Análisis del contenido génico de dos inversiones polimórficas del cromosoma U en *Drosophila subobscura*

Jorge Martínez Jordán

MU Bioinf. y Bioest.
Análisis de datos ómicos

Nombre Tutor/a de TF
Dorcas Orengo Ferriz

Profesor/a responsable de la asignatura
David Merino Arranz

Fecha Entrega
Enero 2023

Universitat Oberta de Catalunya

UNIVERSITAT DE BARCELONA

Agradecimientos

A Dorcas por su gran labor como docente, su implicación y su ayuda estos últimos meses.
A Pedro por su apoyo incondicional y por animarme y acompañarme siempre.

FICHA DEL TRABAJO FINAL

Título del trabajo:	Análisis del contenido génico de dos inversiones polimórficas del cromosoma U en *Drosophila subobscura*
Nombre del autor:	Jorge Martínez Jordán
Nombre del consultor/a:	Dorcas Orengo Ferriz
Nombre del PRA:	David Merino Arranz
Fecha de entrega (mm/aaaa):	01/2023
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	*Drosophila subobscura, inversiones cromosómicas, Ontología genética*

Resumen del Trabajo

La especie *Drosophila subobscura* se caracteriza por presentar un elevado grado de polimorfismo cromosómico por inversiones, cuya frecuencia se ha relacionado con las condiciones ambientales. Este hecho sugiere que este tipo de polimorfismos constituyen un mecanismo genético de adaptación al ambiente. Así, para el cromosoma U, se ha observado una mayor frecuencia de la ordenación U_{ST} en climas fríos y de la ordenación U_{1+2} en climas cálidos. Se especula que estas inversiones han fijado conjuntos de genes que contienen alelos coadaptados a las condiciones climáticas. Sin embargo, el contenido génico de éstas no ha sido estudiado.

En este contexto, se desarrolla este trabajo, con el objetivo de localizar y caracterizar los puntos de rotura de las inversiones U_1 y U_2 en *D. subobscura* y rastrear el contenido génico de éstas mediante un análisis de ontología genética (GO).

Los resultados obtenidos han permitido determinar que el mecanismo originario de la inversión U_2 es la rotura y posterior reparación por unión de extremos no homólogos. Para la inversión U_1 no ha sido posible establecer el mecanismo por el cual se generó.

El análisis del contenido génico de las inversiones muestra un enriquecimiento de genes relacionados con el metabolismo lipídico y morfosíntesis de nefronas en U_1, y de genes relacionados con el metabolismo de ácidos nucleicos en U_2.

Adicionalmente, la comparación del genoma de *D. subobscura* con la especie cercana *D. guanche* ha desvelado una serie de microinversiones, probablemente consecuencia de artefactos en el ensamblado de esta última.

Abstract

The species *Drosophila subobscura* exhibits a rich inversion polymorphism, the frequency of which has been related to environmental conditions. This fact suggests that this type of polymorphisms constitute a genetic mechanism of adaptation to different climatic conditions. Thus, for the U chromosome, a higher frequency of the U_{ST} arrangement has been observed in cold climates whereas a higher frequency of the U_{1+2} arrangement has been observed in warm climates. It is hypothesized that these inversions have fixed sets of genes containing alleles coadapted to climatic conditions. However, the genetic content of these inversions has not been studied.

In this context, this work was carried out with the aim of locating and characterizing inversion breakpoints for U_1 and U_2 in *D. subobscura* and examining the gene content of these inversions performing a gene ontology (GO) enrichment analysis

The results have allowed us to determine that U_2 was originated through non-homologous end joining. Nevertheless, the results did not allow the establishment of the mechanism that originated the U_1 inversion.

Analysis of the genetic content of the inversions shows an enrichment of genes related to lipid metabolism and nephron morphosynthesis in U_1, and genes related to nucleic acids metabolism in U_2.

Additionally, comparison of the genome of *D. subobscura* with the closely related species *D. guanche* has revealed a series of microinversions, probably caused by artifacts in the assembly of the latter.

Índice

Lista de figuras

Lista de Tablas

1. Introducción

1.1. Contexto y justificación del Trabajo

Las diferentes especies del género *Drosophila* son utilizadas como organismo modelo en la investigación biológica y genética evolutiva. Dentro del género, la especie *D. subobscura* se caracteriza por presentar un elevado grado de polimorfismo cromosómico por inversiones que afecta a todos sus cromosomas (1). La frecuencia de estas inversiones se ha relacionado con las condiciones ambientales, específicamente con gradientes de temperatura espaciales y temporales, existiendo así diferentes clinas latitudinales en sus áreas de distribución (2,3), fluctuaciones estacionales en la frecuencia de estas (4) y también variaciones temporales a más largo plazo paralelas con cambios del clima (5,6). Estos hechos sugieren que este tipo de polimorfismos constituyen un mecanismo genético de adaptación al ambiente (3,6,7).

La cepa Küsnacht de *D. subobscura* fue la primera en ser caracterizada y la ordenación de sus cromosomas se designó como estándar (ST). A partir de ahí, las diferentes ordenaciones descubiertas se designaron con un subíndice numérico en orden de descubrimiento (8). En el caso del cromosoma U, se ha visto que la frecuencia de las distintas ordenaciones está relacionada con la temperatura, las lluvias y la humedad, siendo más frecuente la ordenación U_{ST} en climas fríos y húmedos mientras que existe una mayor frecuencia de las ordenaciones U_{1+2} y $U_{\underline{1+8}+\underline{2}}$ en climas más cálidos.

Se han propuesto dos mecanismos para el origen de las inversiones cromosómicas. El primer mecanismo es la recombinación homóloga no alélica (NAHR), el cual se genera por la existencia de duplicaciones que permiten esta recombinación entre los extremos de los fragmentos que quedarán invertidos. El segundo mecanismo consiste en la rotura y reparación mediante unión de extremos no homólogos (NHEJ), pudiendo producirse en el corte un extremo "limpio" o un extremo cohesivo (quedando los extremos de forma escalonada). Este último mecanismo es el más prevalente en la especie *D. subobscura* y genera duplicaciones en los extremos del estado invertido (9) Estas mutaciones pueden alterar la estructura o el nivel de expresión de genes localizados en los puntos de rotura de las inversiones o en las regiones adyacentes a estos. De manera indirecta, provocan una reducción de la recombinación entre cromosomas homólogos de distinta ordenación, favoreciendo la diferenciación genética de las distintas ordenaciones cromosómicas (8,10).

La caracterización de los puntos de rotura de las inversiones permite determinar el mecanismo por el que han sido originados, permitiendo conocer su significado evolutivo. En este sentido y usando técnicas de *chromosome walking* e hibridación *in situ*, se han podido determinar los puntos de rotura de algunas inversiones polimórficas de *D. subobscura* (11–13), así como de las inversiones fijadas entre *D. subobscura* y su

especie gemela *D. guanche* (14). Además, comparando los genomas completos de estas dos especies se han determinado los puntos de rotura de las inversiones U_1 y U_2, puesto que *D. guanche* es monomórfica para la ordenación U_{1+2} (15). Sin embargo, aunque se especula que estas inversiones han fijado conjuntos de genes que contienen alelos coadaptados a las condiciones climáticas (6), el contenido de las regiones invertidas aún no ha sido estudiado.

En el presente trabajo se propone el estudio comparativo del contenido del cromosoma U de dos especies cercanas, *D. subobscura* (de amplia distribución geográfica) y *D. guanche* (endémica de las Islas Canarias). Para ello, se realizará el análisis bioinformático de los genomas de referencia disponibles *online*, que en el caso de *D. guanche* presenta la ordenación U_{1+2} (16). En el caso de *D. subobscura*, se dispone de un ensamblado de alta calidad a nivel de cromosoma, pero se desconoce la ordenación cromosómica secuenciada. En primer lugar, se realizará un análisis de sintenia para determinar su ordenación cromosómica. Posteriormente se realizará un análisis comparativo de los genomas de ambas especies para localizar y caracterizar los puntos de rotura de las inversiones existentes en el cromosoma U del genoma de *D. guanche*. Finalmente, se propone la realización de un estudio de ontología genética (GO) para encontrar genes candidatos de selección para ambientes fríos en las inversiones U_1 y U_2 que presenta el genoma disponible online de *D. guanche.*

1.2. Objetivos del Trabajo

1.2.1. Objetivos generales

1. Investigar los genomas disponibles *online* de *D. subobscura* para obtener uno que presente la ordenación U_{ST} que permita comparar con el genoma disponible de *D. guanche* que presenta las inversiones U_1 y U_2.

2. Encontrar los puntos de rotura de las inversiones U_1 y U_2, anotar sus características y rastrear el contenido génico de las inversiones.

1.2.2. Objetivos específicos

1. Investigar los genomas disponibles online de *D. subobscura* para obtener uno que presente la ordenación U_{ST} que permita comparar con el genoma disponible de *D. guanche* que presenta las inversiones U_1 y U_2.

    1.1. Recopilar los datos de los genomas de *D. subobscura* y de *D. guanche* y sus anotaciones.
    1.2. Hacer un estudio de los programas para identificar la ordenación de los cromosomas.
    1.3. Identificar la ordenación del cromosoma U en los genomas disponibles de *D. subobscura* mediante comparación con el genoma de *D. guanche* de ordenación conocida.

2. Encontrar los puntos de rotura de las inversiones U_1 y U_2, anotar sus características y rastrear el contenido génico de las inversiones.

    2.1. Caracterizar los puntos de rotura de las inversiones presentes en el cromosoma U de *D. guanche* y las regiones adyacentes a estos.
    2.2. Hacer un análisis de GO para determinar los genes que podrían estar implicados en el carácter adaptativo de las inversiones.

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

La década de 2010-2019 ha sido la más calurosa registrada hasta la fecha (17). El calentamiento global impulsó los estudios en las poblaciones de *Drosophila* y los posibles efectos de este fenómeno en la frecuencia de inversiones genéticas en esta especie. Aunque no es el único factor relevante para la variación de la frecuencia de las distintas ordenaciones cromosómicas, se han observado gradientes en la frecuencia de inversiones en relación con gradientes de temperatura.

La tendencia de aumento de temperatura global puede estar afectando a las poblaciones de *Drosophila*. Un mayor entendimiento de los genes contenidos en las inversiones cromosómicas y su posible implicación en la adaptación climática de la especie, podría arrojar luz en el impacto que tiene el calentamiento global en la genética poblacional de la especie y, a pesar de no contribuir directamente en las metas de los objetivos de desarrollo sostenible establecidos por la Organización de las Naciones Unidas, sí que constituye un argumento más a favor de la necesidad de abordar urgentemente la emergencia climática.

El impacto del presente trabajo en aspectos ético-sociales o de diversidad es mínimo, debido a su carácter científico-técnico. No obstante, se han tenido en cuenta todas las aportaciones con independencia del género, raza, etnia, o condición de la autora o autor.

1.4. Enfoque y método seguido

El trabajo se dividirá en dos partes diferenciadas: una primera sección dedicada a determinar la ordenación cromosómica de los genomas disponibles de *D. subobscura* y una segunda sección centrada en la caracterización de los puntos de rotura de las inversiones U_1 y U_2 y el rastreo del contenido génico de estas.

En primer lugar, se realizará un estudio de colinearidad (estudio genético comparativo para determinar la ordenación cromosómica) entre los cromosomas U de los genomas de *D. subobscura* y *D. guanche*. Se comenzará con un estudio de los programas disponibles para realizar esta clase de estudios. De los genomas disponibles, se comenzará con la determinación de la ordenación del cromosoma U de la cepa de *D. subobscura* de la Universidad de California (publicado en NCBI el 27-08-2019 UC Berkeley), ensamblado a nivel de cromosoma y de acceso público y que es el genoma de

referencia de la especie (18). En caso de presentar la ordenación U_{ST} podríamos usarlo para comparar con el genoma de *D. guanche*. En caso de que el genoma de referencia de *D. subobscura* no presente la ordenación U_{ST} se buscará entre el resto de los genomas accesibles *on-line* para obtener uno que presente la ordenación U_{ST} que permita comparar con el genoma disponible de *D. guanche* que presenta las inversiones U_1 y U_2.

En segundo lugar, una vez obtenido un genoma de *D. subobscura* que contenga la ordenación U_{ST}, se procederá a la anotación de los puntos de rotura y las regiones adyacentes de aproximadamente 5Kb a lado y lado de modo similar a como lo hacen Orengo y colaboradores(14). Posteriormente, se realizará un análisis de los genes contenidos en las regiones invertidas mediante un análisis de GO, en búsqueda de aquellos genes que puedan estar implicados en su carácter adaptativo a los factores climáticos.

1.5. Planificación del trabajo

A. Tareas:

Descripción	**Fecha de inicio**	**Fecha de fin**
Definición y plan de trabajo - PEC1	**28/09/2022**	**17/10/2022**
Búsqueda y estudio de Bibliografía	28/09/2022	7/10/2022
Definición de la metodología a seguir y plan de trabajo	7/10/2022	10/10/2022
Redacción de la introducción y estado del arte	11/10/2022	17/10/2022
Desarrollo del trabajo. Fase 1. PEC2	**18/10/2022**	**21/11/2022**
Búsqueda y descarga de los genomas y sus anotaciones	18/10/2022	19/10/2022
Revisión de programas que permitan identificar la ordenación cromosómica	19/10/2022	28/10/2022
Instalación y puesta a punto del programa seleccionado	28/10/2022	31/10/2022
Extracción del genoma del cromosoma U para su estudio	1/11/2022	2/11/2022
Determinación de la ordenación del cromosoma U en *D. subobscura*	2/11/2022	4/11/2022
Localización y caracterización de los puntos de rotura de las inversiones U_1 y U_2	6/11/2022	16/11/2022
Documentación de todo el trabajo realizado para la memoria	16/11/2022	21/11/2022
Desarrollo del trabajo. Fase 2. PEC3	**22/11/2022**	**24/12/2022**
Correcciones y *Feedback* de la PEC 2	22/11/2022	28/11/2022

Análisis del contenido génico y selección de genes con carácter adaptativo	28/11/2022	7/12/2022
Análisis de resultados y discusión	7/12/2022	23/12/2022
Cierre de la memoria y de la presentación. **PEC 4**	**27/12/2022**	**15/01/2023**
Revisión de la memoria y preparación de la presentación	27/12/2022	11/01/2023
Grabación de la exposición oral	11/01/2023	13/01/2023
Defensa pública. PEC5	**23/01/2023**	**03/02/2023**

B. Calendario:

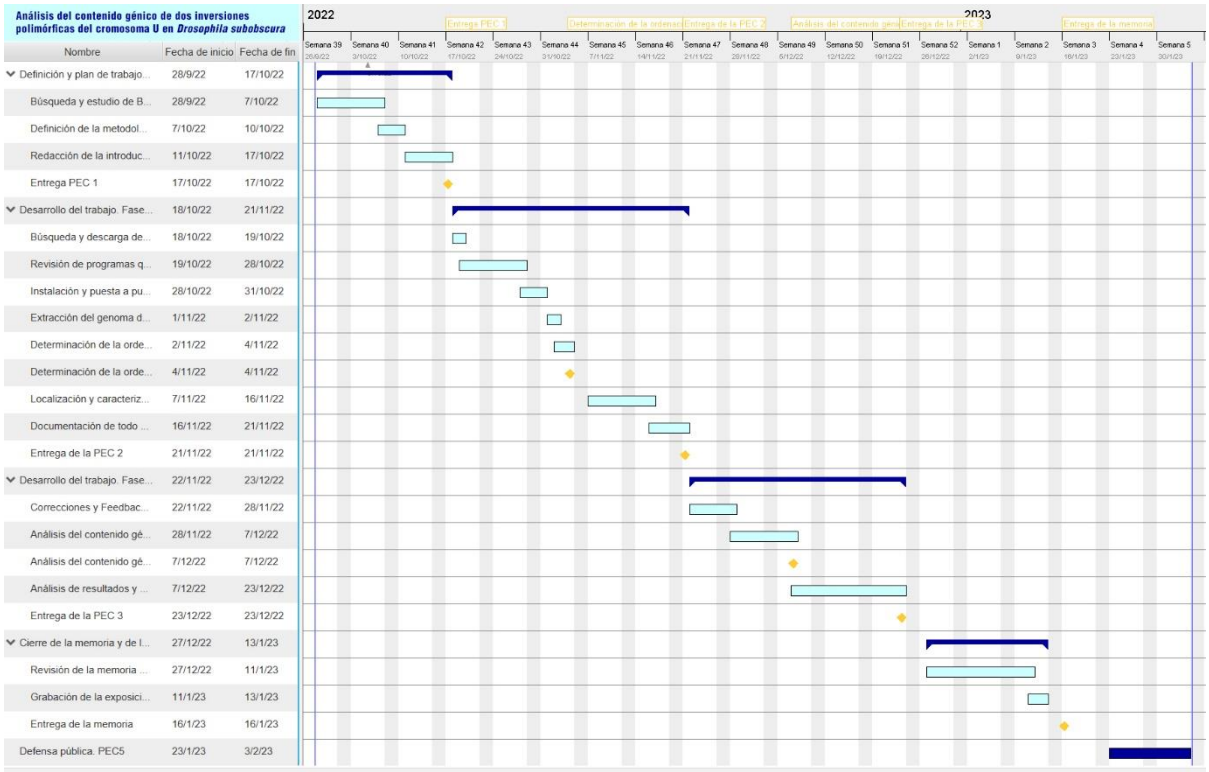

Figura 1 - Diagrama de Gantt

C. Hitos:

- **Definición y plan de trabajo** - 17/10/2022
- **Determinación de la ordenación del cromosoma U en *D. subobscura*** - 4/11/2022
- **Entrega desarrollo del trabajo. Fase 1** - 21/11/2022
- **Análisis del contenido génico y selección de genes con carácter adaptativo** - 7/12/2022
- **Entrega desarrollo del trabajo. Fase 2** - 23/12/2022
- **Entrega de la memoria** - 15/01/2023

D. Análisis de riesgos:

Descripción del riesgo	Severidad	Probabilidad	Mitigación
Problemas con el programa de identificación de ordenación cromosómica	Media	Baja	Investigar programas alternativos
Que la ordenación cromosómica del genoma ensamblado a nivel de cromosoma no sea el esperado	Media	Media	Investigar otros genomas, aunque el ensamblado sea de menor calidad
Mala planificación temporal de las tareas	Media	Baja	Se han fijado tiempos y margen para la corrección de errores y finalización de tareas en previsión a posible subestimación del tiempo necesario para finalizar tareas
Imprevistos de tipo profesional o personal	Baja	Baja	Se han reducido los compromisos profesionales y se cuenta con los fines de semana y días festivos en caso de necesidad

1.6. Breve sumario de productos obtenidos

1. Plan de trabajo

Descripción detallada y temporalización de los objetivos, tareas que se llevarán a cabo durante el TFM.

2. Memoria

Presentación detallada por escrito de los resultados y conclusiones obtenidas durante el TFM. El documento final contendrá los siguientes apartados:
Introducción, estado del arte, métodos, resultados, discusión, conclusiones, glosario, bibliografía y anexos.

3. Presentación en diapositivas y virtual

Diapositivas y vídeo de 20 minutos que contendrá una presentación oral resumiendo los resultados más interesantes y las conclusiones.

1.7. Breve descripción de los otros capítulos de la memoria

Capítulos del cuerpo de la memoria:

2. Estado del arte: Revisión de las últimas publicaciones relacionadas con el tema de estudio del presente trabajo.

3. Materiales y métodos: Descripción de la metodología usada en la elaboración del trabajo. A su vez se divide en los siguientes apartados:

> **3.1. Recopilación de los genomas:** Obtención de los genomas utilizados en el estudio comparativo
> **3.2. Análisis de sintenia:** Se indica el programa utilizado para llevar a cabo el análisis
> **3.3. Localización y Caracterización de los puntos de rotura:** En este apartado se desarrolla la metodología planteada para la acotación de los puntos de rotura.
> **3.4. Análisis de enriquecimiento GO:** Se define qué es un término GO y se detalla la estrategia seguida para el análisis de enriquecimiento de términos GO.

4. Resultados y discusión: Se exponen y analizan los resultados obtenidos del estudio. El capítulo se divide a su vez en tres apartados:

> **4.1. Localización de los puntos de rotura:** Se presentan los resultados del análisis de sintenia.
> **4.2. Localización de los puntos de rotura:** Se precisa la localización de los puntos de rotura de las inversiones y se evalúan los posibles mecanismos por los que se originaron.
> **4.3. Análisis del contenido génico de las inversiones:** En este apartado se detallan los resultados obtenidos del análisis de enriquecimiento GO del contenido de las inversiones y se discute su significado en comparación con estudios previos.

5. Conclusiones y trabajos futuros: Dividido en dos partes:

> **5.1. Conclusiones**: Breve enumeración de las conclusiones sacadas del estudio.
> **5.2. Líneas de trabajo futuro:** Sugerencia de posibles líneas de investigación para continuar y complementar el presente trabajo
> **5.3. Seguimiento de la planificación:** Análisis del desarrollo del trabajo en relación con la planificación inicial. Se detallan los imprevistos surgidos.

2. Estado del arte

Con el fin de determinar el estado actual en el estudio del polimorfismo de inversiones en el género *Drosophila* se ha realizado una búsqueda en Pubmed con las palabras clave "*Drosophila*" e "*inversions"* y se han revisado las publicaciones más recientes. Se ha observado que esta modificación estructural es clave en la especiación y en la adaptación a diversas condiciones ambientales del género *Drosophila* (10). Es especialmente relevante el caso de *D. subobscura*, en el que se ha establecido una relación entre la frecuencia de estas inversiones en relación a diferentes condiciones ambientales, principalmente la temperatura, la humedad y la cantidad de lluvias (19,20).

Las especies *D. madeirensis* y *D. guanche* son endémicas en Madeira y las Islas Canarias, respectivamente, y surgieron de la colonización por un ancestro común de *D. subobscura* de manera independiente. A diferencia de *D. subobscura*, ambas especies son monomórficas a nivel cromosómico. La caracterización de los puntos de rotura de las inversiones fijadas desde la divergencia entre *D. subobscura y D. guanche* ha permitido establecer aquellas que se encontraban presentes en los antecesores de *D. subobscura* que colonizaron las islas, y aquellas que surgieron posteriormente. Asimismo, se pudo determinar en algunas de las inversiones que el mecanismo por el cual fueron originadas es la rotura y reparación cromosómicas(14). Esto concuerda con los resultados obtenidos en otras investigaciones y sugiere que es el principal mecanismo por el que se producen estas inversiones en esta especie (9).

Además de la relación entre las condiciones ambientales y la frecuencia de las inversiones, también se han establecido relaciones con características fenotípicas (tamaño corporal y forma de las alas) y tolerancia térmica (21,22). En relación a esto, se han estudiado los niveles de expresión de la proteína de choque térmico HSP70, una proteína involucrada en la tolerancia térmica en *D. melanogaster* cuya expresión se induce en respuesta a incrementos repentinos de temperatura (23).

La expresión diferencial de genes en *D. subobscura* también ha sido estudiada. En el experimento llevado a cabo por Laayouni y colaboradores (24) se encontraron principalmente diferencias en la expresión de genes implicados en procesos metabólicos entre poblaciones de *D. subobscura* tras tres años de condicionamiento térmico.

La calidad de los genomas comparados constituye una limitación para la aproximación genómica al estudio del polimorfismo de inversiones. El reciente desarrollo de ensamblados de alta calidad para *D. guanche* (16) y *D. subobscura* (18), permite el uso de herramientas bioinformáticas para su estudio. Aunque se especula que las regiones invertidas han fijado conjuntos de genes que contienen alelos coadaptados a las condiciones climáticas (6,22), no se han encontrado hasta la fecha estudios sobre el contenido génico de estas regiones.

El concepto de sintenia tiene su origen en el estudio de la conservación del contenido genético de los cromosomas entre especies. Se puede definir un bloque sinténico como un segmento cromosómico que abarca una serie de genes, ortólogos, que conservan su ordenación comparados con otro genoma.

Existen diferentes herramientas y aproximaciones disponibles para la detección de bloques sinténicos en un genoma o entre diferentes especies. Un factor que se debe tener en cuenta en un estudio de sintenia es la distancia filogenética entre los genomas a comparar. Para genomas estrechamente relacionados se puede optar por un estudio a nivel de ADN, mientras que para genomas más distantes se recomienda un análisis a nivel de proteína, puesto que la mayor divergencia a nivel de ADN dificulta su estudio (25–27).

En este caso, el estudio sinténico se centrará a nivel de ADN, puesto que los genomas comparados corresponden a dos especies muy cercanas *D. guanche* y *D. subobscura*.

En el artículo elaborado por Lallemand y colaboradores (25), se resumen las características de diversas herramientas disponibles para el estudio de sintenia. Para el estudio a nivel de ADN recomiendan las herramientas *Symap* (28) y *Satsuma* (29). Otras herramientas encontradas que podrían ser de utilidad son el paquete *LAST* para *bioconda* (30) y *NGSEP* (31). Las principales características de estas herramientas se resumen en la tabla 1.

Herramienta	*Input*	*Output*	Características	Lenguaje de programación
Symap	Secuencias en formato fasta y anotaciones en formato gff3	Genes homólogos, bloques de sintenia	Dispone de interfaz de usuario gráfica. Proporciona visualizaciones interactivas	No se requiere
Satsuma	Secuencias nucleotídicas	Archivos de texto tabulado, múltiples gráficos interactivos	Encuentra emparejamientos por correlación cruzada	C++, en Linux
LAST	Secuencias en formato fasta	Alineamientos en formato maf	Requiere de librerías adicionales para la visualización gráfica	Phyton en Linux o MacOS
NSPEG	Secuencias en formato fasta y anotaciones en formato gff3	Archivos de texto con la id y localización de genes ortólogos y parálogos entre los genomas	Incluye una interfaz gráfica de usuario para la visualización de los resultados	Comandos en Bash

Tabla 1 - Resumen de las características de los programas de estudio de sintenia valorados

El programa *Symap* fue utilizado con éxito en el artículo de Karageorgiou y colaboradores (7) para realizar un análisis de sintenia entre *D. subobscura* y las especies cercanas relacionadas *D. guanche, D. pseudobscura D. melanogaster*. Además, no requiere el uso de lenguajes de programación y permite la visualización de los resultados mediante su interfaz gráfica de usuario.

La frecuencia de las inversiones cromosómicas en relación con la temperatura sugiere la existencia de genes contenidos en las inversiones implicados en la adaptación a los cambios climáticos. (7,24). El presente estudio podría apoyar esa hipótesis mediante el estudio del enriquecimiento de términos GO en los genes anotados en las inversiones U_1 y U_2, en los que se ha observado una diferencia significativa de la frecuencia de las ordenaciones invertidas entre cepas adaptadas a climas cálidos y fríos.

3. Materiales y métodos

3.1. Recopilación de los genomas

El principal factor limitante en caracterización molecular de los puntos de rotura de las inversiones mediante análisis genético es la calidad de los genomas a comparar (14)

Para poder caracterizar los puntos de rotura de las inversiones U_1 y U_2 disponemos del genoma de *D. guanche* (DGUA_6) que contiene dichas inversiones. Este genoma, que fue publicado en 2018, se obtuvo a partir de *reads* de Illumina y los *scaffolds* se ordenaron en cromosomas mediante hibridaciones *in situ* (16).

En el caso de *D. subobscura*, el genoma de referencia que se encuentra en GenBank corresponde al ensamblado UCBerk_Dsub_1.0 obtenido con tecnología de Oxford Nanopore, que se encuentra a nivel de cromosoma (32). Sin embargo, no hemos encontrado información sobre la ordenación cromosómica de la cepa. Además, GenBank recoge otras cuatro secuencias genómicas de distintas cepas de *D. subobscura* y a distinto nivel de ensamblado. En particular, el ensamblado dsubES12G5 del grupo de Genómica, Bioinformática y Biología evolutiva de la Universidad Autónoma de Barcelona (UAB) a nivel de *scaffolds* corresponde a la cepa *cherry—curled* (*ch-cu*), que presenta la ordenación estándar para el cromosoma U (33).

En el género *Drosophila*, el contenido génico de los cromosomas está muy conservado entre especies, siendo las inversiones la principal causa de diferenciación entre las mismas (34). Para la optimización del trabajo, el análisis del presente estudio se ha centrado en el cromosoma U.

3.2. Análisis de sintenia

Para el análisis se ha utilizado la herramienta *Symap* en su última versión (v5.1.0) utilizando los parámetros por defecto (28).

3.3. Localización y Caracterización de los puntos de rotura

El algoritmo de *Symap* utiliza *MUMmer* para calcular las coincidencias en bruto entre los dos genomas. Con el fin de localizar los puntos de rotura, se tomó la secuencia localizada entre las coincidencias más cercanas al final e inicio de cada uno de los bloques de sintenia para el cromosoma U.

Una vez extraídas las secuencias, se utilizó la herramienta BLAST (Basic Local Alignment Search Tool) de Centro Nacional para la Información Biotecnológica (NCBI) (35) para delimitar la ubicación de los puntos de rotura de manera más precisa. Esta herramienta permite el alineamiento local de secuencias de ADN, ARN o de proteínas. Es posible comparar una secuencia problema (*query*) frente a secuencias disponibles (*subject*) en una base de datos o aportar dos secuencias a comparar directamente entre sí.

En ambos ensamblados, el alineamiento con BLAST se realizó para cada región de una ordenación contra los 3 fragmentos de la otra ordenación. Esta estrategia permite observar posibles duplicaciones nuevas. Asimismo, se alinearon frente al transposón *SGM* (36), la secuencia repetitiva Sat290 (37) y se utilizó la herramienta *Repeatmasker* (38) con el fin de localizar elementos repetitivos.

3.4. Análisis de enriquecimiento GO

Los avances en las tecnologías biológicas y, en particular, la secuenciación de genomas completos, producen datos a gran escala. El proyecto Gene Ontology (GO) es una iniciativa que tiene como fin unificar y estandarizar la terminología utilizada para definir los genes y sus atributos, permitiendo así la interoperabilidad entre bases de datos.
El proyecto proporciona un lenguaje estructurado que se compone de términos descriptivos para función molecular, proceso biológico al que se asocian o la localización celular de los productos génicos, así como la relación jerárquica entre los diversos términos, de más general a más específico (39,40).

El análisis de enriquecimiento de ontología genética es ampliamente utilizado para el estudio de expresión génica diferencial. Su estudio permite determinar aquellos términos relativos a procesos biológicos, localizaciones celulares o funciones moleculares que se encuentran sobrerrepresentados o infrarrepresentados en los genes objeto del estudio (41).

En primer lugar, para el análisis es necesaria una lista de genes de interés sobre la que se desea investigar si existe enriquecimiento de algún término. Los genes utilizados en este estudio son genes codificantes de proteína anotados en las regiones U_1 y U_2 del genoma del ensamblado UCBerk_Dsub_1.0 de *D. subobscura*.

El análisis de enriquecimiento se ha realizado comparando los genes codificantes de proteína contenidos en las inversiones con la totalidad del genoma. El paquete utilizado para el análisis es el paquete *topGO* (42) de *bioconductor* en *R*. El test utilizado para evaluar la significación estadística fue la prueba exacta de Fisher, basado en contaje de genes. Los test que analizan puntuaciones basadas en datos de expresión génica fueron descartados, puesto que no se cuenta con tales datos en el presente estudio. Se utilizó el algoritmo por defecto para la puntuación de términos GO.

El código utilizado para el análisis de enriquecimiento GO con *topGO* se adjunta en el Anexo I.

4. Resultados y discusión

4.1. Localización de los puntos de rotura

En la comparación del cromosoma U del ensamblado UCBerk_Dsub_1.0 de *D. subobscura* con el mismo cromosoma del ensamblado de *D. guanche* DGUA_6 se observa que sus cromosomas U mantienen la colinearidad de los marcadores en toda su región central (Figura 2A), desvelando que comparten su ordenación cromosómica. Por tanto, la cepa secuenciada de *D. subobscura* es U_{1+2} al igual que la especie canaria.

Como alternativa, se realizó el análisis de sintenia con el ensamblado dsubES12G5 del grupo de Genómica, Bioinformática y Biología evolutiva de la Universidad Autónoma de Barcelona (UAB), ensamblado a nivel de *scaffolds*.

El resultado de este análisis muestra dos grandes inversiones en la parte central del cromosoma (Figura 2B), como era de esperar, pues la cepa *ch-cu* usada por la UAB para su ensamblado es homocariotípica para la ordenación estándar en todos sus cromosomas excepto para el cromosoma O que presenta la configuración O_{3+4}.

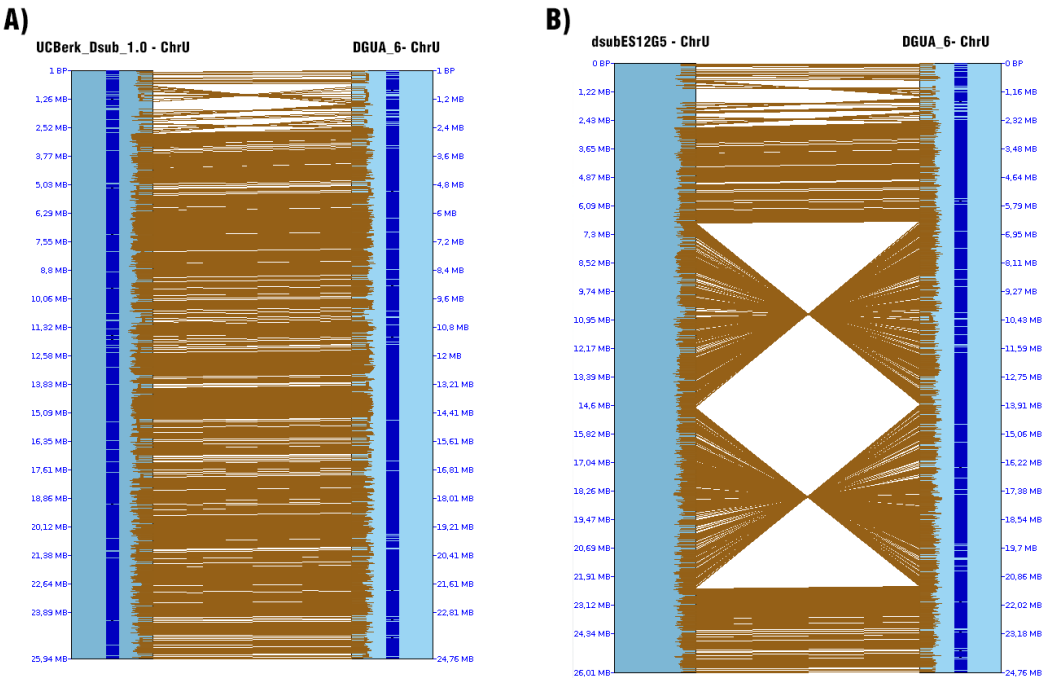

Figura 2 - Análisis comparativo de sintenia entre ensamblado DGUA_6 de *D. guanche* y los ensamblados UCBerk_Dsub_1.0 (A) y dsubES12G5 (B) de *D. subobscura*.

Al comparar los genomas de las dos especies, se observan además unas microinversiones al inicio del cromosoma U presentes en las dos comparativas entre las cepas de *D. subobscura* y *D. guanche*. En la ampliación de la Figura 3 se pueden observar mejor hasta 3 microinversiones entre las dos especies. Para comprobar si se tratan de

inversiones reales o se trata de un artefacto del ensamblado se han rastreado los genomas implicados.

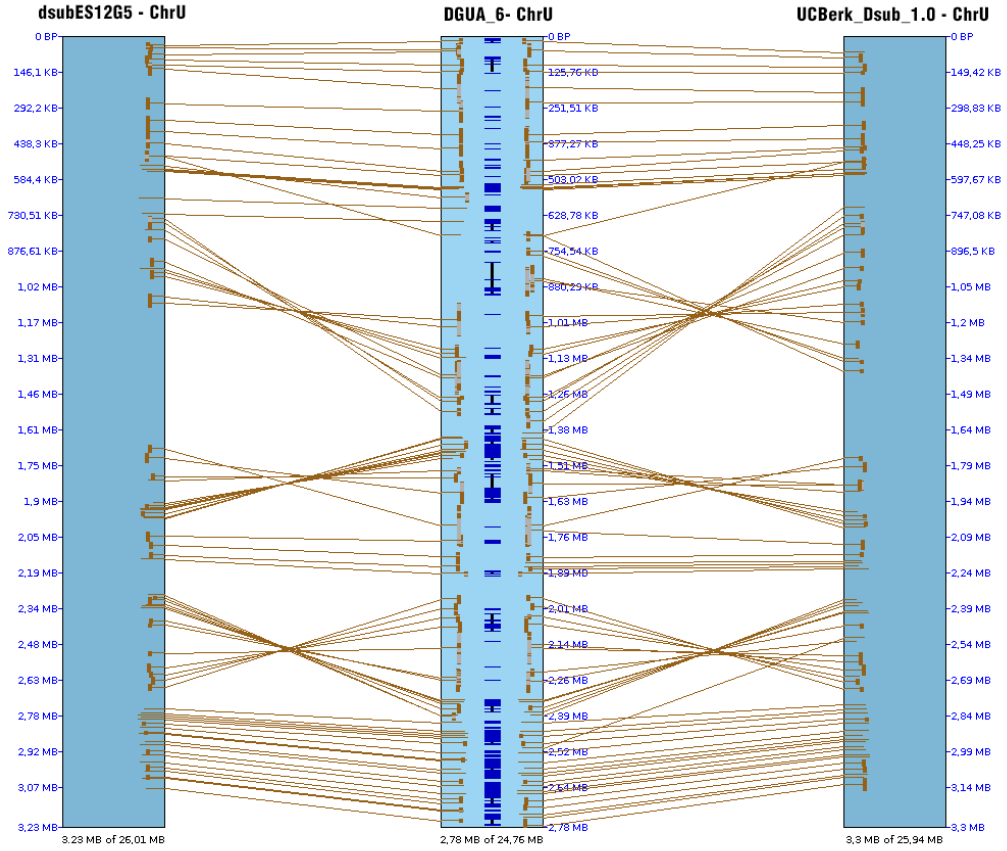

Figura 3 - Ampliación del segmento inicial del cromosoma U del análisis de sintenia mediante *Symap* entre el ensamblado DGUA_6 de *D. guanche* (centro) y los ensamblados dsubES12G5 (izquierda) y UCBerk_Dsub_1.0 (derecha) de *D. subobscura*. Se observan microinversiones presentes en ambos ensamblados de *D. subobscura* en comparación con el de *D. guanche*.

El ensamblado D_GUA6 de *D. guanche* contiene una serie de huecos (gaps) entre *scaffolds* en la región inicial del cromosoma (Tabla 2). La existencia de estos huecos de mapeo en el genoma de *D. guanche* podría explicar el origen de las microinversiones. Es posible que las microinversiones detectadas no sean inversiones reales, sino que sean artefactos del ensamblado.

De hecho, en el artículo del ensamblado se indica que existe cierta incertidumbre en la orientación de los *scaffolds* próximos al centrómero, dada la tendencia natural a la rotura de los cromosomas politénicos en las preparaciones citológicas (16). Este es el caso para el cromosoma U, en el que no se pudo determinar la orientación de dos de los *scaffolds* de la región próxima al centrómero.

Gaps entre *scaffolds*	*scaffold* anterior	*scaffold* posterior
chrU: 567840-567939	dgua6_s00022	dgua6_s00020
chrU: 1408149-1408248	dgua6_s00020	dgua6_s00049
chrU: 1529667-1529766	dgua6_s00049	dgua6_s00034
chrU: 1750137-1750236	dgua6_s00034	dgua6_s00041
chrU: 1899031-1899130	dgua6_s00041	dgua6_s00024
chrU: 2402946-2403045	dgua6_s00024	dgua6_s00010

Tabla 2 - Gaps entre *scaffolds* del ensamblado de *D. guanche* DGUA_6 localizados en la región proximal del cromosoma U. No se conoce la longitud total de cada gap, pero, por defecto, tienen asignada una longitud de 100 bases. Se indican también los *scaffolds* entre los que se encuentra el gap con los colores utilizados en la figura 4.

Los resultados obtenidos en este trabajo sugieren que la orientación de los *scaffolds* dgua6_s00020, dgua6_s00024, dgua6_s00049 y dgua6_s00034 podría estar invertida (Figura 4). Además, la posición de los *scaffolds* dgua6_s00049 y dgua6_s00034 parece estar intercambiada.

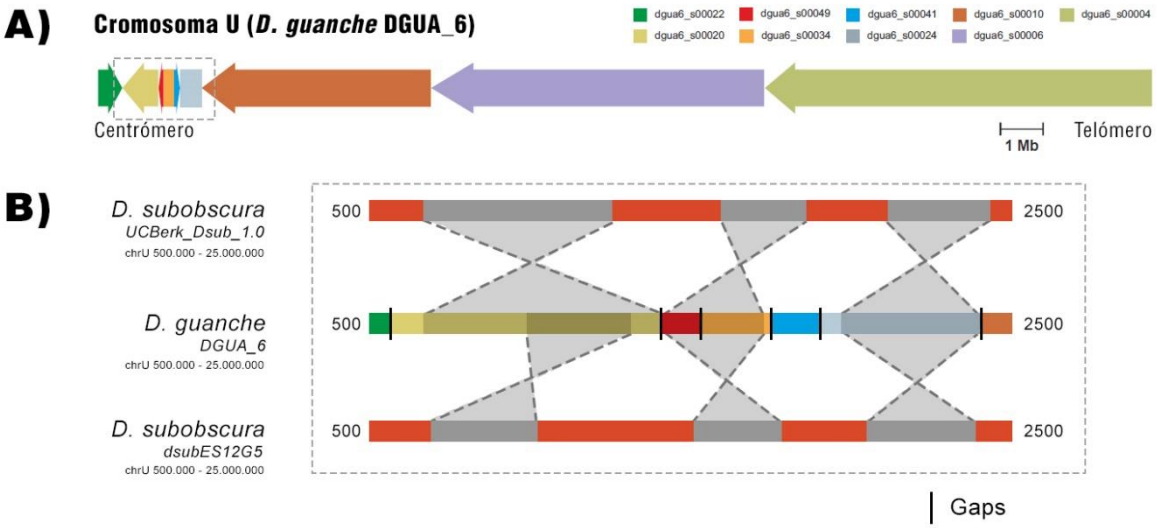

Figura 4 – A: Ensamblado a nivel de scaffolds del cromosoma U de *D. guanche*. Modificado de Orengo y colaboradores (16)
B: Representación esquemática de las microinversiones y los gaps presentes en el ensamblado D_GUA6 de *D. guanche* en comparación con los ensamblados UCBerk_Dsub_1.0 y dsubES12G5 de *D. subobscura*. Región proximal del cromosoma U entre 500 y 2500 kb.

Puesto que los genomas de *D. subobscura* de los ensamblados dsubES12G5 y UCBerk_Dsub_1.0 presentan la configuración U_{ST} y U_{1+2} respectivamente, se prosigue el estudio de los puntos de rotura mediante la comparación de dichos genomas.

4.2. Caracterización de los puntos de rotura

Los puntos de rotura distal de la inversión U_1 y proximal de la inversión U_2 de las inversiones U_1 y U_2 se encuentran muy próximos entre sí. La herramienta *Symap* no ha detectado coincidencias (*hits*) en la región intermedia entre ambas inversiones. Para acotar la localización de los puntos se han extraído 6 secuencias, comprendidas entre los *hits* identificados por *Symap* que delimitan las inversiones. Para la identificación del punto de rotura se ha utilizado la herramienta BLAST comparando todos los fragmentos entre sí.

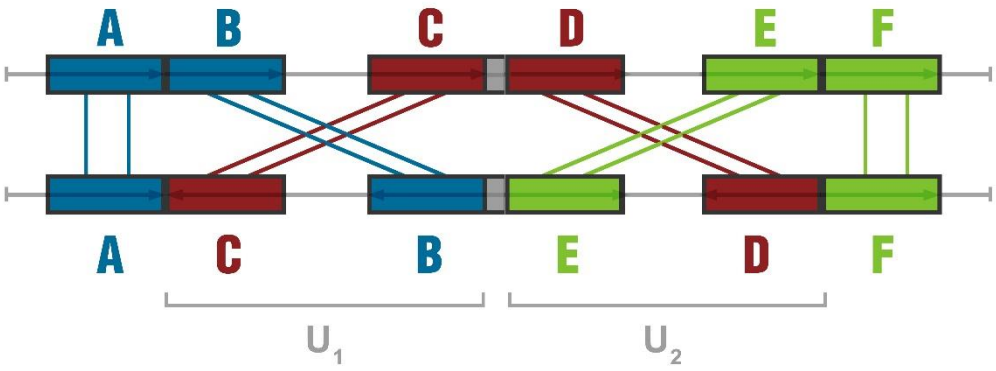

Figura 5 - Representación esquemática de los fragmentos tomados para la delimitación de los puntos de rotura. En los fragmentos denominados CD y BE se deben de encontrar dos puntos de rotura: El terminal de la inversión U_1 y el proximal de la inversión U_2

En algunos casos, para la correcta caracterización del punto de rotura, ha sido necesario tomar fragmentos más largos. Los resultados de los alineamientos se resumen en las figuras 6 y 7.

Cepa	Fragmento	Inicio	Final	Longitud (bases)
UCBerk_Dsub_1.0	AB	6.840.275	6.850.676	10.402
	CD	14.576.127	14.639.318	63.192
	EF	22.262.835	22.337.675	74.841
dsubES12G5	AC	6.818.568	6.860.996	42.429
	BE	14.666.318	14.704.611	38.294
	BE (ext)	14.616.457	14.776.125	159.669
	DF	22.386.937	22.423.420	36.484

Tabla 3 - Fragmentos utilizados en el alineamiento local (BLAST) para la acotación de los puntos de rotura. Para la caracterización del punto distal de la inversión U_2 en UCBerk_Dsub_1.0 y proximal de U_2 en dsubES12G5 ha sido necesario tomar un fragmento de mayor tamaño.

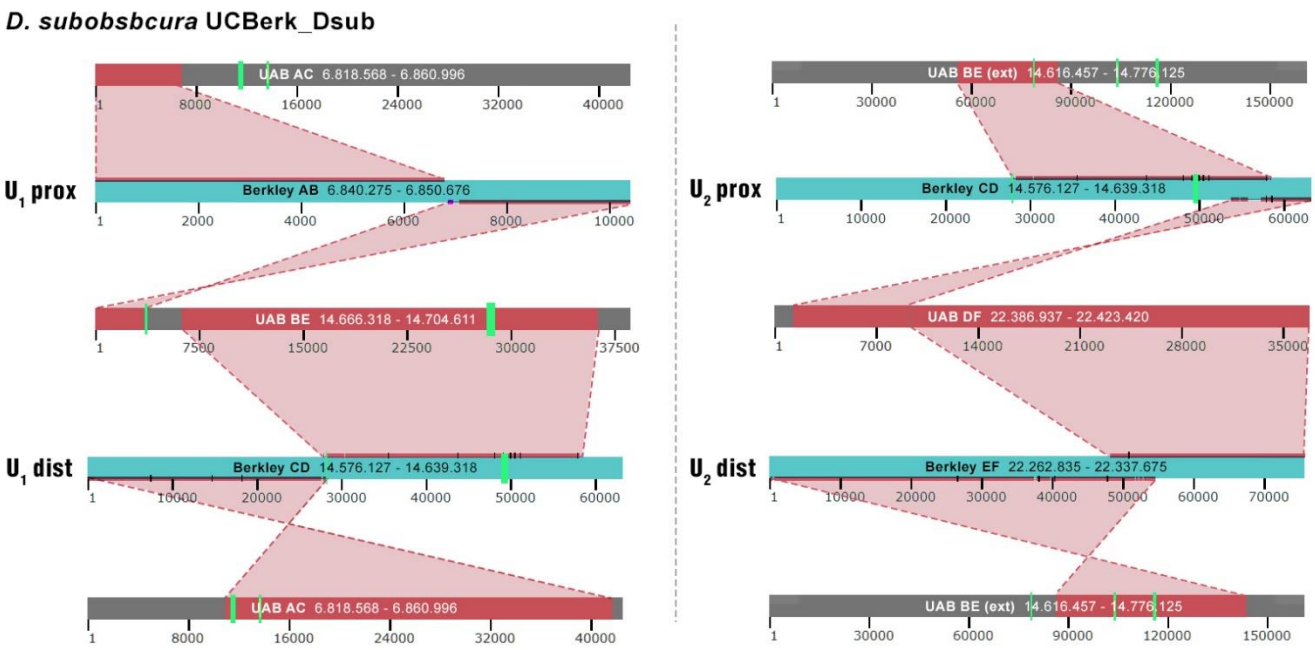

Figura 6 – Representación esquemática de los alineamientos entre los ensamblados UCBerk_Dsub_1.0 (azul) y dsubES12G5 (rojo) de *D. subobscura*. A la izquierda se observan los puntos de rotura proximal y distal de la inversión U_1 del ensamblado UCBerk_Dsub_1.0, localizados en los fragmentos denominados AB y CD. A la derecha los puntos U_2 proximal en el fragmento CD y U_2 distal en la región EF. Se muestran en verde los alineamientos de las secuencias con el transposón SGM.

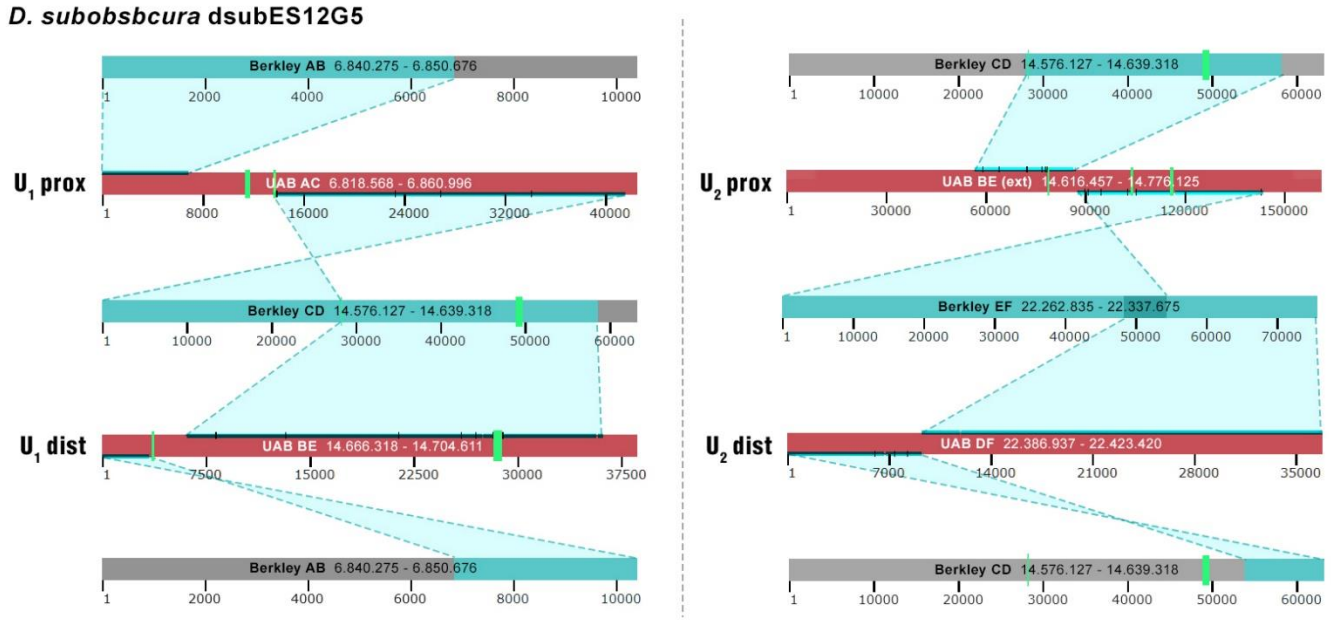

Figura 7 - Representación esquemática de los alineamientos entre los ensamblados UCBerk_Dsub_1.0 (azul) y dsubES12G5 (rojo) de *D. subobscura*. A la izquierda se observan los puntos de rotura proximal y distal de la inversión U_1 del ensamblado dsubES12G5, localizados en los fragmentos denominados AC y BE. A la derecha los puntos U_2 proximal en el fragmento BE (ext) y U_2 distal en la región DF. Se muestran en verde los alineamientos de las secuencias con el transposón SGM.

Los alineamientos muestran un solapamiento en los extremos de la inversión U_2 en el ensamblado UCBerk_Dsub_1.0 en su alineamiento con dsubES12G5. Este hecho indica la existencia de duplicaciones en los extremos de los puntos de rotura de dicha inversión en dsubES12G5. El alineamiento de las regiones BE y DF muestra dos duplicaciones con orientación opuesta flanqueando los extremos de la reordenación (Figura 8). Esto indicaría que se ha originado por un mecanismo de NHEJ (Figura 9) y que el ordenamiento ancestral es el U_2 y la ordenación estándar se ha originado a partir de ésta, como ya había sido propuesto (15).

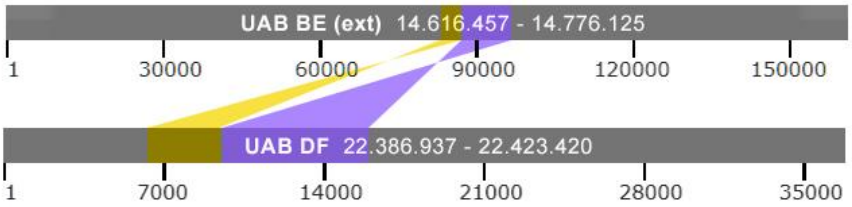

Figura 8 - Alineamiento entre los fragmentos BE y DF del ensamblado dsubES12G5 en el que se observan dos alieamientos correspondientes a duplicaciones

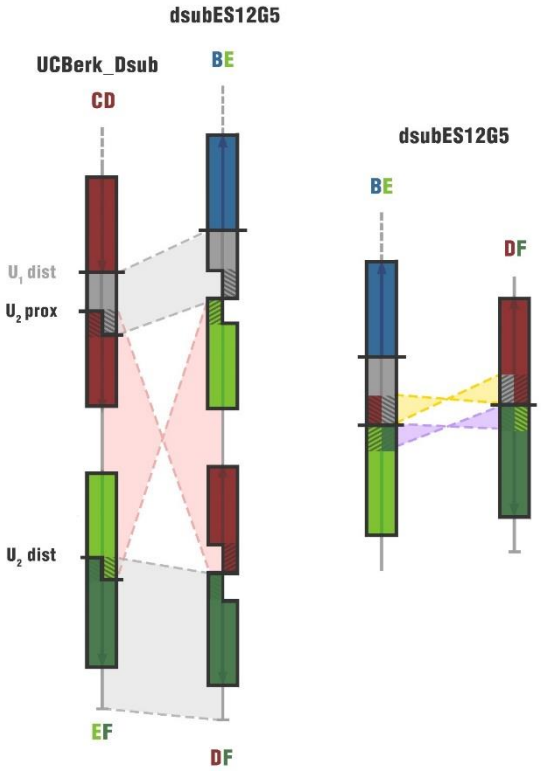

Figura 9 - Esquema del posible origen de las duplicaciones observadas en los fragmentos BE y DF del ensamblado dsubES12G5. En amarillo representada la duplicación originada por la rotura escalonada del fragmento CD (U_2 proximal) y posterior reparación de extremos no homólogos. De manera análoga, se indica en lila la duplicación originada por la rotura en el fragmento EF (U_2 distal).

Respecto a la ordenación U_1, los puntos de rotura quedan acotados en las posiciones chrU:6.847.076..6.847.348 del ensamblado de Berkeley para la rotura proximal (fragmento AB) y chrU:14.604.464..14.604.478 para la rotura distal (fragmento CD).

Al comparar el fragmento AB consigo mismo en BLAST, se aprecia una duplicación en tándem muy cercana a la región del punto de rotura proximal (Figura 10). Adicionalmente, existe alineamiento entre la secuencia CD y el transposón SGM en el punto de rotura distal de U_1. No se observan regiones duplicadas ni solapamientos en el alineamiento con los fragmentos del ensamblado de la UAB. Al no encontrar fragmentos duplicados a ambos lados de la inversión en ninguna de las dos ordenaciones no se puede inferir cual ha sido el mecanismo de su origen. Que no se detecten no significa que no hayan existido y posteriormente se hayan eliminado por un acúmulo de mutaciones. Así, el mecanismo de rotura que provocó el reordenamiento tanto podría ser por NHEJ como por NAHR y tampoco se puede descartar que fuera por acción de los transposones.

La herramienta *Repeatmasker* no ha detectado más secuencias repetitivas en los puntos de rotura de las inversiones.

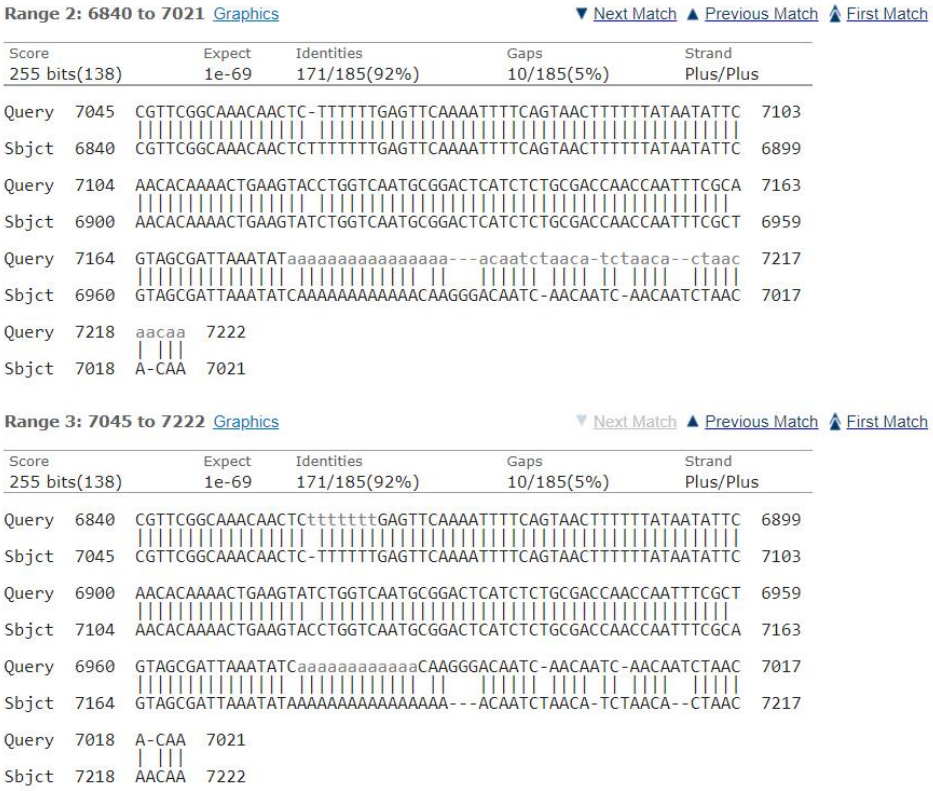

Figura 10 – Resultado parcial del BLAST del fragmento AB (UCBerk_Dsub_1.0) consigo mismo. Se aprecia una duplicación en tándem en el mismo fragmento AB 6840-7021 = 7045-7222.

En los resultados del BLAST, también se observan unas regiones en los fragmentos AC y BE que no producen alineamiento con ninguna de las regiones con los que se han comparado. Utilizando el paquete *Biostrings* (43) se han extraído las secuencias que no producen alineamiento para realizar un BLAST con la base de datos de la suite.

En la secuencia extraída del fragmento BE se obtiene un alineamiento con un 98% de identidad con un gen del componente 4 del complejo de exocitosis de *D. subobscura*:

PREDICTED: Drosophila subobscura exocyst complex component 4 (LOC117897130)

Por otro lado, para el fragmento AC, los resultados incluyen alineamientos con regiones genéticas que contienen el transposón SGM, las secuencias repetitivas *GEM1730*, *GEM871* y con una secuencia parcial del gen *GA16100*, de función desconocida.

4.3. Análisis del contenido génico de las inversiones

El genoma del ensamblado UCBerk_Dsub_1.0 contiene anotados 14.381 genes, de los cuales 13.440 corresponden a genes codificantes de proteína (32). Sin embargo, estas anotaciones no contienen términos GO para los distintos genes. Para la anotación de los mismos se ha recurrido a la aplicación web *eggNOG-mapper* (44). La aplicación permite obtener los términos GO de un genoma suministrando secuencias de proteínas, secuencias codificantes o secuencias genómicas.

Se han utilizado las secuencias de proteínas de la cepa de Berkeley para la anotación de términos GO. Se han conseguido anotar 21.413 proteínas de 22.475 secuencias de proteínas suministradas (en algunos casos existe más de una isoforma proteica por gen anotado). Los resultados de *eggNOG-mapper* se pueden descargar en el Anexo II.

Para el estudio de enriquecimiento génico de la inversión U_1, se han tomado los genes codificantes de proteína comprendidos entre las posiciones 6.419.392 y 13.846.098 del cromosoma U de la cepa de Berkeley. Se encuentran anotados en esta región 825 genes. La región tomada para el estudio de la inversión U_2 es la comprendida entre las posiciones 13.869.464 y 21.253.679 del cromosoma U, con un total de 829 genes anotados.

Para la inversión U_1, los términos GO más enriquecidos (para proceso biológico) se relacionan con el catabolismo de ácidos grasos y morfogénesis de nefronas (Tabla 4, Figura 11).

En el estudio elaborado por Laayouni y colaboradores (24), se evalúa la expresión diferencial de genes midiendo la abundancia relativa de ARN mensajero en poblaciones de *D. subobscura* adaptadas durante 3 años a regímenes de temperatura fríos (13ºC) y cálidos (22ºC) en comparación con una población adaptada a su temperatura óptima de crecimiento (18ºC). Sus resultados muestran una sobrerrepresentación de genes diferencialmente expresados entre las poblaciones adaptadas a clima cálido frente a frío de genes implicados en metabolismo de carbohidratos, metabolismo de ácidos nucleicos y regulación de la transcripción. En cambio, categorías relacionadas con el metabolismo de ácidos orgánicos y ácidos carboxílicos aparecen infrarrepresentadas.

En el presente estudio no se evalúa la expresión diferencial, sino el enriquecimiento de términos GO en los genes contenidos en las inversiones. Es llamativo que en la inversión U_1 existe un enriquecimiento génico de dos categorías que para las que no se ha encontrado expresión diferencial en el estudio de Laayouni y colaboradores.

En el caso de la inversión U_2 se observa enriquecimiento de términos relativos a metabolismo y recuperación de bases nitrogenadas (Tabla5, Figura 12). Como se ha mencionado previamente, se ha observado una expresión diferencial de genes relacionados con el metabolismo de ácidos nucleicos entre poblaciones adaptadas a climas cálidos y fríos, lo que sugiere la implicación de las inversiones en la regulación de la expresión de dichos genes.

El resto de los resultados obtenidos en el análisis de enriquecimiento se adjuntan en el Anexo III.

Término GO	Función biológica	Anotados	Encontrados	Esperados	Puntuación Fisher
GO:0009062	fatty acid catabolic process	89	17	5.40	2.0e-05
GO:0006635	fatty acid beta-oxidation	60	13	3.64	4.8e-05
GO:0019395	fatty acid oxidation	83	15	5.03	0.00012
GO:0034440	lipid oxidation	84	15	5.10	0.00014
GO:0072329	monocarboxylic acid catabolic process	105	17	6.37	0.00018
GO:0060675	ureteric bud morphogenesis	35	9	2.12	0.00018
GO:0072171	mesonephric tubule morphogenesis	35	9	2.12	0.00018
GO:0044242	cellular lipid catabolic process	147	21	8.92	0.00020
GO:0016042	lipid catabolic process	204	26	12.37	0.00025
GO:0072078	nephron tubule morphogenesis	37	9	2.24	0.00028
GO:0072088	nephron epithelium morphogenesis	37	9	2.24	0.00028
GO:0016267	O-glycan processing, core 1	12	5	0.73	0.00045
GO:0072202	cell differentiation involved in metanep...	12	5	0.73	0.00045
GO:0001658	branching involved in ureteric bud morph...	32	8	1.94	0.00050
GO:0072028	nephron morphogenesis	40	9	2.43	0.00052
GO:0060231	mesenchymal to epithelial transition	13	5	0.79	0.00069
GO:0001657	ureteric bud development	50	10	3.03	0.00070
GO:0072163	mesonephric epithelium development	50	10	3.03	0.00070
GO:0072164	mesonephric tubule development	50	10	3.03	0.00070
GO:0001823	mesonephros development	51	10	3.09	0.00083

Tabla 4 - Términos GO para proceso biológico con un enriquecimiento significativo para la inversión U_1, la función biológica que desempeñan; Anotados, número de genes anotados en el genoma que contienen el término; Encontrados, número de genes encontrados en la inversión que contienen el término ; Esperados, número de genes que se esperaría encontrar si no hubiera un enriquecimiento; y puntuación obtenida en el test exacto de Fisher. En amarillo se muestran el conjunto de términos relacionados con el metabolismo de ácidos grasos y en rojo aquellos implicados en la morfogénesis de nefronas. Muchos genes pertenecen a más de una catergoría (presentan más de un término GO asociado). Se muestran los 20 resultados más significativos en función del test exacto de Fisher.

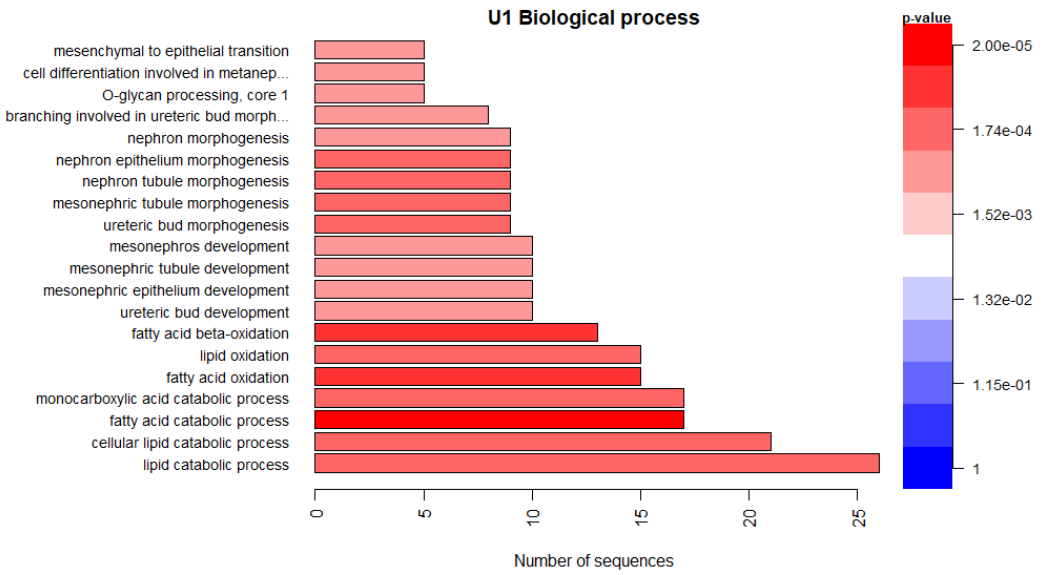

Figura 11 - Análisis de enriquecimiento de términos GO para proceso biológico de los genes codificantes de proteína anotados en la inversión U_1 de *D. subobscura* UCBerk_Dsub_1.0. En el eje X se indica el número de genes encontrados que contienen el término anotado.

Término GO	Función biológica	Anotados	Encontrados	Esperados	Puntuación Fisher
GO:0008655	pyrimidine-containing compound salvage	14	7	0.89	9.7e-06
GO:0043097	pyrimidine nucleoside salvage	14	7	0.89	9.7e-06
GO:0009313	oligosaccharide catabolic process	7	5	0.45	2.0e-05
GO:0043174	nucleoside salvage	17	7	1.09	4.7e-05
GO:0021891	olfactory bulb interneuron development	9	5	0.58	0.00011
GO:0006013	mannose metabolic process	10	5	0.64	0.00020
GO:0046134	pyrimidine nucleoside biosynthetic proce...	21	7	1.34	0.00022
GO:0006216	cytidine catabolic process	3	3	0.19	0.00026
GO:0006217	deoxycytidine catabolic process	3	3	0.19	0.00026
GO:0009972	cytidine deamination	3	3	0.19	0.00026
GO:0046087	cytidine metabolic process	3	3	0.19	0.00026
GO:0046109	uridine biosynthetic process	3	3	0.19	0.00026
GO:0046127	pyrimidine deoxyribonucleoside catabolic...	3	3	0.19	0.00026
GO:2001252	positive regulation of chromosome organi...	174	24	11.12	0.00028
GO:0043412	macromolecule modification	2238	178	143.06	0.00040
GO:0061983	meiosis II cell cycle process	37	9	2.37	0.00041
GO:0006213	pyrimidine nucleoside metabolic process	30	8	1.92	0.00044
GO:1905269	positive regulation of chromatin organiz...	118	18	7.54	0.00047
GO:1902275	regulation of chromatin organization	192	25	12.27	0.00051
GO:0021889	olfactory bulb interneuron differentiati...	12	5	0.77	0.00057

Tabla 5 - Términos GO para proceso biológico con un enriquecimiento significativo para la inversión U_2, la función biológica que desempeñan; Anotados, número de genes anotados en el genoma que contienen el término; Encontrados, número de genes encontrados en la inversión que contienen el término; Esperados, número de genes que se esperaría encontrar si no hubiera un enriquecimiento; y puntuación obtenida en el test exacto de Fisher. Muchos genes pertenecen a más de una categoría (presentan más de un término GO asociado). Se muestran los 20 resultados más significativos en función del test exacto de Fisher.

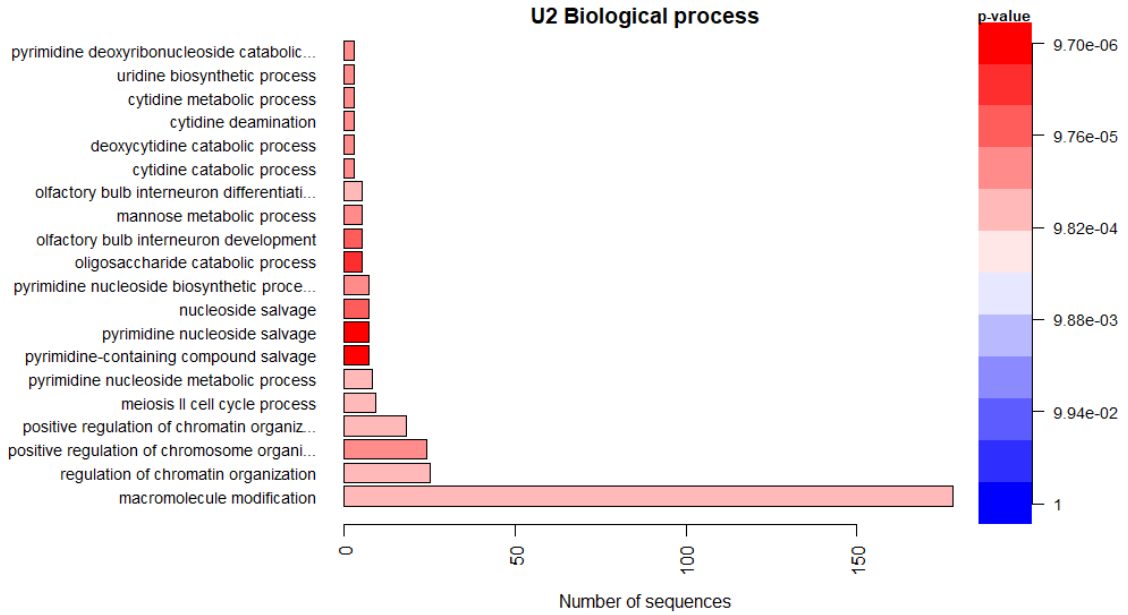

Figura 12 - Análisis de enriquecimiento de términos GO para proceso biológico de los genes codificantes de proteína anotados en la inversión U_2 de *D. subobscura* UCBerk_Dsub_1.0. En el eje X se indica el número de genes encontrados que contienen el término anotado.

5. Conclusiones y trabajos futuros

5.1. Conclusiones

En el presente trabajo se ha realizado un análisis de sintenia y caracterización de puntos de rotura de los reordenamientos U_1 y U_2 del cromosoma U de *D. subobscura*. Además, se ha pretendido ahondar en el carácter adaptativo a los cambios de temperatura de dichas inversiones mediante el análisis de enriquecimiento de términos GO en los genes contenidos en las regiones invertidas.

Uno de los mayores focos de dificultad del trabajo ha sido el análisis de enriquecimiento de términos GO. En este tipo de análisis se suelen comparar niveles de expresión génica midiendo la abundancia relativa de ARN mensajero en las poblaciones a comparar. En el caso de este trabajo no se contaba con datos de expresión génica, sino una lista de genes contenidos en las regiones invertidas. Además, las anotaciones disponibles de los ensamblados de *D. subobscura* no disponían de términos GO anotados y no se dispone de una base de datos de la especie. La solución encontrada para realizar el análisis sobre un organismo *no modelo* y sin datos de expresión génica consistió en obtener anotaciones GO para el genoma completo de la cepa y comparar el enriquecimiento de términos GO en los genes contenidos en las regiones invertidas frente a la totalidad del genoma.

A la vista de los resultados obtenidos se pueden extraer las siguientes conclusiones:

1. El estudio de sintenia entre los genomas de referencia de *D. subobscura* y de *D. guanche* revela que el cromosoma U de la cepa utilizada en el ensamblado UCBerk_Dsub_1.0 de *D.subobscura* tiene la ordenación U_{1+2}. Posteriormente a este resultado se ha encontrado este dato (18).
2. Adicionalmente, el estudio de sintenia entre estas dos especies ha desvelado una serie de microinversiones en el ensamblado de *D. guanche* probablemente originados por artefactos del ensamblado que eran desconocidos.
3. El análisis de sintenia seguido del uso de BLAST ha permitido acotar la localización de los puntos de rotura de las inversiones U_1 y U_2. Los resultados revelan que el mecanismo originario de la inversión en U_2 ha sido por NHJE con extremos escalonados, generándose la ordenación U_{ST} a partir de la ordenación ancestral U_2 como era esperado. Sin embargo, no ha sido posible determinar con seguridad el mecanismo por el cual se generó la inversión en U_1.
4. El análisis de enriquecimiento de términos GO ha permitido determinar que existe una abundancia relativa de genes con términos asociados al metabolismo lipídico y con la morfogénesis de nefronas en la inversión U_1. Parece razonable que los cambios adaptativos a las condiciones climáticas incluyan cambios en el metabolismo. Sin embargo, en estudios previos de expresión diferencial entre cepas adaptadas a climas cálidos y a climas fríos, no se ha encontrado expresión diferencial de genes implicados en las categorías para las que se ha encontrado enriquecimiento en la inversión. Sería necesario evaluar si estos genes tienen algún papel en la adaptación climática de *D. subobscura*.

5. Los datos obtenidos en el análisis de términos GO para la inversión U_2 sugieren la implicación de las inversiones en la regulación de la expresión de los genes contenidos en la inversión. Éstos presentan un enriquecimiento de términos relativos a metabolismo de ácidos nucleicos para los que se ha observado diferencias en la expresión génica entre cepas adaptadas a climas cálidos y a climas fríos.

5.2. Líneas de trabajo futuro

El estudio podría continuarse analizando aquellos genes que presentan enriquecimiento de términos GO. Se podría determinar su distribución dentro de las inversiones, si pertenecen a familias génicas aparecidas por duplicaciones de un gen o si pertenecen a genes que actúan en distintos niveles de una misma cadena o red de reacciones.

También sería interesante comparar los ensamblados UCBerk_Dsub_1.0 de *D. subobscura* y D_GUA6 de *D. guanche*, ambos con la ordenación U_{1+2}, para evaluar si las regiones de los puntos de rotura de las inversiones han divergido mucho.

5.3. Seguimiento de la planificación

En el diseño del trabajo se propuso realizar los análisis comparativos con los genomas de referencia de *D. guanche* y *D. subobscura*. Sin embargo, se desconocía la ordenación del cromosoma U de esta última. El análisis demostró que el ensamblado presenta la ordenación U_{1+2} (información que se encontró posteriormente), por lo que se utilizó alternativamente el ensamblado de la UAB con la ordenación U_{ST} para el estudio comparativo, como estaba previsto en el análisis de riesgos. De haber sabido inicialmente la ordenación del ensamblado de referencia de *D. subobscura*, podría haberse hecho el análisis comparativo directamente con los dos ensamblados de *D. subobscura*. No obstante, y a pesar del retraso, el análisis de sintenia ha permitido encontrar las microinversiones presentes en el ensamblado de *D. guanche*.

Una dificultad no prevista fue la ausencia de anotaciones de términos GO en los ensamblados de *D. subobscura*, complicando el análisis de enriquecimiento al tener que investigar posibles soluciones. Tampoco existe una base de datos de términos GO para la especie, lo que dificulta aún más el análisis.

A pesar de las dificultades encontradas, ha sido posible la consecución de todos los objetivos planteados en la propuesta de trabajo.

6. Glosario

Bioconductor – Proyecto de código abierto para el análisis de datos en Genómica basado en el lenguaje de programación R.

BLAST – *Basic Local Alignment Search Tool*. Herramienta para el alineamiento de secuencias de tipo local, ya sea de ADN, ARN o de proteínas. El programa permite comparar una secuencia problema (*query*) frente a una base de datos o con otra secuencia o secuencias suministradas (*subject*).

D. subobscura – *Drosophila subobscura* especie de mosca de la fruta de la familia *Drosophilidae*.

dsubES12G5 – Ensamblado de *Drosophila subobscura* a nivel de *scaffolds* de la Universidad Autónoma de Barcelona (UAB) a partir de una cepa *cherry—curled*.

DGUA_6 – Ensamblado de *Drosophila guanche* a nivel de cromosomas del Centro Nacional de Análisis Genómico de Barcelona.

GO – *Gene Ontology*. Ontología Génica. Términos descriptivos de un gen y los atributos del producto génico en cualquier organismo clasificados en función molecular, proceso biológico al que se asocian o la localización celular de los productos génicos. También contienen información sobre la relación jerárquica entre los diversos términos

Inversión cromosómica – Modificación estructural por el cual una región cambia de sentido dentro del propio cromosoma. Se modifica por tanto la ordenación de elementos contenidos en la región invertida con relación a la considerada estándar.

NAHR – *Nonallelic Homologous Recombination*. Recombinación homóloga no alélica.

NCBI – Centro Nacional para la Información Biotecnológica.

NHEJ – *Non-homologous DNA end joining*. Rotura y reparación mediante unión de extremos no homólogos.

Scaffold – Porción de secuencia génica resultado de secuenciación obtenida por la ordenación y ensamblado de *contigs* (secuencias continuas obtenidas por el solapamiento de pequeños fragmentos génicos).

Sintenia – Conservación del orden génico y la orientación a lo largo del cromosoma.

SyMap – *Synteny Mapping and Analysis Program*. Programa para la detección y visualicazión de relaciones sinténicas entre genomas.

topGO – Paquete de *Bioconductor* para el análisis de enriquecimiento de términos GO.

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

7. Bibliografía

1. Menozzi P, Krimbas CB. The inversion polymorphism of *D. subobscura* revisited: Synthetic maps of gene arrangement frequencies and their interpretation. J Evol Biol [Internet]. 1 de julio de 1992;5(4):625-41. Disponible en: https://onlinelibrary.wiley.com/doi/full/10.1046/j.1420-9101.1992.5040625.x
2. Prevosti A, Serra L, Ribo G, Aguade M, Sagarra E, Monclus M, et al. The Colonization of *Drosophila subobscura* in Chile. II. Clines in the Chromosomal Arrangements. Evolution (N Y). julio de 1985;39(4):838.
3. Prevosti A, Ribo G, Serra L, Aguade M, Balaña J, Monclus M, et al. Colonization of America by *Drosophila subobscura*: Experiment in natural populations that supports the adaptive role of chromosomal-inversion polymorphism. Proc Natl Acad Sci U S A [Internet]. 1 de agosto de 1988;85(15):5597-600. Disponible en: https://europepmc.org/articles/PMC281806
4. Rodríguez-Trelles F, Alvarez G, Zapata C. Time-Series Analysis of Seasonal Changes of the O Inversion Polymorphism of *Drosophila Subobscura*. Genetics [Internet]. enero de 1996;142(1):179. Disponible en: /pmc/articles/PMC1206946/?report=abstract
5. Orengo DJ, Prevosti A. Temporal changes in chromosomal polymorphism of *Drosophila subobscura* related to climatic changes. Evolution (N Y). 1996;50(3):1346-50.
6. Orengo DJ, Puerma E, Aguadé M. Monitoring chromosomal polymorphism in *Drosophila subobscura over* 40 years. Entomol Sci [Internet]. 1 de julio de 2016;19(3):215-21. Disponible en: https://onlinelibrary.wiley.com/doi/full/10.1111/ens.12189
7. 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 [Internet]. 2019;20(223). Disponible en: https://doi.org/10.1186/s12864-019-5590-8
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 [Internet]. 9 de octubre de 2020;11:565836. Disponible en: /pmc/articles/PMC7584159/
9. Puerma E, Orengo DJ, Aguadé M. The origin of chromosomal inversions as a source of segmental duplications in the Sophophora subgenus of *Drosophila*. Sci Rep [Internet]. 29 de julio de 2016;6. Disponible en: https://pubmed.ncbi.nlm.nih.gov/27470196/
10. Hoikkala A, Poikela N. Adaptation and ecological speciation in seasonally varying environments at high latitudes: *Drosophila virilis* group Anneli Hoikkala & Noora Poikela. Fly (Austin) [Internet]. 2022;16(1):85-104. Disponible en: https://www.tandfonline.com/action/journalInformation?journalCode=kfly20
11. Papaceit M, Segarra C, Aguadé M. Structure and population genetics of the breakpoints of a polymorphic inversion in *Drosophila subobscura*. Evolution (N Y) [Internet]. 1 de enero de 2013;67(1):66-79. Disponible en: https://onlinelibrary.wiley.com/doi/full/10.1111/j.1558-5646.2012.01731.x
12. 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. Hered 2015 1146 [Internet]. 25 de febrero de 2015;114(6):610-8. Disponible en: https://www.nature.com/articles/hdy20154

13. 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 [Internet]. 1 de septiembre de 2014;31(9):2331-41. Disponible en: https://academic.oup.com/mbe/article/31/9/2331/2925711
14. 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]. 1 de diciembre de 2019;9(1). Disponible en: /pmc/articles/PMC6368638/
15. Krimbas CB, Loukas M. Evolution of the obscura group drosophila species. I. Salivary chromosomes and quantitative characters in *D. subobscura* and two closely related species. Hered 1984 533 [Internet]. 1984;53(3):469-82. Disponible en: https://www.nature.com/articles/hdy1984109
16. Puerma E, Orengo DJ, Cruz F, Jèssica Gómez-Garrido, 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;10(8):1956-69. Disponible en: http://www.bioinformatics.babraham.ac.uk/
17. U.S. Global Change Research Program. Climate science special report: Fourth national climate assessment, volume I. Wuebbles DJ, Fahey DW, Hibbard KA, Dokken DJ, Stewart BC, Maycock TK, editores. US Glob Chang Res Progr [Internet]. 2018;1:470. Disponible en: http://www.climate.gov/news-features/understanding-climate/climate-change-global-temperature
18. Bracewell R, Chatla K, Nalley MJ, Bachtrog D. Dynamic turnover of centromeres drives karyotype evolution in *Drosophila*. Elife. 1 de septiembre de 2019;8.
19. Galludo M, Canals J, Pineda-Cirera L, Esteve C, Rosselló M, Balanyà J, et al. Climatic adaptation of chromosomal inversions in *Drosophila subobscura*. Genetica. 1 de octubre de 2018;146(4-5):433-41.
20. Zivanovic G, Arenas C, Mestres F. Adaptation of *Drosophila subobscura* chromosomal inversions to climatic variables: the Balkan natural population of Avala. Genetica. 1 de junio de 2021;149(3):155-69.
21. Simões P, Fragata I, Lopes-Cunha M, Lima M, Kellen B, Bárbaro M, et al. Wing trait-inversion associations in *Drosophila subobscura* can be generalized within continents, but may change through time. J Evol Biol [Internet]. 1 de diciembre de 2015;28(12):2163-74. Disponible en: https://pubmed.ncbi.nlm.nih.gov/26302686/
22. Simões P, Pascual M. Patterns of geographic variation of thermal adapted candidate genes in *Drosophila subobscura* sex chromosome arrangements. BMC Evol Biol [Internet]. 24 de abril de 2018;18(1). Disponible en: /pmc/articles/PMC5921438/
23. Puig Giribets M, Santos M, García Guerreiro MP. Basal hsp70 expression levels do not explain adaptive variation of the warm- and cold-climate O3 + 4 + 7 and OST gene arrangements of *Drosophila subobscura*. BMC Evol Biol [Internet]. 31 de enero de 2020;20(1). Disponible en: /pmc/articles/PMC6995229/

24. Laayouni H, García-Franco F, Chávez-Sandoval BE, Trotta V, Beltran S, Corominas M, et al. Thermal evolution of gene expression profiles in *Drosophila subobscura*. BMC Evol Biol [Internet]. 2007;7. Disponible en: https://pubmed.ncbi.nlm.nih.gov/17371595/
25. Lallemand T, Leduc M, Landès C, Rizzon C, Lerat E. An Overview of Duplicated Gene Detection Methods: Why the Duplication Mechanism Has to Be Accounted for in Their Choice. Genes (Basel) [Internet]. 2020;11(1046). Disponible en: www.mdpi.com/journal/genes
26. Schaeffer SW. Muller "Elements" in *Drosophila*: How the Search for the Genetic Basis for Speciation Led to the Birth of Comparative Genomics. Genetics [Internet]. 1 de septiembre de 2018;210(1):3. Disponible en: /pmc/articles/PMC6116959/
27. Farmacéutica Mexicana A, México López-López A, Gutiérrez L, Ulises A, Espuñes S, del Rosario T, et al. ¿Qué sabe usted acerca de...Genómica? Rev Mex Ciencias Farm [Internet]. 2005;36(1):42-4. Disponible en: https://www.redalyc.org/articulo.oa?id=57936107
28. Soderlund C, Bomhoff M, Nelson WM. SyMAP v3.4: a turnkey synteny system with application to plant genomes. Nucleic Acids Res [Internet]. 1 de mayo de 2011;39(10):e68-e68. Disponible en: https://academic.oup.com/nar/article/39/10/e68/1310457
29. Satsuma2 – Whole-genome Synteny Aligner – My Biosoftware – Bioinformatics Softwares Blog [Internet]. Disponible en: https://mybiosoftware.com/satsuma-whole-genome-synteny-aligner.html
30. Kiełbasa SM, Wan R, Sato K, Horton P, Frith MC. Adaptive seeds tame genomic sequence comparison. Genome Res [Internet]. 1 de marzo de 2011;21(3):487-93. Disponible en: https://genome.cshlp.org/content/21/3/487.full
31. Tello D, Natalia Gonzalez-Garcia L, Gomez J, Camilo Zuluaga-Monares J, Garcia R, Angel R, et al. NGSEP 4: Efficient and Accurate Identification of Orthogroups and Whole-Genome Alignment. bioRxiv [Internet]. 28 de enero de 2022;2022.01.27.478091. Disponible en: https://www.biorxiv.org/content/10.1101/2022.01.27.478091v1
32. *Drosophila subobscura* genome assembly UCBerk_Dsub_1.0 - NCBI - NLM [Internet]. Disponible en: https://www.ncbi.nlm.nih.gov/data-hub/genome/GCF_008121235.1/
33. *Drosophila subobscura* genome assembly dsubES12G5 - NCBI - NLM [Internet]. Disponible en: https://www.ncbi.nlm.nih.gov/data-hub/genome/GCA_903684685.1/
34. Haudry A, Laurent S, Kapun M. Population genomics on the fly: Recent advances in *Drosophila*. Methods Mol Biol [Internet]. 2020;2090:357-96. Disponible en: https://link.springer.com/protocol/10.1007/978-1-0716-0199-0_15
35. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol [Internet]. 1990;215(3):403-10. Disponible en: https://pubmed.ncbi.nlm.nih.gov/2231712/
36. 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. Disponible en: https://pubmed.ncbi.nlm.nih.gov/11070048/

37. Bachmann L, Raab M, Sperlich D. Satellite DNA and speciation: A species specific satellite DNA of *Drosophila guanche*. J Zool Syst Evol Res [Internet]. 1 de mayo de 1989;27(2):84-93. Disponible en: https://onlinelibrary.wiley.com/doi/full/10.1111/j.1439-0469.1989.tb00333.x
38. RepeatMasker Web Server [Internet]. Disponible en: https://www.repeatmasker.org/cgi-bin/WEBRepeatMasker
39. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. Nat Genet [Internet]. mayo de 2000;25(1):25. Disponible en: /pmc/articles/PMC3037419/
40. Camayd Viera I, Sautié Castellanos M, Zardón Navarro MA, Martínez Ortiz C, Hernández Cáceres JL. Un acercamiento a la ontología de genes y sus aplicaciones. Rev Cuba Informática Médica. 2010;2.
41. Tomczak A, Mortensen JM, Winnenburg R, Liu C, Alessi DT, Swamy V, et al. Interpretation of biological experiments changes with evolution of the Gene Ontology and its annotations. Sci Rep [Internet]. 1 de diciembre de 2018;8(1). Disponible en: /pmc/articles/PMC5865181/
42. Alexa A, Rahnenführer J. Gene set enrichment analysis with topGO. 2022; Disponible en: http://www.mpi-sb.mpg.de/~alexa
43. Pagès H, Aboyoun P, Gentleman R, DebRoy S. Biostrings: Efficient manipulation of biological strings. R package version 2.66.0 [Internet]. 2022. Disponible en: https://bioconductor.org/packages/Biostrings
44. Cantalapiedra CP, Herṇandez-Plaza A, Letunic I, Bork P, Huerta-Cepas J. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. Mol Biol Evol. 2021;38(12):5825-9.

8. Anexos

8.1. **Anexo I.** Código para el análisis de enriquecimiento GO en R.

```
# Se indica el directorio de trabajo donde se encuentran los archivos que se usarán en el anális
is

setwd("C:/Users/jorge/Desktop/UOC/TFM BM/R GO")

# Lectura de las anotaciones y obtención de la lista de genes en las inversiones U1 y U2 a parti
r de las anotaciones de la cepa UCBerk_Dsub_1.0 de D. subobscura
# https://www.ncbi.nlm.nih.gov/genome/64269?genome_assembly_id=666799


library(readr)
proteins_table <- read_csv("proteins_64269_666799.csv")

# Los genes de la inversión U1 se encuentran en el cromosoma U entre las posiciones 6419392-1384
6098

proteinsU1 <- proteins_table[proteins_table[1] == "chromosome U" & proteins_table$Start > 641939
2 & proteins_table$Stop < 13846098,]

# Para el análisis se necesita una lista de la ID de los genes contenidos en la inversión. Con e
l fin de obtener sólo una isoforma por gen, pero no perder posibles duplicaciones génicas, se fi
ltra para eliminar aquellos genes cuyas posiciones de inicio o final se encuentresn dulpicadas

U1list <- proteinsU1[!duplicated(proteinsU1$Start), ]
U1list <- U1list[!duplicated(U1list$Stop), ]

# Se sigue el mismo procedimiento para U2

proteinsU2 <- proteins_table[proteins_table[1] == "chromosome U" & proteins_table$Start > 138694
64 & proteins_table$Stop < 21253679,]

U2list <- proteinsU2[!duplicated(proteinsU2$Start), ]
U2list <- U2list[!duplicated(U2list$Stop), ]

library(readr)

# Se carga el archivo obtenido de eggNOG-mapper

eggnog_data <- read_tsv("eggnog_data.tsv", skip = 4)
```

EggNOG-mapper es una aplicación web disponible en http://eggnog-mapper.embl.de/ La aplicación permite obtener los términos GO de un genoma suministrando secuencias de proteínas, secuencias codificantes o secuencias genómicas. Para obtener el archivo usado en este estudio se han utilizado las secuencias de proteínas de la cepa de Berkeley. En los resultados no viene indicado a qué gen pertenencen, por lo que se procede a identificar el gen a partir del número de acceso de la proteína con la información de las anotaciones de la cepa.

```
library(readr)
library(dplyr)

# FIltramos en las anotaciones de la cepa aquellas que se encuentran en el resultado de eggNOG-m
apper

proteins_table2 <- filter(proteins_table, proteins_table$`Protein product` %in% eggnog_data$`#qu
ery`)


# Hay 36 proteínas que no han podido ser anotadas con la aplicacción eggNOG-mapper

# Se añade la información de la ID del gen a los resultados de eggNOG en una nueva columna

eggnog_data2 <- filter(eggnog_data, eggnog_data$`#query` %in% proteins_table2$`Protein product`)

eggnog_data2 <- eggnog_data2[order(eggnog_data2$`#query`),]
```

```
proteins_table2 <- proteins_table2[order(proteins_table2$`Protein product`),]

eggnog_data2$GeneID <- proteins_table2$GeneID


# Eliminamos los id de genes duplicados para quedarnos sólo con un tránscrito por gen

eggnog_data2 <- eggnog_data2[!duplicated(eggnog_data2$GeneID), ]
```

La herramienta topGO permite realizar un análisis de ontología genética aportando una lista de anotaciones de términos GO personalizada, permitiendo así realizar el análisis con organismos no-modelo de los que no se dispone de bases de datos con anotaciones GO públicas.

```
# Instalación del paquete y carga de los paquetes necesarios
if (!require("BiocManager", quietly = TRUE))
    install.packages("BiocManager")
BiocManager::install("topGO")


library(tidyverse)
library(topGO)
library(lintr)
library(lattice)
library(dplyr)

# Los datos de entrada para topGO consisten en una lista de genes de interés y el resto de genes
con los que queremos comparar con los términos GO asociados a esos genes (universo de genes).
# Se genera un objeto que contenga la lista de los genes con sus términos GO asociados.

geneUniverse <- eggnog_data2[c(22,10)]
write.table(geneUniverse, file = "geneUniverse.txt", append = FALSE, sep = "\t", row.names = FAL
SE, col.names = T)

# Modificar archivo geneUniverse.txt para que tenga el formato correcto, eliminando comillas y e
l encabezado y reemplazando las comas  por  coma más un espacio.

# Para la lectura del archivo generado con el universo de genes se utiliza la función readMappin
gs. El universo de genes se corresponde con todos aquellos de los que se dispone de anotaciones
GO

geneID2GO <- readMappings(file = "geneUniverseMod.txt")
geneUniverse <- names(geneID2GO)

# Posteriormente, se proporciona la lista de genes de interés.

U1genes <- as.character(U1list$`GeneID`)
geneListU1 <- factor(as.integer(geneUniverse %in% U1genes))
names(geneListU1) <- geneUniverse

U2genes <- as.character(U2list$GeneID)
geneListU2 <- factor(as.integer(geneUniverse %in% U2genes))
names(geneListU2) <- geneUniverse

# Una vez definidos los genes de interés y el universo de genes, se une todo en un objeto. En es
te punto se puede indicar la ontología genética de nuestro interés: "BP": Proceso biológico, "CC
": Localización en componentes celulares o "MF": Función molecular

U1GOdata <- new("topGOdata", description="Dsub U1", ontology="BP", allGenes=geneListU1,  annot =
annFUN.gene2GO, gene2GO = geneID2GO)

# Realización del análisis de enriquecimiento de términos GO

resultFisherU1 <- runTest(U1GOdata, algorithm="classic", statistic="fisher")

# Es posible generar una tabla con los resultados obtenidos. En nuestro caso, se genera una tabl
a con los 20 primeros resultados por orden de significancia

U1Res <- GenTable(U1GOdata, classicFisher = resultFisherU1, orderBy = "resultFisher", ranksOf =
"classicFisher", topNodes = 20)

write.table(U1Res, file = "U1Res.tsv", append = FALSE, sep = "\t", row.names = F, col.names = T)
```

```
# Configuración para el resultado gráfico

dev.off()

layout(t(1:2), widths=c(8,1))
par(mar=c(4, .5, .7, .7), oma=c(3, 15, 3, 4), las=1)

# Se genera una paleta de color continua

library(colorRamps)
rbPal <- colorRampPalette(c ('red', 'white', 'blue' ))
pvalue <- as.numeric (gsub ("<", "", U1Res$classicFisher)) # Elimina los símnolos '<'

max_value <- as.integer(max(-log(pvalue)))+1
pv_range <- exp(-seq(max_value, 0, -1))
U1Res$Color <- rbPal(max_value)[cut(pvalue, pv_range)]

# Genera la figura

o <- order (U1Res$Significant, decreasing = TRUE)
barplot (U1Res$Significant[o], names.arg=U1Res$Term[o], las=2, horiz= TRUE, col=U1Res$Color[o],x
lab= "Number of sequences", main="U1 Biological process", sub=geneUniverse, cex.names=0.85)
```

```
# Genera la leyenda

image (0, seq(1, max_value), t(seq_along(seq(1, max_value) )), col=rev (rbPal(max_value) ), axes
=FALSE, ann=FALSE)
pv_label <- exp(-seq(log(1), -log(min(pvalue)), l=6))
pv_label <- formatC(pv_label, format = "e", digits = 2)
axis(4, at=seq(1, max_value, length=6), labels=c(1, pv_label[2:6]), cex.axis=0.85)
title ("p-value", cex.main = 0.8)
```

Se realiza el mismo análisis para la inversión U_2

```
U2GOdata <- new("topGOdata", description="Dsub U2", ontology="BP", allGenes=geneListU2,  annot =
annFUN.gene2GO, gene2GO = geneID2GO)

resultFisherU2 <- runTest(U2GOdata, algorithm="classic", statistic="fisher")

U2Res <- GenTable(U2GOdata, classicFisher = resultFisherU2, orderBy = "resultFisher", ranksOf =
"classicFisher", topNodes = 20)

write.table(U2Res, file = "U2Res.tsv", append = FALSE, sep = "\t", row.names = F, col.names = T)

dev.off()

layout(t(1:2), widths=c(8,1))
par(mar=c(4, .5, .7, .7), oma=c(3, 15, 3, 4), las=1)

library(colorRamps)
rbPal <- colorRampPalette(c ('red', 'white', 'blue' ))
pvalue <- as.numeric (gsub ("<", "", U2Res$classicFisher))
max_value <- as.integer(max(-log(pvalue)))+1
pv_range <- exp(-seq(max_value, 0, -1))
U2Res$Color <- rbPal(max_value)[cut(pvalue, pv_range)]

o <- order (U2Res$Significant, decreasing = TRUE)
barplot (U2Res$Significant[o], names.arg=U2Res$Term[o], las=2, horiz= TRUE, col=U2Res$Color[o],x
lab= "Number of sequences", main="U2 Biological process", sub=geneUniverse, cex.names=0.85)

image (0, seq(1, max_value), t(seq_along(seq(1, max_value) )), col=rev (rbPal(max_value) ), axes
=FALSE, ann=FALSE)
pv_label <- exp(-seq(log(1), -log(min(pvalue)), l=6))
pv_label <- formatC(pv_label, format = "e", digits = 2)
axis(4, at=seq(1, max_value, length=6), labels=c(1, pv_label[2:6]), cex.axis=0.85)
title ("p-value", cex.main = 0.8)
```

```
# U1 MF

U1GOdataMF <- new("topGOdata", description="Dsub U1", ontology="MF", allGenes=geneListU1,  annot
= annFUN.gene2GO, gene2GO = geneID2GO)

resultFisherU1MF <- runTest(U1GOdataMF, algorithm="classic", statistic="fisher")

U1ResMF <- GenTable(U1GOdataMF, classicFisher = resultFisherU1MF, orderBy = "resultFisher", rank
sOf = "classicFisher", topNodes = 20)

write.table(U1ResMF, file = "U1ResMF.tsv", append = FALSE, sep = "\t", row.names = F, col.names
= T)

# U1 CC

U1GOdataCC <- new("topGOdata", description="Dsub U1", ontology="CC", allGenes=geneListU1,  annot
= annFUN.gene2GO, gene2GO = geneID2GO)

resultFisherU1CC <- runTest(U1GOdataCC, algorithm="classic", statistic="fisher")

U1ResCC <- GenTable(U1GOdataCC, classicFisher = resultFisherU1CC, orderBy = "resultFisher", rank
sOf = "classicFisher", topNodes = 20)


write.table(U1ResCC, file = "U1ResCC.tsv", append = FALSE, sep = "\t", row.names = F, col.names
= T)

# U2 MF

U2GOdataMF <- new("topGOdata", description="Dsub U1", ontology="MF", allGenes=geneListU2,  annot
= annFUN.gene2GO, gene2GO = geneID2GO)

resultFisherU2MF <- runTest(U2GOdataMF, algorithm="classic", statistic="fisher")

U2ResMF <- GenTable(U2GOdataMF, classicFisher = resultFisherU2MF, orderBy = "resultFisher", rank
sOf = "classicFisher", topNodes = 20)

write.table(U2ResMF, file = "U2ResMF.tsv", append = FALSE, sep = "\t", row.names = F, col.names
= T)

# U2 CC

U2GOdataCC <- new("topGOdata", description="Dsub U2", ontology="CC", allGenes=geneListU2,  annot
= annFUN.gene2GO, gene2GO = geneID2GO)

resultFisherU2CC <- runTest(U2GOdataCC, algorithm="classic", statistic="fisher")

U2ResCC <- GenTable(U2GOdataCC, classicFisher = resultFisherU2CC, orderBy = "resultFisher", rank
sOf = "classicFisher", topNodes = 20)

write.table(U2ResCC, file = "U2ResCC.tsv", append = FALSE, sep = "\t", row.names = F, col.names
= T)
```

8.2. Anexo II. Resultado de la anotación de términos GO del genoma ensamblado UC_Berk_dsub de *D. subobscura*.

Los resultados se encuentran disponibles para su descarga en el repositorio:

https://github.com/Jorge-Jordan/TFM-BM-UOC

En el repositorio se incluyen los siguientes archivos:

1. DsubGO.Rmd - Código utilizado para el análisis de enriquecimiento de términos GO

2. GCF_008121235.1_UCBerk_Dsub_1.0_protein.faa – Secuencias proteicas del ensamblado UCBerk_Dsub_1.0 de *Drosophila subobscura*

3. eggnog_data.rar - Resultados de la anotación de términos GO para las proteínas del ensamblado UCBerk_Dsub_1.0 de *Drosophila subobscura* obtenido con eggNOG-mapper

4. proteins_64269_666799.csv - Anotaciones del ensamblado UCBerk_Dsub_1.0 de *Drosophila subobscura*

8.3. Anexo III. Resultados del análisis de enriquecimiento de términos GO para los genes contenidos en las inversiones U_1 y U_2 de *D. subobscura*.

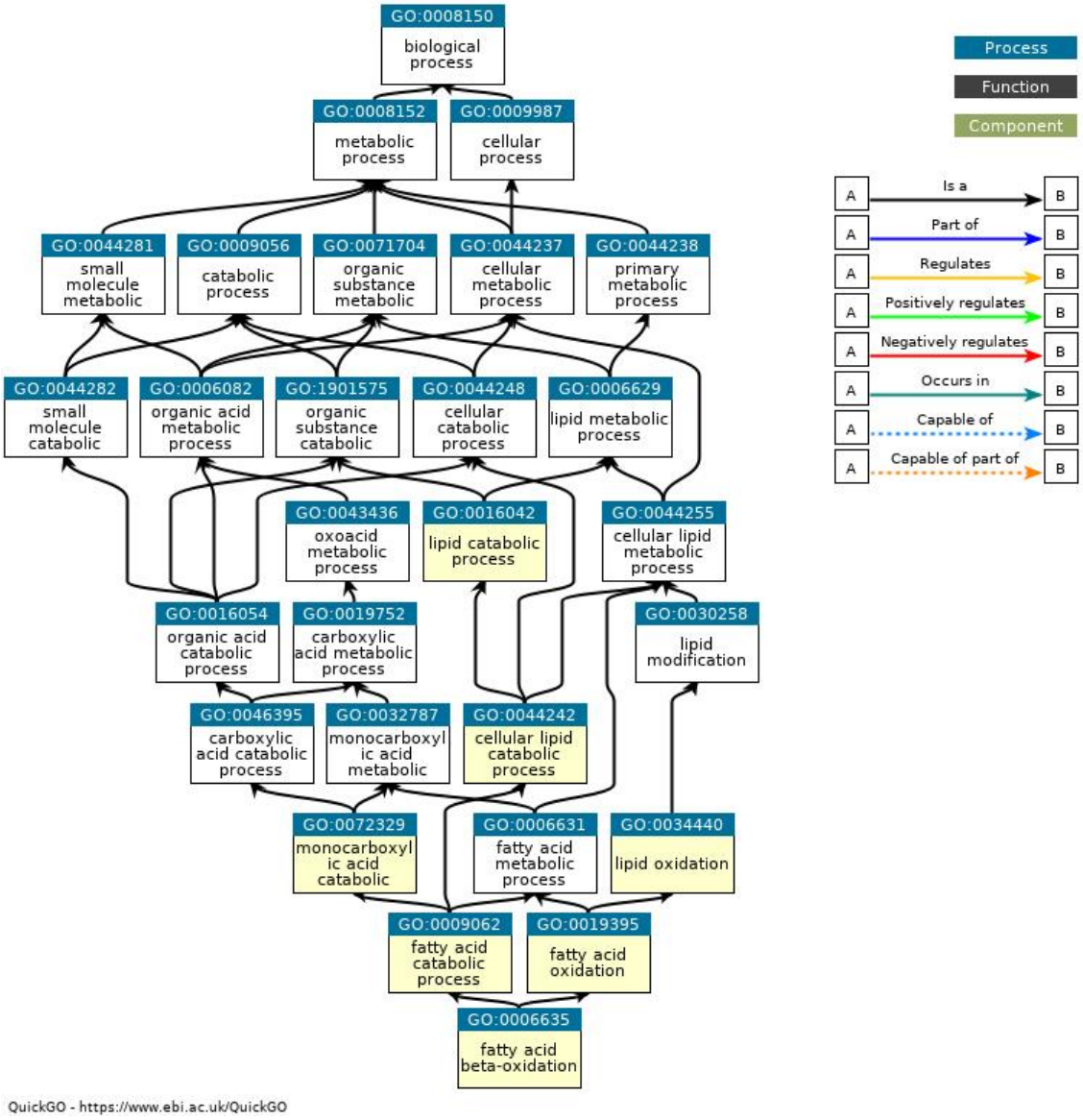

Figura 13 - Relación de términos GO para el término **GO:0006635 fatty acid beta-oxidation**. Se remarcan en amarillo los términos relacionados para los que se ha encontrado enriquecimiento en la inversión U_1.

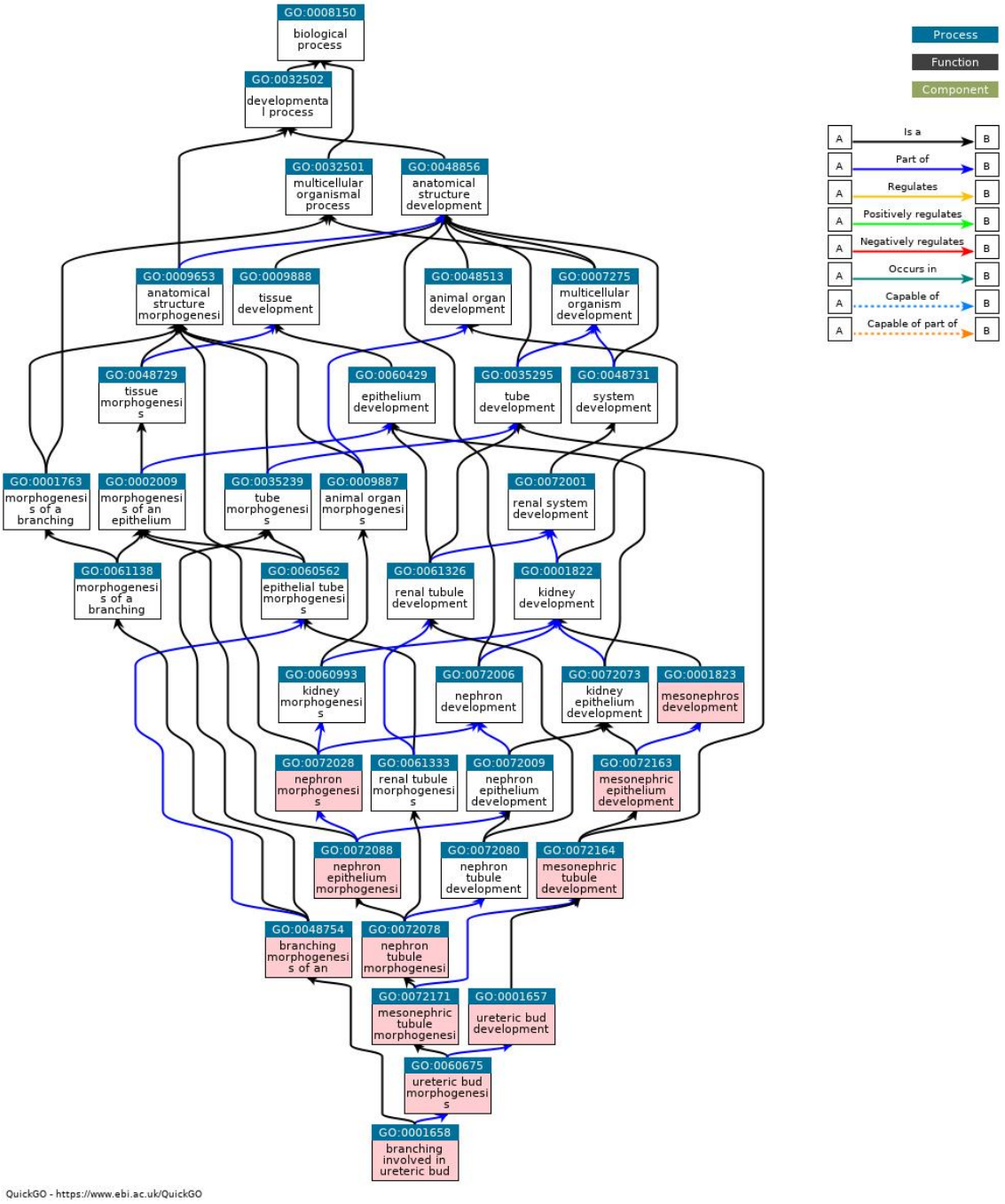

Figura 14 - Relación de términos GO para el término **GO:0001658 branching involved in ureteric bud morphogenesis**. Se remarcan en rojo los términos relacionados para los que se ha encontrado enriquecimiento en la inversión U_1.

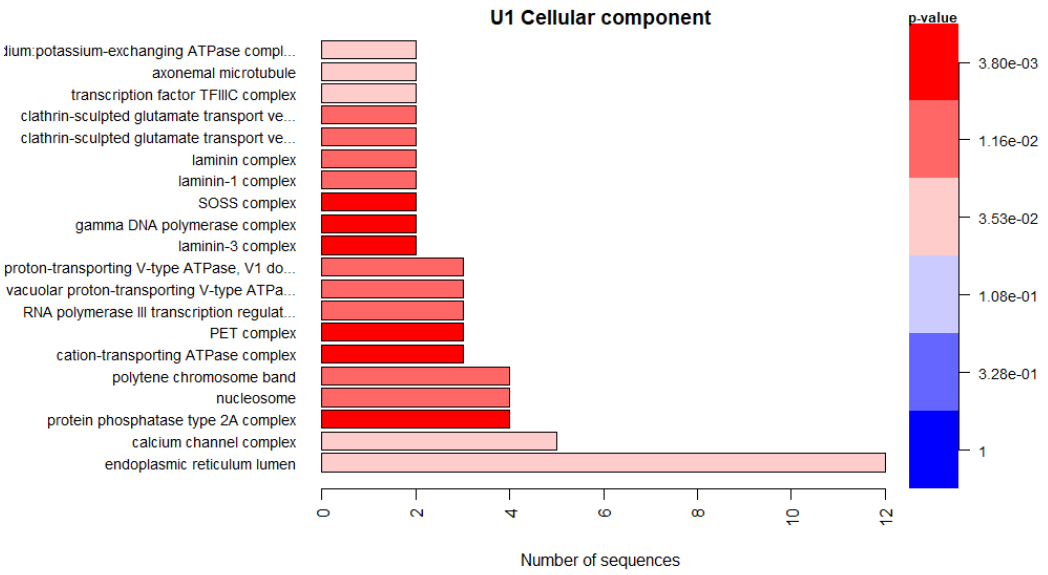

Figura 15 - Análisis de enriquecimiento de términos GO para componente celular de los genes codificantes de proteína anotados en la inversión U_1 de *D. subobscura*_Dsub_1.0. En el eje X se indica el número de genes encontrados que contienen el término anotado.

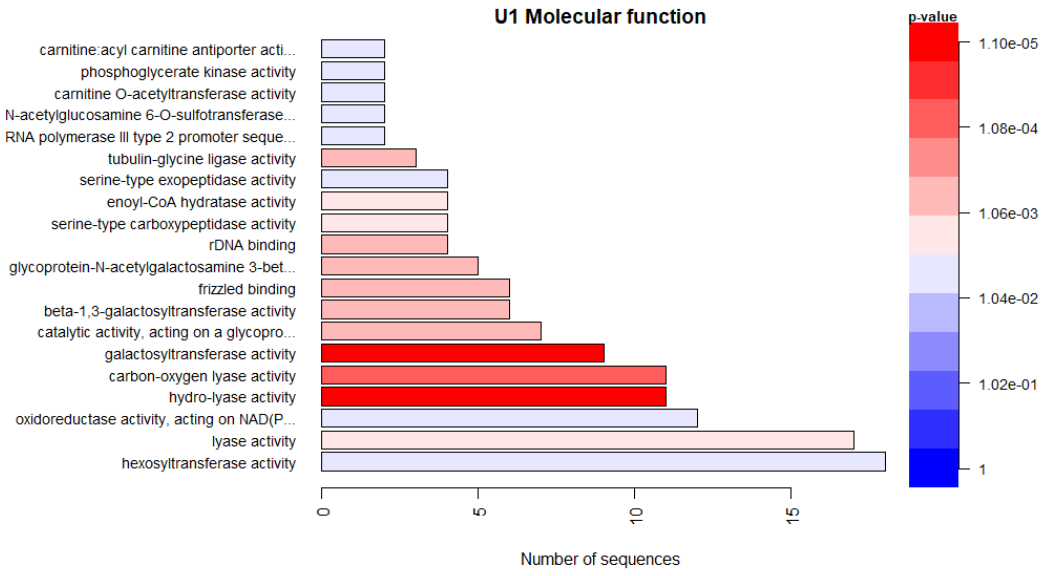

Figura 16 - Análisis de enriquecimiento de términos GO para función molecular de los genes codificantes de proteína anotados en la inversión U_1 de *D. subobscura*_Dsub_1.0. En el eje X se indica el número de genes encontrados que contienen el término anotado.

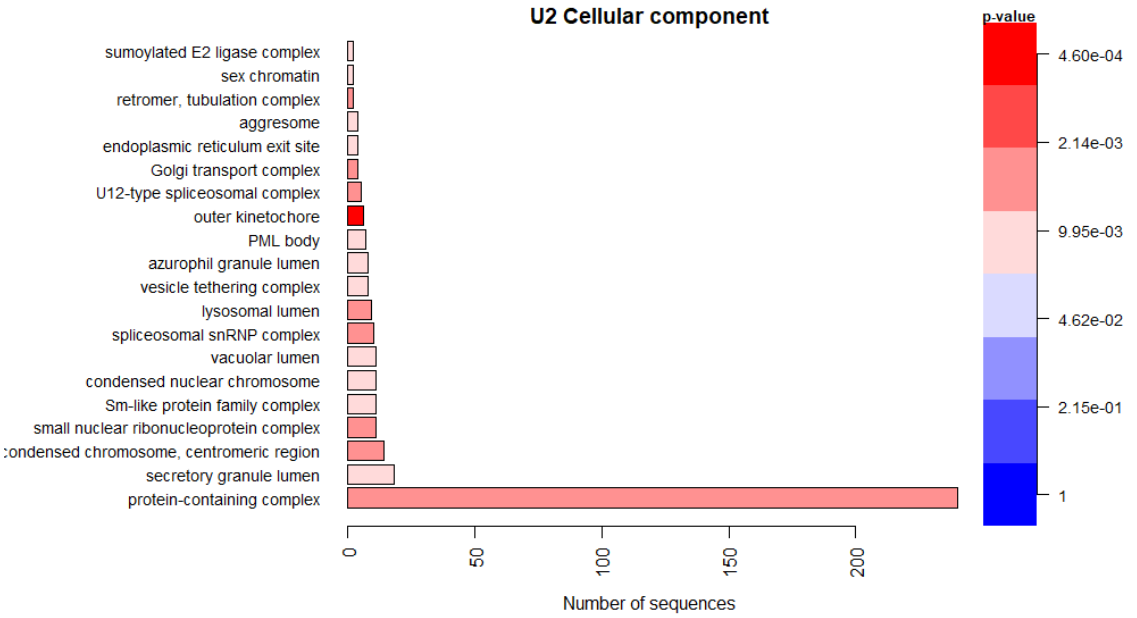

Figura 17 - Análisis de enriquecimiento de términos GO para componente celular de los genes codificantes de proteína anotados en la inversión U_2 de *D. subobscura*_Dsub_1.0. En el eje X se indica el número de genes encontrados que contienen el término anotado.

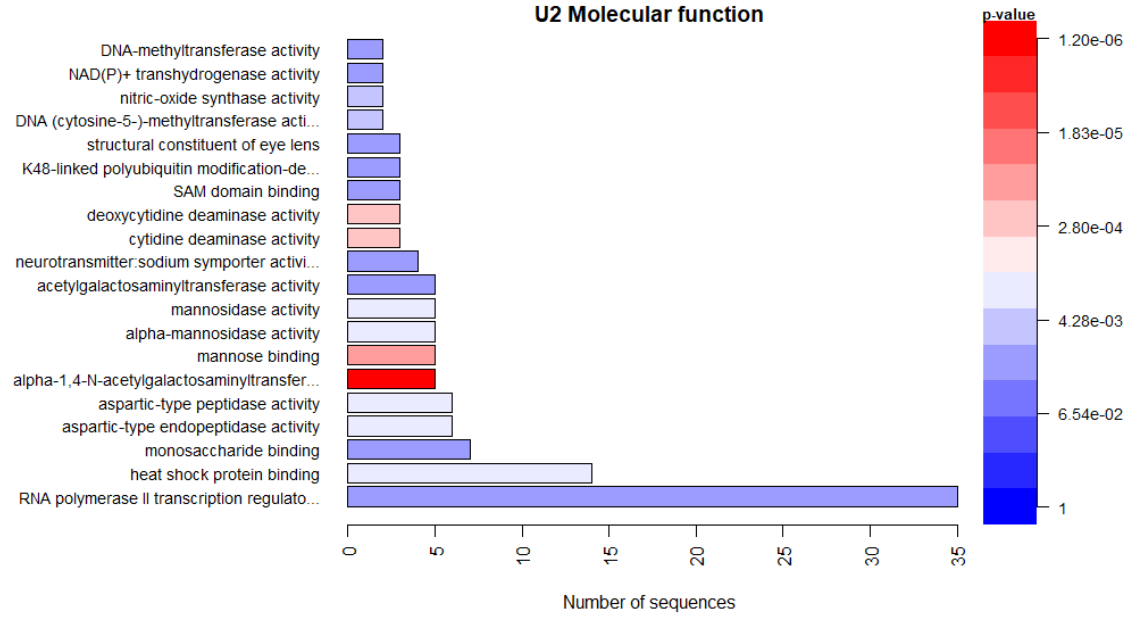

Figura 18 - Análisis de enriquecimiento de términos GO para función molecular de los genes codificantes de proteína anotados en la inversión U_2 de *D. subobscura*_Dsub_1.0. En el eje X se indica el número de genes encontrados que contienen el término anotado.