Caracterización de inversiones cromosómicas en *Drosophila subobscura*

Kenia Mª Delgado Pérez

MU Bioinf. y Bioest.
Análisis de datos ómicos

Nombre Tutor/a de TF
Dorcas Orengo Ferriz

Fecha Entrega
Junio 2023

Universitat Oberta de Catalunya

UNIVERSITAT DE BARCELONA

Esta obra está sujeta a una licencia de Reconocimiento-NoComercial-SinObraDerivada 3.0 España de Creative Commons

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Caracterización de inversiones cromosómicas en Drosophila subobscura*
Nombre del autor:	*Kenia Mª Delgado Pérez*
Nombre del consultor/a:	*Dorcas Orengo Ferriz*
Nombre del PRA:	*David Merino Arranz*
Fecha de entrega (mm/aaaa):	*06-2023*
Titulación o programa:	*Máster Universitario en Bioinformática y Bioestadística UOC-UB*
Área del Trabajo Final:	*Análisis de datos ómicos*
Idioma del trabajo:	*Castellano*
Palabras clave	*Inversiones cromosómicas, Drosophila subobscura, puntos de rotura.*

Resumen del Trabajo

Las inversiones cromosómicas son variaciones estructurales producidas por la rotura de un fragmento cromosómico y su reinserción en orientación invertida. Drosophila ha sido el género más estudiado, siendo *D. subobscura* particularmente interesante por su abundante polimorfismo por inversión. Se ha relacionado las inversiones cromosómicas con el carácter adaptativo de la especie.

El objetivo de este estudio fue la caracterización de inversiones cromosómicas por mapeo de *reads paired-end* de Illumina sobre el genoma de referencia. Para ello, se creó un protocolo bioinformático que permitiera obtener los puntos de rotura de la inversión J_1, de especial interés por su carácter adaptativo, y los puntos de rotura de la inversión U_6.

Se caracterizaron los dos puntos de rotura de la inversión U_6 y los resultados revelaron que la inversión fue originada por NHEJ y rotura escalonada. El estudio de los genes codificantes flanqueantes determinó que la inversión no afectó a ninguno de ellos. No obstante, sería necesario corroborar dichos puntos de rotura mediante PCR para afirmar este hecho.

No se pudo caracterizar molecularmente los puntos de rotura de la inversión J_1. Este hecho muestra la dificultad de obtener puntos de rotura debido a zonas muy repetitivas en el genoma que dificultan el proceso de ensamblaje de datos NGS.

A pesar de no obtener los resultados esperados, se obtuvo un protocolo bioinformático que permitirá al resto de investigadores determinar los puntos

de rotura de inversiones y abrir puertas a una elevada cantidad de estudios de adaptación.

Abstract

Chromosomal inversions are structural variations produced by the breakage of a chromosomal fragment and its reinsertion in an inverted orientation. Drosophila has been the most studied genus, with *D. subobscura* being particularly interesting for its abundant inversion polymorphism. Chromosomal inversions have been related to the adaptive character of the species.

The aim of this study was the characterization of chromosome inversions by Illumina paired-end reads mapping to a reference genome. For this purpose, a bioinformatics protocol was created to obtain the breakpoints of the J_1 inversion, of special interest due to its adaptive nature, and the breakpoints of the U_6 inversion.

The two break points of the U_6 inversion were characterized and the results revealed that the inversion was caused by NHEJ and staggered break. The study of the flanking coding genes determined that the inversion did not affect any of them. However, it would be necessary to corroborate said breakpoints by PCR to confirm this fact.

The breaking points of the J_1 inversion could not be characterized molecularly. This fact shows the difficulty of obtaining breakpoints due to highly repetitive areas in the genome that hinder the NGS data assembly process.

Despite not obtaining the expected results, a bioinformatics protocol was obtained that will allow future researchers to determine the breakpoints of inversions and will be a starting point for a large number of adaptation studies.

Índice

Lista de figuras

Lista de tablas

1. Introducción

1.1. Descripción general

En el presente trabajo se pretende caracterizar inversiones cromosómicas de *Drosophila subobscura*, donde se procederá a localizar los puntos de rotura de las inversiones J_1 y U_6 a partir de datos NGS. Para ello se desarrollará un protocolo bioinformático que deberá permitir identificar los genes contenidos en dichas inversiones y buscar evidencias biológicas que expliquen el carácter adaptativo de estas inversiones.

1.2. Contexto y justificación del Trabajo

A comienzos del siglo XX, Alfred H. Sturtevant, inmerso en un estudio con *Drosophila melanogaster*, observó alteraciones en la proporción de los fenotipos en la descendencia derivado de un factor cromosómico previamente desconocido, el cual, provocaba la supresión de la recombinación. Años más tarde, el propio Sturtevant descubrió que este fenómeno era producido por un tipo de polimorfismo cromosómico, al cual se le denominó, inversión cromosómica (1).

Las inversiones cromosómicas son un tipo de reordenamiento genómico, en el cual se produce la rotura de un fragmento cromosómico y la reinserción de éste en la orientación invertida (2). Estos polimorfismos están muy extendidos en la naturaleza, ya que se han observado en diversas especies animales y vegetales. Estudios recientes han mostrado una fuerte evidencia sobre el papel de estas mutaciones en procesos biológicos como la adaptación, la especiación y la evolución genética (3).

Este polimorfismo ha sido estudiado principalmente en dípteros, ya que estas especies presentan unos cromosomas gigantes, denominados cromosomas politénicos, formados por rondas repetidas de duplicación de ADN sin que haya división celular (4). En la era pre-genómica, estas estructuras eran estudiadas con laboriosas técnicas citológicas, las cuales requerían que las inversiones fueran suficientemente grandes para poder detectar diferencias estructurales (5).

La especie *Drosophila subobscura* es originaria de la región paleártica donde habita, desde el sur de Escandinavia hasta el norte de África. Además, a finales del siglo XX colonizó amplias zonas de América (6). El estudio del polimorfismo cromosómico por inversión de esta especie tiene un interés particular debido a que es muy abundante, ya que se han descrito más de 80 ordenaciones cromosómicas diferentes entre sus cinco cromosomas acrocéntricos (7).

Esta especie cuenta con el cariotipo ancestral del género Drosophila: cinco cromosomas acrocéntricos (que en esta especie reciben los nombres de A=X, J, U, E y O) y un cromosoma puntiforme (denominado dot).

La cepa “Küsnacht” fue la primera cepa de *D. subobscura* caracterizada y la ordenación de sus cromosomas fueron establecidos como estándar (ST). Las diferentes ordenaciones caracterizadas a partir de ese momento se designaron con subíndices numéricos siguiendo el orden de descubrimiento (8). Así los nombres J_1, U_{1+2} o $U_{\underline{1+2+6}}$ indican la existencia de la inversión 1 en el cromosoma J, las inversiones independientes 1 y 2 sobre el cromosoma U y las inversiones 1, 2 junto con la 6 que se solapa con la 1 y la 2 sobre el cromosoma U, respectivamente.

Existen dos mecanismos principales de origen de las inversiones cromosómicas. El primer mecanismo, la recombinación homóloga no alélica (NAHR), genera inversiones cuando recombinan dos secuencias que están repetidas en el cromosoma y se encuentran en orientación invertida. Así, en este mecanismo se observan repeticiones invertidas en los extremos de la inversión tanto de la ordenación original como de la ordenación invertida. El segundo mecanismo, rotura cromosómica y reparación ectópica a través de la unión de extremos no homólogos (NHEJ), se produce debido a dos roturas aleatorias y simultáneas en un mismo cromosoma y la posterior reparación errónea. Este segundo mecanismo, no requiere de duplicaciones invertidas en la ordenación original y puede generar una ordenación invertida sin duplicaciones con cortes rectos, o con duplicaciones invertidas en sus extremos a través de la reparación de cortes escalonados (9–11).

La identificación y caracterización precisa de los puntos de rotura de una inversión es un objetivo muy importante desde el punto de vista del estudio de la evolución de las poblaciones, debido a que las regiones adyacentes al punto de rotura experimentan muy poca o ninguna recombinación entre las distintas ordenaciones y permite indagar en la historia evolutiva de las inversiones (12).

Por ello, el presente trabajo se dispone a caracterizar molecularmente los puntos de rotura de dos inversiónes para poder estudiar los genes contenidos en ellas y buscar evidencias biológicas que expliquen el carácter adaptativo que se ha observado para ellas.

1.3. Objetivos del Trabajo

1.3.1. Objetivo general

Caracterizar inversiones cromosómicas en *Drosophila subobscura* por mapeo de *reads pair-ends* de Illumina sobre un genoma de referencia.

1.3.2. Objetivos específicos

- Desarrollar un protocolo bioinformático que permita identificar inversiones cromosómicas a partir de datos NGS. Para ello se usarán los *reads* apareados de Illumina de la línea OF28 de *D. subobscura* que es homocariotípica para las ordenaciones J_1, U_{1+2+6}, E_{st}, O_{st} y A_{st}.

- Caracterizar los puntos de rotura de la inversión J_1. Esta inversión es particularmente interesante puesto que diversos estudios indican que puede estar sometida a selección.

- Caracterizar los puntos de rotura de la inversión U_6. Este objetivo es un objetivo secundario que sólo se abordará en caso de que el tiempo de ejecución del TFM lo permita.

- Identificar los genes contenidos en la inversión J_1 y buscar evidencias que expliquen su carácter adaptativo.

1.4. Impacto en sostenibilidad, ético-social y de diversidad

Una vez asumida la importancia del cambio climático que el planeta está experimentando, muchos científicos han tenido la necesidad de estudiar cómo este cambio en la temperatura puede afectar a las poblaciones. Para ello, se han llevado a cabo diversas líneas de investigación de diferentes especies (13,14).

El estudio de inversiones cromosómicas en *D. subobscura* puede ayudar a entender cómo las poblaciones de esta especie responden a cambios ambientales y cómo se adaptan a ellos (15). Los primeros informes sobre el calentamiento global apuntaban posibles cambios en la frecuencia de las inversiones, no obstante, estudios recientes han demostrado que existen otros factores que influyen en estos cambios (16,17).

El estudio del cambio climático está siendo un revulsivo para estudiar en profundidad los mecanismos genéticos de adaptación, así como los genes contenidos en las inversiones cromosómicas que pueden explicar el carácter

adaptativo de estas y ayudar a crear estrategias de estudios que expliquen la adaptación de otras especies al calentamiento global (15).

También es importante considerar que la manipulación genética y las investigaciones en animales pueden plantear preocupaciones éticas en torno al bienestar animal. Sin embargo, los estudios en *Drosophila* utilizan métodos que minimizan el daño de estos animales, por lo tanto, el impacto ético-social es mínimo.

Por otro lado, la secuenciación de genomas completos por NGS, permite que, utilizando un mínimo número de animales, se puedan realizar múltiples estudios. Este es el caso del presente trabajo donde se han utilizado las secuencias obtenidas previamente para otros fines.

Respecto al impacto en diversidad, este es mínimo ya que se trata de un trabajo de carácter científico-técnico. No obstante, se han tenido en cuenta todas las aportaciones con independencia del género, raza, etnia, o condición del autor o la autora.

1.5. Enfoque y método seguido

La identificación y caracterización de inversiones en *D. subobscura* han dado muchas claves sobre la evolución y especiación de las especies de su subgrupo. Hoy en día se han desarrollado herramientas bioinformáticas para que su estudio sea más sencillo y permita el estudio en profundidad. Así pues, en el presente trabajo se está estudiando diferentes *pipelines* para llevar a cabo la identificación de los puntos de rotura, y se ha elegido la que muestra la Tabla 1:

Pasos del análisis	Programas	Descripción
Control de calidad y recorte de *reads*	FastQC	Señala anomalías producidas en la creación de las bibliotecas o durante la secuenciación (18).
	Trimmomatic	Es un recortador flexible de datos de secuenciación de Illumina, la cual recorta *reads* de baja calidad (19).
Mapeo	BWA-MEM	Algoritmo de alineación para alinear secuencias a un genoma de referencia. Admite *reads* de extremos emparejados y muestra uno de los mejores rendimientos en comparación con otros alineadores hasta la fecha (20).
Clasificación, conversión de formato y filtrado (21)	Samtools	Implementa el post-procesamiento de alineaciones, como la indexación, la llamada de variantes o la visualización de alineaciones (22).
Localización de puntos de corte putativos	Breakdancer	Herramienta de detección de variantes estructurales en todo el genoma a partir de *reads* paired-end de NGS (23)
Ensamblado de *novo*	SPades	Conjunto de herramientas de ensamblaje capaz de proporcionar ensamblajes híbridos utilizando *reads* cortos o largos de diferentes tecnologías como PacBio, Illumina o Oxford Nanopore (24)
Anotación de inversiones cromosómicas	BLAST	Algoritmo para búsqueda de similitud de secuencias de ADN y proteínas en bases de datos (25)

Tabla 1: *Herramientas bioinformáticas para la caracterización de inversiones cromosómicas.*

En un principio, estos son los programas elegidos para el análisis siguiendo la metodología de varios autores (12,26,27). A lo largo del proyecto se irá detallando los programas finalmente utilizados y en el caso de cambio o adición se indicará los detalles y la explicación de dicho cambio.

1.6. Planificación del Trabajo

En este apartado se realizará una temporalización del trabajo, dividiendo el trabajo por tareas y organizando los objetivos marcados en el tiempo.

1.6.1. Tareas

- Búsqueda bibliográfica sobre la metodología empleada para caracterizar inversiones cromosómicas a partir de datos NGS.

- Control de calidad de los datos crudos de secuenciación y eliminación de nucleótidos de baja calidad.

- Mapeo de los *reads* con el genoma de referencia.

- Localización de *clúster de reads* discordantes para identificar posibles puntos de rotura en los cromosomas J y U.

- Ensamblado de *novo* de los *clústers* de *reads* discordantes para obtener los puntos de rotura en las ordenaciones invertidas.

- Caracterización de las regiones de corte en ambas ordenaciones (identificación de genes cercanos y de posibles secuencias repetitivas).

1.6.2. Calendario

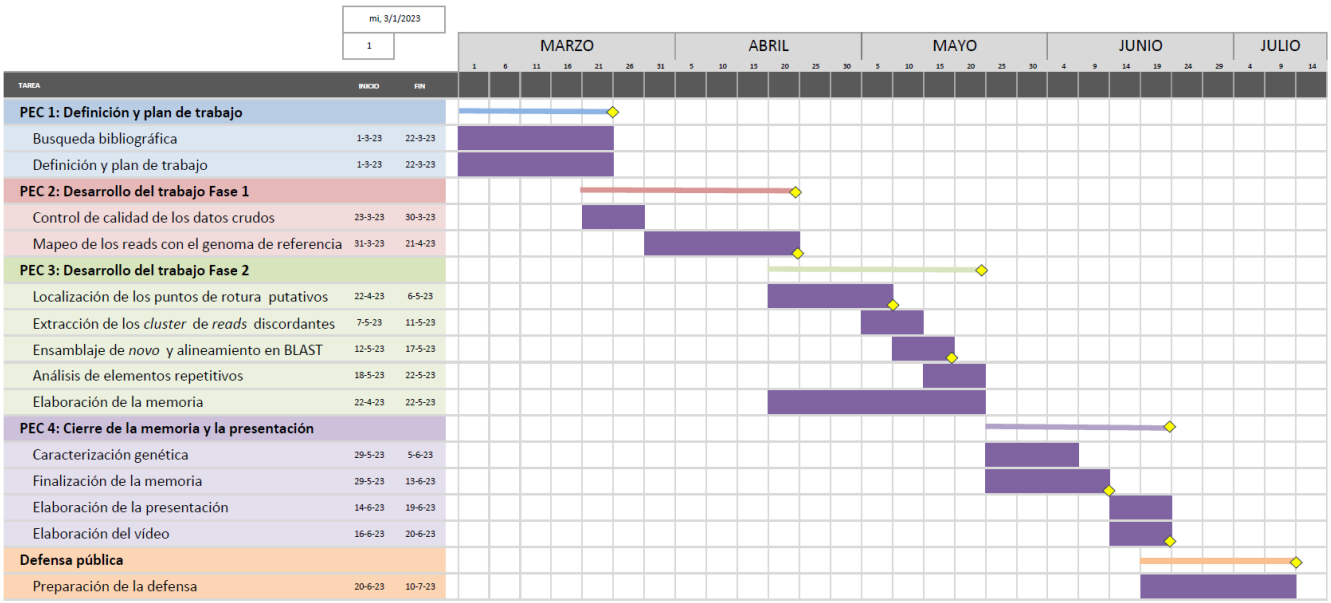

***Figura 1**: Diagrama de Gantt con la planificación y las tareas realizadas a lo largo del estudio. Cada color representa una fase del estudio. Los rombos en amarillo representan los hitos definidos en el apartado siguiente.*

1.6.3. Hitos.

- **Hito 1:** Finalización del plan de trabajo.
- **Hito 2:** Finalización del mapeo de los datos NGS de *D. subobscura*.
- **Hito 3:** Finalización de la PEC 2 con los resultados obtenidos hasta el momento.
- **Hito 4:** Localización de los puntos de rotura putativos de las inversiones J_1 y U_6
- **Hito 5:** Finalización del ensamblaje de *novo* y el alineamiento de los *contig* obtenidos en BLAST.
- **Hito 6:** Finalización de la PEC 3 con los resultados del análisis.
- **Hito 7:** Finalización de la memoria del Trabajo Final de Máster.
- **Hito 8:** Finalización de la presentación y del vídeo del trabajo
- **Hito 9:** Finalización de la PEC 4 y cierre de la memoria.
- **Hito 10:** Defensa pública.

1.6.4. Análisis de riesgos.

El trabajo planteado cuenta con algunos factores que pueden influir negativamente en el progreso del estudio, como los siguientes:

- La necesidad de un ordenador personal con una gran capacidad de memoria RAM que permita el procesamiento de los datos en un periodo de tiempo adecuado.
- La necesidad de un sistema operativo de software libre para poder realizar los pasos necesarios para el análisis.

La solución planteada para estos problemas es la siguiente:

- La utilización de un ordenador con una memoria de RAM de 16Gb que permita el análisis y la instalación de un subsistema de Windows el cual permita la utilización de Linux (software libre).

Otro factor limitante que puede aparecer a lo largo del estudio es:

- La lentitud al mapear datos NGS de gran volumen, el cual puede retrasar los análisis posteriores y con ello, no obtener los resultados deseados.

Si esto ocurre, se pedirá el uso remoto de un servidor de la UB.

1.7. Resultados esperados.

1.7.1. Plan de trabajo.

El plan de trabajo es un entregable donde se explicará la temática del estudio y se plantearán los objetivos que se deberán alcanzar para llevar a cabo el estudio con éxito. Para ello, se definirán las líneas generales del estudio, se marcará el objetivo general y los objetivos específicos de manera clara y concisa, se llevará a cabo una planificación de las tareas e hitos que se deberán alcanzar y por último se valorarán los posibles riesgos que podrían acontecer.

1.7.2. Memoria.

La memoria será el escrito final con todos los datos del estudio. Para su realización se han planteado cuatro entregables donde se irán desarrollando los diferentes apartados del trabajo final. El entregable 4 será la memoria finalizada que constará de los siguientes apartados: introducción, materiales y métodos, resultados y discusión, conclusión y trabajos futuros, glosario y bibliografía.

1.7.3. Producto.

El producto obtenido al final de este trabajo será un primer estudio de los puntos de rotura de las inversiones J_1 y U_6. En caso de obtener los puntos de rotura de la ordenación J_1 y siempre que el tiempo de ejecución del TFM lo permita, se abordará la posible implicación biológica de genes contenidos en dicha inversión sobre la adaptación de *D. subobscura* a diferentes entornos. Si los resultados son satisfactorios, este estudió se podrá utilizar para posteriores publicaciones.

1.7.4. Presentación virtual.

Se realizará una presentación que consistirá en un resumen del proyecto realizado. En ella se explicará la temática del proyecto, los análisis llevados a cabo y los resultados obtenidos, con una breve conclusión. Para la exposición oral de dicha presentación, se realizará un vídeo.

2. Estado del arte

Las inversiones cromosómicas son una herramienta valiosa en la investigación genética y evolutiva en *Drosophila*. Cada vez hay más evidencias que demuestran el papel de los polimorfismos de inversión en la adaptación y la especiación (28).

D. subobscura lleva años siendo ampliamente estudiada por sus ricos polimorfismos de inversión. Diversos estudios han revelado que la frecuencia alélica de algunas inversiones cromosómicas está relacionada con factores ambientales como la temperatura, la lluvia o la humedad (15,29,30).

Además, se ha demostrado, que estos polimorfismos no solo han respondido a señales ambientales, sino que también han logrado tener la capacidad de adaptarse rápidamente a nuevos hábitats cambiando las frecuencias de las ordenaciones cromosómicas para aportar alelos favorecidos en el nuevo entorno (30).

Los alelos contenidos en las inversiones pueden interactuar de forma epistática o aditiva para mantener un fenotipo poligénico complejo, como el tamaño corporal, la resistencia al estrés o la fecundidad (21)

Los puntos de roturas en las inversiones también han sido ampliamente estudiados, ya que la alteración de la estructura de los cromosomas puede afectar directa o indirectamente a la expresión del genoma. Los puntos de rotura de las inversiones pueden afectar directamente al cambiar la función y expresión de un gen situado cerca de la región fragmentada. Por otro lado, los puntos de

rotura pueden afectar indirectamente a la expresión al mantener unidas las combinaciones de variantes genéticas, debido a la reducción de la recombinación. Rozas y colaboradores (31) estudiaron el gen *rp49* en diferentes ordenaciones de *D. subobscura* y encontraron un nivel de intercambio genético más elevado cuando el gen se encontraba en una posición más central del fragmento invertido que cuando se encontraba cerca de los puntos de rotura.

Esto hace plantearse la importancia de encontrar aquellos genes dentro de las inversiones y estudiar cómo se comportan. Por ello, se busca contribuir al estudio de *D. subobscura* para alcanzar un mayor conocimiento de su genoma. En este contexto, se propone identificar los puntos de rotura de las inversiones J_1 y U_6 e identificar los genes contenidos en la inversión J_1, lo cual nos podría dar claves para entender mejor la adaptación climática de esta especie.

El polimorfismo cromosómico por inversiones en *D. subobscura* ha sido ampliamente investigado en relación con su distribución geográfica y desde un punto de vista evolutivo. *D. subobscura* es una especie originaria de la región paleártica, sin embargo, en los años 80, la especie fue descubierta en el norte y sur de América. A los pocos años de su introducción en América, se observó clinas latitudinales en el mismo sentido que en la zona Paleártica. Estos resultados apoyan estudios sobre el carácter adaptativo de la especie, los cuales han revelado diferencias en las frecuencias de los ordenamientos genómicos de los cromosomas J y U relacionados con factores microclimáticos y cambios temporales (30,32,33).

El análisis de las inversiones ha permitido distinguir las ordenaciones cromosómicas en dos grupos, adaptadas a clima frío y adaptadas a clima cálido. La ordenación J_1 es considerada de clima cálido y disminuye de frecuencia a medida que la latitud aumenta. La ordenación J_1 es común del área mediterránea, sin embargo, la ordenación U_{1+2+6} es poco frecuente. No obstante, su estudio nos ampliará el conocimiento sobre los mecanismos de generación de inversiones de esta especie altamente polimórfica (6,33)

D. subobscura ha sido durante mucho tiempo un modelo central en el estudio de la genética evolutiva. Hasta hace relativamente poco, su uso se ha visto obstaculizado por la falta de un genoma de referencia. Su estudio era llevado a cabo con técnicas de comparación de genomas de especies cercanas del género Drosophila como *D. pseudoobscura* y *D. persimilis* (34,35). Para romper esta brecha, diversos grupos de investigación se esforzaron para obtener un genoma de referencia de dicha especie. Karageorgiou y colaboradores (36) presentaron la primera secuencia de lectura larga y alta calidad de la cepa *chcu* de *D. subobscura*, la cual presenta la ordenación cromosómica estándar para todos los cromosomas a excepción de O_{3+4}. Casi simultáneamente, se obtuvo la secuencia de otro genoma de alta calidad por Bracewell y colaboradores (37). Este genoma presenta las mismas ordenaciones cromosómicas que la cepa

chcu excepto para el cromosoma U que presenta la ordenación U_{1+2} que es la ancestral de la especie. Este último, es el genoma de referencia depositado en el Centro Nacional para la Información Biotecnológica o, en inglés, National Center for Biotechnology Information (NCBI) y dispone de archivos de anotación genética, facilitando así su estudio. Por esta razón, será el utilizado para este proyecto. Además, el hecho de poseer la ordenación U_{1+2} facilitará la caracterización de U_6, que se originó sobre un cromosoma con ordenación U_{1+2} y es una inversión que se solapa tanto a U_1 como a U_2.

Este reciente avance nos permite desarrollar un protocolo bioinformático como los llevado a cabo en *D. melanogaster* para la caracterización de puntos de rotura de inversiones cromosómicas (14).

En base a lo anterior, se pretende crear un protocolo bioinformático que permita al resto de investigadores determinar los puntos de rotura de inversiones y abrir puertas a una elevada cantidad de estudios con la especie *D. subobscura*, facilitando un gran avance en la investigación científica de la adaptación de la especie.

3. Material y métodos

3.1. Obtención de las secuencias

Las muestras se obtuvieron a partir de una hembra capturada en los terrenos del Observatorio Fabra, en Barcelona, y tras al menos 13 generaciones de cruces entre hermanos. El objetivo fue conseguir una línea lo más isogénica posible. Tras ello, se obtuvo una línea homocariotípica para las ordenaciones de todos sus cromosomas A_{st}, J_1, U_{1+2+6}, E_{st} y O_{st}. A esta línea se le denominó "*línea OF28*".

Las muestras fueron preparadas y posteriormente secuenciadas por el Centro Nacional de Análisis Genómico de Barcelona. Los *reads* obtenidos de Illumina fueron de tipo *paired end* y de longitud de 101 pb.

Por otro lado, para realizar el mapeo fue necesario disponer de un genoma de referencia. Este lo obtuvimos de la base de datos NCBI, la cual fue obtenida por Bracewell y colaboradores en 2019. El genoma de referencia tiene la ordenación estándar para todos sus cromosomas, a excepción del cromosoma U y O, que presentan la ordenación U_{1+2} y O_{3+4}, respectivamente (37).

A continuación, se detallará el protocolo bioinformático llevado a cabo y los programas utilizados para su desarrollo. Los *scripts* utilizados para el estudio están disponibles en https://github.com/Kenia98/TFM

3.2. Análisis de calidad de las secuencias

El primer paso que se debe realizar tras obtener los datos de ultrasecuenciación es el estudio de calidad de las secuencias obtenidas. FastQC permite señalar anomalías producidas en la creación de las bibliotecas o durante la secuenciación (18).

El control de calidad es presentado mediante una serie de tablas y gráficos que permiten estudiar la calidad de los *reads* y poder solucionar problemas que surjan para los posteriores análisis.

Tras introducir los dos archivos a estudiar, se obtuvieron dos archivos en formato HTML, donde se podía observar estos gráficos. Los gráficos dan la información suficiente para poder comprobar si la calidad de los datos es buena o si por el contrario hay que realizar algún pre-procesado con ellos.

3.3. Limpieza de secuencias

Tras observar que la calidad de los *reads* era baja, se procedió a realizar un recorte de estas para así mejorar la calidad de los datos y poder seguir con el análisis correctamente. Para ello, se utilizó trimmomatic.

Trimmomatic es un recortador flexible de datos de secuenciación de Illumina, encargado de recortar *reads* de baja calidad. Su función principal es la eliminación de adaptadores y cebadores para proporcionar una mejor calidad de lectura. Esta herramienta utiliza *reads* de tipo *paired end* y *single end*. Además, trimmomatic admite datos de calidad de secuencia tanto en formato estándar (phred+33) como en formato Illumina (phred+64) (19).

Para realizar el recorte se utilizan una serie de parámetros que se explicarán a continuación:

- ILLUMINACLIP: elimina los adaptadores usados durante la secuenciación, en caso de que estén presentes.
- LEADING: elimina los nucleótidos del extremo 5’ con una calidad inferior a la indicada.
- TRAILING: elimina los nucleótidos del extremo 3’ con una calidad inferior a la indicada.
- SLIDINGWINDOW: rastrea la secuencia por ventanas y corta cuando la calidad media es inferior a la indicada. En primer lugar, se indica la longitud de la ventana y posteriormente, el umbral de calidad.
- MINLEN: elimina las secuencias que presentan una longitud inferior a la indicada.

Para realizar el recorte se ejecutó la herramienta para los dos archivos *paired end* y se le indicaron los parámetros elegidos para el recorte. Tras su ejecución se obtuvieron cuatro archivos, "*forward paired"*, "*forward unpaired"*, "*reverse paired"* y "*reverse unpaired"*, dos correspondientes a *reads* apareados y otros dos a *reads* no apareados. Solo se utilizaron los *reads* apareados.

Tras el recorte de las secuencias con esta herramienta, se volvió a analizar la calidad de las secuencias con FastQC para comprobar si había mejorado o por si el contrario había que mejorar el recorte.

3.4. Mapeado de *reads*

Tras el procesamiento y comprobar que los datos tenían buena calidad, se procedió a realizar el mapeo. Para ello, se utilizó la herramienta Burrows-Wheeler Aligner (BWA).

BWA es un software diseñado para la alineación de secuencias genéticas poco divergentes con un genoma de referencia grande. Este software presenta tres algoritmos conocidos como BWA-backtrack, BWA-SW y BWA-MEM. BWA-backtrack está diseñado para lecturas cortas, con una longitud máxima de 100 pb. Por el contrario, BWA-SW y BWA-MEM, está diseñado para secuencias más largas comprendidas entre 70 pb y 1Mb. BWA-SW y BWA-MEM presentan características similares, sin embargo, en general se recomienda el uso de BWA-MEM para consultas de alta calidad, por su precisión y su mayor velocidad. BWA-MEM también tiene un mayor rendimiento que BWA-baktrack cuando se trabaja con secuencias de Illumina entre 70 y 100 pb (20,38,39)

Por esa razón, se utilizó BWA-MEM. Los *reads* fueron mapeados siguiendo el método descrito en Kapun y colaboradores (40). El primer paso fue crear un hologenoma, el cual, utilizamos como genoma de referencia. El hologenoma es una unidad genética formada por la combinación del genoma del hospedador y los microorganismos asociados a él (41).

El hologenoma utilizado estaba compuesto por el genoma de *D. subobscura* (NC_048530) y los microorganismos asociados a esta especie, incluyendo *Saccharomyces cerevisiae* (GCF_000146045.2), *Wolbachia pipientis* (NC_002978.6), *Pseudomonas entomophila* (NC_008027.1), *Commensalibacter intestine* (NZ_AGFR00000000.1), *Acetobacter pomorum* (NZ_AEUP00000000.1), *Gluconobacter morbifer* (NZ_AGQV00000000.1), *Providencia burhodogranariea* (NZ_AKKL00000000.1), *Providencia alcalifaciens* (NC_AKKM01000049.1), *Providencia rettgeri* (NZ_AJSB00000000.1) y *Enterococcus faecalis* (NC_004668.1).

Tras obtener el hologenoma, se procedió a su indexación y se inició el alineamiento de secuencias con los archivos obtenidos en el apartado anterior. Tras terminar el proceso de alineación se obtuvieron dos archivos de salida en formato SAM (Sequence Alignment/Map), correspondiente a las alineaciones de las muestras con el genoma de referencia. SAM es un formato de salida común de alineaciones diseñado para facilitar y unificar los datos de alineamientos (42).

3.5. Transformación de archivos y eliminación de duplicados.

Tras realizar el mapeo, se utilizó la herramienta Samtools. Samtools es un paquete de software para analizar y manipular alineaciones en formato SAM/BAM (22). En nuestro caso, se encargó de convertir el archivo SAM a BAM, ordenar el archivo y eliminar los duplicados. BAM es un archivo equivalente, con la misma información solo que almacenada de forma binaria. De este modo, los programas utilizados posteriormente podrán trabajar de forma más eficiente.

Tras convertir el archivo de formato SAM a BAM, se ordenó por coordenadas y se procedió a la eliminación de *reads* duplicados. Para finalizar se indexó el archivo. Este paso proporciona un nuevo archivo de índice, el cual permite una búsqueda rápida de los datos de alineamiento. La indexación del archivo BAM fue necesaria para procesos posteriores.

3.6. Localización de puntos de rotura putativos.

Para obtener los puntos de rotura putativos se utilizó el software BreakDancerMax (43). Este es un software de código abierto que implementa un algoritmo capaz de detectar cinco variantes estructurales a través de todo el genoma: inserciones, deleciones, inversiones y translocaciones intracromosómicas e intercromosómicas. El proceso de detección busca regiones genómicas que presentan el mapeo de más pares de *reads* anómalos de lo esperado, basándose en la distancia de separación y orientación de los pares de *reads* alineados, el umbral especificado por el usuario y la distribución empírica del tamaño de inserto.

El software fue ejecutado siguiendo el método descrito por Fan y colaboradores (44). La ejecución se desarrolló en dos pasos, primero se ejecutó el *script* “bam2cfg.pl” que se encarga de analizar los primeros miles de *reads* de los archivos de entrada bam y crea un archivo de configuración con estadísticas para cada grupo de lectura. Tras ello, se ejecutó el script “breakdancer-max”, el cual utiliza el archivo estadístico anteriormente descrito para identificar y agrupar los pares de *reads* discordantes y genera un archivo con una lista de variantes estructurales (SV, structural variation) putativos (44).

Para aumentar la eficiencia y siempre y cuando no se busquen translocaciones intercromosómicas, es recomendable ejecutar el *script* “breakdancer-max” por cromosomas. Por ello, se ejecutó dos veces, uno por cada cromosoma de interés, J y U. Tras ello, se obtuvieron dos archivos de salida de texto, uno por cada cromosoma, los cuales mostraron todos los SV putativos localizados.

3.7. Identificación de posibles puntos de rotura.

Tras obtener dos archivos con los SV putativos, se filtró para obtener solo las inversiones cromosómicas. Se seleccionaron las inversiones detectadas cuyas coordenadas coincidían con la región acotada por los marcadores flanqueantes en los puntos de rotura de las inversiones estudiadas. La Figura 2 muestra el mapa citológico descrito por Kunze-Mühl and Müller (45) que en el caso del cromosoma U, se ha modificado para tener la ordenación $U_{1+2,}$ y los marcadores seleccionados (26,34,46).

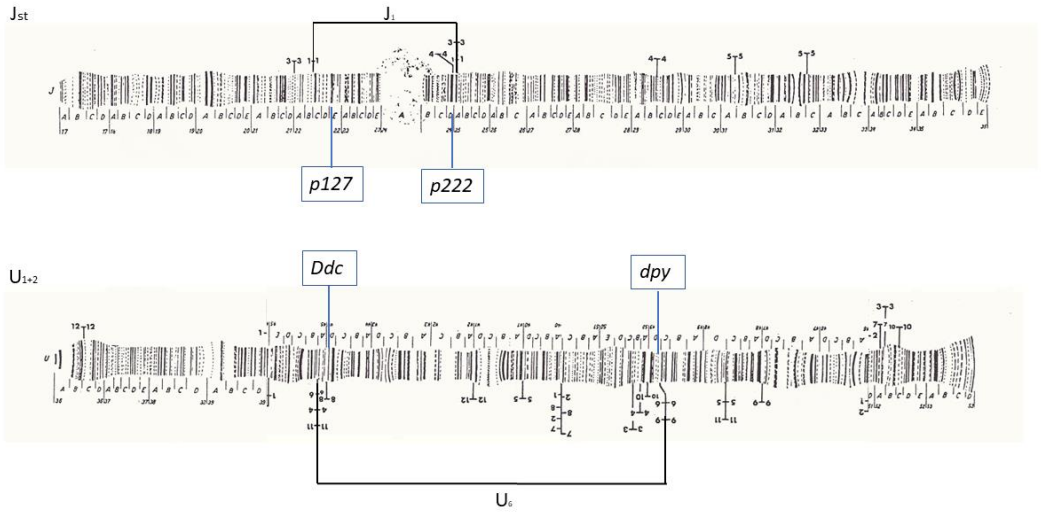

***Figura 2**: Mapa citológico de los cromosomas Jst y U_{1+2} de D. subobscura (Kunze-Mühl and Müller 1958) donde se ha indicado la localización de las inversiones cromosómicas J1 y U6 y los marcadores genéticos más cercanos dichas inversiones.*

Para localizar las inversiones J_1 y U_6 en el archivo de breakdancer, las secuencias de los marcadores genéticos fueron descargadas del NCBI para su posterior alineamiento con el genoma de referencia de *D. subobscura* (NC_048530). Las secuencias de los marcadores del cromosoma J no están depositadas en NCBI. En este caso, se extrajeron las coordenadas alineando la secuencia de los oligos de amplificación usados en la Tesis de Pratdesaba (46). Este paso permitió obtener las coordenadas en las cuales debía estar la inversión y extraer las localizaciones putativas de las inversiones cromosómicas.

3.8. Extracción de los *clúster* de *reads* discordantes

Tras tener localizadas las inversiones cromosómicas putativas, se tuvo que comprobar dichas inversiones y obtener las coordenadas reales de los puntos de corte. Por ello, lo primero fue extraer del archivo bam obtenido con anterioridad, los *reads* que mapeaban alrededor de las coordenadas obtenidas en el apartado anterior. Se obtuvieron cuatro archivos bam, dos para el cromosoma J y dos para el cromosoma U; uno para la coordenada inicial y otro para la coordenada final. Estos archivos fueron ordenados y a continuación, se transformaron en archivos fastq. Se crearon tres archivos fastq por cada archivo bam, uno correspondiente a los *reads forward*, otro a los *reads reverse* y el último a los *reads* no emparejados (*singletons*). Este último era el de mayor interés, pues en él se esperaba encontrar los *reads* cuyas parejas podrían corresponder justo al punto de rotura en la ordenación invertida. En el genoma de referencia, estos *reads* mapearán su inicio junto a uno de los puntos de rotura y su final lo hará junto al otro punto de rotura y, por tanto, los programas de mapeo encuentran un problema y lo descartan. Este archivo contenía los nombres de los *reads* de interés, y con ellos se extrajo de los archivos originales, los *reads* no emparejados y su pareja. El objetivo de extraer la pareja de los *reads* no emparejados es obtener aquellos *reads* no mapeados posiblemente por encontrarse en el punto de corte de la inversión.

Para finalizar el proceso de extracción, se crearon dos archivos por cada cromosoma, uno correspondiente a los *reads forward* y otro a los *reads reverse*. Todo este proceso fue llevado a cabo con la herramienta Samtools (22).

3.9. Ensamblaje de novo y alineamiento en BLAST

El siguiente paso fue el ensamblaje de *novo* con la herramienta SPAdes (47). SPAdes es un software de ensamblaje de genomas que utiliza lecturas cortas y largas para construir secuencias de ADN a partir de datos de secuenciación de alto rendimiento. Utiliza un enfoque de ensamblaje basado en gráficos de De Bruijn y es capaz de manejar datos de alta cobertura y complejidad, como metagenomas y genomas poliploides. Además, es compatible con varias plataformas de secuenciación, como Illumina, Ion Torrent y PacBio. SPAdes es un programa de código abierto ampliamente utilizado en investigación genómica y bioinformática debido a su alta eficiencia y precisión (24).

El proceso de ensamblaje se realizó para construir *contigs* donde se localizan los puntos de rotura de las inversiones estudiadas. Para ello, se ejecutó el *script* “spades.py” dos veces, uno por cada cromosoma, donde en cada ejecución se utilizó los dos archivos fastq creados en el apartado anterior.

Tras el ensamblaje, se llevó a cabo la evaluación del proceso con la herramienta quast (48). Quast es una herramienta de evaluación de calidad de genomas ensamblados (49). Esta herramienta proporciona un archivo de texto con información estadística sobre el ensamblaje como el número de *contigs* creados, la clasificación de los *contigs* por tamaño y una serie de parámetros de ensamblaje, entre otros.

Una vez llevado a cabo el ensamblaje con éxito, se seleccionaron los *contigs* más largos y se llevaron a la plataforma BLAST (en inglés, Basic Local Alignment Search Tool) de alineamiento local. Allí se alinearon los *contigs* con el genoma de referencia de *D. subobscura* (NC_048530) buscando si algún *contig* alineaba un extremo en una región y otro extremo en otra región separadas del genoma, cerca de los puntos de corte putativos seleccionados.

3.10. Análisis de elementos repetitivos

Los elementos repetitivos pueden complicar la detección de puntos de rotura de las inversiones. Por esta razón se utilizaron herramientas para su detección.

RepeatMasker es un software encargado de detectar repeticiones intercaladas en secuencias de ADN y secuencias de ADN de baja complejidad. El programa nos proporciona una anotación detallada con todas las repeticiones presentes en las secuencias de consulta (50).

Se introdujeron en RepeatMasker los *contigs* de longitud mayor a 1kb obtenidos en el ensamblaje de *novo* de *reads* del cromosoma U. De este modo, se estudió los *contigs* en busca de estructuras repetidas.

Por otro lado, se buscó posibles inserciones del transposón SGM (SubobscuraGuancheMaddeirensis). Este transposón es un tipo de elemento transponible encontrado en el grupo de especies *D. subobscura* (*D. subobscura, D. guanche, D. maddeirensis*). Es una familia de ADN transponible que se encuentra en el genoma de múltiples especies de Drosophila. Se ha demostrado que los elementos SGM tienen una función reguladora y pueden afectar la expresión génica y la estructura del genoma (51). Por esta razón, la importancia de su estudio. Así pues, se alineó la secuencia del transposón SGM a los *contigs* de longitud mayor a 1kb en busca de este elemento.

Por último, también se alinearon en BLAST los *contigs* de longitud mayor a 1kb entre ellos, en busca de secuencias repetidas entre los *contigs* ensamblados.

3.11. Caracterización genética del punto de rotura U_6

La caracterización de los puntos de rotura de las inversiones a nivel molecular permite conocer los mecanismos, por el cual se originan las inversiones cromosómicas y, además comprender su significado evolutivo a nivel fenotípico (5).

Se accedió al listado en NCBI de las regiones de codificación de genes (CDS, en inglés Coding Sequences) del genoma de referencia de *D. subobscura* para buscar los genes de proteínas flanqueantes a los puntos de rotura de la inversión U_6. Se filtró por coordenadas los genes cercanos al punto de rotura proximal obtenidos por el ensamblaje de *novo*, U:7.714.784 y al punto de rotura distal, U:16.613.622. Tras ello, se evaluaron los genes encontrados y se estudió las posibles consecuencias de la inversión cercana a estos genes.

3.12. Tiempo de ejecución y memoria consumida

Para finalizar informamos de los recursos computacionales utilizados para el desarrollo del protocolo bioinformático para la caracterización de los puntos de rotura de las inversiones cromosómicas (Tabla 2).

El dispositivo utilizado es un portátil que cumple con los requisitos necesarios para llevar a cabo las tareas requeridas en el contexto del estudio.

El portátil está equipado con una memoria RAM de 16 GB, garantizando un rendimiento fluido y eficiente en las tareas de mapeado y análisis de datos. El procesador (CPU, unidad central de procesamiento) utilizado es Intel® Core™ i7 de 11th generación, lo que permitió un rápido procesamiento de los datos. La combinación de estas características garantizó un rendimiento óptimo y una capacidad de respuesta adecuada para abordar los desafíos planteados en el desarrollo del estudio.

Procesos	Tiempo de ejecución(s)	CPU consumida
Control de calidad	>7200	-
Mapeado	2903,726	24.64%
Ensamblado	0.136521	7Gb(max.)

***Tabla 2**: Tiempo de ejecución y CPU consumida por las herramientas principales en el desarrollo del estudio.*

4. Resultados

4.1. Calidad de las secuencias y recorte

La evaluación de la calidad de las secuencias de los datos crudos mostró en primer lugar, la estadística básica de los archivos. Los archivos de secuenciación contenían 41.333.785 pares de secuencias, con una longitud de 101 pb y un contenido de GC del 43%. Las secuencias mostraron unos malos niveles de calidad de la secuencia por base y por secuencia. En la Figura 3B se observa que a partir de la posición 90, algunas de las secuencias tienen un valor de calidad por debajo de 28 phred, lo que indica que la calidad de secuencia no es buena y esta empeora en las bases posteriores a 90. Debido a esto, se tuvo que recortar los *reads* que tenían mala calidad con el programa trimmomatic.

En la Figura 3C “calidad de secuencias por teselas”, se observa que el programa lanza una señal de advertencia, ya que se pueden ver algunas manchas sobre el fondo azul. Esto indica que se han producido interferencias. Estas pueden ser debidas a problemas técnicos o con los reactivos en el proceso de secuenciación. Aunque es cierto, que son manchas muy pequeñas, por lo que pueden ser causadas también por los *reads* de mala calidad.

Todos los demás gráficos muestran buenos resultados, tanto el contenido de GC, la distribución de las secuencias o el nivel de duplicaciones.

Tras la evaluación de las secuencias en crudo, se pasó a utilizar trimmomatic para recortar las secuencias de mala calidad y se vuelve a evaluar las secuencias recortadas.

La estadística básica (figura 4A) muestra que ahora han quedado 30.693.900 pares de secuencias de entre 75 y 101 pb, ya que se decidió eliminar con trimmomatic las secuencias por debajo de 75 pb. Con ello, se pasa a tener un porcentaje de GC de 42%, una bajada sin importancia, ya que solo ha disminuido en un 1%.

El gráfico de calidad de secuencia por base (Figura 4B) muestra una buena calidad de las secuencias. Por otro lado, en el gráfico de calidad de las secuencias por teselas (Figura 4C), se observa una mejora de la calidad, ya que han disminuido las manchas claras. Por lo que se puede concluir que las interferencias podrían estar causadas por esos fragmentos que han sido recortados por mala calidad. Estos pueden haber sido mal secuenciados y de ahí los problemas de calidad.

En este caso, en el análisis de datos post-procesados, se ha activado una advertencia en el gráfico de longitud de distribución por secuencia. Esto es debido a que, al recortar secuencias de mala calidad, las lecturas presentan longitud de bases diferentes, por lo que la distribución no es equilibrada. Sin embargo, se observa que el número de *reads* recortados era muy bajo en comparación con los *reads* que presentan 101 pb, por lo tanto, no fue de suma importancia.

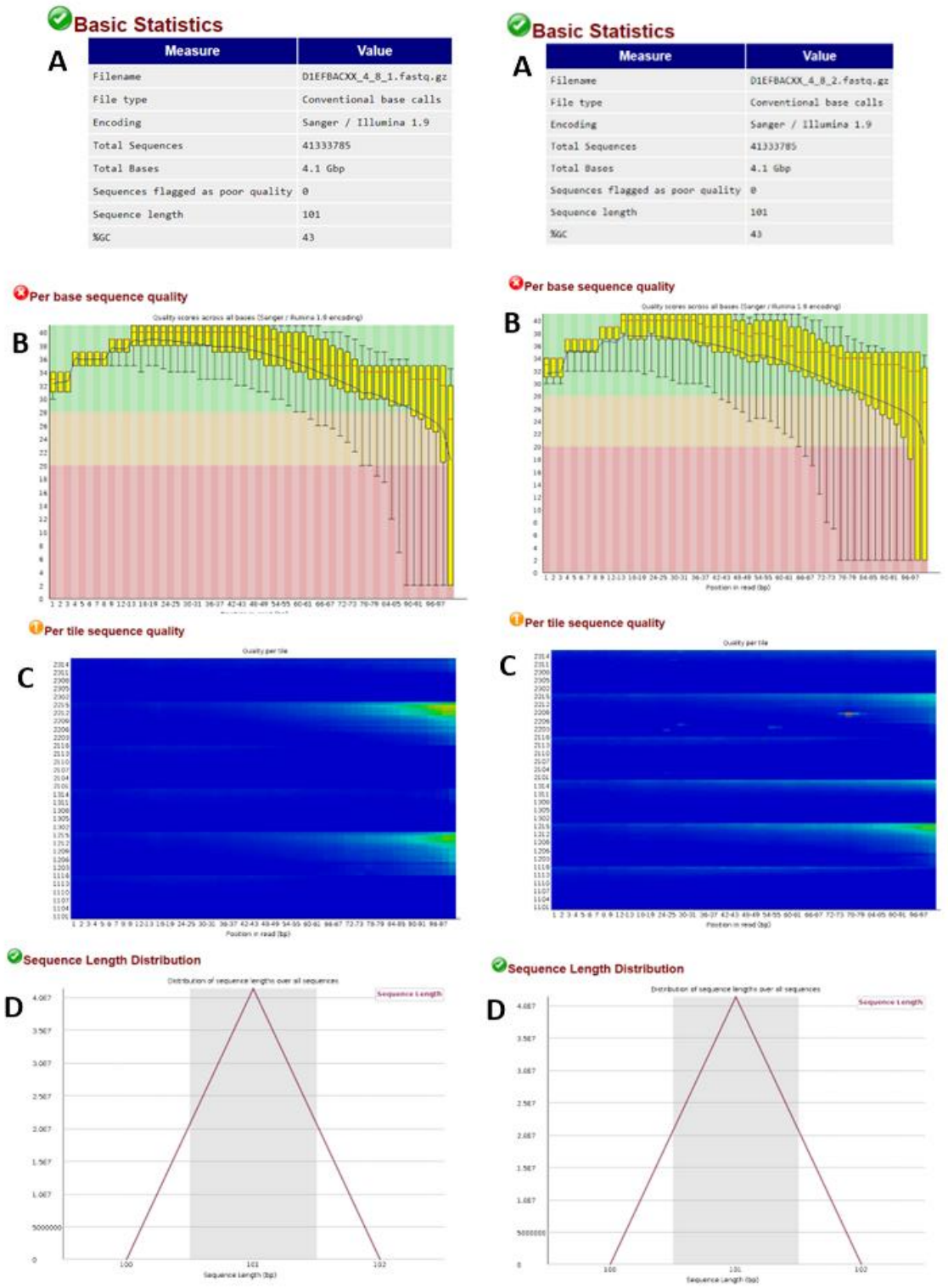

Figura 3: *Evaluación de la calidad de las secuencias crudas NGS tras el uso del programa FastQC. A. Estadísticas básicas. B. Calidad de la secuencia por base. C. Calidad de las secuencias por teselas. D. distribución de la longitud de las secuencias*

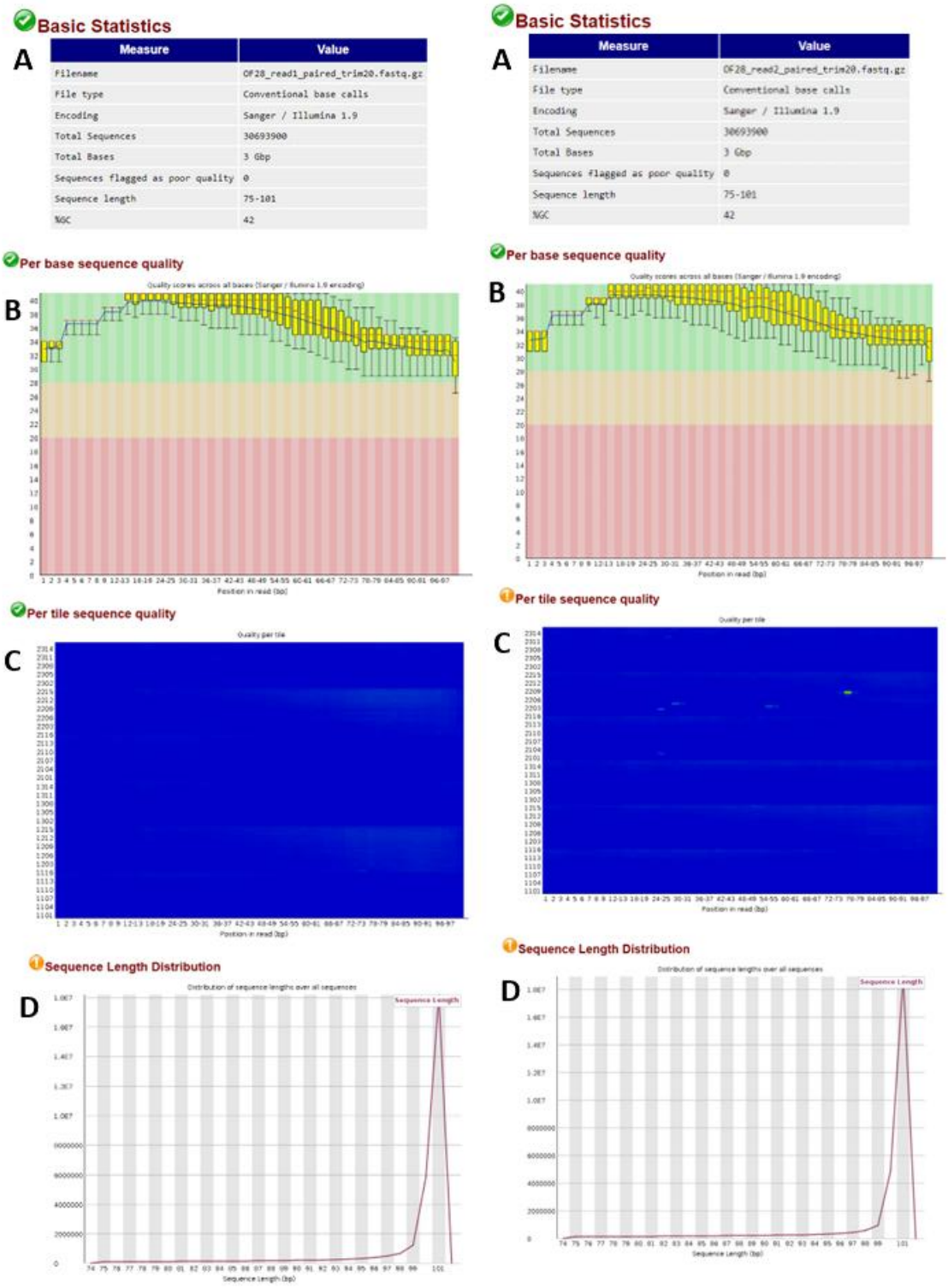

Figura 4: *Evaluación de la calidad de las secuencias NGS recortadas tras el uso del programa FastQC. A. Estadísticas básicas. B. Calidad de la secuencia por base. C. Calidad de las secuencias por teselas. D. Distribución de la longitud de las secuencias.*

4.2. Mapeado

El mapeo de las lecturas NGS contra el genoma de referencia de *D.subobscura* y sus patógenos más comunes presentó una tasa de alineamiento promedio del 97,40% para los 61.685.601 *reads* presentes en nuestro estudio. Tras la eliminación de duplicados y de los reads que han perdido su pareja durante el recorte, se obtuvo una tasa de alineamiento promedio del 96,64% para los 47.500.188 *reads* resultantes. La tabla 3 muestra las estadísticas más relevantes de este mapeo.

Características	N.º secuencias	Porcentaje
***Reads* totales**	47.500.188	100%
***Reads* mapeados**	45.904.235	96.64%
***Reads* pareados correctamente**	44.991.954	94.92%
***Reads* cuyo compañero no mapea (*singletons*)**	80.918	0.17%
***Reads* con su compañero mapeado en otro cromosoma**	401.603	0.85%
***Reads* con su compañero mapeado en otro cromosoma con mapQ>=5**	181.925	0.38%

Tabla 3: *Estadísticas del mapeo de las lecturas NGS contra el genoma de referencia de D. subobscura y sus principales patógenos tras usar la herramienta de estadísticas "Samtools flagstats".*

4.3. Identificación de los puntos de rotura de la inversión J_1

La herramienta Breakdancer proporcionó un archivo de texto del cromosoma analizado con todos los SV encontrados.

Para el cromosoma J, se obtuvo un total de 75 posibles inversiones de las cuales 45 mostraron un *score* igual o superior a 90.

Los puntos de rotura de la inversión J_1 están ubicados entre las secciones 22E y 24D en el mapa citológico J_{st} de Kunze-Mühl y Müller (45). Según los marcadores citológicos del cromosoma J cercanos al punto de rotura, *p127*

cercano al punto de corte proximal y *p222* cercano al punto de corte distal, la inversión J_1 debía abarcar como mínimo la región entre J:7.959.888 y J:10.815.307. Sin embargo, ninguna de las inversiones detectadas por el programa abarcaba la región delimitada por los marcadores citológicos. No obstante, una de las inversiones propuestas por el programa mostraba un tamaño compatible con la inversión J_1, aunque en una posición no esperada. La posición detectada fue J:12.814.771 y J: 15.108.152. Ante la posibilidad de que, al indexar la secuencia genómica, se hubiese tomado en sentido contrario a la del genoma original, se procedió a investigar esta inversión putativa en profundidad.

Marcador	Banda citológica	Posición física
p127	22E	7.959.888 – 7.961.367
p222	24D	10.814.137 – 10.815.307

Tabla 4: *Localización citológica en D. subobscura de las regiones de estudio del cromosoma J.*

Para el ensamblaje de *novo*, se utilizaron los *reads* del cromosoma J que abarcaban 3 Kb a ambos lados de los puntos de rotura indicados por Breakdancer para asegurar su localización. Se generaron 18 *contigs* de los cuales, 3 eran mayores a 1Kb.

Para comprobar las coordenadas de dichos *contigs* en el genoma de referencia, se realizó un alineamiento local en BLAST. Se comprobó que los *contigs* mapean en las mismas posiciones indicadas por Breakdancer (Figura 5). Esto indicó que la hipótesis propuesta, en la que se supuso una indexación del genoma de referencia en sentido contrario, no era válida.

Figura 5: *Resultado parcial del alineamiento en BLAST de un contig del ensamblaje de novo de reads del cromosoma J. Se aprecia que la localización de los reads corresponde a la región indicada por el programa Breakdancer.*

Ninguna de las otras inversiones propuestas por Breakdancer se ajustaba al tamaño mínimo indicado por los marcadores moleculares, por lo que se abandonó esta línea de estudio.

Tras analizar los SV de los cromosomas J, no se obtuvieron puntos de corte putativos válidos para la inversión J_1. En ocasiones esta herramienta presenta

limitaciones para detectar inversiones cromosómicas, debido a la presencia de regiones en tándem o repeticiones invertidas que dificultan el análisis. Cuando esto sucede, se requiere de análisis y filtrado adicionales o el uso de lecturas más largas que permitan mitigar esta dificultad (43).

4.4. Identificación y caracterización de los puntos de rotura de la inversión U_6.

La herramienta Breakdancer proporcionó un archivo de texto del cromosoma analizado con todos los SV encontrados.

Para el cromosoma U, se obtuvo un total de 113 posibles inversiones de las cuales 66 obtuvieron un *score* superior a 90.

Los puntos de rotura de la inversión U_6 están ubicados entre las secciones 44D y 49D en el mapa citológico U_{st} de Kunze-Mühl y Müller (45). Según los marcadores citológicos del cromosoma U cercanos al punto de rotura, *Ddc* cercano al punto de corte proximal y *dpy* cercano al punto de corte distal, la inversión U_6 debía abarcar como mínimo la región entre U:7.888.695 y U:16.420.251. El programa Breakdancer nos mostró una posible inversión entre U: 7.714.784 y U: 16.613.873, con lo que situaría a la inversión muy cercana a los marcadores elegidos. Por lo tanto, se procedió a investigar esta inversión putativa en profundidad.

Marcador	Banda citológica	Posición física
Ddc	44D	7.888.695 – 7.890.041
dpy	49D	16.420.251 – 16.503.087

Tabla 5: *Localización citológica en D. subobscura de las regiones de estudio del cromosoma U.*

El ensamblaje de *novo* se realizó con los *reads* que mapeaban cerca de los puntos obtenidos por el programa Breakdancer. En concreto, para el ensamblaje del cromosoma U se usaron los *reads* que mapearon hasta 10 Kb a ambos lados de los puntos de rotura indicados por Breakdancer para asegurar su localización. Se obtuvieron 53 *contigs* de los cuales, 7 eran mayores a 1Kb.

Para comprobar las coordenadas de dichos *contigs* en el genoma de referencia, se realizó un alineamiento local en BLAST. Al introducir el *contig* uno a BLAST, se observó que una parte del *contig* mapea en la localización U: 7.714.441-7.725.201 y otra parte en la localización U: 16.613.891-16.613.622 en dirección

contraria. La Figura 7 muestra cómo la secuencia problema *(query*) se alinea con el genoma de referencia (*subject*), en dos posiciones distintas del genoma de referencia y en direcciones opuestas.

Figura 6: *Alineamiento parcial en BLAST del primer contig obtenido tras el ensamblaje de novo de los reads del cromosoma U.*

El resultado obtenido evidencia la existencia de la inversión U_6, cuyo punto de rotura proximal se encuentra cercano a la posición U: 7.714.441 y punto de rotura distal cercano a la posición U:16.613.622 (Figura 7).

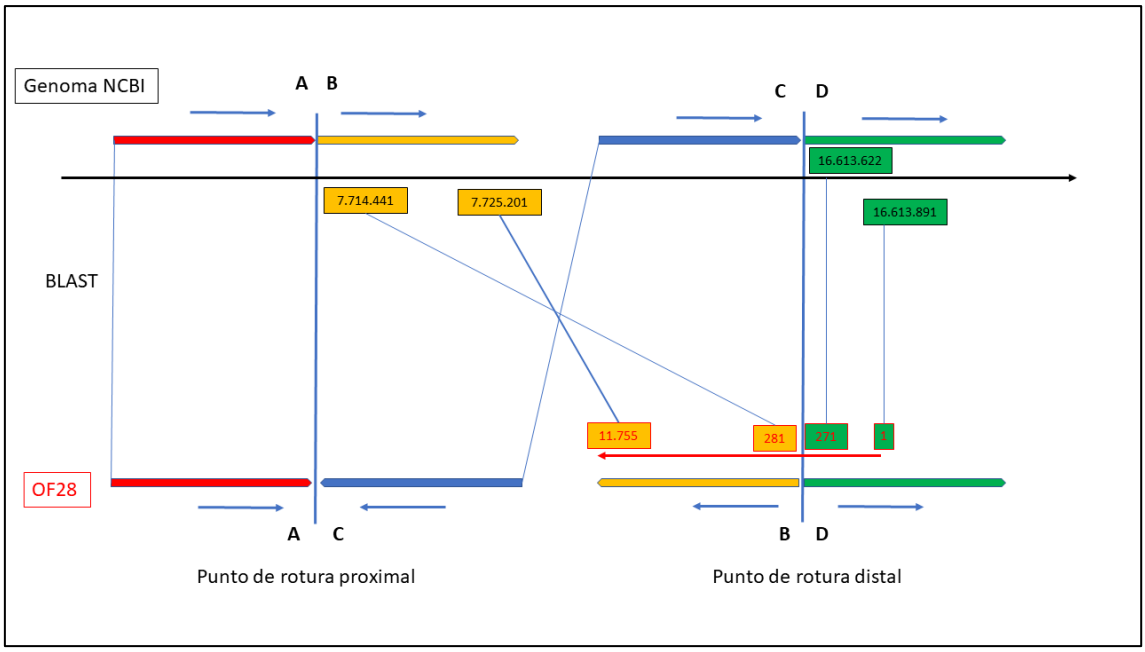

Figura 7: *Representación esquemática de la identificación del punto de rotura distal de la inversión U_6. Las líneas roja, amarilla, azul y verde respectivamente corresponden a los fragmentos alineados del genoma de referencia de D. subobscura UCBerk_Dsub_1.0 con la secuencia de análisis OF28. La flecha roja corresponde al contig número uno del ensamblaje de novo alineado con el genoma de referencia. Las líneas cruzadas unen las secuencias homólogas.*

El ensamblaje de genomas casi completos y altamente contiguos permite estudiar a fondo regiones genómicas de alta divergencia (52). El ensamblaje de *novo* de los *reads* de la cepa OF28 (U_{1+2+6}) ha permitido detectar el punto de rotura distal de la inversión U_6, ya que ha construido un *contig* que atraviesa el punto de rotura y mapea un fragmento en cada extremo de la inversión en el genoma de referencia (U_{1+2}). Sin embargo, se esperaba encontrar otro *contig* que atravesase el segundo punto de rotura, el cual, no ha sido detectado. Diversos estudios que han caracterizado puntos de rotura para la especie *D. subobscura* (34,35,53–55), han observado regiones duplicadas como consecuencia de mecanismos de rotura escalonada y posterior reparación. Cuando una inversión se forma por el mecanismo de rotura escalonada, la reparación de los extremos cohesivos forma duplicaciones en orientación opuesta flanqueante a la inversión (11). Este mecanismo crea zonas con duplicaciones que pueden dificultar el ensamblado de los *reads* de Illumina en los genomas con inversiones.

Aunque en *D. melanogaster* se ha descrito la caracterización de algunas inversiones por este sistema (14), *D. subobscura* presenta un nivel de polimorfismos mucho mayor y, en consecuencia, los mecanismos de generación de inversiones podrían ser ligeramente distintos implicando una mayor dificultad en su detección.

Por esta razón se analizaron los *contigs* en busca de zonas repetidas. La herramienta RepeatMasker detectó repeticiones simples. No se apreció

problemas por duplicaciones en los extremos de los *contigs* que expliquen dificultad en unirse entre ellos.

También se compararon con BLAST los *contigs* obtenidos en el ensamblaje de *novo* consigo mismos para comprobar si existían zonas repetidas. Se observó que el *contig* uno se alineaba con el *contig* siete en la región correspondiente a U:16.613.622-16.613.891, región D del punto de rotura distal de la inversión U_6 (Figura 8). A continuación, se alineó este *contig* siete en el genoma de referencia *D. subobscura UCBerk_Dsub_1.0* y se comprobó que esta secuencia se alineaba de forma continua en la región CD del genoma de referencia. Esto parece indicar, que como mínimo, el fragmento U:16.613.622-16.613.891 se ha duplicado a ambos lados de la inversión. La duplicación detectada a ambos lados del punto de rotura evidencia la posibilidad del origen de la inversión a través del mecanismo NHEJ (rotura escalonada). No se pudo detectar la longitud del fragmento duplicado, debido a la pequeña longitud ensamblada en la región D.

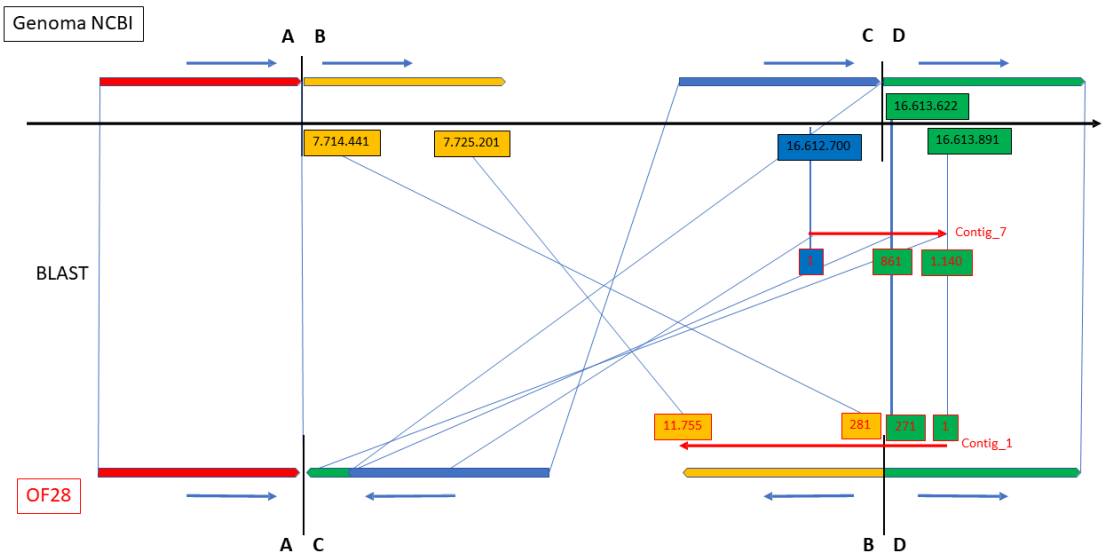

***Figura 8**: Representación esquemática de la identificación del punto de rotura distal de la inversión U_6. Las flechas rojas representan los contig 1 y 7 alineados. Corresponde a secuencias duplicadas que señalan la evidencia del origen de la inversión por NHEJ y rotura escalonada.*

A continuación, se buscó posibles inserciones del elemento transponible SGM y se detectó alineamiento parcial de SGM con varios de los *contigs* obtenidos. También se alineó con las regiones AB y CD de genoma de referencia *D. subobscura* UCBerk_Dsub_1.0. En este caso se observó la presencia de un elemento prácticamente intacto justo a 40 nucleótidos del punto de rotura proximal (AB) de la inversión U_6.

Estos resultados podrían explicar el hecho de no haber obtenido un *contig* que atravesase el punto de corte proximal de la inversión U_6. El *contig* que se esperaba encontrar como AC podría presentar secuencias SGM en el extremo A y duplicaciones resultantes del mecanismo de rotura escalonada de la región D en su extremo C, que hayan dificultado el ensamblado de *novo* de esta región.

Las secuencias genómicas repetitivas, a menudo dificultan la determinación de los puntos de corte precisos mediante secuenciación de lecturas cortas y es necesario utilizar otras estrategias que faciliten su detección como secuenciación de lecturas largas (56).

4.5. Caracterización genética de la región flanqueante de los puntos de rotura de la inversión U_6

Se estudiaron los genes codificantes que flanquean a los puntos de rotura. Para el punto de rotura proximal de la inversión U_6, en la región flanqueante fuera de la inversión se encontró el gen CG31177 que codifica a la proteína aminopeptidasa N. En la región flanqueante dentro de la inversión se encontró el gen CG14196 que codifica a la proteína transportadora de monocarboxilato 13. Ambos genes se encuentran cercanos al punto de rotura proximal, localizado en la región 7.714.441, pero no sufren ninguna rotura y, por consiguiente, no se espera la pérdida de su función (Tabla 6).

Para el punto de rotura distal de la inversión U_6, en la región flanqueante fuera de la inversión se encontró el gen *sip3* que codifica a la proteína ubiquitina-proteína ligasa E3 HDR1A con degradación asociada al retículo endoplasmático (ERAD). En la región flanqueante dentro de la inversión se encontró el gen *alphaTry* que codifica a la proteína trypsina-1. Ambos genes se encuentran cercano al punto de rotura distal, localizado en la región 16.613.622 tras el ensamblaje de *novo*, pero no sufren ninguna rotura, ya que se encuentran alejados de dicho punto (Tabla 6). Sin embargo, la región del punto de rotura que detecta Breakdancer difiere en algunos nucleótidos con respecto al encontrado tras el ensamblaje de *novo* y posterior alineamiento en BLAST. La posición detectada por Breakdancer fue 16.613.873 y dicha posición se encontraría dentro del gen *sip3*, por lo tanto, la inversión podría provocar una posible rotura del gen y la posible inactivación de este.

Nombre de la proteína/Gen	Posición física	Componente celular	Función molecular	Proceso biológico
Amilopeptidasa N (CG31177)	7.709.001-7.712.193	Citoplasma	Actividad metaloaminopeptidasa unión de péptidos unión de iones de zinc	Proceso catabólico de péptidos Proteólisis
Transportador de monocarboxilato 13 (CG14196)	7.718.659-7.721.361	Membrana	Actividad del transportador transmembrana de ácido monocarboxílico	Transporte de ácido monocarboxílico
Tripsina-1 (AlphaTry)	16.585.786-16.586.867	Espacio extracelular	Actividad serina-hidrolasas	Proteólisis
Ubiquitina proteína ligasa E3 HRD1A asociada a ERAD (Sip3)	16.613.805-16.614.348	Citoplasma Retículo endoplasmático Mebrana del retículo endoplasmático Membrana	Actividad de la proteína ligasa de la ubiquinina Unión de iones zinc Fijación de la proteína p53	Respuesta de proteína desplegada del RE Homeostasis de células madre intestinales Ubiquitinación proteica Regulación de traslación de la respuesta al estrés del RE

***Tabla 6**: Características generales en términos GO de los genes flanqueantes a la inversión U6.*

Los puntos de corte de las inversiones cromosómicas han sido ampliamente estudiados debido a que alteraciones en la estructura de los cromosomas pueden afectar a la expresión del genoma. Se han estudiado las regiones flanqueantes a las inversiones caracterizadas y ningún punto de rotura corta ningún gen. Si se observa una proteína flanqueante al punto de rotura distal, que según el programa Breakdancer podría haberse fragmentado. Sin embargo, se observan características esenciales, lo cual nos da a pensar, que la inactivación de dicho gen provocaría la letalidad recesiva.

Diversos estudios han evaluado las consecuencias fenotípicas asociadas a la rotura de genes como consecuencias de los polimorfismos por inversión y mostraron que mayoritariamente, los fenotipos no se ven gravemente afectados. Sin embargo, si estos puntos de rotura fragmentan un gen, esto puede provocar su inactivación y como consecuencia provocar letalidad recesiva o esterilidad.

No obstante, para sorpresa de muchos investigadores, existen roturas de genes que provocan pérdidas de función, pero no letalidad para el individuo (56).

Por esta razón sería interesante seguir el estudio, pues a pesar de ser una inversión poco frecuente, su estudio podría explicar la razón de su frecuencia y ayudar a entender el carácter evolutivo de la especie *D. subobscura*.

5. Conclusiones y trabajos futuros

5.1. Conclusiones

Los resultados obtenidos en el presente estudio han permitido desarrollar una *pipeline* válida para la detección de puntos de rotura de inversiones cromosómicas. Se ha conseguido caracterizar los puntos de rotura de la inversión U_6 y estudiar a nivel molecular su estructura. Por el contrario, no se pudo identificar los puntos de rotura de la inversión J_1, ya que Breakdancer no proporcionó ninguna coordenada coincidente con la región delimitada por los marcadores caracterizados previamente. A pesar de ello, se ha llevado a cabo un estudio que podrá ser de interés para la comunidad científica ya que podrá facilitar la caracterización de polimorfismos por inversión a nivel bioinformático.

En relación a los resultados obtenidos, se puede concluir lo siguiente:

1. La caracterización de los puntos de rotura de las inversiones J_1 y U_6 ha sido un proceso dificultoso debido a que *D. subobscura* es una especie altamente polimórfica. No obstante, se han obtenido resultados satisfactorios para la inversión U_6, la cual ha podido ser caracterizada molecularmente.

2. El análisis de la estructura de la inversión U_6 ha revelado que el mecanismo por el cual se ha originado la inversión fue NHEJ y rotura escalonada.

3. No se ha detectado ningún gen fragmentado como consecuencia del origen de la inversión U_6. No obstante, será necesario realizar análisis en el laboratorio que corroboren los puntos de rotura localizados y caracterizar con precisión dichos puntos para poder asegurar este hecho.

4. No se pudo caracterizar los puntos de rotura de la inversión J_1 y como consecuencia no se pudo estudiar los genes contenidos en dicha inversión y su carácter adaptativo relacionado con factores microclimáticos. Resulta complicado caracterizar a nivel molecular los puntos de rotura debido a que están compuestos por regiones altamente

repetitivas, lo cual dificulta el proceso de ensamblaje utilizando tecnologías de secuenciación de lectura corta.

5.2. Trabajos futuros

No se ha conseguido obtener ningún punto de rotura para el cromosoma J, pero se plantean posibles estudios futuros:

- Repetir el procedimiento bioinformático, pero utilizando otras líneas secuenciadas, para comprobar si el problema reside en la secuenciación de la línea OF28.

- Repetir el procedimiento bioinformático, pero utilizando datos de ultrasecuenciación de lecturas largas que ayuden al ensamblado de regiones conflictivas.

Solo se consiguió obtener uno de los puntos de rotura para la inversión U_6 en la ordenación invertida, pero se plantean estudios posteriores para comprobar la validez del punto de corte detectado:

- Realizar un ensayo de PCR donde diseñar oligonucleótidos flanqueando los puntos putativos reflejados por Breakdancer para intentar amplificar la secuencia flanqueada por PCR usando secuencias conservadas entre el genoma de referencia y los *contigs* que se han obtenido al ensamblar de *novo* los *reads* de Illumina y usando distintas parejas.

5.3. Seguimiento de la planificación

La planificación del estudio no se llevó a cabo tal y como se había planteado al comienzo del estudio. El calendario se retrasó debido a la falta de bibliografía disponible para la realización del protocolo bioinformático. Al inicio del estudio, se utilizó como modelo el artículo publicado por Corbett-Detig y colaboradores (14), pero al iniciar la búsqueda de la *pipeline* para el proceso de localización de puntos de rotura, esta no se encontraba disponible, por lo tanto, se tuvo que buscar otros artículos relacionados (44). Se encontraron varios programas de detección de SNVs. Finalmente se utilizó Breakdancer, una herramienta para la detección de variaciones estructurales, que presentaba buenos resultados en la detección de inversiones cromosómicas y una fácil instalación. La búsqueda de programas alternativos retrasó el calendario.

El resto del proceso de caracterización de inversiones se siguió como el planteado al comienzo del proyecto, sin embargo, no se obtuvieron los resultados esperados y, por lo tanto, se plantearon tareas nuevas no planteadas en la

planificación inicial. Se utilizó la herramienta RepeatMasker y se realizó un alineamiento con el elemento transponible SGM en busca de secuencias repetidas que explicasen los motivos por lo que no se obtuvieron los resultados esperados.

A pesar de estos cambios, el estudio se ha llevado a cabo correctamente, obteniéndose resultados prometedores e interesantes, aunque no fueran los esperados.

6. Glosario

BLAST – *Basic Local Alignment Search Tool*. Herramienta básica de búsqueda de alineación local.

CDS – *Coding Sequence*. Región de codificación de un gen son regiones de ADN que codifican a proteínas.

Contig – Conjunto de secuencias que se superponen de tal forma que de forma continua representan una región genómica.

D. subobscura- *Drosophila subobscura*. Especie de mosca de la fruta de la familia Drosophilidae

NAHR – *Non-allelic homologous recombination*. Recombinación de extremos no homólogos.

NCBI – *The National Center for Biotechnology Information*. Centro Nacional de Información Biotecnológica.

NGS – *Next Generation Sequencing*. Término para identificar una nueva tecnología de secuenciación de última generación.

NHEJ – *Non-homologous end joining*. Unión de extremos no homólogos.

OF28 – Línea homocariotípica de *D. subobscura* para todos sus cromosomas con ordenaciones A_{st}, J_1, $U_{\underline{1+2+6}}$, E_{st} y O_{st}.

Read – Lectura en español, es una secuencia de pares de bases que corresponde a la totalidad o a parte de un fragmento de ADN, formado mediante secuenciación.

Read paired-end – Secuencia de extremos emparejados son secuencias que se obtienen mediante un mecanismo de secuenciación por el cual se secuencian ambos extremos de un fragmento y genera datos de secuenciación alineables de alta calidad.

SV – *Structural Variation*. Variación estructural.

UCBerk_Dsub_1.0 – Ensamblado de *Drosophila subobscura* a nivel de scaffolds de la Universidad de California en Berkeley.

7. Bibliografía

1. Kapun M, Flatt T. The adaptive significance of chromosomal inversion polymorphisms in *Drosophila melanogaster*. Mol Ecol [Internet]. 2019 Mar 1 [cited 2023 Mar 21];28(6):1263–82. Available from: https://onlinelibrary.wiley.com/doi/full/10.1111/mec.14871
2. Puig Giribets M, García Guerreiro MP, Santos M, Ayala FJ, Tarrío R, Rodríguez-Trelles F. Chromosomal inversions promote genomic islands of concerted evolution of Hsp70 genes in the *Drosophila subobscura* species subgroup. Mol Ecol. 2019 Mar 1;28(6):1316–32.
3. Simões P, Calabria G, Picão-Osório J, Balanyà J, Pascual M. The Genetic Content of Chromosomal Inversions across a Wide Latitudinal Gradient. PLoS One [Internet]. 2012 Dec 18 [cited 2023 Mar 21];7(12):e51625. Available from: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0051625
4. Hoffmann AA, Rieseberg LH. Revisiting the Impact of Inversions in Evolution: From Population Genetic Markers to Drivers of Adaptive Shifts and Speciation? Annu Rev Ecol Evol Syst [Internet]. 2008 Dec 12 [cited 2023 Mar 21];39:21. Available from: /pmc/articles/PMC2858385/
5. Orengo DJ, Puerma E, Aguadé M. The molecular characterization of fixed inversions breakpoints unveils the ancestral character of the *Drosophila guanche* chromosomal arrangements. Sci Rep [Internet]. 2019 Dec 1 [cited 2023 Mar 21];9(1). Available from: /pmc/articles/PMC6368638/
6. Prevosti A, Ribo G, Serra L, Aguade M, Balara J, Monclus M, et al. Colonization of America by Drosophila subobscura: Experiment in natural populations that supports the adaptive role of chromosomal-inversion polymorphism (evolution/natural selection/adaptation/clines) [Internet]. Vol. 85, Proc. Natl. Acad. Sci. USA. 1988. Available from: https://www.pnas.org
7. Orengo D, Puerma E, Aguadé M. A new spontaneous chromosomal inversion in a classical laboratory strain of *Drosophila subobscura*. 2015.
8. Karageorgiou C, Tarrío R, Rodríguez-Trelles F. The Cyclically Seasonal Drosophila subobscura Inversion O7 Originated From Fragile Genomic Sites and Relocated Immunity and Metabolic Genes. Front Genet. 2020 Oct 9;11.
9. Orengo DJ, Puerma E, Cereijo U, Aguadé M. The molecular genealogy of sequential overlapping inversions implies both homologous chromosomes of a heterokaryotype in an inversion origin. Sci Rep. 2019 Dec 1;9(1).
10. Torres OC. Dinámica evolutiva de las reordenaciones cromosómicas y coincidencia de los puntos de rotura: análisis molecular de las inversiones fijadas en el cromosoma 2 de <i>Drosophila Buzzatii<i/>. 2010.
11. Ranz JM, Maurin D, Chan YS, Von Grotthuss M, Hillier LDW, Roote J, et al. Principles of genome evolution in the *Drosophila melanogaster* species group. PLoS Biol. 2007 Mar;5(6):1366–81.
12. Corbett-Detig RB, Said I, Calzetta M, Genetti M, McBroome J, Maurer NW, et al. Fine-Mapping Complex Inversion Breakpoints and Investigating Somatic Pairing in the *Anopheles gambiae* Species Complex Using Proximity-Ligation Sequencing. Genetics [Internet]. 2019 [cited 2023 Mar 21];213(4):1495. Available from: /pmc/articles/PMC6893396/
13. Ye J, Xiao Z, Li C, Wang F, Liao J, Fu J, et al. Past climate change and recent anthropogenic activities affect genetic structure and population

demography of the greater long-tailed hamster in northern China. Integr Zool [Internet]. 2015 Sep 1 [cited 2023 Apr 5];10(5):482–96. Available from: https://onlinelibrary.wiley.com/doi/full/10.1111/1749-4877.12150
14. Corbett-Detig RB, Cardeno C, Langley CH. Sequence-based detection and breakpoint assembly of polymorphic inversions. Genetics. 2012 Sep 1;192(1):131–7.
15. Mestres F, Zivanovic G, Arenas C. How do organisms adapt to climate change? Chromosomal inversions in 'Drosophila subobscura': The case of serbian populations. Metode. 2016;(6):46–53.
16. Galludo M, Canals J, Pineda-Cirera L, Esteve C, Rosselló M, Balanyà J, et al. Climatic adaptation of chromosomal inversions in *Drosophila subobscura*. Genetica. 2018 Oct 1;146(4–5):433–41.
17. Zivanovic G, Arenas C, Mestres F. Short- and long-term changes in chromosomal inversion polymorphism and global warming: *Drosophila subobscura* from the balkans. Isr J Ecol Evol. 2012 Jan 1;58(4):289–311.
18. Van Verk MC, Hickman R, Pieterse CMJ, Van Wees SCM. RNA-Seq: revelation of the messengers. Trends Plant Sci. 2013 Apr 1;18(4):175–9.
19. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics [Internet]. 2014 Aug 1 [cited 2023 Mar 21];30(15):2114–20. Available from: https://academic.oup.com/bioinformatics/article/30/15/2114/2390096
20. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. 2013 Mar 16; Available from: http://arxiv.org/abs/1303.3997
21. McBroome J, Liang D, Corbett-Detig R. Fine-Scale Position Effects Shape the Distribution of Inversion Breakpoints in *Drosophila melanogaster*. Genome Biol Evol [Internet]. 2020 [cited 2023 Mar 21];12(8):1378. Available from: /pmc/articles/PMC7487137/
22. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics [Internet]. 2009 Aug 8 [cited 2023 Mar 21];25(16):2078. Available from: /pmc/articles/PMC2723002/
23. Breakdancer. https://breakdancer.sourceforge.net/ (accedido abr 28, 2023).
24. Prjibelski A, Antipov D, Meleshko D, Lapidus A, Korobeynikov A. Using SPAdes De Novo Assembler. Curr Protoc Bioinformatics [Internet]. 2020 Jun 1 [cited 2023 Jun 9];70(1):e102. Available from: https://onlinelibrary.wiley.com/doi/full/10.1002/cpbi.102
25. Mount DW. Using the Basic Local Alignment Search Tool (BLAST). Cold Spring Harb Protoc [Internet]. 2007 Jul 1 [cited 2023 Jun 25];2007(7):pdb.top17. Available from: http://cshprotocols.cshlp.org/content/2007/7/pdb.top17.full
26. Puerma E, Orengo DJ, Cruz F, Gómez-Garrido J, Librado P, Salguero D, et al. The High-Quality Genome Sequence of the Oceanic Island Endemic Species *Drosophila guanche* Reveals Signals of Adaptive Evolution in Genes Related to Flight and Genome Stability. Genome Biol Evol [Internet]. 2018 Aug 1 [cited 2023 Mar 21];10(8):1956–69. Available from: https://academic.oup.com/gbe/article/10/8/1956/5045874
27. Corbett-Detig RB, Hartl DL. Population Genomics of Inversion Polymorphisms in *Drosophila melanogaster*. PLoS Genet [Internet]. 2012

Dec [cited 2023 Mar 21];8(12):1003056. Available from: /pmc/articles/PMC3527211/
28. Simões P, Pascual M. Patterns of geographic variation of thermal adapted candidate genes in *Drosophila subobscura* sex chromosome arrangements. BMC Evol Biol. 2018 Apr 24;18(1).
29. Orengo DJ, Prevosti A. Temporal Changes in Chromosomal Polymorphism of *Drosophila subobscura* Related to Climatic Changes [Internet]. Vol. 50, Source: Evolution. 1996. Available from: https://www.jstor.org/stable/2410676
30. Orengo DJ, Puerma E, Aguadé M. Monitoring chromosomal polymorphism in *Drosophila subobscura* over 40 years. Entomol Sci. 2016 Jul 1;19(3):215–21.
31. Rozas J, Segarra C, Ribó G, Aguadé M. Molecular Population Genetics of the rp49 Gene Region in Different Chromosomal Inversions of *Drosophila subobscura* [Internet]. 1999. Available from: https://academic.oup.com/genetics/article/151/1/189/6032938
32. Zivanovic G, Milanovic M, Andjelkovic M. Chromosomal inversion polymorphism of *Drosophila subobscura* populations from Jastrebac Mountain shows temporal and habitat-related changes. J Zool Syst Evol Research. 1995;33:81–3.
33. Balanyà J, Solé E, Oller JM, Sperlich D, Serra L. Long-term changes in the chromosomal inversion polymorphism of *Drosophila subobscura*. II. European populations*. Journal of Zoological Systematics and Evolutionary Research [Internet]. 2004 Aug 1 [cited 2023 May 2];42(3):191–201. Available from: https://onlinelibrary.wiley.com/doi/full/10.1111/j.1439-0469.2004.00274.x
34. Papaceit M, Segarra C, Aguadé M. Structure and population genetics of the breakpoints of a polymorphic inversion in *Drosophila subobscura*. Evolution (N Y). 2013 Jan;67(1):66–79.
35. Puerma E, Orengo DJ, Salguero D, Papaceit M, Segarra C, Aguadé M. Characterization of the breakpoints of a polymorphic inversion complex detects strict and broad breakpoint reuse at the molecular level. Mol Biol Evol. 2014;31(9):2331–41.
36. Karageorgiou C, Gámez-Visairas V, Tarrío R, Rodríguez-Trelles F. Long-read based assembly and synteny analysis of a reference *Drosophila subobscura* genome reveals signatures of structural evolution driven by inversions recombination-suppression effects. BMC Genomics 2019 20:1 [Internet]. 2019 Mar 18 [cited 2023 Apr 25];20(1):1–21. Available from: https://link.springer.com/articles/10.1186/s12864-019-5590-8
37. Bracewell R, Chatla K, Nalley MJ, Bachtrog D. Dynamic turnover of centromeres drives karyotype evolution in drosophila. Elife. 2019 Sep 1;8.
38. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics [Internet]. 2009 Jul 15 [cited 2023 May 10];25(14):1754–60. Available from: https://academic.oup.com/bioinformatics/article/25/14/1754/225615
39. Li H, Durbin R. Fast and accurate long-read alignment with Burrows–Wheeler transform. Bioinformatics [Internet]. 2010 Mar 1 [cited 2023 May

1];26(5):589–95. Available from: https://academic.oup.com/bioinformatics/article/26/5/589/211735
40. Kapun M, Fabian DK, Goudet J, Flatt T. Genomic Evidence for Adaptive Inversion Clines in *Drosophila melanogaster*. Mol Biol Evol [Internet]. 2016 May 1 [cited 2023 Apr 28];33(5):1317–36. Available from: https://academic.oup.com/mbe/article/33/5/1317/2579921
41. Moran NA, Sloan DB. The Hologenome Concept: Helpful or Hollow? PLoS Biol [Internet]. 2015 Dec 4 [cited 2023 Apr 28];13(12):e1002311. Available from: https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1002311
42. Sakib MN, Tang J, Zheng WJ, Huang CT. Improving Transmission Efficiency of Large Sequence Alignment/Map (SAM) Files. PLoS One [Internet]. 2011 Dec 2 [cited 2023 Jun 20];6(12):e28251. Available from: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0028251
43. Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, et al. BreakDancer: An algorithm for high-resolution mapping of genomic structural variation. Nat Methods. 2009;6(9):677–81.
44. Fan X, Abbott TE, Larson D, Chen K. BreakDancer: Identification of genomic structural variation from paired-end read mapping. Curr Protoc Bioinformatics. 2014;(SUPPL.45).
45. Kunze-Mühl E, Müller E. Weitere Untersuchungen über die chromosomale Struktur und die natürlichen Strukturtypen von *Drosophila subobscura* Coll. Chromosoma. 1958;9(6):559–70.
46. Pratdesaba R. Anàlisi multilocus del polimorfisme nucleotídic al llarg del cromosoma J de *Drosophila subobscura* [Internet]. 2014 [cited 2023 Jun 9]. Available from: www.tdx.cat
47. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. Journal of Computational Biology. 2012 May 1;19(5):455–77.
48. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: Quality assessment tool for genome assemblies. Bioinformatics. 2013 Apr 15;29(8):1072–5.
49. Quast. Quality Assessment Tool for Genome Assemblies. https://quast.sourceforge.net/quast. (accedido may 31, 2023).
50. RepeatMasker. https://www.repeatmasker.org/ (accedido, jun 3, 2023).
51. Miller WJ, Nagel A, Bachmann J, Bachmann L. Evolutionary Dynamics of the SGM Transposon Family in the *Drosophila obscura* Species Group. Mol Biol Evol [Internet]. 2000;17(11):1597–609. Available from: https://academic.oup.com/mbe/article/17/11/1597/1167645
52. Christmas MJ, Wallberg A, Bunikis I, Olsson A, Wallerman O, Webster MT. Chromosomal inversions associated with environmental adaptation in honeybees. Mol Ecol [Internet]. 2019 Mar 1 [cited 2023 Jun 5];28(6):1358–74. Available from: https://onlinelibrary.wiley.com/doi/full/10.1111/mec.14944
53. Puerma E, Orengo DJ, Aguadé M. The origin of chromosomal inversions as a source of segmental duplications in the Sophophora subgenus of Drosophila. Scientific Reports 2016 6:1 [Internet]. 2016 Jul 29 [cited 2023

May 2];6(1):1–8. Available from: https://www.nature.com/articles/srep30715
54. Puerma E, Orengo DJ, Aguadé M. Inversion evolutionary rates might limit the experimental identification of inversion breakpoints in non-model species. Sci Rep. 2017 Dec 1;7(1).
55. Orengo DJ, Puerma E, Papaceit M, Segarra C, Aguadé M. A molecular perspective on a complex polymorphic inversion system with cytological evidence of multiply reused breakpoints. Heredity (Edinb) [Internet]. 2015 Jun 14 [cited 2023 Jun 5];114(6):610–8. Available from: https://pubmed.ncbi.nlm.nih.gov/25712227/
56. Miller DE, Kahsai L, Buddika K, Dixon MJ, Kim BY, Calvi BR, et al. Identification and characterization of breakpoints and mutations on *Drosophila melanogaster* balancer chromosomes. G3: Genes, Genomes, Genetics. 2020 Nov 1;10(11):4271–85.