Efecto y estimación de la homoplasia en el crecimiento poblacional

Licenciatura en Ciencias Genómicas

Vicente Diego Ortega Del Vecchyo

Tutor: Joost van Heerwaarden

Instituto de Ecología / Centro de Ciencias Genómicas

Universidad Nacional Autónoma de México / Generación 2005

Índice

Resumen

En la ecología existe un gran interés en el estudio del crecimiento histórico del tamaño poblacional. Dicho crecimiento poblacional puede reflejar varias causas, como el inicio de un periodo interglacial cuyo clima favorece la proliferación de la población. El primer paso para hacer inferencias sobre las causas del crecimiento poblacional es la estimación del tiempo hacia el pasado en el que comenzó la expansión poblacional.

En las secuencias de ADN no recombinantes, el análisis de la distribución de las diferencias pareadas, o distribución *mismatch*, ha sido el método más usado para inferir el tiempo en el que comenzó el crecimiento poblacional. En el caso de las plantas, cuya tasa de sustitución en el cloroplasto es baja, la distribución *mismatch* se ha utilizado con microsatélites, marcadores moleculares que tienen una tasa de mutación que va de 10^{-6} a 10^{-2} mutaciones por generación y que es mayor a la tasa de sustitución de los nucleótidos. Sin embargo, al usar microsatélites se puede subestimar el tiempo en que comenzó el crecimiento poblacional ya en dichos marcadores es muy frecuente la homoplasia por efecto de la alta tasa de mutación de los microsatélites y se ha utilizado la distribución *mismatch* bajo el supuesto de que los microsatélites mutan según un modelo de sitios infinitos, donde no se espera homoplasia.

Esta tesis estudia el efecto de diferentes medidas de homoplasia en la estimación del tiempo de la expansión. Se demuestra que el promedio de la homoplasia de tamaño por microsatélite y una nueva medida llamada homoplasia por distancia son las dos medidas de homoplasia más asociadas con la subestimación del tiempo en el que inició la expansión poblacional. Se desarrolló un algoritmo bayesiano aproximado que incorpora un modelo de mutación realista para los microsatélites, el modelo

de mutación *stepwise*, y que estima la homoplasia por distancia y el promedio de la homoplasia de tamaño por microsatélite en datos de microsatélites. Gracias a la ponderación del efecto de la homoplasia, se obtienen estimaciones más precisas del tiempo en que comenzó el crecimiento poblacional con el algoritmo bayesiano aproximado que con la distribución *mismatch,* usando la distribución *mismatch* bajo el supuesto de que los microsatélites evolucionan según un modelo de sitios infinitos. Este nuevo método se aplicó en datos de *Pinus caribaea*. Así pudo comprobarse que debido a la homoplasia, el uso de la distribución *mismatch* produce una subestimación de alrededor del 25% en el tiempo del crecimiento poblacional en dichos datos. Esta subestimación sitúa el tiempo del inicio de la expansión poblacional en un ciclo glacial posterior al estimado con el algoritmo bayesiano aproximado.

Abstract

In ecology there is a big interest in the study of historical population growth. Population growth can reflect several causes, such as the beginning of the interglacial period which has climatic conditions that favor population proliferation. The first step to make inferences about the causes of population growth is the estimation of the time in which the population expansion began.

In no recombining DNA sequences, the analysis of the *mismatch* distribution has been the most used method to infer the time when the population growth began. In plants, which possess a low substitution rate in the chloroplast, the analysis of population growth has been carried out using the *mismatch* distribution with microsatellites, molecular markers with a mutation rate ranging from 10^{-6} to 10^{-2} mutations per generation, a mutation rate higher than the nucleotide substitution rate. However, the use of microsatellites may underestimate the time when the population growth began, since microsatellites are very susceptible to homoplasy.

This thesis examines the effect of different measures of homoplasy to estimate the time of the expansion. I demonstrate that mean size homoplasy and a new measure called distance homoplasy are the two homoplasy measures more associated with the underestimation of the time when the population expansion began. I used approximate bayesian computation to estimate the distance homoplasy and the mean size homoplasy in microsatellite's data. Thanks to the weighting of the homoplasy effect, we could obtain more precise estimates of the time in which the population growth began using bayesian approximate computation rather than with the use of the *mismatch* distribution under the assumption that microsatellites evolve using an infinite sites model. This new method was applied on Pinus caribaea's data. It was proven that, because of the homoplasy, the use of

the *mismatch* distribution for population growth estimation produces an underestimation of about 25% in the time of population growth in those data. This underestimation provokes that the onset of population expansion time starts in a different glacial cycle, previous to the one estimated using approximate bayesian computation.

Agradecimientos

A mi tutor Joost van Heerwaarden, por su paciencia y ayuda en la creación, desarrollo y discusión de esta tesis. Gracias a sus consejos he desarrollado una mente y visión más crítica de la ciencia que me ayudará para realizar mejores proyectos en el futuro.

A Daniel Piñero, por recibirme en el Laboratorio de Genética y Evolución en el Instituto de Ecología de la UNAM, y por su ayuda para integrarme en una red de trabajo con otras personas del Instituto. Gracias también por aceptarme como asistente en el curso de Genética de Poblaciones de la Licenciatura de Ciencias Genómicas durante tres semestres.

A Rodolfo Salas-Lizana, Luis Eguiarte y Pablo Vinuesa, por sus comentarios y sugerencias para desarrollar esta tesis.

A Lluvia Flores, por enseñarme a trabajar en el laboratorio y por invitarme a participar en su proyecto de investigación.

A Alejandra Vázquez-Lobo, por sus enseñanzas sobre genética de poblaciones y filogenética.

A Lev Jardón, por facilitarme sus datos de *Pinus caribaea*.

A Jérôme Verleyen y Lorenzo Segovia, por enseñarme a usar el cluster del Instituto de Biotecnología.

Al CONACYT, por otorgarme una beca para la realización en este trabajo a través del SNI y al proyecto SEP-CONACYT.

A Alejandra Moreno-Letelier, por ayudarme a comprender mejor la dinámica de los ciclos glaciales.

A mi padre, por su ayuda para la edición de esta tesis y por su apoyo a lo largo de la carrera.

A mi madre, que demasiadas cosas hizo por mí.

A mi bella novia Roxana Blancas por todas las porras, terapias y sacrificios que me permitieron terminar esta tesis en un estado mental aceptable. Te amo mucho.

A mi carnal por las tardes de Nintendo y varias recomendaciones musicales.

A mis profesores y compañeros en la licenciatura.

A los compañeros del Laboratorio de Genética y Evolución: Alicia, Ana, Miroslava, Alejandra Ortiz, Nadia, Bianca, Mónica, Mariana, Adán, Brian, Paty, Gaby y Valeria.

A la banda del Zeldatón y Zombietón: Jáuregui, Heiblum, Grobet, Gonzalo, Chayanne y Charandín.

A Onaki, el DT Cerritos y todo el equipo de fútbol de Ecología.

A tres Pérez Gay y una Juárez.

A la FEFA (Federación de Estudiantes de la Fila de Atrás): Jorge, Julián (se te extraña), Pale y Turing.

Al sector prángana: Jero, Pano, Gano, Quique y Toñón.

A Martagón y a Lina.

A toda la familia Del Vecchyo: Chelis, Bernardo (chico y grande), Maité, Valeria, Caro, Chucho, Silvia, Chino, Tito, Mario, Gina, Beto (también chico y grande), Olimpia, Mimi.

A mi familia cancunense: Gaby, Alejandra, Yeya, Juan Luis, Renato y Juan.

A la banda del taller de Azar: Carlos Azar, Daniela, Jimena, Adriana, Anna Lee y todos los demás que alguna vez pasaron por ahí.

A la banda de la Facultad de Ciencias: Judith, Marduk, Noriko y Asaf.

Acrónimos y parámetros importantes

ABC.- Algoritmos bayesianos aproximados (ver página 32).

CORAGHE.- Coalescent based rejection algorithm for population growth and homoplasy estimation (ver página 44).

CH.- Homoplasia basada en el coalescente (ver página 25).

Distribución *mismatch*.- Distribución de diferencias pareadas (ver página 29).

HD.- Homoplasia por distancia (ver página 25).

HS.- Homoplasia por sitios (ver página 25).

IAM.- Modelo de alelos infinitos (ver página 24).

ISM.- Modelo de sitios infinitos (ver página 24).

MASH.- Homoplasia de tamaño molecularmente accesible (ver página 25).

MSH.- Promedio de la homoplasia de tamaño por microsatélite (ver página 25).

SASH.- Homoplasia estructuralmente accesible (ver página 25).

SMM.- Modelo de mutación *stepwise* (ver página 24).

TMRCA.- Tiempo al ancestro común más reciente (ver página 32).

SH.- Homoplasia por tamaño (ver página 25).

θ_0.- (ver página 29).

θ_1.- (ver página 29).

τ.- (ver página 29).

S.- (ver página 32).

s.- (ver página 32).

ε.- (ver página 32).

Introducción

Las poblaciones suelen cambiar su tamaño a través del tiempo. Los cambios en el tamaño poblacional pueden deberse a eventos históricos importantes para la especie, por ello es de interés conocer el tiempo y la magnitud de los cambios en el tamaño poblacional. A través del análisis de expansiones poblacionales podemos inferir hace cuánto tiempo sucedieron eventos de colonización, como el poblamiento humano de América (Bonatto y Salzano, 1997) o la migración de aves de continentes hacia islas (Estoup y Clegg, 2003). También podemos analizar el efecto de fenómenos geológicos en la proliferación de las especies. Se puede estudiar cómo la dinámica de los ciclos glaciales afecta la expansión poblacional de las especies y la migración hacia nuevos territorios (Moreno-Letelier y Piñero, 2009; Hewitt, 2004; Emerson y Hewitt, 2005). La inferencia correcta de los eventos de expansión poblacional nos ayuda a elaborar una historia más completa sobre las razones causantes de la proliferación de una especie.

Existen varias metodologías que nos pueden ayudar a detectar eventos de expansión poblacional. En las plantas, podemos analizar qué lugares habitaban en el pasado mediante el estudio del polen en las capas del suelo. Con el registro del polen podemos inferir expansiones poblacionales (Hewitt, 2000). Sin embargo, el estudio del polen puede ser ineficaz por las diferencias en la productividad y la dispersión del polen, que provocan que los registros no reflejen la diversidad de las plantas en el pasado (Vad Odgaard, 1999). Además del polen, la distribución de otro tipo de fósiles también puede ayudarnos a inferir expansiones poblacionales (Betancourt *et al.*, 1991). El problema es que el registro de fósiles muchas veces es pobre y no tiene resolución para inferir eventos de expansión poblacional. Una alternativa para inferir el tiempo y la magnitud del

crecimiento poblacional es el estudio de las secuencias de ADN. En las secuencias de ADN pueden quedar registrados eventos demográficos que dejan una firma detectable mediante el uso de varias pruebas estadísticas (Tajima, 1989; Fu y Li, 1993).

Las secuencias de ADN que se emplean para analizar el crecimiento poblacional necesitan un conjunto de cualidades que les permitan inferir de manera correcta la dinámica de la población en el pasado. Estas secuencias deben no deben poseer recombinación (las secuencias de ADN están ligadas) como ocurre en la mitocondria, el cloroplasto y en la región no recombinante del cromosoma Y humano (Tilford *et al.*, 2001) o deben tener escasa recombinación, como en algunas regiones nucleares cortas. En el núcleo puede haber distintas tasas de recombinación tanto entre especies, como en bacterias donde hay bacterias como *Escherichia coli* con una baja tasa de recombinación o como en *Neisseria meningitidis* donde existe una alta tasa de recombinación (Posada *et al.*, 2002), como dentro de un mismo genoma, por efecto de los *hotspots* de recombinación (Pineda-Krch y Redfield, 2005). Por ello, cuando se utilizan secuencias nucleares para realizar análisis de expansión poblacional es importante cerciorarse de que dentro de esta secuencia no encontraremos una alta tasa de recombinación. La ausencia de recombinación en las secuencias que utilizamos nos permite afirmar que todo el compartimento se puede tomar como un solo locus y, por lo tanto, compartirá la misma genealogía (Garrigan y Hammer, 2006) y las firmas de crecimiento poblacional que encontremos serán correctas. Las pruebas que sirven para detectar expansión parten del supuesto de que todos los sitios comparten la misma genealogía y la recombinación es débil o ausente (Fu y Li, 1993; Fu, 1997; Tajima, 1989; Rogers y Harpending, 1992). Además, las secuencias de ADN que usemos deben provenir de compartimentos genómicos que presenten variación genética, ya que éstas proveen una señal más clara de expansión poblacional. Por ejemplo, en

plantas es preferible el uso del cloroplasto sobre la mitocondria en estudios de expansión, ya que los cloroplastos tienen más variación genética. Se ha estimado que en gimnospermas la tasa de sustituciones sinónimas de mitocondrias respecto al cloroplasto es de 1:2, mientras que en angiospermas es de 1:3 (Drouin *et al.*, 2008).

Para realizar los estudios de expansión poblacional se requiere un marcador molecular que mute rápido, ya que de la tasa de mutación depende la antigüedad del proceso evolutivo a estudiarse, que en este caso es un proceso poblacional (Kuhner, 2008). Unos marcadores que pueden servir para este fin son los microsatélites, marcadores moleculares que consisten en pequeñas secuencias de ADN con un motivo, de una a seis pares de bases, que se repite consecutivamente. Los microsatélites tienen una tasa de mutación que oscila entre 10^{-6} y 10^{-2} mutaciones por generación, más alta que la de las sustituciones de nucleótidos (Schlötterer, 2000), lo cual convierte a los microsatélites en un marcador molecular muy eficaz en estudios de genética de poblaciones. Por ello, en las plantas podemos emplear microsatélites de cloroplasto para estudiar el crecimiento poblacional (Navascués *et al.*, 2006; Echt *et al.*, 1998). En esta tesis se trabajará exclusivamente con microsatélites y se evaluará su efectividad para realizar estudios de crecimiento poblacional.

Un problema común en varios estudios biológicos es la homoplasia, a la cual son propensos todos los marcadores moleculares. La homoplasia surge cuando en dos o más especies (o individuos en una población) se comparte el mismo estado para algún caracter, aunque éste no sea heredado de un ancestro común (Butler y Saidel, 2000) (Figura 1). Por ello, la homoplasia ocasiona que en nuestras secuencias observemos menos mutaciones de las que realmente ocurrieron. Esto provoca una pérdida de información que puede ser útil para obtener mejores estimaciones de los parámetros que definen el crecimiento poblacional. La homoplasia tiene un

efecto en los estimados del tiempo en el que comenzó un crecimiento poblacional. Utilizando microsatélites y la distribución de las diferencias entre pares de secuencias (distribución *mismatch*) (Rogers y Harpending, 1992), ampliamente usada para inferir el tiempo y la magnitud del crecimiento, se ha observado que el tiempo en el que se lleva a cabo una expansión poblacional se subestima conforme la expansión es más antigua (Navascués *et al.*, 2006). Este resultado se relaciona con la homoplasia porque cuando las expansiones son más antiguas, es probable que ocurran más mutaciones en la población y, por tanto, existe una probabilidad mayor de que surjan caracteres homoplásicos. Si el estimado del tiempo en el que comienza la expansión poblacional está sesgado, elaboraremos una historia ecológica incorrecta y no comprenderemos las condiciones con las que logró proliferar la especie bajo estudio.

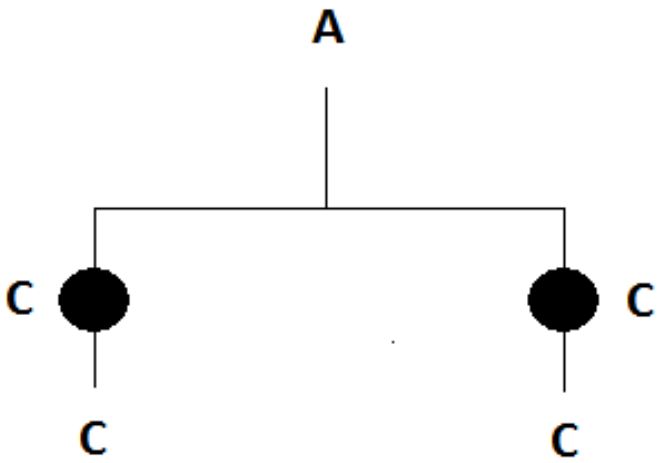


Figura 1.- Homoplasia. En un sitio segregante, una especie ancestral tenía una adenina. Las dos especies derivadas de la especie ancestral poseen una citosina en dicho sitio por dos mutaciones que ocurrieron posteriormente a la divergencia de las dos especies. Por lo tanto, como las dos especies poseen la misma base pero ésta no es heredada de un ancestro común, tenemos dos caracteres homoplásicos.

En esta tesis estudié el efecto de la homoplasia en el análisis del crecimiento poblacional usando microsatélites y la distribución *mismatch*. Evalúe cómo cambia la forma de la distribución *mismatch,* y por tanto la inferencia del tiempo de inicio del crecimiento poblacional, a partir de datos generados con simulaciones computacionales bajo diferentes escenarios de

crecimiento poblacional. Relacioné la homoplasia con el tiempo de inicio del crecimiento poblacional. Planteé una metodología capaz de inferir la homoplasia en datos de microsatélites reales junto con estimados del tiempo y la magnitud del crecimiento poblacional. Y por último, estudié la efectividad de esta metodología en un conjunto de datos de microsatélites reales.

1.1. Modelos de crecimiento poblacional

Para analizar las expansiones poblacionales, debemos elegir un modelo de crecimiento poblacional. En la Figura 2 se ilustran los tres modelos demográficos más usados. También existe una metodología, llamada *Bayesian skyline plot*, que permite inferir las fluctuaciones del tamaño poblacional a través del tiempo, pero tiene la desventaja de que sólo se puede usar cuando se tiene muy buena información sobre la historia de la población (Drummond *et al.*, 2005).

La elección de un modelo de crecimiento poblacional que se ajuste poco a nuestros datos puede sesgar nuestros resultados. Sin embargo, no existe un criterio riguroso para elegir el modelo demográfico que ilustra mejor las fluctuaciones en el tamaño de una población. No obstante, si se cuenta con secuencias que tengan mucha señal filogenética, lo cual no es frecuente, hay un método que puede ayudar a decidir cuál es el modelo demográfico que mejor se ajusta a nuestro conjunto de datos (Pybus y Rambaut, 2002). Las tres razones por las que trabajaremos con el modelo de crecimiento *stepwise* son: 1) si la magnitud del crecimiento es muy grande, todos los modelos de crecimiento poblacional generarán distribuciones *mismatch* muy similares (Rogers y Harpending, 1992); 2) es el modelo más simple para estudiar el crecimiento poblacional; por lo que, si todos los modelos de crecimiento poblacional generan las mismas

distribuciones *mismatch*, es preferible emplear el modelo más simple; 3) una metodología para inferir el crecimiento poblacional con las distribuciones *mismatch* basado en el modelo de crecimiento *stepwise* se encuentra desarrollada en el *software* Arlequin (Excoffier *et al.*, 2005).

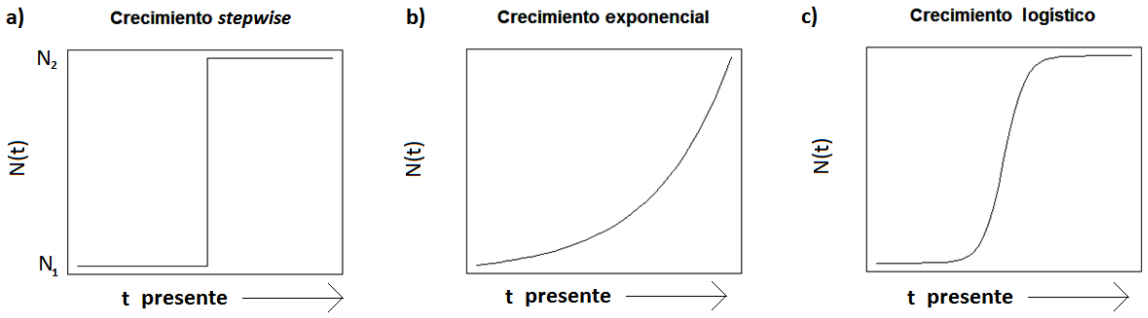


Figura 2.- Modelos de crecimiento poblacional. a) En el modelo de crecimiento *stepwise*, la población tiene un tamaño N_1 en el pasado y en un cierto tiempo t, la población aumenta de manera súbita su tamaño a N_2 y este tamaño se mantiene hasta el presente. N_1, N_2 y t son los parámetros a estimar. b) El modelo de crecimiento exponencial predice que la población aumenta su tamaño de acuerdo a la ecuación $N(t) = N(0) \exp(-rt)$. En este modelo tenemos dos parámetros a estimar: N(0) que es el tamaño de la población en el presente y r que es la tasa de crecimiento exponencial. c) El modelo de crecimiento logístico modela el crecimiento poblacional de acuerdo a la ecuación $N(t) = N(0) [(1+c) / (1 + c(\exp(rt)))]$. Este modelo depende de tres parámetros: el tamaño poblacional en el presente (N0), la tasa de crecimiento exponencial (r), y un parámetro de forma logística (c) (Pybus y Rambaut, 2002b).

1.2. Coalescencia

Para generar grupos de datos tomando en cuenta el modelo de crecimiento poblacional, necesitamos un modelo estocástico. La herramienta más útil para este propósito es la coalescencia, que describe la genealogía de un conjunto de genes muestreados en una población desde el presente hacia el pasado (Marjoram y Tavaré, 2006; Kingman, 1982).

La coalescencia utiliza como base un modelo neutral. El de uso más común y que utilizaremos en esta tesis es el modelo Wright-Fisher (Wright, 1931; Fisher, 1930 en Hein *et al.*, 2005), que describe la transmisión de los

genes en una población a través del tiempo, aunque existen otros modelos neutrales como el modelo de Moran cuya diferencia con el modelo Wright-Fisher es que permite generaciones sobrelapadas (Moran, 1958). En el modelo Wright-Fisher, cada generación *t* presenta un número 2*N* de genes y se escogen de forma aleatoria con reemplazo 2*N* genes de la generación *t* que pasarán a la generación *t* + 1 (hacia el presente). Esto se puede repetir a través de muchas generaciones y si nosotros escogemos algunos genes en el presente, podemos trazar su genealogía, como se ilustra en la Figura 3 (Hein *et al.*, 2005).

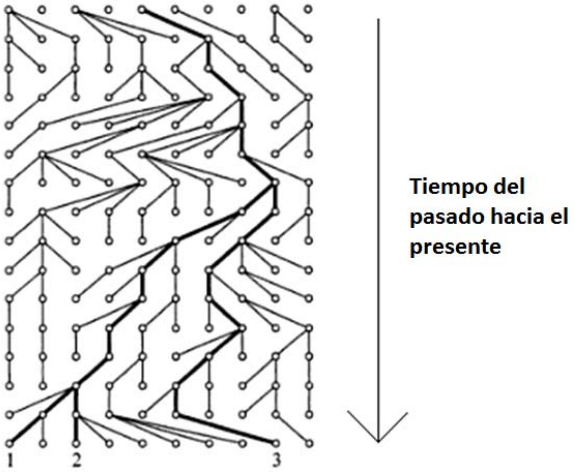


Figura 3.- Reconstrucción de la genealogía de tres genes usando el modelo Wright-Fisher.

En los procesos coalescentes, se puede construir la genealogía de *k* genes de nuestra población haploide de tamaño *2N*. Esta genealogía se crea en base a los principios del modelo Wright Fisher: dos genes coalescen con una probabilidad de 1/2*N* en la generación anterior y no coalescen en la generación anterior con una probabilidad de (1 - 1/2*N*). Con base en estos principios, es posible deducir la probabilidad de que dos de los *k* genes en esta muestra coalezcan hace un número *j* de generaciones, $P(T_k = j)$, que es aproximadamente:

$$P(T_k = j) \approx \left[1 - \binom{k}{2}\left(\frac{1}{2N}\right)\right]^{j-1} \binom{k}{2}\left(\frac{1}{2N}\right) \qquad (1)$$

T_k es el tiempo en generaciones hacia el pasado en el que se lleva a cabo una coalescencia. T_k se distribuye de manera geométrica con media $\binom{k}{2}/2N$. La distribución geométrica es discreta y computacionalmente es más sencillo modelar procesos bajo un modelo continuo. La distribución exponencial puede aproximar a la distribución geométrica en una escala continua, por lo que podemos aproximar la distribución de T_k por medio de una distribución exponencial con media $\binom{k}{2}/2N$. Para facilitar algunos cálculos podemos ajustar la distribución de los tiempos a una escala donde 1 = 2*N* generaciones. Para que T_k se ajuste a esta escala, debe tener una distribución exponencial con media $\binom{k}{2}$. A partir de la distribución de T_k puede reconstruirse una genealogía como la que se muestra en la Figura 4. La distribución de T_k está afectada por expansiones poblacionales, de manera que conforme nos recorremos hacia el pasado, T_k tiene valores más bajos y las ramas son más cortas, debido a que en el pasado el tamaño de la población era menor y los eventos de coalescencia eran más frecuentes (Hein *et al.*, 2005).

Después de que se construyó la genealogía, podemos colocar mutaciones en las ramas. Para realizar este propósito, recurrimos al parámetro θ, el cual nos dice el número de mutaciones esperadas entre cada par de linajes. Para secuencias provenientes de organelos haploides como el cloroplasto o la mitocondria, θ está definido como $\theta = 2Nu$, donde *N* es el tamaño efectivo de la población[1] y *u* es la tasa de mutación por

[1] El tamaño efectivo de la población se refiere al número de individuos que necesitaríamos en una población virtual panmíctica y de un tamaño estable para que la deriva génica ejerza la misma fuerza que en la población bajo estudio. (Futuyma, 2005)

generación de la secuencia. En una genealogía, hay dos ramas entre cada dos linajes, por lo que en cada rama se esperarían en promedio $\theta/2$ mutaciones. Si las ramas tienen una longitud *t* (*t* es el tiempo medido en una escala de *N* generaciones; Si *t* = 1, entonces ya han transcurrido *N* generaciones hacia el pasado en la coalescencia), esperaríamos que las mutaciones se distribuyeran de manera Poisson con una media de $t^*(\theta/2)$ mutaciones en cada rama. Después de colocar mutaciones en las ramas, es posible determinar un estado para el ancestro común de la genealogía y, a partir de un modelo de mutación, registrar la evolución a través de la genealogía, hasta derivar en los estados de los genes en el presente (Hein *et al.*, 2005).

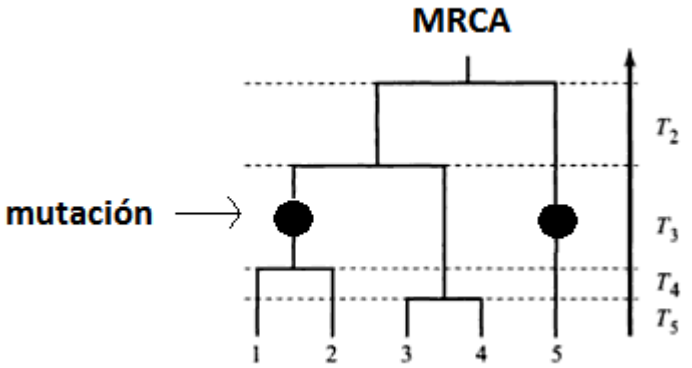


Figura 4.- Construcción de una genealogía. Tenemos 5 genes, se simula un tiempo T_5 al azar de una distribución exponencial con media $\binom{5}{2}$, y este será el tiempo de la primera coalescencia. Después se seleccionan al azar dos genes, en este caso 3 y 4, y hacemos que estos genes coalezcan. Después generamos otro tiempo (T_4) de una distribución exponencial con media $\binom{4}{2}$, y seleccionamos otro par de genes (1 y 2) y los hacemos coalescer. El proceso se repite hasta que todos los genes hayan coalescido a un gene que representa al ancestro común más reciente (MRCA). En cada rama se insertará un número de mutaciones que tiene una media de $t^*(\theta/2)$. (Modificada de Hein *et al.*, 2005).

Las genealogías que presentan una expansión poblacional poseen largas ramas terminales y ramas internas muy cortas (Slatkin y Hudson, 1991) (Figura 5). Esto quiere decir que la mayoría de las mutaciones caería

en las ramas terminales, ya que el número de mutaciones en cada rama depende de la longitud de la rama, lo cual provoca que la distribución *mismatch* tenga una forma unimodal. En contraposición, cuando el tamaño poblacional es constante se crea una distribución *mismatch* multimodal. Cuando hay una reducción en el tamaño poblacional, las ramas terminales serán mucho más cortas que las ramas internas, lo que provocará un mayor número de mutaciones en las ramas internas. Las distribuciones *mismatch* producto de una reducción en el tamaño poblacional presentan muchas crestas en valores altos del eje “x”. En el siguiente apartado explicaré más a fondo las propiedades de la distribución *mismatch*.

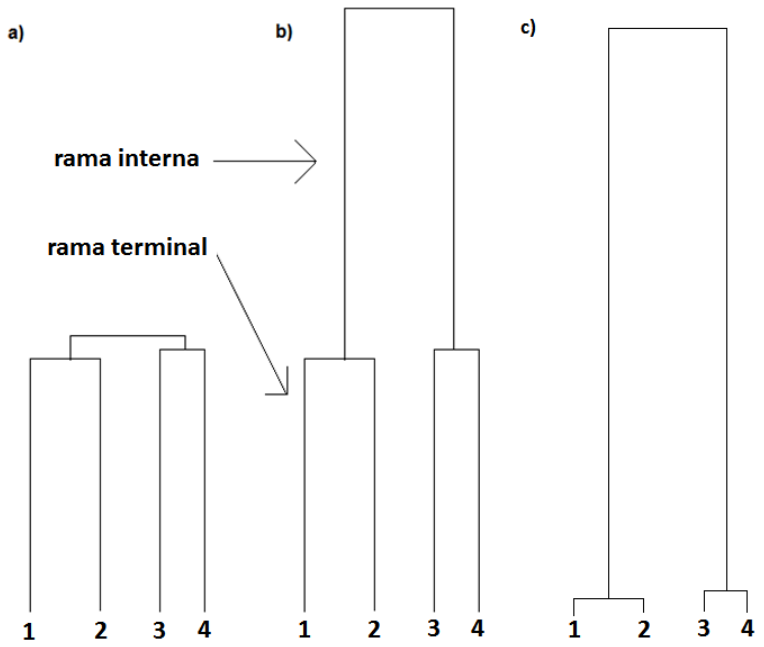


Figura 5.- Efecto del crecimiento poblacional en la coalescencia.

a) Genealogía con crecimiento poblacional.- Las ramas internas son mucho más cortas que las ramas terminales, a diferencia de las genealogías con tamaño constante a través del tiempo.
b) Genealogía con tamaño constante a través del tiempo.
c) Genealogía con reducción en el tamaño poblacional.

1.3. Modelos de mutación de microsatélites

Mediante el modelo coalescente, podemos reconstruir la evolución de una población a partir de algún marcador molecular. El cambio de estado en los marcadores moleculares se lleva a cabo por mutaciones registradas en las genealogías. Dichas mutaciones deben interpretarse de acuerdo a un modelo de mutación, que debe ser el más adecuado para representar la evolución del marcador molecular usado.

Podemos estudiar la evolución de los microsatélites con el modelo de sitios infinitos (*infinite sites model*, ISM). El ISM es un modelo derivado del modelo de alelos infinitos (*infinite allele model*, IAM). Tanto en el ISM como en el IAM, cada mutación provoca la aparición de un nuevo alelo. Sin embargo, en el ISM se asume que cada mutación cae un sitio que previamente no había mutado (Kimura, 1971). Por definición, en el ISM no ocurre la homoplasia porque dos mutaciones no permiten que se generen dos secuencias con el mismo estado. Por lo tanto, generaré datos donde las mutaciones se codifiquen de acuerdo con el ISM, para analizar la estimación del tiempo de inicio del crecimiento poblacional que obtendríamos si no existiera homoplasia.

Sin embargo, sabemos que en los microsatélites existe homoplasia. Por ello necesitamos un modelo que tome en cuenta la homoplasia para explicar la evolución de los microsatélites. Se han desarrollado varios modelos que explican la dinámica de los microsatélites: el *two phase model* modelo da una cierta probabilidad (p) de que se aumente o se disminuya en una base el microsatélite y otra probabilidad ($1 - p$) de que se aumente o disminuya más de una base en el microsatélite (Di Rienzo *et al.*, 1994); en el *K-allele model* cuando el microsatélite muta, existe la misma probabilidad de que el microsatélite mute a alguno de K estados posibles (Crow y Kimura, 1970 en Estoup *et al.*, 2002); también se han desarrollado modelos que

explican la dinámica de los microsatélites mediante cadenas de Markov (Kruglyak *et al.*, 1998). El modelo más simple para explicar la dinámica de los microsatélites es el *stepwise mutation model* (SMM) (Kimura *et al.*, 1978; Valdes *et al.*, 1993). Bajo el SMM, una mutación aumenta o disminuye en una unidad la longitud del microsatélite. Está comprobado que el modelo SMM puede ser útil para estudiar los microsatélites. Por ejemplo, en datos genómicos de humanos y chimpancés se ha encontrado que el modelo de mutación más adecuado para estudiar la evolución de sus microsatélites es el SMM (Saiduniin *et al.*, 2004). En este trabajo, modelaremos la evolución de los microsatélites bajo el SMM.

1.4. Homoplasia

El SMM es un modelo de evolución de microsatélites en el cual ocurren eventos homoplásicos frecuentes. Pueden generarse microsatélites que tengan la misma longitud y que no sean idénticos por descendencia. Necesitamos establecer formas para cuantificar la homoplasia en microsatélites.

La homoplasia suele analizarse mediante el parámetro de homoplasia por tamaño (*size homoplasy*, SH) (Estoup *et al.*, 2002). SH nos dice la probabilidad de que dos genes idénticos por estado (en microsatélites podemos considerar que el tamaño del microsatélite es el estado) no sean idénticos por descendencia. SH se define como:

$$SH = 1 - \left(\frac{Homocigosis\ bajo\ el\ modelo\ de\ alelos\ infinitos\ (IAM)}{Homocigosis\ bajo\ el\ SMM}\right) \quad (2)$$

El IAM (Kimura y Crow, 1964) establece que cada vez que ocurre una mutación, ésta crea un alelo que previamente no se encontraba en la

población. Dos microsatélites son idénticos por estado bajo el modelo de alelos infinitos si del ancestro común de ambos microsatélites hacia el presente no ha surgido ninguna mutación. La homocigosis bajo el modelo de alelos infinitos expresa la probabilidad de que dos microsatélites en la población sean idénticos por estado bajo el modelo de alelos infinitos.

Si dos microsatélites tienen el mismo tamaño, decimos que son idénticos por estado bajo el modelo SMM. La homocigosis bajo el modelo SMM nos dice la probabilidad de que dos microsatélites de la población tengan el mismo tamaño.

La homoplasia tiene un efecto en el análisis de varios parámetros poblacionales. Se ha probado que la homoplasia por tamaño subestima el número de alelos y la heterocigosis (van Oppen *et al.*, 2000) y también reduce el poder de la prueba Fs, que se basa en la proporción de haplotipos raros, para detectar expansión (Navascués *et al.*, 2006).

La SH según Estoup *et al.* (2002) está definida para un solo microsatélite. Cuando se quieren estudiar varios microsatélites ligados (sin recombinación entre ellos) en un haplotipo, se ha propuesto considerarlos como un solo microsatélite sumando los tamaños de cada microsatélite (Wimberger *et al.*, en preparación). Por tanto, SH se puede dividir en la homoplasia de tamaño molecularmente accesible (*molecularly accesible size homoplasy*, MASH) (Estoup *et al.,* 2002) y la homoplasia de tamaño estructuralmente accesible (*structurally accesible size homoplasy*, SASH). Donde MASH es una fracción de la homoplasia que puede ser encontrada en el laboratorio y SASH es otra fracción de la homoplasia que es inestimable en datos reales. SH, MASH y SASH se relacionan de la siguiente manera:

$$SH = 1 - \left(\frac{1 - H_E^I}{1 - H_E^{SI}}\right);\ SASH = 1 - \left(\frac{1 - H_E^I}{1 - H_E^{MU}}\right);\ MASH = 1 - \left(\frac{1 - H_E^{MU}}{1 - H_E^{SI}}\right) \quad (3, 4 \text{ y } 5)$$

Donde $1 - H_E^I$ es la homocigosis esperada[2] bajo el modelo de alelos infinitos.

$1 - H_E^{SI}$ es la homocigosis esperada por el tamaño total. Para obtener el tamaño total se suma la longitud de todos los microsatélites en el haplotipo y esto se toma como un estado. La homocigosis esperada por el tamaño total es la probabilidad de que dos haplotipos en la población tengan el mismo tamaño total.

$1 - H_E^{MU}$ es la homocigosis esperada multilocus, la probabilidad de que dos haplotipos tengan el mismo estado en cada uno de sus microsatélites.

La homoplasia también provoca que observemos una reducción en el número de haplotipos respecto a la que observaríamos si no existiera homoplasia. Mediremos dicho efecto con la homoplasia basada en el coalescente (*coalescent homoplasy*, CH):

$$Homoplasia\ basada\ en\ el\ coalescente = 1 - \left(\frac{número\ de\ haplotipos\ diferentes\ por\ estado\ bajo\ SMM}{número\ de\ haplotipos\ diferentes\ por\ estado\ bajo\ IAM}\right) \quad (6)$$

Adicionalmente, en este trabajo, propondré otras medidas para cuantificar la homoplasia. La primera toma en cuenta el promedio de la homoplasia observada entre microsatélites (*mean size homoplasy*, MSH):

$$Promedio\ de\ la\ homoplasia\ de\ tamaño\ por\ microsatélite = \frac{\sum_{i=1}^{número\ de\ microsatélites}\left(\frac{Homocigosis\ bajo\ IAM\ del\ microsatélite}{Homocigosis\ bajo\ SMM\ del\ microsatélite}\right)}{número\ de\ microsatélites} \quad (7)$$

[2] La heterocigosis esperada se calcula como $H_E = \left(\frac{N}{N-1}\right)(1 - \sum_{i=1}^{n} p_i^2)$. Donde N es el número de individuos en el estudio, n es el número de estados, p_i^2 es la frecuencia al cuadrado del estado i.

La homoplasia también afecta el número de sitios segregantes que tenemos cuando hay homoplasia a cuando no la hay. Para cuantificar el efecto de la homoplasia en el número de sitios segregantes, adopté la medida de homoplasia de sitios (*site homoplasy*, HS):

$$Homoplasia\ de\ sitios = 1 - \frac{Número\ de\ sitios\ segregantes\ bajo\ el\ SMM}{Número\ de\ sitios\ segregantes\ bajo\ el\ ISM} \qquad (8)$$

Esperaríamos que la homoplasia disminuyera el número de diferencias que observamos entre pares de secuencias, ya que la homoplasia provoca un aumento del número de haplotipos iguales por estado. Por ello, calcularemos una medida de homoplasia que usa el número de diferencias promedio entre todas las secuencias (*distance homoplasy*, HD):

$$Homoplasia\ de\ distancia = \frac{\pi_{ISM} - \pi_{SMM}}{\pi_{ISM}} \qquad (9)$$

donde:

$$\pi = \left(\frac{número\ de\ haplotipos}{número\ de\ haplotipos - 1}\right) \sum_{i=1}^{número\ de\ sitios\ segregantes} 2pq \qquad (10)$$

π es un estimado de la proporción del número de diferencias entre pares de secuencias o de la heterocigosis esperada (Hedrick, 2000). Para calcular π, tomamos en cuenta que en cada sitio hay dos alelos; *p* es la frecuencia de un alelo y *q* es la frecuencia del otro alelo.

Usando datos de microsatélites generados con simulaciones coalescentes, estudié la relación de estas medidas de homoplasia con la estimación del crecimiento poblacional. Identificaré qué medidas de homoplasia tienen mayor relación con un sesgo en el estimado del tiempo del crecimiento poblacional.

1.5. Métodos de detección de expansión poblacional

En el ADN podemos encontrar varias firmas de un crecimiento poblacional. Uno de los trabajos pioneros de búsqueda de firmas de expansión poblacional fue realizado por Tajima en 1987. Tajima descubre que la resta del número de diferencias entre las secuencias de ADN menos el número de sitios segregantes es significativamente menor cuando se tienen eventos de expansión poblacional a cuando la población mantiene un tamaño constante. Posteriormente han surgido otras pruebas que infieren crecimiento poblacional (Fu y Li, 1993; Fu, 1997). De forma paralela, se han buscado métodos para inferir los parámetros del tiempo y la magnitud de la expansión poblacional con base en su información genética. Los métodos para inferir estos parámetros pueden requerir que estimemos las genealogías que mejor explican nuestros datos y utilizan diferentes estrategias para explorar el espacio de genealogías. La estrategia más comúnmente usada es el *correlated sampling*, que a partir de una genealogía arbitraria realiza ligeros cambios para buscar las genealogías que mejor expliquen los datos (Kuhner *et al.*, 2008). Después estas genealogías pueden evaluarse según un enfoque bayesiano (Drummond *et al*., 2007; Kuhner *et al.*, 2008) o un enfoque de máxima verosimilitud (Kuhner *et al.*, 2006).

También podemos utilizar la distribución *mismatch*, esta metodología provee la forma más simple para estudiar el crecimiento poblacional, ya que no requiere que estimemos las genealogías que explican mejor nuestros datos. Cuando tenemos una población en crecimiento, la forma de la distribución *mismatch* tiene una fórmula analítica de la cual partimos para analizar el crecimiento poblacional.

Rogers y Harpending (1992) derivaron los valores que determinan los parámetros del crecimiento poblacional en la distribución *mismatch* a partir de una fórmula que utiliza Li (1977) para inferir el número de diferencias que esperaríamos entre dos secuencias tomadas al azar en una población.

$$F_i(\theta_0, \theta_1, \tau) = F_i(\theta_1) + (\exp\left(-\tau\left(\frac{\theta_1+1}{\theta_1}\right)\right))\,(\sum_{j=0}^{i}\frac{\tau^j}{j!}\left(F_{i-j}(\theta_0) - F_{i-j}(\theta_1)\right)) \quad (11)$$

Donde $F_i(\theta_0, \theta_1, \tau)$ muestra la probabilidad de encontrar i diferencias entre dos genes tomados al azar en una población que ha experimentado un crecimiento poblacional. $\tau = 2ut$, donde *u* es la tasa de mutación por generación de la secuencia y *t* es el tiempo en generaciones hacia el pasado en que se llevó a cabo el crecimiento poblacional. $\theta_0 = 2N_0$, donde N_0 es el tamaño efectivo de la población antes del crecimiento poblacional (antes de τ). $\theta_1 = 2N_1$, donde N_1 es el tamaño efectivo de la población después de τ. Tanto $F_i(\theta_0)$ como $F_i(\theta_1)$ muestran la probabilidad de encontrar i diferencias entre dos genes en una población con tamaño constante y parámetros de θ_0 y θ_1, respectivamente. Watterson (1975) define $F_i(\theta)$ como:

$$F_i(\theta) = \frac{\theta^i}{(\theta+1)^{i+1}} \quad (12)$$

Para encontrar los valores de θ_0 y θ_1 y τ, Rogers y Harpending (1992) usan el método no lineal de mínimos cuadrados, que trata de minimizar el valor de SSD, suma de las diferencias al cuadrado o *sum of square differences*, de la siguiente ecuación:

$$SSD = \sum_{i=0}^{n}(F_i\ datos\ - F_i\ modelo)^2 \quad (13)$$ Tomado de Schneider y Excoffier (1999)

Donde $F_i\ datos$ representa la frecuencia del número de veces que un par de secuencias tienen i diferencias en los datos. $F_i\ modelo$ es la frecuencia del número de veces que un par de secuencias tienen i diferencias en los datos según la ecuación 11. La idea es encontrar los parámetros θ_0, θ_1 y τ tales que $F_i(\theta_0, \theta_1, \tau)$ según la ecuación 11, minimicen el valor SSD de la ecuación 13. Esta metodología es la que desarrolla el *software* Arlequin (Excoffier *et al.*, 2005).

Otra contribución valiosa de Rogers y Harpending (1992) es la explicación del efecto que tienen los parámetros θ_0, θ_1 y τ en la distribución *mismatch*. Los autores deducen que:

1) El parámetro τ depende del lugar donde se encuentra la cresta de la distribución *mismatch*. Valores mayores de τ implican que la cresta de la distribución *mismatch* se encuentra más a la derecha.
2) El parámetro θ_1 depende del lugar donde la distribución atraviesa al eje "y". Valores mayores de θ_1 implican que la distribución atravesará el eje "y" en valores más bajos. Cuando la distribución *mismatch* atraviesa al eje "y" en valores cercanos a cero, es muy difícil determinar el verdadero valor de θ_1. El valor de θ_1 que se obtiene del estudio de las distribuciones *mismatch* suele estar sesgado hacia valores más altos (Schneider *et al.*, 1999).
3) El parámetro θ_0 determina el efecto de la pendiente de la distribución. Mientras más pequeño sea θ_0 , la pendiente de la distribución *mismatch* será más inclinada.

Para analizar la distribución *mismatch*, Rogers y Harpending se basan en el modelo de sitios infinitos (ISM), que funciona cuando se tienen secuencias de ADN con una tasa de mutación baja, pero no para microsatélites, donde

se sabe que existe una tasa de mutación alta. Se sabe que la homoplasia causa una subestimación de τ cuando se utiliza la distribución *mismatch* (Navascués *et al.*, 2006). Sin embargo, no se ha establecido una relación precisa en la subestimación que esperaríamos de τ con diferentes valores de homoplasia. En esta tesis investigaré dicha relación y propondré el uso de una metodología que emplea un algoritmo bayesiano aproximado (*Approximate bayesian computation*, ABC) para estimar la homoplasia. Esto podrá ayudarnos a estimar el sesgo esperado en τ cuando se utiliza la distribución *mismatch* en presencia de diferentes niveles de homoplasia.

1.6. Algoritmos bayesianos aproximados (ABC)

La homoplasia es un problema al estimar los parámetros que definen el crecimiento poblacional. Sería deseable un método de inferencia del crecimiento poblacional que pudiera estimar la homoplasia y a la vez corregir su efecto en la estimación del tiempo del crecimiento poblacional. Los ABC (Pritchard *et al.*, 1999; Estoup y Clegg, 2003) proveen una buena alternativa para inferir el crecimiento poblacional porque pueden producir estimados del tiempo y la magnitud del crecimiento poblacional dado un conjunto de genealogías y un modelo de evolución elegido por el investigador. Ya que en los ABC podemos registrar las genealogías que mejor explican nuestros datos, podemos cuantificar parámetros adicionales a partir de estas genealogías, como la homoplasia o el tiempo al ancestro común más reciente (TMRCA). Además, los ABC pueden aplicarse para modelar escenarios demográficos complejos (Estoup *et al.*, 2001).

En los ABC se establece un escenario demográfico (por ejemplo, los modelos de crecimiento poblacional explicados en la sección 1.2) y se trata de estimar el valor más probable para los parámetros que definen el escenario al cual ajustamos los datos.

Posteriormente se estima la distribución posterior *f(P|D)*, donde *P* son los parámetros que definen nuestro modelo demográfico y *D* representa a nuestro conjunto de datos.

Existen varias metodologías para realizar algoritmos bayesianos aproximados (Marjoram *et al.*, 2003). Un método consiste en reducir un conjunto de datos a un conjunto de estadísticos de resumen *S* que son informativos sobre un escenario demográfico, como el crecimiento poblacional, en nuestros datos. Un trabajo anterior ha utilizado el promedio de la varianza en el número de repeticiones por microsatélite, el promedio de la heterocigosis esperada por microsatélite y el número de haplotipos diferentes por estado como estadísticos de resumen para analizar el crecimiento poblacional (Pritchard *et al.*, 1999). Se ha justificado que la media de la varianza del número de bases por microsatélite y la media de la heterocigosis por microsatélite usados en conjunto son informativos de cambios en el tamaño en la población (Kimmel *et al.*, 1998). El número de haplotipos diferentes por estado se relaciona con el número de mutaciones que caen en las ramas terminales. En las expansiones poblacionales habrá un mayor número de mutaciones en las ramas terminales y, por tanto, un mayor número de haplotipos diferentes por estado (Pritchard *et al.*, 1999). En este trabajo propondré estadísticos de resumen que pueden mejorar la estimación del crecimiento poblacional. Estos estadísticos tratan de aprovechar una firma característica del crecimiento poblacional: una mayor proporción de mutaciones en las ramas terminales respecto a las ramas internas. Los estadísticos son: el número de sitios segregantes en todos los microsatélites, el número de *singletons* en todos los microsatélites (los singletons son estados que sólo se encuentran en uno de los linajes de la muestra que utilizamos), la varianza en el número de sitios segregantes por locus y la heterocigosis esperada tomando en cuenta todo el haplotipo.

Al emplear el conjunto de estadísticos de resumen *S*, podemos generar un algoritmo bayesiano aproximado de la siguiente manera (Marjoram *et al.*, 2003):

1) De un conjunto de datos *D*, calcular un conjunto de estadísticos de resumen *S*.
2) Generar un conjunto de parámetros *P* de una distribución *a priori*.
3) Simulamos un conjunto de datos *d* a partir de un modelo estocástico *M* (usaremos a la coalescencia como nuestro modelo estocástico) donde aplicamos el conjunto de parámetros *P*. De los datos *d* que obtenemos, calculamos un conjunto de estadísticos de resumen *s*.
4) Calculamos $|S-s|$.
5) Si $|S-s| < \varepsilon$, aceptamos los valores de *P*. ε representa el valor del umbral. Si para todos los estadísticos de resumen el valor de la diferencia entre los datos simulados y los datos reales es menor a ε, se guardan los valores de *P*.
6) Regresamos a 2 (continuamos regresando a 2 el número n de veces deseado).

Todo el conjunto de valores *P* aceptados generan la distribución $f(P|S)$; si el conjunto de estadísticos de resumen son suficientes (que sean suficientes se refiere a que no existe otro estadístico que pueda ser calculado de los datos que otorgue más información para el valor del parámetro que se desea estimar) (Fisher, 1922), se cumplirá que $f(P|S) = f(P|D)$. En este trabajo se estudiarán varios estadísticos para probar el conjunto de estadísticos que satisface el criterio de suficiencia para explicar al modelo de crecimiento poblacional.

A partir de los algoritmos bayesianos aproximados con base en la coalescencia y en un modelo de mutación SMM, mostraré cómo estimar el

tiempo y la magnitud del crecimiento poblacional. Adicionalmente, propondré cómo medir la homoplasia con ABC y evaluaré la efectividad de los ABC para medir la homoplasia en varios escenarios de crecimiento poblacional.

Objetivos

El objetivo general de este trabajo es realizar una metodología que nos permita estimar la homoplasia en datos reales de microsatélites y analizar cómo nuestros estimados del crecimiento poblacional son afectados por la homoplasia. Para realizar esto, cumpliremos con cuatro objetivos particulares:

1) Evaluar el efecto de la homoplasia en la estimación del crecimiento poblacional mediante la distribución *mismatch* cuando se utilizan microsatélites.
2) Encontrar el conjunto de estadísticos de resumen que mejor estiman el tiempo y la magnitud del crecimiento poblacional junto con la homoplasia en un modelo *stepwise*.
3) Estudiar el desempeño de los algoritmos bayesianos aproximados para estimar la homoplasia y el crecimiento poblacional en un conjunto de datos simulados.
4) Probar la utilidad de los algoritmos bayesianos aproximados en un conjunto de datos reales.

Métodos

3.1. Simulación de microsatélites

Se realizó un programa para generar datos de microsatélites de una población teórica donde puede existir crecimiento poblacional. La evolución de cada conjunto de datos de microsatélites creados con este programa se modela según un SMM, donde existe homoplasia, y un ISM, donde no hay homoplasia. Los datos creados con este programa serán usados en análisis posteriores para ponderar el efecto de la homoplasia en la inferencia del crecimiento poblacional.

Se modificó el código fuente del programa *msHOT* (Hellenthal, 2007) para hacerlo capaz de generar datos de microsatélites. *msHOT* realiza un cambio al programa *ms* (Hudson, 2002). *ms* genera secuencias e incorpora mutaciones de acuerdo con el ISM, usando un modelo coalescente (Hudson, 2002). *msHOT* utiliza casi íntegramente el código fuente de *ms* y lo modifica para incorporar *hotspots* de recombinación. Me apoyé en *msHOT* como base para tener la herramienta diseñada en caso de que en el futuro se deseen generar datos de microsatélites con *hotspots* de recombinación.

El programa *msHOT* simula genealogías con base en el modelo Wright-Fisher. En *msHOT* se supone que tenemos un tamaño efectivo N de individuos diploides; por tanto, tendremos $2N$ genes en la población. El tiempo hacia el pasado en *msHOT* se ubica en una escala donde $t = 1$ significa que ya transcurrieron $4N$ generaciones hacia el pasado y el parámetro θ está definido como $4Nu$. Los parámetros de *msHOT* serán interpretados de otra manera para ajustarlos a un modelo donde tenemos organelos haploides. Por lo tanto, si tenemos un tamaño efectivo de N individuos haploides, tendremos N genes en la población. Esto causará que

la escala se ajuste de manera que *t* = 1 signifique que ya transcurrieron 2*N* generaciones hacia el pasado y que el parámetro θ sea igual a 2*Nu*.

msHOT crea genealogías con el siguiente algoritmo:

1) Se comienza con un número k de linajes.
2) Se simula un tiempo de espera (T_k) que se distribuye de acuerdo a una exponencial con media $\binom{k}{2}(2)$.
3) Se seleccionan dos linajes al azar y se hacen coalescer a un solo linaje, de manera que k disminuye su número en 1.
4) Si k no es igual a 1, se repite el paso 2. Si k es igual a 1, entonces significa que en la genealogía ya se ha encontrado el ancestro común más reciente. Por tanto, el algoritmo se detiene, puesto que ya se ha terminado de construir la genealogía.

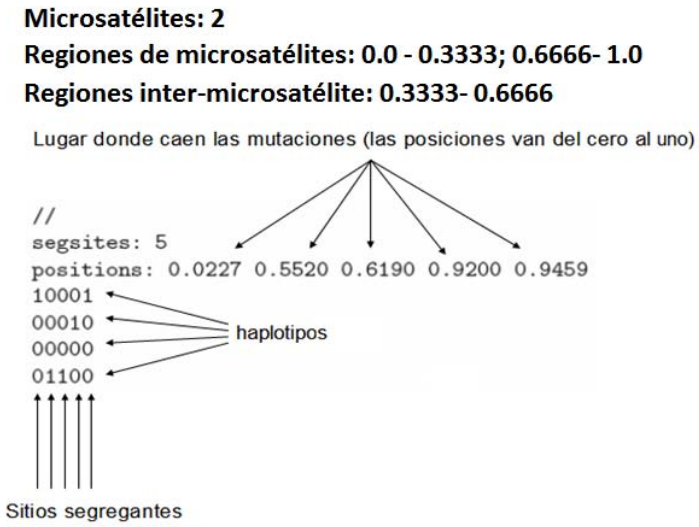
Una vez que se construye la genealogía, las mutaciones se agregan en cada rama de acuerdo a una distribución Poisson con media t * θ, donde t es la longitud de la rama.

En *msHOT*, las mutaciones se colocan de manera aleatoria en una secuencia continua que va del cero al uno (Figura 6). Para codificar los microsatélites, se dividirá la secuencia en dos tipos de regiones: 1) regiones donde hay un microsatélite y 2) regiones inter-microsatélites (en donde se puede asignar recombinación si el usuario lo requiere). Para codificar microsatélites, siempre se colocará uno en los dos extremos de la secuencia y se tendrá una región intermicrosatélite entre cada dos microsatélites. Como ejemplo, si se quieren codificar tres microsatélites, la secuencia se divide en 5. De 0 a 0.2, 0.4 a 0.6 y 0.8 a 1 tendremos una región donde hay un microsatélite. De 0.2 a 0.4 y de 0.6 a 0.8 habrá una región inter-microsatélites.

msHOT supone un ISM para codificar las mutaciones. En cada sitio segregante, cada una de las secuencias pudieron o no haber mutado. *msHOT* coloca un “1” si la secuencia tuvo una mutación y un “0” si la secuencia no mutó (ver Figura 6). Codificaré un SMM simétrico, donde cada mutación ocasiona que el microsatélite tenga una base menos o más con la misma probabilidad. Primero se revisa cuáles mutaciones cayeron en una región donde hay un microsatélite. Para los sitios segregantes que se encuentren en una región donde haya un microsatélite, el algoritmo asignará un -1 ó un +1 de forma aleatoria y con la misma probabilidad para cada sitio segregante. Después se revisan todos los haplotipos para ver en cuáles ocurrió una mutación; donde haya una mutación, el microsatélite disminuye en 1 base su tamaño (si al sitio segregante se le asignó un -1) o aumenta en 1 base su tamaño (si al sitio segregante se le asignó un +1). Este proceso se repite para todas las regiones donde hay un microsatélite hasta codificar todos los microsatélites.

Por cada genealogía simulada, se generan cinco archivos:

1) Los microsatélites creados a partir de un SMM de cada genealogía se imprimen a un archivo de salida en un formato que puede ser leído por otro programa que utiliza un ABC.
2) Los microsatélites generados con el SMM se imprimen en un formato que puede ser leído por Arlequin (Excoffier, 2005).
3) Se imprimen las mutaciones que ocurrieron en cada microsatélite de acuerdo con un ISM en un formato que puede ser leído por el programa que usa el ABC. Los microsatélites se crean con un ISM para analizar la estimación del crecimiento poblacional que esperaríamos si no existiera homoplasia.
4) Los microsatélites creados de acuerdo a un ISM se imprimen en un formato que puede leer Arlequin.



Mutaciones en regiones de microsatélite: 0.0227 para el primer microsatélite (se asigna al azar +1); 0.9200 y 0.9259 para el segundo microsatélite (se asigna al azar un -1 y un +1 respectivamente).

Haplotipos	microsatélite 1	microsatélite 2
1	1	1
2	0	-1
3	0	0
4	0	0

Figura 6.- Codificación de mutaciones en msHOT. Se realizó una genealogía con cuatro genes. En esta genealogía surgieron cinco mutaciones. La primera secuencia tiene dos mutaciones (ubicadas en 0.0227 y 0.9456 en la secuencia).
Se localizaron las mutaciones que cayeron en una región de microsatélite (3) y para cada una de estas mutaciones se añadió un -1 o un +1 al azar. Si un haplotipo contiene un microsatélite con una cierta mutación, se resta una base si la mutación es igual a -1 y se suma una base si la mutación tiene un +1.

5) Se guardan valores de medidas de homoplasia y de estadísticos de resumen de los datos. Las medidas de homoplasia que se imprimen son: HS, CH, HD, MSH, SH, MASH y SASH (ver la sección 1.4 y acrónimos). También se imprimen los valores de los siguientes estadísticos de resumen: a) promedio de la varianza en el número de repeticiones por microsatélite; b) promedio de la heterocigosis esperada por microsatélite; c) número de haplotipos diferentes por estado; d) número de sitios segregantes en todos los microsatélites (la suma de la diferencia entre el tamaño mayor menos el tamaño menor en cada microsatélite); e) número de *singletons* en todos los microsatélites; f) varianza en el número de sitios segregantes por microsatélite; y g) heterocigosis esperada tomando en cuenta todo el haplotipo.

3.2. Distribuciones *mismatch*

Para analizar el efecto de la homoplasia en la distribución *mismatch* y, por tanto, en el análisis del crecimiento poblacional, se plantearon 46 grupos de datos con distintas combinaciones de los parámetros que definen el crecimiento *stepwise* (ver Tabla suplementaria 1). En estos datos se define un valor para θ_1 (5, 15 ó 30) y θ_0 se ajusta para que su valor sea una milésima parte de θ_1 para simular una expansión poblacional muy clara. Los valores de θ_1 y de τ fueron elegidos porque representaban valores de θ_1 similares a los encontrados en datos de *Pinus caribaea*. Para cada grupo de datos se aplican diferentes valores de τ que van de 1.5 a 30 en intervalos de 1.5 cuando θ_1 es igual a 15 o a 30. Si θ_1 es igual a 5, τ va de 1.5 a 9 en intervalos de 1.5 (Cuando θ_1 = 5, si la población mantuviera un tamaño constante, el tiempo de coalescencia promedio de todos los genes sería τ = 10. Por lo tanto, si el cambio en el tamaño de la población ocurre cuando τ es mayor que diez, es probable que todos los genes ya hayan coalescido y que los cambios en el tamaño poblacional no afecten a la genealogía ya trazada). Se plantearon seis grupos de datos donde no hay expansión y θ es igual a 5, 15 y 30 (ver Tabla suplementaria 1). Para cada grupo de datos, se simularon cien genealogías con 150 individuos para obtener cien juegos de datos de microsatélites con sus medidas de homoplasia.

Para analizar el efecto de la homoplasia en la forma de la distribución *mismatch,* se graficó la distribución *mismatch* para todos los juegos de datos de microsatélites generados bajo el ISM y el SMM. Se promediaron las distribuciones *mismatch* de los 100 juegos de datos realizados con cada combinación de parámetros de θ_0, θ_1 y τ, a fin de generar una distribución *mismatch* promedio distintiva de cada conjunto de parámetros bajo el ISM y el SMM. Para analizar la similitud entre distribuciones *mismatch*, se usó una

medida de bondad de ajuste (*goodness of fit*) entre distribuciones (Ricci, 2005):

$$\delta = \sum_{i=1}^{Diferencias\ entre\ pares\ de\ secuencias} |f(i)_{distribución\ A} - f(i)_{distribución\ B}| \quad (15)$$

Las medidas de bondad de ajuste se calcularon en función de: a) Las distribuciones *mismatch* promedio generadas para cada combinación de parámetros de crecimiento poblacional contra la distribución *mismatch* promedio cuando no hay crecimiento poblacional bajo el SMM; b) Las distribuciones *mismatch* promedio generadas bajo cada combinación de parámetros de crecimiento poblacional contra la distribución *mismatch* promedio cuando no hay crecimiento poblacional bajo el ISM; y c) La distribución *mismatch* promedio bajo el SMM contra la distribución *mismatch* promedio bajo el ISM para cada una de las combinaciones de parámetros de crecimiento poblacional.

Se obtuvieron estimados de θ_0 , θ_1 y τ ($\widehat{\theta_0}$, $\widehat{\theta_1}$ y $\hat{\tau}$) usando la distribución *mismatch* mediante el *software* Arlequin para todos los conjuntos de datos de microsatélites creados bajo el ISM y el SMM. Se calculó la media y la varianza de los estimados de $\widehat{\theta_0}$, $\widehat{\theta_1}$ y $\hat{\tau}$ para cada grupo de datos (Tabla suplementaria 1) y se estimaron intervalos de confianza para los valores de $\widehat{\theta_0}$, $\widehat{\theta_1}$ y $\hat{\tau}$ en Arlequin. Para crear los intervalos de confianza, con los valores estimados de $\widehat{\theta_0}$, $\widehat{\theta_1}$ y $\hat{\tau}$ se generan un número B de datos (se usaron mil datos). Para cada uno de éstos, se estiman los parámetros del tiempo y la magnitud del crecimiento poblacional ($\dot{\theta_0}$, $\dot{\theta_1}$ y $\dot{\tau}$). Para un nivel de confianza α (que se fijaron en 0.05) se obtuvieron los valores aproximados de los límites de los intervalos de confianza como los valores que están en los percentiles α/2 y 1 - α/2 de las distribuciones de $\dot{\theta_0}$, $\dot{\theta_1}$ y $\dot{\tau}$. Se calculó el número de veces que τ cae en los intervalos de

confianza generados por Arlequin. No se realizaron pruebas de neutralidad como la D de Tajima (Tajima, 1989) ni de Fu (Fu, 1997) porque se partió del supuesto de que en nuestros microsatélites existe crecimiento poblacional, pero en un análisis estricto debía realizarse alguna de las dos pruebas de neutralidad para aceptar la hipótesis de que existe crecimiento poblacional.

Para estudiar la relación entre la homoplasia y el sesgo en la estimación de $\hat{\tau}$ se realizaron dos análisis. Con las cien simulaciones efectuadas dentro de un grupo de datos, se establecieron correlaciones entre los valores de las medidas de homoplasia y la siguiente medida del sesgo relativo en la estimación de τ :

$$Sesgo\ relativo\ en\ la\ estimación\ de\ \hat{\tau} = (\hat{\tau}_{ISM} - \hat{\tau}_{SMM})/\hat{\tau}_{ISM} \quad (16)$$

También se tomaron todos los datos de microsatélites donde existe crecimiento poblacional con θ_1 = 15 y, por separado, todos los datos con θ_1 = 30 y se calculó el sesgo relativo en $\hat{\tau}$. No se analizaron los datos generados con un valor de θ_1 que no permitiera encontrar señales de expansiones poblacionales antiguas. Después se eliminaron todas las simulaciones donde el sesgo relativo en la estimación de $\hat{\tau}$ fuera menor que -1, porque son simulaciones donde se encuentra que en la distribución *mismatch,* supuestamente, la cresta del número de diferencias bajo el SMM es mayor que usando el ISM. Esto no es posible, ya que siempre que existe homoplasia se encuentra un mayor número de diferencias bajo el ISM. El efecto de encontrar más diferencias aparentes bajo el ISM que bajo el SMM se debe a que, cuando se trata de estimar la función $F_i(\theta_0, \theta_1, \tau)$ que más se aproxima a la distribución *mismatch* de estos datos (dado que la distribución *mismatch* de los datos no forma una distribución unimodal clara), la función ajustada $F_i(\theta_0, \theta_1, \tau)$ no se ajusta bien a la distribución *mismatch* de nuestros datos. Como la función $F_i(\theta_0, \theta_1, \tau)$ no tiene un buen ajuste a la distribución

mismatch de nuestros datos, el valor de $\hat{\tau}$ puede tomar valores muy altos y volátiles, donde se puede estimar un valor de $\hat{\tau}$ mayor bajo el SMM que bajo el ISM. Lo cual no esperaríamos puesto que siempre esperamos más diferencias bajo el ISM que bajo el SMM.

Se graficó la relación entre el sesgo relativo en la estimación de $\hat{\tau}$ y todas las medidas de homoplasia. También se calculó la relación lineal entre estas medidas.

Para detallar el efecto que puede tener la homoplasia en la forma de la distribución *mismatch* se generaron mil juegos de datos con una θ_1 = 15 y una τ = 6. Se graficaron cuatro distribuciones *mismatch* promedio con microsatélites creados bajo el SMM y bajo el ISM con: 1) Los cien juegos de datos con una menor HD; 2) Los cien juegos de datos con una mayor HD; 3) Los cien juegos de datos con una menor MSH, 4) los cien juegos de datos con una mayor MSH.

3.3. Algoritmo bayesiano aproximado

Desarrollé un algoritmo bayesiano aproximado con el propósito de estimar la homoplasia, además del tiempo y la magnitud del crecimiento poblacional.

Los algoritmos bayesianos aproximados requieren de una lista de tres valores aleatorios de θ_0, θ_1 y τ. Para este propósito se hizo un programa en C, el cual genera valores aleatorios para θ_0 (que se escribe como una proporción del valor de θ_1), θ_1 y τ. El valor de τ se obtiene al multiplicar el valor que crea el generador de números aleatorios por 4*Nu* (*N* se obtiene de despejar θ_1 = 2*Nu*, donde *u* es la suma de la tasa de mutación de todos los microsatélites; como el valor de N se obtiene a partir de θ_1, es importante estimar bien θ_1 cuando se usan los ABC para estimar τ correctamente). Esos tres valores se toman de una distribución uniforme aleatoria donde se definen los valores límites para los números aleatorios. El programa genera

un número *N* de números aleatorios para cada uno de los tres parámetros y los imprime a un archivo que puede ser leído por un programa que desarrolle en esta tesis y que ejecuta el algoritmo bayesiano aproximado.

Para ejecutar el algoritmo bayesiano aproximado se modificó el programa que genera datos de microsatélites. El ABC usa el siguiente procedimiento:

1) Se lee un conjunto de datos de microsatélite y se calcula un conjunto de estadísticos *S* que son elegidos por el usuario (estos estadísticos son: el promedio de la varianza en el número de repeticiones por microsatélite, el promedio de la heterocigosis esperada por microsatélite, el número de haplotipos diferentes por estado, el número de sitios segregantes en todos los microsatélites, el número de singletones en todos los microsatélites, la varianza en el número de sitios segregantes por locus y la heterocigosis esperada tomando en cuenta todo el haplotipo).
2) Se toman tres valores aleatorios para los parámetros de θ_0, θ_1 y τ de una lista creada con el generador de números aleatorios.
3) Se simula un conjunto de datos de microsatélites usando el modelo coalescente y codificando las mutaciones de los microsatélites de acuerdo a un SMM simétrico.
4) Se calcula un conjunto de estadísticos *s* a partir de los datos de microsatélites generados.
5) Calculamos |*S*-*s*| para cada uno de los estadísticos.
6) Si |*S*-*s*| < ε para todos los estadísticos (ε representa el valor del umbral, sección 1.6). Aceptamos los valores de θ_0, θ_1, τ, SH, MASH, SASH, MSH, HD, homoplasia basada en el coalescente y homoplasia por sitios.

7) Regresamos a 2 (Continuamos regresando a 2 un número n de veces deseado).

Todos los valores de θ_0, θ_1, τ, SH, MASH, SASH, MSH, HD, CH y HS aceptados generan una distribución posterior para cada parámetro. Los parámetros pueden estimarse tomando la moda (Ross-Ibarra *et al.*, 2009) o la media (Pritchard *et al.*, 1999) de la distribución posterior. Pruebas iniciales mostraron que la moda produce mejores resultados.

Llamaré CORAGHE (Coalescent Based Rejection Algorithm for Population Growth and Homoplasy Estimation) a este programa (Todos los programas usados en esta tesis estarán disponibles en http://code.google.com/p/coraghe/. Mientras tanto, los programas pueden ser pedidos al autor de este trabajo).

3.4. Elección de los parámetros que mejor describen el crecimiento stepwise

En esta sección propongo un análisis para encontrar los estadísticos de resumen que mejor estiman los parámetros determinantes del crecimiento poblacional y la homoplasia. Todos los estadísticos propuestos tratan de encontrar una alta proporción de mutaciones en las ramas terminales.

Se simularon diez juegos de datos de microsatélites bajo el modelo SMM (Tabla 1) y se corrió CORAGHE con una ε de 0.1 (este valor de ε fue usado en Pritchard *et al.*, 1999) y nueve diferentes combinaciones de estadísticos hasta obtener cien aceptaciones en los diez juegos de datos (Tabla 2a). Los estadísticos de resumen que se dejaron fijos y con los que se realizó la primera corrida (Tabla 2a: 1) fueron el promedio de la varianza en el número de repeticiones por microsatélite, el promedio de la heterocigosis esperada por microsatélite y el número de haplotipos

diferentes por estado. En conjunto, estos estadísticos de resumen han probado su eficacia para analizar el crecimiento poblacional en análisis previos (Pritchard *et al.*, 1999; Estoup *et al.*, 2003). Se realizaron también corridas con los anteriores tres estadísticos de resumen base con alguno más que pudiera ayudar a mejorar los análisis (Tabla 2a: 2-5) y con todos los estadísticos de resumen menos la varianza en el número de sitios segregantes por locus (ya que dicho estadístico causaba un aumento muy grande del tiempo computacional, porque dicho estadístico tiene una gran varianza, lo cual aumenta el tiempo entre simulaciones aceptadas) (Tabla 2a: 6). También se analizó el efecto de quitar alguno de los tres estadísticos de resumen base (Tabla 2a: 7-9). Para todas esas corridas se usaron distribuciones *a priori* uniformes con valores de 0 a 200 para θ_1, de 0*θ_1 a 0.01*θ_1 para θ_0 y de 0*2θ_1 a 1.0*2θ_1 para τ.

Se calculó la media y el error estándar de los valores estimados de τ, MSH y HD respecto en los diez juegos de datos para cada una de las nueve combinaciones de conjuntos de estadísticos de resumen. Después se graficó la proporción de diferencia de los valores estimados respecto al valor real en valores absolutos como:

$$Proporción\ de\ la\ diferencia\ de\ los\ valores\ estimados\ \ respecto\ al\ valor\ real\ en\ valores\ absolutos = \frac{(\,|\,Valor\ estimado - valor\ real\,|)}{valor\ real} \qquad 17$$

En otro experimento, cuya motivación era encontrar los estadísticos de resumen que dieran buenos estimados de τ, MSH y HD y además no aumentaran sustancialmente el tiempo computacional, se usaron las mismas combinaciones de estadísticos de resumen que en el análisis anterior y se rechazaron aquellas combinaciones que causaran un aumento sustancial en el tiempo computacional y no mejoraran la estimación de τ,

MSH y HD. Se corrió CORAGHE con una ε de 0.1 hasta obtener 1000 aceptaciones. Con estas 1000 simulaciones se graficó la media y el error estándar en valores estimados de τ, MSH y HD y se graficó la proporción de la diferencia de los valores estimados respecto al valor real en valores absolutos. Para cada combinación de estadísticos se recorrió la ε de cada uno de los estadísticos de 0.1 a 0.01, en intervalos de 0.01 y se calculó la media y el error estándar de τ, MSH y HD. Después se graficó la proporción de la diferencia de los valores estimados respecto al valor real en valores absolutos para cada vez que se recorría el valor de ε.

Simulación	θ_1	θ_0	τ
1	30	0.03	1.5
2	30	0.03	3
3	30	0.03	4.5
4	30	0.03	6
5	30	0.03	7.5
6	30	0.03	9
7	30	0.03	10.5
8	30	0.03	12
9	30	0.03	13.5
10	30	0.03	15

Tabla 1.- Simulaciones usadas para elegir los estadísticos y las ε más informativas sobre la homoplasia y los parámetros que definen el crecimiento poblacional.

Tabla 2.- Combinaciones de estadísticos de resumen usados para la búsqueda de los mejores estimados de τ, MSH y HD con CORAGHE.

Combinación	Estadísticos usados
1	El promedio de la varianza en el número de repeticiones por microsatélite, el promedio de la heterocigosis esperada por microsatélite y el número de haplotipos diferentes por estado.

2	El promedio de la varianza en el número de repeticiones por microsatélite, el promedio de la heterocigosis esperada por microsatélite, el número de haplotipos diferentes por estado y el número de sitios segregantes en todos los microsatélites.
3	El promedio de la varianza en el número de repeticiones por microsatélite, el promedio de la heterocigosis esperada por microsatélite, el número de haplotipos diferentes por estado y el número de singletones en todos los microsatélites.
4	El promedio de la varianza en el número de repeticiones por microsatélite, el promedio de la heterocigosis esperada por microsatélite, el número de haplotipos diferentes por estado y la varianza en el número de sitios segregantes por locus.
5	El promedio de la varianza en el número de repeticiones por microsatélite, el promedio de la heterocigosis esperada por microsatélite, el número de haplotipos diferentes por estado y la heterocigosis esperada tomando en cuenta todo el haplotipo.
6	El promedio de la varianza en el número de repeticiones por microsatélite, el promedio de la heterocigosis esperada por microsatélite, el número de haplotipos diferentes por estado, el número de sitios segregantes en todos los microsatélites, el número de singletones en todos los microsatélites y la heterocigosis esperada tomando en cuenta todo el haplotipo.
7	El promedio de la varianza en el número de repeticiones por microsatélite y el promedio de la heterocigosis esperada por microsatélite.
8	El promedio de la varianza en el número de repeticiones por microsatélite y el número de haplotipos diferentes por estado.
9	El promedio de la heterocigosis esperada por microsatélite y el número de haplotipos diferentes por estado.

3.5. Estimación de la homoplasia y del crecimiento poblacional mediante un algoritmo bayesiano aproximado

Se generaron 100 conjuntos de datos, cada uno con 150 haplotipos con 6 microsatélites creados bajo el SMM, para cuatro grupos de datos como se muestra en la Tabla 3:

Tabla 3.- Simulaciones usadas para analizar la estimación de la homoplasia y los parámetros del tiempo y la magnitud del crecimiento poblacional con CORAGHE.

Grupo 1			
Número de juegos de datos de microsatélites	θ_1	θ_0	τ
10	30	0.03	1.5
10	30	0.03	3
10	30	0.03	4.5
10	30	0.03	6
10	30	0.03	7.5
10	30	0.03	9
10	30	0.03	10.5
10	30	0.03	12
10	30	0.03	13.5
10	30	0.03	15
Grupo 2			
Número de juegos de datos de microsatélites	θ_1	θ_0	τ
100	30	.03	3
Grupo 3			
Número de juegos	θ_1	θ_0	τ

de datos de microsatélites			
100	30	0.03	6
Grupo 4			
Número de juegos de datos de microsatélites	θ_1	θ_0	τ
100	30	0.03	9

Se usaron distribuciones *a priori* uniformes de 0 a 200 para θ_1, de 0*θ_1 a 0.01*θ_1 para θ_0 y de 0*2θ_1 a 1.0*2θ_1 para τ. Con dichas distribuciones prior fue posible estimar los valores de τ, MSH y HD con CORAGHE, así como el valor de τ estimado con Arlequin para cada una de las simulaciones dentro de cada grupo de datos. Se calculó la correlación lineal entre todos los valores estimados y los valores reales de τ, MSH y HD dentro de cada grupo de datos. Se calculó la media y el error estándar de los valores estimados de τ, MSH y HD en cada grupo de datos. Se graficó la proporción de la diferencia de los valores estimados respecto al valor real con y sin valores absolutos. La proporción de la diferencia de los valores estimados respecto al valor real sin el uso de valores absolutos se expresa del siguiente modo:

$$Proporción\ de\ diferencia\ de\ los\ valores\ estimados\ respecto\ al\ valor\ real = \frac{(Valor\ estimado - valor\ real)}{valor\ real} \quad 18$$

Con el fin de observar el efecto de una buena predicción de la homoplasia en la estimación de τ para cada juego de datos de los grupos 2, 3 y 4 (Tabla 2), se tomaron las 200 simulaciones con un valor de HD y MSH más cercano al valor real de HD y MSH. A partir de estas simulaciones se estimó $\hat{\tau}$ en cada juego de datos y se promedió el valor de $\hat{\tau}$ en los cien juegos de

datos dentro de cada grupo. También se tomaron las 200 simulaciones con un valor de HD y MSH más lejano al valor real para calcular $\hat{\tau}$ y se promedió el valor de $\hat{\tau}$ en los cien juegos de datos dentro de cada grupo. Por último, con 200 simulaciones tomadas al azar se calculó $\hat{\tau}$ y se promedió el valor de $\hat{\tau}$ para las cien simulaciones dentro de cada grupo. Se graficó la proporción de la diferencia de los valores estimados respecto al valor real con y sin valores absolutos.

Para verificar si otras medidas de homoplasia pueden ser estimadas correctamente con CORAGHE, se calcularon los valores de SASH, SH, CH y HS en todas las simulaciones de los Grupos 1, 2 y 4 de la Tabla 2. Se determinó la relación lineal entre estos estimados y los valores reales de SASH, SH, CH y HS.

3.6. Análisis de la homoplasia y el crecimiento poblacional en datos de *Pinus caribaea*

Se tomaron datos de seis microsatélites (cinco microsatélites simples y uno compuesto de dos microsatélites cuyas partes se separaron y se analizaron como dos microsatélites simples) de un conjunto de individuos de *Pinus caribaea* contenidos dentro de un *cluster* de BAPS (*bayesian analysis for population structure*) (datos de Jardón-Borbolla *et al.*, en preparación). De dichos datos, se omitieron los individuos procedentes de islas del Caribe para evitar que la separación entre los pinos de las islas y el continente sesguen el análisis. Se corrió CORAGHE para los datos de microsatélites con una distribución prior uniforme de 0 a 200 para θ_1, de 0*θ_1 a 0.01*θ_1 para θ_0 y de 0*2θ_1 a 1.0*2θ_1 para τ. Se graficó la distribución posterior de $\hat{\tau}$, $\widehat{\theta_1}$ y diferentes medidas de homoplasia obtenidas con CORAGHE. También se infirió $\hat{\tau}$ con Arlequin. El número de años se estimó a partir de $\hat{\tau}$ como ($\hat{\tau}$/(2**l***u*)) * tiempo generacional en *Pinus caribaea*. El tiempo generacional

en pinos se fijó en 42.5 años porque alrededor de esa edad se alcanza la máxima fecundidad para *Pinus caribaea* (según comunicación personal con López-Almirall en Jardón-Borbolla *et al.,* en preparación). Como en los datos tenemos siete microsatélites, *l* se iguala a 7. La tasa de mutación por microsatélite por generación, *u*, se dejó igual a $5.5 * 10^{-5}$ (se utiliza esa tasa de mutación porque es un valor promedio de la tasa de mutación observada en microsatélites de *Pinus torreyana* en Provan *et al.*, 1999).

Para analizar el efecto del tiempo generacional en el estudio de las expansiones poblacionales, se calculó la subestimación de la expansión poblacional en años utilizando tiempos generacionales que van de 15 a 100 años dejando una *u* fija de $5.5 * 10^{-5}$. Tanto 15 como 100 años son tiempos que previamente habían sido usados en estudios de expansión poblacional en *Pinus ayacahuite* y *Pinus strobiformis* (Moreno-Letelier, 2009).

El efecto de la tasa de mutación se analizó dejando constante el tiempo generacional en 42.5 años y utilizando tasas de mutación de microsatélites que van de $3.2 * 10^{-5}$ a $7.9 * 10^{-5}$ mutaciones por generación, que son las tasas de mutación que podemos encontrar en los microsatélites de cloroplasto de *Pinus torreyana* (Provan, 1999).

Análisis de resultados

4.1. Uso de la distribución *mismatch* con microsatélites para inferir crecimiento poblacional

4.1.1. Forma de la distribución *mismatch* con microsatélites

Se graficaron las distribuciones *mismatch* promedio de cien simulaciones de datos de microsatélites generados bajo el SMM con una cierta combinación de los parámetros θ_0, θ_1 y τ (ver sección 1.5) (Figura 7a, 7c, 7e, 7g, 7i y 7k). De acuerdo con Rogers y Harpending (1992), debe cumplirse que: 1) Valores mayores de θ_1 provoquen que disminuya el valor donde la distribución cruzará al eje de las ordenadas (la distribución cruza al eje de las ordenadas donde el número de diferencias entre pares de secuencias es igual a 0). 2) Valores mayores de τ muevan la cresta de la distribución *mismatch* hacia la derecha. El primer principio se cumple en las distribuciones *mismatch* graficadas. El segundo principio se mantiene con una pequeña variación. A partir de ciertos valores de τ, las distribuciones *mismatch* promedio ya no se mueven a la derecha y su distribución *mismatch* promedio se asemeja mucho a la que esperaríamos cuando no hay crecimiento poblacional. Cuando el valor de θ_1 es menor, en valores más pequeños de τ la distribución *mismatch* promedio de datos simulados bajo crecimiento poblacional toma una forma más parecida a la distribución *mismatch* promedio generada con datos de una población con tamaño estable. Otra forma de visualizar el efecto de θ_1 en la distribución *mismatch* es con la bondad de ajuste. Cuantificando las diferencias entre las distribuciones *mismatch* promedio de datos con crecimiento poblacional contra datos de un modelo sin expansión poblacional, se observa que en

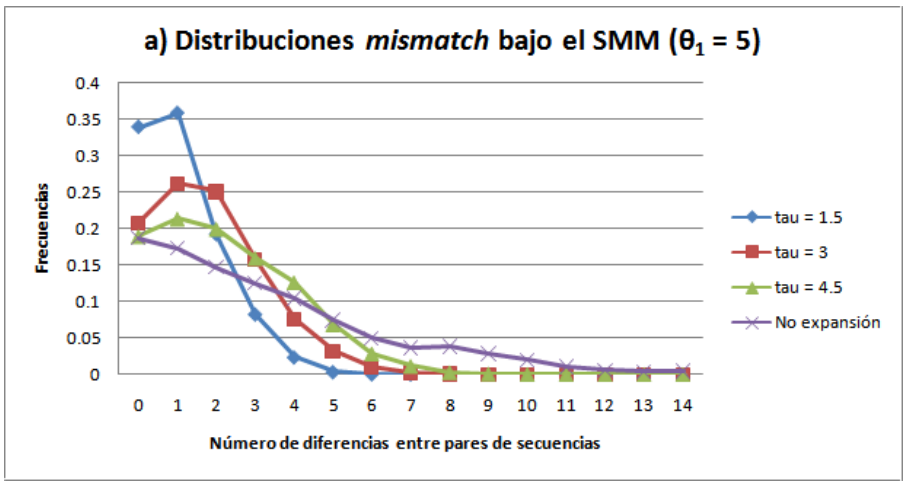
valores más grandes de τ la bondad de ajuste (Figuras 8a, 8b y 8c) tiene valores más bajos cuando los valores de θ_1 son menores. Lo cual indica que con valores más bajos de θ_1 tendremos menos señal para identificar expansiones antiguas, puesto que la distribución *mismatch* que obtendremos se parecerá más a la obtenida cuando no existe crecimiento poblacional a valores más bajos de τ.

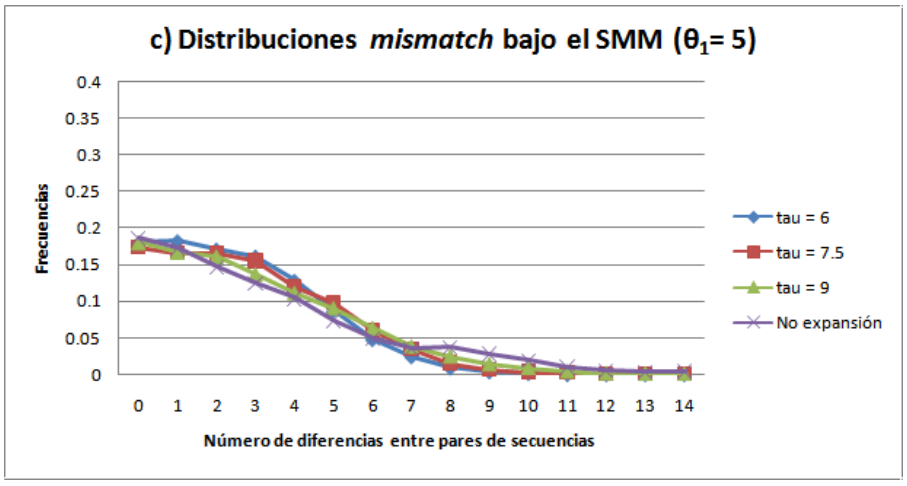
4.1.2. Efecto de la homoplasia en la distribución *mismatch*

Se graficó la distribución *mismatch* de microsatélites creados bajo el ISM a partir de las mismas genealogías usadas en el apartado anterior (Figura 7b, 7d, 7f, 7h, 7j y 7l). Con los microsatélites creados bajo el ISM se registran las mutaciones que ocurrieron en cada microsatélite y no hay posibilidad de que dos mutaciones lleven al mismo estado. Por lo tanto, no existe homoplasia con microsatélites creados bajo el ISM. Es útil analizar las distribuciones *mismatch* de datos bajo el ISM porque nos muestran las distribuciones *mismatch* que esperaríamos si en los microsatélites no existiera homoplasia.

La comparación de las distribuciones *mismatch* con microsatélites bajo el SMM y microsatélites bajo el ISM muestran que, con datos generados con una θ_1 de 15 y de 30, en valores de τ mayores a 6 la cresta de la distribución *mismatch* bajo el SMM no se recorre tanto a la derecha como cuando nos encontramos bajo un ISM. El valor de τ que se estima usando la distribución *mismatch* depende de la cresta de la distribución *mismatch* y la distribución *mismatch* parte del supuesto de que nuestras secuencias están bajo el ISM. Esto nos indica que el valor de τ estimado usando la distribución *mismatch* con microsatélites, cuya evolución puede modelarse con el SMM, debería estar subestimando el valor real de τ.

Figura 7


a) Distribuciones *mismatch* bajo el SMM (θ_1 = 5)
Frecuencias
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14
Número de diferencias entre pares de secuencias
tau = 1.5
tau = 3
tau = 4.5
No expansión


c) Distribuciones *mismatch* bajo el SMM (θ_1= 5)
Frecuencias
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14
Número de diferencias entre pares de secuencias
tau = 6
tau = 7.5
tau = 9
No expansión

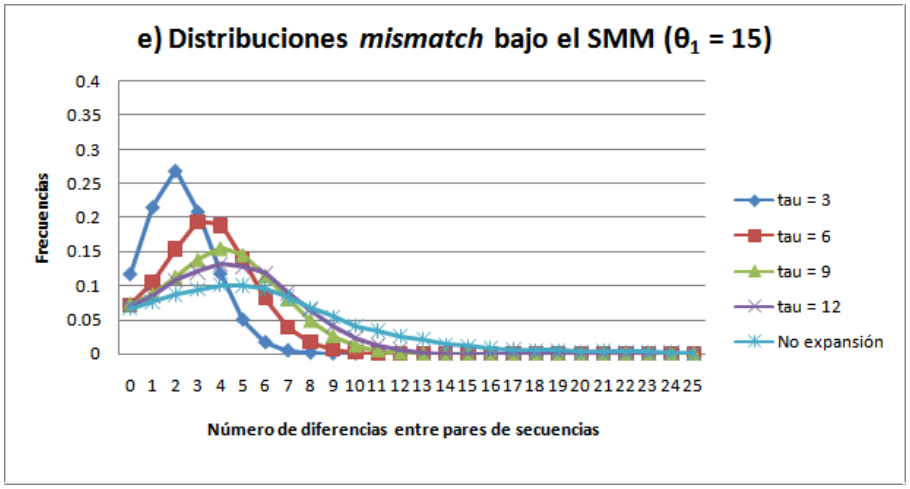
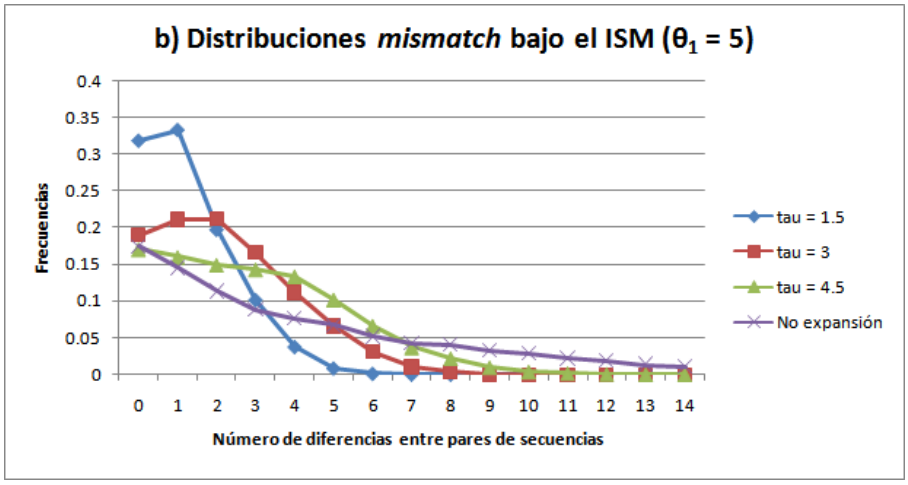
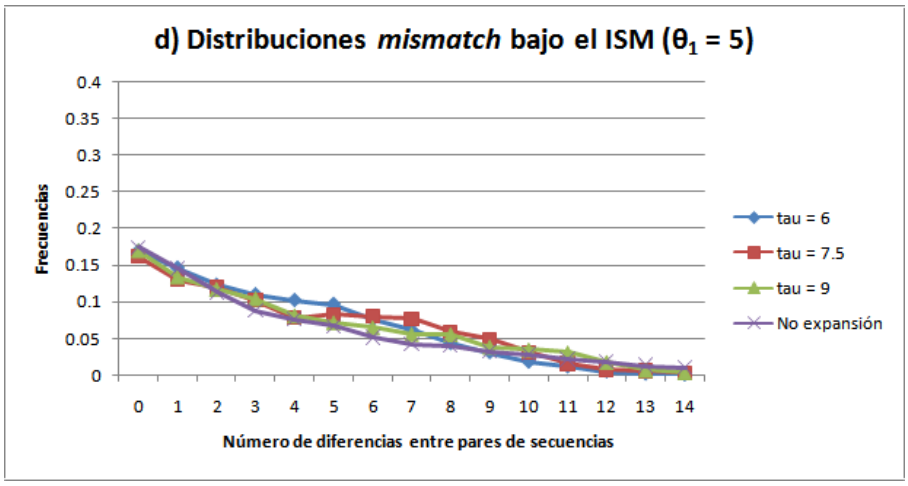

e) Distribuciones *mismatch* bajo el SMM (θ_1 = 15)
Frecuencias
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
Número de diferencias entre pares de secuencias
tau = 3
tau = 6
tau = 9
tau = 12
No expansión

Figura 7 (continuación)





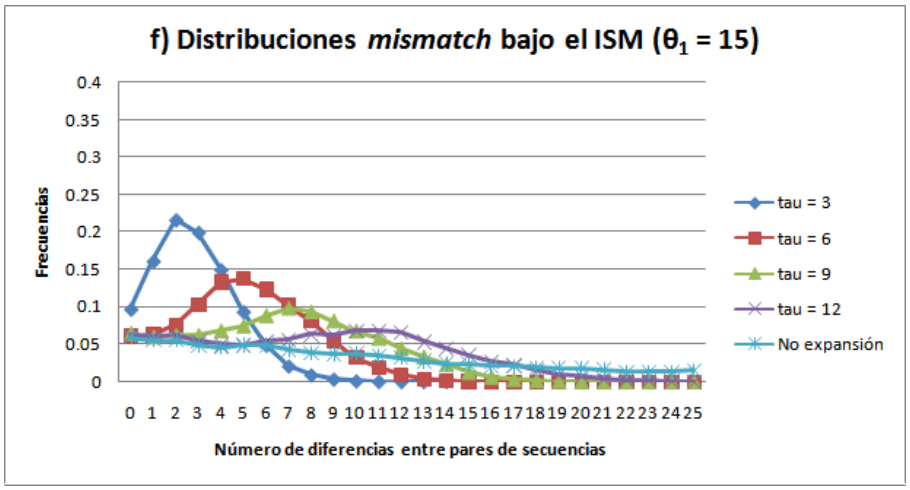
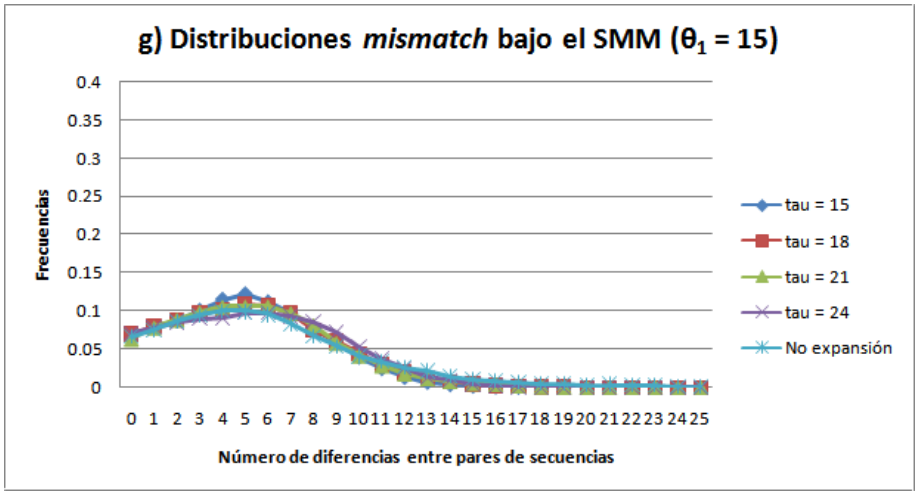
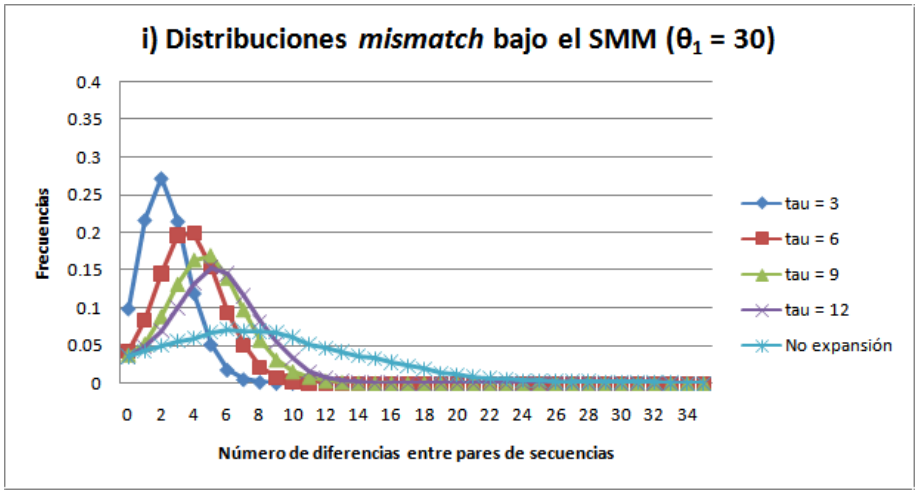
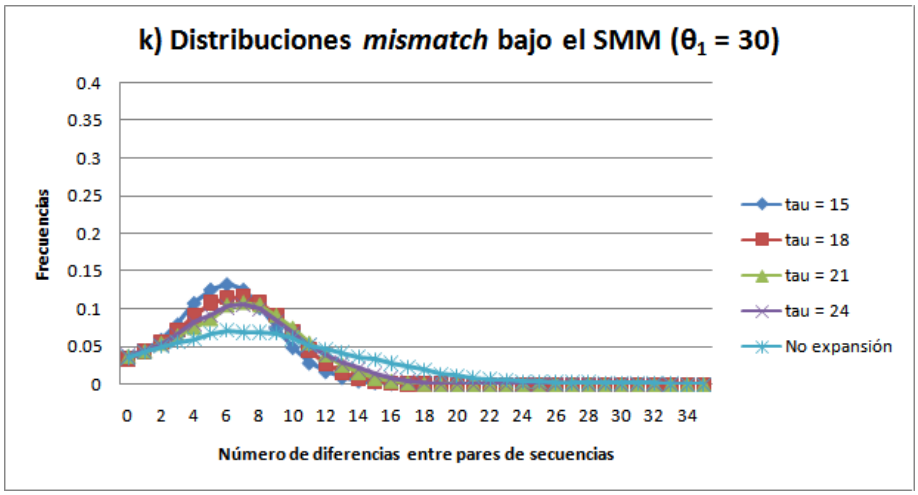


Figura 7 (continuación)







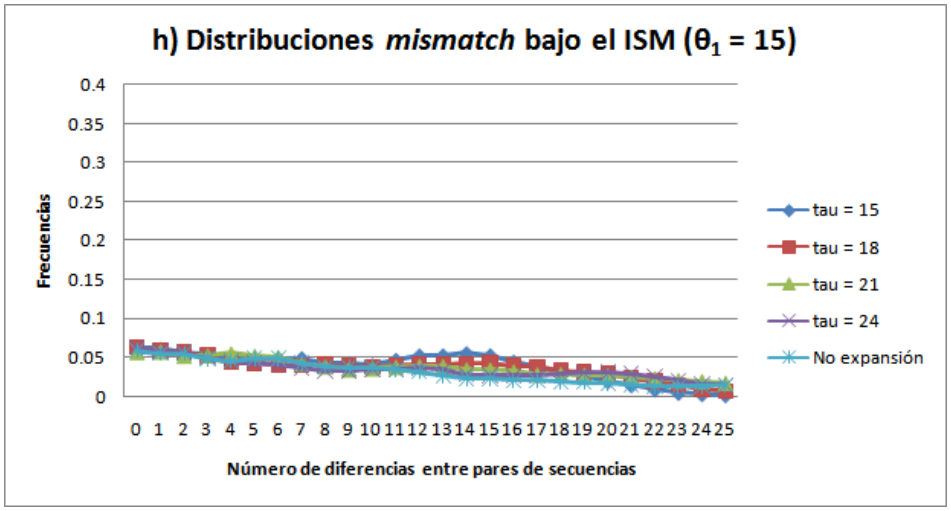
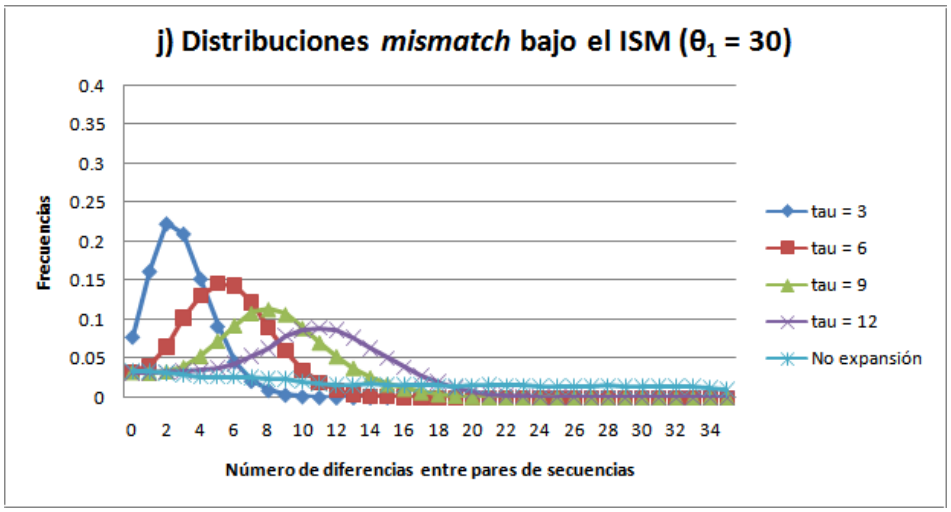


Figura 7 (continuación)



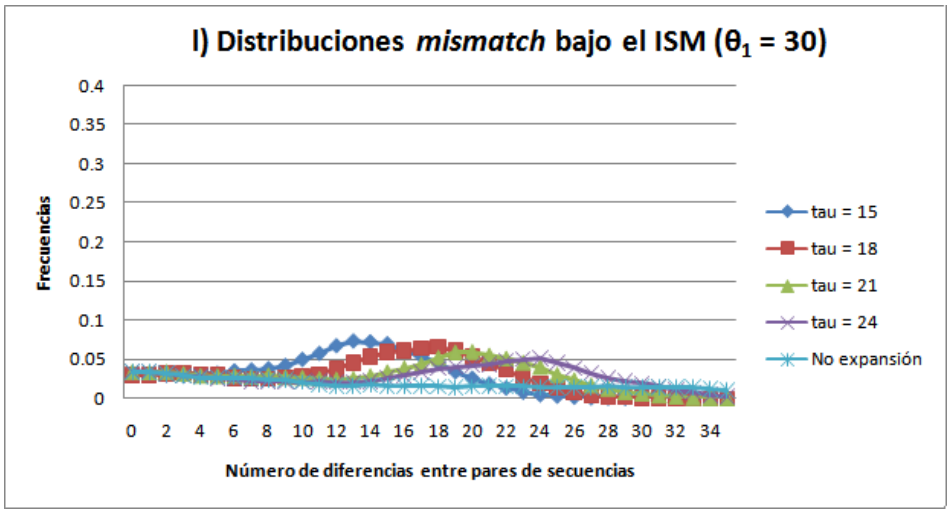


Figura 7.- Distribuciones *mismatch* promedio de cien conjuntos de datos de microsatélites generados bajo el SMM y el ISM con una $\theta_1 = 5$ [a), b), c) y d)], $\theta_1 = 15$ [e), f), g) y h)] y $\theta_1 = 30$ [i), j), k) y l)] con diferentes valores de τ. También se incluye la gráfica de una distribución *mismatch* promedio de cien conjuntos de datos de microsatélites generados con el SMM y el ISM en una población con tamaño estable y una $\theta = 5$ [a), b), c) y d)], $\theta = 15$ [e), f), g) y h)] y $\theta = 30$ [i), j), k) y l)].

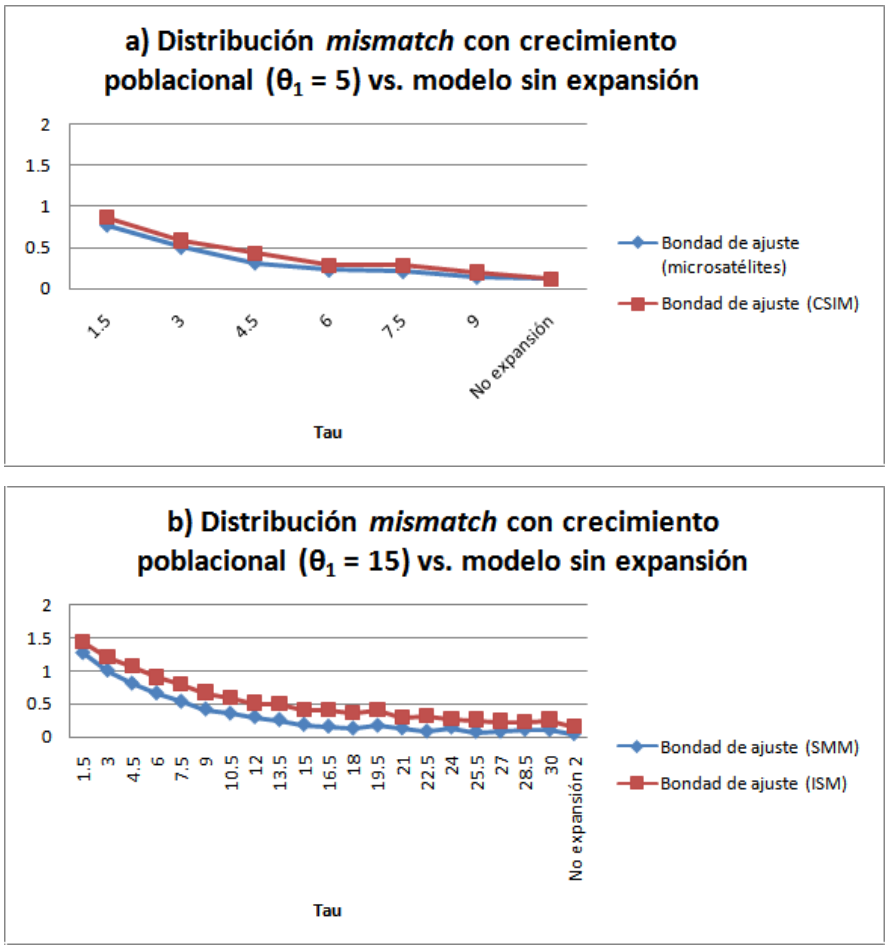
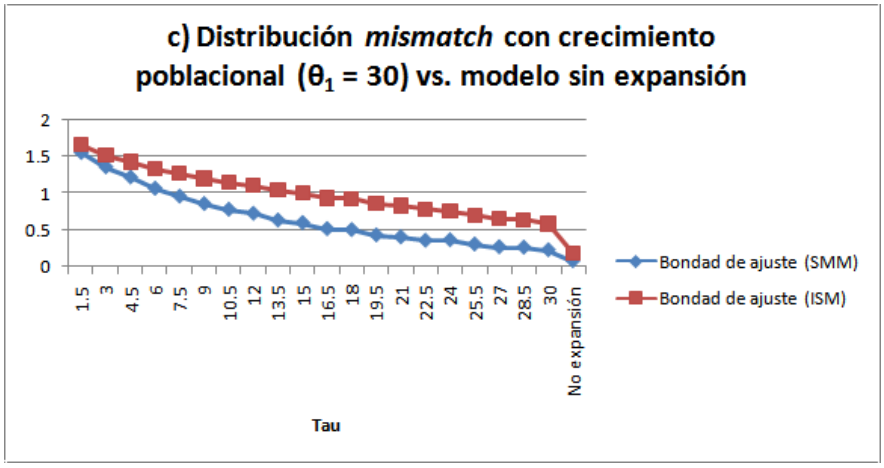
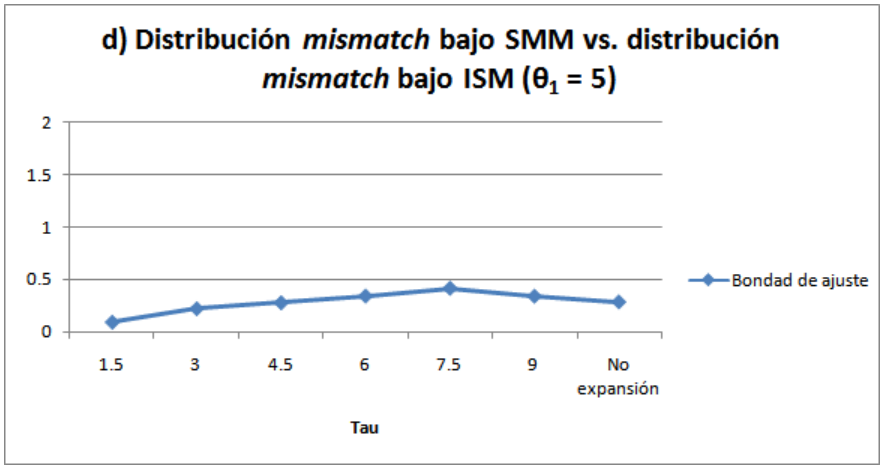


Figura 8





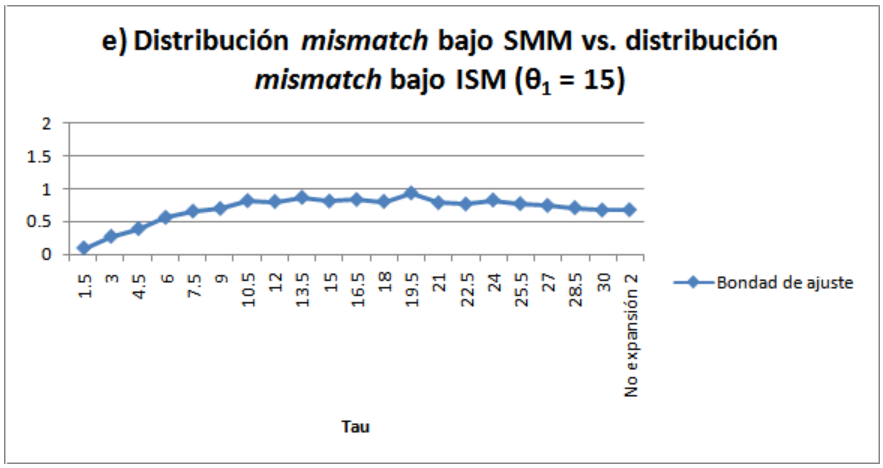


Figura 8 (continuación)

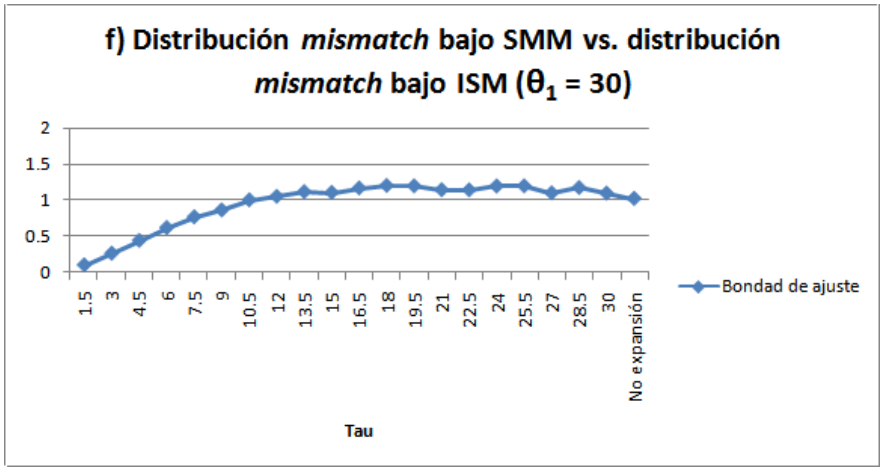


Figura 8.- Gráficas de la bondad de ajuste entre dos distribuciones *mismatch*.
Se compararon las distribuciones *mismatch* promedio de 100 simulaciones con crecimiento poblacional contra una distribución *mismatch* promedio de 100 simulaciones sin crecimiento poblacional con microsatélites creados bajo el ISM y bajo el SMM con diferentes valores de τ y valores de θ_1 = a) 5, b) 15, c) 30.
También se comparan las distribuciones *mismatch* promedio de 100 simulaciones de microsatélites creados bajo el ISM contra las distribuciones *mismatch* promedio de 100 simulaciones de microsatélites creados bajo el SMM con diferentes valores de τ y valores de θ_1 = d) 5, e) 15, f) 30.

En las gráficas de bondad de ajuste (Figura 8a, 8b y 8c) observamos que las diferencias entre la distribución *mismatch* de una población sin expansión poblacional y una población con crecimiento poblacional aumentan bajo un ISM, comparadas a cuando se crean microsatélites bajo un SMM. Esto nos muestra que cuando no existe homoplasia, hay mayores diferencias entre las distribuciones *mismatch* generadas si tenemos o no crecimiento poblacional.

Se compararon las distribuciones *mismatch* usando microsatélites cuyas mutaciones siguen un SMM y un ISM (Figura 8d, 8e y 8f). Si la homoplasia no fuera importante, esperaríamos que ambas distribuciones tuvieran una bondad de ajuste baja. Se observa que al crear datos de microsatélites con valores de τ mayores, hay más diferencias entre las distribuciones *mismatch* usando microsatélites bajo un SMM, respecto a las distribuciones *mismatch* creadas con microsatélites bajo un ISM. Esta

diferencia es producto de la homoplasia. Cuando aumenta el valor de τ, hay más tiempo para acumular mutaciones en la genealogía, dado que se observarán ramas más largas antes de la reducción del tamaño de la población yendo del presente al pasado. Un número mayor de mutaciones incrementa la posibilidad de que se obtengan dos microsatélites homoplásicos. Si vamos del presente al pasado, veremos una reducción en el tamaño de la población en el pasado, lo cual hará difícil acumular mutaciones, puesto que las ramas de la genealogía se harán más cortas en el pasado y las mutaciones que afectan cada rama del árbol dependen de las longitudes de las ramas del árbol.

4.1.3. Exactitud de los valores de $\hat{\tau}$ inferidos con la distribución *mismatch*

Se calculó la media y la varianza de los valores de $\hat{\tau}$ inferidos en cada grupo de datos (Tabla suplementaria 2). La media de $\hat{\tau}$ calculada con microsatélites bajo el SMM resulta siempre menor que la media de $\hat{\tau}$ estimada con microsatélites bajo el ISM. Utilizando la media de cien valores de $\hat{\tau}$, vemos que la $\hat{\tau}$ estimada con microsatélites bajo el SMM se subestima más en la medida que la expansión es más antigua, esto es congruente con resultados obtenidos en estudios anteriores donde se encuentra el mismo resultado (Navascués *et al.*, 2006). El valor real de τ es muy cercano a la media de cien valores estimados usando la distribución *mismatch* con microsatélites bajo el ISM, siempre y cuando el valor de τ no sea muy alto. También se encontró que cuando se generan datos sin crecimiento poblacional, los valores de $\hat{\tau}$ estimados en dichos datos son similares a los valores de $\hat{\tau}$ estimados con datos generados en expansiones antiguas (Tabla suplementaria 2). También los estimados de $\hat{\tau}$ son susceptibles a tener una gran varianza en expansiones antiguas puesto que la señal de

expansión es poco clara y la distribución *mismatch* se asemeja a la esperada en caso de que no hubiera crecimiento poblacional.

Para calcular la exactitud, se analizó el número de veces que el verdadero valor de τ está dentro del intervalo de confianza con una α de 5% (CI 95%) para todas las simulaciones (Tabla suplementaria 3). Hay dos resultados contrastantes para el número de veces que τ cae dentro del intervalo de confianza cuando se usan diferentes modelos de evolución para los microsatélites. Conforme el valor de τ aumenta, es menos probable que $\hat{\tau}$ se encuentre dentro del CI 95% si se utilizan microsatélites bajo el SMM. Estudios anteriores han llegado a este mismo resultado (Navascués *et al.*, 2006). Si usamos microsatélites creados bajo el ISM, conforme el valor de τ aumenta, no se observa un decremento en el número de veces que el verdadero valor de τ cae dentro del CI 95%, hasta que las expansiones son muy antiguas. Para θ_1= 15, con una τ de 15 empieza a disminuir el número de veces que τ se encuentra en los CI 95%. Para una θ_1= 30 y una θ_1= 5, en el rango de valores de τ usados no se nota cuándo empieza a disminuir el número de veces que τ no se encuentra en los CI 95%.

Al combinar los dos análisis de este apartado, se concluye que $\hat{\tau}$ se subestima usando microsatélites en expansiones antiguas. Además, se concluye que la homoplasia no es el único factor que influye en un buen estimado de $\hat{\tau}$. Esto se muestra porque al usar microsatélites que están bajo un ISM no siempre el verdadero valor de τ cae dentro del CI 95%. Por lo tanto, la estocasticidad del proceso genealógico puede causar un mal estimado de $\hat{\tau}$, aunque en esta tesis sólo se analizará cómo la homoplasia puede causar malos estimados de $\hat{\tau}$.

4.1.4. Relación de la homoplasia con el sesgo en el $\hat{\tau}$

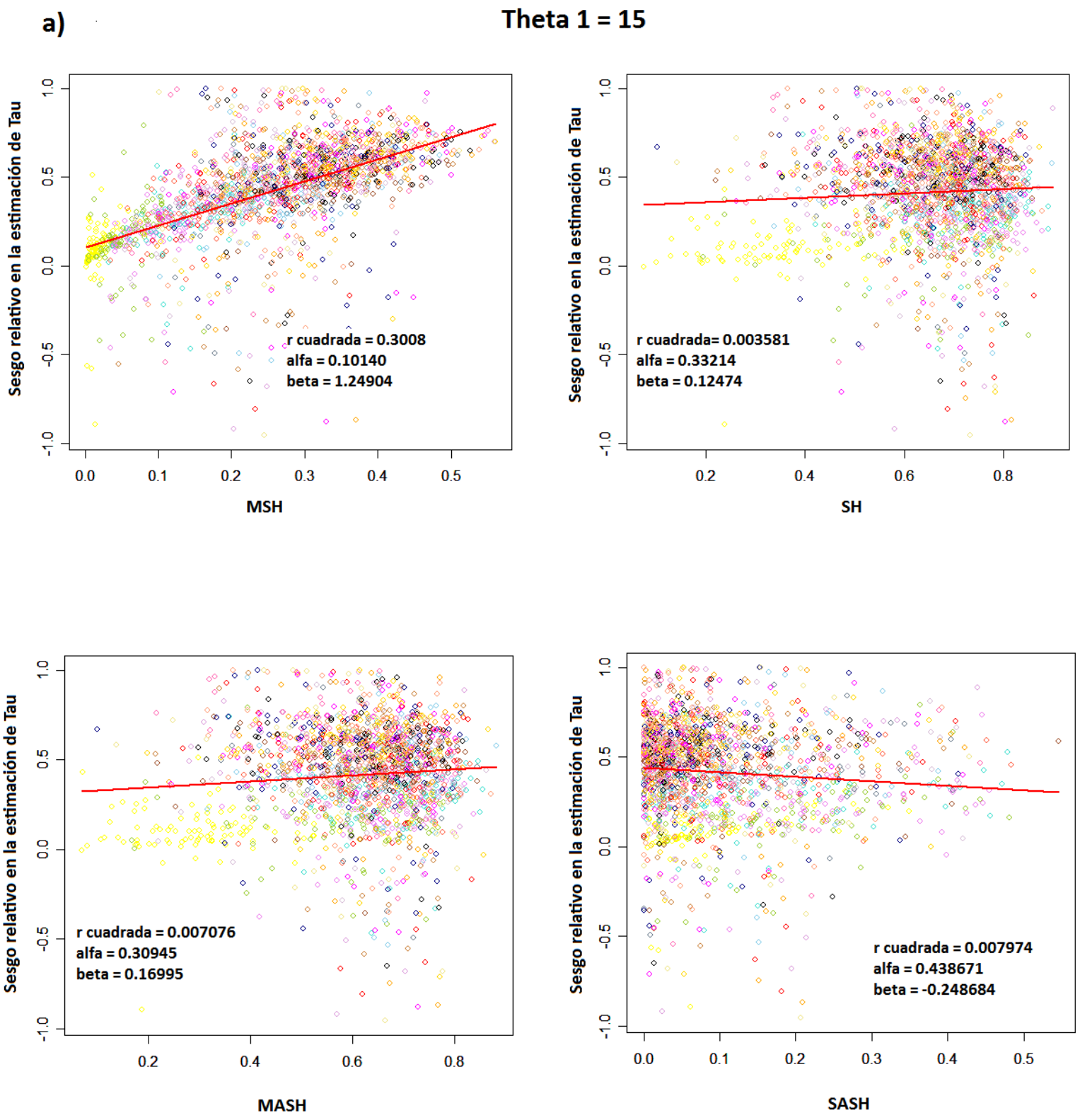
La homoplasia causa una subestimación de $\hat{\tau}$. Es necesario encontrar la medida de la homoplasia más relacionada con el sesgo en la estimación de $\hat{\tau}$. Para ello se estudió la relación lineal que existe entre el sesgo relativo en la estimación de $\hat{\tau}$ y varias medidas de homoplasia (Figura 9).

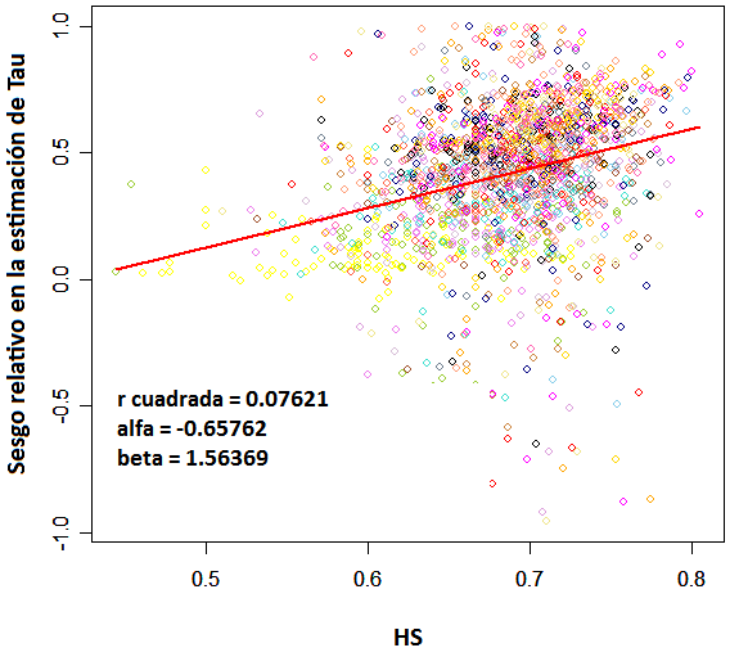
No se analizaron los datos generados con θ_1 = 5 porque se encontró que en datos con valores de θ_1 más bajos es más difícil encontrar señales de expansiones más antiguas (Figura 6). También en el grupo de datos con una θ_1 = 15 se rechazó el 8.2 % (164 de los 2000 juegos de datos) de los datos y en el grupo de datos con una θ_1 = 30 (23 de los 2000 juegos de datos) se rechazó el 1.15 % de los datos porque en dichas simulaciones se obtenía un sesgo relativo en la estimación de $\hat{\tau}$ menor a -1. Esto se debe a que en dichas simulaciones la función ajustada $F_i(\theta_0, \theta_1, \tau)$ no tiene un buen ajuste con la distribución *mismatch* y esto produce valores de $\hat{\tau}$ mucho más altos bajo el SMM que bajo el ISM, lo cual no es lógico porque siempre deberíamos encontrar más diferencias entre secuencias bajo el ISM que bajo el SMM. La mayoría de las simulaciones rechazadas corresponden a datos generados con una τ que es tan antigua que los datos generados son parecidos a los que esperaríamos si no hubiera expansión poblacional (Tabla suplementaria 4).

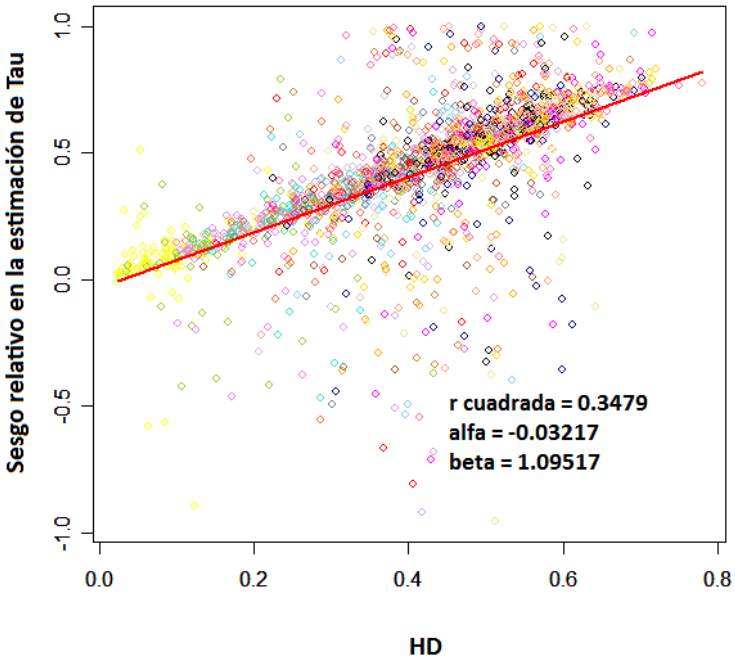
Se encontró que las dos medidas de homoplasia que más relación tenían con una subestimación de $\hat{\tau}$ son HD y MSH, donde HD tiene una mayor relación con el sesgo en la estimación de $\hat{\tau}$ (esto se deduce porque la r^2 entre HD y el sesgo en la estimación de $\hat{\tau}$ es mayor que la r^2 entre MSH y el sesgo en la estimación de $\hat{\tau}$) (Figura 9, Tabla suplementaria 5). También es relevante señalar que a partir de esta gráfica podemos discernir cuál es el sesgo que esperaríamos en τ a partir de un cierto valor de HD o

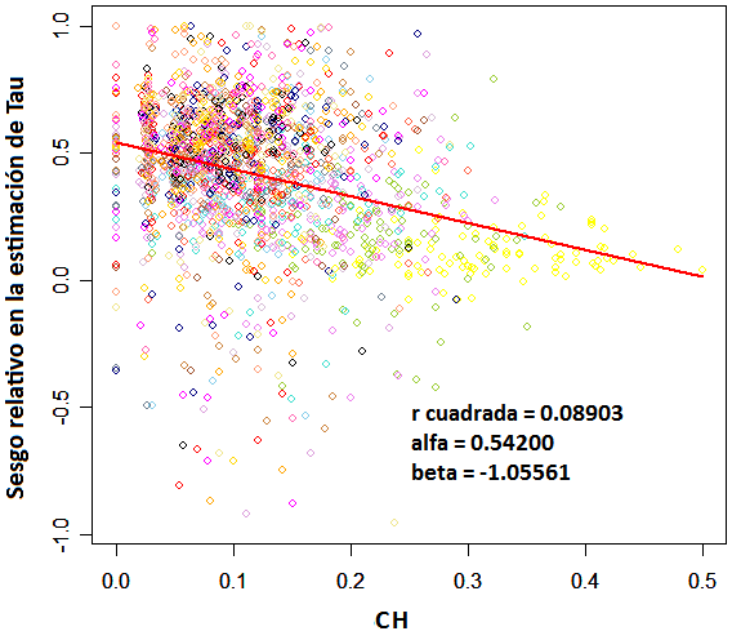
MSH. A lo largo del rango de valores de τ, HD se relaciona mejor con el sesgo en τ que MSH. Las demás medidas de homoplasia no tienen una relación tan fuerte con el sesgo en la estimación de τ. Por lo tanto, nos concentraremos en HD y MSH para cuantificar el sesgo que puede tener $\hat{\tau}$ por efecto de la homoplasia.

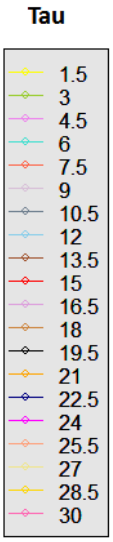
Un ejemplo que nos ilustra el efecto que puede tener la homoplasia se muestra en la forma de ocho distribuciones *mismatch* promedio generadas a partir de mil juegos de datos (Figura 10). Si tomamos los 100 juegos de datos con el valor más bajo de MSH, las distribuciones *mismatch* promedio generadas con microsatélites bajo el SMM y el ISM tienen la cresta en un valor muy parecido, lo cual se refleja en que los valores de τ estimados son muy semejantes (la diferencia promedio entre los cien valores de τ estimados con microsatélites bajo el SMM y bajo el ISM es de 0.49749). Si tomamos los 100 juegos de datos con los valores más altos de MSH, la cresta promedio se ubica en valores visiblemente distintos y la diferencia entre los valores de τ estimados con los microsatélites generados bajo el SMM y el ISM resulta más amplia (3.56717). Si realizamos el mismo análisis con la HD, observamos el mismo efecto en el cambio de las crestas en condiciones de alta y baja HD. Esto también provoca un cambio en la diferencia de los valores de τ estimados con los microsatélites bajo el ISM y el SMM (el promedio de la diferencia en condiciones de bajo HD es de 0.58032 y en condiciones de alta HD es de 3.46445). Por lo tanto, HD y MSH son medidas que pueden ayudarnos a cuantificar el sesgo relativo en $\hat{\tau}$.

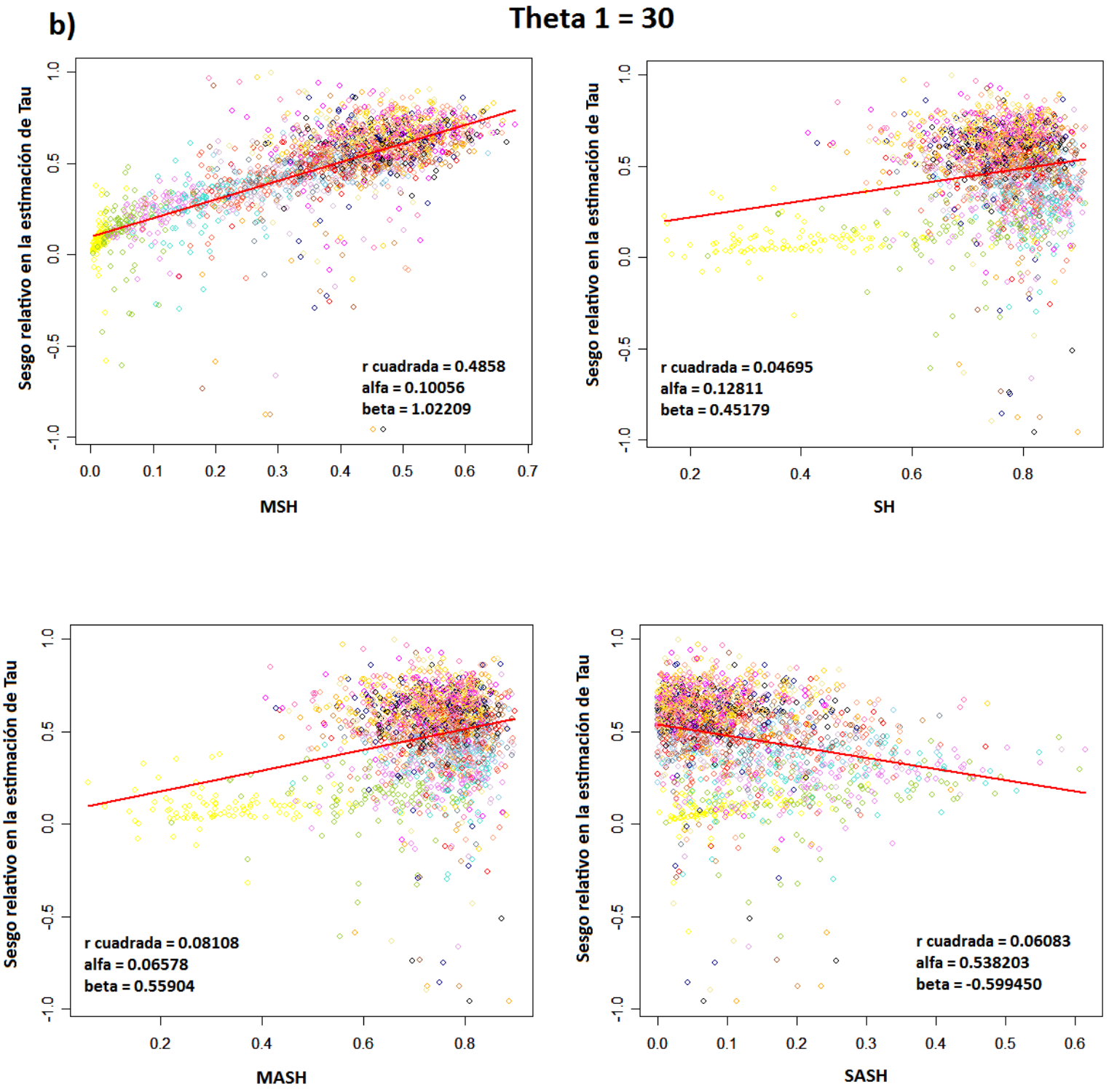

a)
Theta 1 = 15
Sesgo relativo en la estimación de Tau
MSH
r cuadrada = 0.3008
alfa = 0.10140
beta = 1.24904
SH
r cuadrada= 0.003581
alfa = 0.33214
beta = 0.12474
MASH
r cuadrada = 0.007076
alfa = 0.30945
beta = 0.16995
SASH
r cuadrada = 0.007974
alfa = 0.438671
beta = -0.248684

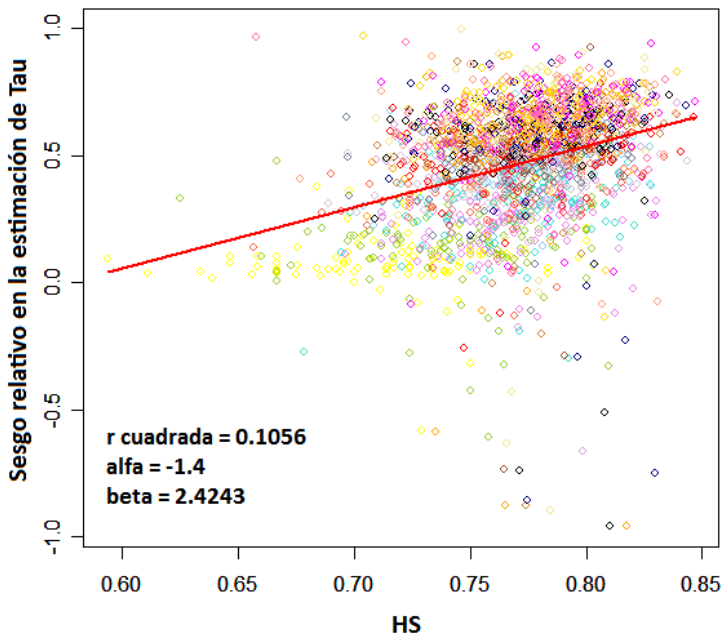

Sesgo relativo en la estimación de Tau
HS
r cuadrada = 0.07621
alfa = -0.65762
beta = 1.56369

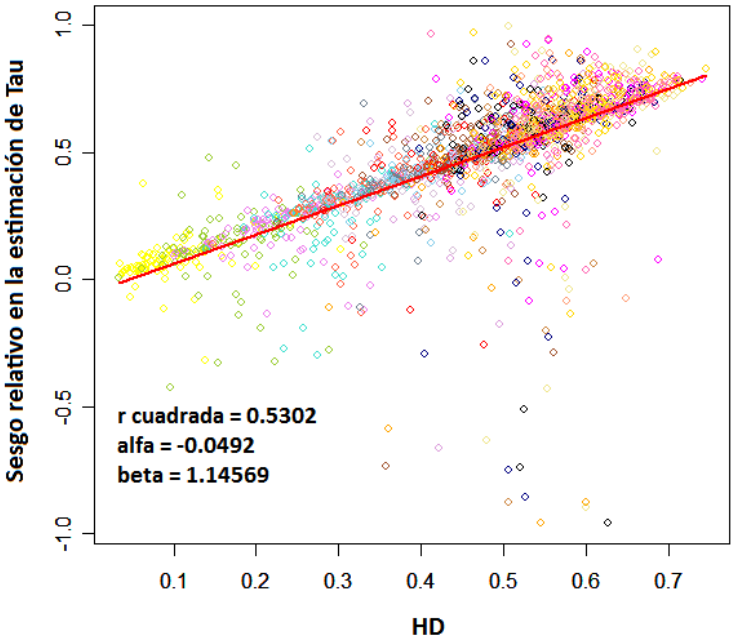

Sesgo relativo en la estimación de Tau
HD
r cuadrada = 0.3479
alfa = -0.03217
beta = 1.09517

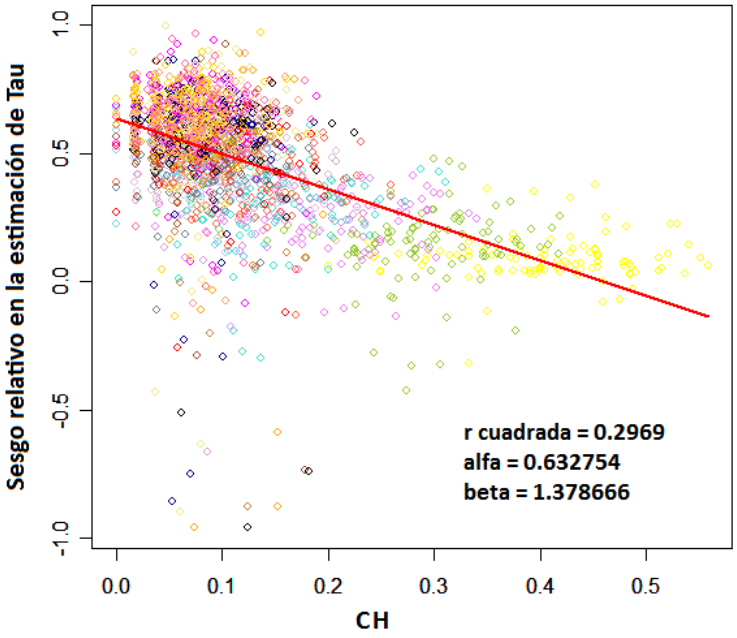

Sesgo relativo en la estimación de Tau
CH
r cuadrada = 0.08903
alfa = 0.54200
beta = -1.05561


Tau
1.5
3
4.5
6
7.5
9
10.5
12
13.5
15
16.5
18
19.5
21
22.5
24
25.5
27
28.5
30


b)
Theta 1 = 30
Sesgo relativo en la estimación de Tau
MSH
r cuadrada = 0.4858
alfa = 0.10056
beta = 1.02209
SH
r cuadrada = 0.04695
alfa = 0.12811
beta = 0.45179
MASH
r cuadrada = 0.08108
alfa = 0.06578
beta = 0.55904
SASH
r cuadrada = 0.06083
alfa = 0.538203
beta = -0.599450


Sesgo relativo en la estimación de Tau
1.0
0.5
0.0
-0.5
-1.0
r cuadrada = 0.1056
alfa = -1.4
beta = 2.4243
0.60
0.65
0.70
0.75
0.80
0.85
HS


Sesgo relativo en la estimación de Tau
1.0
0.5
0.0
-0.5
-1.0
r cuadrada = 0.5302
alfa = -0.0492
beta = 1.14569
0.1
0.2
0.3
0.4
0.5
0.6
0.7
HD


Sesgo relativo en la estimación de Tau
1.0
0.5
0.0
-0.5
-1.0
r cuadrada = 0.2969
alfa = 0.632754
beta = 1.378666
0.0
0.1
0.2
0.3
0.4
0.5
CH

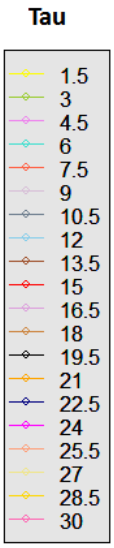

Tau
1.5
3
4.5
6
7.5
9
10.5
12
13.5
15
16.5
18
19.5
21
22.5
24
25.5
27
28.5
30

Figura 9.- Relación lineal entre el sesgo relativo en la estimación de τ y varias medidas de homoplasia en escenarios de crecimiento poblacional con distintos valores de τ y de θ_1.

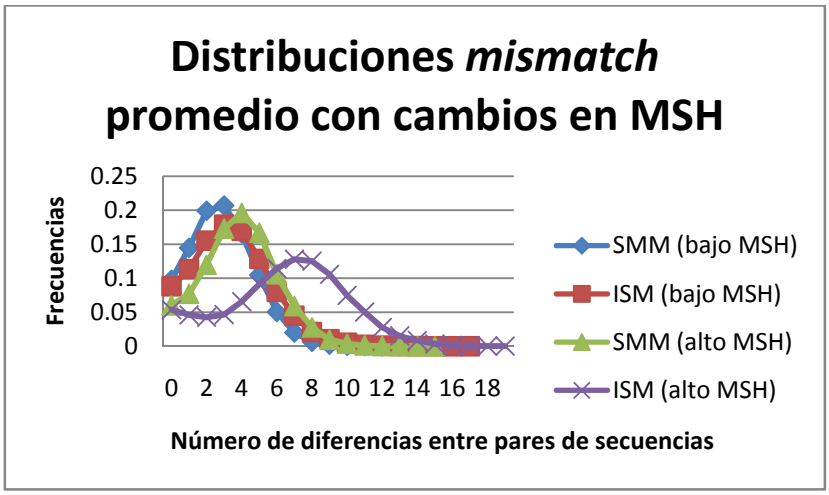


Figura 10

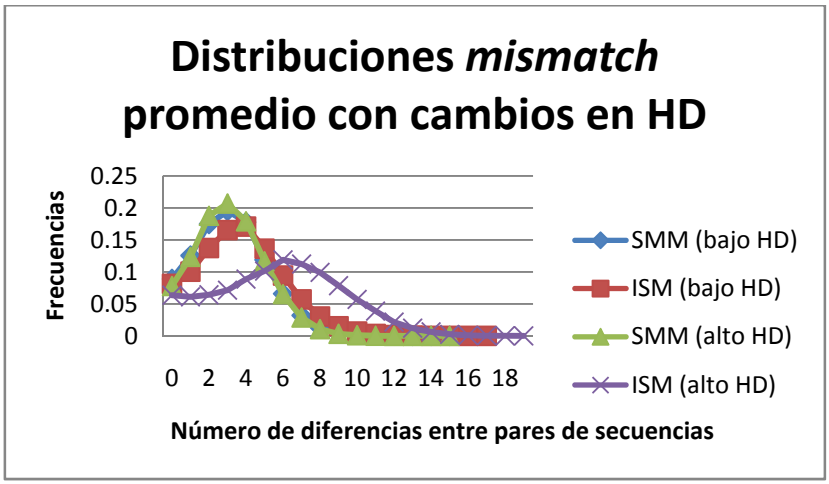
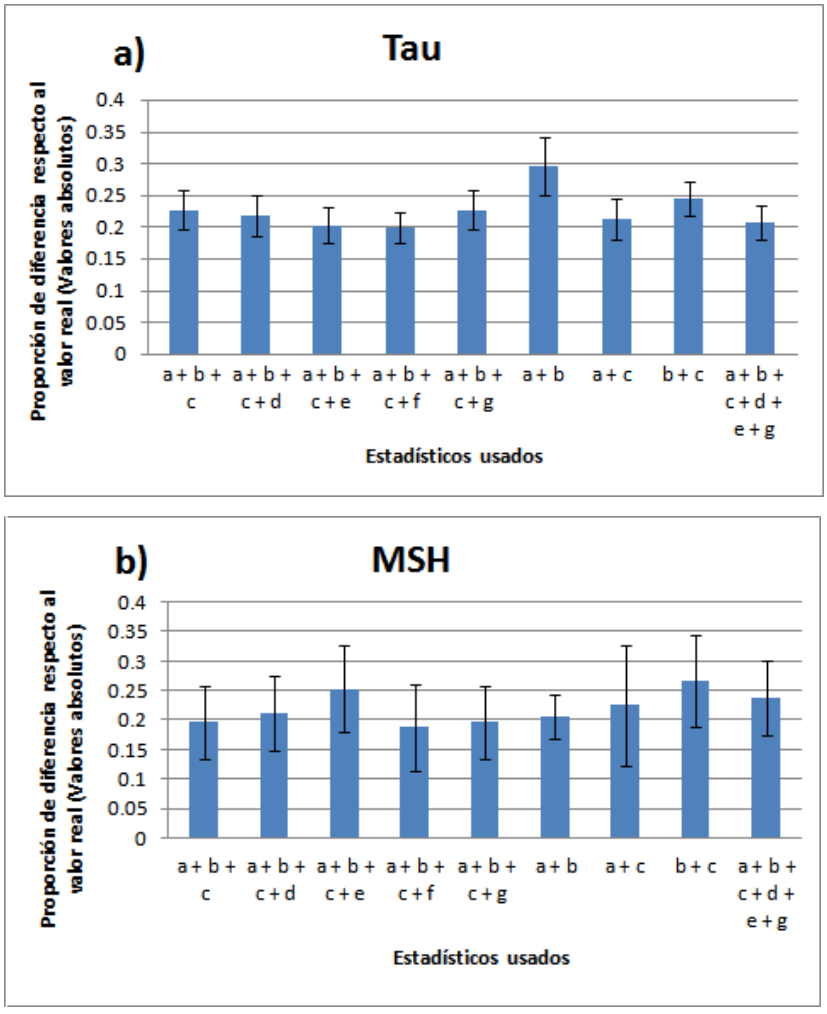


Figura 10.- Distribuciones *mismatch* promedio en diferentes condiciones de MSH y de HD.

4.2. Elección de los mejores estadísticos para correr un algoritmo bayesiano aproximado

4.2.1. Corridas con bajo número de simulaciones aceptadas

Con el fin de elegir los estadísticos más apropiados para cuantificar $\hat{\tau}$, MSH y HD, se estimaron los valores de $\hat{\tau}$, MSH y HD a partir de una distribución posterior generada con 100 simulaciones aceptadas con CORAGHE usando nueve combinaciones diferentes de estadísticos y una ε de 0.1 (ver sección 3.3) (Figura 11) en diez juegos de datos de microsatélites (ver Tabla 1).



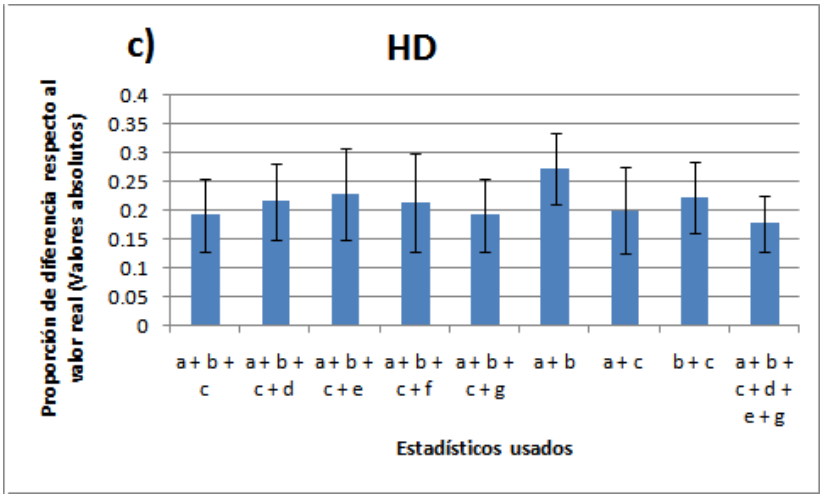


Figura 11.- Sesgo en las estimaciones de τ, MSH y HD con diferentes combinaciones de estadísticos en cien simulaciones aceptadas. Se graficó el promedio y el error estándar de la media (las barras) del porcentaje de la diferencia respecto al valor real de τ, MSH y HD en valores absolutos de diez datos simulados. Las letras debajo de las columnas indican la combinación de estadísticos usados. Las letras significan: a) El promedio de la varianza en el número de repeticiones por microsatélite; b) El promedio de la heterocigosis por microsatélite; c) El número de haplotipos diferentes por estado; d) Número de sitios segregantes en todos los microsatélites; e) Número de singletones en todos los microsatélites; f) La varianza en el número de sitios segregantes por locus y g) La heterocigosis tomando en cuenta todo el haplotipo.

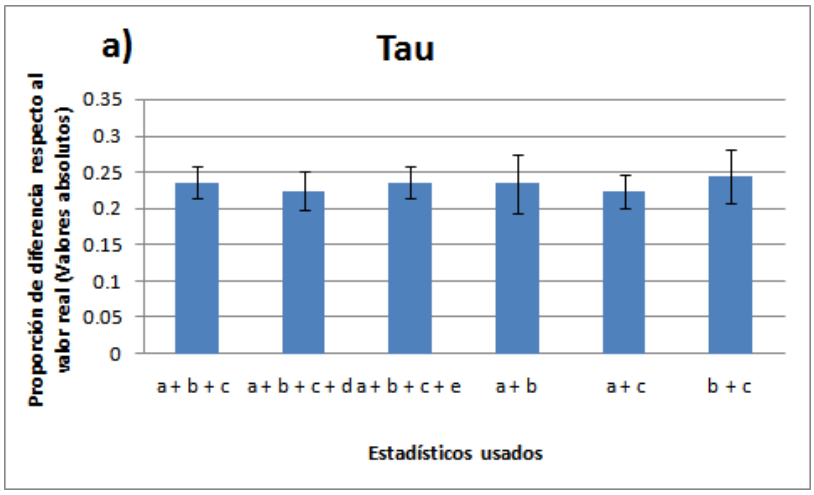
Para seis de las corridas, se dejaron como base tres estadísticos de resumen, que son: el promedio de la varianza en el número de repeticiones por microsatélite, el promedio de la heterocigosis por microsatélite y el número de haplotipos diferentes por estado. Se dejaron fijos porque han probado su utilidad en anteriores estudios de crecimiento poblacional (Pritchard *et al.*, 1999; Estoup *et al.*, 2003; Estoup *et al.*, 2004). Se encontró que la adición de ningún estadístico de resumen propuesto en esta tesis a los tres estadísticos de resumen dejados como base mejoraba significativamente las estimaciones de $\hat{\tau}$, MSH y HD. Se realizó un análisis posterior con las mismas corridas de este apartado, pero se aceptó un número mayor de simulaciones (1000). Para dicho análisis se rechazaron

aquellas corridas que no mejoraban significativamente los estimados de $\hat{\tau}$, MSH y HD, además de aumentar en forma notoria el tiempo computacional (ver Tabla suplementaria 7 para un análisis del número de corridas necesarias para obtener cien aceptaciones). Ya no se utilizó la media de la varianza en el número de sitios segregantes por locus y el número de singletones en los siguientes análisis porque causaban un aumento importante en el tiempo computacional (cuando menos lo triplicaban) y no mejoraban los estimados de $\hat{\tau}$, MSH y HD. Este primer análisis sirvió para descartar al número de sitios segregantes por locus y al número de singletones como estadísticos de resumen que mejoren significativamente los estimados de $\hat{\tau}$, MSH y HD además de producir resultados que no aumenten demasiado el tiempo computacional.

4.2.2. Elección de los mejores estadísticos de resumen para inferir el crecimiento poblacional y la homoplasia

Se corrió CORAGHE usando seis diferentes combinaciones de estadísticos, hasta obtener mil simulaciones aceptadas para los mismos diez juegos de datos de microsatélites usados en la sección anterior. En este experimento se volvieron a dejar fijos los mismos tres estadísticos del apartado anterior. Agregar el número de sitios segregantes y la heterocigosis tomando el haplotipo completo a los tres estadísticos de resumen fijos no ayuda a mejorar los estimados de τ, MSH y HD (Figura 12). Se comprobó que eliminar uno de los tres estadísticos fijos puede tener un efecto en la estimación de $\hat{\tau}$. Eliminar el promedio de la varianza en el número de repeticiones por microsatélite y el número de haplotipos diferentes por estado de los tres estadísticos fijos en el análisis causa una sobreestimación de $\hat{\tau}$. La omisión del promedio de la heterocigosis esperada por microsatélite no provoca un sesgo muy diferente en la estimación de $\hat{\tau}$

respecto a cuando usamos los tres estadísticos base. Se prefirió realizar corridas posteriores utilizando los tres estadísticos base porque: 1) utilizar tres estadísticos base no aumenta mucho el tiempo computacional respecto a los dos estadísticos anteriormente referidos; 2) en la literatura ya se ha argumentado cómo la interacción de la heterocigosis promedio por microsatélite y la varianza en el número de repeticiones por microsatélite contiene información sobre el crecimiento poblacional (Kimmel *et al.*, 1998; King *et al.*, 2000); 3) estudios anteriores fundados en estos tres estadísticos de resumen como base para el estudio del crecimiento poblacional han dado buenos resultados (Pritchard *et al.*, 1999; Estoup *et al.*, 2003; Estoup *et al.*, 2004).



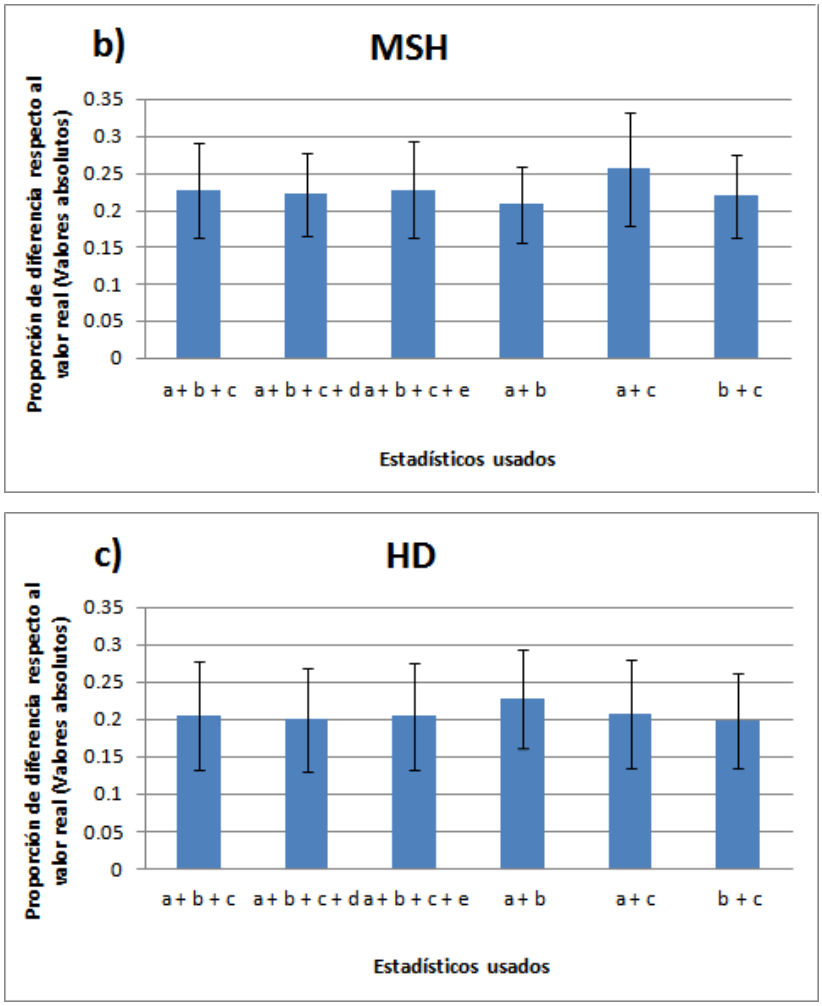


Figura 12.- Sesgo en las estimaciones de τ, MSH y HD con diferentes combinaciones de estadísticos en mil simulaciones aceptadas. Se graficó el promedio y el error estándar de la media del porcentaje de la diferencia del valor estimado respecto al valor real de τ, MSH y HD en valores absolutos de diez datos simulados. Las letras debajo de las columnas indican la combinación de estadísticos usados. Las letras significan: a) El promedio de la varianza en el número de repeticiones por microsatélite; b) El promedio de la heterocigosis por microsatélite; c) El número de haplotipos diferentes por estado; d) Número de sitios segregantes en todos los microsatélites y e) La heterocigosis tomando en cuenta todo el haplotipo.

4.2.3. Valores de épsilon más efectivos para estimar el crecimiento poblacional y la homoplasia

Para las mil aceptaciones obtenidas de cada combinación de parámetros en el apartado anterior, se analizó el efecto de cambiar los valores de ε en los estimados de τ, MSH y HD. Se observó que cambiar el valor de ε para cada uno de los estadísticos dentro de las tres combinaciones de parámetros no mejora mucho los estimados de τ, MSH y HD (ver Figura suplementaria 1). Concluimos que la mejor manera de estimar τ, MSH y HD es mediante el uso de los tres estadísticos fijos con una ε de 0.1.

4.3. Estimación de la homoplasia y el crecimiento poblacional con CORAGHE

4.3.1. Estimación de los parámetros que definen el crecimiento poblacional

Se simularon varios conjuntos de datos con diferentes valores de τ (Grupo 1 de la Tabla 2) para analizar si CORAGHE podía detectar cambios en los estimados de los parámetros que definen el crecimiento poblacional. No se analizaron los valores de θ_0 porque se obtienen malas distribuciones posteriores de θ_0 con CORAGHE. Esto quiere decir que el valor estimado de θ_0 se asocia a muchos valores y la distribución posterior no se define dentro de la distribución prior impuesta, a pesar de que el valor verdadero está contenido dentro de la distribución prior (ver Figura 13 para un ejemplo).

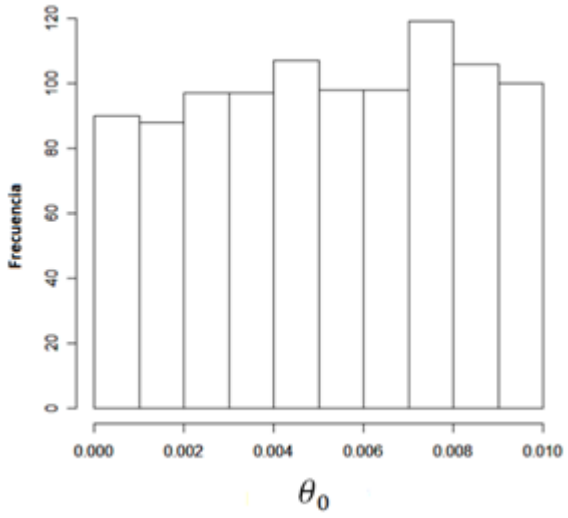


Figura 13.- Distribución posterior de θ_0 usando CORAGHE.

Como se muestra en la Tabla suplementaria 8, con datos generados a través de varios valores de τ se observan altos valores de r^2 para los valores estimados y reales de τ, recordando que r^2 nos dice qué tanto de la variación en los datos se explica por una recta de regresión lineal (Figura 14).

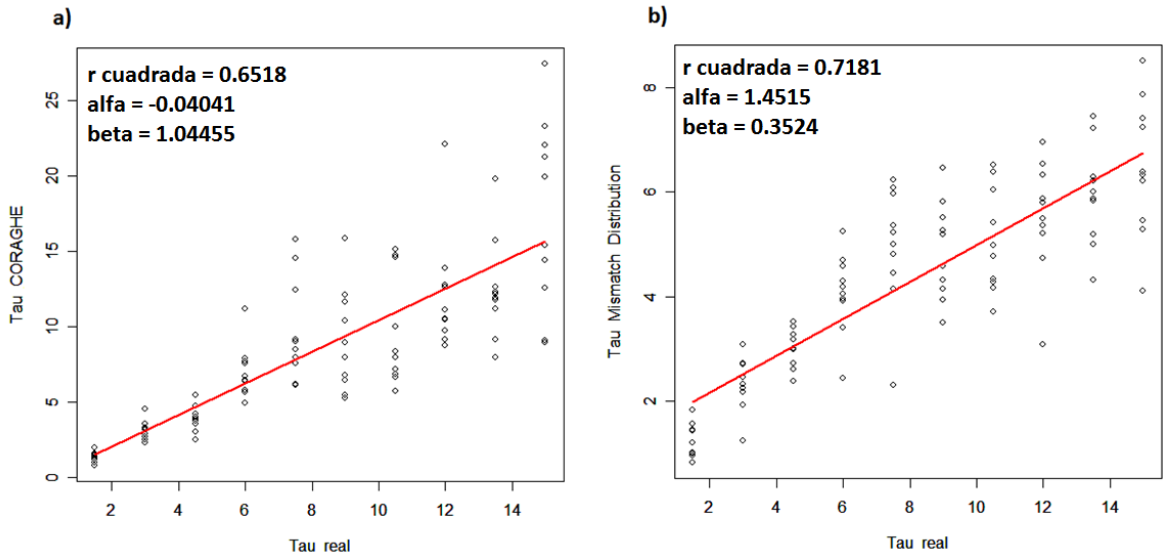
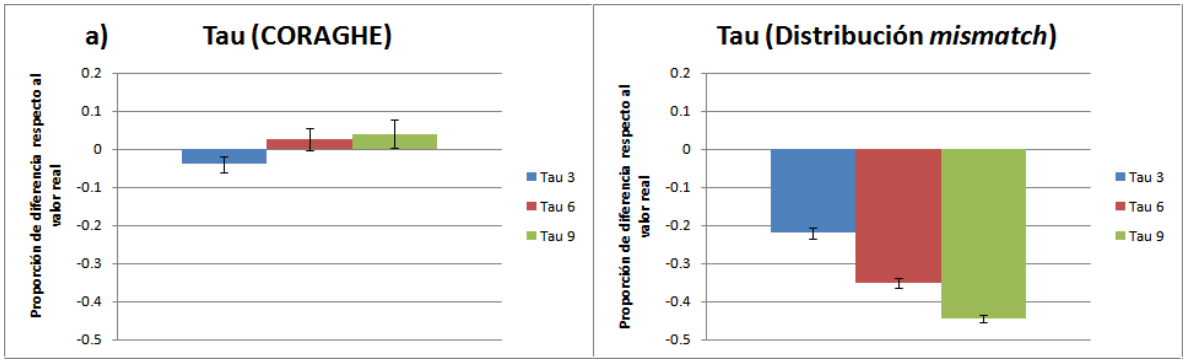


Figura 14.- Diagrama de dispersión y la ecuación de regresión para la relación lineal entre: a) el valor real de τ y su valor estimado con CORAGHE y b) el valor real de τ y su valor estimado con la distribución *mismatch*. Se usó una θ_0 = 0.03 y una θ_1 = 30.

Para investigar el sesgo y la precisión de los valores estimados de $\hat{\tau}$ y de $\widehat{\theta_1}$, usé la proporción de la diferencia del valor estimado respecto al valor real para dicho parámetro poblacional, a partir de valores con signo y con valores absolutos respectivamente. Midiendo el sesgo en los valores estimados de $\hat{\tau}$ con datos generados a partir de un mismo valor de τ, observamos que el valor real de $\hat{\tau}$ se subestima más cuando se utiliza la distribución *mismatch* y en promedio se sobreestima ligeramente con CORAGHE (Figura 15a). La precisión de los estimados de $\hat{\tau}$ obtenidos con CORAGHE es mejor que los valores estimados de $\hat{\tau}$ obtenidos con la distribución *mismatch* (Figura 15b). No se analizó el valor estimado de $\widehat{\theta_1}$ con la distribución *mismatch* porque dicho valor puede estar sesgado a valores más altos (Schneider *et al.*, 1999), pero dicho valor se analizó con CORAGHE y se obtuvieron estimados precisos y ligeramente subestimados de $\widehat{\theta_1}$ (Figura 15c).



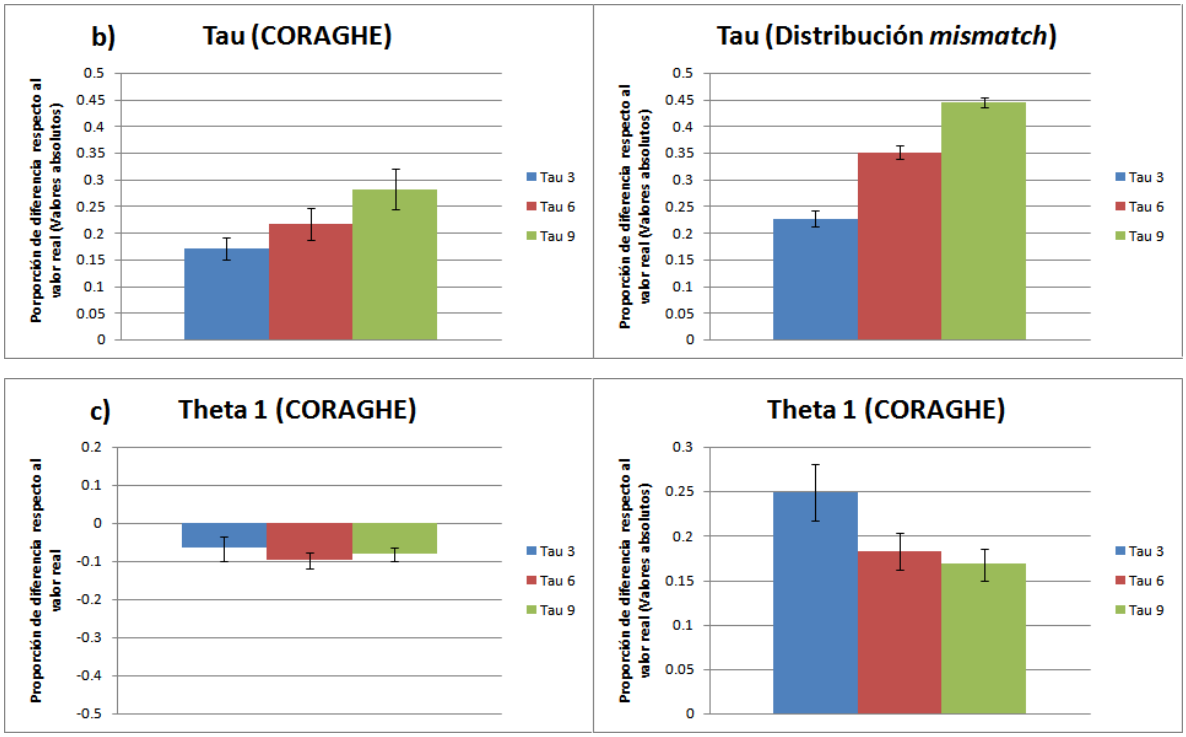


Figura 15.- Para tres grupos de datos donde se define un valor de τ se grafica la proporción de la diferencia del valor estimado respecto al valor real para la τ estimada con CORAGHE y la τ estimada mediante la distribución *mismatch* con a) valores con signo, para ver si existe un sesgo en las estimaciones; b) valores absolutos, para cuantificar la precisión del estimado. También se cuantificó c) la proporción de la diferencia del valor estimado respecto al valor real para la θ_1 estimada con CORAGHE. En todos los casos se uso una θ_0 = 0.03 y una θ_1 = 30.

4.3.2. Estimación de la homoplasia

Se analizó el uso de CORAGHE para estimar varias medidas de homoplasia, con especial atención en las medidas MSH y HD, ya que dichas medidas están más asociadas con el sesgo en $\hat{\tau}$.

Con datos generados a través de un conjunto de valores de τ, se encontró que existe una relación lineal entre los valores reales y los valores estimados de MSH y HD (Tabla suplementaria 8). También se observan valores altos de r^2 para las medidas estimadas y reales de MSH, HD y de τ (Figura 16).

MSH y HD aparentemente pueden ser estimados con CORAGHE. Para analizar si CORAGHE es capaz de diferenciar entre condiciones tanto de alta como baja MSH y HD en datos generados con los mismos parámetros de crecimiento poblacional, se simularon cien conjuntos de datos bajo tres diferentes condiciones de crecimiento poblacional donde se cambiaba el valor de τ (grupos 2, 3 y 4 de la Tabla 2) a fin de analizar los estimados de homoplasia obtenidos con CORAGHE.

En promedio, se encontró que HD se sobreestima en menos del 15% en los tres grupos de datos (Figura 17a). También, en los tres grupos de datos podemos rechazar la hipótesis de que existe una relación lineal significativa entre los valores reales y los valores estimados de HD, y esto se refleja también en sus bajos valores de r^2 (Tabla suplementaria 9). En los diagramas de dispersión para HD, se muestra que los valores estimados de HD y los valores reales de HD no guardan relación alguna (Figuras 16c, 16e y 16g). Esto nos dice que CORAGHE no puede distinguir entre condiciones de alta y baja HD en conjuntos de datos generados con los mismos parámetros de crecimiento poblacional.

MSH tiende a sobreestimarse ligeramente conforme aumentan los valores de τ (Figura 17b). Para MSH hay una relación lineal significativa entre los valores reales y los valores estimados de MSH y, por tanto, encontramos un *p-value* significativo en la ecuación de regresión lineal construida (Tabla suplementaria 9). El problema es que los valores de r^2 no son tan altos como desearíamos (sus valores van de 0.2 a 0.33), lo cual nos dice que a pesar de que hay una cierta relación lineal entre la estimación y los valores reales de MSH, muy poco de la variación en los datos reales se explica por los valores estimados de MSH.

En este conjunto de experimentos se probó que las medidas de homoplasia MSH y HD parecen bien estimadas a lo largo de diferentes valores de τ. Sin embargo, dentro de un mismo valor de τ no se puede

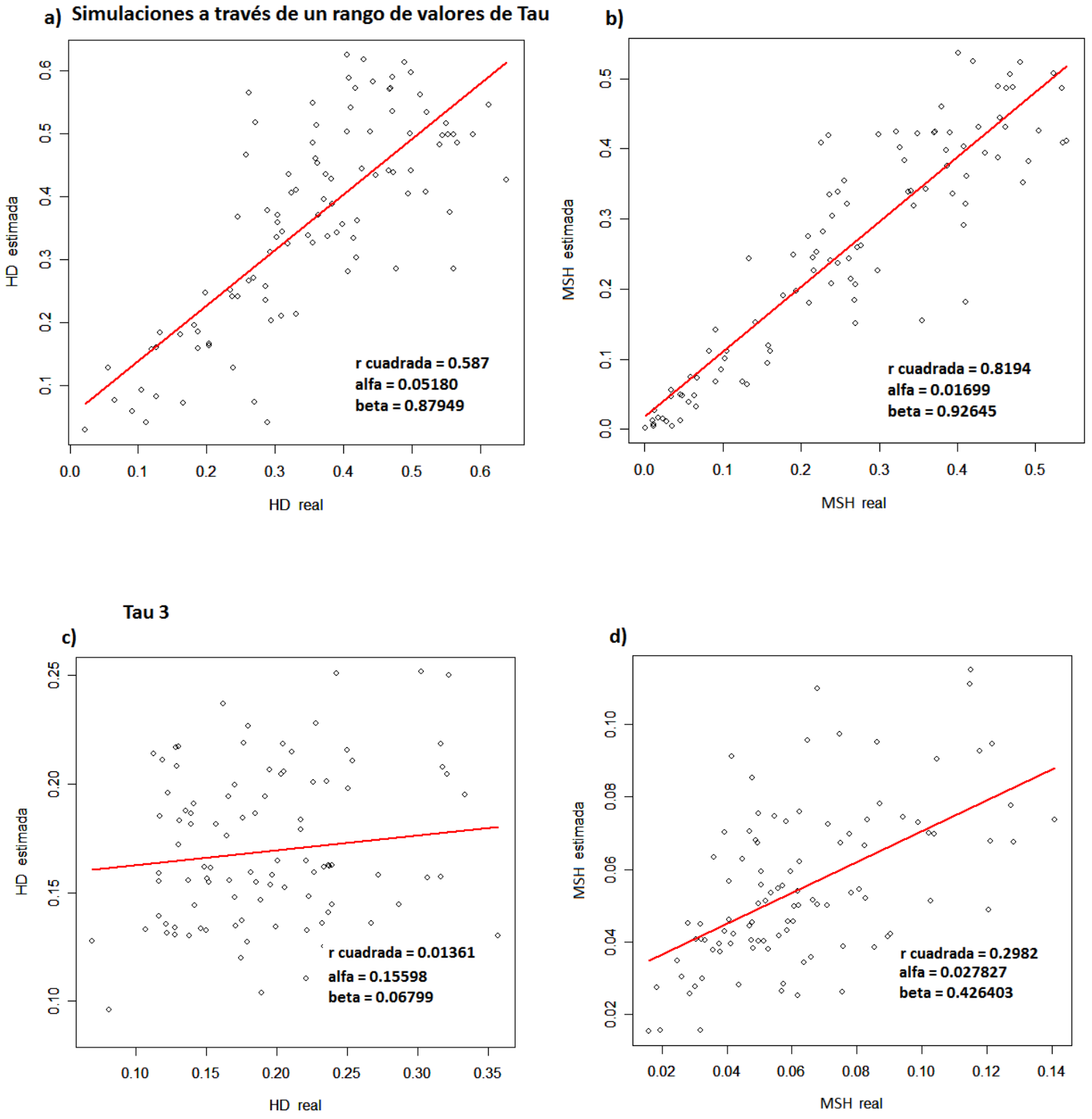
distinguir entre condiciones de alta y baja HD. CORAGHE puede distinguir, aunque de manera poco precisa, entre condiciones de alta y baja MSH. Esto indica que al usar CORAGHE para estimar MSH o HD en los datos, obtendremos un estimado de MSH y HD que será más bien un promedio representativo de los valores de MSH y HD que esperaríamos encontrar en poblaciones que comenzaron a crecer hace un cierto valor de τ.

Se exploró la estimación de otras medidas de homoplasia (HS, SH SASH y CH) en simulaciones donde se tiene un rango de valores de τ, τ es reciente y τ es antigua (Grupos 1, 2 y 4 de la Tabla 2 respectivamente) y se compararon los valores estimados con sus valores reales (Figura 18). Dentro de un rango de diferentes valores de τ se encontró que las medidas de homoplasia CH y SH estimadas tienen una relación lineal importante con sus valores reales (Figura 18a). Tanto SASH como HS no tienen una relación lineal importante con sus valores reales, lo cual demuestra por su baja r^2.

Al analizar datos generados con un mismo valor de τ (Figura 18b y 18c) se encuentra que la medida de homoplasia con una relación lineal más fuerte entre los valores reales y los valores estimados es SH, pero aun así esta medida muestra valores de r^2 bajos, 0.442 y 0.2573, lo cual establece que la relación lineal no es tan fuerte (Tabla suplementaria 10). Las medidas restantes no muestran una relación lineal de importancia entre los valores reales y los valores estimados de SASH, HS y CH.

Concluimos que con CORAGHE sólo se pueden obtener estimados confiables de los valores reales que SH y CH tendrían en promedio en datos creados con una cierta τ. También se concluye que CORAGHE no puede distinguir entre condiciones de alta y baja SH, SASH, HS y CH dentro de datos generados con la misma τ.

Figura 16


a) Simulaciones a través de un rango de valores de Tau
HD estimada
HD real
r cuadrada = 0.587
alfa = 0.05180
beta = 0.87949
b)
MSH estimada
MSH real
r cuadrada = 0.8194
alfa = 0.01699
beta = 0.92645
Tau 3
c)
HD estimada
HD real
r cuadrada = 0.01361
alfa = 0.15598
beta = 0.06799
d)
MSH estimada
MSH real
r cuadrada = 0.2982
alfa = 0.027827
beta = 0.426403

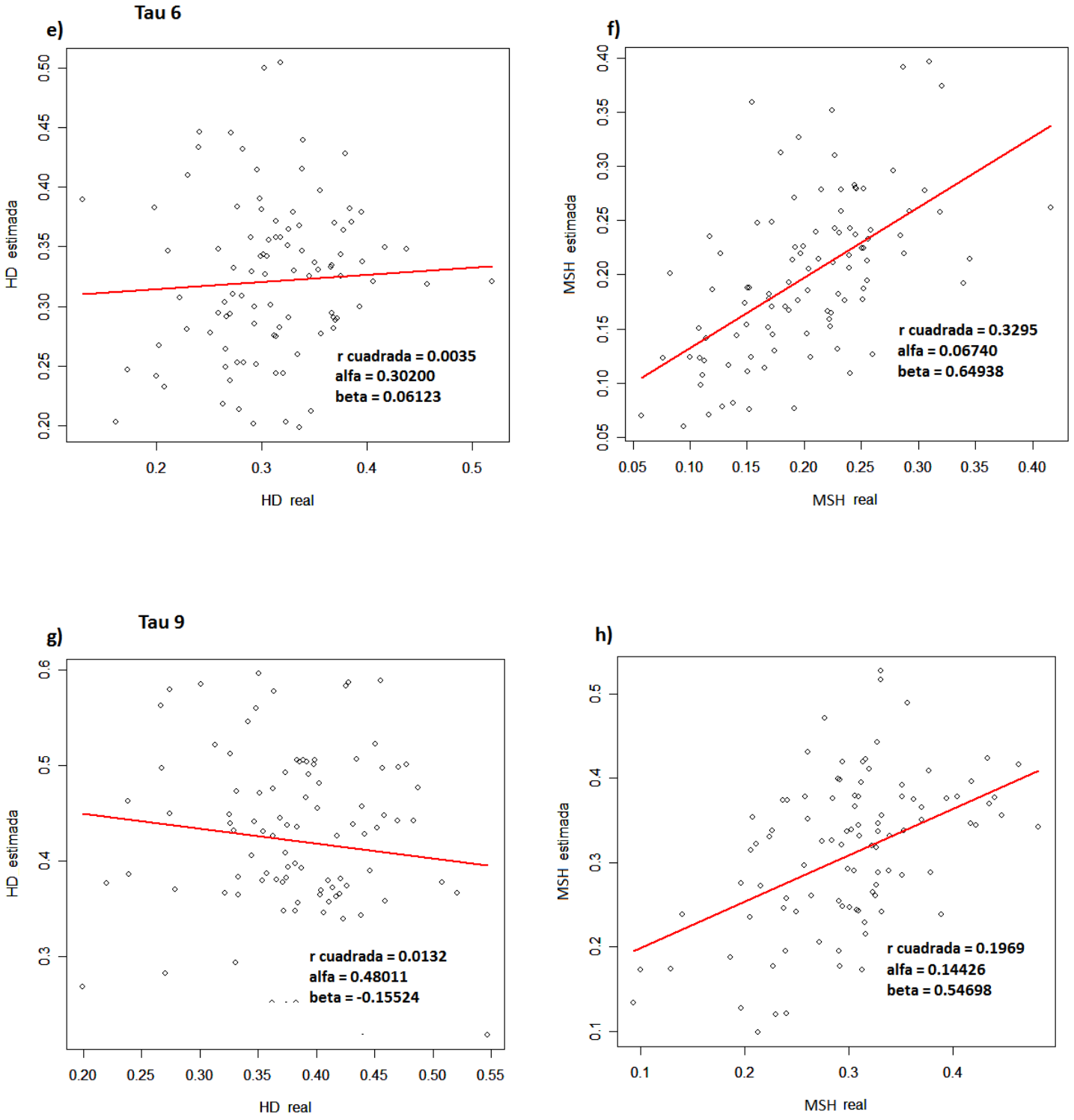


Figura 16.- Relación lineal de los valores estimados con CORAGHE y los valores reales de MSH y HD con datos generados a través de un rango de valores de τ [a y b], con una τ = 3 [c y d], una τ = 6 [e y f] y d) una τ = 9 [g y h]. Se usó una θ_0 = 0.03 y una θ_1 = 30.

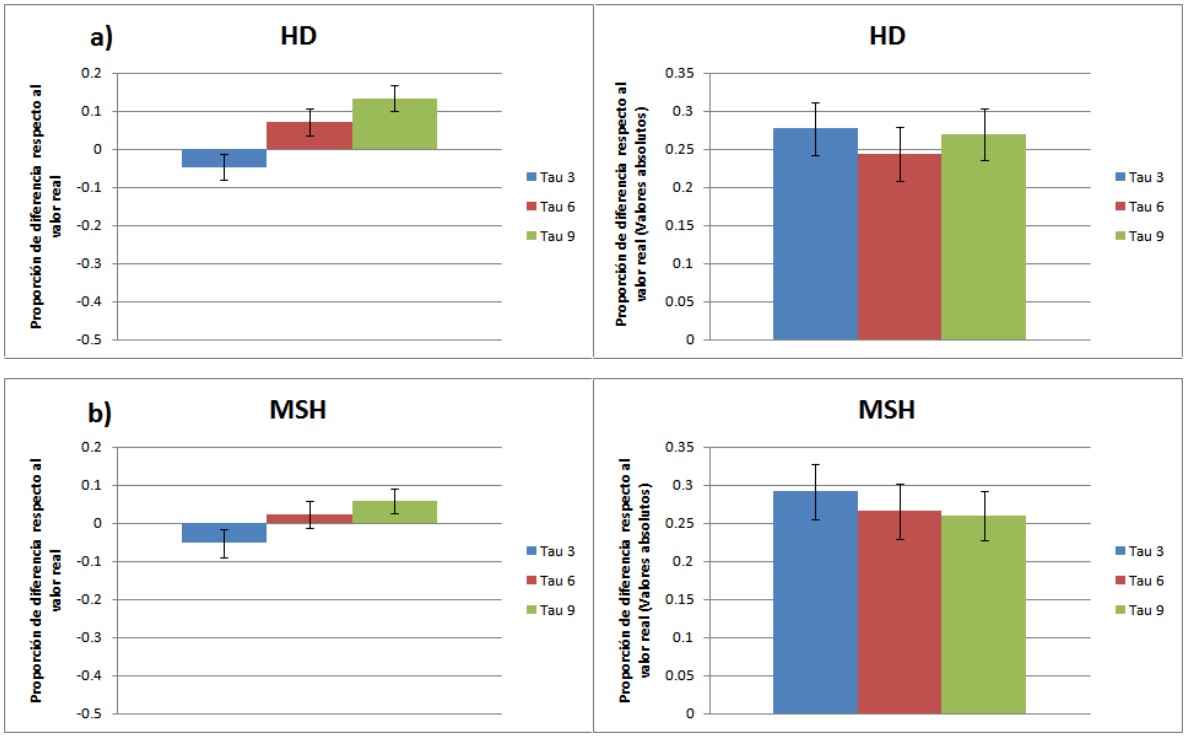
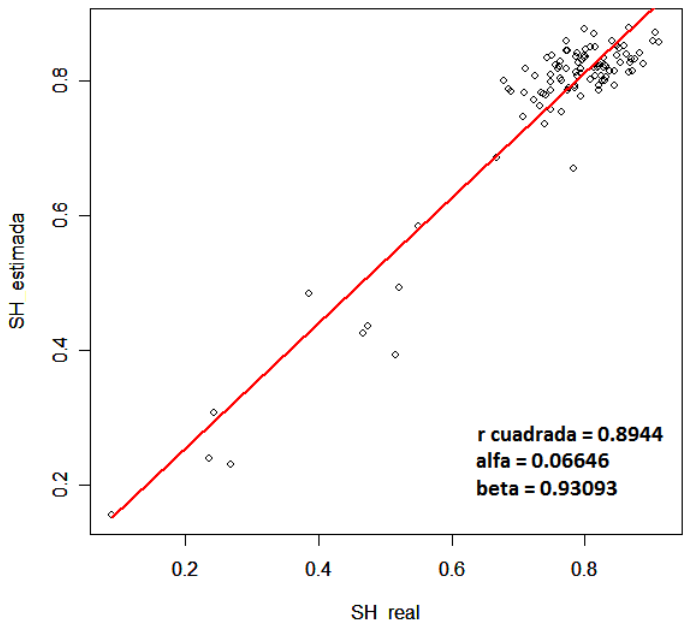
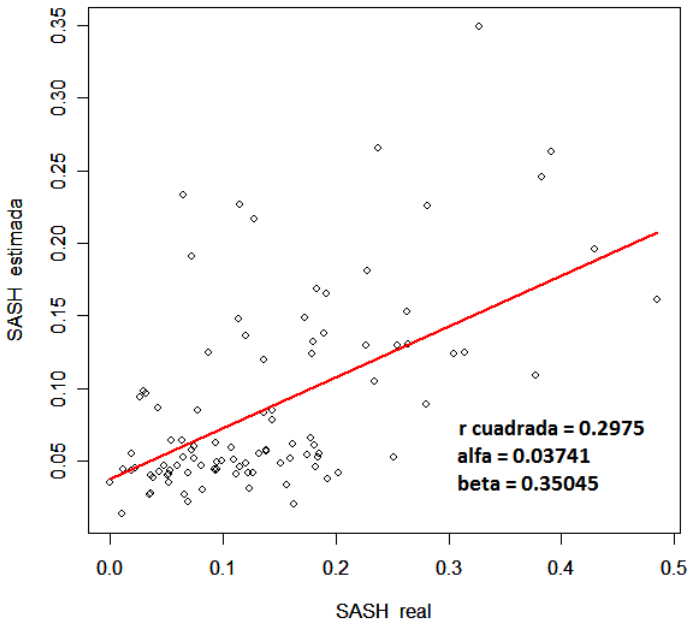


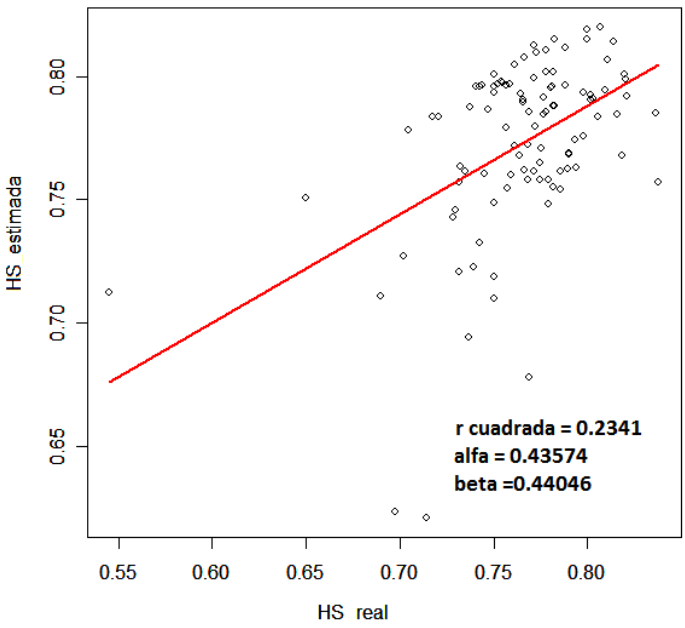
Figura 17.- Para tres grupos de datos donde se define un valor de τ se grafica la proporción de la diferencia del valor estimado respecto al valor real para a) MSH y b) HD. En todos los casos se usó una $\theta_0 = 0.03$ y una $\theta_1 = 30$.

Figura 18

a) Simulaciones a través de un rango de valores de Tau







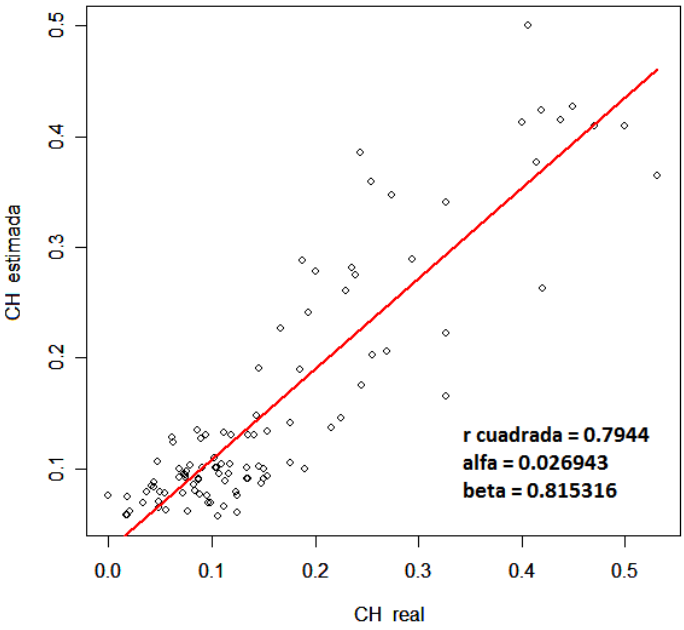
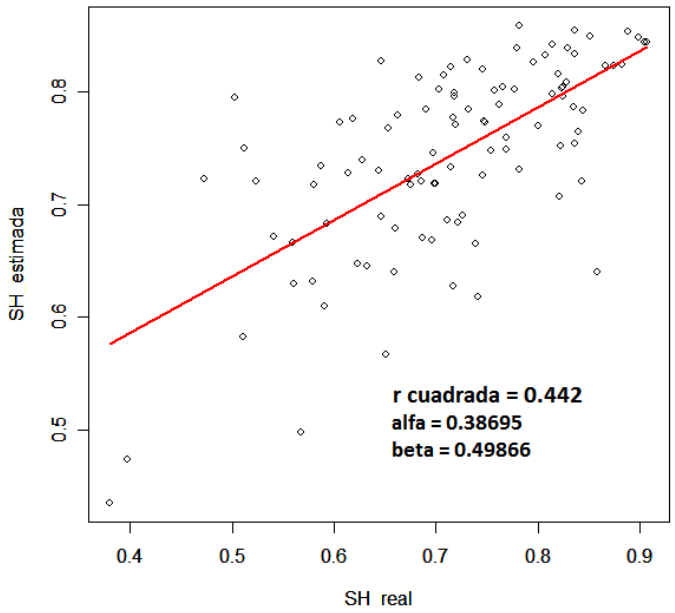
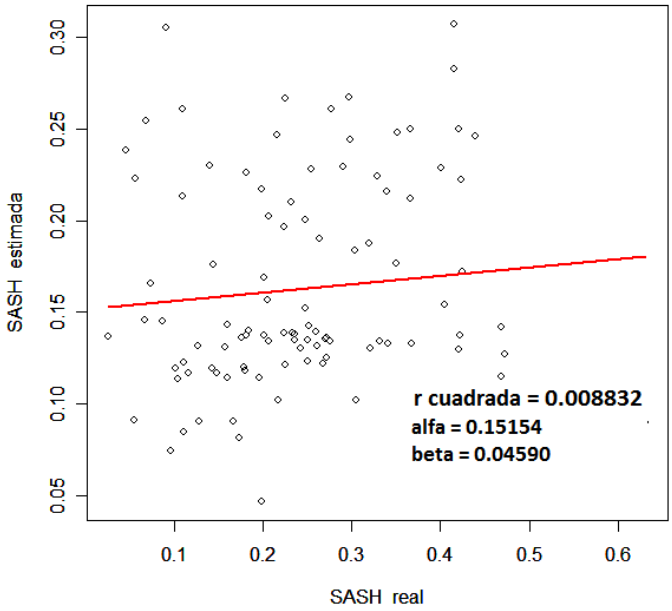
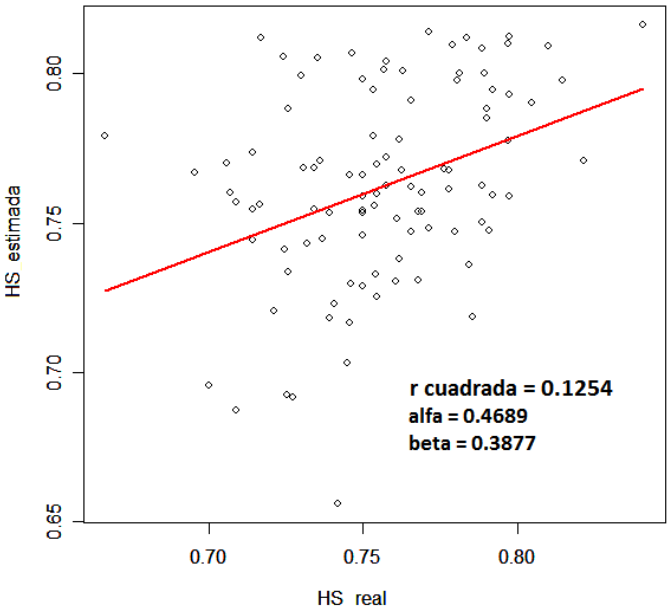


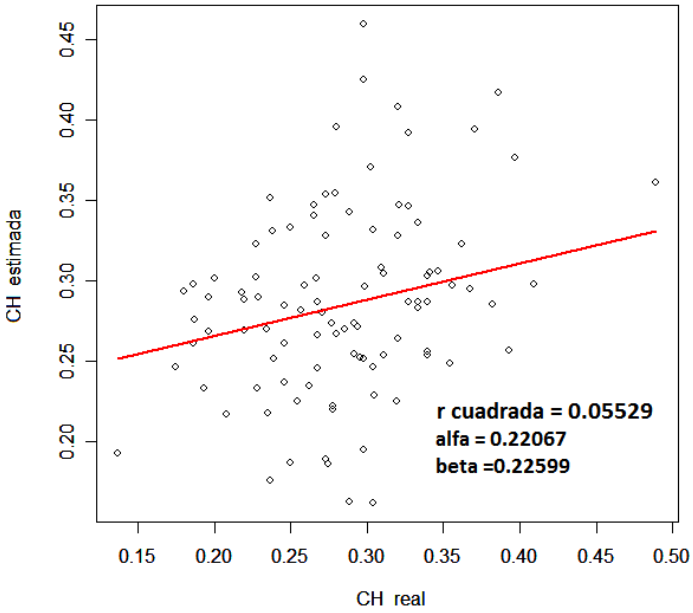
Figura 18 (continuación)

b) Tau 3









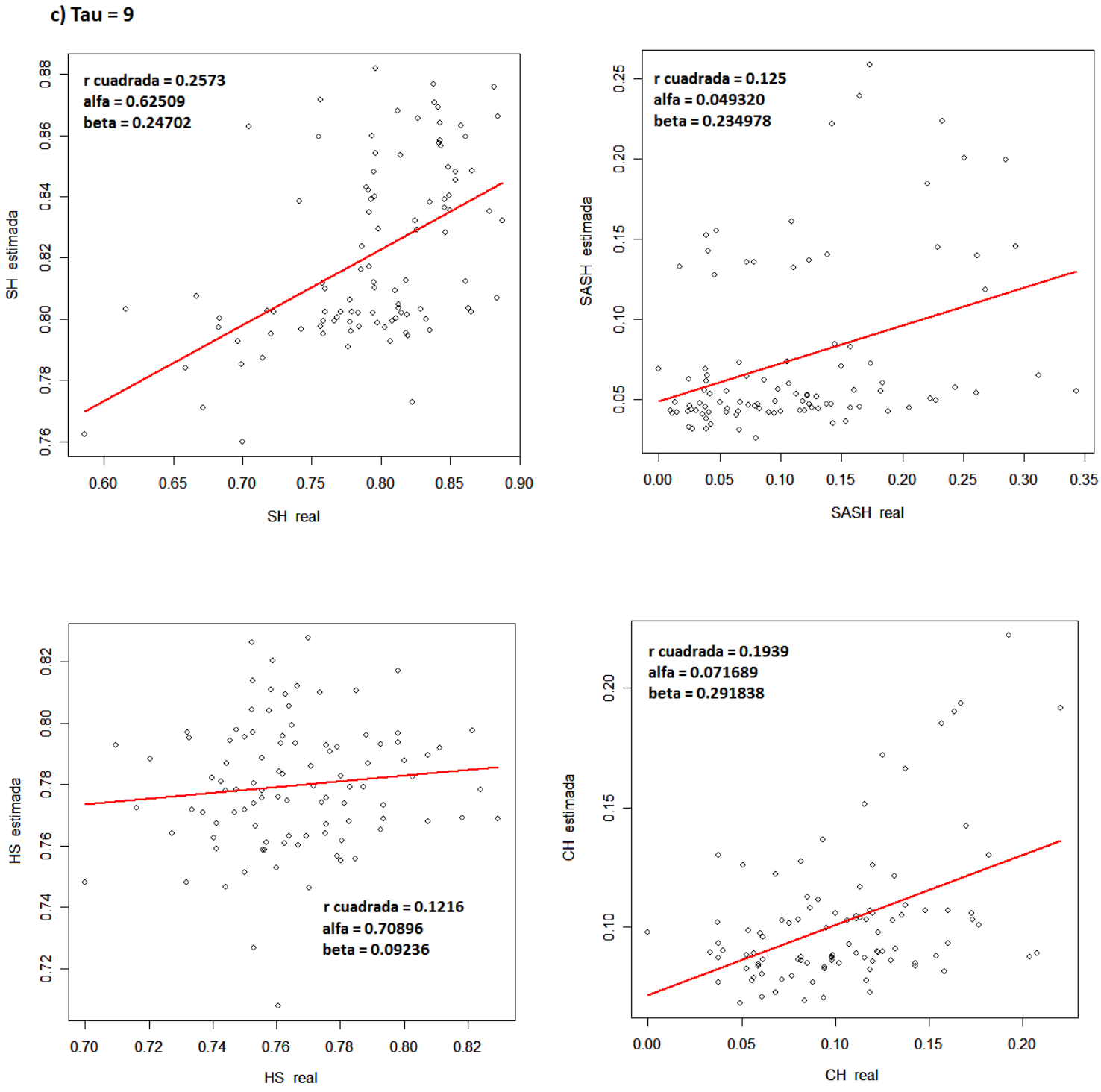
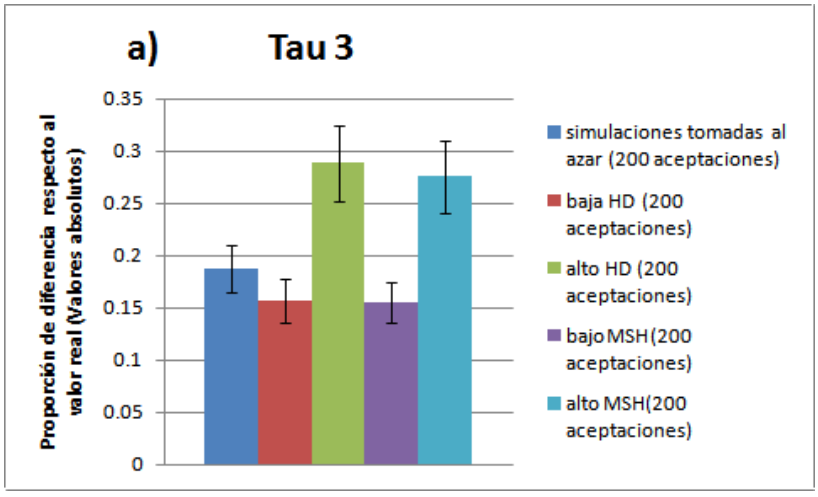


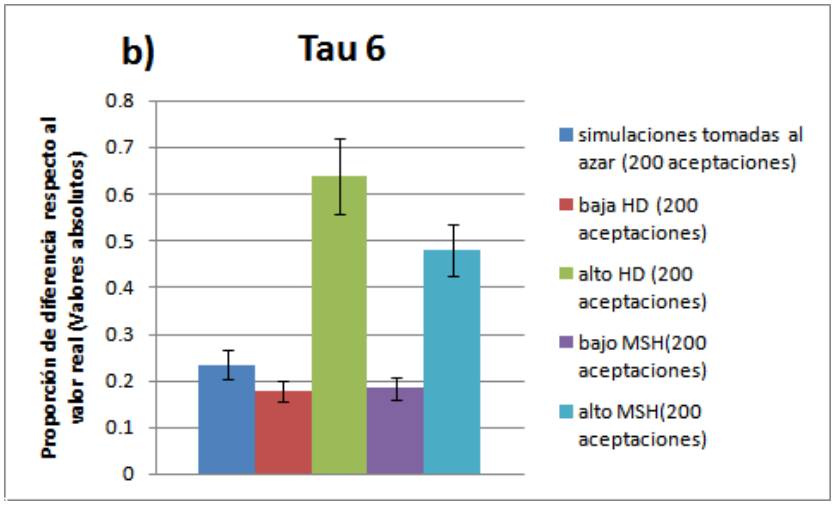
Figura 18.- Estimación de SH, SASH, HS y CH en tres grupos de cien simulaciones con: a) diferentes valores de τ; b) una τ = 3; y c) τ = 9. En todos los casos se usó una θ_0 = 0.03 y una θ_1 = 30.

4.3.3. Efecto de la estimación de la homoplasia en la estimación de τ

Los estimados de $\hat{\tau}$ obtenidos con CORAGHE son más cercanos en promedio al valor real de τ que cuando se usa la distribución *mismatch*. Esto se debe a que CORAGHE pondera el efecto de la homoplasia. Eso sugiere que la estimación de $\hat{\tau}$ se podría mejorar si se obtuvieran mejores estimados de MSH y HD, porque se ponderaría mejor el efecto de la homoplasia.

Para establecer si en los ABC una mejora en la estimación de la homoplasia, medida por MSH o HD, puede llevar a una mejora en la estimación de $\hat{\tau}$, se compararon las 200 simulaciones que tienen un valor de homoplasia más cercano al que tienen los datos simulados y las 200 simulaciones con un valor de homoplasia más lejano al de los datos. Se encontró que una mejora en la estimación de la homoplasia produce una ligera mejora en la estimación de $\hat{\tau}$ (Figura 19). Las 200 simulaciones que tienen una estimación menos precisa también producen una estimación de $\hat{\tau}$ menos precisa.





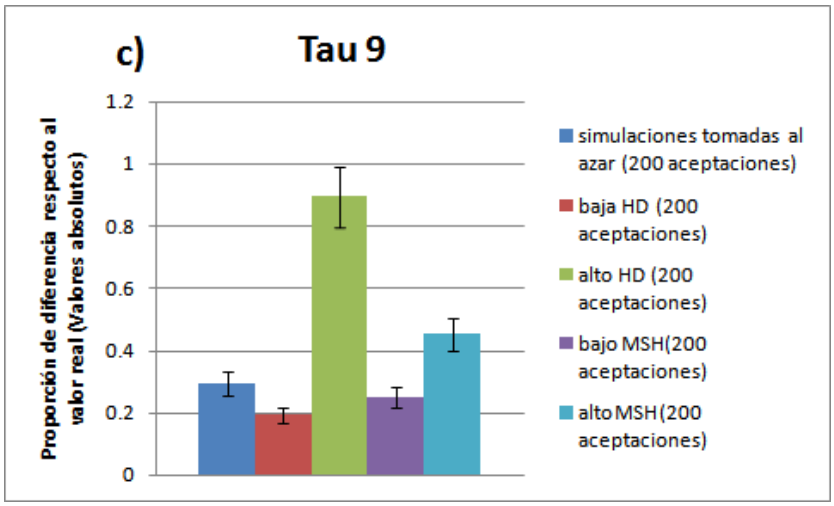


Figura 19.- Promedio de la proporción de la diferencia en la estimación de $\hat{\tau}$ respecto al valor real en valores absolutos tomando 200 aceptaciones al azar, las 200 aceptaciones de las mil obtenidas con CORAGHE donde el valor de HD y el valor de MSH son más cercanos al valor real de HD y MSH respectivamente, y las 200 aceptaciones de las mil obtenidas con CORAGHE donde el valor de HD y el valor de MSH son más lejanas al valor real de HD y MSH respectivamente, en cien conjuntos de datos con una a) τ = 3, b) τ = 6 y c) τ = 9. Se usó una θ_0 = 0.03 y una θ_1 = 30.

4.4. Estimación del crecimiento poblacional y la homoplasia con datos de Pinus caribaea

Se analizaron datos de microsatélites de *Pinus caribaea* en donde se sabe que existe crecimiento poblacional para probar la efectividad de CORAGHE en la estimación de la homoplasia y el crecimiento poblacional. Se graficó la distribución posterior de los valores de $\hat{\tau}$, $\hat{\theta}_1$, MSH y HD obtenida con CORAGHE (Figura 20). También se estimó el valor de las medidas de homoplasia SH y CH, además del valor de τ que se obtendría usando la distribución *mismatch*. El valor de $\hat{\tau}$ estimado con la distribución *mismatch* es menor que el obtenido con CORAGHE (Tabla 4). Según la τ obtenida con Arlequin, hace 224,900 años comenzó el crecimiento poblacional. Por lo tanto, el crecimiento poblacional comenzó en un periodo glacial: el Estadio Isotópico Marino 7.4 (Martinson *et al.*, 1987). A su vez, con la τ obtenida en CORAGHE, inferimos que el crecimiento poblacional comenzó hace 299,900 años. Entonces, podemos concluir que el crecimiento poblacional inició en un periodo cercano a un periodo interglacial (cercano al Estadío Isotópico Marino 9) (Gibbard y Van Kolfschoten, 2004; http://www.quaternary.stratigraphy.org.uk/correlation/ POSTERSTRAT_BOREAS_v2005c.pdf, Martinson *et al.*, 1987).

El valor de $\hat{\tau}$ estimado con la distribución *mismatch* es 25% menor que el estimado con CORAGHE. En secciones anteriores de esta tesis se encontró que en expansiones antiguas (valores de $\hat{\tau}$ >3 aproximadamente) la subestimación de $\hat{\tau}$ usando la distribución *mismatch* es mayor por efecto de la homoplasia y también que el valor de $\hat{\tau}$ estimado por CORAGHE es más cercano al valor real de τ, lo cual es congruente con lo encontrado en estos datos de *Pinus caribaea*.

En este ejemplo utilizamos un tiempo generacional de 42.5 años y una tasa de mutación de $5.5 * 10^{-5}$ mutaciones por generación. Para ver

cuál es el efecto que podía tener la tasa de mutación por generación y el tiempo generacional en el tiempo estimado de expansión en años, se graficó la diferencia entre el estimado de $\hat{\tau}$ obtenido con CORAGHE y el estimado de $\hat{\tau}$ obtenido usando la distribución *mismatch* ($\hat{\tau}$ = 1.3589) al ser transformado en años usando distintas tasas de mutación por generación. Se encontró que la tasa de mutación tiene una relación inversamente proporcional con el sesgo en años de la estimación del tiempo de inicio de la expansión poblacional a causa de la homoplasia (Figura 21a). El tiempo generacional en cambio tiene una relación directamente proporcional con el sesgo en años de la estimación del tiempo del inicio de la expansión poblacional por efecto de la homoplasia (Figura 21b). Por ello, cuando estamos calculando el sesgo en nuestro cálculo del tiempo de inicio de la expansión poblacional tenemos que tomar en cuenta tanto *u* como el tiempo generacional para saber que tanto podría subestimarse el tiempo de inicio de la expansión poblacional por efecto de la homoplasia.

Tabla 4.- Análisis de los datos de *Pinus caribaea*. Estimaciones de HD, MSH, τ y θ_1 obtenidas con CORAGHE y el estimado de τ obtenido con Arlequin.

HD estimada	0.2388463
MSH estimada	0.11589683
Tau (CORAGHE)	5.43287979
Tau (Arlequin)	4.074
Theta 1 estimada	15.6458882
CH	0.06684169
SH	0.7890094

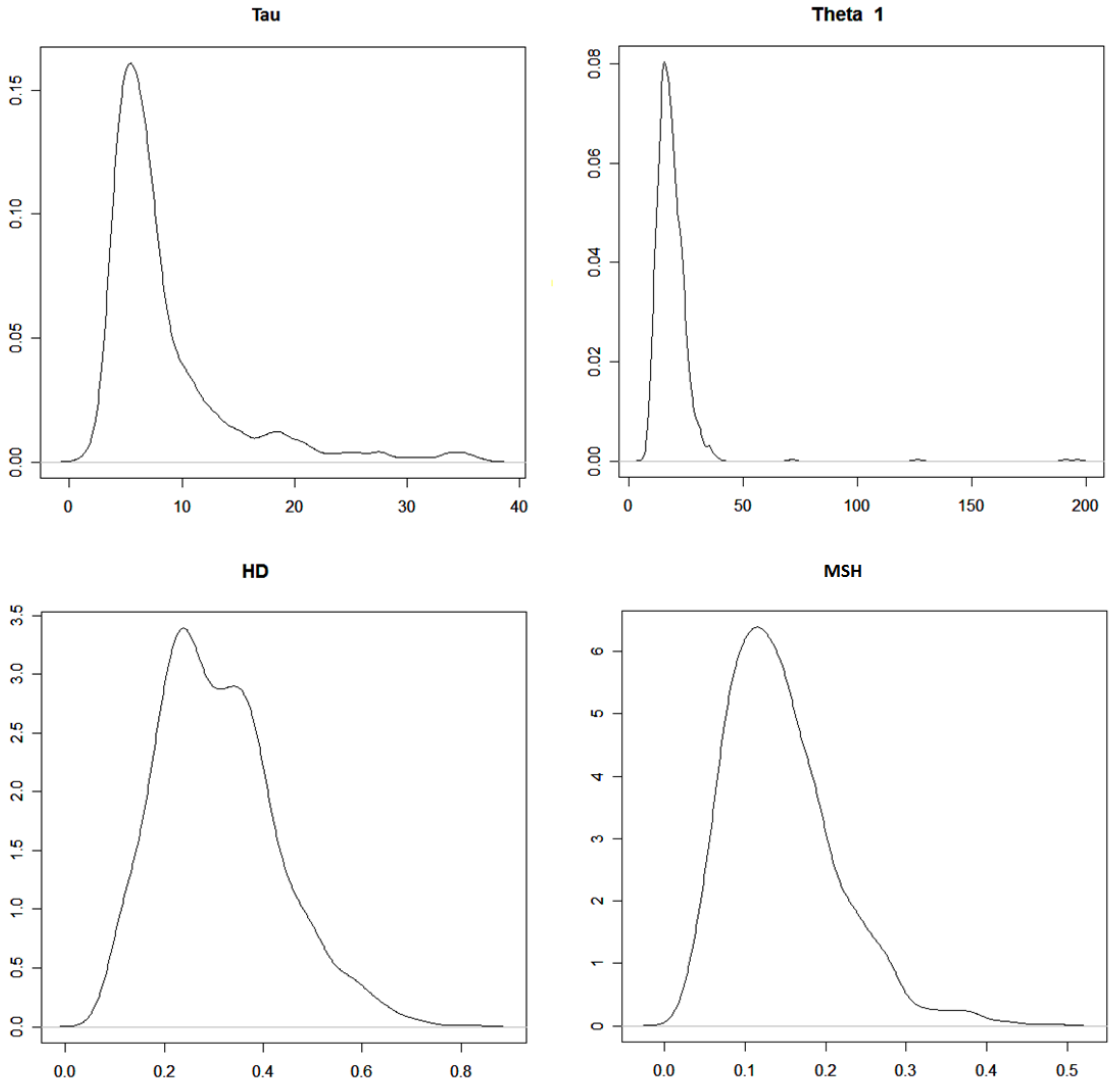


Figura 20.- Distribución posterior de los valores de τ, θ_1, MSH y HD estimados con CORAGHE. La moda de cada distribución posterior es nuestro mejor estimado de $\hat{\tau}$, $\hat{\theta}_1$, MSH y HD.

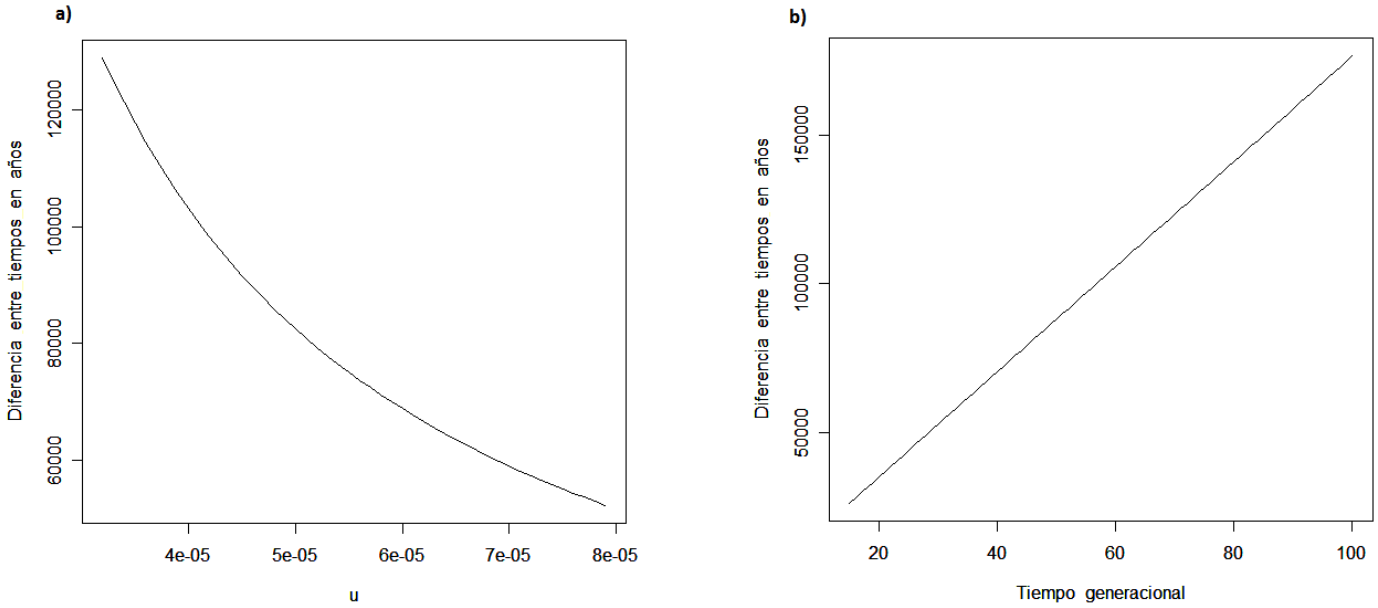


Figura 21.- Efecto del sesgo en la estimación del tiempo de inicio de la expansión poblacional. Para ver el sesgo se graficó la diferencia entre el estimado de $\hat{\tau}$ obtenido con CORAGHE y el estimado de $\hat{\tau}$ obtenido usando la distribución *mismatch* ($\hat{\tau}$ = 1.3589). Se evalúo el efecto que puede ocasionar dicho sesgo en la diferencia en años de los tiempos a la expansión poblacional mediante la variación de: a) la tasa de mutación y b) el tiempo generacional.

Discusión

5.1. Efecto de la homoplasia en la forma de la distribución *mismatch*

La homoplasia provoca una subestimación del tiempo en que comenzó el crecimiento poblacional cuando utilizamos la distribución *mismatch* bajo el supuesto de que el marcador molecular que utilizamos evoluciona de acuerdo a un ISM (Navascués et al., 2006; Navascués et al., 2009). Bajo el ISM, no deberíamos encontrar homoplasia en nuestras secuencias. Usando microsatélites, el supuesto de que nuestras secuencias evolucionan bajo un ISM es falible si se tiene una alta tasa de mutación por microsatélite o tamaños poblacionales grandes, dado que ambos parámetros propician un aumento del parámetro θ y esto causa un aumento en la homoplasia (Estoup *et al.*, 2002).

Bajo el modelo coalescente, cuando la τ es muy reciente la disminución en el tamaño de las ramas provoca la aparición de un menor número de mutaciones que caen en cada rama y, por lo tanto, se espera que la homoplasia tenga un menor efecto. A su vez, en expansiones más antiguas, se acumula un número mayor de mutaciones y se espera un efecto mayor de la homoplasia. Estos efectos son consistentes con mis resultados (Figuras 7 y 8), donde se observa que la homoplasia afecta más la forma de la distribución *mismatch* en expansiones más antiguas. Este efecto de la homoplasia en la distribución *mismatch* provoca una subestimación del tiempo en que comenzó el crecimiento poblacional ($\hat{\tau}$) en expansiones más antiguas (congruente con los resultados de Navascués *et al.*, 2006). La estimación de $\hat{\tau}$ también depende de un valor alto de $\widehat{\theta_1}$, valores más altos de $\widehat{\theta_1}$ nos permitirán obtener estimados más precisos de $\hat{\tau}$ en expansiones más antiguas (Figura 7).

Es necesaria una forma de evaluar la subestimación de $\hat{\tau}$ por efecto de la homoplasia. En esta tesis se propusieron varias medidas de homoplasia para definir cuál tenía una relación más fuerte con el sesgo en la estimación de τ. Se encontró que el sesgo está relacionado con dos medidas: MSH y HD, siendo más fuerte la relación del sesgo con HD. Que HD sea la medida más relacionada con el sesgo en $\hat{\tau}$ no es tan sorprendente, ya que la distribución *mismatch* utiliza las diferencias entre pares de haplotipos y HD cuantifica las diferencias entre pares de haplotipos que se pierden por efecto de la homoplasia. MSH también probó ser una medida con una relación importante con el sesgo en $\hat{\tau}$, lo cual muestra que el promedio de la homoplasia entre microsatélites puede darnos información sobre el sesgo en $\hat{\tau}$. Sin embargo, la introducción de las medidas MSH y HD es valiosa porque se demostró que no todas las medidas de homoplasia se relacionan con el sesgo en $\hat{\tau}$ (Figura 9). En efecto, la medida que más se relaciona con el sesgo en $\hat{\tau}$, HD, no ha sido la medida usada para cuantificar el sesgo en $\hat{\tau}$ en estudios anteriores (Navascués *et al.*, 2006).

La gravedad del sesgo en $\hat{\tau}$ depende de los valores de MSH y HD. En las simulaciones realizadas con una $\widehat{\theta_1}$ = 30, se encontró que si los valores de MSH y HD son 0, la subestimación de $\hat{\tau}$ será cercana al 10% y al 0%, respectivamente. Pero si los valores de MSH y HD son altos (por ejemplo, 0.5), la subestimación de $\hat{\tau}$ será del 61% y del 52%, respectivamente (Figura 9).

5.2. Análisis de los mejores estadísticos de resumen para correr CORAGHE

En este trabajo se empleó un ABC (algoritmo bayesiano aproximado) para estimar HD y MSH en conjunto con $\hat{\tau}$ ponderando el efecto de la homoplasia, ya que estas medidas contribuyen a tener un mejor estimado del error que podría tener $\hat{\tau}$. Anteriormente, los ABC han sido empleados para calcular

parámetros relacionados con modelos de crecimiento poblacional (Pritchard *et al.*, 1999) o de divergencia entre poblaciones (Ross-Ibarra *et al.*, 2009). Los algoritmos bayesianos aproximados son una herramienta eficaz para comparar escenarios demográficos complejos y ya se ha desarrollado *software* para implementarlos de forma sencilla (Cornuet *et al.*, 2008). Pero nunca se han empleado para estimar la homoplasia.

Los ABC son dependientes de los estadísticos de resumen y los valores de ε. En este trabajo realicé una búsqueda de estadísticos de resumen que pudieran mejorar la estimación de la homoplasia y los parámetros que definen el crecimiento poblacional. Los estadísticos de resumen propuestos buscaban explotar la firma de las genealogías sometidas a crecimiento poblacional, es decir, una alta proporción de mutaciones en las ramas terminales y una baja proporción de mutaciones en las ramas internas. Ninguno de los estadísticos propuestos mejoró de forma importante la estimación de τ (Figuras 11 y 12). La elección de los estadísticos para un ABC más apropiados siguen siendo un tema de discusión (Beaumont *et al.*, 2002), ya que no existe una metodología estricta para elegir los estadísticos que mejor describen las propiedades de un modelo demográfico al calcular los parámetros poblacionales que definen al modelo. La elección de los valores de ε tampoco se basa en un criterio rígido y los valores de ε varían de un estudio a otro (Pritchard *et al.*, 1999, utiliza una ε de 0.1; Estoup *et al.*, 2003, una ε de 0.12; en Estoup *et al.*, 2001, se usan valores de 0.08 y de 0.14).

5.3. Estimación del tiempo del crecimiento poblacional y la homoplasia con CORAGHE

Se evalúo la efectividad de CORAGHE para estimar los parámetros del tiempo y la magnitud del crecimiento poblacional. Se encontró que los

valores de $\hat{\tau}$ estimados con CORAGHE son más precisos que los estimados de $\hat{\tau}$ obtenidos con la distribución *mismatch,* aunque los valores de $\hat{\tau}$ estimados con CORAGHE muestran una ligera sobreestimación en expansiones antiguas (Figuras 14 y 15). Esto se debe a que al simular datos con CORAGHE se toma en cuenta la homoplasia usando un modelo de mutación de microsatélites más realista, el SMM. Entonces, en los datos simulados puede existir homoplasia y se pondera el hecho de su existencia, lo cual no es tomado en cuenta por la distribución *mismatch*, donde se parte del supuesto de que nuestros datos evolucionan de acuerdo a un ISM.

Se analizó la capacidad de CORAGHE para estimar la homoplasia. En los ABC se aceptan aquellas genealogías donde se produjo un nivel de variación genética similar al que encontramos en datos reales. Dichas genealogías deben tener información importante de los procesos que condujeron a los datos a tener un cierto nivel de variación genética. Por ello, calcular la homoplasia a partir de las genealogías aceptadas resulta una estrategia viable, ya que obtenemos la distribución posterior P(homoplasia | datos). En teoría, esta estrategia podría emplearse para calcular la homoplasia en cualquier programa basado en un conjunto de genealogías para calcular parámetros poblacionales de interés, como son IM (Hey y Nielsen, 2007), BEAST (Drummond *et al*. 2007) ó LAMARC (Kuhner *et al.*, 2006). Pero a diferencia de estos programas y del uso de métodos analíticos, los ABC son más flexibles y pueden adaptarse para reproducir modelos demográficos más complejos.

Calcular medidas de la homoplasia en nuestros datos no es tan importante si no se relaciona con los problemas que la homoplasia puede causar en la estimación de parámetros poblacionales, como es el caso de la subestimación de $\hat{\tau}$ en el crecimiento poblacional. Por ejemplo, se ha mostrado que la homoplasia tiene un efecto débil en la F_{st} (Estoup *et al.*, 2002) y en el estimado de Nei de la diversidad haplotípica, H_E (Nei, 1978),

aunque sí tiene un efecto en el estimado de Goldstein de la distancia genética (Goldstein *et al.*, 1995), una medida que usa las diferencias pareadas entre microsatélites (Navascués y Emerson, 2005). Para escenarios demográficos complejos, que no pueden modelarse fácilmente, sería de interés cuantificar cuál es efecto que podría tener la homoplasia en los parámetros poblacionales que definen nuestro modelo demográfico.

Uno de los objetivos más importantes de este trabajo es mostrar que puede lograrse una buena estimación de los valores de HD y MSH para obtener un mejor estimado de $\hat{\tau}$ ponderando la homoplasia. En esta tesis se mostró que con los ABC es posible establecer buenas aproximaciones de los valores de HD y de MSH (Figuras 16 y 17). Gracias a esto, se obtienen mejores estimados de $\hat{\tau}$ que utilizando la distribución *mismatch* bajo el supuesto de que los microsatélites evolucionan según un ISM. A diferencia de HD, MSH se puede estimar en cierta medida dentro de datos generados con el mismo valor de τ. Esto se manifiesta en una mejor estimación de MSH comparado con HD.

El hecho de que HD se estima con menor precisión que MSH es un resultado desafortunado, puesto que HD es la medida de homoplasia que más se relaciona con un sesgo en la estimación de $\hat{\tau}$. Sin embargo, también MSH se relaciona de modo muy estrecho con la calidad del estimado de $\hat{\tau}$, lo cual es bastante alentador (Figura 9). Además, se concluyó que CORAGHE permite una estimación de los valores promedio de CH y SH. Es la primera vez que estas medidas se han podido estimar en datos de microsatélites (Figura 18).

5.4. Uso de CORAGHE para estimar el crecimiento poblacional y la homoplasia en datos de *Pinus caribaea*

Se comprobó que CORAGHE puede utilizarse en datos reales para calcular tanto $\hat{\tau}$ como las medidas más relacionadas con el sesgo en su estimación, MSH y HD (Tabla 4). En datos de *Pinus caribeae,* se probó que el valor de $\hat{\tau}$ estimado con la distribución *mismatch* es menor que el estimado con CORAGHE en una proporción del 25%. Esto es congruente con los resultados que muestran que al usar microsatélites, la distribución *mismatch* subestima el verdadero valor de $\hat{\tau}$. El valor real de τ es lo suficientemente alto para que la homoplasia cause una subestimación del valor real de $\hat{\tau}$. Podemos relacionar el sesgo que esperaríamos en estos estimados con el valor de MSH y HD con una regresión lineal. Al usar la regresión lineal con un valor de θ_1 = 15 (Figura 9A) y sustituir los valores estimados de MSH y HD, vemos que el sesgo esperado de $\hat{\tau}$ calculado con la distribución *mismatch* presenta una subestimación situada entre el 23 y el 25% del valor real. ¿Qué tan importante es esta subestimación? En este caso la diferencia en el valor de de $\hat{\tau}$ obtenido con la distribución *mismatch* respecto al valor de $\hat{\tau}$ estimado con CORAGHE es de 1.3589. Recordemos que $\tau = 2ut$, donde u es la tasa de mutación por generación de la secuencia y *t* es el tiempo en generaciones hacia el pasado en que se llevó a cabo el crecimiento poblacional. El efecto de la subestimación del tiempo en años en que se llevó a cabo el crecimiento poblacional en el pasado depende de u y del tiempo generacional. Si u es grande y el tiempo generacional es pequeño el sesgo en la estimación del tiempo de inicio de la expansión poblacional, medido en años, no es tan importante (Figura 21). De acuerdo a los microsatélites de cloroplasto de *Pinus caribaea* usados en esta tesis, un cambio de una unidad en τ significa que el tiempo a la expansión se recorre 55,195 años, lo cual es un tiempo lo suficientemente grande para

mover el tiempo en que se lleva a cabo el crecimiento poblacional a otro periodo glacial. En estos datos de *Pinus caribaea* la subestimación del valor real de $\hat{\tau}$ fue lo suficientemente grande para provocar que el cálculo del tiempo en el que comenzó el crecimiento poblacional se estimara en un periodo glacial diferente, lo cual tiene repercusiones importantes en nuestras inferencias demográficas.

5.5 Perspectivas para el uso de los ABC en modelos demográficos

Este trabajo ha demostrado que los microsatélites en conjunción con un ABC tienen gran utilidad para estimar $\widehat{\theta_1}$ y $\hat{\tau}$, con resultados más precisos que los que se obtienen del uso de la distribución *mismatch* bajo el supuesto de que los microsatélites evolucionan de acuerdo a un ISM. También los ABC poseen la ventaja sobre el uso de las distribuciones *mismatch* de que con el uso de los ABC es posible calcular la homoplasia. Este trabajo no pretende invalidar el uso de las distribuciones *mismatch*, sólo pretende señalar que los estimados de $\hat{\tau}$ se encuentran subestimados por el supuesto erróneo de que los microsatélites mutan según un ISM. Sin embargo, recientemente se ha propuesto una manera para mejorar las estimaciones de $\hat{\tau}$ utilizando la distribución *mismatch* asumiendo un SMM (Navascués *et al.*, 2009). En el futuro, queda pendiente verificar si los ABC pueden producir estimados de $\hat{\tau}$ tan acertados como los del método recien propuesto. También queda pendiente ver si se puede mejorar la estimación de MSH y HD por medio de nuevas estadísticas de resumen. Mientras tanto, concluimos que los ABC constituyen una herramienta útil para calcular tanto parámetros poblacionales como valores de homoplasia en un modelo de crecimiento poblacional, cuya aplicación puede extenderse a otros modelos demográficos.

Conclusiones

1) La homoplasia provoca una subestimación del tiempo hacia el pasado en que sucedió el crecimiento poblacional ($\hat{\tau}$).

2) La subestimación de $\hat{\tau}$ se relaciona de forma lineal con las medidas de homoplasia MSH y HD para las condiciones de crecimiento poblacional exploradas en esta tesis.

3) La mejor forma de correr un algoritmo bayesiano aproximado (ABC) en un modelo de crecimiento *stepwise* es utilizando una ε de 0.1 y con los siguientes estadísticos de resumen: el promedio de la varianza en el número de repeticiones por microsatélite, el promedio de la heterocigosis esperada por microsatélite y el número de haplotipos diferentes por estado.

4) Los ABC pueden estimar $\hat{\tau}$ y $\widehat{\theta_1}$ de forma más precisa que cuando se usa la distribución *mismatch* bajo un ISM.

5) Los ABC pueden estimar la MSH y la HD.

6) Se demostró la aplicación de los ABC en datos de *Pinus caribaea* para la estimación de la homoplasia, medida por MSH y HD, junto con el tiempo y la magnitud del crecimiento poblacional.

Bibliografía

Beaumont M. A., Zhang W. y Balding D. J. (2002) “Approximate bayesian computation in population genetics”, *Genetics*, **162**: 2025-2035.

Betancourt, J.L., Schuster, W.S., Mitton, J.B. y Anderson, R.S. (1991) “Fossil and genetic history of a pinyon pine (*Pinus edulis*) isolate”, *Evolution*, **72**: 1685-1697.

Bonatto, S.L. y Salzano, F.M. (1997) “Diversity and age of the four major mtDNA haplogroups and their implications for the peopling of the New World”, *American Journal of Human Genetics*, **61**: 1413-1423.

Butler, A.B. y Saidel, W.M. (2000) “Defining sameness: historical, biological, and generative homology”, *BioEssays*, **22**: 846-853.

Crow J.F., Kimura M. (1970) *An Introduction to Population GeneticsTheory,* New-York, Harper, Row, pp. 591.

Cornuet J. M., Santos, F., Beaumont M. A., Robert C. P., Marin J. M., Balding D. J., Guillemaud T. y Estoup A. (2008) “Inferring population history with DIY ABC: a user-friendly approach to approximate Bayesian computation”, *Bioinformatics*, **24**: 2713-2719.

Di Rienzo, A., Peterson, A. C., Garza, J. C., Valdes, A. M., Slatkin, M. y Freimer, N. B. (1994) “Mutational processes of simple-sequence repeat loci in human populations”, *Proceedings of the Nacional Academy of Sciences of the United States of America*, **91**: 3166-3170.

Drouin, G., Daoud, H. y Xia, J. (2008) “Relative rates of synonymous substitutions in the mitochondrial, chloroplast and nuclear genomes of seed plants”, *Molecular Phylogenetics and Evolution*, **49**: 827-831.

Drummond, A.J. y Rambaut, A. (2007) “BEAST: Bayesian evolutionary analysis by sampling trees”, *BMC Evolutionary Biology*, **7**:214.

Drummond, A.J., Rambaut, A., Shapiro, B. y Pybus, O.G. (2005) “Bayesian coalescent inference of past population dynamics from molecular sequences”, *Molecular Biology and Evolution*, **22**:1185-1192.

Echt, C.S., Deverno, L.L., Anzidei, M. y Vendramin, G.G. (1998) “Chloroplast microsatellites reveal population genetic diversity in red pine, *Pinus resinosa* Ait.”, *Molecular Ecology*, **7**: 307-316.

Emerson, B.C. y Hewitt, G.M. (2005) “Phylogeography”, *Current Biology*, **15**: 368-371.

Estoup, A. y Clegg, S. M. (2003) “Bayesian inferences on the recent island colonization history by the bird *Zosterops lateralis lateralis*”, *Molecular Ecology,* **12**: 657-674.

Estoup, A., Jarne, P. y Cornuet, J.M. (2002) “Homoplasy and mutation model at microsatellite loci and their consequences for population genetics analysis”, *Molecular Ecology*, **11**: 1591-1604.

Estoup, A., Wilson, I. J., Sullivan, C., Cornuet, J.M. y Moritz, C. (2004) “Inferring population history from microsatellite and enzyme data in serially introduced cane toads, *Bufo marinus*”, *Genetics*, **159**: 1671-1687.

Excoffier, L., Laval, G. y Schneider, S. (2005) “Arlequin (version 3.0): An integrated software package for population genetics data analysis”, *Evolutionary Bioinformatics Online*, **1**: 47-50.

Fisher, R. (1922) “On the mathematical foundations of theoretical statistics”, *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering sciences*, **222**: 309-368.

Fisher, R. (1930) “The genetical theory of natural selection”, Oxford, Clarendon Press.

Fu, Y.X. (1997) “Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection”, *Genetics*, **147**: 915-925.

Fu, Y.X. y Li, W.H. (1993) “Statistical tests of neutrality of mutations”, *Genetics*, **133**: 693-709.

Futuyma, D. (2005) *Evolution*, Massachusetts, Sinauer Associates, pp. 545-553.

Garrigan, D. y Hammer, M.F. (2006) “Reconstructing human origins in the genomic era”, *Nature Reviews Genetics*, **7**: 669-680.

Gibbard, P. y van Kolfschoten, T. (2004) “The pleistocene and holocene epochs” en *A geologic time scale*, Cambridge, Cambridge University Press, pp. 441-452.

Goldstein D.B., Linares A.R., Cavalli-Sforza L.L., Feldman M.W. (1995) “An evaluation of genetic distances for use with microsatellite loci” *Genetics*, **139**, 463–471.

Hedrick, P. (2000) *Genetics of populations*, United States of America, Jones and Bartlett Publishers, Inc., pp. 81.

Hein, J., Schierup, M.H. y Wiuf, C. (2005) *Gene genealogies, variation and evolution: a primer in coalescent theory*, New York, Oxford University Press, pp. 1- 171.

Hellenthal, G. y Stephens, M. (2007) “msHOT: modifying Hudson’s ms simulator to incorporate crossover and gene conversion hotspots”, *Bioinformatics*, **23**: 520-521.

Hewitt, G.M. (2000) “The genetic legacy of the Quaternary ice ages”, *Nature*, **405**: 907-913.

Hewitt, G.M. (2004) “Genetic consequences of climatic oscillations in the Quaternary”, *Philosophical Transactions of the Royal Society B: Biological Sciences*, **359**: 183-195.

Hey, J. and Nielsen, R. (2007) “Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics”,

Proceedings of the Nacional Academy of Sciences of the United States of America, **104**: 2785-2790.

http://www.quaternary.stratigraphy.org.uk/correlation/POSTERSTRAT_BOREAS_v2005c.pdf (9 de abril del 2009).

Hudson, R.R. (2002) “Generating samples under a Wright-Fisher neutral model of genetic variation”, *Bioinformatics*, **18**: 337-338.

Jardón-Borbolla, L., Delgado-Valerio, P., Geada-López, G., Vázquez-Lobo, A. y Piñero, D., “Phylogeography of subsection *Australes* pines in the Caribbean basin”, en prensa.

Kimmel, K., Chakraborty, R., King, J.P., Bamshad, M., Watkins, W.S. y Jorde, L.B. (1998) “Signatures of population expansion in microsatellite repeat data”, *Genetics*, **148**: 1921-1930.

Kimura, M. (1971) “Theoretical foundation of population genetics at the molecular level”, *Theoretical Population Biology*, **2**: 174-208.

Kimura M. y Crow J. (1964) “The number of alleles that can be mantained in a finite population”, *Genetics*, **49**: 725-738.

Kimura, M. y Ohta, T. (1978) “Stepwise mutation model and distribution of allelic frequencies in a finite population”, *Proceedings of the Nacional Academy of Sciences of the United States of America,* **6**: 2868-2872.

King, J.P., Kimmel M. y Chakraborty R. (2000) “A power analysis of microsatellite-based statistics for inferring past population growth”, *Molecular Biology and Evolution*, **17**: 1859-1868.

Kingman, J.F.C. (1982) “On the genealogy of large populations”, *Journal of Applied Probability*, **19**: 27-43.

Kruglyak, S., Durrett, R.T., Schug, M.D. y Aquadro, C.F. (1998) “Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations”, *Proceedings of the Nacional Academy of Sciences of the United States of America,* **95**: 10774-10778.

Kuhner, M.K. (2006) “LAMARC 2.0: maximum likelihood and Bayesian estimation of population parameters”, *Bioinformatics*, **22**: 768-770.

Kuhner, M.K., (2008) “Coalescent genealogy samplers: windows into population history”, *Trends in Ecology and Evolution*, **24**: 86-93.

Li, W.H. (1977) “Distribution of nucleotide differences between two randomly chosen cistrons in a finite population”, *Genetics*, **85**: 331-337.

Marjoram, P., Molitor, J., Plagnol, V. y Tavaré, S. (2003) “Markov chain monte carlo without likelihoods”, *Proceedings of the Nacional Academy of Sciences of the United States of America*, **100**: 15324-15328.

Marjoram, P. y Tavaré, S. (2006) “Modern computational approaches for analysing molecular genetic variation data”, *Nature Reviews Genetics*, **7**: 759-770.

Martinson, D., Pisias, N.G., Hays, J.D., Imbrie, J., Moore Jr., T.C. y Shackleton, N.J. (1987) “Age dating and the orbital theory of the ice ages: development of a high-resolution 0 to 300,000 year chronostratigraphy”, *Quaternary Research*, **27**: 1-29.

Moran, P.A. (1958) “A General Theory of the Distribution of Gene Frequencies. I. Overlapping Generations”, *Proceedings of the Royal Society of London. Series B, Biological Sciences*, **149 (934)**: 102-112.

Moreno-Letelier, A. y Piñero, D. (2009) “Phylogeographic structure of *Pinus strobiformis* Engelm. across the chihuahuan desert filter-barrier”, *Journal of Biogeography*, **36**: 121-131.

Moreno-Letelier, A. (2009) “Estructura filogeográfica de Pinus strobiformis y su relación con los cambios climáticos durante el Pleistoceno”, Tesis de Doctorado, Instituto de Ecología.

Navascués, M. y Emerson B.C. (2005) “Chloroplast microsatellites: measures of genetic diversity and the effect of homoplasy”, *Molecular Ecology*, **14**: 1333-1341.

Navascués, M., Vaxevanidou, Z., González-Martínez, S.C., Climent, J., Gil, L. y Emerson, B.C. (2006) “Chloroplast microsatellites reveal colonization and metapopulation dynamics in the Canary Island pine”, *Molecular Ecology*, **15**: 2691-2698.

Navascués, M., Hardy, O.J. y Burgarella, C. (2009) “Characterization of demographic expansions from pairwise comparisons of linked microsatellite haplotypes”, *Genetics*, **181**: 1013-1019.

Nei, M. (1978) “Estimation of average heterozygosity and genetic distance from a small number of individuals”, *Genetics*, **89**: 583-590.

Pineda-Krch, M. y Redfield, R.J. (2005) “Persistence and Loss of Meiotic Recombination Hotspots,” *Genetics,* **169**: 2319-2333.

Pritchard, J.K., Seielstad, M.T., Perez-Lezaun, A. y Feldman, M.W. (1999) “Population growth of human Y chromosomes: A study of Y chromosome microsatellites”, *Molecular Biology and Evolution*, **16**: 1791-1798.

Provan, J., Soranzo, N., Wilson, N.J., Goldstein, D.B. y Powell, W. (1999) “A low mutation rate for chloroplast microsatellites”, *Genetics*, **153**: 943-947.

Pybus, O.G. y Rambaut, A. (2002) “Genie: estimating demographic history from molecular phylogenies”, *Bioinformatics*, **18**: 1404-1405.

Pybus, O.G. y Rambaut, A. (2002b) “GENIE v3.0 User Manual” en http://evolve.zoo.ox.ac.uk/Evolve/Genie.html (19 de mayo del 2008).

Pybus, O.G., Rambaut, A. y Harvey, P.H. (2000) “An integrated framework for the inference of viral population history from reconstructed genealogies”, Genetics, 155: 1429-1437.

Ricci V. (2005) “Fitting distributions with R” en http://cran.r-project.org/doc/contrib/Ricci-distributions-en.pdf (30 de marzo del 2009).

Rogers, A.R. y Harpending, H. (1992) “Population growth makes waves in the distribution of pairwise genetic differences”, *Molecular Biology and Evolution*, **9**: 552-569.

Ross-Ibarra J., Tenaillon M. y Gaut B.S. (2009) “Historical divergence and gene flow in the genus Zea”, *Genetics*, **181**: 1399-1341.

Sainudiin, R., Durrett, R.T., Aquadro, C.F. y Nielsen, R. (2004) “Microsatellite mutation models: insights from a comparison of humans and chimpanzees”, *Genetics*, **168**: 383-395.

Schlötterer, C. (2000) “Evolutionary dynamics of microsatellite DNA”, *Chromosoma*, **109**: 365-371.

Schneider, S. y Excoffier, L. (1999) “Estimation of past demographic parameters from the distribution of pairwise differences when the mutation rates vary among sites: Application to human mitochondrial DNA”, *Genetics*, **152**: 1079-1089.

Slatkin, M. y Hudson, R.R. (1991) “Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations”, *Genetics*, **129**: 555-562.

Tajima, F. (1989) “The effect of change in population size on DNA polymorphism”, *Genetics*, **123**: 597-601.

Tilford, C.A., Kuroda-Kawaguchi, T., Skaletsky, H., Rozen, S., Brown, L. G., Rosenberg, M., McPherson, J.D., Wylie, K., Sekhon, M., Kucaba, T.A., Waterston, R.H. y Page, D.C. (2001) “A physical map of the human Y chromosome” *Nature,* **409**: 943-945.

Vad Odgaard, B. (1999) “Fossil pollen as a record of past biodiversity”, *Journal of Biogeography*, **26**: 7-17.

Valdes A.M., Slatkin M. y Freimer N.B. (1993) “Allele frequencies at microsatellite loci: the stepwise mutation model revisited”, *Genetics*, **133**: 737-749.

Van Oppen, M.J.H., Rico, C., Turner, G.F. y Hewitt, G.M. (2000) “Extensive homoplasy, nonstepwise mutations, and shared ancestral polymorphism at a complex microsatellite locus in Lake Malawi cichlids”, *Molecular Biology and Evolution,* **17**: 489–498.

Watterson, G.A. (1975) “On the number of segregating sites in genetical models without recombination”, *Theoretical Population Biology*, **7**: 256-276.

Wimberger, D., Alavez, V., Moreno-Letelier, A. y Piñero, D., “Effect of chloroplast microsatellite structure and linkage on population parameter estimation in conifers: a simulation study”. En preparación.

Wright, S. (1931) “Evolution in mendelian populations”, *Genetics*, **16**: 97-159.

Material suplementario

Tabla suplementaria 1.- Grupos de datos usados para evaluar propiedades de la distribución *mismatch*.

Grupo de datos	θ_1	θ_0	τ
1	5	0.005	1.5
2	5	0.005	3
3	5	0.005	4.5
4	5	0.005	6
5	5	0.005	7.5
6	5	0.005	9
7	15	0.015	1.5
8	15	0.015	3
9	15	0.015	4.5
10	15	0.015	6
11	15	0.015	7.5
12	15	0.015	9
13	15	0.015	10.5
14	15	0.015	12
15	15	0.015	13.5
16	15	0.015	15
17	15	0.015	16.5
18	15	0.015	18
19	15	0.015	19.5
20	15	0.015	21
21	15	0.015	22.5
22	15	0.015	24
23	15	0.015	25.5
24	15	0.015	27
25	15	0.015	28.5
26	15	0.015	30
27	30	0.03	1.5
28	30	0.03	3
29	30	0.03	4.5
30	30	0.03	6
31	30	0.03	7.5

Tabla suplementaria 1 (continuación)

32	30	0.03	9
33	30	0.03	10.5
34	30	0.03	12
35	30	0.03	13.5
36	30	0.03	15
37	30	0.03	16.5
38	30	0.03	18
39	30	0.03	19.5
40	30	0.03	21
41	30	0.03	22.5
42	30	0.03	24
43	30	0.03	25.5
44	30	0.03	27
45	30	0.03	28.5
46	30	0.03	30
Sin expansión			
47	$\theta = 5$		
48	$\theta = 15$		
49	$\theta = 30$		
50	$\theta = 5$		
51	$\theta = 15$		
52	$\theta = 30$		

Tabla suplementaria 2.- Promedio del valor de $\hat{\tau}$ inferido a partir de cien simulaciones con las combinaciones de parámetros usados en la Tabla suplementaria 1. Entre paréntesis se indica la varianza de $\hat{\tau}$.

	θ_1 = 5 (SMM)	θ_1 = 5 (ISM)	θ_1= 15 (SMM)	θ_1 = 15 (ISM)	θ_1 = 30 (SMM)	θ_1 = 30 (ISM)
tau 1.5	1.35771 (0.390918)	1.50883 (0.5542974)	1.39298 (0.1794741)	1.54501 (0.2731524)	1.40744 (0.0967729)	1.54923 (0.1421494)
tau 3	2.12242 (0.9398241)	2.70278 (2.071611)	2.33948 (0.3740755)	2.86146 (0.8440209)	2.36151 (0.1782917)	2.82665 (0.4738111)
tau 4.5	2.8915 (1.820014)	3.99875 (3.623544)	3.08095 (0.4962294)	4.17859 (1.308598)	3.16253 (0.2935232)	4.21947 (0.8173148)
tau 6	3.38643 (3.200807)	5.11805 (7.933139)	3.80312 (0.6978153)	5.70578 (2.736306)	3.92559 (0.544495)	5.63641 (1.725364)
tau 7.5	3.77996 (4.320428)	6.18963 (9.847857)	4.25289 (1.087585)	7.04565 (3.477891)	4.47721 (0.673883)	7.35244 (3.208723)
tau 9	4.07201 (5.969718)	8.56456 (210.3873)	4.80007 (1.957866)	9.2552 (69.84245)	4.98445 (0.9314403)	8.49221 (3.667235)
tau 10.5			5.03817 (2.637466)	10.1922 (75.93474)	5.44286 (1.051565)	10.46572 (4.74839)
tau 12			5.40411 (4.03691)	12.15574 (137.0442)	5.70827 (1.179086)	11.52924 (5.550702)
tau 13.5			5.75807 (4.607185)	10.56923 (24.93429)	6.2551 (1.321505)	13.81116 (64.19116)
tau 15			6.81282 (71.82145)	14.11327 (268.1286)	6.27367 (2.125988)	13.72272 (17.35128)
tau 16.5			6.3444 (6.982973)	14.68424 (206.434)	6.90533 (2.790999)	15.46432 (20.72808)
tau 18			6.30222 (8.330808)	14.56973 (224.9602)	6.81183 (2.76535)	16.33199 (24.46909)
tau 19.5			7.48544 (70.91363)	15.91318 (164.2788)	7.3275 (3.490862)	17.96117 (30.63531)
tau 21			6.10593 (6.462271)	17.74449 (1263.801)	7.6359 (4.321115)	19.36441 (31.90478)
tau 22.5			7.29034 (12.84904)	15.64027 (144.2302)	7.43304 (5.8188)	19.9167 (46.40007)
tau 24			7.98787 (72.91172)	22.81549 (2281.188)	7.39477 (5.835471)	22.83099 (144.0852)
tau 25.5			6.2972 (8.912077)	17.43631 (519.7498)	8.20482 (5.11073)	24.28072 (29.73251)
tau 27			6.14659	14.34409	7.95713	22.63848

Tabla suplementaria 2 (continuación)

			(10.3455)	(103.0912)	(9.077692)	(78.25527)
tau 28.5			6.81972 (10.85167)	15.08805 (136.2972)	8.47847 (6.724889)	26.46776 (61.931)
tau 30			8.0992 (142.8986)	19.42173 (1318.523)	8.46689 (10.48184)	24.41092 (108.4387)
No expansión (SMM, simulación 1)	4.57072 (10.71184)		9.74375 (461.4394)		8.41951 (25.46356)	
No expansión (SMM, simulación 2)	4.3283 (7.348222)		6.95799 (16.20364)		7.68522 (16.26235)	
No expansión (ISM simulación 1)	9.17606 (182.1498)		21.211 (1563.480)		24.49782 (780.6915)	
No expansión (ISM simulación 2)	10.57028 (266.3176)		25.27987 (2645.576)		19.73299 (405.1332)	

Tabla suplementaria 3.- Número de veces que el verdadero valor de τ cae dentro del intervalo de confianza con un α = 0.05.

a) θ_1 = 5

$\hat{\tau}$	$\hat{\tau}$ dentro del intervalo bajo SMM e ISM	$\hat{\tau}$ dentro del intervalo bajo ISM y no está dentro bajo SMM	$\hat{\tau}$ dentro del intervalo bajo SMM y no está dentro bajo ISM	$\hat{\tau}$ no está dentro del intervalo bajo ISM ni bajo SMM	$\hat{\tau}$ dentro del intervalo bajo ISM	$\hat{\tau}$ dentro del intervalo bajo SMM
1.5	60	4	4	32	64	64
3	48	17	5	30	65	53
4.5	54	23	0	23	77	54
6	50	22	0	28	72	50
7.5	33	45	1	21	78	34
9	31	35	3	31	66	34

b) θ_1 = 15

$\hat{\tau}$	$\hat{\tau}$ dentro del intervalo bajo SMM e ISM	$\hat{\tau}$ dentro del intervalo bajo ISM y no está dentro bajo SMM	$\hat{\tau}$ dentro del intervalo bajo SMM y no está dentro bajo ISM	$\hat{\tau}$ no está dentro del intervalo bajo ISM ni bajo SMM	$\hat{\tau}$ dentro del intervalo bajo ISM	$\hat{\tau}$ dentro del intervalo bajo SMM
1.5	78	2	9	11	80	87
3	42	37	7	14	79	49
4.5	36	40	0	24	76	36
6	36	50	1	13	86	37
7.5	24	63	0	13	87	24
9	22	68	1	9	90	23
10.5	17	69	2	12	86	19
12	14	71	1	14	85	15
13.5	11	74	0	15	85	11
15	5	66	3	26	71	8
16.5	9	70	0	21	79	9
18	4	64	0	32	68	4
19.5	1	68	0	31	69	1
21	0	52	0	48	52	0

22.5	5	60	1	34	65	6
24	4	48	0	48	52	4
25.5	1	51	0	48	52	1
27	0	40	0	60	40	0
28.5	0	42	1	57	42	1
30	0	39	0	61	39	0

c)	θ_1 = 30					
$\hat{\tau}$	$\hat{\tau}$ dentro del intervalo bajo SMM e ISM	$\hat{\tau}$ dentro del intervalo bajo ISM y no está dentro bajo SMM	$\hat{\tau}$ dentro del intervalo bajo SMM y no está dentro bajo ISM	$\hat{\tau}$ no está dentro del intervalo bajo ISM ni bajo SMM	$\hat{\tau}$ dentro del intervalo bajo ISM	$\hat{\tau}$ dentro del intervalo bajo SMM
1.5	81	7	5	7	88	86
3	44	36	6	14	80	50
4.5	17	63	1	19	80	18
6	9	73	3	15	82	12
7.5	14	74	0	12	88	14
9	11	76	0	13	87	11
10.5	7	83	0	10	90	7
12	7	84	0	9	91	7
13.5	3	90	0	7	93	3
15	5	83	0	12	88	5
16.5	1	91	0	8	92	1
18	1	91	0	8	92	1
19.5	2	88	0	10	90	2
21	1	87	0	12	88	1
22.5	1	86	0	13	87	1
24	0	89	0	11	89	0
25.5	2	88	0	10	90	2
27	0	83	0	17	83	0
28.5	1	80	0	19	81	1
30	0	80	0	20	80	0

Tabla suplementaria 4.- Número de simulaciones rechazadas en el análisis de la relación entre el sesgo relativo en la estimación de τ y varias medidas de homoplasia.

a) θ_1 = 15

Tau	Simulaciones removidas
1.5	2
3	2
4.5	0
6	0
7.5	0
9	0
10.5	4
12	4
13.5	10
15	15
16.5	8
18	13
19.5	10
21	9
22.5	13
24	9
25.5	13
27	11
28.5	20
30	21
Porcentaje de simulaciones removidas	0.082

b) θ_1 = 30

Tau	Simulaciones removidas
1.5	1
3	0
4.5	0

Tabla suplementaria 4 (continuación)

6	0
7.5	0
9	0
10.5	0
12	0
13.5	0
15	1
16.5	2
18	2
19.5	4
21	0
22.5	1
24	1
25.5	0
27	3
28.5	1
30	7
Porcentaje de simulaciones removidas	0.0115

Tabla suplementaria 5.- Relación lineal entre el sesgo relativo en la estimación de τ y varias medidas de homoplasia.

a) θ_1 = 15

	r^2	p value de ecuación de regresión múltiple
MSH	0.3008	< 2.2e-16
SH	0.003581	0.01033
MASH	0.007076	0.0003082
SASH	0.007974	0.0001275
HS	0.07621	< 2.2e-16
HD	0.3479	< 2.2e-16
HBC	0.08903	< 2.2e-16

b) θ_1 = 30

	r^2	p value de ecuación de regresión múltiple
MSH	0.4858	< 2.2e-16
SH	0.04695	< 2.2e-16
MASH	0.08108	< 2.2e-16
SASH	0.06083	< 2.2e-16
HS	0.1056	< 2.2e-16
HD	0.5302	< 2.2e-16
HBC	0.2969	< 2.2e-16

Tabla suplementaria 6.- Promedio de la correlación del sesgo relativo en la estimación de $\hat{\tau}$ a través de datos de microsatélites creados con distintos valores de τ.

a) θ_1 = 15

Tau	MSH	SH	MASH	SASH	HS	HD	HBC
1.5 – 6	0.41928942	0.15681796	0.08495089	0.19650594	0.12609336	0.43442741	-0.0128671
7.5 – 12	0.45874715	0.103116	0.09363937	0.04241175	0.2395662	0.6224882	0.06914131
13.5 - 18	0.3941279	-0.05442318	-0.04059453	-0.05486024	0.13828047	0.45349482	-0.01626866
19.5 - 24	0.30603833	-0.10963212	-0.1080538	-0.05040566	0.09211728	0.35130905	-0.04862073
25.5 - 30	0.38484641	-0.18377569	-0.16831263	-0.09006262	0.16469287	0.46026045	-0.09578011

b) θ_1 = 30

Tau	MSH	SH	MASH	SASH	HS	HD	HBC
1.5 – 6	0.39324222	0.19211355	0.07491722	0.30575913	0.12152641	0.4223666	0.05856685
7.5 – 12	0.50942269	0.08868609	0.0995812	-0.03872244	0.25550584	0.6525844	-0.05077809
13.5 - 18	0.3805954	0.04111882	0.0630191	-0.05453528	0.15000494	0.41746088	-0.13022627
19.5 - 24	0.31616585	-0.0809222	-0.03416755	-0.16455551	0.05565574	0.31999539	-0.15588729
25.5 - 30	0.1881421	-0.06696298	-0.05347104	-0.05188513	0.07076277	0.36875351	-0.02218556

Tabla suplementaria 7.- Número de simulaciones realizadas antes de obtener cien aceptaciones.

Simulación	Tau	a + b + c	a + b + c + d	a + b + c + e	a + b + c + f	a + b + c + g	a + b	a + c	b + c	a + b + c + e + f + g
1	1.5	132701	171945	714556	612516	132701	79734	88296	29150	1568294
2	3	83730	183922	357135	327706	83730	44134	69628	14044	1458973
3	4.5	71625	157473	388490	1136645	71625	19364	54030	10303	605162
4	6	71226	117710	305752	835043	71226	30869	61003	10609	529887
5	7.5	69889	135690	213588	7431229	69889	18693	55546	8944	423747
6	9	79791	92200	225928	432213	79791	27411	68793	8787	309386
7	10.5	41553	59737	123201	158131	41553	7594	37655	5048	150444
8	12	22984	28376	72602	217774	22984	4423	22106	4167	99959
9	13.5	45209	55633	205243	578698	45209	11662	27054	6448	309511
10	15	34285	48690	118240	318224	34285	8894	24924	5696	152291

Tabla suplementaria 8.- Relación lineal entre la estimación y el valor real de HD, MSH y τ.

	R cuadrada	p value de ecuación de regresión lineal
HD	0.587	< 2.2e-16
MSH	0.8194	< 2.2e-16
τ (CORAGHE)	0.6518	< 2.2e-16
τ (*Mismatch* distribution)	0.7181	< 2.2e-16

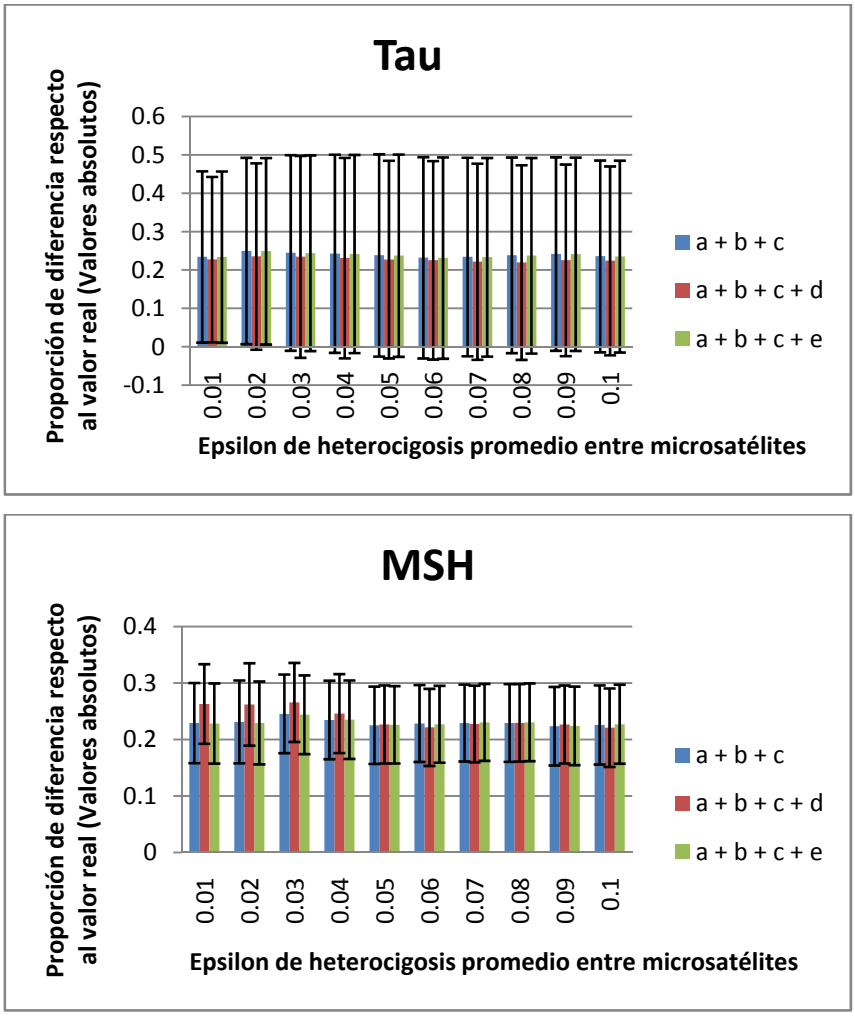
Tabla suplementaria 9.- Relación lineal entre la estimación y el valor real de HD, MSH y τ.

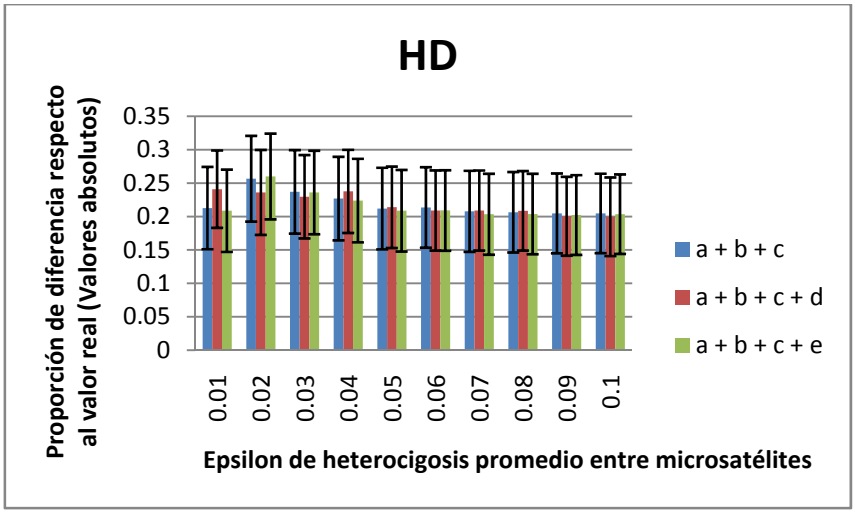
τ = 3		
	R cuadrada	p value de ecuación de regresión lineal
HD	0.01361	0.2478
MSH	0.2982	4.20E-09
τ = 6		
	R cuadrada	p value de ecuación de regresión lineal
HD	0.0035	0.5536
MSH	0.3295	4.28E-10
τ = 9		
	R cuadrada	p value de ecuación de regresión lineal
HD	0.0132	0.2542
MSH	0.1969	3.77E-06

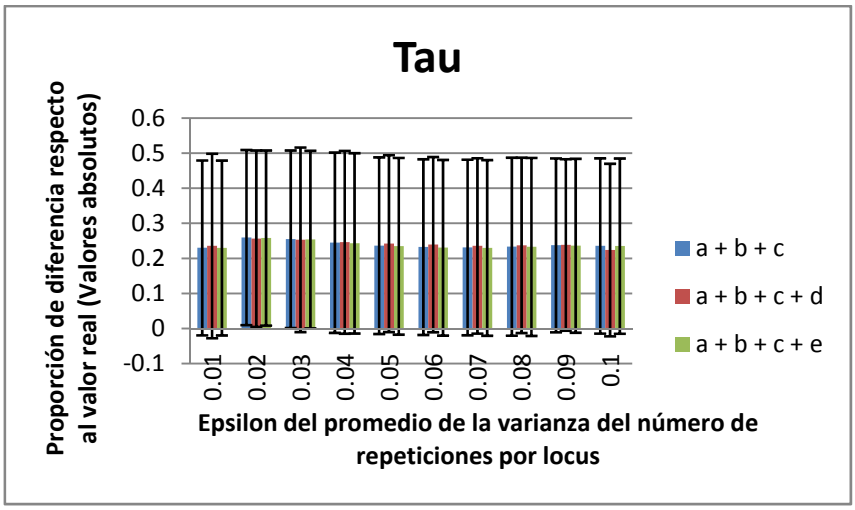
Tabla suplementaria 10.- Relación lineal entre la estimación y el valor real de SH, SASH, HS y CH.

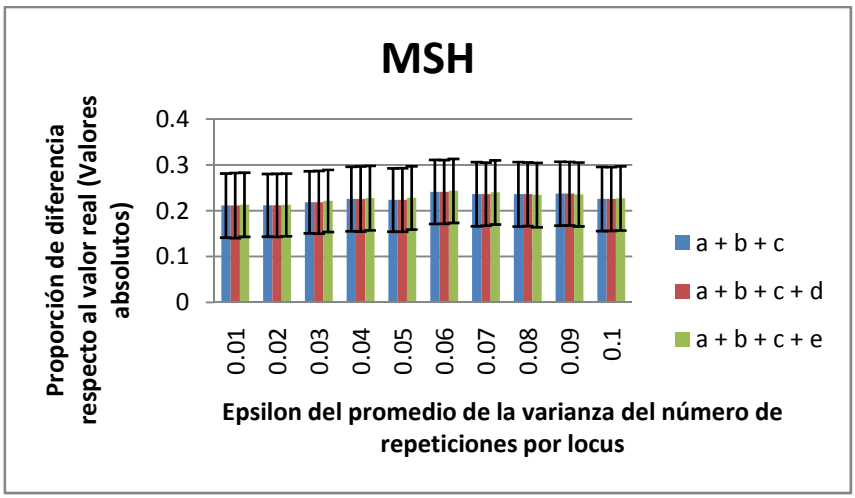
Diferentes valores de τ		
	R cuadrada	p value de ecuación de regresión lineal
SH	0.8944	< 2.2e-16
SASH	0.2982	4.43E-09
HS	0.2341	3.40E-07
CH	0.7944	< 2.2e-16
$\tau = 3$		
	R cuadrada	p value de ecuación de regresión lineal
SH	0.442	4.61E-14
SASH	0.008832	3.52E-01
HS	0.1254	0.0003004
CH	0.05529	0.01853
$\tau = 9$		
	R cuadrada	p value de ecuación de regresión lineal
SH	0.2573	7.20E-08
SASH	0.125	3.08E-04
HS	0.01216	0.2748
CH	0.1939	4.54E-06

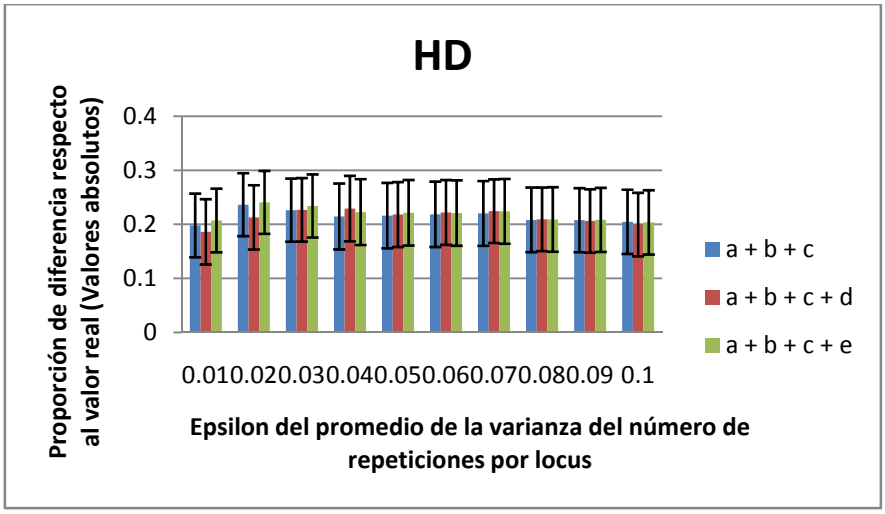
Figura suplementaria 1.- Cambio en los valores de ε. El valor debajo de cada columna indica el valor de epsilón usado para cada estadístico. Cada letra indica uno de los estadísticos usados: a) el promedio de la varianza en el número de repeticiones por microsatélite; b) el promedio de la heterocigosis por microsatélite; c) el número de haplotipos diferentes por estado; d) el número de sitios segregantes en todos los microsatélites y e) la heterocigosis tomando en cuenta todo el haplotipo. Las barras son el error estándar de la media de la proporción de diferencias respecto al valor real.

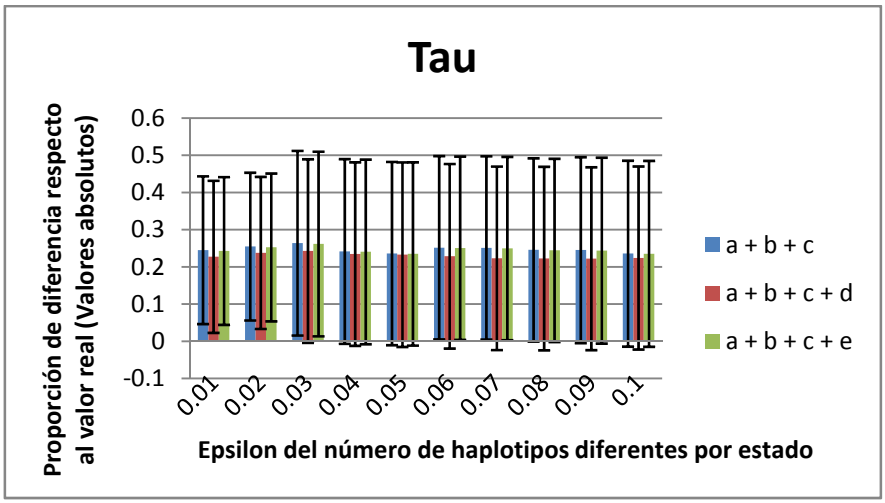


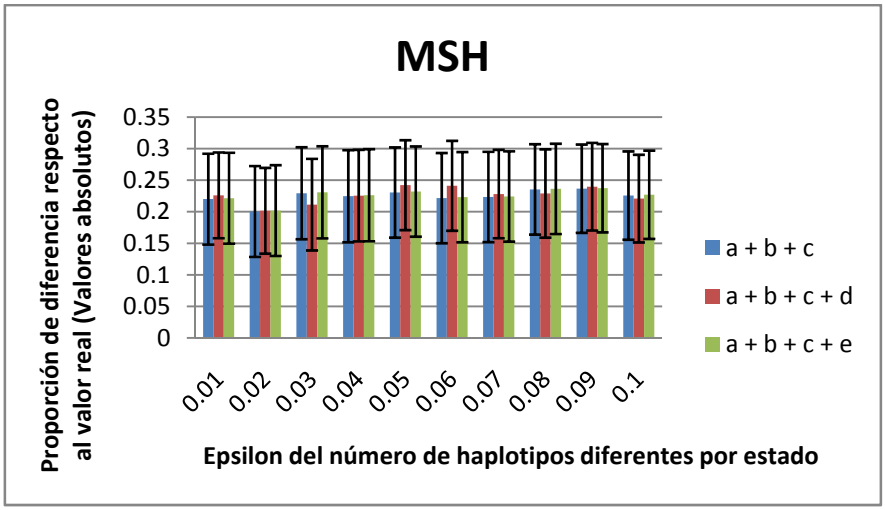

HD
Proporción de diferencia respecto al valor real (Valores absolutos)
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
0.09
0.1
a + b + c
a + b + c + d
a + b + c + e
Epsilon de heterocigosis promedio entre microsatélites

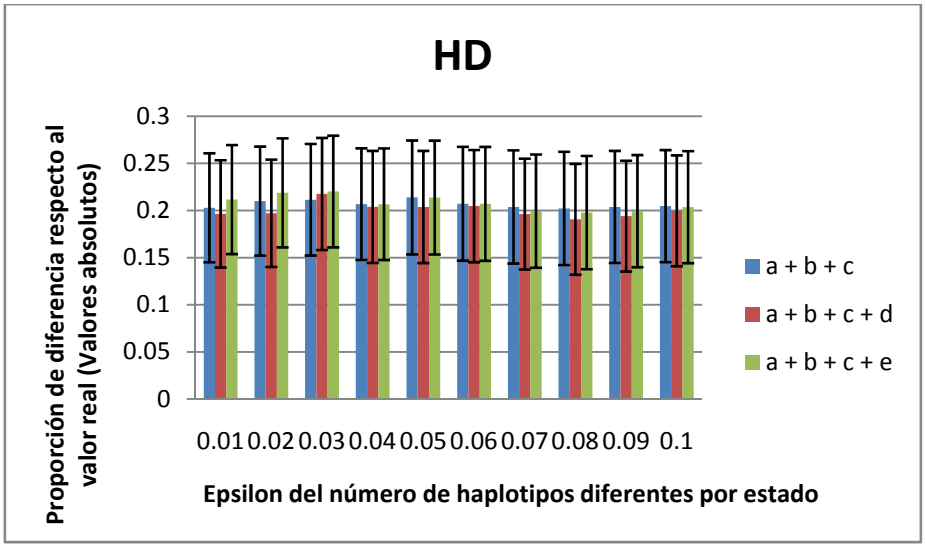

Tau
Proporción de diferencia respecto al valor real (Valores absolutos)
0.6
0.5
0.4
0.3
0.2
0.1
0
-0.1
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
0.09
0.1
a + b + c
a + b + c + d
a + b + c + e
Epsilon del promedio de la varianza del número de repeticiones por locus

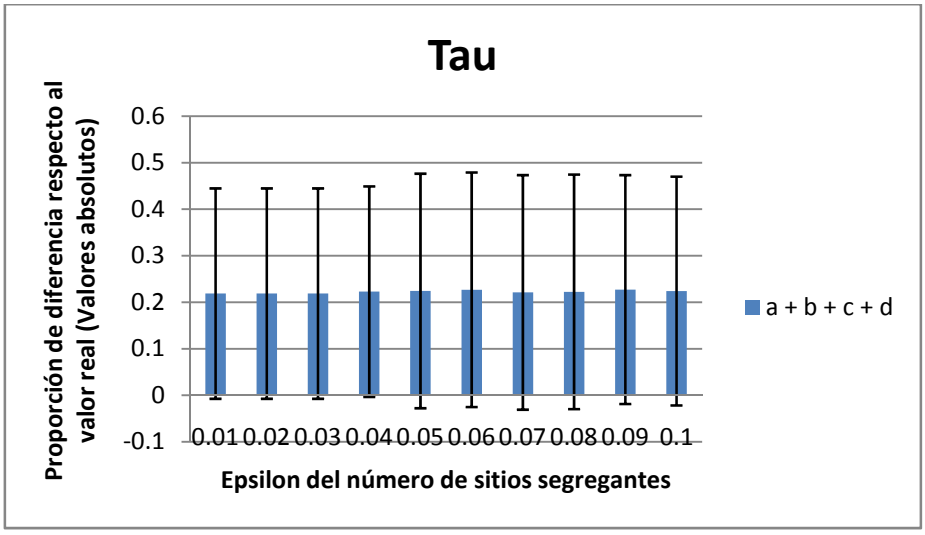

MSH
Proporción de diferencia respecto al valor real (Valores absolutos)
0.4
0.3
0.2
0.1
0
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
0.09
0.1
a + b + c
a + b + c + d
a + b + c + e
Epsilon del promedio de la varianza del número de repeticiones por locus

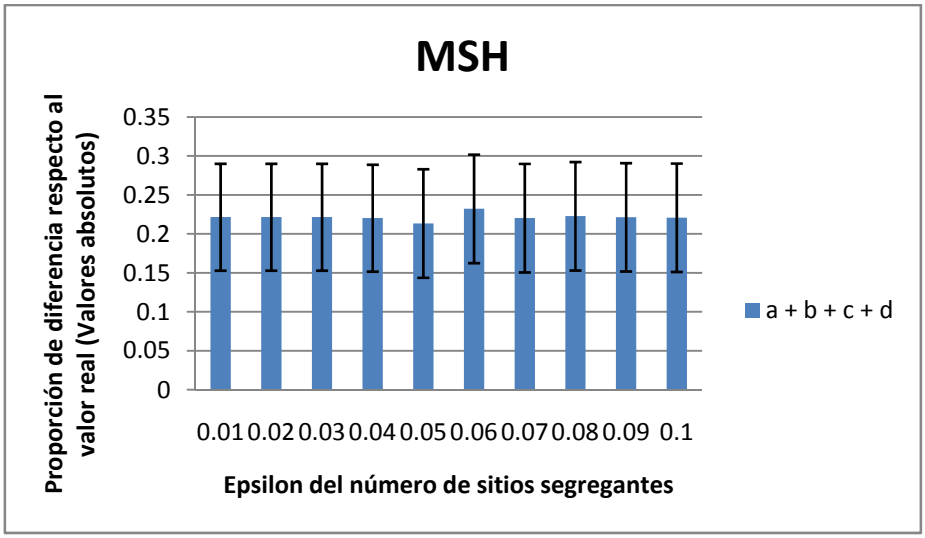

HD
Proporción de diferencia respecto al valor real (Valores absolutos)
0.4
0.3
0.2
0.1
0
0.01 0.02 0.03 0.04 0.05 0.06 0.07 0.08 0.09 0.1
a + b + c
a + b + c + d
a + b + c + e
Epsilon del promedio de la varianza del número de repeticiones por locus

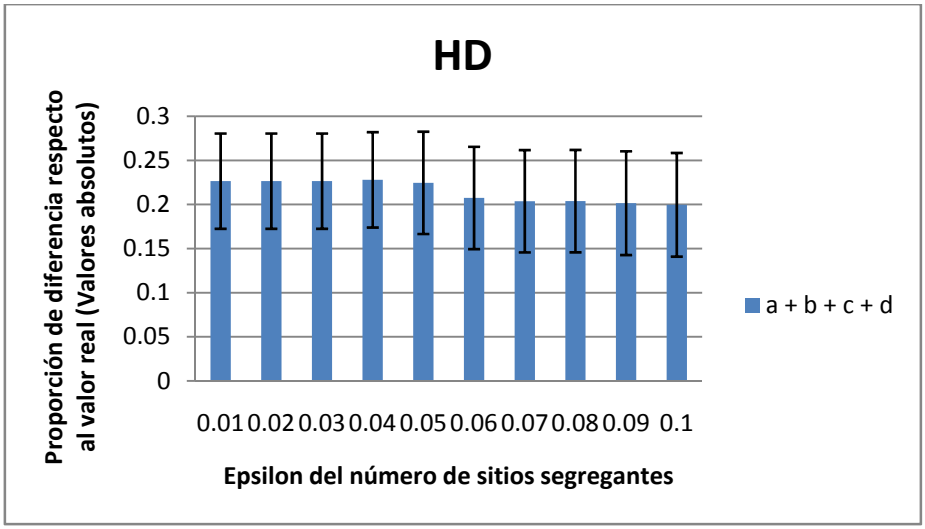

Tau
Proporción de diferencia respecto al valor real (Valores absolutos)
0.6
0.5
0.4
0.3
0.2
0.1
0
-0.1
0.01 0.02 0.03 0.04 0.05 0.06 0.07 0.08 0.09 0.1
a + b + c
a + b + c + d
a + b + c + e
Epsilon del número de haplotipos diferentes por estado

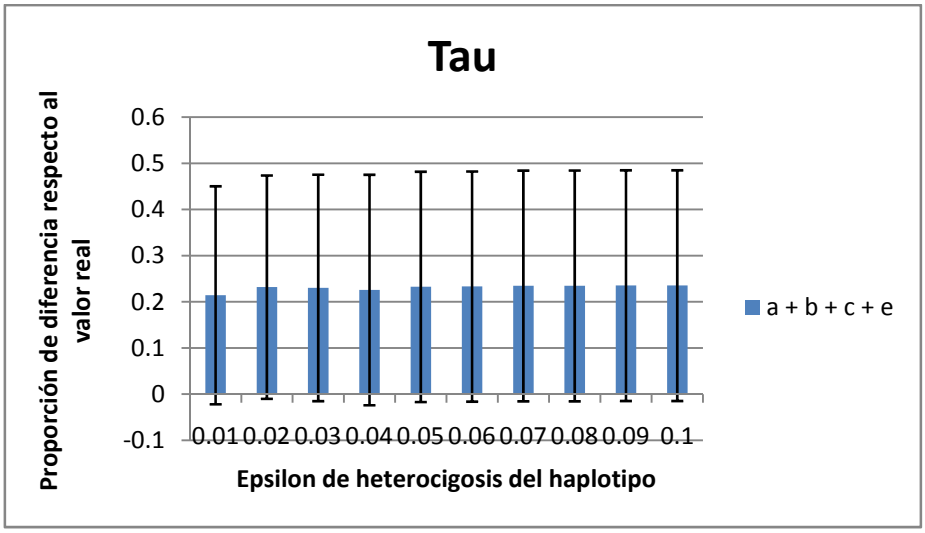

MSH
Proporción de diferencia respecto al valor real (Valores absolutos)
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0.01 0.02 0.03 0.04 0.05 0.06 0.07 0.08 0.09 0.1
a + b + c
a + b + c + d
a + b + c + e
Epsilon del número de haplotipos diferentes por estado

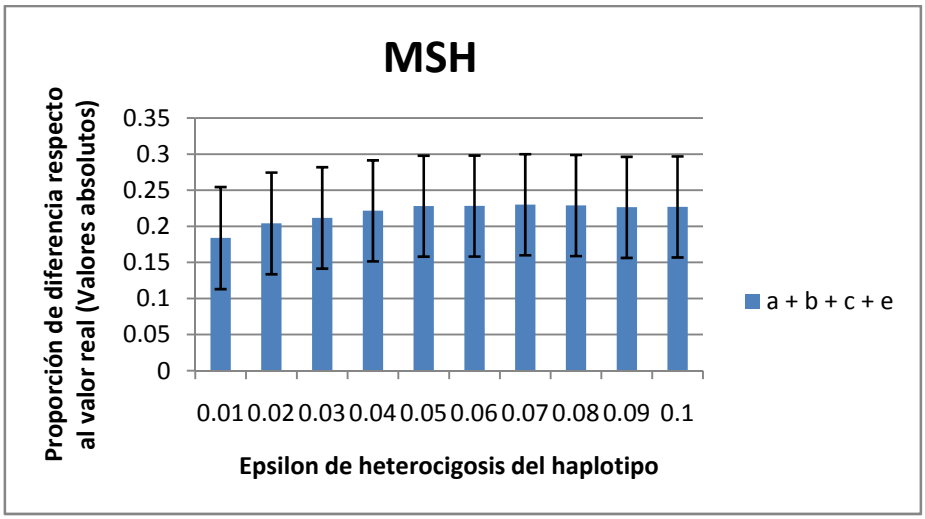

HD
Proporción de diferencia respecto al valor real (Valores absolutos)
0.3
0.25
0.2
0.15
0.1
0.05
0
0.01 0.02 0.03 0.04 0.05 0.06 0.07 0.08 0.09 0.1
Epsilon del número de haplotipos diferentes por estado
a + b + c
a + b + c + d
a + b + c + e

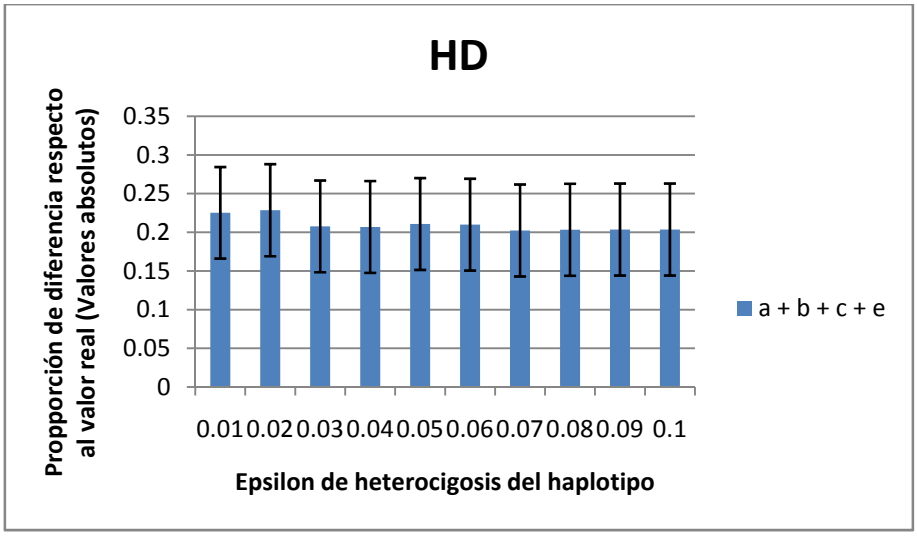

Tau
Proporción de diferencia respecto al valor real (Valores absolutos)
0.6
0.5
0.4
0.3
0.2
0.1
0
-0.1
0.01 0.02 0.03 0.04 0.05 0.06 0.07 0.08 0.09 0.1
Epsilon del número de sitios segregantes
a + b + c + d


MSH
Proporción de diferencia respecto al valor real (Valores absolutos)
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0.01 0.02 0.03 0.04 0.05 0.06 0.07 0.08 0.09 0.1
Epsilon del número de sitios segregantes
a + b + c + d


HD
Proporción de diferencia respecto al valor real (Valores absolutos)
0.3
0.25
0.2
0.15
0.1
0.05
0
0.01 0.02 0.03 0.04 0.05 0.06 0.07 0.08 0.09 0.1
Epsilon del número de sitios segregantes
a + b + c + d


Tau
Proporción de diferencia respecto al valor real
0.6
0.5
0.4
0.3
0.2
0.1
0
-0.1
0.01 0.02 0.03 0.04 0.05 0.06 0.07 0.08 0.09 0.1
Epsilon de heterocigosis del haplotipo
a + b + c + e


MSH
Proporción de diferencia respecto al valor real (Valores absolutos)
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0.01 0.02 0.03 0.04 0.05 0.06 0.07 0.08 0.09 0.1
Epsilon de heterocigosis del haplotipo
a + b + c + e


HD
Propporción de diferencia respecto al valor real (Valores absolutos)
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0.01 0.02 0.03 0.04 0.05 0.06 0.07 0.08 0.09 0.1
Epsilon de heterocigosis del haplotipo
a + b + c + e