Universitat Oberta de Catalunya

uoc.edu

Análisis de la variación nucleotídica de la región génica que incluye el gen *phantom* en poblaciones de *Drosophila melanogaster* de África y Norteamérica.

Mikel Zarandona Garai
Máster en Bioinformática y Bioestadística
Evolución molecular

Directora del TFM:
Dorcas Orengo Ferriz
Profesor responsable de la asignatura:
Ferran Prados Carrasco

01/2021

FICHA DEL TRABAJO FINAL

Título del trabajo:	Análisis de la variación nucleotídica de la región génica que incluye el gen *phantom* en poblaciones de *Drosophila melanogaster* de África y Norteamérica.
Nombre del autor:	*Mikel Zarandona Garai*
Nombre del consultor/a:	*Dorcas Orengo Ferriz*
Nombre del PRA:	*Ferran Prados Carrasco*
Fecha de entrega (mm/aaaa):	01/2021
Titulación:	*Máster en bioinformática y bioestadística*
Área del Trabajo Final:	*TFM- Bioinformatica y Bioestadística Area 3*
Idioma del trabajo:	*Castellano*
Palabras clave	*Drosophila melanogaster, selección natural y variación nucleotídica.*

Resumen del Trabajo (máximo 250 palabras): *Con la finalidad, contexto de aplicación, metodología, resultados y conclusiones del trabajo.*

La especie *Drosophila melanogaster* es originaria de África subsahariana y mediante el "comensalismo" con los humanos se ha expandido por todo el mundo. Debido a la relativamente reciente expansión, los efectos de la adaptación al nuevo clima además del efecto demográfico se deberían de reflejar en el genoma. En una población europea, se habían encontrado huellas de la selección en una región del cromosoma X que incluye el gen *phantom*. En este trabajo se analiza la variación nucleotídica en esta misma región en busca de huellas de la selección en cuatro poblaciones de *D. melanogaster* (una norteamericana y tres africanas. Asimismo, se comparan las poblaciones africanas para ver si se han diferenciado. Para ello, las secuencias se ubican en el genoma, se alinean y se analizan mediante los programas bioinformáticos BLAST, MEGA y DnaSp. Con los resultados obtenidos se llega a las siguientes conclusiones: En el caso de las tres poblaciones africanas se observa que se rechaza la neutralidad, pero probablemente este efecto se genere en otra región cercana. En el caso de la población estadounidense la diversidad nucleotídica es muy baja. Aunque se rechaza la neutralidad también se deben de realizar más análisis para comprobar con total certeza la existencia de selección en esta región. Por otro lado, las poblaciones africanas se han diferenciado entre sí, pero al realizar comparaciones en grupos de dos en dos la única diferenciada es la población etíope

Abstract (in English, 250 words or less):

Drosophila melanogaster is a specie originated in central Africa that moved to warmer zones. The relatively recent expansion must reflect the effect of the adaptation to the new climate in addition to the demographic effect. The intents of this document are to make a polymorphism analysis in order to uncover the footprint left by selection in the X chromosome where the *phantom* gene is located and to compare three African populations. The footprint left by selection in this region was uncovered in a European population, so now the survey´s aim is to uncover this effect in a North American and three African populations. Firstly, the sequences are located in the genome, aligned and analyzed using bioinformatic tools such as BLAST, MEGA and DnaSP. As seen in the results, in the case of the three African populations neutrality is rejected, but this effect is probably generated in another nearby region. In the case of the American population, nucleotide diversity is very low. Neutrality is rejected also, but further analysis must be performed to fully verify if selection exists in this region. To finish, African populations have been differentiated from each other, but when making comparisons in groups of two, the only differentiated population is the Ethiopian one.

Índice

Lista de figuras

1. Introducción

1.1 Contexto y justificación del Trabajo

Cambios bióticos y abióticos en el ambiente generalmente suscitan cambios adaptativos debido a la selección de las mutaciones que ofrecen una ventaja en el nuevo ambiente[1]. Por lo que el estudio de poblaciones derivadas después de una expansión de las especies es ideal para observar los efectos que genera la selección natural en la población debido a los nuevos ambientes. Bien es cierto que no solo se impondrán los efectos de la selección natural en el genoma, sino que la historia demográfica generada por la expansión también se verá reflejada en el genoma de esa especie. Diferentes procesos evolutivos afectan también a los patrones de variación y, a veces, imitan el efecto de la selección positiva[2] haciendo más difícil encontrar las huellas de la selección.

Drosophila melanogaster es una especie original de África subsahariana que se expandió de zonas tropicales a zonas más templadas tras la última glaciación hace 10000-15000 años[3] a través del "comensalismo" con los humanos. Esta expansión ha generado una reducción de su variación nucleotídica en las regiones colonizadas tanto por el efecto fundador como por el efecto de la selección natural ante el desafío de enfrentarse a nuevos ambientes. Diferentes estudios de variación nucleotídica de loci individuales del cromosoma X demuestran que en las poblaciones no africanas los niveles de variación son más bajos que en el caso de las poblaciones africanas[4-5].

Las huellas demográficas y de la selección natural en la variación nucleotídica se han encontrado mediante diferentes estudios independientes de patrones de variación en una o varias regiones[6-7] y en estudios multilocus de variación de las secuencias de una población en poblaciones ancestrales y derivadas de *D. melanogaster* [8].

Drosophila melanogaster es uno de los organismos modelo en genética y evolución molecular. El patrón de variación nucleotídica de las poblaciones de las distintas regiones geográficas indica la historia evolutiva que ha tenido la especie. En un estudio realizado por Orengo y Aguadé[9] se analizaron 17 fragmentos a lo largo de una región de ~190 kb y se encontraron huellas de la selección natural. El fragmento focal (fragmento 0) se encontró en otro estudio de Orengo y Aguadé[2] en el que se realizó un estudio multilocus de patrones de variación. Se hallo una región con una variabilidad nucleotídica muy baja y se observó que el foco de la selección natural se centraba en una región de ~32 kb en las cuales se hallan los fragmentos a analizar. En esta región de 32 ~kb se hallan el gen *phantom* y el gen *Cyp18a1.* El gen *phantom* codifica la proteína CYP306A1, una enzima del citocromo P450 en la vía

esteroidogénica, que podría estar relacionado con la resistencia a pesticidas.

Figura 1: Mapa del mundo con la ubicación de las poblaciones que se han analizado.

Por ello se analizarán tres poblaciones ancestrales (África) y una población derivada (Norteamérica) para analizar la región genómica en la cual se han encontrado huellas de la selección en una población catalana. Además, se compararán las poblaciones ancestrales entre ellas para comprobar las diferencias que puedan existir entre las poblaciones de la región geográfica ancestral.

1.2 Objetivos del Trabajo

A continuación, se exponen los objetivos que se buscan del trabajo.

1.2.1 Objetivos generales

1. Analizar la variación nucleotídica de poblaciones ancestrales y derivadas en *Drosophila melanogaster* en una región que incluye el gen *phantom*.

2. Comparar tres poblaciones del área de origen de la especie.

1.2.2 Objetivos específicos

1. Analizar la variación nucleotídica de poblaciones ancestrales y derivadas en *Drosophila melanogaster* en una región que incluye el gen *phantom*.

1.1 Localizar en el genoma de referencia las secuencias a analizar.

1.2 Elegir las poblaciones y los individuos idóneos.

1.3 Obtener las secuencias correspondientes.

1.4 Analizar las secuencias mediante indicadores estadísticos.

1.5 Comparar los resultados con los resultados de los trabajos anteriores.

2. Comparar tres poblaciones del área de origen de la especie.

2.1 Realizar un análisis de diferenciación genética.

2.2 Buscar posibles razones de la posible diferenciación genética.

1.3 Enfoque y método seguido

Primero se debe localizar la región genómica que ha sido analizada en el estudio realizado en la población de Barcelona en las poblaciones que se van a analizar. Para ello, se utilizará la herramienta BLAST una herramienta útil para buscar similitudes entre secuencias biológicas como ADN, ARN o proteínas[10].

Después se escogerán varios ejemplares de cada población de una base de datos. La base de datos que se utilizará será Drosophila Genome Nexus (DGN)[11] ya que hay muchos individuos de las regiones que se quieren analizar. Las secuencias se conseguirán del browser Popfly[12]. La elección de los individuos no será al azar, sino que se escogerán los individuos que en las secuencias genómicas que se quieren analizar tengan la menor cantidad de bases sin determinar y procurando que las secuencias se hayan obtenido de cigotos haploides o por otro método que asegure que la secuencia corresponde a un único alelo.

Finalmente, se analizarán las secuencias mediante varios indicadores para poder proceder al análisis y después comparar los resultados con el estudio anterior. Para realizar el análisis se pueden utilizar distintos programas, pero en este caso se usará DnaSP, una herramienta popular para realizar análisis genéticos poblacionales exhaustivos en alineaciones de múltiples secuencias[13]. Después se comprobará la significación estadística de los indicadores estadísticos mediante simulaciones realizadas mediante el programa mlcoalsim[14].

1.4 Planificación del Trabajo

Ahora se explicarán las tareas que se realizarán en el Trabajo de Fin de Máster. También se mostrará un diagrama con las fechas de inicio y final de cada tarea y entrega y se analizarán los posibles riesgos que puedan afectar a la realización del TFM.

1.4.1. Tareas

Las tareas son las siguientes:

- Localización en el genoma de referencia de las secuencias a analizar
- Elección de los individuos más idóneos de la base de datos.
- Obtención de las secuencias nucleotídicas de cada individuo para todas las secuencias a analizar.
- Análisis de las secuencias mediante los indicadores *D*, *H*, y π.
- Estudio de la diferenciación genética de las poblaciones africanas mediante *Snn.*
- Simulación de las poblaciones.
- Comparación de los resultados con los resultados de estudios anteriores.
- Corrección de errores y finalización de la memoria

1.4.2. Calendario

A continuación, se muestra el calendario de trabajo del TFM. Se tiene en cuenta la disponibilidad del autor que realizará el trabajo (todos los días de la semana por la mañana y por la tarde), compaginándolo con el seguimiento de 3 asignaturas más del máster.

El inicio del TFM lo marca la primera PEC, el 16 de septiembre. La primera tarea será el 14 de octubre, después de entregar las dos primeras PEC. Esta tarea es la de identificar la localización de las secuencias nucleotídicas en el genoma. Después se realizarán dos tareas más en octubre, elegir los individuos de cada población y conseguir las secuencias nucleotídicas para su posterior análisis.

A finales del mes de octubre se comenzará con el análisis de las secuencias nucleotídicas mediante indicadores y se acabará el 24 de noviembre. Después, Se realizarán las simulaciones de las poblaciones para observar la significación de los indicadores. En la primera quincena de diciembre se leerán artículos para comparar con los resultados que se obtienen en este trabajo. Este mes también se comenzará la redacción del trabajo final.

En la segunda quincena de diciembre se comenzará el inicio del cierre de la memoria. Con esta tarea se finalizará la memoria y será entregada.

Finalmente, en enero se realizarán las dos últimas partes del trabajo: La presentación y la defensa pública.

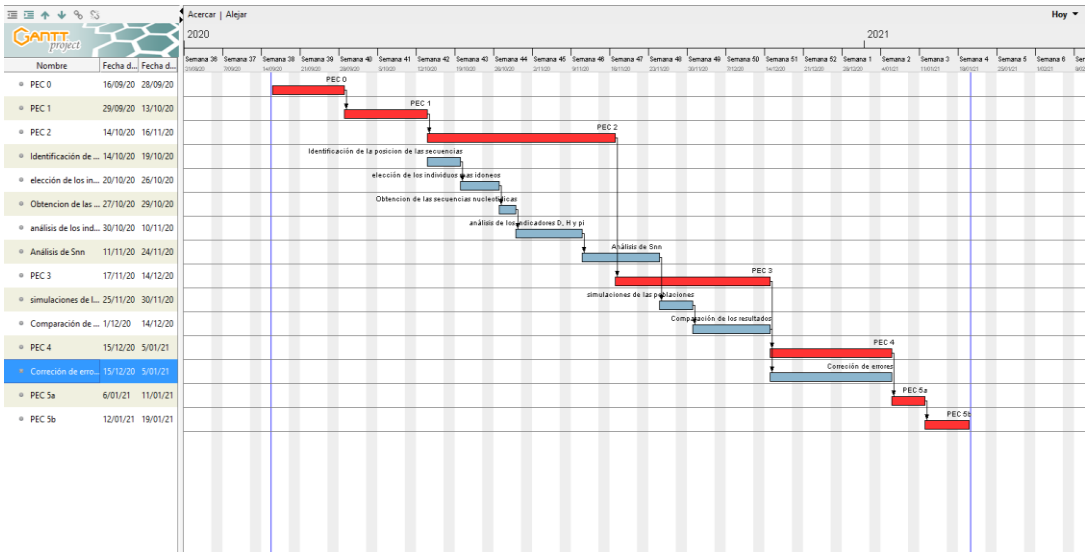


Figura 2: Diagrama de Gantt de las tareas.

1.4.3. Hitos

A continuación, se muestran los hitos, fechas clave para el trabajo, que es importante ser estricto en el cumplimiento de las fechas, ya que repercutiría en las siguientes tareas o entregas.

Tabla 1. Hitos y sus fechas de entrega

Hito	PEC	Fecha de entrega
Plan de trabajo	PEC 1	13/10/2020
Identificar la posición de las secuencias	PEC 2	19/10/2020
Obtención de las secuencias nucleotídicas	PEC 2	28/10/2020
Análisis de las secuencias mediante los indicadores *D*, *H*, y π.	PEC 2	10/11/2020
Estudio de la diferenciación genética de las poblaciones africanas mediante *Snn.*	PEC 3	24/11/2020
Simulación de las poblaciones	PEC 3	30/11/2020
Comparar los resultados	PEC 3	14/12/20
Entregar memoria	PEC 4	05/01/2021
Elaborar presentación	PEC 5a	10/01/2021
Realizar defensa pública	PEC 5b	20/01/2021

1.4.4. Análisis de riesgos

Durante la realización de este trabajo podrían aparecer problemas de distinta índole que afecten negativamente a su progreso temporal.

Estos son varios de los factores para tener en cuenta:

- Gestión inadecuada del tiempo. Para evitarlo se realiza una planificación, con la intención de gestionar bien el trabajo.
- Problemas con la instalación y/o uso de los programas informáticos. Para evitarlo se preverá una cantidad de días para acomodarse a cada programa.
- Obtención de secuencias demasiado incompletas para la región que se quiere analizar que impediría el análisis poblacional de la variabilidad nucleotídica. Para evitarlo se realizará un análisis de cada individuo y se escogerán las mejores opciones para los análisis.
- Reconocimiento tardío de errores y tiempo limitado para su corrección. Para evitarlo se contactará con la tutora del máster frecuentemente para la supervisión de los diferentes procesos y resultados.

1.5 Breve sumario de productos obtenidos

El trabajo se finalizará cuando se realicen los siguientes 4 entregables: Plan de trabajo, memoria, presentación virtual y autoevaluación del proyecto.

1.5.1. Plan de trabajo

El objetivo de este documento es concretar, delimitar y describir el trabajo que se va a realizar. Se especifican los objetivos del TFM y se muestra la planificación de los hitos y la temporización prevista.

1.5.2. Memoria

En este documento se registrará todo el trabajo realizado durante el TFM. Se explicará el contexto del trabajo y el interés que tiene para la sociedad. Se detallarán los objetivos a cumplir y los métodos a seguir y, finalmente los resultados.

1.5.3. Presentación virtual

Se realizará una presentación que sintetizará el trabajo realizado y los resultados obtenidos en el cual se ofrecerá una perspectiva general del TFM y se recogerán los aspectos más relevantes del proyecto. Se realizará con un material de apoyo visual.

1.5.4. Autoevaluación del proyecto

En este documento se mostrará si se han podido cumplir los objetivos planeados al principio del semestre. Se razonará el porqué de cumplir o no cumplir los objetivos.

1.6 Breve descripción de los otros capítulos de la memoria

La memoria está compuesta por 6 apartados sin considerar la introducción:

- Materiales y métodos: En este capítulo se detallan los procesos realizados durante el estudio. De igual manera se informa acerca de las herramientas bioinformáticas empleadas.
- Resultados y discusión: Se presentan y se discuten los resultados obtenidos tras el análisis realizado.
- Conclusiones: Se valoran los resultados obtenidos y la realización personal del trabajo.
- Glosario: Definición de conceptos y aclaración de siglas y acrónimos empleados durante la memoria.
- Bibliografía: Se muestran todas las fuentes de información mencionadas durante el trabajo.
- Anexos: Último apartado en el cual se exponen el listado de los individuos analizados y los archivos de entrada para las simulaciones de las poblaciones.

2. Materiales y métodos

Se realizan los siguientes procesos

2.1. Secuencias analizadas

Para la realización de los análisis se han estudiado 3 fragmentos de una región de la secuencia del genoma de *D. melanogaster* en los que se habían encontrado huellas de la selección natural en un trabajo previo[9] en una población catalana. Estos tres fragmentos tienen el mismo nombre que poseen en el trabajo de Orengo y Aguadé[9] (fragmento 0-20, fragmento 22 y fragmento 32) e indican la posición relativa en kilobases al fragmento 0. El primer nucleótido del fragmento 0-20 se encuentra en la posicion 18.566.173 del cromosoma X en el genoma de la página de Popfly[12]. Los fragmentos tienen longitudes distintas. El fragmento 0-20 consta de 20287 nucleótidos y alberga el gen *phantom* y el gen *Cyp18a1*, el fragmento 22 de 762 nucleótidos y el fragmento 32 de 845 nucleótidos.

Las secuencias de la población catalana están depositadas en EMBL/GenBank Data Libraries bajo el número de acceso AM411681–AM41186.0 y se ha realizado un BLAST[10] para ubicarlos en el genoma. BLAST se realizó en la página de flybase[15].

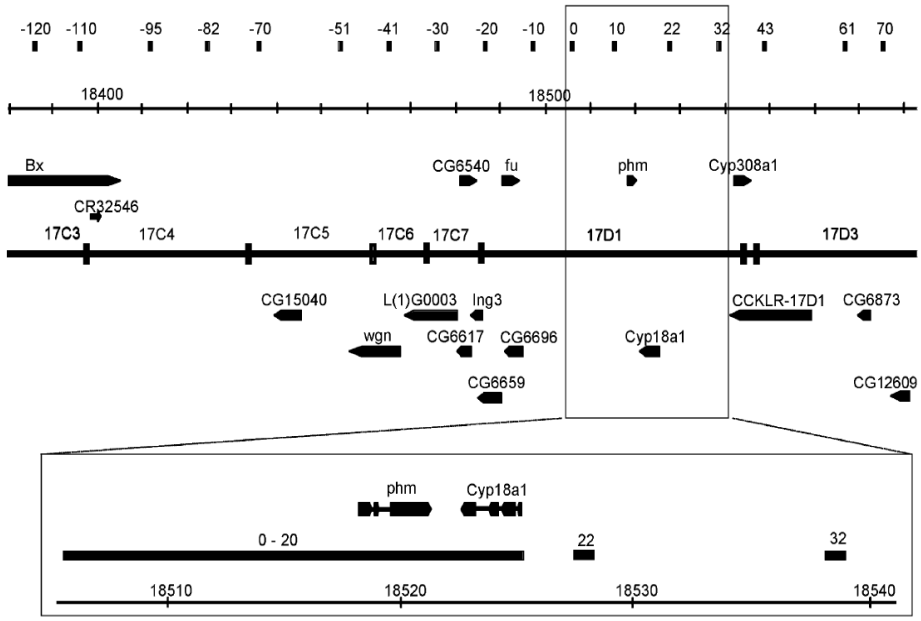


Figura 3: Esquema de la región analizada y sus alrededores [9].

Los fragmentos analizados se han obtenido a partir de las secuencias del cromosoma X de 15 individuos de la especie *Drosophila melanogaster* de 4 poblaciones de áreas geográficas diferentes (Estados Unidos, Etiopia, Ruanda y Zambia) y una secuencia del cromosoma X de un individuo de *Drosophila simulans*. Las secuencias de *D. melanogaster* han sido obtenidas de la base de datos Drosophila Genome Nexus

(DGN) mediante el Gbrowser de la página popfly[16]. La secuencia de *Drosophila simulans* se ha obtenido de la página Ensembl Metazoa[17].

2.2 Eleccion de las poblaciones y los individuos

Para poder realizar un estudio comparativo de las distintas poblaciones, se deben obtener secuencias para un número de individuos suficientes y similar en todas ellas. Para ello, primero se investigó en las bases de datos qué poblaciones de *D. melanogaster* disponen de suficientes individuos secuenciados totalmente. En lo posible, se trató de que las secuencias provinieran de cigotos haploides para asegurar que la secuencia corresponde a un único alelo. Se eligieron 3 poblaciones a lo largo de África Oriental que podrían representar el área de origen de *D. melanogaster* y con suficientes secuencias obtenidas a partir de cigotos haploides (Zambia, Ruanda y Etiopia en las localidades de Siavonga, Gikongoro y Fiche, respectivamente)[18-19]. Una población de Estados Unidos también fue escogida debido a que las poblaciones de *D. melanogaster* experimentaron dos procesos de colonización con el fin de establecerse, primero al salir de África y después en la llegada al continente americano. En el caso de los individuos de Estados Unidos la población elegida fue la de Raleigh (Carolina del Norte)[20] donde las secuencias se habían obtenido mediante línea endogámica.

Cada una de estas poblaciones cuenta con un gran número de individuos secuenciado. Para ello, se buscaron los individuos con las secuencias más completas en los tres fragmentos y se escogieron al azar 15 de esos individuos. El listado de los individuos se indica en el anexo 1.

2.3 Análisis de secuencias

Las secuencias se han alineado mediante el programa MEGA-X usando el algoritmo MUSCLE[21]. Se ha estimado la diversidad nucleotídica (π[22]) y varios indicadores estadísticos (*D* de Tajima[23] y *H* de Fay y Wu[24]) con el programa DnaSP[13]. Ésta es una herramienta popular para realizar análisis genéticos poblacionales exhaustivos en alineaciones de múltiples secuencias. Tanto *D* como *H* se utilizan para detectar desviaciones de escenarios de neutralidad. Mientras que *D* es más eficaz rechazando la neutralidad cuando hay un exceso de variantes a baja frecuencia, *H* muestra mayor certeza rechazando la neutralidad cuando hay un exceso de variantes a alta frecuencia. Una diferencia importante para el cálculo de estos indicadores es que *D* utiliza solo los datos del polimorfismo, mientras que para calcular *H* se necesita además la secuencia de otra especie que permita clasificar las variantes como ancestrales o derivadas. Para ello se ha utilizado la secuencia de *D. simulans*.

El fragmento 0-20 se ha partido en varios fragmentos más pequeños, ya que la secuencia es larga y de esta manera se puede analizar como varían los indicadores estadísticos que se obtendrán teniendo en cuenta

la distancia a los dos genes alojados en el fragmento. Al realizar el análisis para el indicador estadístico *H* en sliding windows usando ventanas de mil no solapadas se han definido las coordenadas de cada subfragmento. De esta manera, se logra que los subfragmentos tengan más de mil bases ya que los subfragmentos no se encuentran limitados por las coordenadas del alineamiento, sino por las posiciones analizadas. Esto se debe a que, al alinear las dos especies, *D. melanogaster* y *D. simulans,* deben introducirse algunos InDels. Estas coordenadas se han utilizado para obtener los fragmentos en los que estudiar los otros indicadores (*D* y π).

La significación estadística ha sido obtenida mediante simulaciones utilizando el software mlcoalsim[14]. Estas simulaciones se han realizado con los datos de las posiciones segregantes y recombinación. El parámetro de recombinación de la población (R= 4Nr) ha sido estimado por la tasa de recombinación del gen *phantom* (r= 2,27 X 10^{-8} tasa de recombinación/pares de bases/generación[25]) en el fragmento 0 y la población efectiva de *D. melanogaster* (N= 10^6) que se ha utilizado en el trabajo de Orengo y Aguadé[2]. Se ha utilizado el parámetro no_rec_males = 1 puesto que los machos de Drosophila no recombinan y también el parámetro factorn_chr = 0,75 por tratarse de secuencias del cromosoma X. Para contrarrestar el problema de las comparaciones múltiples se aplica la corrección de Bonferroni para cada conjunto de tests de una misma población[26].

Para comprobar si existe diferenciación genética entre las poblaciones del área de origen de la especie se ha realizado un análisis de diferenciación genética también con la herramienta DnaSP. Para ello se ha analizado el estimador estadístico *Snn de Hudson*[27]. Se ha obtenido la significación por el método de las permutaciones tanto para el conjunto de las 3 poblaciones como para cada pareja de poblaciones.

3. Resultados y Discusión

Se obtienen los siguientes resultados.

3.1 Variación nucleotídica

Como indicadores de la variación nucleotídica de los distintos fragmentos y poblaciones analizadas se ha usado el número de posiciones segregantes y la diversidad nucleotídica. Para testar si esta variación nucleotídica se ajusta a un escenario de neutralidad se han utilizado dos estadísticos distintos, la *D* de Tajima[23] y la *H* de Fay y Wu[24]. Valores significativos de estos estadísticos permiten rechazar la neutralidad. *D* lo rechaza cuando hay un exceso de variantes en frecuencia baja y *H* cuando esto sucede a alta frecuencia. Es posible, que uno de los dos rechace la neutralidad aunque el otro no sea capaz

La tabla 2 muestra los resultados para los polimorfismos de los fragmentos 0-20, 22 y 32 de las 4 poblaciones analizadas.

Tabla 2: Polimorfismo de los fragmentos analizados

Población	Fragmento	Posición	*S*	*π*	*D*	*p-D*	*H*	*p-H*
Zambia	0-20	1-20992	528	0,00664	-1,02	0,001*	-6,057	>0,05
Zambia	22	22808-23570	18	0,00583	-1	>0,05	0,404	>0,05
Zambia	32	32787-33631	26	0,00828	-0,644	>0,05	-0,573	>0,05
Ruanda	0-20	1-20998	479	0,00671	-0,654	0,001*	-8,990	>0,05
Ruanda	22	22808-23570	8	0,00274	-0,889	>0,05	-0,323	>0,05
Ruanda	32	32787-33631	21	0,00665	-0,844	>0,05	-1,109	>0,05
Etiopia	0-20	1-20953	251	0,00404	0,0239	>0,05	-22,65	0,005*
Etiopia	22	22808-23570	7	0,00267	-0,351	>0,05	0,296	>0,05
Etiopia	32	32787-33631	6	0,00321	1,522	0,011*	0,630	>0,05
EE. UU.	0-20	1-20953	89	0,00134	-0,128	>0,05	-15,21	0,001*
EE. UU.	22	22808-23570	4	0,00103	-1,220	>0,05	0,647	>0,05
EE. UU.	32	32787-33631	2	0,00106	1,117	>0,05	0,493	>0,05

Nota. Población, población a la que pertenece la secuencia; Fragmento, posición relativa al fragmento 0 en kilobases; Posición, Nucleotidos

exactos de la secuencia analizada; *S*, numero de sitios segregantes; *π*, diversidad nucleotídica; *D*, estadístico *D* de Tajima; *p-D*, valor *p* del *D* de Tajima; *H*, indicador estadísticos *H* de Fay y Wu; *p-H*, valor *p* del estadístico *H*. * indica que mantiene la significación al 0,05 tras realizar la corrección de Bonferroni.

En el caso de Zambia y Ruanda se observan resultados parecidos. En el fragmento 0-20 (tabla 2) se halla un valor de *D* negativo y significativo indicando que en este fragmento se halla un exceso de variantes a baja frecuencia que no se ajusta a un escenario de neutralidad. Esta situación puede deberse a varios motivos tales como una expansión tras un reciente cuello de botella, selección natural… En los fragmentos 22 y 32 no se encuentran ningún tipo de rastro para rechazar la neutralidad.

En la población africana restante, es decir, Etiopia, la neutralidad también es rechazada en el fragmento 0-20 (tabla 2) pero en este caso debido a un exceso de variantes a alta frecuencia, ya que *H* resulta ser negativa y significativa. En el fragmento 22 se asume la neutralidad y en el fragmento 32 se rechaza debido a una falta de alelos a baja frecuencia en el fragmento con un valor de *D* positivo y significativo (p= 0, 011).

En la población estadounidense, al igual que en la de Etiopia, se rechaza la neutralidad en el fragmento 0-20 (Tabla 2) por un exceso de variantes derivadas a alta frecuencia con una *H* negativa y significativa (p= 0,001). En los fragmentos 22 y 32 no se percibe que los fragmentos se alejen de la neutralidad.

A continuación, se presenta la tabla 3 que muestra los resultados para los polimorfismos de los subfragmentos del fragmento 0-20.

Tabla 3: Polimorfismo de los subfragmentos analizados del fragmento 0-20 de las 4 poblaciones

Población	Sub	Posición	*S*	*π*	*D*	*p-D*	*H*	*p-H*
Zambia	1	1-1242	31	0,00676	-1,158	0,013	0,543	>0,05
Zambia	2	1243-2325	10	0,00185	-1,519	0,002*	1,161	>0,05
Zambia	3	2326-3400	25	0,00688	-0,439	>0,05	3,181	>0,05
Zambia	4	3401-4545	36	0,0082	-1,025	0,033	0,133	>0,05
Zambia	5	4546-5655	21	0,00452	-1,128	0,014	2,324	>0,05
Zambia	6	5656-6999	31	0,00711	-0,594	>0,05	0,848	>0,05
Zambia	7	7000-8066	23	0,00432	-1,586	0,002*	2,333	>0,05

Población	Sub	Posición	*S*	*π*	*D*	*p-D*	*H*	*p-H*
Zambia	8	8067-9289	32	0,00624	-1,280	0,008	1,095	>0,05
Zambia	9	9290-10573	48	0,01049	-0,676	>0,05	-3,486	>0,05
Zambia	10	10574-11708	23	0,00536	-1,076	0,018	-0,343	>0,05
Zambia	11	11709-12909	43	0,01099	-0,756	>0,05	-0,361	>0,05
Zambia	12	12910-14036	22	0,00499	-1,071	0,024	2,038	>0,05
Zambia	13	14037-15084	32	0,00732	-1,149	0,009	-3,009	>0,05
Zambia	14	15085-16267	45	0,01006	-1,139	0,006	-2,847	>0,05
Zambia	15	16268-17377	31	0,00748	-0,888	0,033	-1,409	>0,05
Zambia	16	17378-18400	18	0,00476	-0,567	>0,05	2,095	>0,05
Zambia	17	18401-19409	21	0,00555	-0,74	>0,05	-1,686	>0,05
Zambia	18	19410-20528	21	0,00555	-0,676	>0,05	3,571	>0,05
Zambia	19	20529-20992	15	0,00684	-1,305	0,026	-0,124	>0,05
Ruanda	1	1-1277	32	0,00797	-0,651	>0,05	4,8	0,021
Ruanda	2	1278-2348	13	0,00263	-1,347	0,015	1,638	>0,05
Ruanda	3	2349-3428	31	0,00962	0,038	>0,05	4,543	0,024
Ruanda	4	3429-4628	22	0,00581	-0,525	>0,05	-1,533	>0,05
Ruanda	5	4629-6020	19	0,00303	-1,685	0,001*	2	>0,05
Ruanda	6	6021-7131	41	0,01056	-0,477	>0,05	0,838	>0,05
Ruanda	7	7132-8225	15	0,00438	-0,102	>0,05	-2,686	>0,05
Ruanda	8	8226-9472	23	0,00602	-0,327	>0,05	3,733	0,023
Ruanda	9	9473-10786	45	0,0092	-0,901	0,03	-6,809	0,05
Ruanda	10	10787-11975	19	0,0047	-0,755	>0,05	1,809	>0,05
Ruanda	11	11976-13122	37	0,00977	-0,459	>0,05	-1,819	>0,05
Ruanda	12	13123-14257	18	0,00386	-1,224	0,009	0,686	>0,05

Población	Sub	Posición	*S*	*π*	*D*	*p-D*	*H*	*p-H*
Ruanda	13	14258-15407	49	0,01233	-0,800	0,044	-5,705	>0,05
Ruanda	14	15408-16508	31	0,0083	-0,635	>0,05	1,267	>0,05
Ruanda	15	16509-17599	28	0,00772	-0,527	>0,05	3,924	0,04
Ruanda	16	17600-18627	24	0,00712	-0,145	>0,05	0,105	>0,05
Ruanda	17	18628-19638	10	0,00315	0,107	>0,05	1,486	>0,05
Ruanda	18	19639-20759	16	0,00445	-0,500	>0,05	-0,257	>0,05
Ruanda	19	20760-20998	6	0,00598	-0,971	>0,05	0,971	>0,05
Etiopia	1	1-1232	6	0,00199	0,474	>0,05	0,352	>0,05
Etiopia	2	1233-2288	5	0,00107	-1,033	>0,05	0,495	>0,05
Etiopia	3	2289-3333	9	0,00336	0,817	>0,05	-2,038	>0,05
Etiopia	4	3334-4411	9	0,00337	0,869	>0,05	1,104	>0,05
Etiopia	5	4412-5512	3	0,00092	0,096	>0,05	0,257	>0,05
Etiopia	6	5513-6758	9	0,00263	0,327	>0,05	-1,5524	>0,05
Etiopia	7	6759-7838	12	0,00356	0,045	>0,05	-1,0857	>0,05
Etiopia	8	7844-9052	16	0,00568	1,037	0,035	0,9524	>0,05
Etiopia	9	9053-10166	11	0,00258	-0,859	>0,05	-6,1714	0,004
Etiopia	10	10167-11445	22	0,00385	-1,408	0,001*	-8	0,003
Etiopia	11	11446-12622	18	0,00719	1,430	0,004	-1,7905	>0,05
Etiopia	12	12623-13696	10	0,00339	0,389	>0,05	0,781	>0,05
Etiopia	13	13697-14770	15	0,00318	-1,207	0,019	-5,114	0,012
Etiopia	14	14771-15959	45	0,01395	-0,005	>0,05	-3,191	>0,05
Etiopia	15	15960-17028	22	0,00526	-0,885	0,037	0,82	>0,05
Etiopia	16	17029-18056	13	0,0044	0,395	>0,05	0,371	>0,05
Etiopia	17	18057-19060	8	0,00264	0,280	>0,05	0,438	>0,05

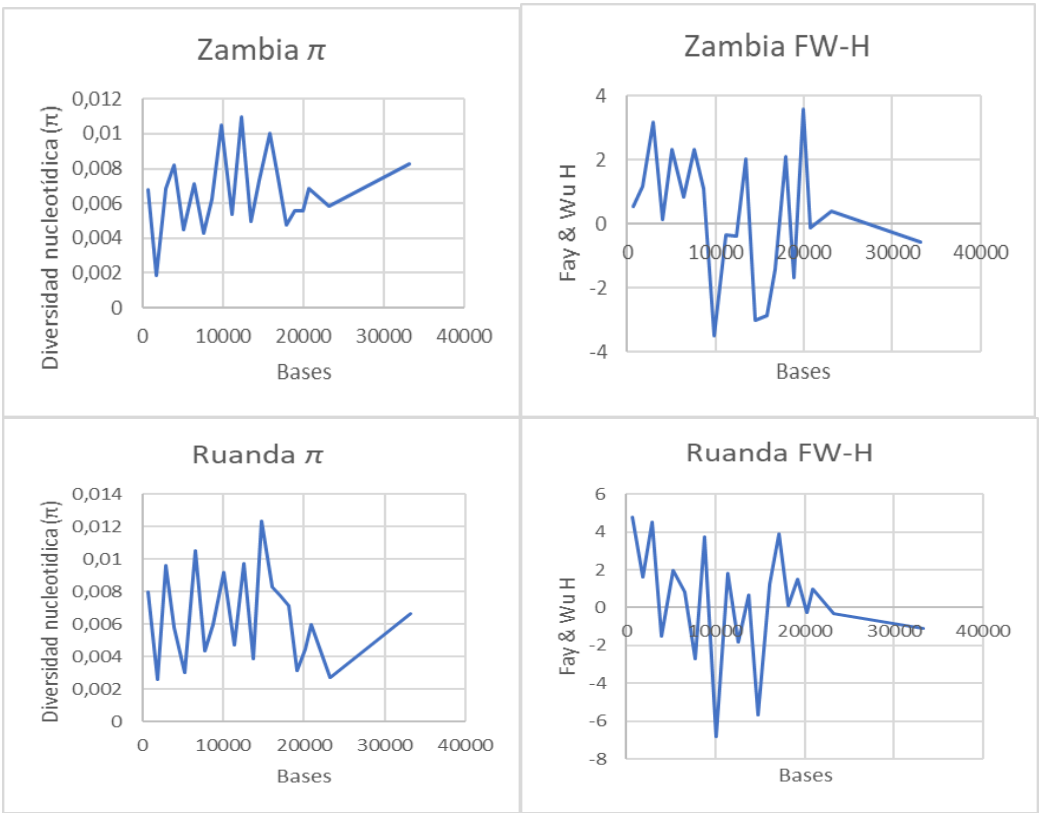
Población	Sub	Posición	*S*	*π*	*D*	*p-D*	*H*	*p-H*
Etiopia	18	19061-20194	11	0,00444	1,223	0,027	1,076	>0,05
Etiopia	19	20195-20953	7	0,00279	-0,319	>0,05	-0,362	>0,05
EE. UU	1	1-1163	8	0,00242	0,1092	>0,05	-1,219	>0,05
EE. UU	2	1164-2209	5	0,00185	0,722	>0,05	-2,533	0,023
EE. UU	3	2210-3234	13	0,00413	0,152	>0,05	-1,124	>0,05
EE. UU	4	3235-4305	7	0,00149	-1,084	>0,05	-4,686	0,005
EE. UU	5	4306-5430	7	0,00254	0,796	>0,05	-2,238	0,049
EE. UU	6	5431-6657	12	0,00384	0,688	>0,05	0,847	>0,05
EE. UU	7	6658-7752	13	0,00451	0,826	>0,05	-4	0,029
EE. UU	8	7753-8858	2	0,00044	-0,594	>0,05	0,381	>0,05
EE. UU	9	8859-10008	1	0,00013	-1,159	>0,05	0,124	>0,05
EE. UU	10	10009-11248	1	0,00011	-1,159	>0,05	0	
EE. UU	11	11249-12366	5	0,00091	-1,284	0,049	-1,095	>0,05
EE. UU	12	12367-13406	4	0,00084	-0,972	>0,05	0,714	>0,05
EE. UU	13	13408-14463	0	0	n.d.		0	
EE. UU	14	14464-15649	2	0,00026	-1,490	>0,05	0,123	>0,05
EE. UU	15	15650-16666	3	0,0004	-1,685	0,034	-1,486	>0,05
EE. UU	16	16667-17715	5	0,00132	-0,406	>0,05	0,848	>0,05
EE. UU	17	17716-18727	0	0	n.d.		0	
EE. UU	18	18728-19749	0	0	n.d.		0	
EE. UU	19	19751-20854	1	0,00013	-1,159	>0,05	0,124	>0,05
EE. UU	20	20855-20953	0	0	n.d.		0	

Nota. Población, población a la que pertenece la secuencia; Sub, Numero de identificación del subfragmento; Posición, nucleótidos exactos de la secuencia analizada; *S*, numero de sitios segregantes; *π*,

diversidad nucleotídica; *D*, estadístico *D* de Tajima; *p-D*, valor *p* del *D* de Tajima; *H*, indicador estadísticos *H* de Fay y Wu; *p-H*, valor *p* del estadístico *H*. * indica que mantiene la significación al 0,05 tras realizar la corrección de Bonferroni. n.d., no determinado.

En varios de los subfragmentos del fragmento 0-20 (tabla 3) se halla también un valor de *D* significativo. Mediante la corrección de Bonferroni solo se puede asumir esta significación en dos casos: fragmentos 2 y 7 en Zambia y Fragmento 4 en Ruanda. De igual manera, la corrección de Bonferroni niega la significación de los únicos subfragmentos (1, 3, 8, 9 y 15) que muestran *H* significativa en la población de Ruanda. Los subfragmentos en los que se rechaza la neutralidad están ubicados al inicio de la secuencia lejos de los genes *phantom* y *Cyp18a1* que son los genes seleccionados en Europa.

En el caso de *π,* los gráficos (figura 4) de las dos poblaciones los perfiles muestran gran semejanza y en el fragmento 0-20 la diversidad nucleotídica es muy similar (Zambia *π*= 0,00664, Ruanda *π*= 0,00671). No se percibe ninguna región de la secuencia que posea una reducción drástica de la diversidad nucleotídica que pueda significar que haya algún efecto que rompa el equilibrio. En la posición donde se ubican los genes *phantom* y *Cyp18a1* y sus alrededores no se ve una reducción de la variación nucleotídica. Además, los subfragmentos en los cuales se ha rechazado la neutralidad están ubicados en la parte izquierda de la secuencia completa y a cierta distancia de los genes. Esto demuestra que, dicho rechazo de la neutralidad podría deberse a genes ubicados cerca de esta región.



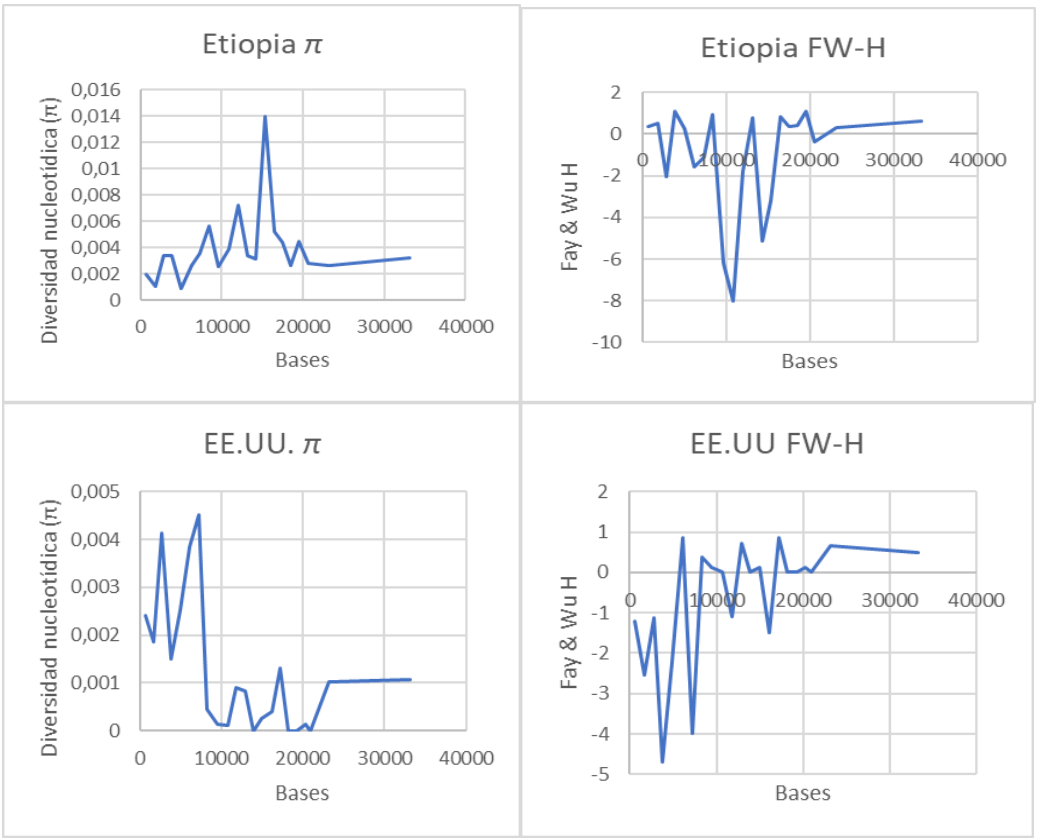


Figura 4: Gráficos de la distribución de π y *H* en las secuencias de las 4 poblaciones

En el caso de Etiopia, se ven 3 subfragmentos (9, 10 y 13) (Tabla 3) con valores de *H* significativos pero rechazados por la corrección de Bonferroni. En el caso de la *D* no hay significación global en el fragmento 0-20, sin embargo, hay varios subfragmentos con valores positivos y negativos (8, 10, 11, 13, 15 y 18). Después, la significación estadística de estas secuencias es rechazada por la aplicación de Bonferroni menos en el caso del subfragmento 10, ya que muestra un exceso de variantes a baja frecuencia. En este caso los subfragmentos con *p* significativa pero rechazados por la aplicación de la corrección de Bonferroni están ubicados en los entornos de los dos genes.

En el gráfico de Etiopia de π se observa un perfil diferente a los gráficos de las otras dos poblaciones africanas. Esto se debe a que en este caso π (fragmento 0-20 π= 0,00404) es algo menor que en las otras dos poblaciones generando máximos menores. No se observa tampoco ningún segmento en la secuencia con diversidad nucleotídica muy reducida. Por lo que, hay un efecto que rechaza la neutralidad en la región 0-20 y posiblemente la región causante de este efecto esté fuera de esta región genómica, como en el caso de las otras dos poblaciones africanas.

En lo referente a los subfragmentos en la población estadounidense (Tabla 3), en 4 (2, 4, 5 y 7) se obtiene un valor *H* significativo, aunque tras la aplicación de la corrección de Bonferroni no se mantiene su significación. En los demás subfragmentos no se puede apreciar nada, ya que, al haber una diversidad nucleotídica tan baja, las pruebas pierden potencia. Esta región de baja variabilidad es donde se ubican los genes y sus alrededores.

En cuanto al gráfico de π se perciben dos cosas.
Por una parte, la variabilidad nucleotídica es mucho menor que en las poblaciones africanas (por debajo de 0,002 la mayor parte del tiempo), y esto puede deberse a las expansiones de las poblaciones generadas después del cuello de botella que debió experimentar al colonizar América. En la población de Raleigh, la π global en el cromosoma X de *D. melanogaster* es igual a 0,00393, con lo cual además del efecto de la expansión, es probable que en esta región donde π es más baja que la media, haya otro efecto que reduzca la variabilidad nucleotídica.[28]

Por otra parte, se aprecia una región en la secuencia en la cual la diversidad nucleotídica es extremadamente baja. Esta región de baja diversidad nucleotídica es de unos ~12 kb mínimo y podría a llegar a ser de unos ~25 kb. Podría decirse la longitud exacta de la secuencia si se estudiase la secuencia que hay en los fragmento 22 y 32. La meritada región se menciona en el trabajo de Orengo y Aguadé[9] (figura 5) en una población de Cataluña y se confirmó mediante diferentes pruebas que esta región era el foco del efecto de la selección natural. Dentro de ella están los genes *phantom* y *Cyp18a1* y podrían ser los causantes de que en la población estadounidense se rechace la neutralidad en la secuencia analizada.

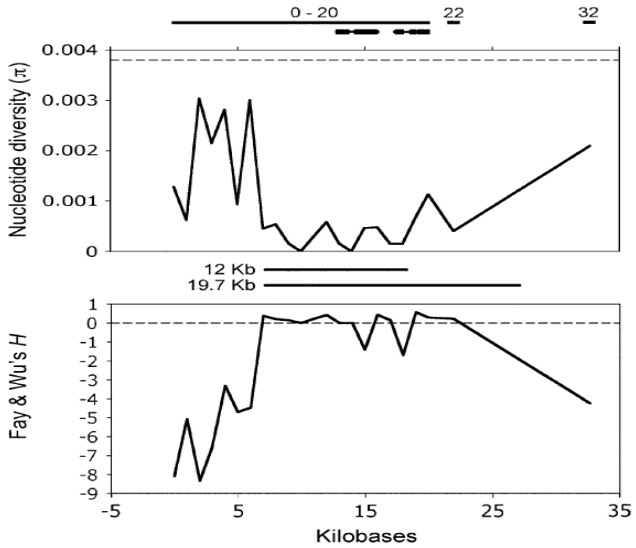


Figura 5: Diversidad nucleotídica e indicador estadístico *H* de Fay-Wu de una población catalana [9]

Sería necesaria la realización de pruebas complementarias para asegurar que la huella proviene de esta región de π baja. También se debería tener en cuenta que la muestra analizada es de 15 individuos de

una población de un área geográfica extensa por lo que para comprobar si en esta región se hallan huellas de la selección natural también se deberían de realizar análisis con más individuos o analizar distintas poblaciones.

3.2. Diferenciación genética entre las poblaciones africanas

La tabla 4 muestra los resultados para diferenciación genética entre las poblaciones africanas.

Tabla 4: Diferenciación genética entre las poblaciones africanas

Fragmento	Poblaciones	*Snn*	*p*
0-20	Etiopia-Ruanda-Zambia	0,66667	<0,001
0-20	Etiopia-Ruanda	0,9	<0,001
0-20	Etiopia-Zambia	0,9	<0,001
0-20	Ruanda-Zambia	0,63333	0,068
22	Etiopia-Ruanda-Zambia	0,50713	<0,001
22	Etiopia-Ruanda	0,65746	0,005
22	Etiopia-Zambia	0,65833	0,008
22	Ruanda-Zambia	0,6027	0,089
32	Etiopia-Ruanda-Zambia	0,6795	<0,001
32	Etiopia-Ruanda	0,96667	<0,001
32	Etiopia-Zambia	0,84206	<0,001
32	Ruanda-Zambia	0,58889	0,113

Nota. Fragmento, el fragmento al que corresponde el análisis; Poblaciones, Las poblaciones que se compara; *Snn, Snn* de Hudson; *p*, probabilidad.

Al analizar la diferenciación genética entre las tres poblaciones ancestrales mediante *Snn* se rechaza la hipótesis nula de igualdad en los tres fragmentos (tabla 4). Sin embargo, al analizar las poblaciones de dos en dos, se ha descubierto que no todas las poblaciones se han diferenciado entre ellas. Mientras que Etiopia se ha diferenciado de las otras dos. No obstante, Zambia y Ruanda no se han diferenciado genéticamente, a pesar de que el valor *Snn* esté más cerca de 1 que de 0 en los tres fragmentos, lo que sugiere una diferencia genética. En el caso de Etiopia, el fragmento 22 se ubica genéticamente más cerca de las otras dos poblaciones con valores de Etiopia-Ruanda *Snn*= 0,65746 y Etiopia-Zambia *Snn*= 0,65833 aunque los valores *p* están muy lejos de no ser significativos.

La población de Ruanda y la de Etiopía están a 1741,52 km de distancia y la de Ruanda y Zambia a 1563,95 km. las diferencias observadas entre estos pares de poblaciones no parecen responder a un mero efecto de distancia geográfica, debido a que las distancias son similares.

Esta diferenciación observada puede ser debido a la ubicación geográfica de la población etíope ya que está ubicada en el macizo etíope, una región montañosa. Dicha población etíope está ubicada en la localidad de Fiche, a más de 3000m de altura. Esta “barrera natural” puede ser el motivo por el cual existe esta diferenciación genética ya que evita el flujo génico entre ambas poblaciones.

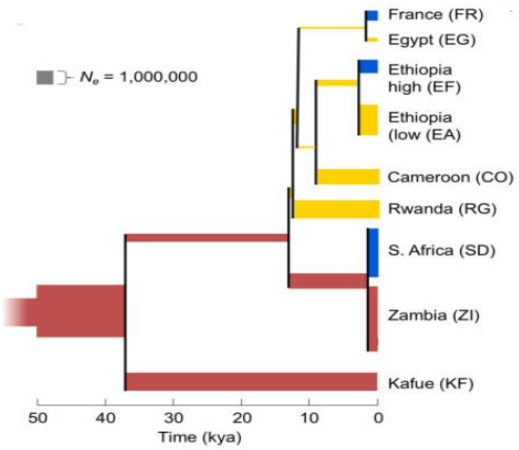


Figura 6: Historia estimada de unas poblaciones analizadas en la región ancestral [29]

La historia demográfica de *D. melanogaster* en África-subsahariana se estimó tras analizar varias poblaciones africanas distintas[29]. En la figura 6 se puede observar que las poblaciones de Zambia y Ruanda se separaron hace ~ 12887 años (11091-17113 años) y que Camerún se separó de Ruanda también alrededor de esos años (120975-16243 años). Etiopia se separó de Camerún hace 9571 años (8916-10547 años. La separación de las tres poblaciones hace aproximadamente el mismo tiempo, refuerza la explicación de que la diferenciación genética existe debido a la región montañosa en la que se ubica la población etíope.

4. Conclusiones

El trabajo realizado permite concluir lo siguiente:

- Al analizar la secuencia del fragmento 0-20 para las tres poblaciones africanas se rechaza la neutralidad en la región genómica analizada (En Ruanda y Zambia por el estadístico *D* y en Etiopia por el estadístico *H*). No obstante, este rechazo podría deberse al efecto que genera otra región cercana, ya que al analizar la región por ventanas de 1000 nucleótidos, no hay indicios de selección natural en esta misma región.

- En la población estadounidense se observa una diversidad nucleotídica extremadamente baja que hace que los estadísticos usados no tengan potencia, sobre todo en el estudio de los subfragmentos. A pesar de ello, hay diversos indicios de que pueda haber estado sometido a selección. Por un lado, la diversidad nucleotídica es inferior a la observada en un estudio a nivel genómico del cromosoma X. Asimismo, se observa una zona con una reducción en la diversidad nucleotídica muy acusada en la misma zona en la que se halló en Europa. Por otro lado, Se obtiene un valor *H* negativo y altamente significativo en la región de 20 kb que indica la acción de la selección natural.

- Al analizar las tres poblaciones africanas conjuntamente se observa que hay diferenciación genética entre ellas. No obstante, al comparar las poblaciones de dos en dos se observa que solo la población de Etiopia se ha diferenciado. Esto no se debe únicamente al efecto distancia, ya que las poblaciones de Zambia y Ruanda se encuentran a una distancia similar a las de Ruanda y Etiopia. Es posible que la población etíope haya quedado más aislada de las otras dos debido a la orografía y que las condiciones climáticas a las que esté sometida sean algo distintas, favoreciendo que haya evolucionado independientemente

Al principio del semestre se plantearon unos objetivos generales, los cuales se han ido concretando a medida que avanzaba el curso. Desde el comienzo se trabajó para intentar llegar a los objetivos establecidos y todos ellos han sido alcanzados, ya que, aunque en el caso de la población estadounidense no se haya confirmado del todo la existencia de la huella de la selección natural se han hallado grandes indicios de que pueda llegar a existir.

La planificación y la metodología fueron establecidos al comienzo del curso. Debido al desconocimiento de las herramientas necesarias para el desarrollo del proyecto se fijaron las fechas con cierto margen para

iniciarse en las herramientas. Se cometió un error en una de las tareas en la cual se utilizó una herramienta informática, pero el error se pudo corregir gracias a los días planificados para la corrección de errores.

Hay varias líneas posibles de trabajo futuro que no se han explorado. Sería interesante, por un lado, estudiar las inmediaciones de la región analizada en el caso de las 3 poblaciones africanas con el fin de definir la procedencia de ese rechazo de la neutralidad en la región analizada. Y Por otro lado, realizar diferentes pruebas para confirmar la selección natural en la población estadounidense. La bibliografía indica que en esta misma región se hallaron huellas de la selección natural en una población catalana, así como patrones de diversidad nucleotídica similares a los obtenidos en la población estadounidense. Sin embargo, en este caso se realizaron más pruebas para obtener mayor certeza. También sería interesante realizar análisis de diferentes poblaciones dentro del continente americano.

Lamentablemente, debido a la exigencia de estudio que requieren todas las cuestiones recientemente planteadas, no se ha podido profundizar en el análisis de las mismas.

5. Glosario

BLAST: Basic Local Alignment Search Tool. Herramienta informatica que busca regiones parecidas entre secuencias biológicas.

D: Estadístico *D* de Tajima que se utiliza para distinguir si una secuencia evoluciona al azar o no.

DnaSP: DNA Sequence Polymorphism. Herramienta bioinformatica que se utiliza para varios analisis con secuencia de ADN

H: Estadístico *H* de Fay y Wu que sirve para distinguir entre una secuencia que evoluciona al azar o que evoluciona bajo el efecto de la selección positiva.

InDel: Contracción de “inserción o deleción”, dos tipos de mutaciones genéticas.

MEGA: Molecular Evolutionary Genetycs Analysis. Herramienta para el análisis de ADN y proteínas.

Snn: test Estadístico que sirve para detectar diferenciaciones genéticas.

π: Diversidad de nucleótidos. Número promedio de diferencias por nucleótido.

6. Bibliografía

1. Gillespie, J. H. The causes of molecular evolution. *Oxford Univ. Press* (1991).

2. Orengo, D. J. & Aguadé, M. Detecting the footprint of positive selection in a European population of *Drosophila melanogaster*: Multilocus pattern of variation and distance to coding regions. *Genetics* **167**, 1759–1766 (2004).

3. Lachaise, D. *et al.* Historical biogeography of the *Drosophila melanogaster* species subgroup. *Evol. Biol.* 159–225 (1988).

4. Begun, D. J. & Aquadro, C. F. African and North American populations of *Drosophila melanogaster* are very different at the DNA level. *Nature* **365**, 548–550 (1993).

5. Andolfatto, P. Contrasting patterns of X-linked and autosomal nucleotide variation in *Drosophila melanogaster* and *Drosophila simulans*. *Mol. Biol. Evol.* **18**, 279–290 (2001).

6. Andolfatto, P. & Przeworski, M. A genome-wide departure from neutrality from the standard neutral populations of *Drosophila*. *Genetics* **156**, 257–268 (2000).

7. Przeworski, M., Wall, J. D. & Andolfatto, P. Recombination and the frequency spectrum in *Drosophila melanogaster* and *Drosophila simulans*. *Mol. Biol. Evol.* **18**, 291–298 (2001).

8. Glinka, S., Ometto, L., Mousset, S., Stephan, W. & De Lorenzo, D. Demography and natural selection have shaped genetic variation in *Drosophila melanogaster*: A multi-locus approach. *Genetics* **165**, 1269–1278 (2003).

9. Orengo, D. J. & Aguadé, M. Genome scans of variation and adaptive change: Extended analysis of a candidate locus close to the *phantom* gene region in *Drosophila melanogaster*. *Mol. Biol. Evol.* **24**, 1122–1129 (2007).

10. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).

11. Drosophila Genome Nexus. https://www.johnpool.net/genomes.html. 10/2020

12. Hervas, S., Sanz, E., Casillas, S., Pool, J. E. & Barbadilla, A. PopFly: the *Drosophila* population genomics browser. *Bioinformatics* **33**, 2779–2780 (2017).

13. Rozas, J. *et al.* DnaSP 6: DNA sequence polymorphism analysis of large data sets. *Mol. Biol. Evol.* **34**, 3299–3302 (2017).

14. Ramos-Onsins, S. E. & Mitchell-Olds, T. Mlcoalsim: Multilocus coalescent simulations. *Evol. Bioinforma.* **3**, 41–44 (2007).

15. Blast. http://flybase.org/blast/. 10/2020

16. popfly. https://popfly.uab.cat/. 10/2020

17. Ensembl Metazoa. http://metazoa.ensembl.org/index.html. 10/2020

18. Pool, J. E. *et al.* Population Genomics of Sub-Saharan *Drosophila melanogaster*: African Diversity and Non-African Admixture. *PLoS Genet.* **8**, (2012).

19. Lack, J. B. *et al.* The drosophila genome nexus: A population genomic resource of 623 *Drosophila melanogaster genomes*, including 197 from a single ancestral range population. *Genetics* **199**, 1229–1241 (2015).

20. Mackay, T. F. C. *et al.* The *Drosophila melanogaster* Genetic Reference Panel. *Nature* **482**, 173–178 (2012).

21. Kumar, S., Stecher, G., Li, M., Knyaz, C. & Tamura, K. MEGA X: Molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.* **35**, 1547–1549 (2018).

22. Nei, M. *Molecular evolutionary genetics*. (New York: Columbia University Press, 1987).

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

24. Fay, J. C. & Wu, C. I. Hitchhiking under positive Darwinian selection. *Genetics* **155**, 1405–1413 (2000).

25. Hey, J. & Kliman, R. M. Interactions between natural selection, recombination and gene density in the genes of *Drosophila*. *Genetics* **160**, 595–608 (2002).

26. Bonferroni, C. Teoria statistica delle classi e calcolo delle probabilita. *Pubbl. del R Ist. Super. di Sci. Econ. e Commericiali di Firenze* **8**, 3–62 (1936).

27. Hudson, R. R. A new statistic for detecting genetic differentiation. *Genetics* **155**, 2011–2014 (2000).

28. Campo, D. *et al.* Whole-genome sequencing of two North American *Drosophila melanogaster* populations reveals genetic differentiation and positive selection. *Mol. Ecol.* **22**, 5084–5097 (2013).

29. Sprengelmeyer, Q. D. *et al.* Recurrent Collection of *Drosophila melanogaster* from Wild African Environments and Genomic Insights into Species History. *Mol. Biol. Evol.* **37**, 627–638 (2020).

7. Anexos

Anexo 1: Listado de los individuos analizados

Zambia	Ruanda	Etiopia	EE. UU.
ZI176	RG10	EF103N	RAL-31
ZI197N	RG11N	EF115N	RAL-360
ZI207	RG13N	EF11N	RAL-382
ZI239	RG18N	EF130N	RAL-391
ZI250	RG21N	EF15N	RAL-530
ZI26	RG22	EF24N	RAL-559
ZI320	RG24	EF32N	RAL-589
ZI341	RG2	EF39N	RAL-630
ZI368	RG32N	EF46N	RAL-774
ZI370	RG33	EF64N	RAL-802
ZI384	RG34	EF75N	RAL-837
ZI402	RG36	EF81N	RAL-843
ZI429	RG39	EF83N	RAL-850
ZI443	RG8	EF98N	RAL-853
ZI446	RG9	EF9N	RAL-913

Anexo 2: Archivo de entrada para la simulación de los fragmentos 0-20, 22 y 32 de la población estadounidense.

```
seed1 7354
print_matrixpol 0
print_neuttest 2

n_iterations 1000
n_loci 3
n_sites 19894 737 830
n_samples 15 15 15
npop 1

recombination 1806.3752 66.9196 75.364
mutations 89 4 2
factorn_chr 0.75 0.75 0.75
no_rec_males 1
likelihood_line 0
```

Anexo 3: Archivo de entrada para la simulación de los subfragmentos del fragmento 0-20 de la población estadounidense.

```
seed1 7354
print_matrixpol 0
print_neuttest 2

n_iterations 1000
n_loci 20
n_sites 1048 1007 1005 1013 1033 1130 1072 1072 1021 1221 1041 1026 1006 1039 1005 1024 1002 1004 1035 90
n_samples 15 15 15 15 15 15 15 15 15 15 15 15 15 15 15 15 15 15 15 15
npop 1

recombination 95.1584 91.4356 91.254 91.9804 93.7964 102.604 97.3376 97.3376 92.7068 110.8668 94.5228 93.1608 91.3448 94.3412 91.254 92.9792 90.9816 91.1632 93.978 8.172
mutations 8 5 13 7 7 12 13 2 1 1 5 4 0 2 3 5 0 0 1 0
factorn_chr 0.75 0.75 0.75 0.75 0.75 0.75 0.75 0.75 0.75 0.75 0.75 0.75 0.75 0.75 0.75 0.75 0.75 0.75 0.75 0.75
no_rec_males 1
likelihood_line 0
```