PONTIFICIA UNIVERSIDAD CATÓLICA DEL ECUADOR

FACULTAD DE INGENIERÍA

MAESTRÍA EN BIOLOGÍA COMPUTACIONAL

# Ensamble de novo y caracterización del genoma en *Drosophila rucux* (*Diptera*, *Drosophilidae*)

Kevin Josue Casal Arroyo

LABORATORIO DE GENÉTICA EVOLUTIVA

Declaración de Derechos de Autor

© [2024] [KEVIN JOSUÉ CASAL ARROYO]. Todos los derechos reservados.

Ninguna parte de esta tesis puede ser reproducida, almacenada en un sistema de recuperación, o transmitida en ninguna forma o por ningún medio, ya sea electrónico, mecánico, fotocopia, grabación, o de otro tipo, sin el permiso previo por escrito del autor. Los datos fueron obtenidos gracias al laboratorio de genética evolutiva, encargado por la Dra. Doris Vela.

Quito, 22 de agosto de 2024

Sr. Mgt.

Charles Escobar

DECANO FACULTAD DE INGENIERÍA

PONTIFICIA UNIVERSIDAD CATÓLICA DEL ECUADOR

De mis consideraciones:

Se envía el Informe correspondiente a la tutoría realizada al Trabajo de Titulación que se detalla a continuación:

| | | |
|---|---|---|
| TITULO DEL TRABAJO DE INTEGRACIÓN CURRICULAR | Ensamble de novo y caracterización del genoma en Drosophila rucux (Diptera, Drosophilidae) | |
| DIRECTOR | Nombre | Cédula |
| | Doris Vela | 1712651197 |
| ESTUDIANTE(S) | Nombre | Cédula |
| | Kevin Josue Casal Arroyo | 1718724345 |

Se informa que el trabajo ha cumplido con todos los parámetros establecidos, mediante el cual el Maestrante demuestra el desarrollo de competencias en el campo de conocimiento de su profesión y presenta una propuesta en el área de conocimiento, con un nivel de argumentación coherente.

Dando por concluida esta tutoría de trabajo de titulación, CERTIFICO, para los fines pertinentes, que el Maestrante está apto para continuar con el proceso de LECTURA-EVALUACIÓN.

Atentamente,

Dra. Doris Vela

Directora de Trabajo de titulación

# Turnitin Informe de Originalidad

Procesado el: 22-ago.-2024 08:51 -
Identificador: 2436108426
Número de palabras: 6912
Entregado: 1

tesis MBC Por Casal Kevin

Índice de similitud

0%

Similitud según fuente

Internet Sources: 0%
Publicaciones: 0%
Trabajos del estudiante: 0%

PONTIFICIA UNIVERSIDAD CATÓLICA DEL ECUADOR FACULTAD DE INGENIERÍA MAESTRÍA EN BIOLOGÍA COMPUTACIONAL Ensamble de novo y caracterización del genoma en Drosophila rucux (Diptera, Drosophilidae) Kevin Casal LABORATORIO DE GENÉTICA EVOLUTIVA Declaración de Derechos de Autor © [2024] [KEVIN JOSUÉ CASAL ARROYO]. Todos los derechos reservados. Ninguna parte de esta tesis puede ser reproducida, almacenada en un sistema de recuperación, o transmitida en ninguna forma o por ningún medio, ya sea electrónico, mecánico, fotocopia, grabación, o de otro tipo, sin el permiso previo por escrito del autor. Los datos fueron obtenidos gracias al laboratorio de genética evolutiva, encargado por la Dra. Doris Vela. Aprobación del Director del Trabajo de Titulación Hoja de evidencia antiplagio (Informe Turnitin) Agradecimientos Quiero expresar mi más sincero agradecimiento a mi madre, cuyo apoyo incondicional y confianza en mis decisiones han sido fundamentales para alcanzar mis metas. Su aliento constante y amorosa orientación me han dado la fuerza para seguir adelante en cada paso de este viaje académico. Agradezco profundamente a la Doctora Doris Vela por aceptarme en su laboratorio y brindarme la oportunidad de desarrollarme como investigador. Su ejemplo como científica y su guía experta han sido invaluables en mi formación y en el progreso de este proyecto. Mi gratitud también va para Jhael Ortega, una mujer ejemplar que me desafió a superarme y no rendirme. Su apoyo en los momentos más críticos y sus palabras de aliento fueron cruciales para continuar adelante y enfrentar los desafíos con determinación. A los miembros del Laboratorio de Genética Evolutiva, agradezco su colaboración, su apoyo técnico y sus valiosas discusiones que han enriquecido este trabajo. Su dedicación y conocimiento han sido esenciales para la realización de esta investigación. Agradezco también a la Pontificia Universidad Católica del Ecuador por proporcionarme los recursos y el entorno académico que hicieron posible este estudio. Su infraestructura y apoyo institucional han sido fundamentales para el desarrollo de mi investigación. Finalmente, agradezco a mis lectores por su interés en mi trabajo. Su atención y lectura son el reconocimiento más gratificante para este esfuerzo académico. Dedicatoria Dedico esta tesis a las tres mujeres que han tenido un gran impacto en mi vida: a mi madre, por su amor incondicional y su fe constante en mis capacidades; a la Doctora Doris Vela, por ser un modelo a seguir y ofrecerme la oportunidad de crecer como investigador; y a Jhael Ortega, por su inspiración y apoyo inquebrantable en los momentos más difíciles. Esta obra es un testimonio de su influencia y apoyo en mi vida, y su contribución ha sido esencial para el logro de este objetivo académico. Índice General Resumen

## Agradecimientos

Quiero expresar mi más sincero agradecimiento a mi madre, cuyo apoyo incondicional y confianza en mis decisiones han sido fundamentales para alcanzar mis metas. Su aliento constante y amorosa orientación me han dado la fuerza para seguir adelante en cada paso de este viaje académico.

Agradezco profundamente a la Doctora Doris Vela por aceptarme en su laboratorio y brindarme la oportunidad de desarrollarme como investigador. Su ejemplo como científica y su guía experta han sido invaluables en mi formación y en el progreso de este proyecto.

Mi gratitud también va para Jhael Ortega, una mujer ejemplar que me desafió a superarme y no rendirme. Su apoyo en los momentos más críticos y sus palabras de aliento fueron cruciales para continuar adelante y enfrentar los desafíos con determinación.

A los miembros del Laboratorio de Genética Evolutiva, agradezco su colaboración, su apoyo técnico y sus valiosas discusiones que han enriquecido este trabajo. Su dedicación y conocimiento han sido esenciales para la realización de esta investigación.

Agradezco también a la Pontificia Universidad Católica del Ecuador por proporcionarme los recursos y el entorno académico que hicieron posible este estudio. Su infraestructura y apoyo institucional han sido fundamentales para el desarrollo de mi investigación.

Finalmente, agradezco a mis lectores por su interés en mi trabajo. Su atención y lectura son el reconocimiento más gratificante para este esfuerzo académico.

## Dedicatoria

Dedico esta tesis a las tres mujeres que han tenido un gran impacto en mi vida: a mi madre, por su amor incondicional y su fe constante en mis capacidades; a la Doctora Doris Vela, por ser un modelo a seguir y ofrecerme la oportunidad de crecer como investigador; y a Jhael Ortega, por su inspiración y apoyo inquebrantable en los momentos más difíciles. Esta obra es un testimonio de su influencia y apoyo en mi vida, y su contribución ha sido esencial para el logro de este objetivo académico.

Índice General

Índice de Figuras

Índice de Tablas

## Resumen

El género Drosophila es un excelente modelo para la investigación en ecología y evolución, debido a su amplia diversidad y adaptaciones en diferentes ecosistemas. Sin embargo, la información disponible sobre la importancia de las especies dentro de *Drosophila*, a excepción de *D. melanogaster*, es limitada. Este estudio se centra en la obtención y análisis del genoma de *D. rucux*, una especie menos conocida, con el propósito de contribuir a una mejor comprensión de la diversidad genética y los patrones de adaptación en este grupo. Utilizando la herramienta de ensamblaje Masurca, se logró realizar un ensamblaje de novo del genoma (182 mb), y mediante el pipeline de FunAnnotate y Augustus, se predijeron y anotaron 19,805 secuencias codificadoras de proteínas. El análisis de elementos móviles reveló que aproximadamente el 17.5% del genoma se clasifica como secuencias repetitivas, incluyendo elementos de Clase I (como LINE, SINE, LTR y TRIM) y elementos de Clase II (como Helitron, MITE, MAVERICK y TIR). Los elementos de Clase I representaron el 25.8% de la muestra, con los elementos LTR siendo los más abundantes. La comparación con datos de *D. melanogaster* y otras especies del género indica que *D. rucux* presenta un porcentaje de elementos transponibles similar al de otras especies en el género.

Los hallazgos de este estudio proporcionan una base valiosa para explorar las características ecológicas de *D. rucux*. Además, el análisis filogenético que se puede realizar con el genoma ensamblado permitirá investigar las relaciones evolutivas dentro del género *Drosophila* y su contexto evolutivo más amplio. Así, este trabajo no solo amplía nuestro conocimiento sobre la biología de *D. rucux*, sino que también sienta las bases para futuras investigaciones que busquen entender mejor las dinámicas ecológicas y evolutivas de este género fascinante.

**Palabras clave:** *Drosophila rucux, Genome assembly, Secuenciación Illumina, Anotaciones*

## Abstract

The genus *Drosophila* is an excellent model for research in ecology and evolution due to its wide diversity and adaptations in different ecosystems. However, the available information regarding the importance of species within *Drosophila*, with the exception of *D. melanogaster*, is limited. This study focuses on obtaining and analyzing the genome of *D. rucux*, a lesser-known species, with the aim of contributing to a better understanding of genetic diversity and adaptation patterns within this group. Using the Masurca assembly tool, a de novo genome assembly (182 Mb) was successfully performed, and through the FunAnnotate and Augustus pipeline, 19,805 protein-coding sequences were predicted and annotated. The analysis of mobile elements revealed that approximately 17.5% of the genome is classified as repetitive sequences, including Class I elements (such as LINE, SINE, LTR, and TRIM) and Class II elements (such as Helitron, MITE, MAVERICK, and TIR). Class I elements accounted for 25.8% of the sample, with LTR elements being the most abundant. Comparison with data from *D. melanogaster* and other species in the genus indicates that *D. rucux* exhibits a similar percentage of transposable elements as other species within the genus.

The findings of this study provide a valuable foundation for exploring the ecological characteristics of *D. rucux*. Additionally, the phylogenetic analysis that can be performed with the assembled genome will allow for the investigation of evolutionary relationships within the *Drosophila* genus and its broader evolutionary context. Thus, this work not only expands our knowledge of *D. rucux* biology but also lays the groundwork for future research aimed at better understanding the ecological and evolutionary dynamics of this fascinating genus.

**Key Words:** *Drosophila rucux, Genome assembly, Secuenciación Illumina, Anotaciones*

# 1. Introducción

## 1.1. Diversidad de *Drosophila* en el Ecuador y en la región neotropical

La familia *Drosophilidae* (*Diptera*) comprende alrededor de 3962 especies divididas en más de 70 géneros (O'Grady & DeSalle, 2018). Aunque la clasificación de algunos géneros puede generar controversias, se suele aceptar que todos se distribuyen en dos subfamilias, Steganinae y Drosophilinae (Brake & Bächli, 2008). El género *Drosophila* se subdivide en 8 subgéneros, entre los cuales destacan el subgénero *Sophophora*, que alberga especies como *Drosophila melanogaster*, y el subgénero *Drosophila*, considerado el subgénero más abundante (Figuero, 2017). Según la propuesta de Throckmorton (1975), que posteriormente fue corroborada por O'Grady y DeSalle (2018), en el género *Drosophila* se han identificado dos radiaciones evolutivas significativas: la radiación immigrans-Hirtodrosophila y la radiación virilis-repleta, ambas con origen en regiones tropicales.

En el Neotrópico, el subgénero Drosophila es el grupo más numeroso dentro de la familia Drosophilidae, comprendiendo al menos 46 grupos de especies (O'Grady & DeSalle, 2018). En Ecuador, se ha documentado una notable diversidad de especies de la familia Drosophilidae, con un registro aproximado de 150 especies (Figuero, 2017). Estas especies se distribuyen desde las zonas costeras hasta las regiones amazónicas y de alta montaña (Rafael & Figuero, 2013). La diversidad de Drosophilidae en el país resalta la importancia de realizar estudios de biodiversidad y taxonomía para comprender mejor estas especies en su hábitat natural.

## 1.2. *D. rucux* y el grupo *mesophragmatica*

El grupo *mesophragmatica*, descrito por primera vez por Brncic y Santibañez en 1957, se destaca por ser exclusivamente neotropical y endémico de América del Sur. Actualmente está conformado por 17 especies (O'Grady & DeSalle., 2018), todas ellas presentes en bosques andinos y páramos de la región. Un ejemplo de esta diversidad es *Drosophila rucux*, la cual se distribuye en las estribaciones del Volcán Ruco Pichincha, (Pichincha, Ecuador), y también se ha registrado en otras localidades como el Antisana, Cerro Peñas Blancas y Papallacta, ubicados en la provincia del Napo (Rafael & Figuero, 2013). Parte de la característica distintiva de este grupo es la presencia de cerdas escutelares divergentes, lo que contribuye a su identificación y estudio dentro del contexto de la biodiversidad de la región.

*Drosophila rucux* Céspedes & Rafael, 2012 es una especie de mosca frutícola, de tamaño entre 3.0 a 3.5 mm, de color marrón claro y marrón oscuro (Figura 1A-D), lejanamente relacionada con el modelo de laboratorio *Drosophila melanogaster*. Es una de las especies actualmente descritas que habitan en los Andes y, al igual que muchas de las especies pertenecientes al grupo *mesophragmatica*, se caracterizan por su morfología externa similar y su amplia distribución en los neotrópicos (Brncic & Santibañez, 1957; Céspedes & Rafael, 2012). *D. rucux* es una especie nativa, los adultos pueden ser recogidos y criados fácilmente con redes de mano al igual que con trampas de botella y medios de levadura, gelatina, limón y nipagina (Céspedes & Rafael, 2012).

### 1.3. Información sobre el genoma de especies neotropicales

El estudio del genoma de especies neotropicales ha cobrado especial relevancia en los últimos años debido a la rica diversidad biológica que albergan estas regiones (Brown, 2014). Investigaciones genómicas han revelado una gran variedad de adaptaciones genéticas que permiten a estas especies sobrevivir en entornos tropicales, incluyendo resistencia a enfermedades, tolerancia a condiciones extremas de temperatura y humedad, y capacidades de reproducción y dispersión únicas (Li et al., 2022). Además, el análisis comparativo de genomas de especies neotropicales ha arrojado luz sobre la evolución y la historia natural de estos organismos, así como sobre su papel en la conservación de la biodiversidad y en la comprensión de los impactos del cambio climático y la deforestación en sus poblaciones (O'Grady & DeSalle., 2018).

A pesar de la gran diversidad y endemismo de las especies de Drosophila en el Neotrópico, existe una falta de estudios que aborden aspectos moleculares, especialmente en lo que respecta a especies ecuatorianas. La información disponible sobre el genoma de especies neotropicales es limitada, aunque se han logrado importantes avances en ciertas especies como *Drosophila willistoni*, que fue una de las primeras especies neotropicales de Drosophila cuyo genoma fue secuenciado (Bergman, 2002). Estudio que proporcionó valiosa información sobre la genómica de la adaptación y evolución en esta especie. Otras especies, como *Drosophila tropicalis* y *Drosophila mojavensis,* han sido objeto de análisis genómicos con el propósito de comprender su diversidad genética y los mecanismos de adaptación a entornos tropicales. En particular, *D. virilis* ha sido estudiada debido a su capacidad para adaptarse a una variedad de entornos, lo que ha revelado información valiosa sobre sus características biológicas y su habilidad para colonizar nuevos hábitats (Miral, 2008). En la actualidad, la investigación genómica sigue siendo una

herramienta fundamental para la conservación y manejo de las especies neotropicales, así como para la formulación de estrategias de adaptación y mitigación frente a los desafíos ambientales actuales.

El objetivo principal de este estudio es llevar a cabo un ensamblaje de novo del genoma de *Drosophila rucux* (Diptera) y caracterizarlo a través de la identificación y anotación de los genes y elementos funcionales presentes en el genoma.

## 2. Materiales y métodos

### 2.1. Análisis de calidad de las secuencias y ensamble de novo

Se obtuvieron las librerías de estudios previos, las cuales fueron generadas a través de Illumina True-Seq Nano DNA. Las lecturas crudas (Raw reads) fueron procesadas para remover las secuencias de los adaptadores y remover las secuencias de baja calidad mediante Platanus_trim_v1.2.4 (Kajitani *et al*., 2014), se utilizó FastQC para poder comprobar la calidad de las librerías y corroborar la remoción de adaptadores y secuencias de baja calidad. El ensamble del genoma se lo realizó mediante Masurca_v4.1.1 (Zimin *et al*., 2013) y los gaps que se encontraban en los contigs fueron sellados por GapFiller_v1.1.0 (Boetzer et al., 2012). Las estadísticas del ensamble fueron generadas por QUAST (Quality genoma assesment tool) (Gurevich et al., 2013). La calidad del ensamble fue corroborada mediante el uso de minimap2 (Li, 2018) y BUSCO (v4.0.5) (Manni *et al*., 2021) análisis usando el dataset del linaje diptera_odb10 para evaluar la cobertura y calidad del genoma.

### 2.2. Alineamiento de los genomas

El genoma de *D. rucux* fue alineado con *D. mojavensis (a*ccession: GCF_018153725.1) y *D. melanogaster (a*ccession: GCF_000001215.4) mediante el uso de MUMmer3's nucmer (Marçais et al., 2018). El plot de los alineamientos fueron generados mediante Dot tool (Nattestad, 2020).

### 2.3. Anotaciones estructurales y funcionales

Se utilizó la versión 2.0.4 de RepeatModeler para identificar secuencias repetidas de novo en el genoma de *D. rucux* a través de una búsqueda auto-blast". Luego, se utilizó la versión 4.0 de RepeatMasker para buscar secuencias repetidas conocidas utilizando un programa de cruce-match con una biblioteca de secuencias repetidas derivada de Repbase (versión 20140131) y las secuencias repetidas de novo se construyeron utilizando RepeatModeler.

Para la anotación del genoma se utilizaron tres sets de programas para corroborar su eficacia y calidad de la anotación. Partimos con Augustus v 3.5.0 (Stanke et al., 2006), es un programa de predicción de genes que puede ser

utilizado como un programa ab initio, lo que significa que se basa puramente en la secuencia para hacer sus predicciones. De igual forma se utilizó EvidenceModeler (Haas et al., 2008) que es un programa que combina diferentes fuentes de información (evidencia) para generar una anotación completa y precisa de la estructura de los genes en organismos eucariotas. En el caso de Funannotate v 1.8.15 (Palmer et al., 2023) es un software que hace predicción y anotación de genes mediante el uso de herramientas como GlimmerHMM, SNAP y tRNAScan-SE. En el caso de la anotación funcional, se utilizó InterProScan (Jones et al., 2014) para asignar dominios y motivos de proteínas a secuencias mediante comparación contra una variedad de bases de datos (TIGRFam, ProDom, SMART, HAMAP, Prosite Patterns, Superfamily, PRINTS, Panther, Gene3D, PIRSF, PfamA y Prosite Profiles). Las anotaciones se almacenaron como términos de Ontología Genética (GO) para cada secuencia y se utilizó BioMart de Ensembl (Kinsella et al., 2011) para obtener los nombres y los IDs de cada uno de ellos. Posteriormente estos fueron analizados en gProfiler para análizar el enriquecimiento de genes y observar la sobre-representación de información procedente de términos de Ontología genómica, vías biológicas, elementos reguladores del ADN, anotaciones genéticas y redes de interacciones proteína-proteína, esto se lo hizo con los GO predeterminados para *D. mojavensis*.

Finalmente, se emplearon los identificadores de Ensembl junto con el programa ShinyGO v8.0 (Ge et al., 2020) para comparar el conjunto de genes de *D. rucux* con el conjunto de genes de *D. melanogaster* y mapear estas secuencias en diversas posiciones cromosómicas, con un p-value < 0.005. Para las especies con genomas completamente secuenciados, ShinyGO localiza la distribución cromosómica de todos los genes incluidos en la lista del usuario y realiza un análisis estadístico de las características genómicas. El programa determina si los genes están distribuidos aleatoriamente en los cromosomas mediante una prueba de chi-cuadrado, en comparación con el conjunto de genes de fondo del genoma (Ge et al., 2020). Por último, se llevan a cabo pruebas t para evaluar diferencias significativas entre los genes de interés y los demás genes de fondo en el genoma.

## 3. Resultados y Discusión

### 3.1. Análisis de calidad de las secuencias y ensamble de novo

La secuenciación Illumina TrueSeq de *D. rucux* generó 7.5 Gb de datos de secuencia. El filtrado de adaptadores de Illumina y secuencias de baja calidad eliminó el 0,077% de los 138.9 millones de lecturas sin procesar de extremos emparejados. Las estadísticas de la lectura se proporcionan en la Tabla 1. La

evaluación del ensamble con QUAST, reveló un genoma de 182.7 Mb distribuidos en 183 millones de lecturas filtradas de alta calidad en 39.305 contigs. Al realizar los scaffolds y el gap-filling, se produjo un borrador final de 30.439 contigs, con un total de 182.5 millones de lecturas, un valor N50 de 56.237 kb para los scaffolds y un porcentaje de 37.73 % GC. Las estadísticas de ensamblaje se presentan en la Tabla 2, y el porcentaje de GC fue similar a otras especies del género Drosophila (*D. melanogaster* y *D. mojavensis*).

***Tabla 1.- Estadísticas de las lecturas del secuenciamiento de Illumina en D. rucux***

| **Estadísticas de la lectura** | | **Raw reads** | **Filtered reads** |
|---|---|---|---|
| Porcentaje de nucleótido (%) | Recuento | 138.933.106 | 138.825.046 |
| | A | 31.25 | 31.18 |
| | T | 30.79 | 31.09 |
| | G | 18.91 | 18.87 |
| | C | 19.04 | 18.86 |
| Porcentaje de dinucleótidos (%) | AT | 62.04 | 62.27 |
| | GC | 37.95 | 37.73 |

***Tabla 2.- Estadísticas para el ensamblaje de novo de D. rucux (Masurca v4.1.1)***

| | **Initial conti assembly** | **Final draft assembly** |
|---|---|---|
| Assembled bases (pb) | 183.088.892 | 182.582.952 |
| Number of contigs | 33.484 | 30.439 |
| N50 (pb) | 55.054 | 56.237 |
| GC Content (%) | 37.01 | 37.73 |

El ensamblaje de novo, realizado con Masurca, fue evaluado mediante BUSCO utilizando genes ortólogos del linaje Diptera. Se identificó un 98.6% de genes ortólogos completos, de los cuales el 98.2% correspondían a genes completos de una sola copia, mientras que el 0.4% eran genes completos pero duplicados. Además, el 0.6% de los genes ortólogos evaluados se encontraron fragmentados y el 0.8% no se identificaron. Al comparar estos resultados con los genomas de Drosophila previamente reportados, se observó que el genoma de *D. rucux* presenta una similitud notable con *D. mojavensis* (99% de genes completos) y *D. melanogaster* (99.3% de genes completos) según las accesiones en GenBank: GCF_018153725.1 y GCF_000001215.4,

respectivamente. Finalmente, el análisis de evaluación comparativa predijo la presencia de 3225 genes eucarióticos conservados de los 3285 genes conocidos en el conjunto de datos diptera_odb10 (Figura 2).

Las diferencias observadas entre los tamaños de los genomas de *D. melanogsater, D. mojavensis* y *D. rucux* ~ 20 Mb, pueden estar relacionadas a diferentes tasas de acumulación de pequeñas deleciones e inserciones a lo largo del genoma (Moriyama *et al*., 1998), aunque Boulesteix y colaboradores (2005) mencionan que esto se debe principalmente al contenido de secuencias repetitivas, incluidos los elementos móviles (TEs) y que el porcentaje de secuencias RTRS (Reverse transcriptase–related sequence) está correlacionado con el tamaño del genoma para la mayoría de especies, pertenecientes al grupo *Sophophora*, y que se puede asumir modelos similares para otros grupos de *Drosophila.* También se debe considerar que el tipo de tecnología de secuenciación utilizada para obtener la secuencia del genoma, así como el proceso de ensamblado, puede influir en, cierta medida, en los tamaños génicos y procesos de anotación asociados con especies del género Drosophila u otros (Valencia et al., 2020).

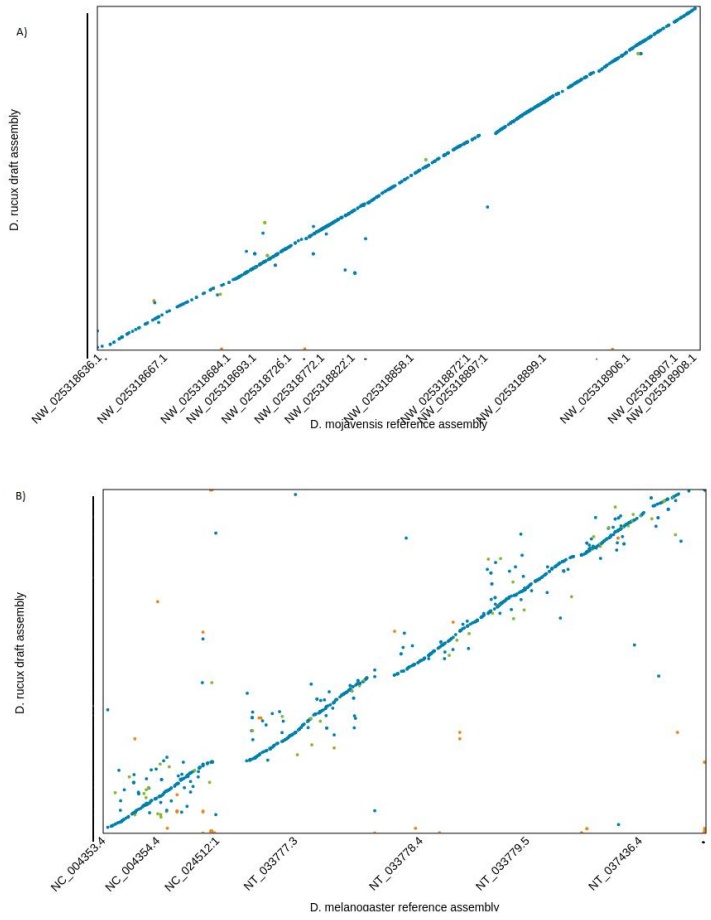


**Figura 1.- Dotplot mostrando el alineamiento del genoma ensamblado de *D. rucux* scaffolds (a lo largo del eje Y) en comparación al genoma de referencia de (A) *D. mojavensis* (B) *D. melanogaster*. Los puntos de colores representan alineaciones directas únicas (azul), alineaciones inversas únicas (verde) y alineaciones repetitivas (naranja).**


**Figura 2.- Representación esquemática de los resultados de Benchmarking Universal Single-Copy Ortholgs (BUSCO) para determinar la integridad del ensamblaje del genoma. Los colores indican grados de integridad de los genes predichos en el ensamblaje. 'n' indica el número total de genes en el conjunto de datos diptera_odb10.**

### 3.2. Anotaciones estructurales y funcionales

El estudio reveló la presencia de elementos móviles pertenecientes a las Clases I y II, así como secuencias de Replicaciones de Secuencias Simples (SSR) y Genes Potenciales (PG). Aproximadamente el 17.5% del genoma se clasificó como secuencias repetitivas mediante RepeatMasker, incluyendo elementos de Clase I, como LINE, SINE, LTR y TRIM, junto con elementos de Clase II, como Helitron, MITE, MAVERICK y TIR. Se identificaron un total de 346 secuencias repetitivas, abarcando tanto elementos móviles como no móviles. Sin embargo, un 45.45% de estas secuencias no pudo ser clasificado. En segundo lugar, los elementos de Clase I representaron el 25.8% de la muestra, siendo los elementos con LTR (Long Terminal Repeat) los más abundantes (Tabla 3).

Barron y colaboradores (2014), reportan un 20% de regiones repetitivas en el genoma de *D. melanogaster*, que es un valor aproximado reportado para el genoma de *D. rucux.* Sin embargo, Barron y colaboradores (2014) reportaron contrastes con la información establecida para *D. melanogaster*, incluyendo la

contenida en Kapitonov y colaboradores (2002) en donde mencionan porcentajes entre el 22% y el 25% de elementos móviles. Para otras especies como *D. mojavensis* y *D. arizonae*, Banho y colaboradores (2021) mencionan porcentajes de 25% y 26% respectivamente. En consecuencia, se puede inferir que los elementos transponibles presentes en *D. rucux* se encuentran en un porcentaje similar a otras especies del género *Drosophila*.

***Tabla 3. Resultados de los ETs para el genoma ensamblado de D. rucux (RepeatModeler)***

| **Class 1** | **Class 2** | **From not TEs:** | **Unclassified** |
|---|---|---|---|
| LINE total (RIX): 27 (7.76%) | Helitron total (DHX): 2 (0.57%) | PHG total: 4 (1.15%) | 158 (45.40%) |
| LTR total (RLX): 43 (12.36%) | MITE total (DXX-MITE): 2 (0.57%) | SSR total: 85 (24.43%) | |
| SINE total (RSX): 8 (2.30%) | Maverick total (DMX): 1 (0.29%) | Not TEs total: 89 (25.57%) | |
| TRIM total (RXX-TRIM): 8 (2.30%) | TIR total (DTX): 4 (1.15%) | | |
| ClassI + unclassified order: 4 (1.15%) | ClassII total: 9 (2.59%) | | |
| Class 1 total: 90 (25.86%) | | | |

***Tabla 4. Resumen de los genes predichos para el genoma ensamblado de D. rucux***

| Herramienta | Tipo de predicción | Total genes |
|---|---|---|
| AUGUSTUS | Predicted genes | 15,920 |
| AUGUSTUS | HiQ Predicted genes | 856 |
| GlimmerHMM | Predicted genes | 27,332 |
| EVM/FAN | Predicted genes | 15,435 |
| SNAP | Predicted genes | 28,865 |
| tRNAScan-SE | Predicted tRNA genes | 222 |
| | Final protein coding genes | 19,805 |

La herramienta tRNAScan-SE identificó 222 genes que codifican para tRNA en el ensamble. Un total de 15,920 modelos génicos fueron predichos mediante AUGUSTUS, de los cuales 856 presentaron un valor >99% de evidencia de exones. Los resultados de anotación de genes analizados en otras herramientas se encuentran resumidos en la tabla 4. Mediante los modelos de genes de EvidenceModeler (EVM) y Funannotate (FAN) se obtuvo un total de 15,435 modelos de genes, otorgando valores similares a los ya descritos por Augustus tanto para *D. melanogaster y D. mojavensis*.

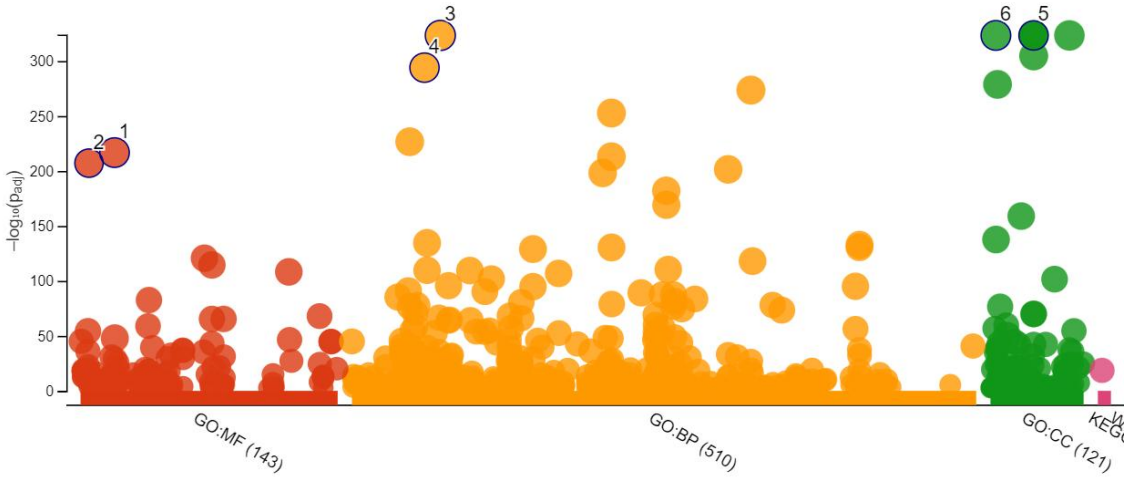
Después de filtrar los modelos de genes que contenían gaps, elementos transponibles y proteínas de longitud inferior a 100 aminoácidos, se retuvieron finalmente 19,805 genes que codifican proteínas. Esta anotación genómica exhibe un número más alto de secuencias codificadas de proteínas que aquellos obtenidos para *D. mojavensis* y *D. melanogaster* (que tienen entre 13.270 y 13.962 secuencias codificadas de proteínas respectivamente). Es importante destacar que este resultado podría deberse a dos factores posibles. En primer lugar, se utilizaron tres bases de datos -SwissProt, UniProt y FlyBase- para maximizar el número de coincidencias, lo que incrementó la probabilidad de identificar un mayor número de proteínas. En segundo lugar, se utilizó el conjunto predeterminado "fly" de Augustus para realizar las comparaciones, lo que se basa en un conjunto de entrenamiento que se aplica a dípteros en general. Como este conjunto de entrenamiento está diseñado para reconocer regiones características de dípteros, incluyendo promotores y CDS, es probable que incluya patrones comunes en estos insectos, lo que podría explicar el aumento en la cantidad de proteínas identificadas. Casi todos los genes (98,7%) devolvieron un match significativo (evalue=1e-5) con secuencias de proteínas conocidas, de alta similitud. La mayoría de los top hits para cada gen coincidieron con secuencias de Dipteros como *Drosophila melanogaster* (79,9%) y *Anopheles albimanus* (14,1%).

La tabla 5 resume las comparativas en las anotaciones de los genomas entre *D. mojavensis* y *D. melanogaster* obtenidas en el NCBI bajo los códigos de acceso descritos previamente.

***Tabla 5. Tabla comparativa para el genoma ensamblado de D. rucux frente a los genomas de D. mojavensis y D. melanogaster.***

| **Categorías** | ***D. rucux*** | ***D. mojavensis*** | ***D. melanogaster*** |
|---|---|---|---|
| Tecnología de secuenciación | Illumina TrueSeq | Oxford Nanopore MinION; Illumina HiSeq | PacBio; Illumina HiSeq |
| Tamaño del genoma | 182.7 Mb | 163,2 Mb | 143,7 Mb |
| Porcentaje GC | 37.73% | 39.5 % | 41.9 % |
| Augustus genes | 15,920 | 15,130 | 17,872 |
| Codifican proteinas | 19,805 | 13,270 | 13,962 |
| Busco análisis | 98,2% | 99.3% | 99.4% |

***Los datos de D. mojavensis y D. melanogaster se los obtuvo a partir de los ensambles del NCBI, códigos de acceso (GCF_018153725.1; GCF_000001215.4 respectivamente)***




| ID | Source | Term ID | Term Name | $p_{adj}$ (query_1) |
|---|---|---|---|---|
| 1 | GO:MF | GO:0005488 | binding | $1.979\times10^{-217}$ |
| 2 | GO:MF | GO:0003824 | catalytic activity | $7.383\times10^{-208}$ |
| 3 | GO:BP | GO:0009987 | cellular process | $4.941\times10^{-324}$ |
| 4 | GO:BP | GO:0008152 | metabolic process | $1.009\times10^{-294}$ |
| 5 | GO:CC | GO:0043226 | organelle | $4.941\times10^{-324}$ |
| 6 | GO:CC | GO:0005622 | intracellular anatomical structure | $4.941\times10^{-324}$ |

**Figura 3. Análisis GO (Gene Ontology) para determinar componentes biológicos y rutas metabólicas mediante gProfiler, *D. rucux*.**

### 3.3. GO analysis

A través de InterProScan se identificaron los dominios conservados de proteínas y se clasificaron los genes de *D. rucux* en categorías funcionales definidas por Gene Ontology. Se agregaron anotaciones funcionales a 13,372 secuencias, representando el 84.0% del total de genes predichos para Augustus. Posteriormente, se obtuvieron los nombres de los genes asociados a estos IDs y se utilizaron para realizar un análisis de enriquecimiento funcional, conocido como análisis de conjuntos de genes, mediante la plataforma gProfiler (Reimand et al., 2016). Los términos de la Ontología Génica (GO) describen los atributos de los productos génicos, incluyendo su función molecular, los procesos biológicos en los que intervienen y su ubicación subcelular. El análisis de enriquecimiento de términos de GO es crucial para resaltar la relevancia biológica de los datos de secuenciación, por lo que se utilizó gProfiler para buscar los términos de GO enriquecidos en nuestro conjunto de datos. Además, gProfiler ofrece la opción de analizar un conjunto de datos, permitiendo el enriquecimiento de GO en listas clasificadas; así, se realizó el análisis en todas las categorías de GO (proceso biológico, función molecular, componente celular) y comparamos los conjuntos de datos correspondientes.

Como era de esperar, hay una sobrerrepresentación de proteínas relacionadas a "procesos biológicos fundamentales" comunes a todas las células eucariotas.

(figura 3). Por ejemplo, en el ámbito de Funciones Molecures (MF), los valores más destacados correspondieron a funciones relacionadas con proteínas de unión, es decir, cómo estas proteínas se unen a otras moléculas seleccionadas y cómo su actividad depende de dicha unión. También destacaron aquellas proteínas con actividad catalítica, es decir, que aceleran la tasa de una reacción química sin ser consumidas o alteradas en el proceso. En el ámbito de la Procesos Biológicos (BP), los valores más representativos se centraron en procesos celulares, procesos metabólicos primarios y procesos metabólicos de sustancias orgánicas. En cuanto al ámbito de Componentes Celulares (CC), los valores más significativos estuvieron relacionados con organelas y proteínas que participan en la estructura celular. Al comparar estos datos con los obtenidos por Vedelek y colaboradores (2018) para *D. melanogaster*, podemos determinar que, en general, los valores más representativos van a estar relacionados a procesos relacionados a la replicación, transcripción y el metabolismo de los organismos.

***Tabla 6. Análisis de sobrerrepresentación de genes para cada una de las categorías GO***

| GO Terms | ID | Nombre | Descripción | Función |
|---|---|---|---|---|
| Funciones Moleculares | GO:0005488 | Act5C | Actin 5C | Unión ATP |
| | | Act42A | Actin 42A | Unión ATP |
| | GO:0003824 | Act57B | Actin 57B | Unión ATP, actividad hidrolasa |
| | | Act79B | Actin 79B | Unión ATP, actividad hidrolasa |
| Procesos Biológicos | GO:0009987 | abd-A | abdominal A | Apoptosis, diferenciación antero-posterior |
| | | Abd-B | Abdominal B | Apoptosis, diferenciación antero-posterior |
| | GO:0008152 | Abl | Abl tyrosine kinase | Desarrollo del sistema nervioso central |
| | | Ace | Acetylcholine esterase | Proceso catabólico de acetilcolina (ACh) |
| Componentes Celulares | GO:0005622 | bru1 | bruno 1 | Componente celular en el citoplasma |
| | | aurA | aurora A | Componente celular en el núcleo |
| | GO:00043226 | RpLP0 | Ribosomal protein P0 | Parte de la subunidad larga del ribosoma |
| | | SmF | Small ribonucleoprotein particle protein SmF | Parte de la subunidad pequeña del ribosoma |

En el análisis GO (Figura 3), se identificaron numerosas proteínas asociadas a diversas categorías funcionales, procesos biológicos y componentes celulares. Para la presentación de los resultados, se seleccionaron únicamente aquellas proteínas que mostraron una sobrerrepresentación significativa en sus respectivas categorías GO, basándose en criterios de significancia estadística y relevancia biológica. La Tabla 6 muestra que los conjuntos de genes con mayor número de repeticiones incluyen una alta proporción de proteínas actinas, destacando que estas proteínas presentan los porcentajes de expresión más elevados. Estas proteínas son algunas de las más abundantes y conservadas en eucariotas, y desempeñan funciones moleculares cruciales en el organismo, como su papel en la unión al ATP (Dominguez et al., 2011). En la categoría de procesos biológicos, las proteínas abdominal A y abdominal B actúan como reguladoras del complejo bithorax, que es esencial para la segmentación del abdomen en Drosophila, así como para el desarrollo de las genitalias (Foronda et al., 2006). Por otro lado, la tirosina quinasa y la acetilcolinesterasa desempeñan roles importantes en el desarrollo del sistema nervioso central y en la hidrólisis de la colina liberada en la sinapsis (Vaikkakara et al., 2023; Mutero et al., 1992). En cuanto a los componentes celulares, bruno1 y aurora A son fundamentales en el citoplasma y son esenciales para el desarrollo y mantenimiento de los músculos de vuelo en *Drosophila* (Hutter et al., 2006). Finalmente, los componentes RpLP0 y SmF son cruciales para el ribosoma, ya que facilitan la comunicación entre este y los factores de traducción unidos a GTP, como el Factor de Elongación-G (EF-G) y el Factor de Elongación Tu (EF-Tu) (Illag et al., 2005).

### 3.4. ShinyGO v 0.80

ShinyGo es una herramienta que facilita el mapeo genético de una lista de genes en relación con su distribución cromosómica en una especie específica. Esta plataforma permite analizar la ubicación de genes de interés en el contexto de un genoma, proporcionando información valiosa sobre la organización del material genético y su relación con funciones biológicas, variantes fenotípicas y posibles implicaciones evolutivas. Para el análisis, se empleó el genoma ensamblado de *D. melanogaster* (Assembly BDGP6.32; Tax ID Ensembl 7227). Esta elección se justifica porque esta especie es una de las pocas cuyo genoma se encuentra ensamblado a nivel de cromosomas. Al querer utilizar a *D. mojavensis* como especie para el análisis el programa no lo permite debido a que esta especie tiene su genoma ensamblado a nivel de scaffolds.

Al interpretar los datos de la distribución de genes en los cromosomas de *Drosophila rucux*, se deben tomar en cuenta ciertas consideraciones. En primer

lugar, el número de cromosomas presentes en ambas especies es diferente, *D. melanogaster* presenta un 2n= 8, mientras que *D. rucux* presenta un 2n= 10. De igual forma, hay que tomar en cuenta los rearreglos y la configuración cromosómica para cada una de las especies. Los puntos rojos en la imagen indican la localización de los genes en los cromosomas. La densidad y distribución de estos puntos pueden proporcionar información sobre la organización genómica y la posible relación entre los genes.

En la figura 4, de arriba hacia abajo, se observan los siguientes elementos: el par sexual (cromosomas X y Y), el brazo izquierdo del cromosoma 2 (chr 2L), el brazo izquierdo del cromosoma 3 (chr 3L), el brazo derecho del cromosoma 2 (chr 2R), el brazo derecho del cromosoma 3 (chr 3R) y el cromosoma 4 (chr 4). Al analizar el par sexual, se nota que hay pocos genes mapeados en el cromosoma Y. Esto se debe a que la extracción de ADN se realizó a partir de células somáticas en lugar de los testis de un macho. Además, es importante destacar que en *D. rucux*, el cromosoma Y es submetacéntrico y, en orden descendente, ocupa el segundo lugar en longitud (Mafla, 2012). Se observa una alta cobertura de genes mapeados en el cromosoma X y en el brazo izquierdo del cromosoma 2 (chr 2L), mientras que el brazo derecho del cromosoma 2 (chr 2R) presenta un bajo porcentaje de genes mapeados. En comparación, en D. melanogaster, los cromosomas 2 y 3 son metacéntricos (Kaufman, 2017), mientras que en *D. rucux* son submetacéntricos (Mafla, 2012), lo que podría explicar la falta de información mapeada en estas regiones.Al utilizar la prueba de chi-cuadrado, el programa ShinyGo evalúa si la distribución de estos genes es aleatoria o si hay un patrón significativo. Al observar una distribución no aleatoria se puede determinar que ciertos genes están ligados entre sí, lo cual podría ser crucial para entender la genética de la especie (Ge et al., 2020). De igual forma, sería importante observar si hay una concentración de genes en un cromosoma específico o en ciertos brazos de los cromosomas. Lo que podría reflejar características evolutivas o adaptativas de *Drosophila rucux*.

Finalmente, habría que tomar en cuenta el aumento en el número de cromosomas en *D. rucux*, esto sugeriría que ha habido eventos de reestructuración cromosómica, como fusiones o fisiones, que han ocurrido durante la evolución de estas especies (Jabbari et al., 2000). Estos cambios en la estructura cromosómica pueden alterar la distribución y organización de los genes, lo que podría tener un impacto en la expresión génica y la adaptación de cada especie a su entorno, lo que podría haber contribuido a la radiación evolutiva de las especies del grupo *mesophragmatica* en los Andes, permitiéndoles explotar una amplia gama de hábitats y recursos

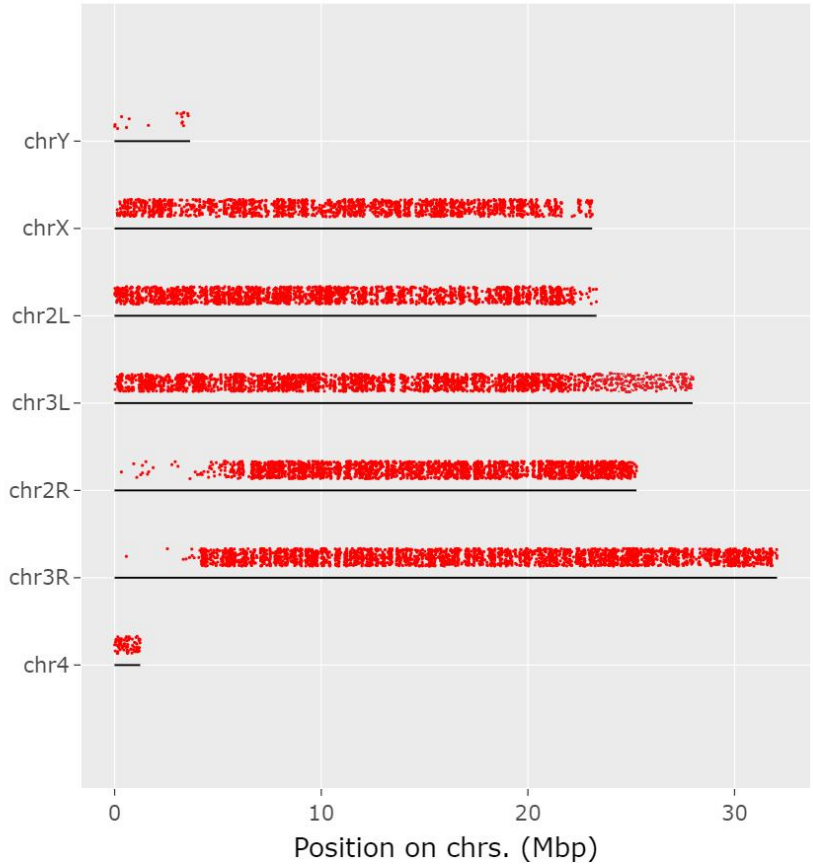



**Figura 4. Mapeo de la lista de genes obtenidos mediante BioMart de Ensembl para *D. rucux*, mediante el programa ShinyGo v8.0**

## 4. Referencias:


Agudelo-Valencia, D., Uribe-Echeverry, P. T., & Betancur-Pérez, J. F. (2020). De novo assembly and annotation of the Ganoderma australe genome. *Genomics*, *112*(1), 930-933.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., … Sherlock, G. (2000). *Gene Ontology: tool for the unification of biology. Nature Genetics, 25(1), 25–29.* doi:10.1038/75556

Banho, C. A., Oliveira, D. S., Haudry, A., Fablet, M., Vieira, C., & Carareto, C. M. A. (2021). Transposable element expression and regulation profile in gonads of interspecific hybrids of Drosophila arizonae and Drosophila mojavensis wrigleyi. *Cells*, *10*(12), 3574.

Barrón, M. G., Fiston-Lavier, A. S., Petrov, D. A., & González, J. (2014). Population genomics of transposable elements in Drosophila. *Annual review of genetics*, *48*(1), 561-581.

Bergman CM, Pfeiffer BD, Rincón-Limas DE, Hoskins RA, Gnirke A, Mungall CJ, Wang AM, Kronmiller B, Pacleb J, Park S, Stapleton M, Wan K, George RA, de Jong PJ, Botas J, Rubin GM, Celniker SE. Assessing the impact of comparative genomic sequence data on the functional annotation of the Drosophila genome. Genome Biol. 2002;3(12):RESEARCH0086. doi: 10.1186/gb-2002-3-12-research0086. Epub 2002 Dec 30. PMID: 12537575; PMCID: PMC151188.

Brake, I., & Bächli, G. (2008). *World catalogue of Insects*.

Boetzer, M., & Pirovano, W. (2012). Toward almost closed genomes with GapFiller. *Genome biology*, *13*, 1-9.

Boulesteix, M., Weiss, M., & Biémont, C. (2006). Differences in genome size between closely related species: the Drosophila melanogaster species subgroup. *Molecular biology and evolution*, *23*(1), 162-167.

Céspedes, D., & Rafael, V. (2012). Cuatro especies nuevas del grupo de especies Drosophila mesophragmatica (Diptera, Drosophilidae) de los Andes ecuatorianos. *Iheringia. Série Zoologia*, *102*, 71-79.

Dominguez, R., & Holmes, K. C. (2011). Actin structure and function. *Annual review of biophysics*, *40*(1), 169-186.

Figuero, M. L., & Rafael, V. (2013). Descripción de tres especies nuevas del género Drosophila (Diptera, Drosophilidae) en el Ecuador. *Iheringia. Série Zoologia*, *103*, 246-254.

Flynn, J. M., Hubley, R., Goubert, C., Rosen, J., Clark, A. G., Feschotte, C., & Smit, A. F. (2020). RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences*, *117*(17), 9451–9457. https://doi.org/10.1073/pnas.1921046117

Foronda, D., Estrada, B., de Navas, L., & Sánchez-Herrero, E. (2006). Requirement of Abdominal-A and Abdominal-B in the developing genitalia of Drosophila breaks the posterior downregulation rule.

Ge SX, Jung D, Yao R. ShinyGO: a graphical gene-set enrichment tool for animals and plants. Bioinformatics. 2020 Apr 15;36(8):2628-2629. doi: 10.1093/bioinformatics/btz931. PMID: 31882993; PMCID: PMC7178415.

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, *29*(8), 1072-1075.

Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, White O, Buell CR, Wortman JR. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biol. 2008 Jan 11;9(1):R7. doi: 10.1186/gb-2008-9-1-r7. PMID: 18190707; PMCID: PMC2395244.

Harrison, P. W., Amode, M. R., Austine-Orimoloye, O., Azov, A. G., Barba, M., Barnes, I., ... & Yates, A. D. (2024). Ensembl 2024. *Nucleic acids research*, *52*(D1), D891-D899.

Hutterer A, Berdnik D, Wirtz-Peitz F, Zigman M, Schleiffer A, Knoblich JA. Mitotic activation of the kinase Aurora-A requires its binding partner Bora. Dev Cell. 2006 Aug;11(2):147-57. doi: 10.1016/j.devcel.2006.06.002. PMID: 16890155.

Ilag LL, Videler H, McKay AR, Sobott F, Fucini P, Nierhaus KH, Robinson CV. Heptameric (L12)6/L10 rather than canonical pentameric complexes are found by tandem MS of intact ribosomes from thermophilic bacteria. Proc Natl Acad Sci U S A. 2005 Jun 7;102(23):8192-7. doi: 10.1073/pnas.0502193102. Epub 2005 May 27. PMID: 15923259; PMCID: PMC1149426.

Jabbari, K., & Bernardi, G. (2000). The distribution of genes in the Drosophila genome. *Gene*, *247*(1-2), 287-292.

Jones P, Binns D, Chang HY, et al. (2014) InterProScan 5: genome-scale protein function classification. Bioinformatics, 30, 1236–1240.

Kapitonov, V. V., & Jurka, J. (2003). Molecular paleontology of transposable elements in the Drosophila melanogaster genome. *Proceedings of the National Academy of Sciences*, *100*(11), 6569-6574.

Kajitani, R., Toshimoto, K., Noguchi, H., Toyoda, A., Ogura, Y., Okuno, M., ... & Itoh, T. (2014). Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome research*, *24*(8), 1384-1395.

Kaufman TC. A Short History and Description of *Drosophila melanogaster* Classical Genetics: Chromosome Aberrations, Forward Genetic Screens, and the Nature of Mutations. Genetics. 2017 Jun;206(2):665-689. doi: 10.1534/genetics.117.199950. PMID: 28592503; PMCID: PMC5499179.

Kinsella, R. J., Kähäri, A., Haider, S., Zamora, J., Proctor, G., Spudich, G., ... & Flicek, P. (2011). Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database*, *2011*, bar030.

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, *34*(18), 3094-3100.

Mafla, A. B. (2012). Cariología beta de tres especies pertenecientes al grupo de especies Drosophila mesophragmatica. *Revista Ecuatoriana de Medicina y Ciencias Biológicas*, *33*(1-2), 38-45.

Manni, M., Berkeley, M. R., Seppey, M., & Zdobnov, E. M. (2021). BUSCO: assessing genomic data quality and beyond. *Current Protocols*, *1*(12), e323.

Marçais, G., Delcher, A. L., Phillippy, A. M., Coston, R., Salzberg, S. L., & Zimin, A. (2018). MUMmer4: A fast and versatile genome alignment system. *PLoS computational biology*, *14*(1), e1005944.

Mirol, P. M., Routtu, J., Hoikkala, A., & Butlin, R. K. (2008). Signals of demographic expansion in Drosophila virilis. BMC Evolutionary Biology, 8, 1-8.

Moriyama, E. N., Petrov, D. A., & Hartl, D. L. (1998). Genome size and intron size in Drosophila. *Molecular biology and evolution*, *15*(6), 770-773.

Mutero A, Pralavorio M, Simeon V, Fournier D. Catalytic properties of cholinesterases: importance of tyrosine 109 in Drosophila protein. Neuroreport. 1992 Jan;3(1):39-42. doi: 10.1097/00001756-199201000-00010. PMID: 1611033.

Nattestad, M (2020). Dot, An interactive Dot Plot Viewer For Genome-Genome Alignments. https://github.com/marianattestad/dot (accessed 05 may 2024)

O'Grady, P. M., & DeSalle, R. (2018). Phylogeny of the genus Drosophila. *Genetics*, *209*(1), 1-25.

Palmer, J. M., & Stajich, J. E. (2023). Funannotate v1.8.9 Disponible en: https://funannotate.readthedocs.io/

Pfeiffer, B.D. (2002.12.2). Drosophila willistoni clone DWIF01_10_L08 (D1413) genomic sequence. *GenBank/EMBL/DDBJ*

Reimand, J., Arak, T., Adler, P., Kolberg, L., Reisberg, S., Peterson, H., & Vilo, J. (2016). g: Profiler—a web server for functional interpretation of gene lists (2016 update). *Nucleic acids research*, *44*(W1), W83-W89.

Smit, A. F. (2004). Repeat-masker open-3.0. *http://www. repeatmasker. org*.

Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., & Morgenstern, B. (2006). AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic acids research*, *34*(suppl_2), W435-W439.

Throckmorton, L. H. (1975). The phylogeny, ecology and geography of Drosophila. Handbook of genetics, 3(17), 422-469.

Thurmond, J., Goodman, J. L., Strelets, V. B., Attrill, H., Gramates, L. S., Marygold, S. J., ... & Calvi, B. R. (2019). FlyBase 2.0: the next generation. *Nucleic acids research*, *47*(D1), D759-D765.

Vaikakkara Chithran A, Allan DW, O'Connor TP. Adult expression of Semaphorins and Plexins is essential for motor neuron survival. Sci Rep. 2023 Apr 11;13(1):5894. doi: 10.1038/s41598-023-32943-4. PMID: 37041188; PMCID: PMC10090137.

Vedelek, V., Bodai, L., Grézal, G., Kovács, B., Boros, I. M., Laurinyecz, B., & Sinka, R. (2018). *Analysis of Drosophila melanogaster testis transcriptome. BMC Genomics, 19(1).* doi:10.1186/s12864-018-5085-z

Xiao, J., Liu, B., Yao, Y., Guo, Z., Jia, H., Kong & Chong, K. (2022). Wheat genomic study for genetic improvement of traits in China. *Science China Life Sciences*, *65*(9), 1718-1775.

Zimin, A. V., Marçais, G., Puiu, D., Roberts, M., Salzberg, S. L., & Yorke, J. A. (2013). The MaSuRCA genome assembler. *Bioinformatics*, *29*(21), 2669-2677.

Zhou, T., Yao, J., & Liu, Z. (2017). Gene ontology, enrichment analysis, and pathway analysis. *Bioinformatics in aquaculture: Principles and methods*, 150-168.