+1)} - p_Ap_B = (1-r)(P(A_1B_1)^{(t)} - p_Ap_B)D^{(t+1)} = (1-r)D^{(t)}$$

y

$$D^{(t)} = (1-r)^t D^{(0)} \tag{1.3}$$

Entonces el desequilibrio de ligamiento entre dos alelos va a disminuir con el paso de las generaciones de manera exponencial, hasta llegar un punto en el cual es posible calcular las frecuencias haplotípicas a partir de las correspondientes frecuencias HW.

Cabe destacar que en el caso general, A y B no tienen porqué estar en el mismo cromosoma. Cuando no lo están, $r = 0{,}5$, y puede haber desequilibrio de ligamiento según la Ecuación (1.3).

El fenómeno de desequilibrio de ligamiento , o LD por sus siglas en inglés, tiene profundos efectos sobre la arquitectura genética de las poblaciones; realísticamente se dan lo que son conocidos como "bloques de LD" , donde frecuencias alélicas se encuentran correlacionadas en bloques discretos, separados por sitios de alta recombinación ("*recombination hotspots*").

Los patrones que forman estos bloques de LD pueden cambiar sustancialmente en presencia de un alelo adaptativo como veremos en la Sección 1.3.1.

1.1.3. Poblaciones subdivididas

En ecosistemas reales, las poblaciones se encuentran subdivididas en grupos o estratos. Ya sea separadas por completo en islas, a través de cordilleras, o siendo los dos extremos de un rango extenso interconectado. Las subdivisiones tienen como resultado la suspensión del flujo génico entre las distintas subpoblaciones. En otras palabras, las subpoblaciones se encuentran aisladas desde el punto de vista genético.

Ponemos por ejemplo el caso de dos poblaciones A y B aisladas entre sí, en las cuales cada una se encuentra en equilibrio HW. Aquí, las frecuencias genotípicas van a estar determinadas por las frecuencias alélicas p_A y p_B, y a su vez las frecuencias alélicas no van a cambiar con el paso de las generaciones.

A pesar de que A y B se encuentran en HW, si las frecuencias alélicas son distintas entre A y B, la metapoblación conceptualizada como la unión de ambas poblaciones no se encontrará en equilibrio HW y las frecuencias genotípicas no van a estar dadas exactamente por las frecuencias alélicas como en la Ecuación (1.1). Este fenómeno se conoce como el efecto Wahlund, y surge cuando existe subdivisión dentro de una población estudiada.

Siendo $\bar{p} = \frac{p_A + p_B}{2}$, las frecuencias genotípicas de la metapoblación van a estar alejadas del equilibrio HW de acuerdo a un factor que es la varianza de las frecuencias alélicas $\sigma^2 = (\bar{p} - p_A)^2 + (\bar{p} - p_B)^2$:

$$
\begin{aligned}
f(A_1A_1) &= \bar{p}^2 + \sigma^2 \\
f(A_1A_2) &= 2\bar{p}(1-\bar{p}) - 2\sigma^2 \\
f(A_2A_2) &= (1-\bar{p})^2 + \sigma^2
\end{aligned}
$$

Entonces, en términos de frecuencias genotípicas podemos ver que el efecto Wahlund está caracterizado por una reducción de la frecuencia de los genotipos heterocigotas, o heterocigosidad, a expensas de un aumento en la homocigosidad (Hahn 2018).

1.1.3.1. El estadístico F_{ST}

Una de las formas más comunes y conceptualmente simples de medir estas diferencias en frecuencias alélicas entre poblaciones (diferenciación) es el estadístico F_{ST}. El estadístico F_{ST}, en este contexto, está definido como

$$
F_{ST} = \frac{\sigma^2}{\bar{p}(1-\bar{p})}
$$

y puede ser interpretado como la proporción de la varianza de la metapoblación ($\bar{p}(1-\bar{p})$) atribuíble a diferencias entre las frecuencias alélicas de las subpoblaciones (σ^2), y se puede relacionar con las frecuencias genotípicas de la metapoblacion de la siguiente manera:

$$
\begin{aligned}
f(A_1A_1) &= \bar{p}^2 + \bar{p}(1-\bar{p})F_{ST} \\
f(A_1A_2) &= 2\bar{p}(1-\bar{p}) - 2\bar{p}(1-\bar{p})F_{ST} \\
f(A_2A_2) &= (1-\bar{p})^2 + \bar{p}(1-\bar{p})F_{ST}
\end{aligned}
$$

Podemos ver que cuando no hay diferenciación y $F_{ST} = 0$, no hay deficiencia de heterocigotas, mientras que cuando la diferenciación es máxima con $F_{ST} = 1$ hay una deficiencia total de heterogicotas.

En el caso mas general de n poblaciones, el estadístico F_{ST} toma la forma

$$F_{ST} = \frac{\sigma^2}{\bar{p}(1-\bar{p})} = \frac{\sum_{i=1}^{n}(\bar{p}-p_i)^2}{\bar{p}(1-\bar{p})}$$

con $\bar{p} = \frac{1}{n}\sum_{i=1}^{n} p_i$.

En este sentido, al estudiar varias poblaciones, podemos interpretar a F_{ST} como medida de la diversidad de un locus, cuanto más cercano a cero, mayor diversidad (mayor heterocigosidad).

1.1.4. Modelo Wright-Fisher

El modelo que utiliza el principio de Hardy-Weinberg tiene muchas derivaciones de acuerdo a las condiciones que son violadas. A nosotros nos interesa imponer una restricción en particular; asumir que el tamaño poblacional no es infinito, sino de tamaño N. Ésta es la principal desviación del modelo Wright-Fisher respecto a HW. En esta introducción, también asumiremos que los organismos son diploides y dióicos. Las últimas dos asunciones son aproximadamente equivalentes a una dinámica de organismos hermafroditas y haploides de tamaño $2N$, siempre y cuando el apareamiento sea al azar y el número de hembras sea el mismo que el numero de machos (Felsenstein 2005).

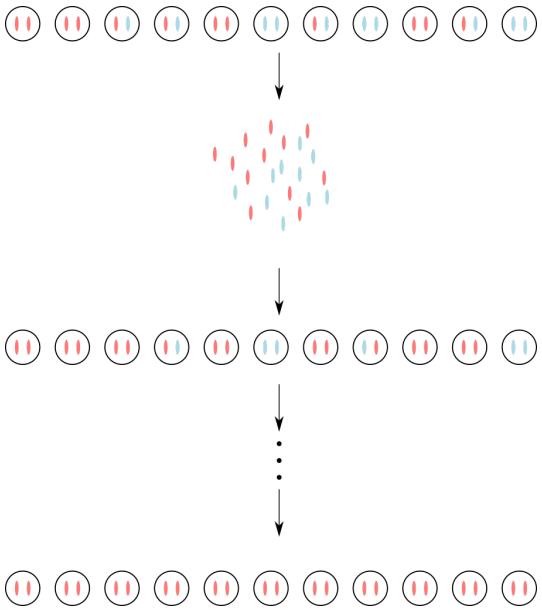

Figura 1.1: Ejemplo de cambios azarosos en las frecuencias alélicas por tamaño finito poblacional. Los individuos contribuyen al acervo genético de alelos, y los alelos son muestreados con reposición para dar lugar a la siguiente generación de individuos.

Al trabajar con una población finita, se introduce un componente azaroso en la dinámica de las frecuencias alélicas a lo largo del tiempo. Los $2N$ alelos en la generación t son elegidos al azar con reposición para dar lugar a los $2N$ alelos en la generación $t+1$ (Figura 1.1). El muestreo al azar causa que las frecuencias alélicas cambien de forma estocástica a lo largo de las generaciones; este fenómeno es conocido como deriva genética.

El cambio estocástico en frecuencias alélicas lo podemos expresar en términos del número de alelos de tipo A_1 en la población. La probabilidad p_{ij} de obtener j alelos de tipo A_1 en la generación $t+1$ va a depender del número de alelos A_1 i en la generación t y del número total de alelos ($2N$). Esto lo podemos expresar en términos de una distribución binomial:

$$p_{ij} = \binom{2N}{j}\left(\frac{i}{2N}\right)^{j}\left(1-\frac{i}{2N}\right)^{(2N-j)} \tag{1.4}$$

Claramente, se puede realizar el mismo razonamiento para los alelos del tipo A_2, donde su número es $k = 2N - i$.

Entonces, dado un tamaño poblacional constante, el modelo WF es una cadena de Markov definida sobre el número de alelos A con una matriz de transición $\mathbf{P}_{(2N\times 2N)} = [p]_{ij}$. Esta cadena posee dos estados absorbentes que van a estar dados por los vectores propios $\mathbf{v}_1 = \begin{bmatrix}1 & 0 & 0 & \dots & 0\end{bmatrix}^T$ y $\mathbf{v}_2 = \begin{bmatrix}0 & 0 & 0 & \dots & 1\end{bmatrix}^T$, que representan los estados en los cuales se fijó el alelo de tipo A_1 o el de tipo A_2, respectivamente. La existencia de estos dos estados absorbentes garantizan que eventualmente uno de los dos alelos será fijado (Felsenstein 2005).

Bajo un modelo de WF, asumiendo que el tipo alélico A_1 tiene una frecuencia $p_1 = \frac{i}{2N}$, A_1 va a ser fijado con probabilidad $P(\text{fijacion}) = p_1$, y su tiempo de persistencia esperado será

$$E(t_{fijacion}) = t(p_1) = -4N\big[p_1 log(p_1) + (1-p_1)log(1-p_1)\big]$$

El tiempo de persistencia puede ser interpretado como el tiempo que un alelo se encontrará segregando en la poblacion, sin fijarse ni perderse. Este resultado se desprende del tratamiento de la deriva genética como un proceso de difusión, desarrollado por Fisher (1958) y Wright (1945), cuya solución a la ecuación diferencial que propone el problema fue descubierta por Kimura (1955).

Lo anterior implica que con suficientes generaciones la deriva llevará a la pérdida o fijación del alelo A. Si lo extendemos a todos los loci del genoma, implica que la deriva llevará a una pérdida de variabilidad genética en los locus segregantes.

Sabemos entonces el destino de las frecuencias alélicas a largo plazo, pero no hemos explorado cómo se comportan las esperanzas de las frecuencias alélicas a corto plazo; de una generación a la siguiente.

Sea p_{t-1} la frecuencia alélica de A_1 a tiempo $t-1$, la frecuencia a tiempo t cambiará por un valor d y será $p_t = p_{t-1} + d$. Podemos calcular la frecuencia esperada a tiempo t condicional a la frecuencia en el tiempo $t-1$:

$$E(p_t|p_{t-1}) = E(p_{t-1} + d|p_{t-1}) = E(p_{t-1}|p_{t-1}) = p_{t-1} \tag{1.5}$$

entonces,

$$E(p_t) = E(E(p_t|p_{t-1})) = E(p_{t-1}) = E(p_{t-2}) = \cdots = p_0 \tag{1.6}$$

La frecuencia alélica esperada va a ser la frecuencia inicial p_0 a un tiempo t_0 arbitrario. Y la varianza de p_t a lo largo del tiempo será

$$\begin{aligned}
\text{Var}(p_t) &= E\big(\text{Var}(p_t|p_{t-1})\big) + Var\big(E(p_t|p_{t-1})\big) \\
&= E\left(\frac{p_{t-1}(1-p_{t-1})}{2N}\right) + \text{Var}(p_{t-1}) \\
&= \frac{1}{2N}E\big(p_{t-1} - p_{t-1}^2\big) + \text{Var}(p_{t-1}) \\
&= \frac{1}{2N}\left(p_0 - E\big(p_{t-1}^2\big)\right) + \text{Var}(p_{t-1}) \\
&= \frac{1}{2N}\left(p_0 - \big(\text{Var}(p_{t-1} + p_0^2)\big)\right) + \text{Var}(p_{t-1}) \\
\text{Var}(p_t) &= \frac{p_0(1-p_0)}{2N} - \frac{\text{Var}(p_{t-1})}{2N} + \text{Var}(p_{t-1})
\end{aligned}$$

Invirtiendo el signo de la ecuación, y sumando $p_0(1-p_0)$ en ambos lados:

$$\begin{aligned}
p_0(1-p_0) - \text{Var}(p_t) &= p_0(1-p_0) - \frac{p_0(1-p_0)}{2N} - \text{Var}(p_{t-1}) + \frac{\text{Var}(p_{t-1})}{2N} \\
&= \left(1 - \frac{1}{2N}\right)\Big(p_0(1-p_0) + \text{Var}(p_{t-1})\Big)
\end{aligned}$$

Siguiendo la relación de recurrencia en t, el comportamiento de $Var(p_t)$ con respecto a la frecuencia inicial p_0, es:

$$\mathrm{Var(p_t)} = \left[1 - \left(1 - \frac{1}{2N}\right)^t\right] p_0(1 - p_0) \tag{1.7}$$

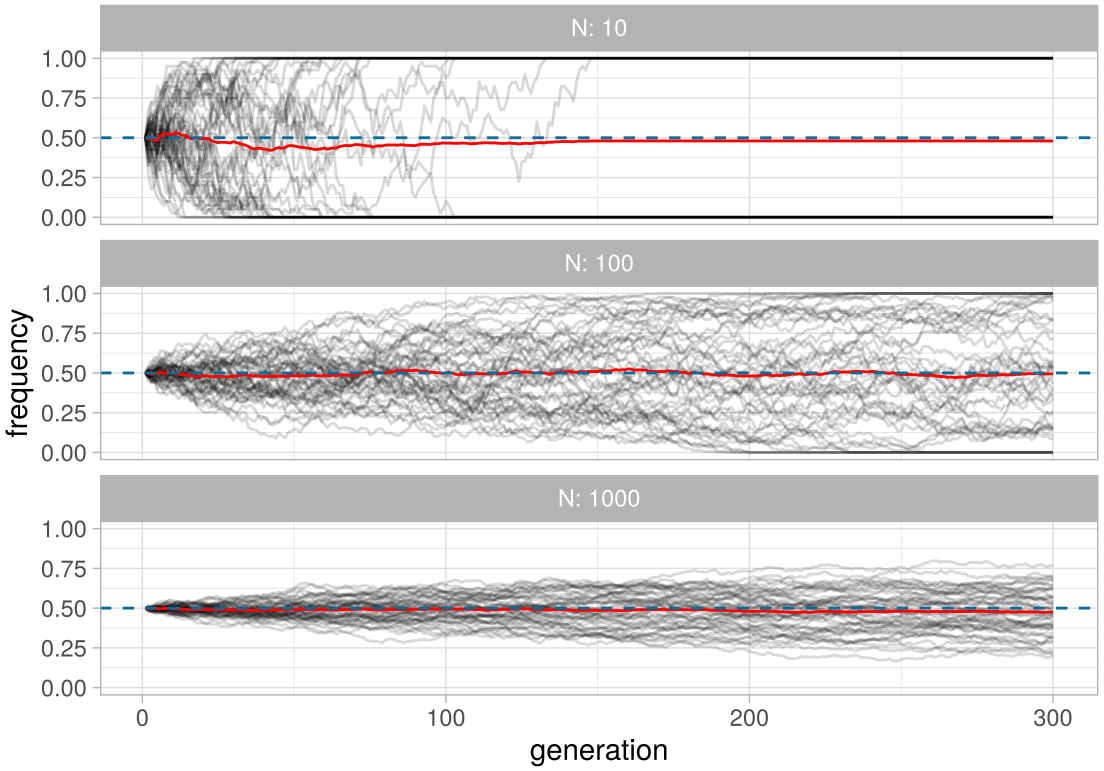

Figura 1.2: Simulaciones de la trayectoria de la frecuencia alélica de poblaciones bajo el modelo WF para distintos valores de tamaño poblacional N durante 300 generaciones. En azul está representada la frecuencia inicial del alelo, en rojo está representada la frecuencia alélica media por generación, y el número de réplicas es 50 para cada N.

Dejamos claro entonces que el valor esperado de la frecuencia alélica a lo largo del tiempo va a estar dado por la frecuencia alélica inicial, mientras que la varianza de la frecuencia alélica aumenta con el paso de las generaciones (Figura 1.2).

1.1.5. Deriva en varias poblaciones

En general, los estudios de genética de poblaciones no se restringen a una población, si no que les compete trabajar con varias poblaciones. Muchas veces estas poblaciones comparten un ancestro en común, que devino en las actuales poblaciones a través de ramificaciones históricas. Estas ramificaciones son suspensiones secuenciales en el flujo génico, que van delimitando las poblaciones a lo largo del tiempo. Gracias a este comportamiento ramificante, podemos representar al proceso a lo largo del tiempo como un árbol filogenético (Figura 1.4), donde cada hoja es una población de muestra, y cada nodo interno es el ancestro común de todas las hojas correspondientes, siendo el largo de ramas el grado de deriva.

Al igual que en el modelo WF, consideramos que las poblaciones van a estar sujetas a fuerzas de deriva determinadas por el tamaño poblacional. Además, de aquí en adelante, asumimos que una vez que la población original se separa en varias poblaciones, los loci estudiados son neutrales (no hay selección) y no hay mutación. Para esta última asunción, nos basamos en un modelo de *"separation-of-timescales"* Wakeley (1999), que asume que la variabilidad genética se origina en una metapoblación ancestral, y luego la fase de división de poblaciones sucede lo suficientemente rápido como para asumir que no hay mutación.

Tomando como ejemplo la Figura 1.3 A, si tenemos una población B que proviene de una población A, podemos formular la frecuencia alélica de un SNP en B de la siguiente manera

$$p_B = p_A + \epsilon_B$$

donde p_A,p_B son las frecuencias del SNP en las poblaciones A y B, respectivamente, y ϵ_B es el cambio de frecuencia alélica debido a deriva, que lo modelamos como $\epsilon_B \sim N\big(0, \delta_{AB} p_A(1-p_A)\big)$ donde δ_{AB} es un coeficiente que refleja la deriva, y es $\delta_{AB} \approx \frac{t_{AB}}{2N}$ donde t_{AB} es el tiempo en generaciones que pasó entre la separación de las poblaciones A y B (modelo desarrollado por Cavalli-Sforza and Piazza (1975) y extendido por Nicholson et al. (2002) para SNPs).

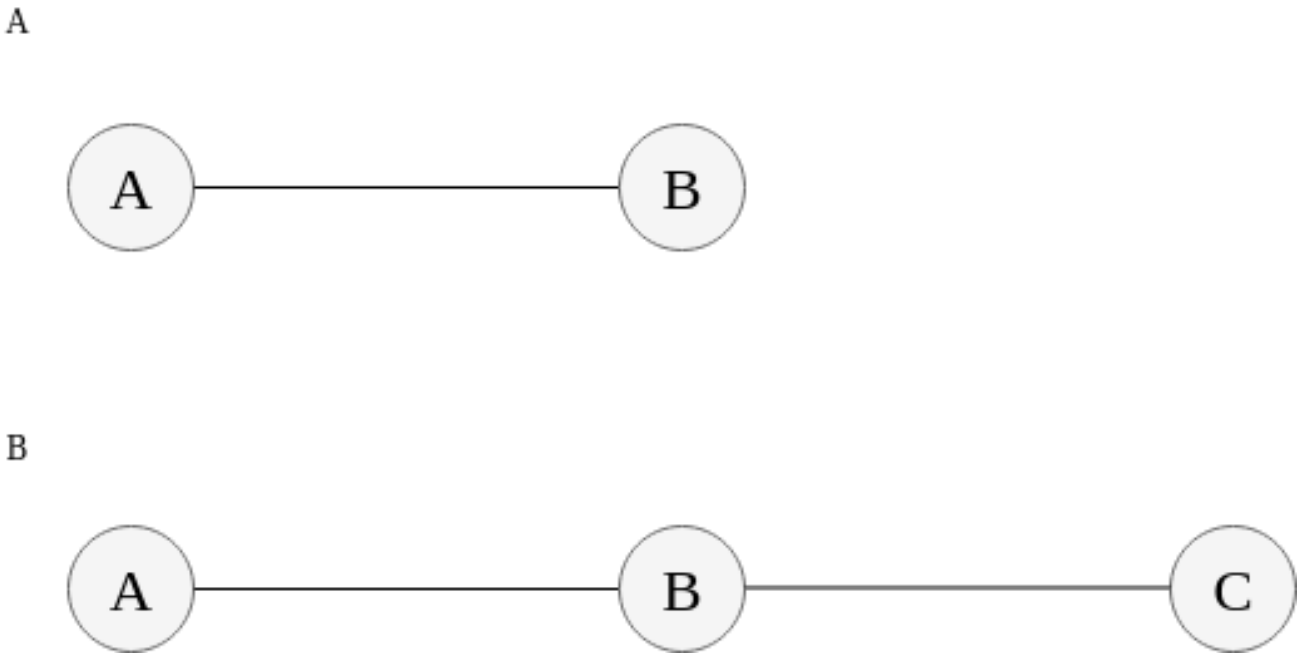

Figura 1.3: Ejemplos de poblaciones relacionadas. En el ejempo superior, la poblacion A da lugar a la población B. En el inferior es similar, con la población B dando lugar a la población C.

Ahora, teniendo en cuenta una tercera población C que desciende de B (Figura 1.3 B), podemos calcular la esperanza y varianza de las frecuencias alélicas de C condicional a la frecuencia alélica de A.

$$E(p_C|p_A) = E(p_A + \epsilon_B + \epsilon_C) = p_A$$

y

$$\begin{aligned} \text{Var}(p_C|p_A) &= \text{Var}(p_A + \epsilon_B + \epsilon_C) \\ &= \text{Var}(\epsilon_B) + \text{Var}(\epsilon_C) + 2\text{Cov}(\epsilon_B, \epsilon_C) \end{aligned}$$

Si asumimos que la deriva entre poblaciones es pequeña, podemos decir que $p_B(1-p_B) \approx p_A(1-p_A)$. Además la deriva entre A y B es independiente de la deriva entre B y C (Cavalli-Sforza and Piazza (1975)). Entonces

$$\begin{aligned} \text{Var}(p_C|p_A) &\approx \text{Var}(\epsilon_B) + \text{Var}(\epsilon_C) \\ &\approx \delta_{AB} p_A(1-p_A) + \delta_{BC} p_B(1-p_B) \\ &\approx (\delta_{AB} + \delta_{BC}) p_A(1-p_A) \end{aligned}$$

y la frecuencia alélica en C sigue aproximadamente una distribución normal con media en la frecuencia alélica de A y varianza que depende de la frecuencia alélica en A y los coeficientes de deriva de la forma

$$p_C \approx N\big(p_A, (\delta_{AB} + \delta_{BC})p_A(1-p_A)\big)$$

Para describir el comportamiento de las frecuencias alélicas en un árbol de poblaciones como, por ejemplo el de la Figura 1.4, seguiremos el modelo anterior, y la covarianza entre dos poblaciones va a ser igual a la varianza del ancestro común de las dos poblaciones.

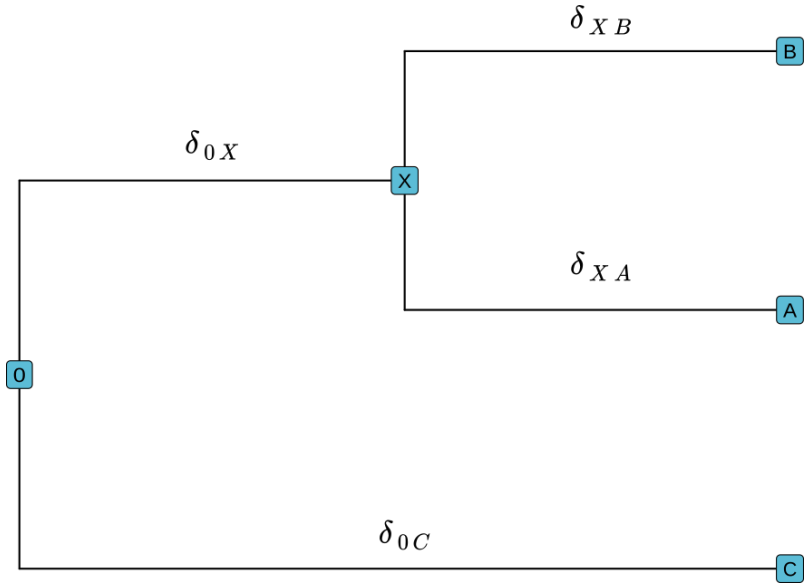

Figura 1.4: Árbol filogenetico de poblaciones con una estructura jerárquica. Cada nodo coresponde con una población, los coeficientes indican la cantidad de deriva sufrida a lo largo del tiempo.

En el caso específico de la Figura 1.4,

$$Cov(p_A, p_B) = \delta_{0X} p_0(1-p_0)$$

y

$$Cov(p_A, p_C) = 0$$

Sabiendo el valor de p_0, podemos modelar el vector de las frecuencias alélicas de las poblaciones como una normal multivariada:

$$\mathbf{p} \sim MVN(\mathbf{p_0}, \mathbf{V})$$

donde

$$\mathbf{p_0} = \begin{bmatrix} p_0 \\ p_0 \\ p_0 \end{bmatrix}$$

y, siguiendo el razonamiento de la ecuación (1.1.5),

$$\begin{aligned} \mathbf{V} &= \begin{bmatrix} \delta_{0X} + \delta_{AB} & \delta_{0X} & 0 \\ \delta_{0X} & \delta_{0X} + \delta_{XA} & 0 \\ 0 & 0 & \delta_{0C} \end{bmatrix} p_0(1-p_0) \\ &= \mathcal{F} p_0(1-p_0) \end{aligned} \tag{1.8}$$

Siendo $\mathcal{F}$ la matriz que describe las correlaciones entre poblaciones debido a la estructura jerárquica.

En síntesis, podemos modelar los cambios en frecuencias alélicas de varias poblaciones conociendo la frecuencia inicial p_0, la forma en la cual se generaron las poblaciones (topología del árbol), y los tamaños poblacionales.

1.2. Migración

La migración entre poblaciones es un fenómeno que induce correlaciones de frecuencias alélicas que no estan únicamente dadas por la topología del árbol, si no que también van a estar dadas por las características del evento de migración en sí. Este caso es de especial interés ya que el fenómeno de *admixture* es esencialmente un fenómeno de migración, de flujo génico entre poblaciones. Modelaremos el fenómeno de migración de acuerdo a Pickrell and Pritchard (2012).

Primero, consideraremos cuál es el comportamiento probabilístico de las frecuencias alélicas a lo largo de un árbol de poblaciones, para luego incluir un evento de migración en el árbol.

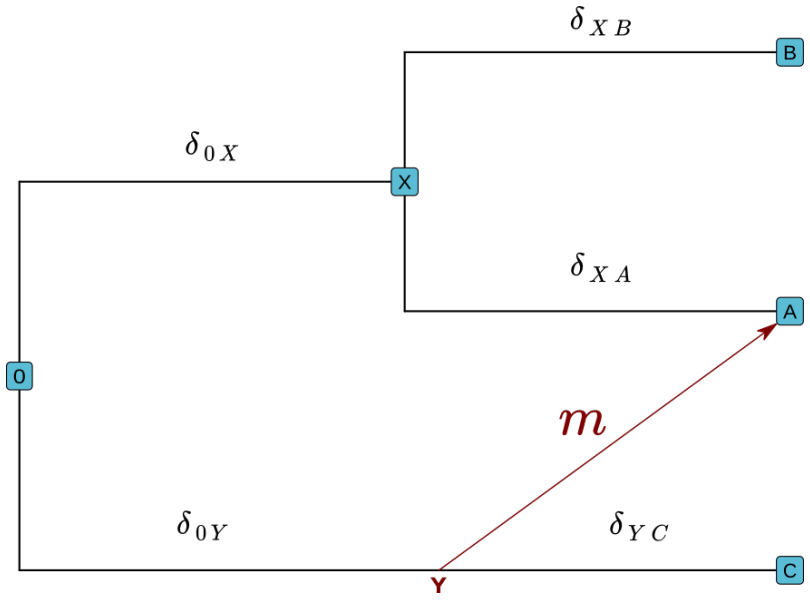

Figura 1.5: Árbol filogenetico de poblaciones con una estructura jerárquica y un pulso de admixture. Cada nodo coresponde con una población, los coeficientes indican la cantidad de deriva sufrida a lo largo del tiempo. m corresponde al peso del pulso de admixture.

Para incorporar un evento de migración usaremos el modelo de deriva visto en la sección anterior, y extenderlo al grafo de *admixture* de la Figura 1.5, donde asumimos que la población A va a estar compuesta por una mezcla de X e Y debido a un pulso migratorio de una única generación.

Modelamos las proporciones de esta mezcla con un peso m que va a definir las proporciones de ancestría de A.

$$p_A = mp_Y + (1-m)(p_X + \epsilon_A)$$

donde $\epsilon_A \sim N\big(0, \delta_{XA} p_0(1-p_0)\big)$.

Entonces $E(p_A) = p_0$ y

$$
\begin{aligned}
\mathrm{Var}(p_A) &= \mathrm{Var}\big(mp_Y + (1-m)(p_X + \epsilon_A)\big) \\
&= m^2\mathrm{Var}(p_Y) + (1-m)^2\Big(\mathrm{Var}(p_X) + \mathrm{Var}(\epsilon_A)\Big) + 2m(1-m)\mathrm{Cov}(p_Y, p_X) \\
&= m^2\mathrm{Var}(p_Y) + (1-m)^2\Big(\mathrm{Var}(p_X) + \mathrm{Var}(\epsilon_A)\Big) \\
&= m^2\delta_{0Y}p_0(1-p_0) + (1-m)^2\big(\delta_{0X}p_0(1-p_0) + \delta_{XA}p_0(1-p_0)\big) \\
&= p_0(1-p_0)\Big(m^2\delta_{0Y} + (1-m)^2(\delta_{0X} + \delta_{XA})\Big)
\end{aligned}
$$

aplicando el criterio anterior, se llega a la matriz de covarianza de las poblaciones que será

$$
\mathbf{V} = \begin{bmatrix} \delta_{0X} + \delta_{XB} & (1-m)\delta_{0X} & 0 \\ (1-m)\delta_{0X} & m^2\delta_{0Y} + (1-m)^2(\delta_{0X} + \delta_{XA}) & \delta_{0Y}m \\ 0 & \delta_{0Y}m & \delta_{0Y} + \delta_{YC} \end{bmatrix} p_0(1-p_0) \tag{1.9}
$$

1.3. Selección

La selección natural es una desviación del principio de HW donde algunos alelos tienen una probabilidad mayor de permanecer en la población, en contraste con los alelos neutrales. Además, visto de una manera más amplia, la selección natural es el proceso que permite a las poblaciones adaptarse a los ecosistemas de los cuales son parte. Dos condiciones son necesarias para que sea posible la selección natural; la presencia de variabilidad fenotípica, y que esta variabilidad sea (al menos parcialmente) heredable.

Los individuos de una población compiten entre sí y con otras especies por los recursos disponibles. Sumado a esto, varios desafíos abióticos pueden presentarse, como por ejemplo cambios climáticos. Estas presiones que surgen de la competencia y del ambiente hacen que los diferentes individuos tengan diferentes probabilidades de dejar descendencia de acuerdo a su fenotipo. En otras palabras, el fenotipo de un individuo dado va a estar asociado con una probabilidad de poder sobrevivir a ciertas condiciones, y dejar descendientes para la siguiente generación. A su vez, la siguiente generación va a estar sujeta a competencia y presiones ambientales y cada uno de sus individuos va a presentar cierta probabilidad de dejar descendencia, siguiendo así a lo largo de generaciones.

Aquí haremos un tratamiento introductorio de los principios básicos que rigen el proceso de selección natural sobre frecuencias alélicas. Asumiremos que el tamaño poblacional es infinito (no hay deriva), y que la selección se da en un único estadío en el ciclo de vida de cada individuo. Además haremos la asunción de que la población bajo estudio es una población hermafrodita. Todas estas asunciones son necesarias para un tratamiento simplificado de la selección natural, pero las conclusiones a las cuales llegaremos son bastante generales. Y si es necesario plantear escenarios más realistas, las asunciones pueden relajarse adoptando modelos más sofisticados.

Siguiendo la presentación y notación de Gillespie (2004), denotemos w_{11}, w_{12} y w_{22} como las probabilidades de supervivencia de los individuos con genotipos A_1A_1, A_1A_2 y A_2A_2, respectivamente. A esta probabilidad le llamaremos viabilidad, que también puede ser interpretada como la proporción de individuos de cada genotipo que van a conformar la siguiente generación debido a presiones selectivas diferenciales. Dadas unas frecuencias genotípicas iniciales dictadas por las frecuencias alélicas de A_1 (p) y de A_2 $(1-p)$ en la generación $t-1$, las nuevas frecuencias genotípicas en la generación t van ser las frecuencias de HW multiplicadas por la viabilidad del genotipo correspondiente y divididas por la viabilidad promedio $\bar{w}$, donde $\bar{w} = p^2w_{11} + 2p(1-p)w_{12} + (1-p)^2w_{22}$ (Ecuación (1.10), fila f_t).

	$\mathbf{A_1A_1}$	$\mathbf{A_1A_2}$	$\mathbf{A_2A_2}$
f_{t-1}	p^2	$2p(1-p)$	$(1-p)^2$
viabilidad	w_{11}	w_{12}	w_{22}
f_t	$p^2\frac{w_{11}}{\bar{w}}$	$2p(1-p)\frac{w_{12}}{\bar{w}}$	$(1-p)^2\frac{w_{22}}{\bar{w}}$
viabilidad relativa	1	$\frac{w_{12}}{w_{11}}$	$\frac{w_{22}}{w_{11}}$
fitness relativo	1	$1-hs$	$1-s$

(1.10)

Es práctico expresar las viabilidades como viabilidades relativas a una de un genotipo en particular. Digamos que el genotipo con mayor viabilidad es A_1A_1, y expresaremos las viabilidades relativas a éste (Ecuación (1.10), viabilidad relativa). Conceptualmente, de esta manera estamos expresando las viabilidades en términos de qué tan viables son los otros dos genotipos con respecto a A_1A_1. Por conveniencia, definimos el *fitness* relativo en relación a la viabilidad relativa como

$$\frac{w_{12}}{w_{11}} = 1 - hs$$
$$\frac{w_{22}}{w_{11}} = 1 - s$$

Donde h le llamamos el coeficiente de dominancia, y s el coeficiente de selección, y ambos pueden tomar cualquier valor entre $-\infty$ y $+\infty$. Conceptualmente, el coeficiente de selección s es el valor que determina la magnitud y el genotipo en el cual se encuentra la presión selectiva; si s es positivo, hay seleccion en contra del genotipo A_2A_2, si es negativo, es a favor. El coeficiente de dominanacia h determina el comportamiento de las presiones selectivas sobre los genotipos heterocigotas. Si $0 < h < 1$, se dice que hay *dominancia incompleta* y el genotipo heterocigota es un intermedio (en terminos de viabilidad) entre los genotipos homocigotas. Si $h < 0$, se da el fenómeno de *sobredominancia*, donde el genotipo más viable es el heterocigota.

Los distintos valores de h y de s van a afectar la dinámica de las frecuencias alélicas a lo largo de las generaciónes. En un modelo determinístico donde el tamaño poblacional es infinito, podemos expresar el cambio de la frecuencia alélica p de una generación a la siguiente como:

$$\begin{aligned}\Delta p &= \frac{p(1-p)\Big[p(w_{11} - w_{12}) + (1-p)(w_{12} - w_{22})\Big]}{p^2 w_{11} + 2p(1-p)w_{12} + (1-p)^2 w_{22}} \\ &= \frac{p(1-p)s[ph + (1-p)(1-h)]}{\bar{w}}\end{aligned}$$

Claramente h y s van a determinar la trayectoria de la frecuencia alélica p ya que van a determinar el signo de Δp. Además cabe notar que Δp también dependerá de p en cada generación anterior. Consideraremos dos casos particularmente relevantes: selección direccional y selección equilibradora.

La selección direccional se da cuando $s > 0$ y hay dominancia incompleta, es decir, $0 < h < 1$. En este caso, el signo de Δp será positivo para todos los valores posibles de p, y la frecuencia alélica de A_1 aumentará hasta fijarse. Éste es un caso clásico en el cual un alelo brinda mayor viabilidad de forma incremental (aditiva si $h = \frac{1}{2}$).

El caso de la selección equilibradora se da cuando $s > 0$ y el genotipo heterocigota A_1A_2 es el más viable ($h < 0$). Esto tiene como resultado que dada cualquier frecuencia inicial, p va a tender a un valor de equilibrio p_{eq}:

$$p_{eq} = \frac{h-1}{2h-1}$$

1.3.1. Selección y ligamiento

Teniendo en cuenta que las frecuencias alélicas a lo largo de un cromosoma se encuentran correlacionadas por la (ausencia de) recombinación, es de esperar que si un alelo adaptativo A_1 surge por mutación en el cromosoma, la frecuencia del haplotipo en el cual apareció va a aumentar, mediada por los efectos de la recombinación.

Este fenómeno en el cual alelos neutros aumentan de frecuencia por proximidad física a un locus bajo selección positiva es conocido como *autoestop genético*, y fue formulado en términos matemáticos por primera vez por Smith and Haigh (1974). En este articulo, se llega a la conclusión de que el efecto del autoestop genético es una reducción de la diversidad (o heterocigosidad) de los loci segregantes aledaños al locus bajo selección. La tasa a la cual la heterocigosidad decrece va a depender de varios parámetros: el coeficiente de selección s, el coeficiente de dominancia h,la frecuencia inicial del alelo adaptativo, el tamaño poblacional N y de la probabilidad de recombinación r entre los loci neutrales y bajo selección (que a su vez va a depender de su distancia física).

Mencionaremos dos escenarios de autoestop que son de particular de interés: *hard sweeps* y *soft sweeps*. Los *hard sweeps* suceden cuando, en una region haplotípica inicialmente neutra, ocurre una mutación que brinda una ventaja adaptativa. Si el coeficiente de seleccion s y el tamaño poblacional N son lo suficientemente altos, el haplotipo en el cual la mutación adaptativa surgió va a aumentar rápidamente en frecuencia, y una vez fijada la mutación adaptativa, va a haber una dramática pérdida de heterocigosidad en los loci aledaños. Debido a la recombinación, no se fija enteramente el haplotipo en el cual surgió la mutación, sino que la heterocigosidad de los loci neutrales va a disminuir progresivamente de acuerdo a la cercanía fisica al locus bajo selección. Este tipo de eventos deja una huella clara en el genoma, ejemplificada en la Figura 1.6.

El caso de los *soft sweeps* es diferente en el sentido de que la mutación no es nueva, si no que ya existía en la población, y por razones ecológicas su *fitness* pasó a ser positivo. Esto implica que antes de que comience la selección, la variante adaptativa se encuentra en varios haplotipos distintos. Como resultado, una vez fijada la variante adaptativa, la reducción en la heterocigosidad no va a ser tan extensa al compararla con un *hard sweep* debido a la presencia de varios haplotipos. Las huellas que dejan los *soft sweeps* en el genoma son más sutiles, y por consecuencia, más difíciles de detectar (Figura 1.6, derecha).

Como hemos visto, la presencia de selección va determinar el destino de las frecuencias alélicas a largo plazo en un modelo determinístico como HW. Una vez que se toma en cuenta la deriva, tratamos con un modelo estocástico, y lidiamos con probabilidades de fijación que van a estar influenciadas por los coeficientes de dominancia y selección. A pesar de convertirse en un proceso estocástico, los principios básicos revisados se mantienen.

Podemos ver que, en particular, la selección direccional va a resultar en una pérdida de la diversidad en el locus que se encuentra bajo selección a medida que aumenta la frecuencia de un alelo que brinda mayor viabilidad. Esta característica de los procesos de selección puede ser explotada por test estadísticos basados en la medición de la diversidad de distintos loci para detectar regiones genómicas que se encuentran, o se encontraron, bajo selección. A continuación veremos algunas de las más relevantes para el objetivo del trabajo.

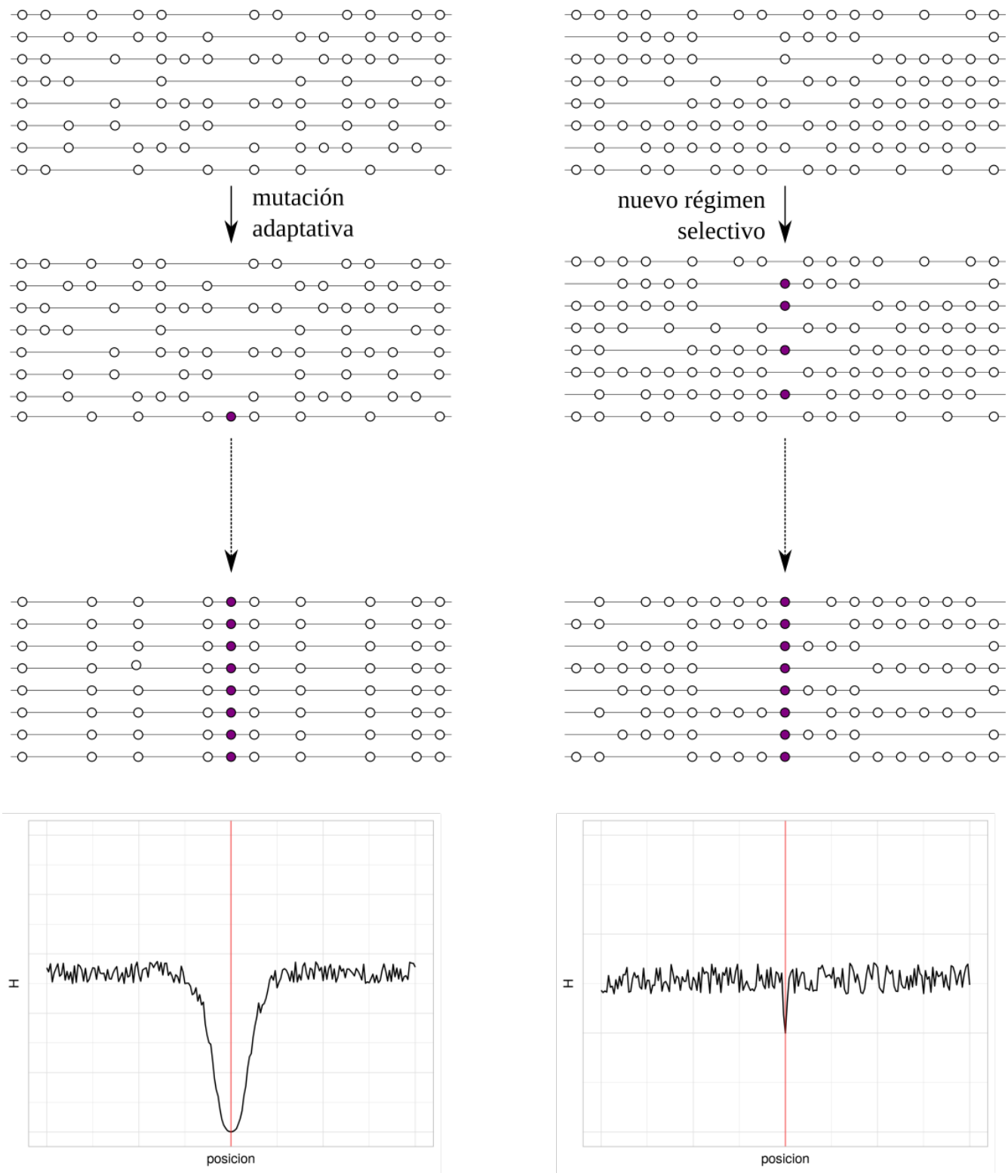

Figura 1.6: Hard sweep (izquierda), soft sweep (derecha) con sus correspondientes señales de heterocigosidad a lo largo del genoma.

Capítulo 2

Métodos

Este estudio pretende determinar la distribución empírica del estadístico hapFLK y su poder para detectar selección usando distintas estimaciones de la matriz de covarianza de poblaciones en presencia de migración.

Se realizaron simulaciones con parámetros demográficos específicos que permitieron obtener una aproximación analítica de la matriz de covarianza *teórica*, que indica el desempeño ideal del estadístico. A partir de los datos genómicos simulados, se calcularon las matrices de covarianza de *kinship* y *empírica* con el software hapFLK, y la matriz de covarianza *treemix* que modela un pulso de migración con el software TreeMix (ver Sección 2.1).

Así se obtuvieron distribuciones de valores de hapFLK que permitieron comparar su desempeño al usar distintas matrices de covarianza en el caso de poblaciones que han sufrido *admixture* recientemente.

El trabajo se dividió en dos secciones principales.

En la primera, llamada Calibración, se estudió el comportamiento empírico de la distribución de hapFLK bajo neutralidad usando las distintas matrices de covarianza. Realizar un estudio de este tipo es relevante para poder determinar tanto la distribución del estadístico bajo la hipótesis nula, como la forma más adecuada de realizar tests de significancia. Se computó hapFLK con los genotipos resultantes de simular dos escenarios demográficos de distinta complejidad, con cien réplicas para cada combinación de coeficiente de *amixture* m y coeficiente de selección s.

En la segunda sección, llamada Evaluación, se determinó el poder estadístico de detección de selección. Se utilizaron las distribuciones de hapFLK bajo neutralidad como distribuciones bajo la hipótesis nula de ausencia de selección. En este caso, se simuló un único escenario demográfico con mil réplicas para cada combinación de s y m.

Conceptualmente, ambas secciones son independientes. En Calibración se estudió únicamente la distribución del estadístico, lo que permitió trabajar con más de un escenario demográfico y con menor número de réplicas, aliviando el costo computacional. En Evaluación, siendo el objetivo el cálculo del poder de detección, se priorizó computar un alto número de réplicas bajo un único escenario demográfico.

En Calibración, se simularon pulsos de *admixture* en el sentido amplio, donde la población mezclada ya tiene asociada cierto grado de deriva. En Evaluación, se simularon pulsos de *admixture* en el sentido estricto, donde el *admixture* genera una nueva población. La razón por la cual se realizaron estos distintos escenarios es que simulando demografías variadas es posible obtener una visión más amplia del comportamiento de hapFLK.

2.1. Estimación de matrices de covarianza

Se utilizaron diferentes matrices de covarianza para calcular el estadístico hapFLK, que se presentan a continuación.

2.1.1. Teórica

Siguiendo el modelo de Pickrell and Pritchard (2012) expuesto en la Sección 1.2, se calcularon las matrices de covarianza de poblaciones teóricas, teniendo en cuenta los tiempos de divergencia, la topología del árbol, y los coeficientes de *admixture* m.

Calibración

De acuerdo a las topologías indicadas, se calcularon las matrices de covarianza teórica para los escenarios A y B (Figuras 2.1 y 2.2, respectivamente). Cabe destacar que el escenario A se encuentra anidado en el escenario B, donde forma parte de un clado dentro de la filogenia completa de B.

Los parámetros indicados como δij representan el coeficiente de deriva acorde al tiempo t entre el nodo padre i y el nodo j, es decir $\delta_{ij} = \frac{t_{ij}}{2N_{ij}}$. El parámetro indicado como m es el coeficiente de *admixture* entre dos poblaciones; tomando el escenario A como ejemplo, indica que en la generación del pulso la nueva generación de `p4` es $100 * m\,\%$ compuesta por cromosomas de `p2` y $100 * (1 - m)\,\%$ de cromosomas de `p4`. Lo mismo sucede en el escenario B.

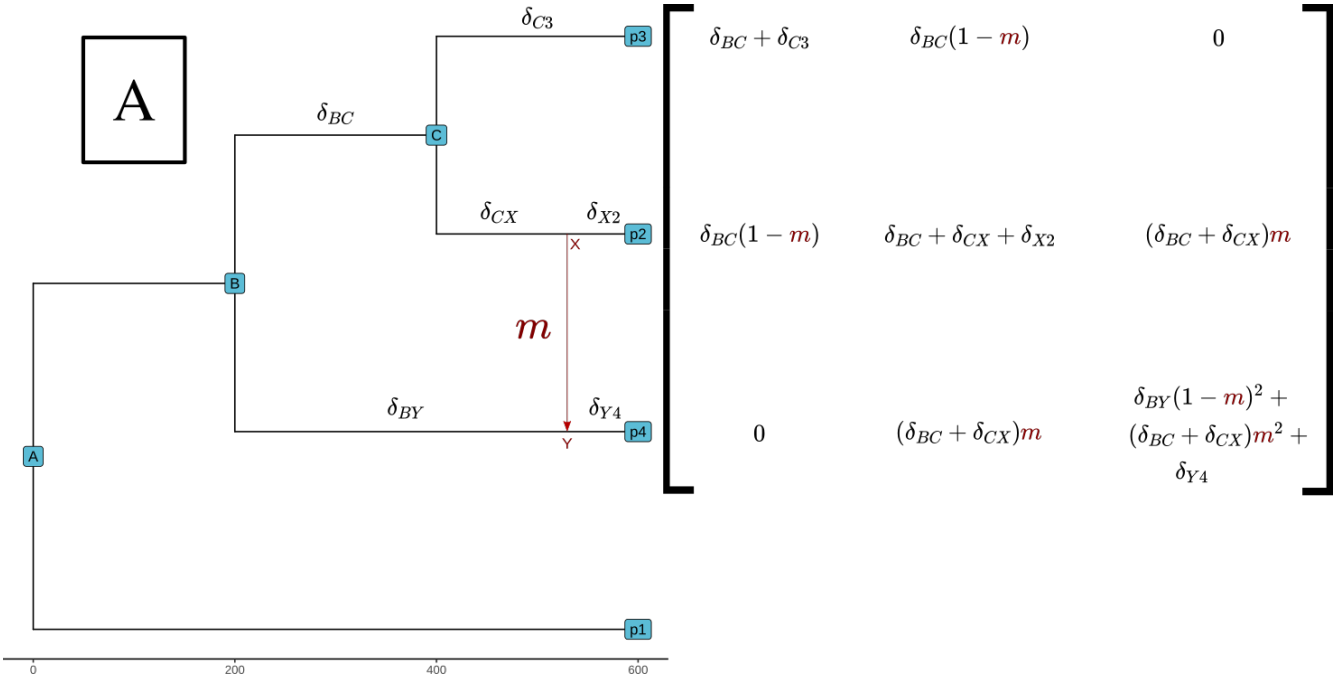

Figura 2.1: Matriz teórica para escenario A. Las letras delta corresponden la cantidad de deriva, mientras que el parámetro indicado como m corresponde al cociente de admixture.

Evaluación

Se utilizó el mismo criterio para el cálculo de la matriz teórica del escenario demográfico de evaluación. En la Figura 2.3 se encuentra la topología en conjunto con la matriz de covarianza teórica. En este caso, la población `p5` se origina a partir de una mezcla de $100 * m\,\%$ cromosomas de `p4` y $100 * (1 - m)\,\%$ cromosomas de `p2`.

2.1.2. Kinship

La matriz de *kinship* $\mathcal{F}$ de la Ecuación (1.8) es estimada construyendo un árbol de poblaciones con el algoritmo Neighbor Joining (NJ)(Saitou and Nei 1987). Éste es calculado a partir de las distancias de Reynolds (Reynolds 1983) de las frecuencias alélicas. Sea $Q_{(n\times l)}$ la matriz de frecuencias alélicas para n poblaciones en l loci bialélicos y $R = QQ^T$, la matriz de distancia de Reynolds D esta dada por

$$D_{ij} = \frac{R_{ii} + R_{jj} - 2R_{ij}}{2(n - R_{ij})}$$

Los largos de ramas del árbol resultante son usados como estimadores de los coeficientes de inbreeding correspondientes a cada entrada de $\mathcal{F}$. Estos coeficientes van a tener en cuenta el N (a menor N, más deriva y mayor largo de ramas), y las correlaciones entre poblaciones debido a la estructura jerárquica. Cada entrada de la matriz (par de poblaciones) va a corresponder a la distancia de la raíz al ancestro común del par de poblaciones. Las diagonales de la matriz serán la distancia desde la raíz hasta el nodo hoja de la población correspondiente (Figura 2.4).

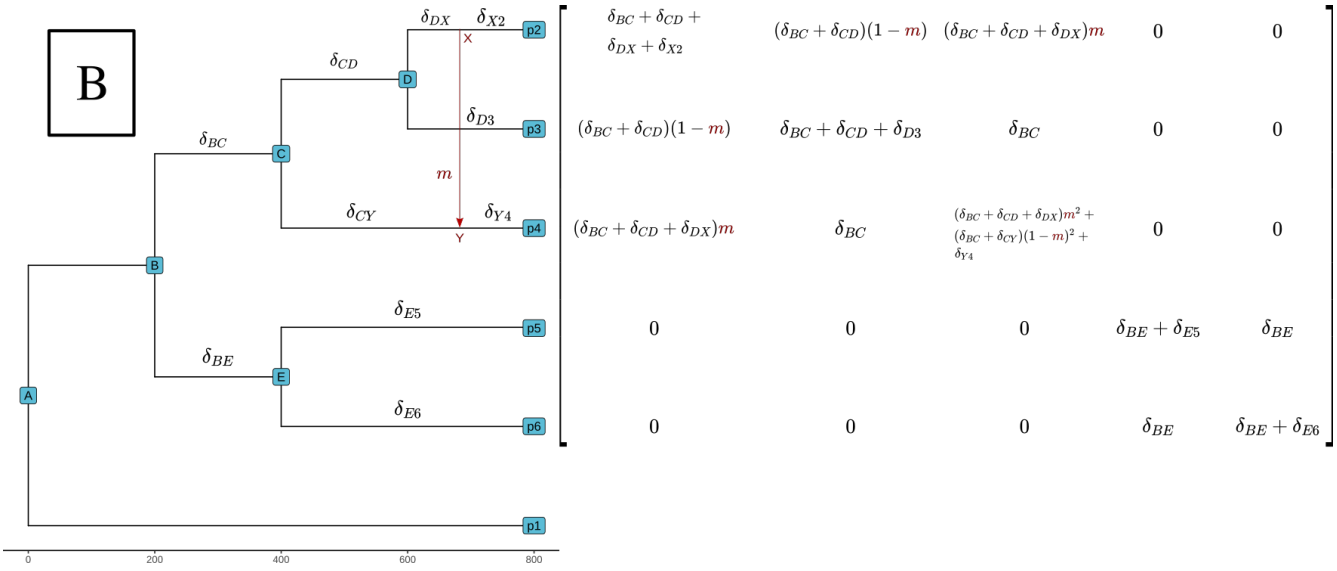

Figura 2.2: Matriz teórica para escenario B. Ibid.

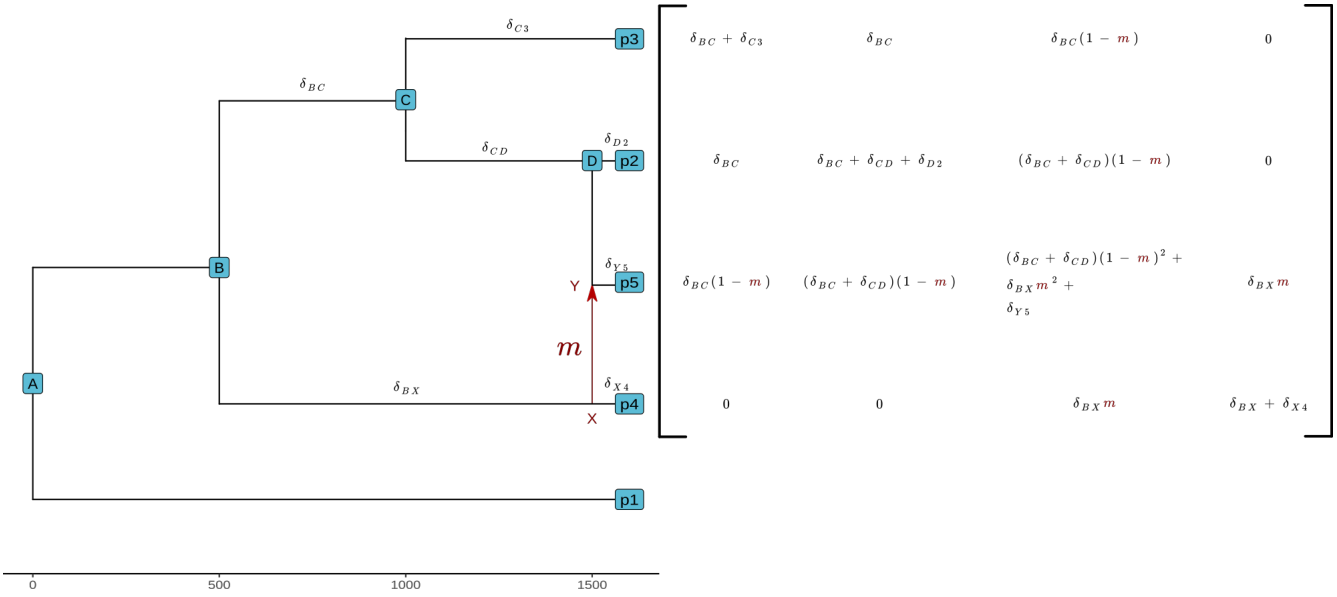

Figura 2.3: Matriz teórica para escenario de evaluación. Ibid

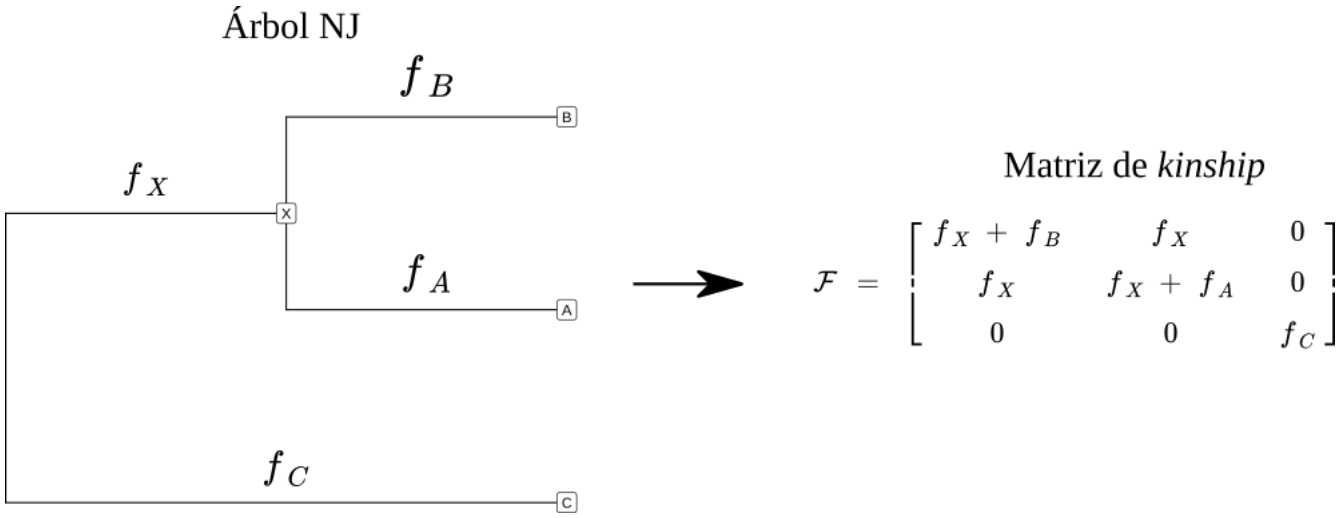

Figura 2.4: Ejemplo de filogenia resultante y matriz correspondiente calculada con el método de kinship.

2.1.3. TreeMix

Se calcularon matrices de covarianza con el algoritmo TreeMix (Pickrell and Pritchard 2012).

2.1.3.1. Modelo

El algoritmo TreeMix tiene como objetivo estimar los patrones de *splits* y *admixture* de poblaciones. En particular, estima las ramificaciones históricas de poblaciones relacionadas, para luego inferir la presencia de un número k de pulsos de admixture, con sus respectivos pesos m. Usa el modelo de migración expuesto en la Sección 1.2 para estimar un grafo de poblaciones G a partir de la covarianza entre frecuencias alélicas.

Usualmente en muestras genómicas no es posible acceder a información sobre las frecuencias alélicas ancestrales p_0. Entonces TreeMix realiza las inferencias sobre la matriz de covarianza muestral $\mathbf{W}$, donde $W_{ij} = E\big((p_i - \bar{p})(p_j - \bar{p})$ y $\bar{p} = \frac{1}{n}\sum_{i=1}^{n} p_i$ para n poblaciones. De hecho, $\mathbf{W}$ se relaciona con $\mathbf{V}$ de la siguiente manera:

$$\begin{aligned} W_{ij} &= E\big((p_i - \bar{p})(p_j - \bar{p}) \\ &= V_{ij} - \frac{1}{n}\sum_{k=1}^{n} V_{ik} - \frac{1}{n}\sum_{k=1}^{n} V_{jk} + \frac{1}{n^2}\sum_{k=1}^{n}\sum_{k'=1}^{n} V_{kk'} \end{aligned} \tag{2.1}$$

El grafo G a estimar es un grafo con raíz, dirigido y acíclico con largo de ramas y pesos de migración definidos. TreeMix utiliza un marco de máxima verosimilitud para maximizar la verosimilitud $\mathcal{L}(\hat{W}|G)$, donde $\hat{W}$ es la matriz de covarianza estimada.

La información a utilizar serán las frecuencias alélicas $\mathbf{P}_{(n\times l)}$ de n poblaciones en l loci. Se estima W_{ij} a partir de esta información de forma eficiente particionando el genoma en p bloques y asumiendo, por el Teorema Central del Límite, que $\hat{W}_{ij}$ va a seguir una distribución normal dada por:

$$\hat{W}_{ij} = N(\hat{W}_{ij}, \sigma_{ij}^2) \tag{2.2}$$

siendo σ_{ij} el error estándar en la estimación de $\hat{W}_{ij}$. Calculando $\hat{\mathbf{W}}$ en cada bloque, los estimadores de la media y varianza son

$$\hat{W}_{ij} = \frac{1}{p}\sum_{k=1}^{p} \hat{W}_{ijk} \tag{2.3}$$

$$\hat{\sigma}_{ij} = \sqrt{\frac{\sum_{k=1}^{p}(\hat{W}_{ijk} - \hat{W}_{ij})^2}{p(1-p)}} \tag{2.4}$$

Tomando de a pares de poblaciones, se computa la verosimilitud compuesta

$$\mathcal{L}(\hat{\mathbf{W}}|\mathbf{W}) = \prod_{i=1}^{n}\prod_{j=1}^{n} N(\hat{W}_{ij}|G, \hat{\sigma}_{ij}^2) \tag{2.5}$$

donde $N(\hat{W}_{ij}|G, \hat{\sigma}_{ij}^2)$ es una densidad Gaussiana con media W_{ij} computada a partir del grafo G definido por la Ecuación (2.1) y varianza σ_{ij}^2 evaluada en $\hat{W}_{ij}$.

Es computacionalmente costoso (si no imposible) enumerar todas las posibles combinaciones de grafos teniendo en cuenta migraciones, así que TreeMix asume una historia de poblaciones ramificantes y construye un árbol con raíz a partir de las frecuencias alélicas usando máxima verosimilitud (Felsenstein 1981), para luego explorar los pares de poblaciones que muestran ajuste pobre, y agregando pulsos de migración para mejorar el ajuste.

Una vez computado el mejor grafo por máxima verosimilitud a partir del árbol inicial y la Ecuación (2.5), se calcula la matriz de valores residuales $\mathbf{R}$:

$$\mathbf{R} = \hat{\mathbf{W}} - \mathbf{W} \tag{2.6}$$

Valores de R_{ij} que se alejen de cero son indicadores de un pobre ajuste al grafo. En el siguiente paso, se determinan los k pares de poblaciones con peor ajuste, y se agregan pulsos de migración en los nodos vecinos hasta lograr un mejor ajuste. En cada paso se maximiza la función de verosimilitud en función del largo de ramas y pesos de migración.

Como resultado se obtiene un grafo G que será un árbol de poblaciones con k ejes de migraciones (k proveído como parámetro). Además, se obtiene $\hat{\mathbf{W}}$, la matriz de covarianza entre poblaciones explicada por G.

2.1.3.2. Implementación

Se usó el software TreeMix (Pickrell and Pritchard 2012) para calcular matrices de covarianzas que modelen explícitamente eventos de migración o *admixture*. Se especificó una única arista de migración (`-m 1`) y se realizaron réplicas de bootstrap en bloques de 10 SNPs (`-bootstrap 10`)

```
treemix -i ${countfile} \
        -m 1 \
        -o ${rep_id}_${params.edges}  \
        -bootstrap -k 10
```

TreeMix devuelve la matriz de covarianza $\mathbf{W}$ estimada a partir del modelo con migración: `*.modelcov.gz`. Esta matriz contiene el outgroup en una de sus entradas, como hapFLK no acepta una matriz de covarianza que contenga la población *outgroup* (`p1`), se retiró la columna y fila correspondiente.

2.1.4. Empírica

La matriz empírica fue calculada automáticamente por el software haFLK a partir de los vectores de frecuencias alélicas de cada población. Para k poblaciones, si p_i es el vector de dimensión k de frecuencias en el locus i, y L es el número total de loci, la matriz de covarianza empírica S es

$$S = \sum_{i=1}^{L} \mathbf{q}_i \mathbf{q}_i^T$$

donde

$$\mathbf{q_i} = \frac{1}{\sqrt{\hat{p}_{0i}(1-\hat{p}_{0i})}}(\mathbf{p}_i - \hat{p}_{0i}\mathbf{1})$$

Siendo p_{0i} la frecuencia alélica ancestral del locus i estimada usando la matriz de kinship $\mathcal{F}$ según lo detallado en la Sección 2.1.2. $\mathbf{q}_i$ puede ser interpretado como un vector de frecuencias alélicas centrado y normalizado de acuerdo a su frecuencia alélica ancestral, al restarse cada frecuencia por el valor ancestral y dividirse por la desviacion estándar esperada bajo una distribución binomial ($\sqrt{\hat{p}_{0i}(1-\hat{p}_{0i})}$).

2.2. Detección de selección

2.2.1. Test LK

Una de las estrategias usadas para detectar selección a partir de los genomas de individuos es determinar la distribución de alguna medida de diversidad a lo largo del genoma, y encontrar loci que se encuentren muy por fuera de la media de esta distribución (outliers).

El razonamiento detrás de estas estrategias asume que los niveles de diversidad de la mayoría de los loci son resultado de procesos demográficos, mientras que sólo unos pocos loci en el genoma muestran la diversidad característica de presiones selectivas. En el caso de selección positiva direccional, se esperaría una reducción de la diversidad, mientras que en el caso de selección equilibradora se esperaría un aumento al comparar con los valores de diversidad del genoma.

Entonces, intuitivamente tienen como objetivo encontrar loci que muestren niveles inusuales de diversidad entre poblaciones, y en general estos niveles inusuales de diversidad locales son interpretados como reflejo de presiones selectivas sobre los loci. Uno de los primeros test estadísticos desarrollados con esta estrategia en mente fue el test de Lewontin y Krakauer (Lewontin and Krakauer 1973). Desarrollado en 1972, el test LK espacial[1] asume que las frecuencias alélicas son realizaciones de una distribución normal, y que el valor de F_{ST} es aproximadamente el mismo para todos los loci. El test consiste en determinar la distribución de valores de F_{ST} para varios loci en el genoma y computar el estadístico

$$T_{LK} = (n-1)\frac{F_{ST}}{\bar{F}_{ST}}$$

para cada locus, donde $\bar{F}_{ST}$ es el F_{ST} promedio de todos los loci, y n es el número de poblaciones estudiadas.

El test estadístico plantea

$$\begin{cases} H_0 : \text{Locus es neutral} \\ H_1 : \text{Locus se encuentra bajo selección} \end{cases}$$

Bajo H_0, T_{LK} sigue aproximadamente una distribución χ^2 con esperanza $E(T_{LK}) = (n-1)$. En el caso de valores altos del estadístico asociados con un p-valor lo suficientemente bajo, se rechaza H_0 y se interpreta que el locus se encuentra posiblemente bajo selección.

Un problema con el test LK es que niveles inusuales de diversidad en un loci específico no son necesariamente fruto de los efectos de presiones selectivas. Apenas fue publicado el test, varias críticas fueron realizadas fundamentando que el test es particularmente sensible a la presencia de poblaciones con estructura (Robertson 1975), migración, o con tamaño poblacional cambiante (Nei and Maruyama 1975). Estos eventos causarían valores sesgados de F_{ST} y por consecuencia falsos positivos de loci bajo selección positiva (observando menos diversidad de la esperada) y selección equilibradora (observando más diversidad de la esperada). Luego de las críticas, algunos autores argumentaron que para evitar los efectos de la estructura poblacional, es posible calcular T_{LK} con pares de poblaciones a cuesta de poder estadístico (Tsakas and Krimbas 1976) debido a la corrección necesaria por testeo múltiple.

A pesar de estas críticas claras, el test LK no fue modificado para tomar en cuenta estructura poblacional hasta el año 2010 con el test FLK (Bonhomme et al. 2010). Sin embargo, varias estrategias se han desarrollado para

[1] Es espacial en el sentido de que usan datos de distintas poblaciones en un *snapshot* temporal. En el mismo artículo desarrollaron un test para tratar con una población única a lo largo del tiempo. En éste texto hacemos referencia únicamente al primero.

tomar en cuenta estructuras jerárquicas basadas en el estadístico F_{ST} (Mark A. Beaumont (1996), Excoffier, Hofer, and Foll (2009), Yi et al. (2010), Günther and Coop (2013)). También se han desarrollado algunas que toman en cuenta variables ambientales para detectar selección (Coop et al. (2010), Gautier (2015)) y el caso mas general de adaptación poligénica (Racimo, Berg, and Pickrell 2018). Se han desarrollado otros métodos basados en el uso de componentes de ancestría estimados (X. Cheng, Xu, and DeGiorgio 2017) y recientemente se han desarrollado métodos para detectar selección en presencia de *admixture* arcaica (J. Y. Cheng, Racimo, and Nielsen 2019) y en historias demográficas arbitrariamente complejas (Refoyo-Martínez et al. 2019).

2.2.2. Test FLK

El test FLK, desarrollado por Bonhomme et al. (2010), utiliza el principio del test LK y lo extiende modelando la variabilidad de N y la covarianza entre poblaciones cuando hay ramificaciones históricas.

El estadístico T_{FLK} es definido como

$$T_{FLK} = (\mathbf{p} - p_0\mathbf{1})^T \mathrm{Var}(\mathbf{p})^{-1}(\mathbf{p} - p_0\mathbf{1})$$

donde $\mathbf{p}$ es el vector de frecuencias alélicas en n poblaciones (dimensión n), p_0 es la frecuencia alelica en la población ancestral, $\mathbf{1}$ es un vector unitario de dimesión n, y la matriz de covarianza de las frecuencias alélicas es modelada como

$$Var(\mathbf{p}) = \mathcal{F} p_0(1 - p_0)$$

siendo $\mathcal{F}$ la matriz de *kinship* de la Ecuación (1.8) donde $\mathcal{F}_{ii}$ es el coeficiente de deriva esperado de la población i, y $\mathcal{F}_{ij}$ es el coeficiente de deriva esperado de la poblacion ancestral de i y j.

Para computar el estadístico, es entonces necesario estimar la matriz de *kinship* $\mathcal{F}$ y la frecuencia alélica ancestral p_0.

2.2.2.1. Estimación de $\mathcal{F}$ y p_0

La matriz de *kinship* $\mathcal{F}$ es estimada de acuerdo a lo expuesto en la Sección 2.1.2. La frecuencia alélica ancestral p_0 es estimada a partir de las frecuencias alélicas de las poblaciones p y la matriz de *kinship*:

$$\hat{p_0} = \frac{\mathbf{1}^T \mathcal{F}^{-1} \mathbf{p}}{\mathbf{1}^T \mathcal{F}^{-1} \mathbf{1}}$$

T_{FLK} tiene esperanza

$$E(T_{FLK}) \approx n - 1$$

y varianza

$$\mathrm{Var}(T_{FLK}) \approx 2(n - 1)$$

suponiendo que $p \sim N(\hat{p}_0, \hat{p}_0(1 - \hat{p}_0))$, T_{FLK} sigue aproximadamente una distribución χ^2_{n-1} (Bonhomme et al. 2010).

2.2.2.2. Alelos múltiples

En el caso de que se trabaje con múltiples alelos en varios loci, es particularmente útil definir un estadístico FLK para loci multialélicos. En el caso de n poblaciones con A alelos en un locus, el vector de frecuencias alélicas es expresado como

$$\mathbf{P} = \begin{bmatrix} p_{11} \\ p_{21} \\ \vdots \\ p_{n1} \\ p_{12} \\ \vdots \\ p_{nA} \end{bmatrix}$$

Entonces, como ya hemos visto, la esperanza del vector $\mathbf{P}$ va a depender de las frecuencias de la población ancestral. En el caso de múltiples alelos, lo podemos expresar de la siguiente manera:

$$E(\mathbf{P}) = \begin{bmatrix} \mathbf{p}_{01}\mathbf{1} \\ \mathbf{p}_{02}\mathbf{1} \\ \vdots \\ \mathbf{p}_{0A}\mathbf{1} \end{bmatrix} = \mathbf{p}_0 \otimes \mathbf{1} = \mathbf{P}_0$$

Donde $\mathbf{p}_{0i}$ es la frecuencia ancestral del alelo i, $\mathbf{p}_0$ es el vector de frecuencias ancestrales, y $\otimes$ representa el producto de Kronecker. La varianza de $\mathbf{P}$ será

$$\begin{aligned} \mathrm{V}(\mathbf{P}) &= \begin{bmatrix} \mathrm{Var}(p_1) & \dots & \mathrm{Cov}(p_1, p_A) \\ \vdots & \mathrm{Var}(p_i) & \vdots \\ \mathrm{Cov}(p_A, p_1) & \dots & \mathrm{Var}(p_A) \end{bmatrix} \\ &= \mathbf{B}_0 \otimes \mathcal{F} \end{aligned}$$

Siendo $\mathbf{B}_0 = \mathrm{diag}(\mathbf{p}_0) - \mathbf{p}_0\mathbf{p}_0^T$.

Entonces la versión multialélica del estadístico T_{FLK} es

$$T_{FLK} = (\mathbf{P} - \mathbf{P}_0)^T (\mathbf{B}_0 \otimes \mathcal{F})^{-1} (\mathbf{P} - \mathbf{P}_0) \tag{2.7}$$

Con

$$E(T_{FLK}) \approx (n-1)(A-1)$$

y

$$\mathrm{V}(T_{FLK}) \approx 2(n-1)(A-1)$$

la versión multialélica de T_{FLK} sigue aproximadamente una distribución $\chi^2_{(n-1)(A-1)}$ bajo la hipótesis nula de que las frecuencias alélicas sólo están determinadas por deriva (Bonhomme et al. (2010) Apéndice B).

2.2.3. Test hapFLK

El test hapFLK aprovecha información sobre la tendencia a que las frecuencias alélicas de loci adyacentes en el genoma se encuentren correlacionadas (desequilibrio de ligamiento). A medida que se da un evento de selección direccional, la variante adaptativa aumenta su frecuencia. Los alelos segregantes de otros loci que inicialmente se encontraban cercanos a la variante adaptativa aumentan en frecuencia a pesar de ser neutrales. Éste fenomeno conocido como autoestop genético (Sección 1.3.1) resulta en tractos de desequilibrio de ligamiento alrededor de la variante adaptativa, y por consiguiente una reducción de la variabilidad genética en el mismo tracto.

Para aprovechar la información local alrededor de cada locus, hapFLK primero realiza clustering de los tractos genómicos presentes en las poblaciones, "pintando" los genomas de K colores (clusters) diferentes implementando

el algoritmo fastPHASE (Scheet and Stephens 2006) (Figura 2.5). El algoritmo fastPHASE es capaz de modelar información que aprovecha el desequilibrio de ligamiento ya que se basa en la implementacion de un modelo de Markov escondido (Rabiner and Juang 1986) en el cual los estados escondidos posibles son los K clusters, mientras que las observaciones son los alelos presentes en cada locus.

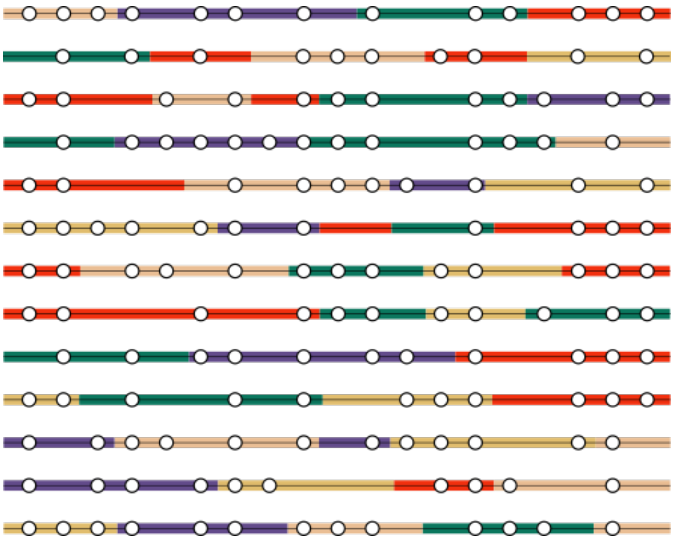

Figura 2.5: Ejemplo de haplotipos clasificados con cinco clusters. Cada linea negra representa la misma región genómica de un individuo. Los círculos blancos representan la presencia de una variante en un locus segregante. Los colores indican la clasificación realizada para un individuo dado en una región genómica dada.

Una vez implementado el algoritmo, a un locus dado en un individuo dado se le asignan las probabilidades de pertenecer a uno de los K clusters para obtener un vector

$$\mathbf{q}^l = \begin{bmatrix} q^l_{11} \\ \vdots \\ q^l_{1n} \\ \vdots \\ q^l_{K1} \\ \vdots \\ q^l_{Kn} \end{bmatrix}$$

donde q^l_{ij} es la probabilidad de que el locus l pertenezca al cluster i en el individuo j. Este vector de probabilidades de pertenencia es tratado como un vector de marcadores multialélicos, y el estadístico hapFLK es calculado como una versión multialélica de FLK (ecuación (2.7)), con una modificación ($\mathbf{B}_0$ es reemplazada por la matriz identidad $\mathcal{I}$ para mayor estabilidad numérica):

$$T_{hapFLK} = (\mathbf{q} - \mathbf{q}_0)^T (\mathcal{I} \otimes \mathcal{F})^{-1} (\mathbf{q} - \mathbf{q}_0) \tag{2.8}$$

Se usó el software hapFLK (Fariello et al. 2013) para calcular los estadísticos hapFLK y FLK. En el caso de la matriz de kinship y empírica, se utilizaron 3000 SNPs para calcular las distancias de Reynolds (`reynolds-snps`). Siempre fue especificada la población `p1` como *outgroup*. El número de *clusters* especificado fue 10 (`K`) y el número de corridas de EM fue 10 (`nfit`).

```
hapflk --ncpu ${task.cpus} \
       --reynolds-snps 3000 \
       --bfile genotypes_${rep_id} \
       --outgroup p1 \
```

```
        -K 10 \
        --nfit 10
```

2.3. Simulaciones

Distintos escenarios demográficos fueron simulados con SLiM (B. C. Haller and Messer 2019). SLiM es un *framework* de simulación *forward-in-time*, es decir, que simula la dinámica de los genomas de individuos a partir de una población inicial y luego somete los individuos a distintos procesos demográficos, adaptativos y ecológicos al pasar las generaciones. Este tipo de simulador permite mayor flexibilidad en cuanto a la complejidad de los procesos evolutivos que se quieren simular en comparación con simuladores basados en el coalescente (Wakeley 2009). En particular SLiM permite obtener las genealogías de genes (Kelleher et al. 2018) y otra información útil a lo largo de la simulación.

En todos los casos las simulaciones se hacen bajo un modelo WF con tamaño poblacional N constante y sin solapamiento de generaciones, en una única región genómica (un cromosoma). Como esquema general, las simulaciones comienzan con lo que es conocido como *burn-in*; se simulan varias generaciones para generar variabilidad genética a lo largo del cromosoma en una única población en condiciones no adaptativas.

Luego de completar el *burn-in*, se modifica la tasa mutacional tal que $\mu = 0$ y se comienzan a simular los *splits* que dan origen a nuevas poblaciones a partir de las preexistentes. Una vez establecidas todas las poblaciones deseadas, se simula un pulso migratorio de una población a otra para generar una población mezclada. Luego de que se da este pulso, se modifica el *fitness* de un único alelo de un polimorfismo, tal que su coeficiente de selección $s = 0{,}1$ y su coeficiente de dominancia $h = 0{,}5$. El momento (en generaciones) en los cuales se da el pulso migratorio y la modificación del *fitness* es específico para cada escenario simulado.

Se simuló un único cromosoma de 10 millones de bases (Mb) con una tasa de mutación inicial $\mu = 10^{-7}$ mutaciones por par de base por generación y una tasa de recombinación $\rho = 10^{-7}$, interpretado como la probabilidad de recombinación entre dos bases adyacentes en una generación (y en un genoma dado). Se mantuvo un $N = 1000$ constante a lo largo de todas las generaciones y en todas las poblaciones. El *burn-in* fue de 5000 generaciones. Como *output* de cada simulación se obtuvieron archivos VCF (Danecek et al. et al. 2011) correspondientes a una muestra de 50 individuos de cada población al final de la simulación, y tablas que contienen la frecuencia alélica de la variante bajo selección para cada población en cada generación.

Se realizaron dos grupos de simulaciones con distintos objetivos.

Primero se realizaron simulaciones con un número bajo de réplicas (100) para observar el comportamiento empírico del estadístico hapFLK bajo H_0 y decidir el mejor método para determinar significancia. A éste grupo se le llamó Calibración, y se realizó con dos escenarios demográficos (Figura 2.6).

En el segundo grupo se realizaron simulaciones en un escenario demográfico concreto (Figura 2.7) con un mayor número de réplicas (1000) para determinar el poder de hapFLK usando las distintas matrices de covarianza. A este grupo se le llamó Evaluación (Figura 2.7). En el caso de $s = 0{,}1$ se descartaron aquellas réplicas cuya frecuencia de `m2` al comienzo de la selección fuera mayor a 0.4, dejando aproximadamente 500 réplicas.

2.3.1. Calibración

Se realizaron simulaciones bajo los dos escenarios demográficos ilustrados en la Figura 2.6. En el escenario más simple (escenario A), se simularon *splits* que dieron lugar a cuatro poblaciones. Se simularon escenarios con y sin pulso de migración ($m \in \{0; 0{,}3\}$), y bajo neutralidad y selección ($s \in \{0; 0{,}1\}$). El escenario B es un escenario un poco más complejo, dando lugar a seis poblaciones. Se simuló con los mismos parámetros de migración y coeficiente de selección. El resultado final es un set de datos de 800 simulaciones independientes (400 para cada escenario).

En el escenario A el pulso de migración se da en la generación 570 desde `p2` hacia `p4`. En la misma generación la población mezclada (`p4`) entra en un régimen selectivo hasta la generación final 700. El escenario B es análogo, con el pulso de migración en la generación 770 y el régimen selectivo hasta la generación final 900.

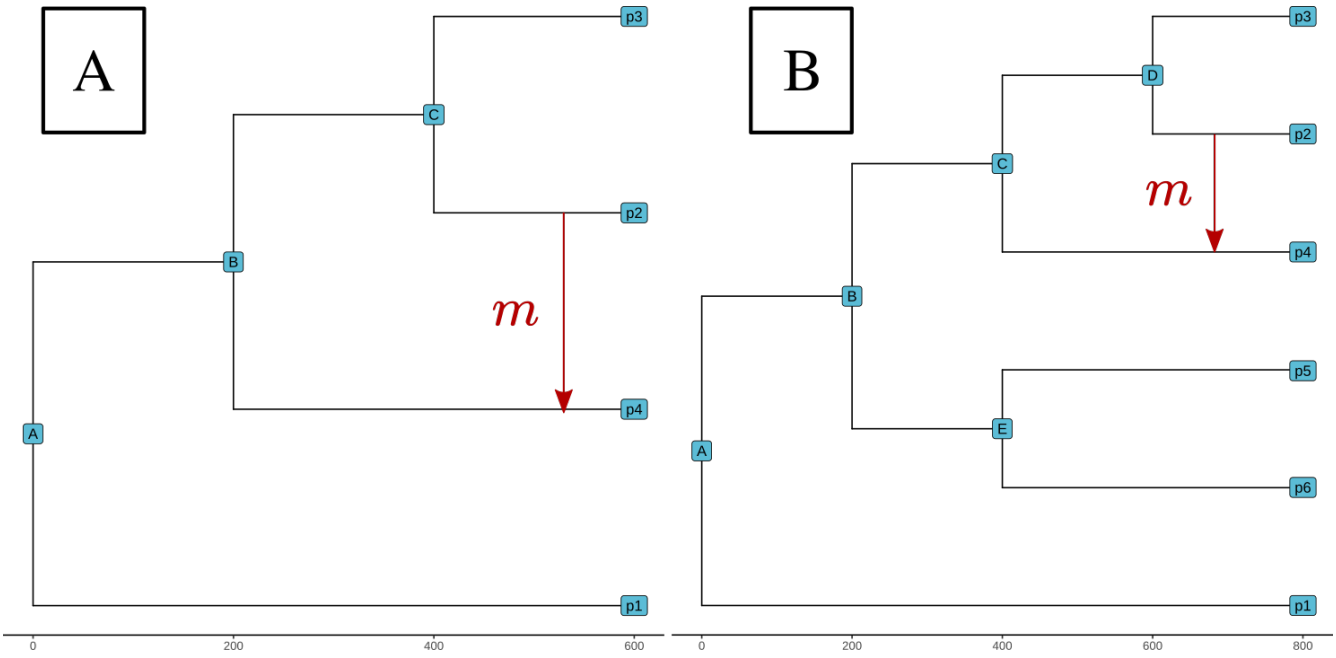

Figura 2.6: Escenarios demográficos de calibración.

2.3.2. Evaluación

Para el análisis de poder del estadístico, se realizaron simulaciones bajo un único escenario demográfico ilustrado en la Figura 2.7. Como resultado se obtuvieron datos genotípicos de cinco poblaciones, junto con las trayectorias de frecuencias alélicas del polimorfismo adaptativo. En este caso, la población `p5` es una población mezclada resultante de un pulso de *admixture* entre `p2` y `p4` en la generación 2500 con su correspondiente tasa de migración m. Se simuló el *admixture* con los parámetros $m \in \{0;0{,}3;0{,}7\}$ y coeficientes de selección $s \in \{0;0{,}1\}$. El régimen selectivo se establece desde la generación 2500 hasta la generación final 2600 en la población `p5`.

2.3.3. Control de establecimiento de mutación adaptativa

SLiM, al ser un simulador *forward*, permite simular los cambios en frecuencias alélicas bajo un modelo WF explícitamente. En la implementación usada en este trabajo, se optó por generar dos clases de mutaciones, una neutral (`m1`), las cuales son agregadas al genoma a una tasa $\mu = 10^{-7}$ en el período de *burn-in*, y una *potencialmente adaptativa* (`m2`). Ésta última clase va a ser inicialmente neutral para luego convertirse en adaptativa en una de las poblaciones y se encuentra en un único locus.

Esto es implementado de la siguiente manera: se realiza el *burn-in*, donde se generan mutaciones neutrales del tipo `m1`. En la generación t,se agrega la mutación *potencialmente adaptativa* del tipo `m2` en un sólo genoma (frecuencia $\frac{1}{2N}$). Luego, en $t + t_1$ se modifica μ tal que sea cero, y se da el primer *split* (fin del *burn-in*). Luego de estos eventos, la simulación sigue bajo un modelo WF sin mutación, con los *splits* correspondientes.

El hecho de que `m2` ingrese a la población con frecuencia $\frac{1}{2N}$ y sea inicialmente neutral hace que sea muy probable su pérdida en el período de tiempo entre las generaciones t y $t + t_1$. Si se perdiera la mutación, no podría ser seleccionada en futuras generaciones. Esto se soluciona agregando cláusulas al código de SLiM que condicionan a que `m2` no sea perdida, y que la simulación sólo continúe una vez se establezca `m2` con cierta frecuencia.

Una vez establecida `m2`, se produce el primer *split* y se deja de utilizar la cláusula, pero es necesario utilizar una cláusula adicional que evite la pérdida de `m2` en las poblaciones ancestrales de la población en la cual `m2` se va a volver adaptativa. De forma similar, el script de SLiM se escribe para controlar de que la mutación `m2` no sea perdida por deriva mientras sea neutral.

Para obtener resultados que sean biológicamente realistas, es esencial no agregar fuerzas evolutivas que causen el establecimiento o que eviten la pérdida de `m2`. Las cláusulas mencionadas funcionan de manera tal que, si se da la

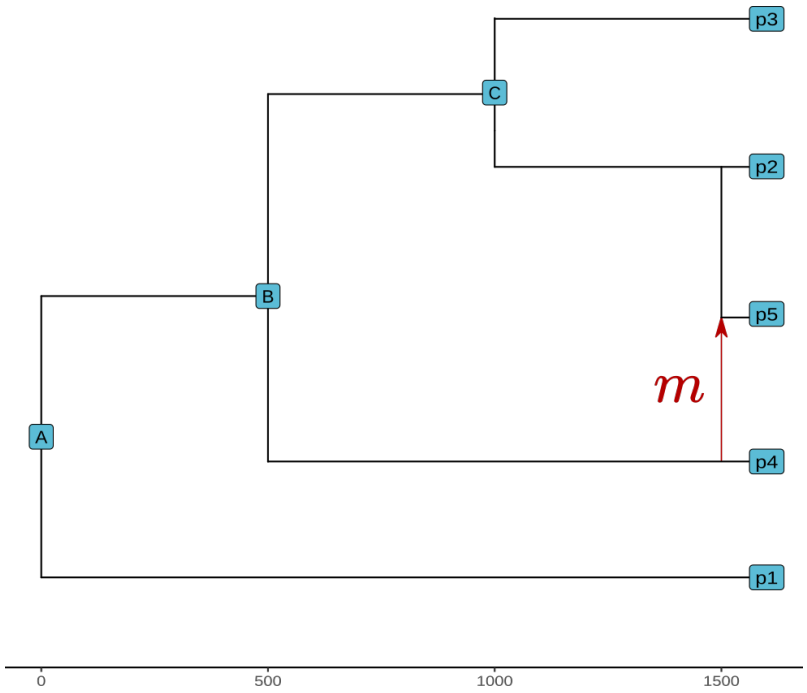

Figura 2.7: Escenario demográfico de evaluación.

condición de pérdida de `m2`, se vuelve a un *checkpoint*[2] en una generación previa en la simulación cuando todavía `m2` no fue perdida, y se comienza desde ese *checkpoint* con una nueva semilla aleatoria. Este proceso es repetido hasta que se logran trayectorias cuyo requerimiento es que exista la mutación `m2` en la población que se encuentra bajo selección.

El uso de este tipo de *checkpoints* genera algunos artefactos en las visualizaciones de las trayectorias de `m2` (por ejemplo en las Figuras 3.2 y 3.14). Esto se debe a como fue escrito el código de SLiM; para una réplica dada, cuando se vuelve a un *checkpoint* se pierde la información sobre la trayectoria pasada dado que las tablas de frecuencia de `m2` son reescritas.

2.3.4. Preprocesamiento de datos

El simulador SLiM permite devolver una muestra de los genotipos de los individuos en una generación dada en un archivo VCF. A partir de estos archivos, se tuvieron que realizar algunos pasos previos a su uso en hapFLK y TreeMix.

Combinación de VCFs: Se escribió y utilizó el script de python `merge-vcf.py` para combinar los VCF de cada una de las poblaciones del *output* de SLiM en un único VCF. El VCF resultante es converido al formato PLINK (Purcell et al. et al. 2007) para uso posterior.

Filtro de MAF. Se realizó un filtro de *Minor Allele Frequency*, que excluye del archivo BED a todos los SNPs cuyo MAF sea menor que 0.1. Se utilizó el comando de PLINK:

```
plink --bfile geno --maf 0.1 --nonfounders --make-bed --out genotypes_${rep_id}
```

Este filtro es necesario ya que loci con valores de MAF extremos afectan negativamente la estimación de las frecuencias ancestrales p_0, sumado a que en situaciones experimentales pueden ser confundidos con errores de secuenciación.

Filtro de LD. Para que TreeMix funcione de forma óptima, es necesario utilizar únicamente SNPs que no se encuentren en desequilibrio de ligamiento. Para ésto, se utilizó el comando de PLINK:

[2] Periódicamente se guarda toda la información de la simulación en disco.

```
plink --bfile genotypes_${rep_id} --indep-pairwise 100 50 0.1 -out ld
plink --bfile genotypes_${rep_id} \
        --extract ld.prune.in --make-bed \
        --out genotypes_${rep_id}_ldpruned
plink --bfile genotypes_${rep_id}_ldpruned --freq --family --out genotypes_${rep_id}
prepare-treemix.py genotypes_${rep_id}.frq.strat genotypes_${rep_id}.counts.gz
```

El script `prepare-treemix.py` fue escrito para convertir de formato PLINK a un formato específico de TreeMix.

Además de devolver los genotipos en formato VCF como *output*, los scripts de SLiM fueron realizados de tal manera que escriban la frecuencia del alelo *potencialmente adaptativo* `m2` en cada una de las poblaciones. Se escribieron y utilizaron los scripts `aggregate-frequencies.py`, que agrega las frecuencias de una simulación de SLiM, y `collect-frequencies.py`, que agrega las frecuencias de varias réplicas en un sólo archivo (Figura 2.8).

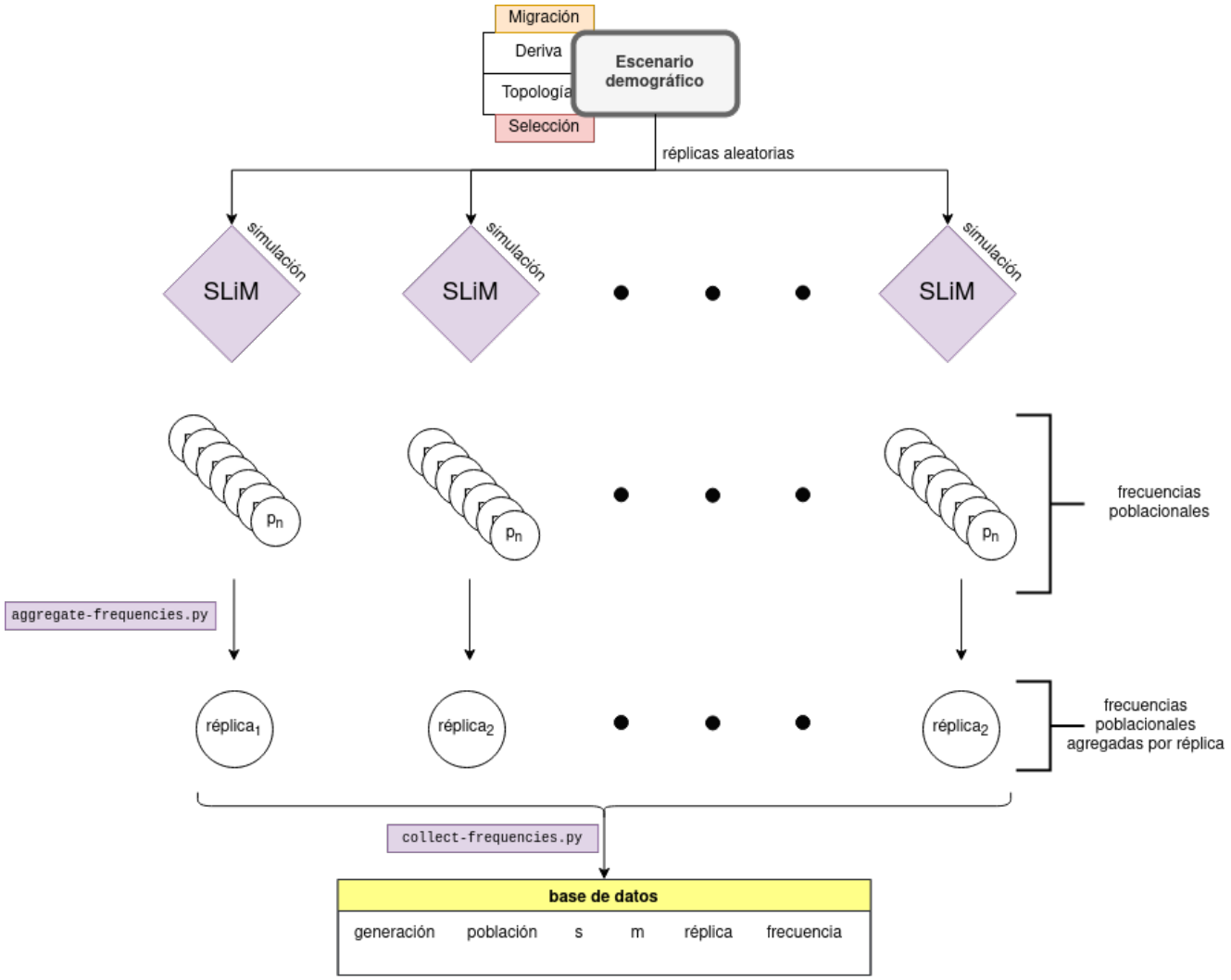

Figura 2.8: Pipeline de procesamiento de frecuencias del alelo adaptativo.

2.4. *Pipeline* de trabajo

El esquema general de trabajo está representado en la Figura 2.9. Se simularon datos genómicos bajo distintos escenarios demográficos con el *framework* de simulaciones SLiM. Los datos genómicos fueron preprocesados de acuerdo con la sección de Preprocesamiento. A partir de los datos genómicos procesados, se estimaron las matrices de covarianza entre poblaciones de cuatro maneras distintas:

- Matriz de covarianza teórica. Computada analíticamente a partir de los parámetros de la simulación (N, número de generaciones, y peso de migración m)(Sección 2.1.1).
- Matriz de *kinship*. Estimada a partir de las distancias de Reynolds en el árbol de poblaciones estimados por *Neighbor Joining* (Sección 2.1.2).
- Matriz de TreeMix. Estimada con el algoritmo de TreeMix (Pickrell and Pritchard 2012) usando un marco de máxima verosimilitud (Sección 2.1.3.1).
- Matriz de covarianza empírica. Estimada a partir de los vectores de frecuencias alélicas, y sus frecuencias alélicas ancestrales correspondientes (Sección 2.1.4).

En los casos de *kinship* y empírica, la matriz fue computada automáticamente por hapFLK. Las matrices de covarianza teórica y de TreeMix fueron proveídas al software hapFLK mediante la opción `--kinship`.

Una vez computados los estadísticos, se construyó una base de datos que contiene información sobre los valores de hapFLK para cada SNP, réplica, covarianza utilizada, y escenario demográfico. Además, se recolectaron las trayectorias de las frecuencias alélicas de las variantes adaptativas en las simulaciones.

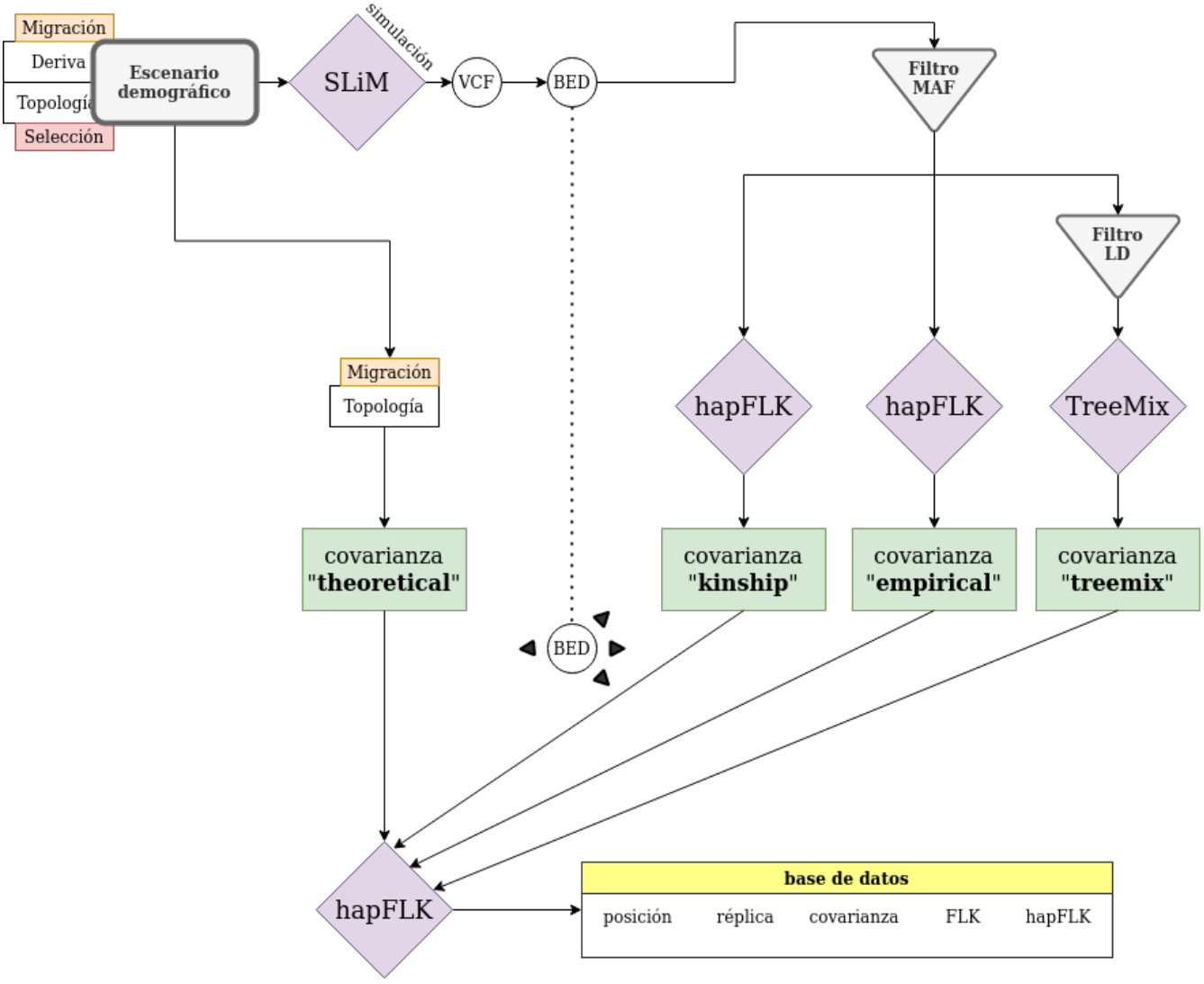

Figura 2.9: Pipeline de trabajo global.

Los análisis y visualizaciones fueron realizados con el software R y los paquetes

- `tidyverse` (Wickham 2017)

- `gridExtra` (Auguie 2017)
- `ggforce` (Pedersen 2021)
- `qqplotr` (Almeida, Loy, and Hofmann 2017)
- `ggtree` (Yu 2020)
- `GGally` (Schloerke et al. 2018)
- `MASS` (Venables and Ripley 2002)
- `wesanderson` (Ram and Wickham 2018)

El *pipeline* de trabajo fue organizado y computado con el lenguaje específico de dominio *Nextflow* (Di Tommaso et al. 2017), que permite un manejo eficiente, ordenado y reproducible de los cómputos. Se escribieron varios scripts de nextflow que manejan los programas utilizados (SLiM, PLINK, hapFLK), y los corren de forma concatenada y paralela (cuando fuese posible).

Los cálculos se realizaron en el servidor de computación del Institut Pasteur de Montevideo y en el cluster del Centro Nacional de Supercomputación (Nesmachnow S. 2019).

El trabajo fue dividido en dos *pipelines*, uno de simulación (`simulation.nf`) que realiza las simulaciones correspondientes a un script de SLiM, y reúne y preprocesa los datos generados (genotipos y frecuencias a lo largo del tiempo). El segundo *pipeline* (`analysis.nf`) toma los archivos BED de un directorio, en conjunto con otros parámetros, y realiza y agrega los cómputos de hapFLK usando las distintas matrices de covarianza.

2.5. Análisis

Se recolectaron los resultados de los estadísticos hapFLK y FLK para cada una de las réplicas y métodos de estimación de la covarianza usando el script `aggregate.py`, obteniendo bases de datos por escenario demográfico con las siguientes variables:

- **posición** de locus (pares de bases)
- identificador de **réplica** simulada
- método de estimación de **covarianza** usado
- estadístico **hapFLK**

Además, se recolectaron las frecuencias alélicas de `m2` a lo largo de las generaciones para cada simulación:

- **generación**
- **población**
- coeficiente de selección **s**
- coeficiente de *admixture* **m**
- **réplica**
- **frecuencia** de `m2`

2.5.1. Reescalado por regresión de cuantiles χ^2

Para obtener valores de hapFLK ajustados a la escala correspondiente de su distribución teórica $\chi^2_{(n-1)(K-1)}$ y realizar el test de hipótesis correspondiente, se realizó una regresión robusta (Venables and Ripley 2002) de los cuantiles teóricos sobre los cuantiles empíricos de distribuciones de hapFLK de cada réplica (por separado).

Sea $\mathbf{y}$ el vector de cuantiles teóricos para una distribución $\chi^2_{(n-1)(K-1)}$ y $\mathbf{x}$ el vector de cuantiles empíricos de hapFLK, se calculan los coeficientes de regresión lineal (robusta) a y b. Entonces, el vector $\hat{\mathbf{y}}$ de los valores hapFLK reescalados será

$$\hat{\mathbf{y}} = a\mathbf{x} + b$$

El vector $\hat{\mathbf{y}}$ resultante tendrá una media $(n-1)(K-1)$ y varianza $2(n-1)(K-1)$.

La implementación del reescalado se encuentra especificada en la Sección 2.6.2.

2.5.2. Cálculo de p-valores

2.5.3. Reescalado por distribución Normal

Se realizó un reescalado de los valores de hapFLK utilizando estimadores robustos de la media y desviación estándar. Sea $\mathbf{x}$ el vector de valores de hapFLK y $\bar{x}$ y σ_s los estimadores robustos de la media y desviación estandar, respectivamente, el vector escalado $\mathbf{z}$ de valores de hapFLK estará dado por:

$$\mathbf{z} = \frac{(\mathbf{x} - \bar{x}\mathbf{1})}{\sigma_s}$$

La implementación en R está especificada en la Sección 2.6.2.

2.5.4. Análisis de Poder

El poder va a ser calculado de forma empírica de la siguiente manera. Dado un escenario demográfico, se computa el valor máximo de hapFLK S^{max} según el valor de s. La distribución de S^{max} de $s = 0$ será la distribución bajo H_0, y la distribución de S^{max} de $s = 0{,}1$ será la distribución bajo H_1.

El poder del estadístico para un valor de error de Tipo I α es la proporción de réplicas bajo H_1 donde $S^{max} > q_\alpha$, siendo q_α el $(1 - q_\alpha)$-ésimo cuantil de la distribución de S^{max} bajo H_0 (Fariello et al. 2013).

El código que implementa el cálculo de poder y gráfico de curvas de poder se encuentra especificado en la Sección 2.6.2.

2.6. Código

El código desarrollado se encuentra en los repositorios:

- En calibration2 se encuentran los pipelines para la generación de datos en los escenarios demográficos de calibración.
- En evaluation se encuentran los pipelines para la generación de datos en los escenarios demográficos de evaluación.
- En masters-analysis se encuentran las jupyter notebooks que contienen el código del análisis de los datos.

2.6.1. Simulaciones

Las simulaciones se escribieron en el lenguaje Eidos (B. Haller 2016), y se encuentran en los repositorios mencionados anteriormente. Específicamente

- `simple_calibration.slim` es el script para simular el escenario demografico de calibración A.
- `complex_calibration.slim` simula el escenario demografico de calibración B.
- `simple_v2.slim` simula el escenario demográfico evaluación.

2.6.2. Análisis

2.6.2.1. Notebooks de análisis

Los análisis se realizaron con Jupyter Notebooks que se pueden encontrar en github. Tres notebooks principales contienen los pipelines:

- Calibración A
- Calibración B
- Evaluación

2.6.2.2. Funciones relevantes

qqreg.R Computa el reescalado por regresión de cuantiles.

normresc.R Computa el reescalado por distribución normal.

power.R Computa el poder, curvas ROC, y curvas de poder en función de la frecuencia de `m2` al comienzo de selección.

2.6.3. Otros

`merge-vcf.py` Une varios archivos VCF en un directorio. Es parte de un paquete de python llamado PEidos, que escribe código de Eidos para realizar simulaciones de SLiM en poblaciones jerárquicas de forma automática. El paquete se encuentra en desarrollo.

`aggregate-frequencies.py` Une archivos de salida de SLiM que contienen las frecuencias alélicas de `m2` en una población en un único archivo.

`collect-frequencies.py` Une archivos de frecuencias de `m2` de distintas réplicas en un único archivo.

`prepare-treemix.py` Modifica un archivo "`frq`" de PLINK para poder ser leído por el software TreeMix.

`prepare-modelcov.py` Modifica el archivo que contiene la matriz de covarianza ajustada (`*.modelcov.gz`) generada por TreeMix para que sea compatible con hapFLK.

`aggregate.py` Agrega resultados del computo de hapFLK de varias réplicas en un único archivo.

2.6.4. Documento

La compilación de este documento fue facilitada con el paquete de R `bookdown` (Xie 2016).

Capítulo 3

Resultados

Siguiendo la estructura general de los métodos, dividiremos los resultados en las secciones de Calibración y Evaluación. En Calibración se evaluó el comportamiento empírico del estadístico hapFLK estimado a partir de distintas matrices de covarianza, y se determinó la forma más adecuada para computar significancia a partir de dichas distribuciones (p-valores). En Evaluación, con más réplicas y bajo una única topología se exploró el efecto que tiene la presencia de *admixture* en el poder estadístico de hapFLK usando las distintas matrices de covarianza.

3.1. Calibración

Se controló que la frecuencia de la mutación adaptativa `m2` final fuera alta para asegurarse de que efectivamente tuvo lugar el *sweep*. Aquellas simulaciones donde la frecuencia de `m2` fuera menor que 0.5 cuando $s = 0{,}1$ fueron descartadas para el análisis subsiguiente (Figura 3.1).

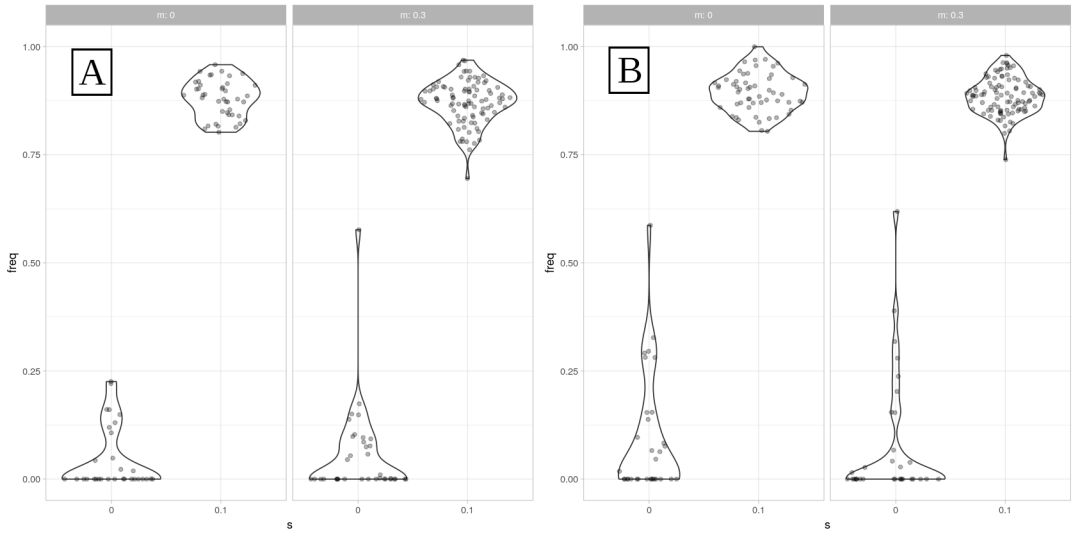

Figura 3.1: Frecuencias de m2 en la generación final. Izquierda, escenario A. Derecha, escenario B.

Al agregar las frecuencias de `m2` en todas las poblaciones a lo largo de las generaciones es posible visualizar el incremento en frecuencia de `m2` únicamente en `p4` cuando $s = 0{,}1$. Cabe destacar que en la generación final, y por ende la generación cuando son muestreados los genotipos, `m2` está mayoritariamente establecida, y el *sweep* está casi completo (Figuras 3.2, 3.3).

En la Tabla 3.1 se encuentra resumido el número de réplicas restantes en ambos escenarios luego del filtro mencionado arriba y el filtro de MAF. El número de loci segregantes al final de las simulaciones fue cercano a 7600 en ambos escenarios (Figura 3.4).

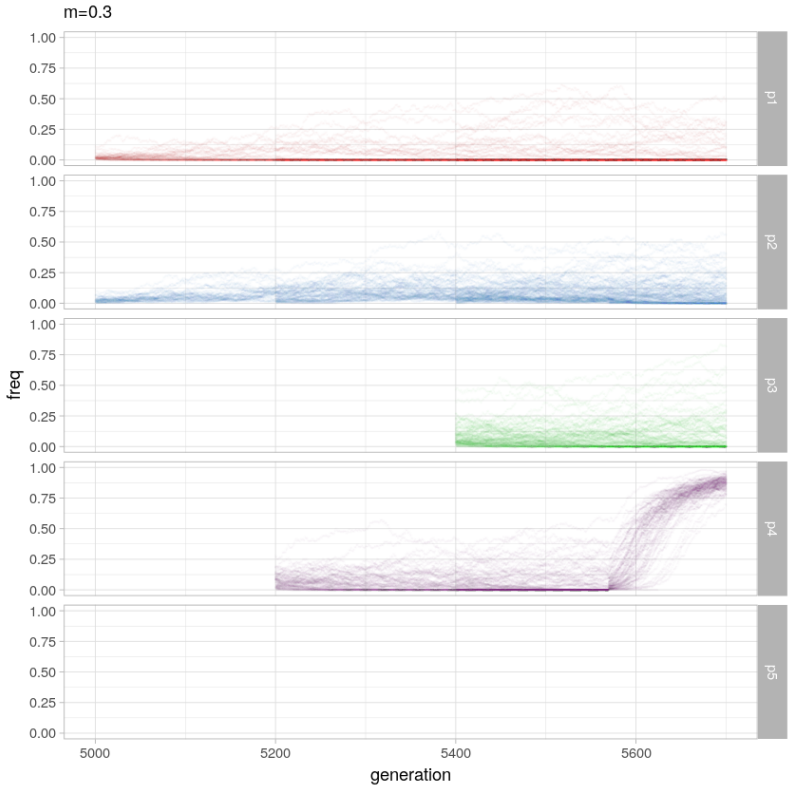

Figura 3.2: Trayectorias de frecuencias de m2 en el escenario A, bajo un régimen selectivo en 'p4'.

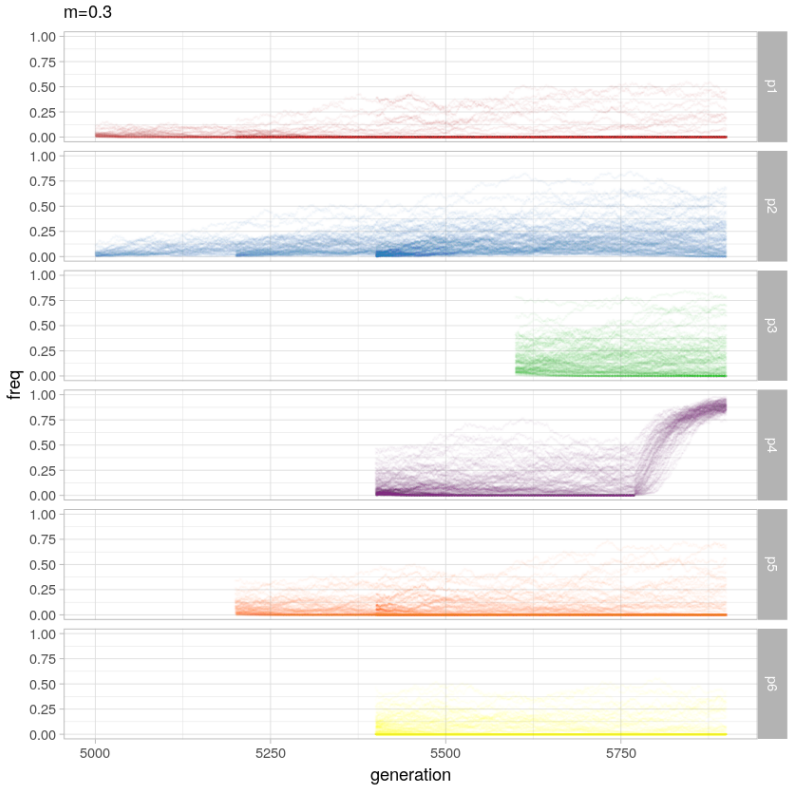

Figura 3.3: Trayectorias de frecuencias de m2 en el escenario B, bajo un régimen selectivo en 'p4'.

Tabla 3.1: Número de simulaciones realizadas para cada escenario según los parámetros s y m.

s	m	A	B
0.0	0.0	100	100
0.0	0.3	100	100
0.1	0.0	47	58
0.1	0.3	92	100

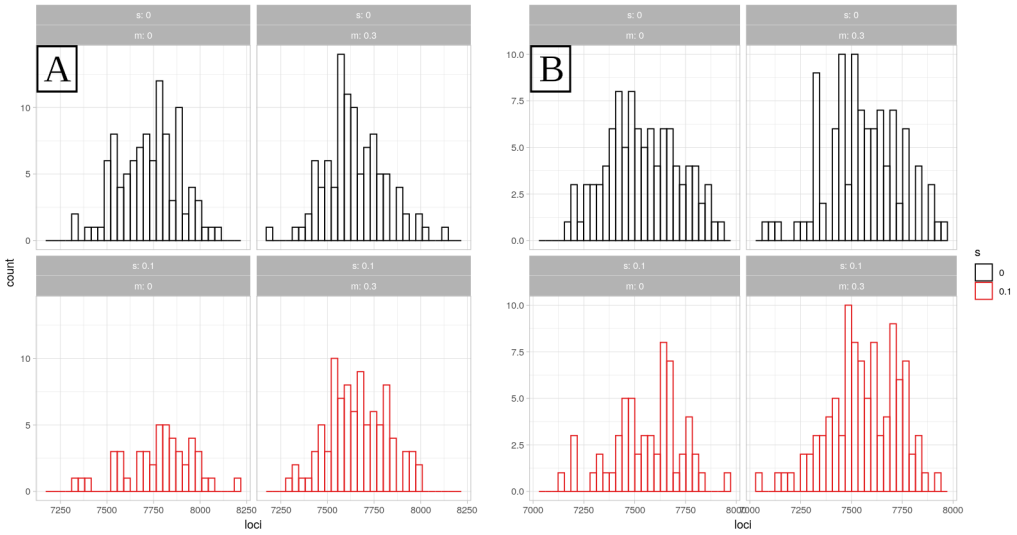

Figura 3.4: Histogramas de números de loci segregantes por réplica utilizados para el análisis. Escenario A (izquierda), escenario B (derecha).

Excepto por las réplicas donde `m2` no pasa el umbral de frecuencia de 0.5, la calidad de las simulaciones es buena. Ésto es indicado por los comportamientos esperados de las trayectorias de frecuencias, las frecuencias de `m2` en la generación final, y la uniformidad del número de loci segregantes por simulación.

3.1.1. Distribución de hapFLK

Una vez computados los valores de hapFLK, se graficaron histogramas agregados por réplicas y separados por método de estimación de covarianza. Podemos ver que tanto el escenario A como el B muestran distribuciones bien definidas, variando en el rango según el método de covarianza. Las distribuciones de valores estimados con las matrices `empirica`, `kinship` y `teorica` muestran medias y varianzas muy similares, mientras que la distribución estimada con `treemix` muestra media y varianza mayor. Éste patrón se mantiene para condiciones con y sin *admixture* ($m = 0$ y $m = 0{,}3$, respectivamente), para condiciones con y sin selección ($s = 0$ y $s = 0{,}1$, respectivamente), y para ambas topologías simuladas (Figuras 3.5 y 3.6).

Hay una sutil diferencia entre distribuciones cuando son comparadas las simulaciones neutras con las simulaciones bajo selección; se observan colas más pesadas con valores más extremos cuando $s = 0{,}1$. Ésto es esperable ya que bajo selección direccional tanto el alelo adaptativo como los polimorfismos cercanos aumentan en frecuencia, mostrando una reducción de diversidad genética (cuando es superado el valor 0.5 para SNPs) que resulta en valores altos de hapFLK.

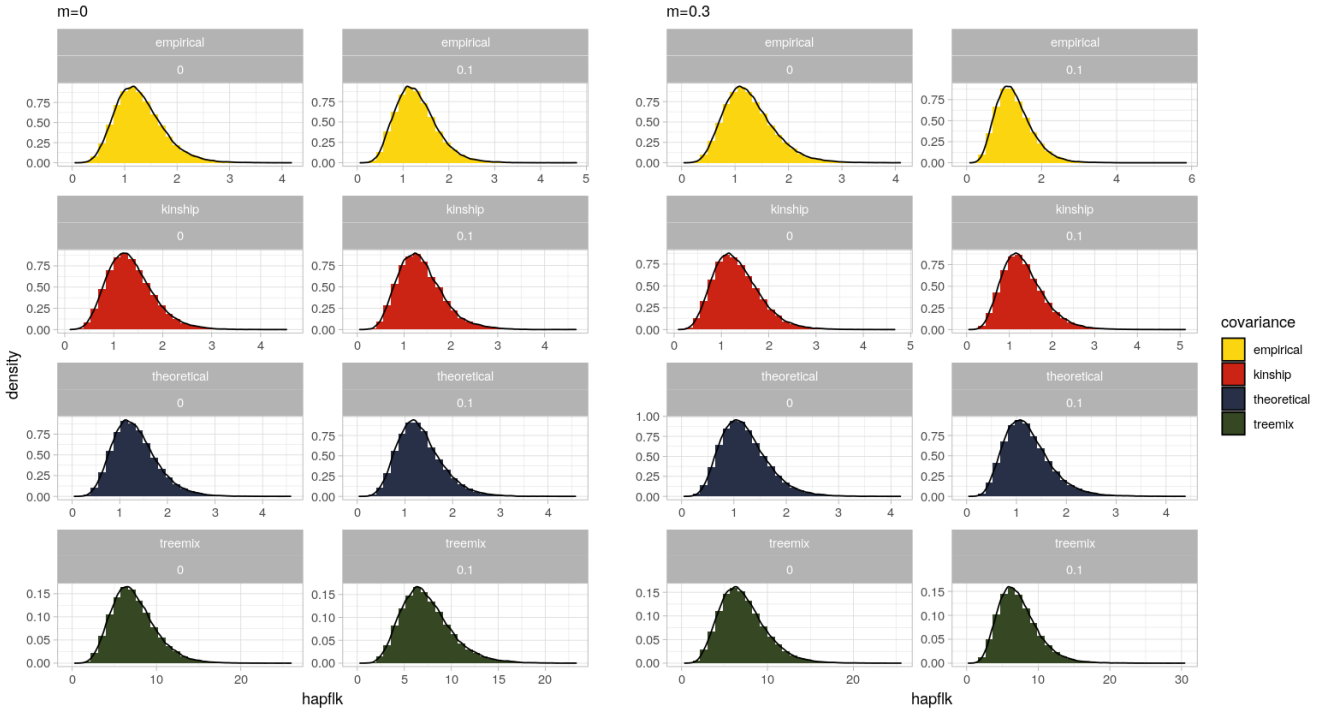

Figura 3.5: Distribuciones del estadístico hapFLK para el escenario A. Valores de coeficiente de *admixture* de $m = 0$ (izquerda) y $m = 0{,}3$ (derecha)

Al observar diagramas de puntos del estadístico en función de la posición genómica, se ve claramente un pico cercano a la posición de `m2` cuando $s = 0{,}1$, a diferencia de $s = 0$ que se observa una distribución relativamente uniforme a lo largo del genoma (Figura 3.7). En concordancia con los histogramas, el rango de valores de hapFLK estimados con la matriz de `treemix` es mayor que los rangos estimados con las otras matrices.

Si se grafican los valores de hapFLK estimados con las distintas matrices de covarianza en diagramas de puntos, se ve una correlación clara entre cada par de grupos. De acuerdo a lo visto en los histogramas, la estimación con `treemix` causa que los valores de hapFLK sean mayores (Figura 3.8).

Claramente la relación entre los valores de hapFLK estimados con `treemix` y con las otras matrices de covarianza es lineal, a pesar de cierta variabilidad. Ésto es consistente con el hecho de que el modelo subyacente de treemix es muy similar al de hapFLK, con la excepción de que la matriz de covarianza estimada con treemix va a estar escalada por

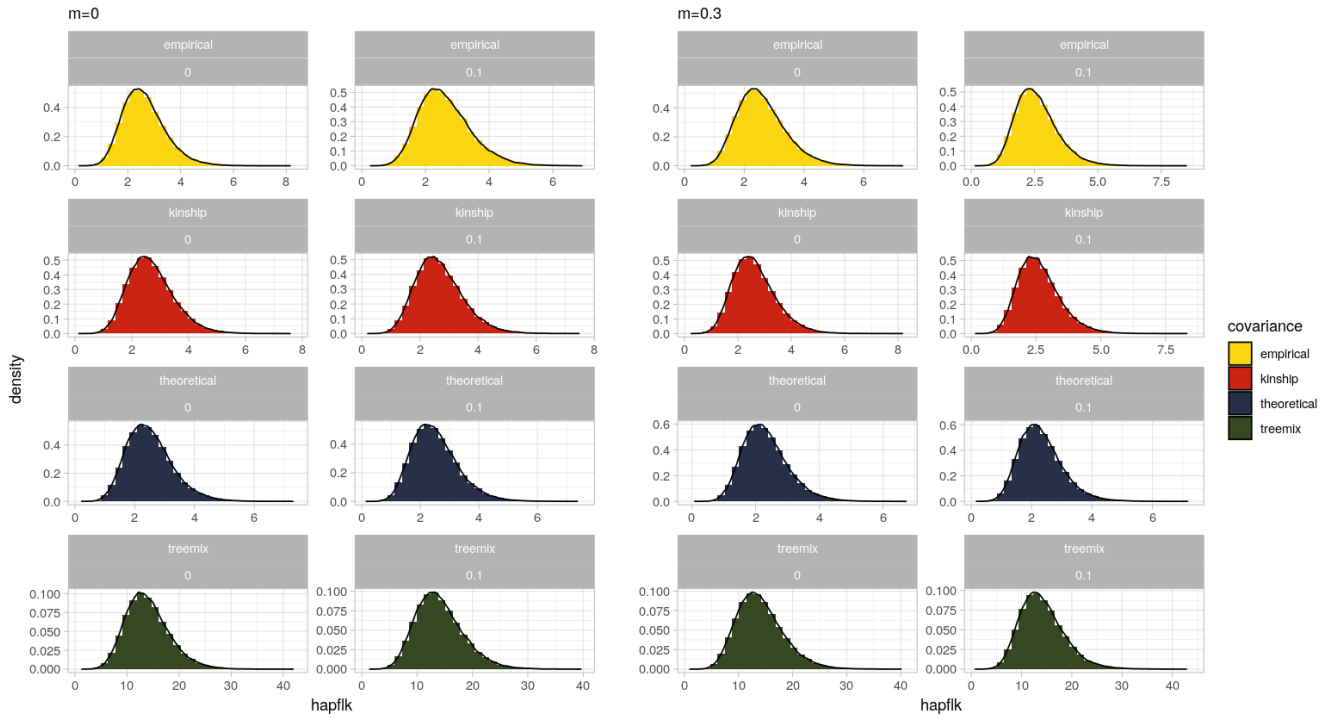

Figura 3.6: Distribuciones del estadístico hapFLK para el escenario A. Valores de coeficiente de *admixture* de $m = 0$ (izquierda) y $m = 0{,}3$ (derecha)

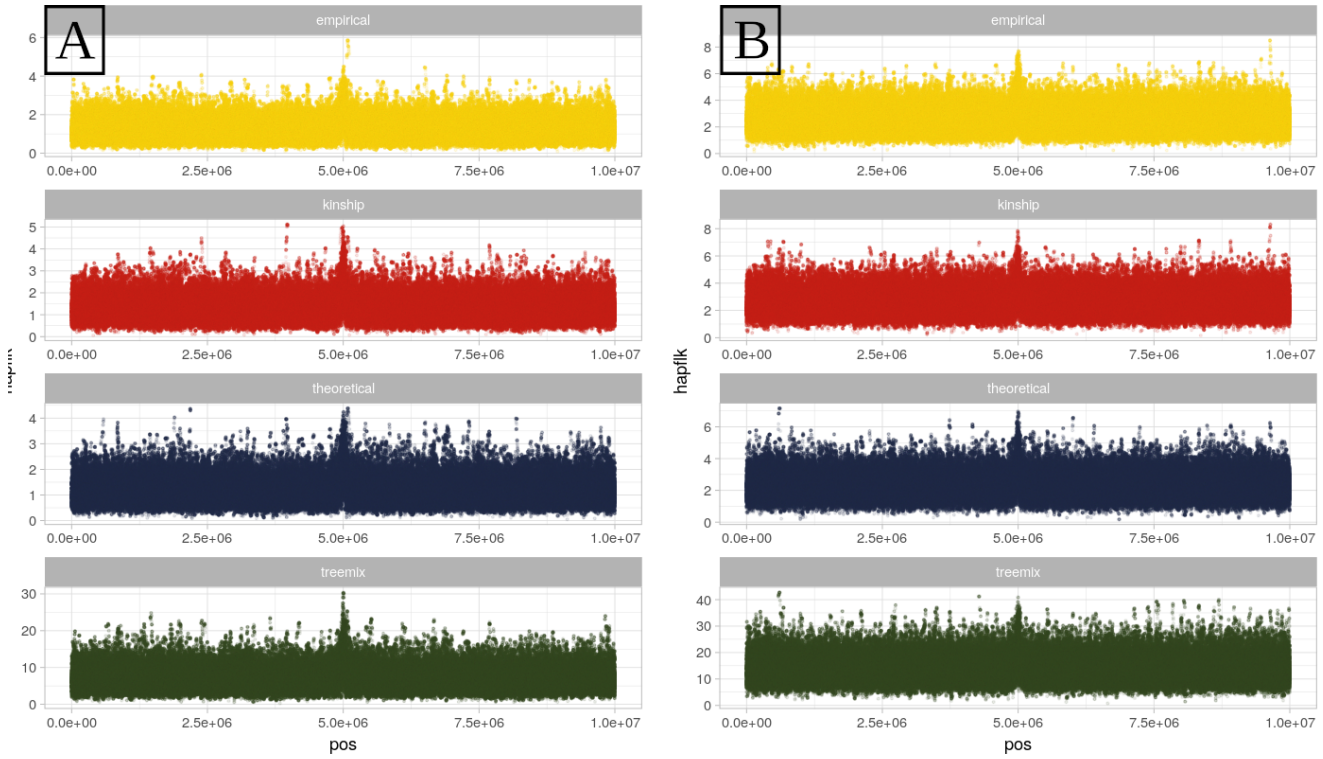

Figura 3.7: Valores de hapFLK en función de la posicion genómica con $s = 0{,}1$. Escenario A (izquierda), escenario B (derecha).

un factor $p_0(1-p_0)$, siendo p_0 la frecuencia ancestral de todas las poblaciones para un alelo dado (Ecuación (1.9)). En Pickrell and Pritchard (2012) hacen explícita la decisión de incluir las p_0 en el estimador de $\mathbf{W}$; argumentan que realizar cualquier tipo de normalización le brinda mayor peso a los loci con MAF extremas, que también son los loci para los cuales la aproximación de difusión se deja de cumplir.

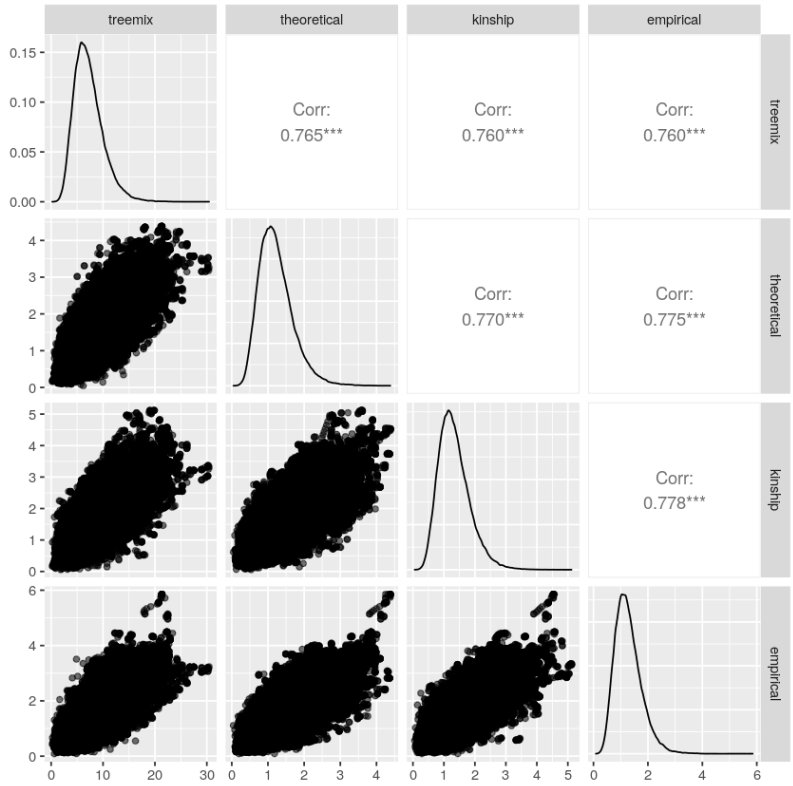

Figura 3.8: Diagramas de puntos y correlaciones de hapFLK estimado con distintas matrices de covarianza. Los datos de ésta figura son un ejemplo del escenario A de calibración, con $m = 0{,}3$ y $s = 0{,}1$.

3.1.2. Ajuste

El estadístico hapFLK sigue teóricamente una distribución $\chi^2_{(n-1)(K-1)}$ para n poblaciones y K clusters. En esta sección, $n = 4$ para el escenario A, y $n = 5$ para el escenario B, con $K = 10$. Con esta base, se crearon gráficos q-q para visualizar el ajuste de las distribuciones obtenidas con la distribución teórica χ^2.

En el escenario A, las distribuciones de hapFLK siguen aproximadamente la forma de una $\chi^2_{(18)}$, donde consistentemente se observan colas más pesadas que la distribución teórica. Esto es más pronunciado cuando $s = 0{,}1$, lo cual es esperable por los reducidos niveles de diversidad que causa el régimen selectivo. No se encuentran diferencias apreciables en presencia o ausencia de pulso de *admixture*, lo cual indica que no afecta la forma de la distribución (Figura 3.9).

De forma similar, en el escenario B, las distribuciones muestran colas pesadas comparadas con la distribución teórica, aunque en el caso de $m = 0$ este efecto se ve reducido para todos los métodos de estimación, y regimenes selectivos. De nuevo, la presencia de *admixture* no parecería afectar sustancialmente la forma de la distribución de hapFLK (Figura 3.10).

Para un alto número de grados de libertad, una distribución χ^2_k puede ser aproximada con una distribución normal $N(k, 4k^2)$. Suponiendo que los valores se podrían ajustar a una distribución normal, se utilizó la misma estrategia para visualizar el ajuste. Se crearon gráficos q-q con los cuantiles asociados a una normal $N(\hat{\mu}, \hat{\sigma^2})$ donde $\hat{\mu}$ y $\hat{\sigma^2}$ son la media y varianza muestral, respectivamente.

Tanto para el escenario A como para el escenario B, el ajuste visual a una distribución normal es malo. Ésto es evidente por las consistentes colas pesadas a la derecha de los gráficos q-q, y por el hecho de que la distribución de

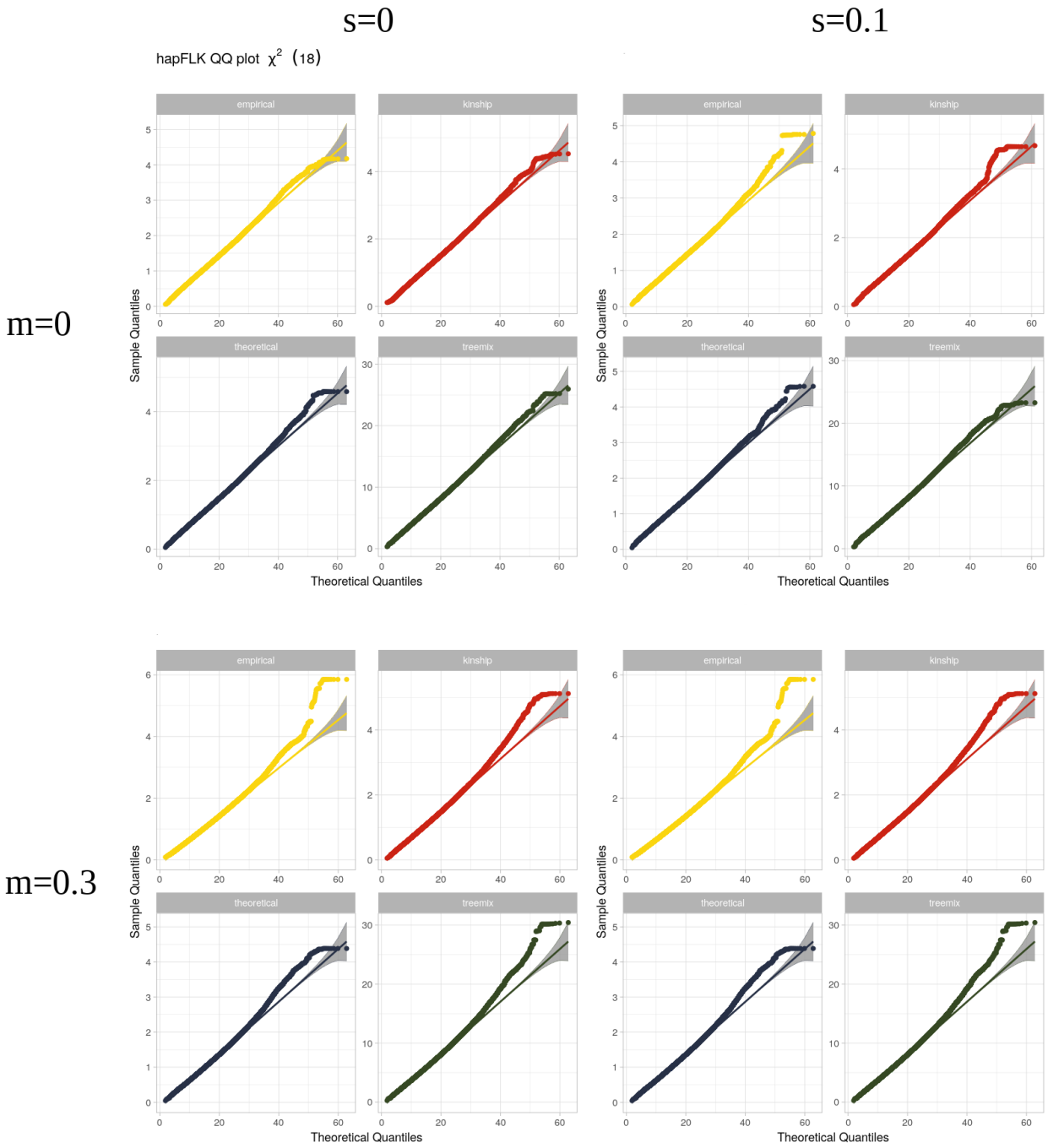

Figura 3.9: Gráficos q-q para el escenario A. Las columnas indican el valor de s, las filas el de m. En las abcisas están los cuantiles para una distribución chi cuadrado con 18 grados de libertad, en las ordenadas los cuantiles de los valores de hapFLK. En amarillo, la distribucion estimada con la covarianza empírica, en rojo kinship, en azul teórica, y en verde estimación por treemix.

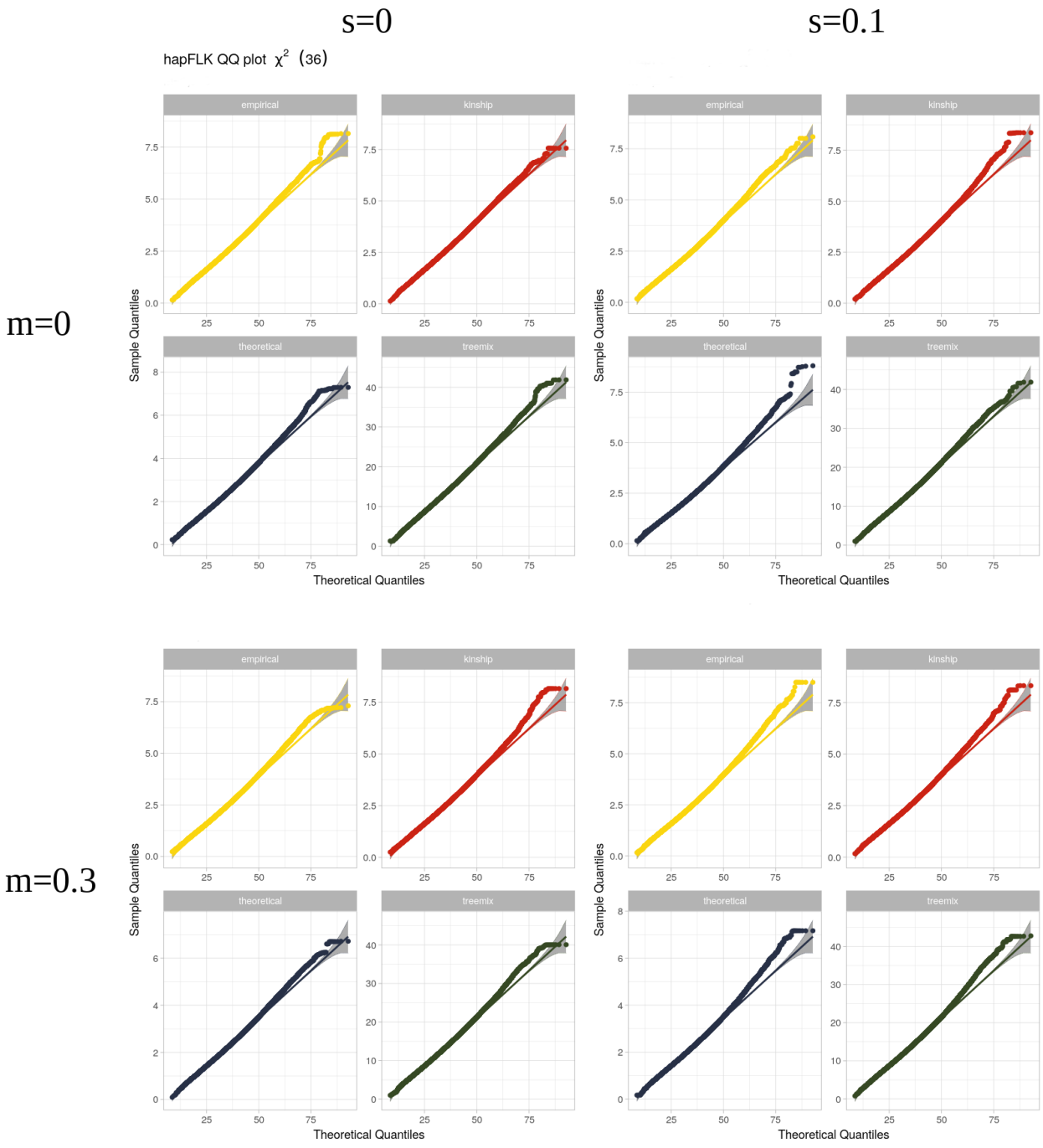

Figura 3.10: Gráficos q-q para el escenario B. Las columnas indican el valor de s, las filas el de m. En las abcisas están los cuantiles para una distribución chi cuadrado con 36 grados de libertad , en las ordenadas los cuantiles de los valores de hapFLK. En amarillo, la distribucion estimada con la covarianza empírica, en rojo kinship, en azul teórica, y en verde estimación por treemix.

hapFLK es por definición no negativa, lo que causa que valores se acumulen en la cola izquierda de las gráficas q-q, en el 0 de la distribucion teórica (Figuras A.1 y A.2).

3.1.3. Cálculo de significancia

Se decidió explorar el mejor método para calcular significancia a partir de los valores del estadístico hapFLK. Se evaluaron los métodos calculando p-valores de acuerdo a la Sección 2.5.2 y observando histogramas de sus distribuciones, donde se espera que sigan una distribución uniforme bajo H_0.

A pesar de mostrar un pobre ajuste a una distribución normal, primero se probó realizar test de hipótesis asumiendo que bajo H_0 hapFLK sigue una distribución normal. Los histogramas obtenidos muestran exceso de valores cercanos a cero en ambos escenarios, ademas de presentar un leve exceso de valores entre 0.75 y 1 (Figuras A.5 y A.6).

El hecho de que bajo selección existan valores extremos de hapFLK causa una inflación en los estimadores de la media y varianza. Se realizó también una estimación robusta de estos parámetros (Sección 2.5.3) para determinar si la presencia de datos atípicos estaba influyendo en el computo de p-valores. Las distribuciones de p-valores muestra que se ganaron valores cercanos a cero a expensas del exceso de valores entre 0.75 y 1 (Figuras A.7 y A.8).

Claramente, asumiendo que hapFLK sigue una distribución normal bajo H_0, los p-valores no siguen una distribución uniforme, lo cual indica que no es un método adecuado para cálculo de significancia. Quizás con un número mas alto de poblaciones la aproximación a una distribución normal sea mas adecuada.

A continuación se realizó el mismo procedimiento asumiendo que H_0 sigue una distribución $\chi^2_{(k)}$, donde $k = 18$ para el escenario A y $k = 36$ para el escenario B. Se tomaron dos estrategias para el cómputo de p-valores.

La primera fue calcular p-valores usando las distribuciones χ^2 con los grados de libertad teóricos. Es decir, 18 y 36. Esta aproximación no dio buenos resultados; las distribuciones de p-valores muestran una cantidad excesiva de valores cerca de uno para ambos escenarios (Figuras A.3 y A.4). Esto es de esperar: si se grafican las distribuciones de hapFLK junto con la distribución teórica $\chi^2_{(k)}$ es clara la falta de solapamiento (Figuras 3.11y A.11) y el corrimiento del estadístico hapFLK empírico hacia valores mas cercanos a cero, lo cual causa que la mayoría de los p-valores sean uno o cercanos a uno.

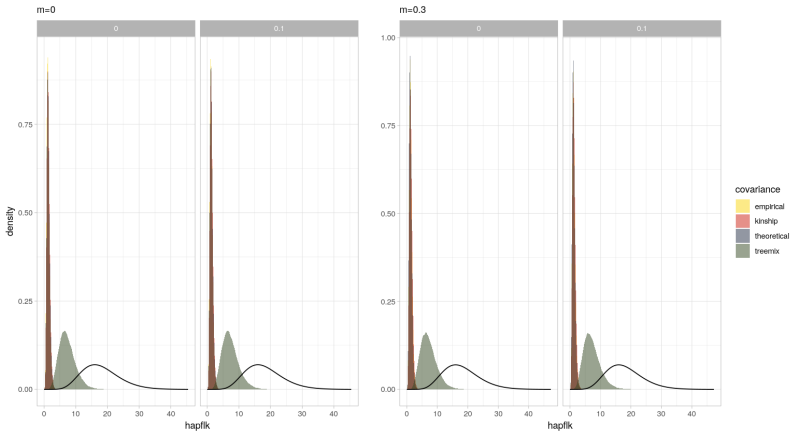

Figura 3.11: Densidades de estadístico hapFLK (curvas transparentes) junto con la densidad chi cuadrado teórica correspondiente para el escenario A.

La segunda estrategia fue reescalar los valores de hapFLK tal que las distribuciones tengan la media y varianza dadas por la distribución teórica $\chi^2_{(k)}$ (detalles en Sección 2.5.1). Esta estrategia fue la más apropiada para el cálculo de p-valores. Los histogramas resultantes muestran una distribucion uniforme bajo H_0 (cuando $s = 0$) (Figuras A.9 y A.10). Cabe destacar que también muestran una distribución uniforme cuando $s = 0{,}1$. Esto se debe a que los p-valores calculados provienen de valores de hapFLK en réplicas variadas con respecto al tipo de *sweep* que sufrió el alelo adaptativo; cuando las frecuencias alélicas de `m2` son bajas en `p4` al comienzo de la selección (*hard sweep*) el estadístico tiene mayor poder que cuando son altas (*soft sweep*)(ver Sección 3.2). En este caso, varias de las réplicas sufrieron un *soft sweep*, lo que causa una pérdida en el poder de detección y que se observe una distribución uniforme bajo H_1 (Figuras A.12 y A.13).

Tabla 3.2: Número de simulaciones realizadas en Evaluación para cada combinación de parámetros s y m.

s	m	replicas
0.0	0.0	1000
0.1	0.0	538
0.0	0.3	1000
0.1	0.3	626
0.0	0.7	1000
0.1	0.7	826

3.2. Evaluación

Una vez determinado el comportamiento empírico del estadístico hapFLK en presencia de poblaciones mezcladas, es de interés realizar un análisis del poder estadístico del algoritmo subyacente usando las distintas estimaciones de la matriz de covarianza. Fariello et al. (2013) mostraron que el poder de hapFLK para detectar selección era comparable a otros software contemporáneos, hasta en el caso de poblaciones donde sucedieron eventos de migración. En esta instancia se evaluó el poder estadístico de detección de loci bajo selección del algoritmo hapFLK en presencia de una población mezclada. Se compararó el poder del algoritmo provisto con las diferentes matrices de covarianza: `teórica`, `empírica`, `kinship` y `treemix`.

Se simuló el escenario demográfico especificado en la Sección 2.3.2 sin *admixture* ($m = 0$) y con *admixture* ($m = 0{,}3$,$m = 0{,}7$), bajo neutralidad $s = 0$, y bajo régimen selectivo en `p5` ($s = 0{,}1$). Para cada combinación de parámetros se realizaron 1000 réplicas para tener mayor confianza en los resultados del análisis de poder.

En los casos que $s = 0{,}1$ las simulaciones cuya frecuencia de `m2` fuese mayor a 0.4 al comienzo de la selección, o menores a 0.5 en la generación final fueron descartadas. En la Tabla 3.2 se encuentran resumidos el número de réplicas simuladas por conjunto de parámetros. Es posible ver que la calidad de las simulaciones es buena observando la frecuencia de `m2` en la generación final (Figura 3.12), y el número de loci segregantes por réplica (Figura 3.13).

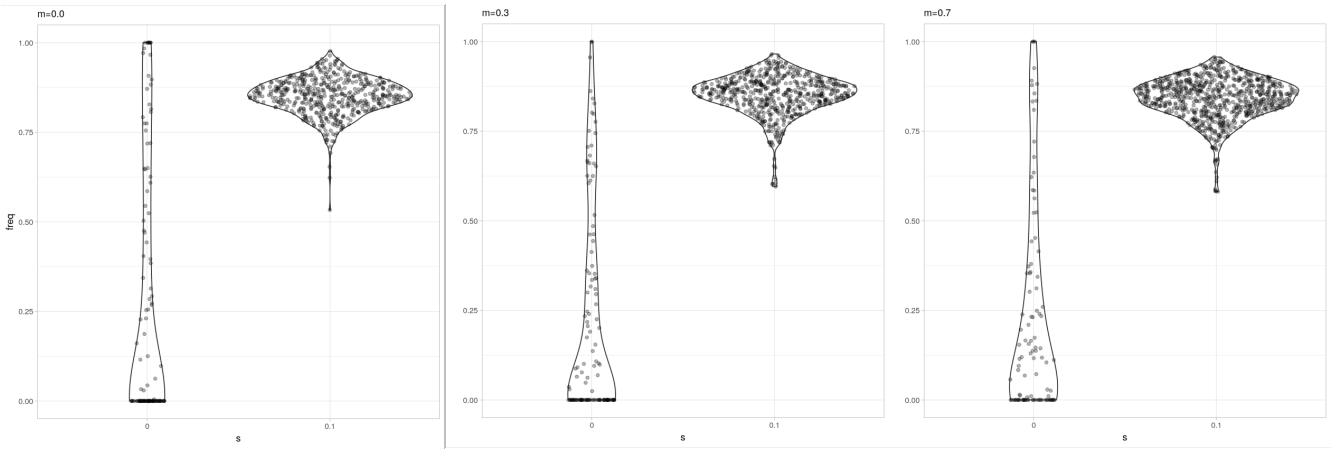

Figura 3.12: Frecuencias alélicas de m2 en la generación final de la simulación en Evaluación. Sin admixture (izquierda), $m = 0{,}3$ (centro), $m = 0{,}7$ (derecha).

De acuerdo a lo observado en las trayectorias de frecuencias alélicas (Figura 3.14), el *sweep* selectivo no logró terminar al llegar la última generación de la simulación. Debido a que el alelo adaptativo `m2` está en camino de fijación, la recombinación todavía no tiene un efecto sustancial en la formación de nuevos haplotipos cercanos a `m2`. Esta es una decisión intencional al momento de escribir la simulación; el poder de hapFLK será maximizado cuando la recombinación no ha roto los haplotipos asociados con `m2`.

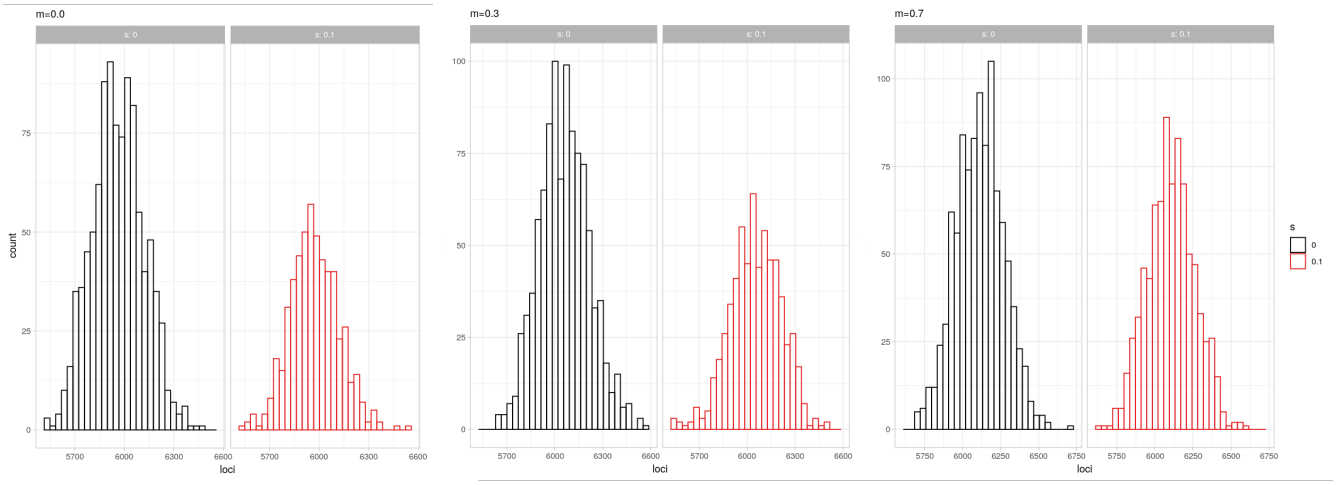

Figura 3.13: Distribución de número de loci segregantes al final de la simulación en Evaluación. Color negro indica que $s = 0$ y rojo $s = 0{,}1$. Sin admixture (izquierda), $m = 0{,}3$ (centro), $m = 0{,}7$ (derecha).

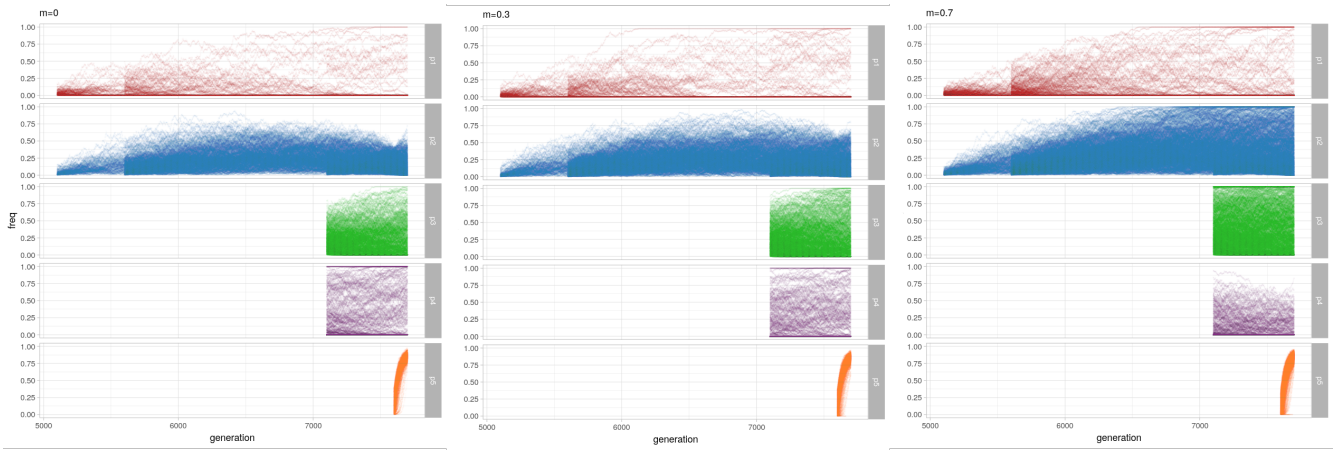

Figura 3.14: Trayectorias de frecuencias alélicas de m2 en simulaciones con $s = 0{,}1$ en Evaluación. Cada color corresponde cada una de las poblaciones simuladas. Sin admixture (izquierda), $m = 0{,}3$ (centro), $m = 0{,}7$ (derecha).

3.2.1. Análisis de poder

El cálculo de poder para un error de Tipo I α se encuentra especificado en la Sección 2.5.4. Para evaluar el poder de los distintos métodos de estimación de covarianza se graficaron curvas ROC condicionadas a ciertas frecuencias de la variante adaptativa `m2` en la generación donde comienza el régimen selectivo.

En primer lugar, para obtener una visión general, se computaron curvas ROC con aquellas réplicas que cumplan con: $f(\texttt{m2}) \leq 0{,}4$ al comienzo de la selección. Estas curvas corresponden a calcular el poder con réplicas donde se han dado *soft sweeps*. Podemos ver que independientemente del coeficiente de *admixture*, el poder es bajo para valores de α cercanos a 0.1. A pesar de esto, es notable la diferencia de poder para distintos valores de m.

Cuando no hay *admixture* ($m = 0$), se observa que los métodos de estimación por `kinship` y covarianza `empírica` se desempeñan de forma muy similar, siendo `treemix` el que muestra peor desempeño (Figura 3.15). El poder es menor cuando hay *admixture*, mostrando curvas ROC con áreas bajo la curva notablemente menores. En el caso de $m = 0{,}3$, la estimación empírica se desempeña mejor que los otros métodos, seguido en este caso de `treemix` (Figura 3.16). En el caso simétrico de $m = 0{,}7$, el poder de todos los métodos baja considerablemente, y se desempeñan de forma muy similar, con áreas bajo la curva similares (Figura 3.17).

Se evaluó el caso de un *hard sweep* calculando el poder usando réplicas donde $f(\texttt{m2}) \leq 0{,}1$ al comienzo de la selección. En ausencia de *admixture*, los métodos de `kinship` y `empírica` se desempeñan muy bien, obteniendo valores de poder comparables al *gold standard* (matriz de covarianza `teórica`), nuevamente, podemos ver que en ausencia de *amixture* la estimación por `treemix` es la que se desempeña peor (Figura 3.18).

En presencia de *admixure* las curvas ROC siguen el patrón general de lo visto anteriormente con el *soft sweep*; para $m = 0{,}3$ la estimación `empírica` se desempeña mejor, seguido de `treemix` y `kinship`, respectivamente. En el caso de $m = 0{,}7$ el poder decae, manteniendo el orden de desempeño; `empírica`, `treemix` y `kinship`.

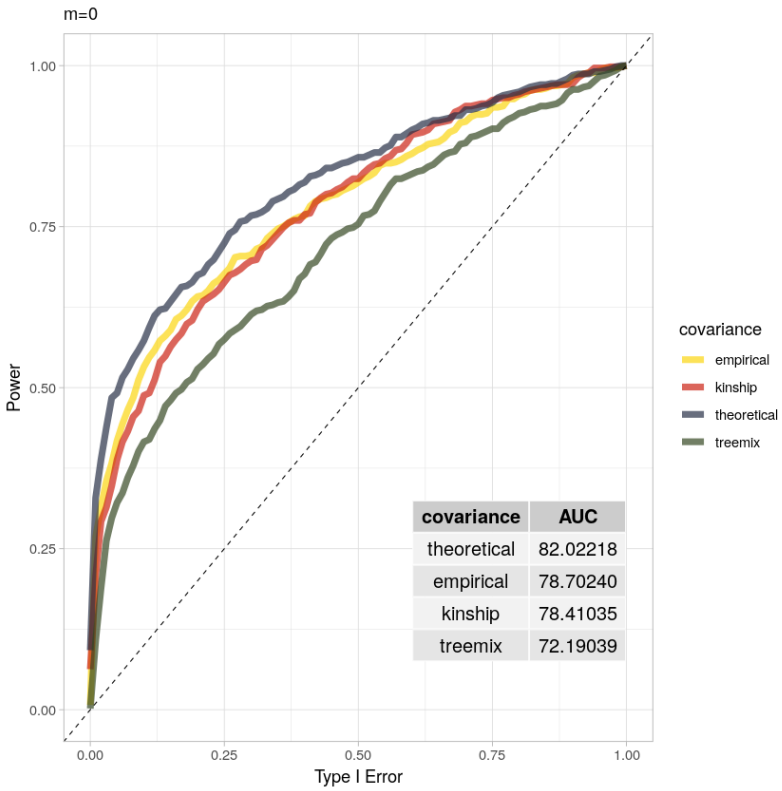

Figura 3.15: Curva ROC para. $m = 0$ y la frecuencia de 'm2' al comienzo de la selección es igual o menor a 0.4

Como se ha visto, la frecuencia alélica de la variante adaptativa al comienzo de selección va a afectar el poder de deteccion del estadístico, independientemente de la matriz de covarianza usada. Esto se debe a que en un *soft sweep*, las señales haplotípicas de selección que explota hapFLK son menores, ya que la variante adaptativa estaba asociada con varios haplotipos previo a la selección.

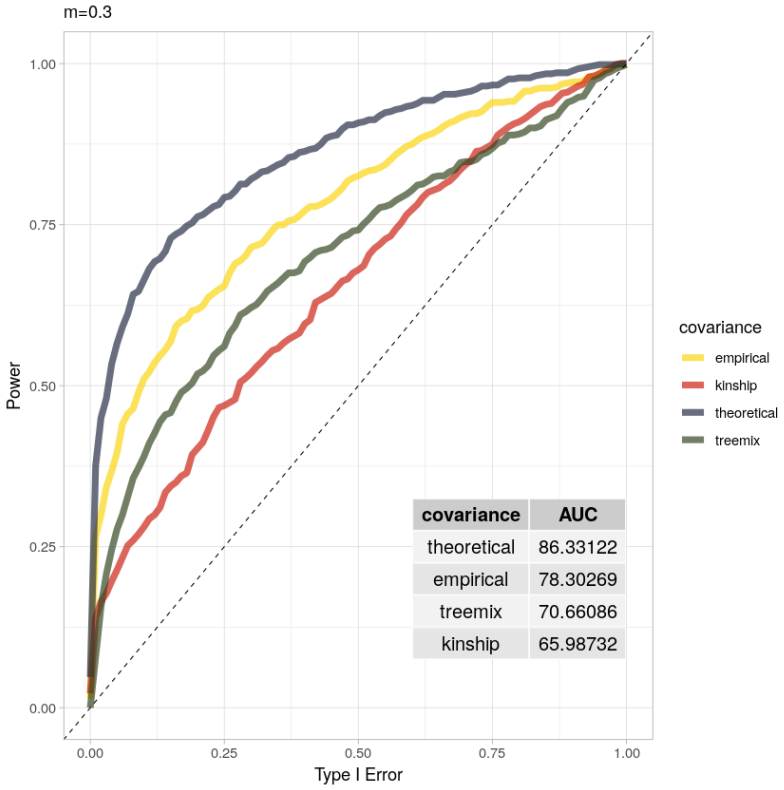

Figura 3.16: Curva ROC para. $m = 0{,}3$ y la frecuencia de 'm2' al comienzo de la selección es igual o menor a 0.4

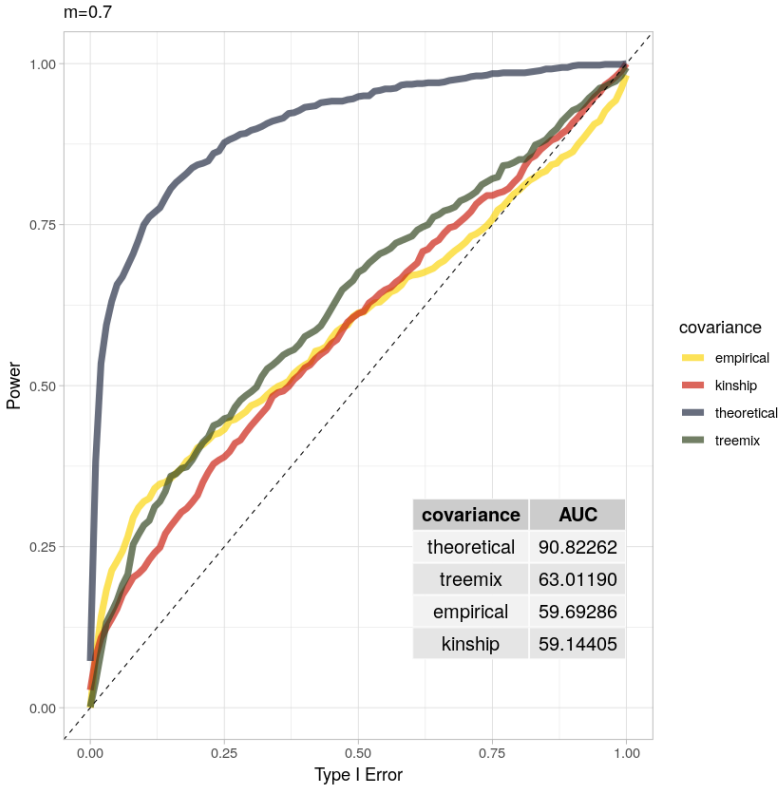

Figura 3.17: Curva ROC para. $m = 0{,}7$ y la frecuencia de 'm2' al comienzo de la selección es igual o menor a 0.4

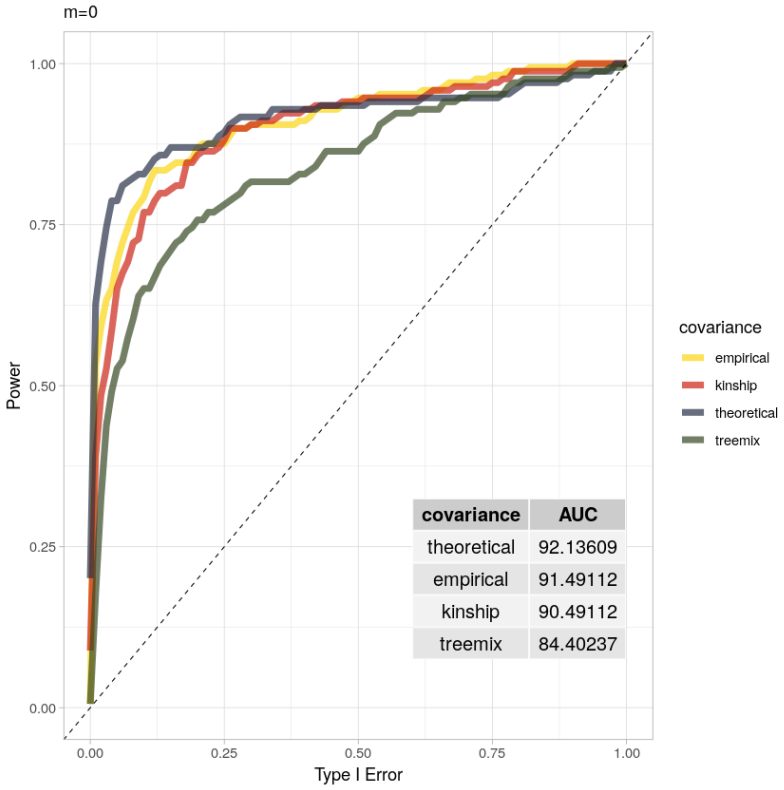

Figura 3.18: Curva ROC para. $m = 0$ y la frecuencia de 'm2' al comienzo de la selección es igual o menor a 0.1

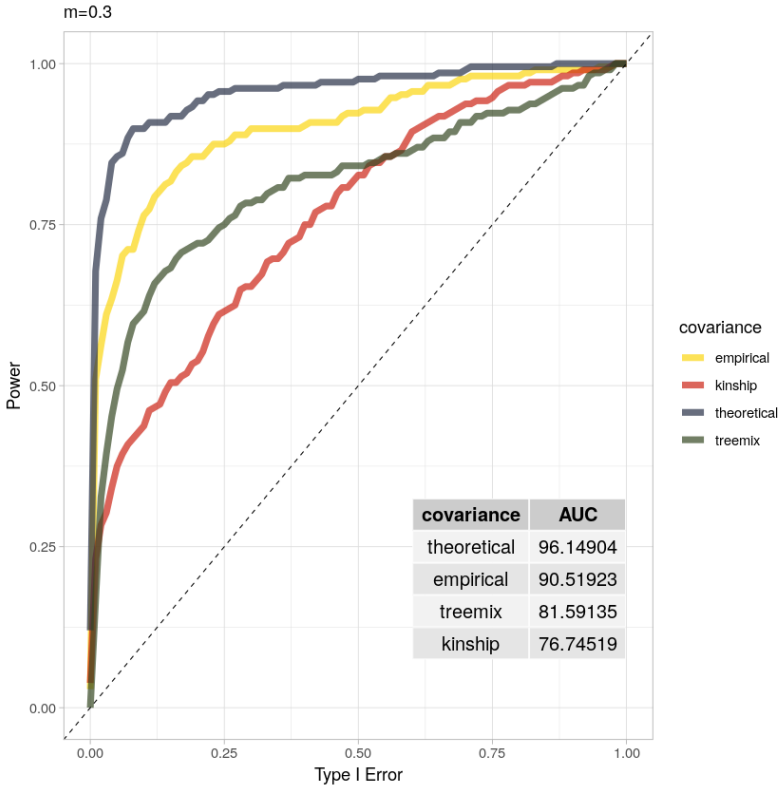

Figura 3.19: Curva ROC para. $m = 0{,}3$ y la frecuencia de 'm2' al comienzo de la selección es igual o menor a 0.1

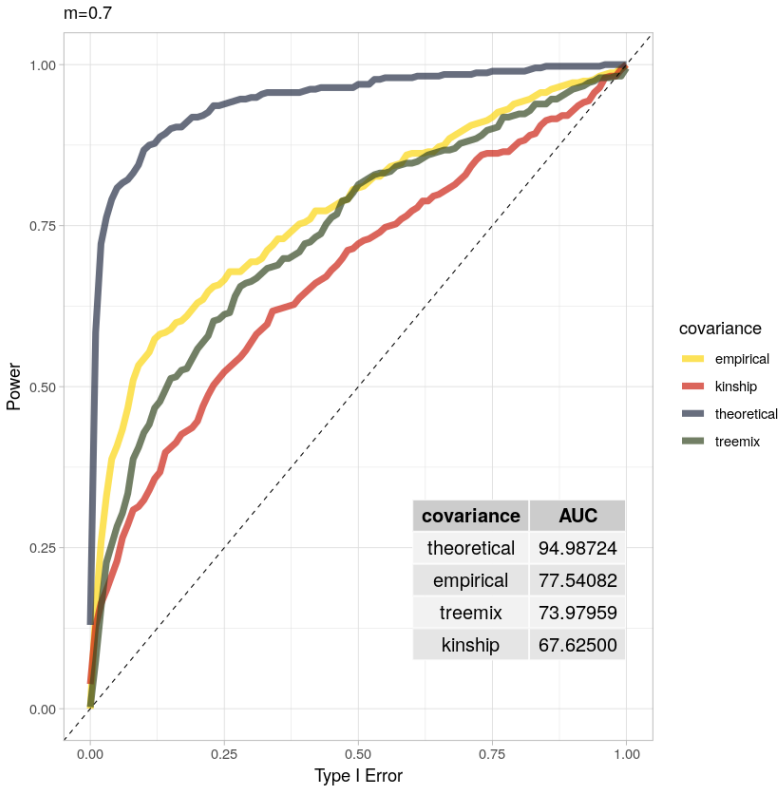

Figura 3.20: Curva ROC para. $m = 0{,}7$ y la frecuencia de 'm2' al comienzo de la selección es igual o menor a 0.1

Para saber en más detalle cómo la frecuencia alélica de `m2` al comienzo de la selección afecta el poder de detección, se evaluó el poder para un error de Tipo I $\alpha = 0{,}1$ en función de la frecuencia de `m2` al comienzo del régimen selectivo para los distintos valores de coeficiente de *admixture* m.

En la Figura 3.21 están resumidos los resultados de este análisis. Fue calculado el poder para un $\alpha = 0{,}1$ usando réplicas cuya frecuencia de `m2` al comienzo de la selección se encontrara en intervalos disjuntos. Los intervalos fueron: $I_1 = [0, 0{,}01)$, $I_2 = [0{,}01, 0{,}05)$, $I_3 = [0{,}05, 0{,}1)$, $I_4 = [0{,}1, 0{,}2)$, $I_5 = [0{,}2, 0{,}3)$, $I_6 = [0{,}3, 0{,}4]$.

Como es de esperar, a medida que la frecuencia al comienzo de selección aumenta, el poder disminuye. Cuando no hay *admixture*, todos los métodos de estimación se desempeñan de forma similar, salvo para los intervalos de frecuencia I_2 y I_4, donde `treemix` es el que peor se desempeña.

Cuando $m = 0{,}3$ se ve una distinción clara en el desempeño de los métodos, con la covarianza `empírica` desempeñandose mejor, seguido de `treemix` y `kinship`. En el caso de $m = 0{,}7$, en acuerdo con las curvas ROC, el poder de detección es muy bajo en rangos de frecuencia distintos a I_2, donde todos los métodos muestran muy bajo poder a lo largo de casi todos los rangos de frecuencias.

Cabe destacar, que bajo cualquier combinación de parámetros, la matriz de covarianza `teórica` se desempeña muy bien en términos de poder. Ésto es un indicador de que la mayor dificultad se encuentra en estimar de forma adecuada la matriz de covarianza en presencia de *admixture*.

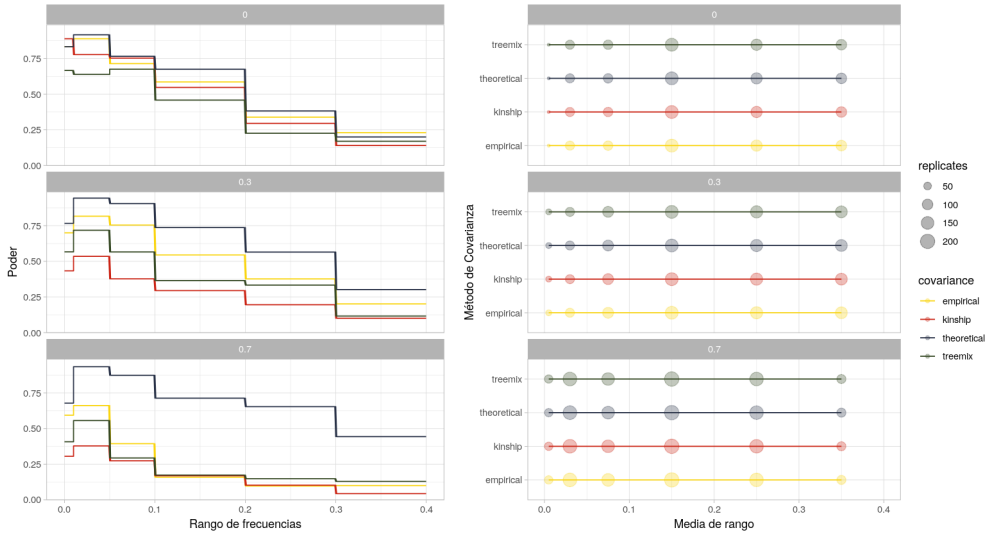

Figura 3.21: Poder para un error de Tipo I de 0.1 en función de la frecuencia al comienzo de la selección. Poder en función de los rangos de frecuencia que se usaron para calcularlo (izquierda). Número de réplicas usado para cada cálculo de poder, situado en el punto medio del rango utilizado (derecha).

Capítulo 4

Conclusiones

Se determinaron las distribuciones del estadístico hapFLK cuando son usadas las matrices de covarianza teórica, empírica, y estimada por treemix. De acuerdo a gráficos q-q, la distribución se ajusta a una χ^2, pero no corresponde con los grados de libertad teóricos $(n-1)(K-1)$. Las distribuciones calculadas con todas las matrices se muestran reescaladas con respecto a una $\chi^2_{(n-1)(K-1)}$ hacia valores cercanos a cero. En particular, la estimación con las matrices teórica, kinship y empírica parecerían sufrir el mismo escalado, mientras que la distribución de hapFLK con estimación con treemix se encuentra en valores intermedios entre las otras distribuciones y la distribución teórica. Esto último es de esperar ya la matriz de covarianza estimada por TreeMix no está normalizada con respecto a las frecuencias alélicas.

Se exploraron distintas estrategias para calcular significancia a partir de las distribuciones observadas. Calculando p-valores asumiendo una distribución normal para hapFLK, no dio buenos resultados, así como asumir una distribución $\chi^2_{(n-1)(K-1)}$, según lo evidenciado por histogramas de p-valores. La mejor estrategia fue realizar un reescalado de los valores de hapFLK a través de una regresión de cuantiles de la distribución teórica sobre la empírica, para luego computar p-valores de acuerdo a los grados de libertad teóricos. Ésta estrategia es la única que tuvo como resultado distribuciones de p-valores uniformes bajo la hipótesis nula de neutralidad.

El análisis de poder reveló que la frecuencia alélica al comienzo de la selección va a afectar drásticamente el poder de detección, independiente del método de esimación de covarianza utilizado. Esto es de esperar, ya que los *soft sweeps* son mas difíciles de detectar que los *hard sweeps* con métodos que usan el desequilibrio de ligamiento como hapFLK. Ésto ya había sido descrito en el articulo original de Fariello et al. (2013), pero los resultados fueron extendidos al uso de la matriz de covarianza empírica, y estimada por treemix, en presencia de migración.

Al comparar los métodos de estimación de covarianza, se observa un patrón consistente; en la población mezclada, el uso de la matriz de covarianza empírica brinda el mayor poder, seguido de la estimada por treemix, y kinship. En el caso de no haber *admixture*, treemix pasa al último lugar, seguida de la matriz de kinship y la matriz de covarianza empírica mostrando el mejor desempeño.

En todos los casos, el uso de la matriz de covarianza teórica mostró el mayor poder. En general, en situaciones experimentales, no es posible acceder a ésta información. Sin embargo, se puede concluir que la mayor dificultad se encuentra en la estimación adecuada de matrices de covariaza, ya que usar hapFLK con la matriz teórica dió buenos resultados.

El hecho de que el uso de la matriz de covarianza empírica muestre consistentemente un mayor poder que el uso de las otras matrices indica que ésta sería la opción preferida para utilizar en datos genómicos experimentales. Además, al mostrar mayor poder tanto en presencia como en ausencia de *admixture*, cuenta con la flexibilidad de poder ser utilizada en estudios preliminares de selección donde no se conoce con seguridad la historia demográfica subyacente.

En síntesis, la mejor práctica de uso sería utilizar la matriz de covarianza empírica reescalando los valores de hapFLK con una regresión de cuantiles, y calcular p-valores de acuerdo a los grados de libertad teóricos, y realizar ajuste por testeo múltiple.

Claramente, los resultados obtenidos están restringidos a los escenarios demográficos y parámetros simulados; un mayor número de réplicas en conjunto con escenarios más variados y realistas ayudarían a identificar aquellos procesos demográficos que pueden disminuir el poder de hapFLK. Por ejemplo, usando valores de s más realistas

($s \leq 0{,}01$) resultaría en una fijación del alelo adaptativo más lenta, generándose mayor cantidad de haplotipos asociados al alelo adaptativo, y potencialmente disminuyendo el poder de hapFLK.

Todavía es necesario determinar cual es el comportamiento del estadístico cuando se trata con datos de genotipado de genomas enteros; en este caso la distribución nula de hapFLK estaría menos sesgada por los valores altos de loci bajo selección, y podría mostrar mayor poder. Mas simulaciones en este contexto son necesarias para responder esta pregunta.

Se desarrolló software ad-hoc que permite la simulación arbitraria de demografías con el framework SLiM y postprocesamiento correspondiente usando el DSL nextflow. Este es un rico recurso para futuros estudios que usen SLiM como simulador, y en particular para el estudio de hapFLK. En conjunto con los scripts de que acompañan los *pipelines* de Nextflow, se construyó un marco sólido para trabajar datos de genética de poblaciones.

Capítulo 5

Referencias

Almeida, Alexandre, Adam Loy, and Heike Hofmann. 2017. *Qqplotr: Quantile-Quantile Plot Extensions for 'Ggplot2'.* https://github.com/aloy/qqplotr.

Auguie, Baptiste. 2017. *gridExtra: Miscellaneous Functions for "Grid"Graphics.* https://CRAN.R-project.org/package=gridExtra.

Bonhomme, Maxime, Claude Chevalet, Bertrand Servin, Simon Boitard, Jihad Abdallah, Sarah Blott, and Magali SanCristobal. 2010. "Detecting Selection in Population Trees: The Lewontin and Krakauer Test Extended," 32.

Cavalli-Sforza, L. L., and A. Piazza. 1975. "Analysis of Evolution: Evolutionary Rates, Independence and Treeness." *Theoretical Population Biology* 8 (2): 127–65. https://doi.org/10.1016/0040-5809(75)90029-5.

Cheng, Jade Yu, Fernando Racimo, and Rasmus Nielsen. 2019. "Ohana: Detecting Selection in Multiple Populations by Modelling Ancestral Admixture Components." Preprint. Bioinformatics. https://doi.org/10.1101/546408.

Cheng, Xiaoheng, Cheng Xu, and Michael DeGiorgio. 2017. "Fast and Robust Detection of Ancestral Selective Sweeps." *Molecular Ecology* 26 (24): 6871–91. https://doi.org/10.1111/mec.14416.

Coop, Graham, David Witonsky, Anna Di Rienzo, and Jonathan K Pritchard. 2010. "Using Environmental Correlations to Identify Loci Underlying Local Adaptation." *Genetics* 185 (4): 1411–23. https://doi.org/10.1534/genetics.110.114819.

Danecek, Petr, Adam Auton, Goncalo Abecasis, Cornelis A Albers, Eric Banks, Mark A DePristo, Robert E Handsaker, et al., et al. 2011. "The Variant Call Format and VCFtools." *Bioinformatics* 27 (15): 2156–58.

Di Tommaso, Paolo, Maria Chatzou, Evan W Floden, Pablo Prieto Barja, Emilio Palumbo, and Cedric Notredame. 2017. "Nextflow Enables Reproducible Computational Workflows." *Nature Biotechnology* 35 (4): 316–19. https://doi.org/10.1038/nbt.3820.

Excoffier, L, T Hofer, and M Foll. 2009. "Detecting Loci Under Selection in a Hierarchically Structured Population." *Heredity* 103 (4): 285–98. https://doi.org/10.1038/hdy.2009.74.

Fariello, María Inés, Simon Boitard, Hugo Naya, Magali SanCristobal, and Bertrand Servin. 2013. "Detecting Signatures of Selection Through Haplotype Differentiation Among Hierarchically Structured Populations." *Genetics* 193 (3): 929–41. https://doi.org/10.1534/genetics.112.147231.

Felsenstein, Joseph. 1981. "Evolutionary Trees from Gene Frequencies and Quantitative Characters: Finding Maximum Likelihood Estimates." *Evolution*, 1229–42.

———. 2005. "Theoretical Evolutionary Genetics Joseph Felsenstein." *University of Washington, Seattle.*

Fisher, Ronald Aylmer. 1958. *The Genetical Theory of Natural Selection.*

Gautier, Mathieu. 2015. "Genome-Wide Scan for Adaptive Divergence and Association with Population-Specific Covariates," 60.

Gillespie, John H. 2004. *Population Genetics: A Concise Guide.* JHU Press.

Günther, Torsten, and Graham Coop. 2013. "Robust Identification of Local Adaptation from Allele Frequencies." *Genetics* 195 (1): 205–20. https://doi.org/10.1534/genetics.113.152462.

Hahn, Matthew William. 2018. *Molecular Population Genetics.* Oxford University Press.

Haller, BC. 2016. "Eidos: A Simple Scripting Language." *URL: Http://Benhaller. Com/Slim/Eidos_Manual. Pdf.*

Haller, Benjamin C, and Philipp W Messer. 2019. "SLiM 3: Forward Genetic Simulations Beyond the Wright–Fisher Model." Edited by Ryan Hernandez. *Molecular Biology and Evolution* 36 (3): 632–37. https://doi.org/10.1093/molbev/msy228.

Kelleher, Jerome, Kevin R. Thornton, Jaime Ashander, and Peter L. Ralph. 2018. "Efficient Pedigree Recording for Fast Population Genetics Simulation." Edited by Sergei L. Kosakovsky Pond. *PLOS Computational Biology* 14

(11): e1006581. https://doi.org/10.1371/journal.pcbi.1006581.
Kimura, Motoo. 1955. "Solution of a Process of Random Genetic Drift with a Continuous Model." *Proceedings of the National Academy of Sciences of the United States of America* 41 (3): 144.
Lewontin, R. C, and J. Krakauer. 1973. "Distribution of Gene Frequency as a Test of the Theory of the Selective Neutrality of Polymorphisms." *Genetics* 74 (1): 175–95. https://www.genetics.org/content/genetics/74/1/175.full.pdf.
Mark A. Beaumont, Richard A. Nichols. 1996. "Evaluating Loci for Use in the Genetic Analysis of Population Structure." *Proceedings of the Royal Society of London. Series B: Biological Sciences* 263 (1377): 1619–26. https://doi.org/10.1098/rspb.1996.0237.
Nei, Masatoshi, and Takeo Maruyama. 1975. "Lewontin-Krakauer Test for Neutral Genes." *Genetics* Comment: 1.
Nesmachnow S., Iturriaga S. 2019. "Cluster-UY: Collaborative Scientific High Performance Computing in Uruguay. In: Torres m., Klapp j. (Eds) Supercomputing." *Communications in Computer and Information Science* 1151. https://doi.org/https://doi.org/10.1007/978-3-030-38043-4_16.
Nicholson, George, Albert V. Smith, Frosti Jonsson, Omar Gustafsson, Kari Stefansson, and Peter Donnelly. 2002. "Assessing Population Differentiation and Isolation from Single-Nucleotide Polymorphism Data." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64 (4): 695–715. https://doi.org/10.1111/1467-9868.00357.
Pedersen, Thomas Lin. 2021. *Ggforce: Accelerating 'Ggplot2'*. https://CRAN.R-project.org/package=ggforce.
Pickrell, Joseph K., and Jonathan K. Pritchard. 2012. "Inference of Population Splits and Mixtures from Genome-Wide Allele Frequency Data." Edited by Hua Tang. *PLoS Genetics* 8 (11): e1002967. https://doi.org/10.1371/journal.pgen.1002967.
Purcell, Shaun, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel AR Ferreira, David Bender, Julian Maller, et al., et al. 2007. "PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses." *The American Journal of Human Genetics* 81 (3): 559–75.
Rabiner, L R, and B H Juang. 1986. "An Introduction to Hidden Markov Models," 13.
Racimo, Fernando, Jeremy J. Berg, and Joseph K. Pickrell. 2018. "Detecting Polygenic Adaptation in Admixture Graphs." *Genetics* 208 (4): 1565–84. https://doi.org/10.1534/genetics.117.300489.
Ram, Karthik, and Hadley Wickham. 2018. *Wesanderson: A Wes Anderson Palette Generator*. https://CRAN.R-project.org/package=wesanderson.
Refoyo-Martínez, Alba, Rute R. da Fonseca, Katrín Halldórsdóttir, Einar Árnason, Thomas Mailund, and Fernando Racimo. 2019. "Identifying Loci Under Positive Selection in Complex Population Histories." *Genome Research* 29 (9): 1506–20. https://doi.org/10.1101/gr.246777.118.
Reynolds, John. 1983. "ESTIMATION OF THE COANCESTRY COEFFICIENT: BASIS FOR A SHORT-TERM GENETIC DISTANCE," 13.
Robertson, Alan. 1975. "GENE FREQUENCY DISTRIBUTIONS AS A TEST OF SELECTIVE NEUTRALITSI." *Genetics*, 11.
Saitou, N, and M Nei. 1987. "The neighbor-joining method: a new method for reconstructing phylogenetic trees." *Molecular Biology and Evolution* 4 (4): 406–25. https://doi.org/10.1093/oxfordjournals.molbev.a040454.
Scheet, Paul, and Matthew Stephens. 2006. "A Fast and Flexible Statistical Model for Large-Scale Population Genotype Data: Applications to Inferring Missing Genotypes and Haplotypic Phase." *The American Journal of Human Genetics* 78 (4): 629–44. https://doi.org/10.1086/502802.
Schloerke, Barret, Jason Crowley, Di Cook, Francois Briatte, Moritz Marbach, Edwin Thoen, Amos Elberg, and Joseph Larmarange. 2018. "Ggally: Extension to Ggplot2." *R Package Version* 1 (0).
Smith, John Maynard, and John Haigh. 1974. "The Hitch-Hiking Effect of a Favourable Gene," 13.
Tsakas, S, and C B Krimbas. 1976. "TESTING THE HETEROGENEITY OF F VALUES: A SUGGESTION AND A CORRECTION." *Genetics* 84 (2): 399–401. https://doi.org/10.1093/genetics/84.2.399.
Venables, W. N., and B. D. Ripley. 2002. *Modern Applied Statistics with s*. Fourth. New York: Springer. https://www.stats.ox.ac.uk/pub/MASS4/.
Wakeley, John. 1999. "Nonequilibrium Migration in Human History." *Genetics* 153 (4): 1863–71.
———. 2009. *Coalescent Theory: An Introduction*. 575: 519.2 WAK.
Wickham, Hadley. 2017. "The Tidyverse." *R Package Ver* 1 (1): 1.
Wright, Sewall. 1945. "The Differential Equation of the Distribution of Gene Frequencies." *Proceedings of the National Academy of Sciences of the United States of America* 31 (12): 382.
Xie, Yihui. 2016. *Bookdown: Authoring Books and Technical Documents with r Markdown*. CRC Press.
Yi, X., Y. Liang, E. Huerta-Sanchez, X. Jin, Z. X. P. Cuo, J. E. Pool, X. Xu, et al. 2010. "Sequencing of 50 Human Exomes Reveals Adaptation to High Altitude." *Science* 329 (5987): 75–78. https://doi.org/10.1126/science.1190

372.
Yu, Guangchuang. 2020. "Using Ggtree to Visualize Data on Tree-Like Structures." *Current Protocols in Bioinformatics* 69 (1): e96.

Apéndice A

Figuras suplementarias

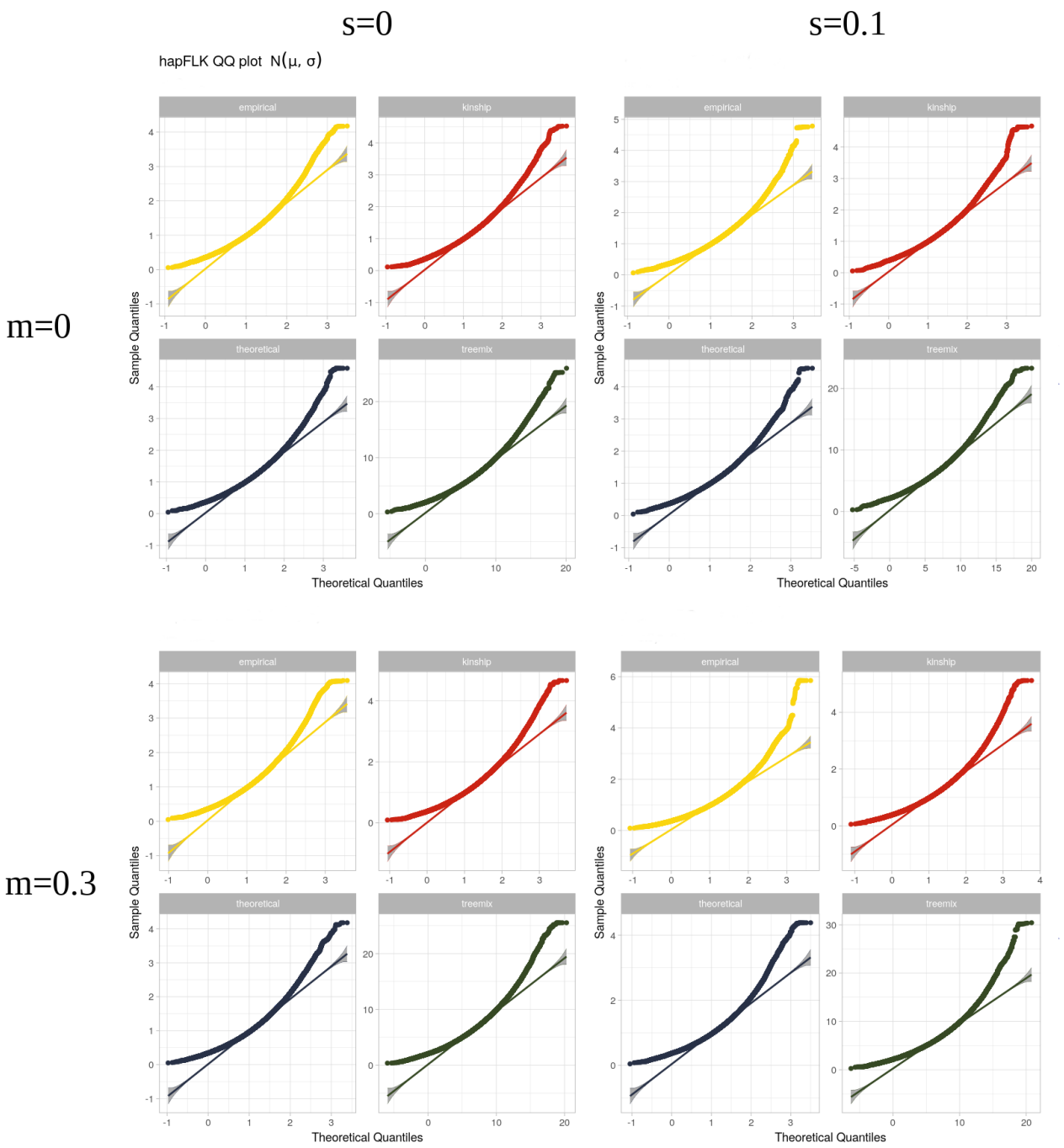

Figura A.1: Gráficos q-q de ajuste a distribución normal para el escenario A. Las columnas indican el valor de s, las filas el de m. En las abcisas están los cuantiles para una distribución normal, en las ordenadas los cuantiles de los valores de hapflk. En amarillo, la distribucion estimada con la covarianza empírica, en rojo kinship, en azul teórica, y en verde estimación por treemix.

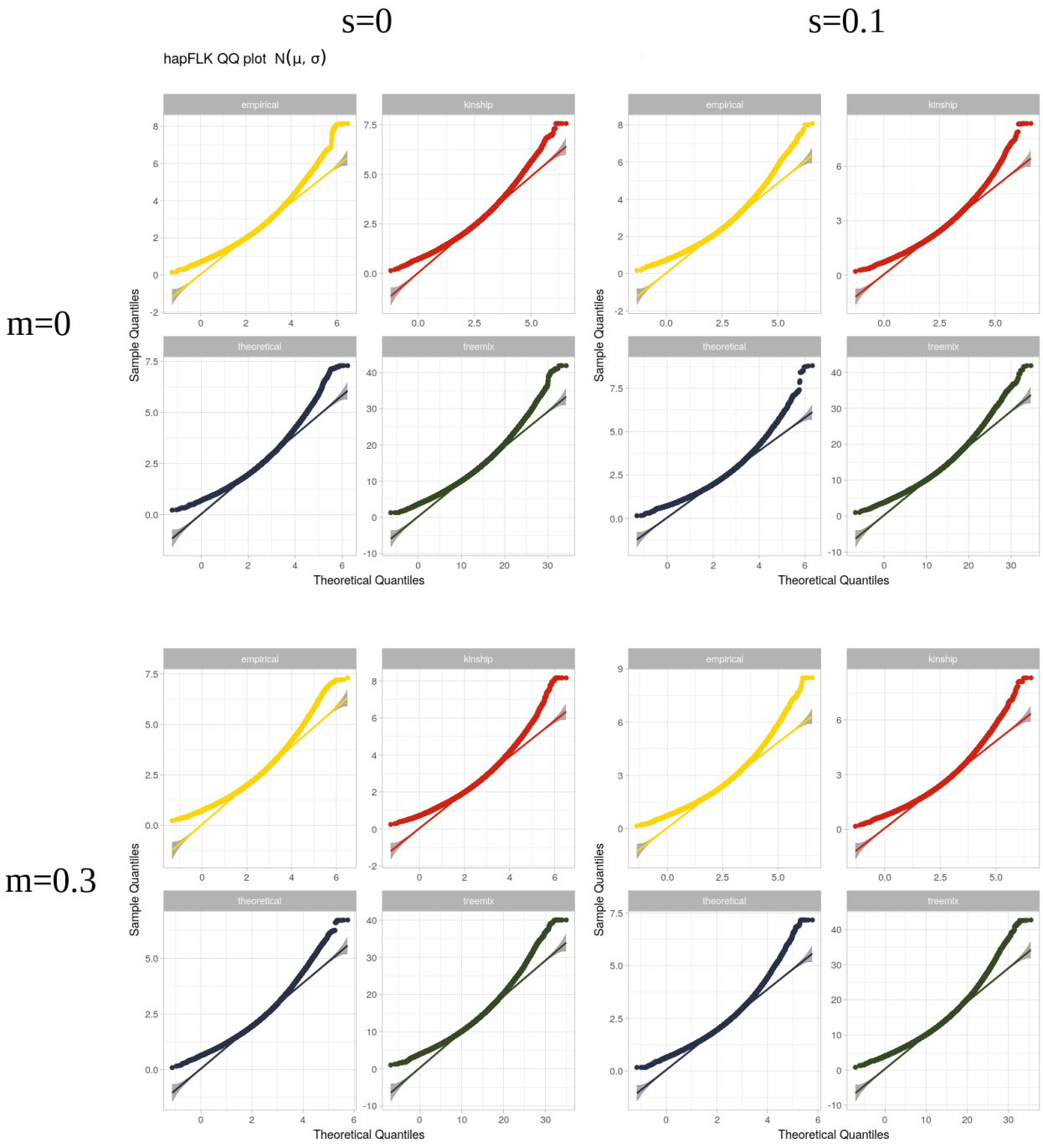

Figura A.2: Gráficos q-q de ajuste a distribución normal para el escenario B. Las columnas indican el valor de s, las filas el de m. En las abcisas están los cuantiles para una distribución normal, en las ordenadas los cuantiles de los valores de hapflk. En amarillo, la distribucion estimada con la covarianza empírica, en rojo kinship, en azul teórica, y en verde estimación por treemix.

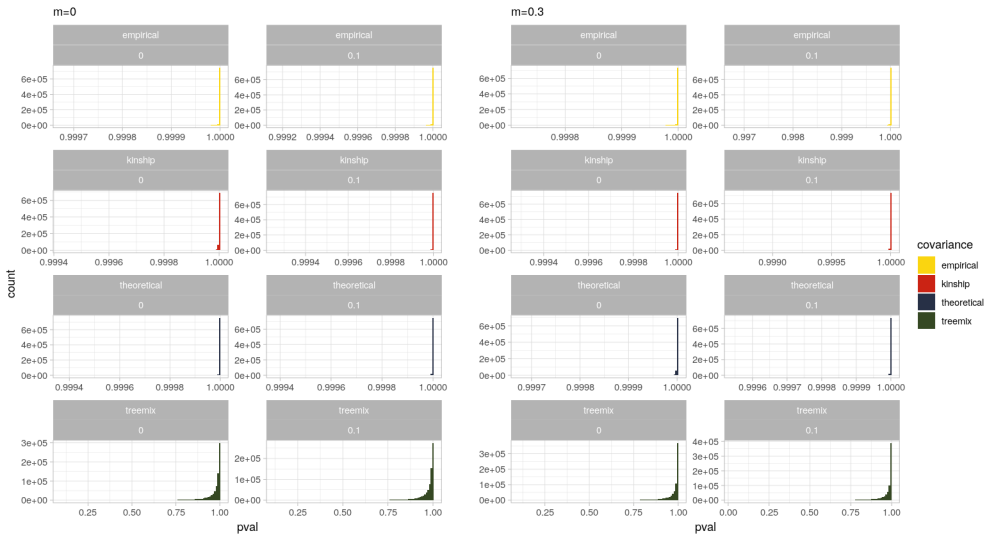

Figura A.3: Histograma de p-valores calculado asumiendo una distribucion chi cuadrado de H_0 para el escenario A. Sin admixture a la izquierda, con admixture a la derecha.

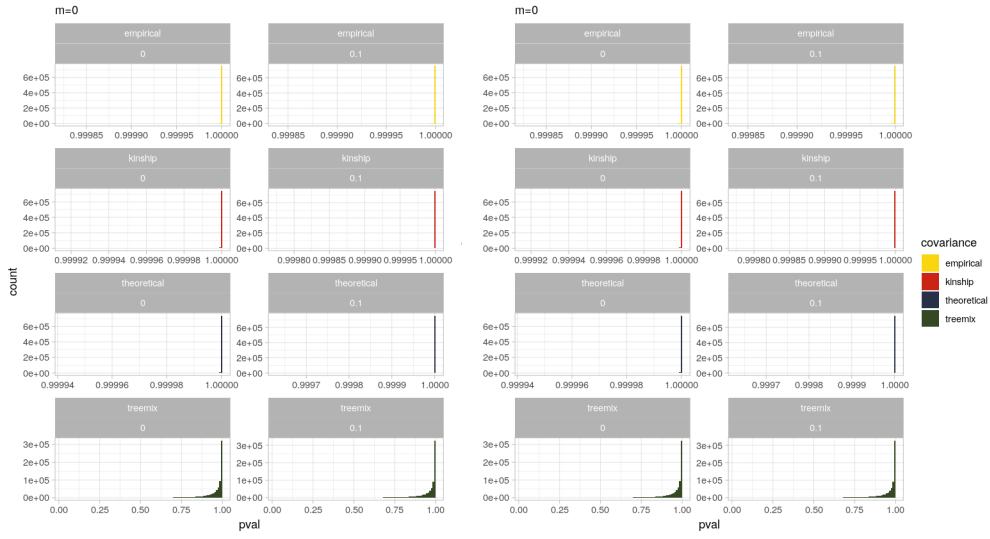

Figura A.4: Histograma de p-valores calculado asumiendo una distribucion chi cuadrado de H_0 para el escenario B. Sin admixture a la izquierda, con admixture a la derecha.

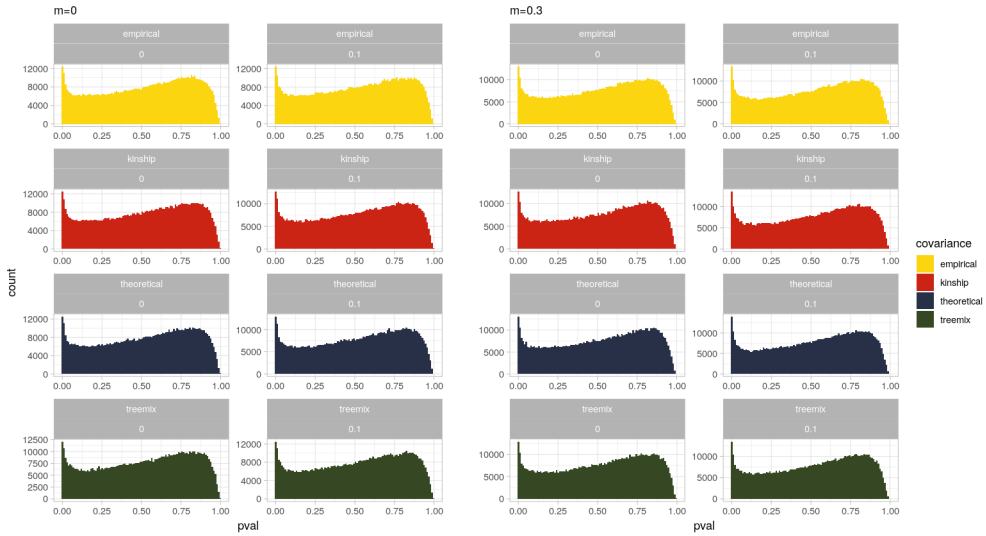

Figura A.5: Histograma de p-valores calculado asumiendo una distribucion normal de H_0 para el escenario A. Sin admixture a la izquierda, con admixture a la derecha.

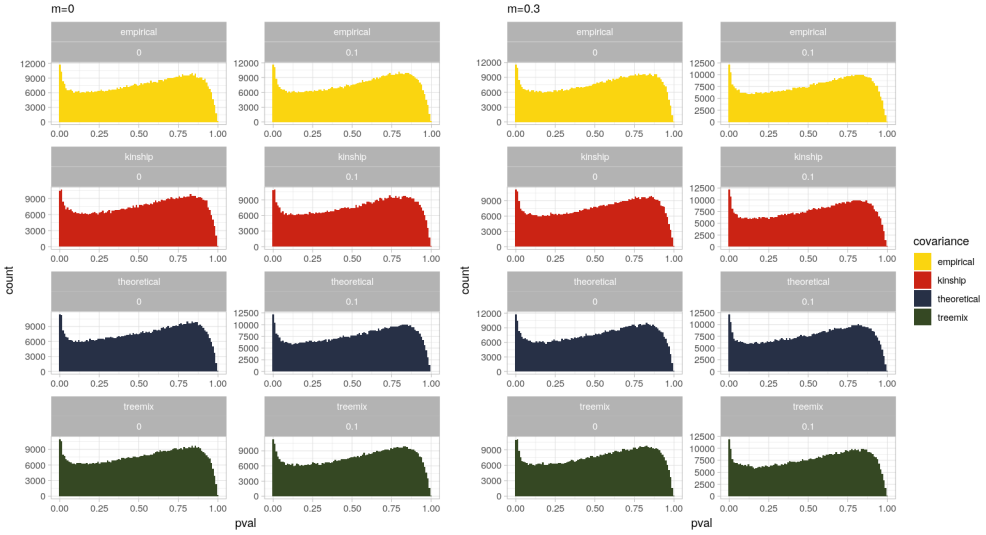

Figura A.6: Histograma de p-valores calculado asumiendo una distribucion normal de H_0 para el escenario B. Sin admixture a la izquierda, con admixture a la derecha.

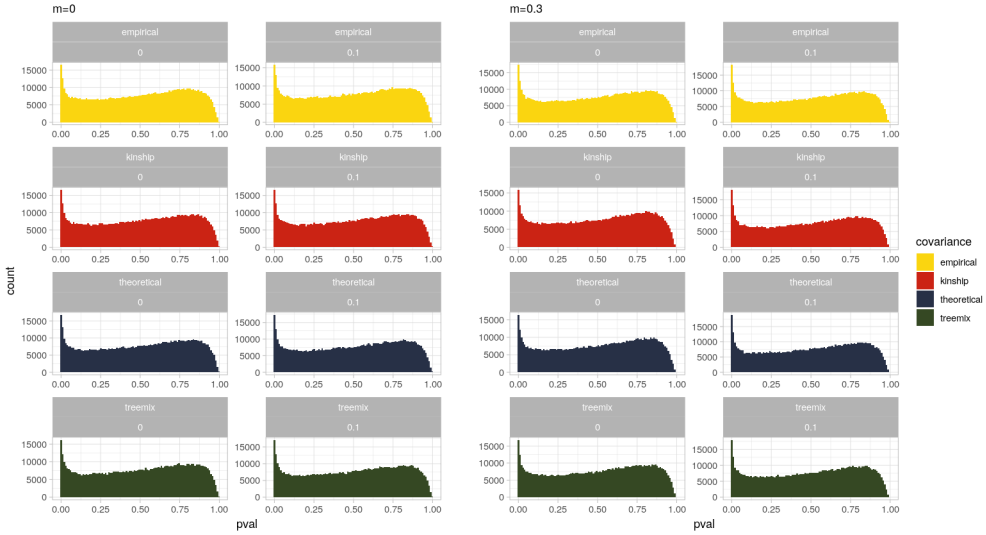

Figura A.7: Histograma de p-valores calculado asumiendo una distribucion normal de H_0 para el escenario A, con estimacion robusta de la media y varianza. Sin admixture a la izquierda, con admixture a la derecha.

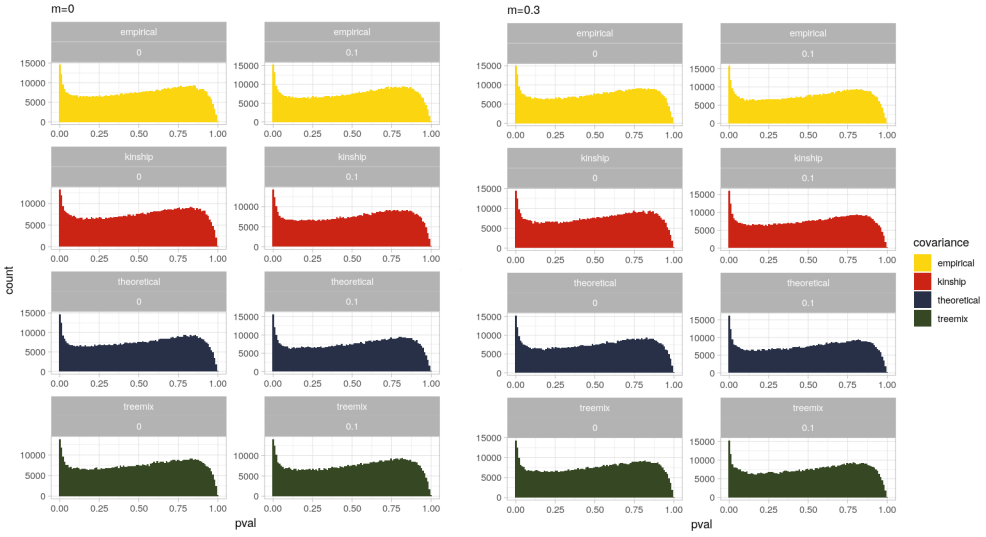

Figura A.8: Histograma de p-valores calculado asumiendo una distribucion normal de H_0 para el escenario B, con estimacion robusta de la media y varianza. Sin admixture a la izquierda, con admixture a la derecha.

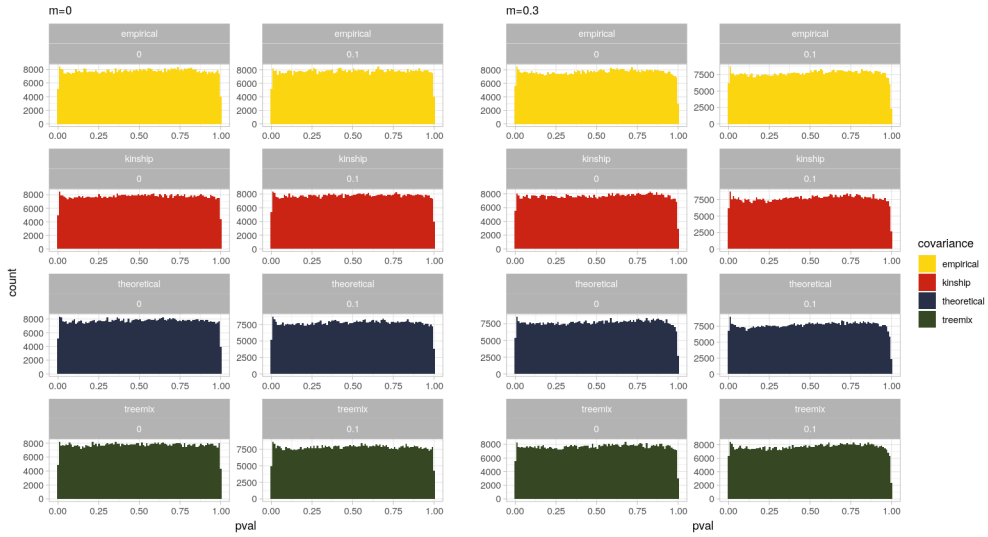

Figura A.9: Histograma de p-valores calculado asumiendo una distribucion chi cuadrado de H_0 para el escenario A, con reescalado por regresión de cuantiles. Sin admixture a la izquierda, con admixture a la derecha.

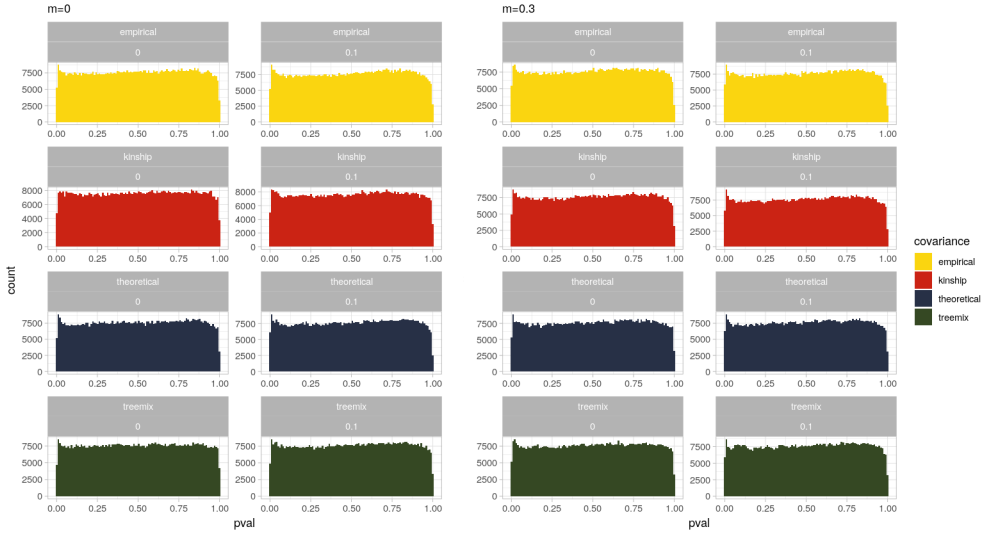

Figura A.10: Histograma de p-valores calculado asumiendo una distribucion chi cuadrado de H_0 para el escenario A, con reescalado por regresión de cuantiles. Sin admixture a la izquierda, con admixture a la derecha.

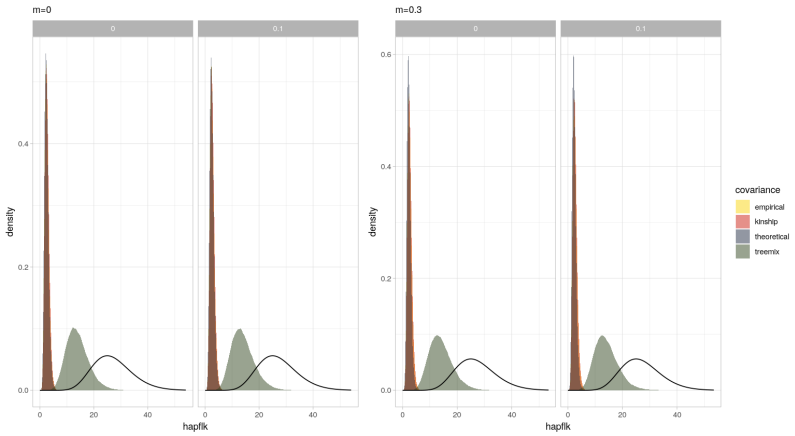

Figura A.11: Densidades de estadístico hapflk (curvas transparentes) junto con la densidad chi cuadrado teórica correspondiente (línea negra), escenario B.

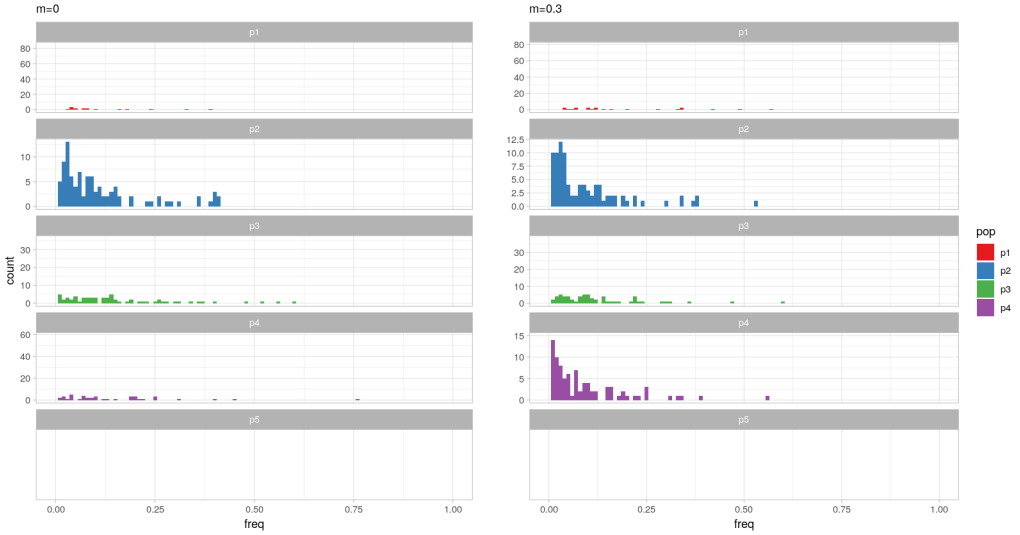

Figura A.12: Histogramas de frecuencias alélicas de m2 al comienzo de la selección para el escenario A. Ausencia de *admixture* (izquierda) y $m = 0{,}3$ (derecha).

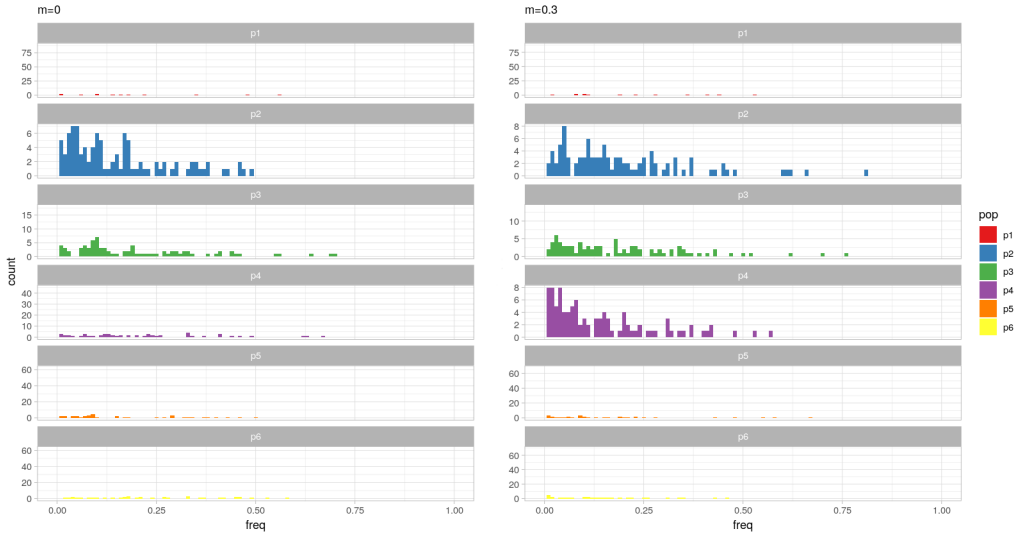

Figura A.13: Histogramas de frecuencias alélicas de m2 al comienzo de la selección para el escenario B. Ausencia de *admixture* (izquierda) y $m = 0{,}3$ (derecha).