Centro de Investigación en Alimentación y Desarrollo, A.C.

CARACTERIZACIÓN GENÓMICA Y TRANSCRIPTÓMICA DE *Skiffia lermae* Y ESPECIES FILOGENÉTICAMENTE CERCANAS (Goodeinae)

Por:

Ana Graciela Carvajal Peraza

TESIS APROBADA POR LA

COORDINACIÓN DE MAZATLÁN

Como requisito parcial para obtener el grado de

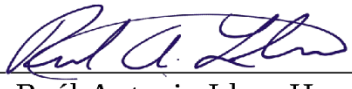
MAESTRA EN CIENCIAS

Mazatlán, Sinaloa Enero 2022

APROBACIÓN

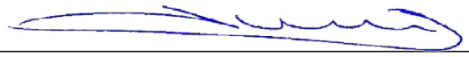
Los miembros del comité designado para la revisión de la tesis de Ana Graciela Carvajal Peraza, la han encontrado satisfactoria y recomiendan que sea aceptada como requisito parcial para obtener el grado de Maestra en Ciencias

Dra. Beatriz Yáñez Rivera
Directora de tesis



Dr. Raúl Antonio Llera Herrera
Co-director de tesis

Dr. Albert Maurits van der Heiden
Integrante del comité de tesis



Dr. Juan Manuel Martínez Brown
Integrante del comité de tesis

DECLARACIÓN INSTITUCIONAL

La información generada en la tesis "Caracterización Genómica y Transcriptómica de *Skiffia lermae* y Especies Filogenéticamente Cercanas (Goodeinae)" es propiedad intelectual del Centro de Investigación en Alimentación y Desarrollo, A.C. (CIAD). Se permiten y agradecen las citas breves del material contenido en esta tesis sin permiso especial de la autora Ana Graciela Carvajal Peraza, siempre y cuando se dé crédito correspondiente. Para la reproducción parcial o total de la tesis con fines académicos, se deberá contar con la autorización escrita de quien ocupe la titularidad de la Dirección General del CIAD.

La publicación en comunicaciones científicas o de divulgación popular de los datos contenidos en esta tesis, deberá dar los créditos al CIAD, previa autorización escrita del manuscrito en cuestión del director(a) de tesis.

CIAD

CENTRO DE INVESTIGACIÓN EN ALIMENTACIÓN Y DESARROLLO, A.C.
Coordinación de Programas Académicos

Dr. Pablo Wong González
Director General

AGRADECIMIENTOS

Le agradezco en primer lugar al Consejo Nacional de Ciencia y Tecnología (CONACYT) por el apoyo económico recibido durante el posgrado que me permitió dedicarme de lleno a las clases y ocupaciones de la maestría.

Agradezco al Centro de Investigación en Alimentación y Desarrollo (CIAD) por permitirme usar sus instalaciones para desarrollar mi proyecto, acudir a clases y estudiar. Especialmente al laboratorio de Ecotoxicología y a su responsable el Dr. Miguel Betancourt Lozano por abrir sus espacios para realizar mis estudios.

Agradezco al Dr. Omar Domínguez y a Ivette Villa de la Universidad Michoacana de San Nicolás de Hidalgo, quienes amablemente me proporcionaron los organismos necesarios para llevar a cabo mi tesis. Igualmente agradezco al laboratorio de ciencias genómicas de la UNAM por las facilidades otorgadas en su instituto y por permitirme hacer uso del servidor Chihuil-ICMYL (http://www.icmyl.unam.mx/mazatlan/).

Le doy gracias a mi comité por trabajar conmigo durante todo el proceso de elaboración de la tesis, por resolver mis dudas y guiarme en los momentos que era necesario, no habría logrado culminar la maestría sin su ayuda.

Gracias a mi familia por apoyarme, orientarme y motivarme cada vez que lo necesité. Gracias a mi mamá por mantenerme enfocada y encontrar formas de volverme fácil el trabajo de tesis. Le agradezco a mi papá por ser un gran ejemplo a seguir y una figura que demuestra que esforzándote consigues lo que sea.

CONTENIDO

CONTENIDO (continuación

LISTA DE FIGURAS

LISTA DE CUADROS

RESUMEN

Una de las familias de peces con mayor endemismo en México es Goodeidae, cuya subfamilia Goodeinae es endémica en su totalidad. Las especies que la conforman presentan viviparismo matrotrófico, condición en la que los embriones presentan un análogo placentario llamado trofotenia que funciona como placenta. Esta subfamilia se considera vulnerable debido a que en 20 años desapareció al menos una especie en la mayoría de sus hábitats (68%) y la distribución de otras cinco se redujo a la mitad. Las causas se relacionan con la contaminación y reducción de los cuerpos de agua por vertimiento de desperdicios domésticos e industriales, deforestación de la cuenca, modificación del hábitat, introducción de especies exóticas y la sobre pesca. En la subfamilia Goodeinae, la mayor parte de la información genética se ha dirigido a entender las relaciones filogenéticas, los patrones biogeográficos y de distribución, así como los aspectos reproductivos. Sin embargo, no se cuenta con datos genómicos de ninguna especie de goodeido, información clave para las estrategias de conservación; ya que permite la evaluación de posibles ventajas adaptativas lo que facilita el éxito de los programas de reintroducción y preservación. Se seleccionaron las cuatro especies del género *Skiffia* (donde *S. lermae* destaca por ser una especie amenazada con algunas poblaciones estables) y a *Girardinichthys viviparus*, ya que son representativas de las reducciones poblacionales de la subfamilia. El objetivo del trabajo es la caracterización genómica de *S. lermae* mediante una aproximación híbrida y el análisis de la presión de selección en genes particulares de cinco especies cercanas con modelos de sustitución de codones. Para *S. lermae* se construyó el mitogenoma completo con una longitud de 16,551 bases y un genoma contiguo de 400 Mb en 8,150 contigs y un contig N50 de 54.7 kb; sin embargo, está incompleto (<40%), porque representa solo un tercio de su magnitud estimada. Se ensambló el transcriptoma de siete tejidos de un ejemplar adulto y una larva completa (>90% de completitud). Se ensamblaron transcriptomas de hígado de *S. bilineata*, *S. francesae*, *S. multipunctata* y *G. viviparus* con calidad de borrador (<60%), lo que permitió la identificación de patrones de selección divergente en los siguientes ortólogos: CYP450, CYP51, HSP90A, HSPB8, NOX1, PIK3R1, PXR, PPAR, SERT, ATP7B y ERB1. Este patrón indicaría que estos genes presentan cambios en los codones que de alguna manera generan una respuesta fenotípica diferencial que permite la adaptación de los organismos a su entorno.

Palabras claves: Recursos genómicos, secuenciación, conservación, transcriptómica.

ABSTRACT

One of the fish families with the most endemism in Mexico is Goodeidae, which contains the Goodeinae subfamily and this one is entirely endemic. Th especies that comprise it present matrotrophic viviparism, a condition in which the embryos have a placental analog called trophoteniae that functions as a placenta. This subfamily is considered vulnerable because, in the past 20 years at least one species has disappeared in most of its habitats (68%) and the distribution of five others has been reduced by half. The causes are related to contamination and reduction of water bodies due to domestic and industrial waste dumping, watershed deforestation, habitat modification, exotic species introduction and overfishing. In the Goodeinae subfamily, most of genetic research has been directed towards understanding phylogenetic relationships, biogeographic and distribution patterns, as well as reproductive aspects. However, no genomic data is available for any goodeid species, which is key information for conservation strategies, as it allows the evaluation of posible adaptive advantages that facilitate the success of reintroduction and preservation programs. The four species of the genus *Skiffia* (where *S. lermae* stands out as a threatened species with some stable populations) and *Girardinichthys viviparus* were selected because they are representative of the population reductions of the subfamily. The aim of the work is the genomic characterization of *S. lermae* using a hybrid approach and the analysis of selection pressure on particular genes of five closely related species with codon substitution models. For *S. lermae*, the complete mitogenome was constructed with a length of 16,551 bases and a 400 Mb contiguos genome with 8,150 contigs and an N50 contig of 54.7 kb; however, it is incomplete (>40%), because it represents only one third of its estimated magnitude. The transcriptome of seven tissues from one adult specimen and a complete larva was assembled (>90% completeness). Liver transcriptomes were assembled from *S. bilineata*, *S. francesae*, *S. multipunctata* and *G. viviparus*, allowing the identification of divergent selection patterns in the following orthologs: CYP450, CYP51, HSP90A, HSPB8, NOX1, PIK3R1, PXR, PPAR, SERT, ATP7B and ERB1. This pattern would indicate that these genes present codon mutations that somehow generate a differential phenotypic response that allows organisms to adapt to their environment.

Key words: Genomic resources, secuencing, conservation, transcriptomics

1. INTRODUCCIÓN

El concepto de biodiversidad integra la variabilidad biológica a través de todos los niveles de organización biológica, desde genes, a especies y desde ecosistemas, hasta paisajes (Walker, 1992). México es un país megadiverso, alberga entre el 10 y 12% de la diversidad de especies mundial (más de 108,000) (Llorente y Ocegueda, 2008). Particularmente los sistemas acuáticos del centro del país son reservorio de diversos grupos de peces con altos niveles de endemismo (Domínguez *et al*., 2010).

Una de las familias con mayor endemismo en el centro del país es Goodeidae (Cyprinodontiformes), cuya subfamilia Goodeinae es endémica en su totalidad con 41 especies clasificadas en 19 géneros las cuales se conocen localmente como mexclapiques (Domínguez y Pérez, 2007; Page *et al*., 2013). Esta subfamilia se considera vulnerable debido a que en 20 años desapareció al menos una especie en la mayoría de sus hábitats (68%) y la distribución de otras cinco se redujo a la mitad (Lyons *et al*., 2019). Las causas se relacionan con la contaminación y reducción de los cuerpos de agua por vertimiento de desperdicios domésticos e industriales, deforestación de la cuenca, modificación del hábitat, introducción de especies exóticas y sobre pesca (Domínguez *et al*., 2005). Actualmente, una especie está extinta (*Characodon garmani*), dos especies sólo se conservan en acuarios (*Skiffia francesae* y *Zoogoneticus tequila*), y 24 especies se encuentran en alguna categoría de protección bajo la NOM-059-SEMARNAT-2010 (SEMARNAT, 2019).

Entre las especies protegidas por el grado de afectación, destaca *Skiffia lermae*. Su distribución original incluía 18 localidades, de las cuáles sólo se mantienen tres con poblaciones estables. Se estima una reducción del 65% desde el año 1999 (Domínguez *et al*., 2008; Lyons *et al*., 2019). Las localidades en las que se distribuye actualmente son: manantial La Mintzita, Lago de Zacapu y Lago de Pátzcuaro, todas en el estado de Michoacán. La cuenca de Pátzcuaro y La Mintzita están salvaguardadas por decretos de Área Natural Protegida (ANP); Pátzcuaro tiene un decreto federal (Cuenca Hidrográfica del Lago de Pátzcuaro) y un estatal de ANP (Cerro del Estribo Grande), mientras que La Mintzita se encuentra en la categoría de Zona Sujeta a Preservación Ecológica de jurisdicción estatal y es sitio Ramsar (SEMARNAP, 2000; RAMSAR, 2009). Por otro lado, el Lago de Zacapu está en condiciones delicadas ya que se encuentra adyacente a asentamientos

urbanos y está en proceso de desecación; la cuenca alta y media presentan deforestación, además del azolvamiento del lago (CONACYT, 2021; Ayuntamiento de Morelia, 2009; Hernández *et al.*, 2013).

Las estrategias de conservación requieren información genómica para implementar programas efectivos de protección de especies. Se ha demostrado que permite la identificación de alelos adaptativos para el rescate evolutivo basado en patrones genómicos de endogamia (Supple y Shapiro, 2018). Particularmente, se ha utilizado el análisis del genoma mitocondrial para la identificación y el manejo de la diversidad genética, así como para dirigir estudios demográficos de poblaciones (Moritz, 1994).

Sin embargo, la única especie de la subfamilia Goodeinae con su genoma mitocondrial secuenciado es *Xenotoca eiseni* y dado el estatus de protección, la generación de datos genómicos tanto mitocondriales como nucleares resulta relevante. Lo cual es frecuente a pesar de la categoría de protección, ya que menos del 1% de todas las especies de flora y fauna enlistadas como amenazadas de acuerdo a la Unión Internacional por la Conservación de la Naturaleza (IUCN) cuentan con la caracterización del genoma (Brandies *et al*., 2019).

Además, al estudiar las variaciones en el genoma es posible inferir los procesos de adaptación que han ocurrido en este grupo con características biológicas particulares como la presencia de una trofotenia. La evaluación de las posibles ventajas adaptativas, que permiten identificar patrones de selección natural, se realiza mediante la tasa de sustituciones no sinónimas sobre las sinónimas, conocida como la proporción d_N/d_S (Jeffares *et al*., 2015). Esta prueba es capaz de describir la dirección y la fuerza de la presión selectiva entre poblaciones (Rocha *et al*., 2006); además ha sido utilizada ampliamente en peces. Por ejemplo, en salmónidos se ha identificado que tienen procesos modificados relacionados con el transporte de hierro en la sangre (Ford, 2001), entre el bacalao del Atlántico y sus especies hermanas se encontró evidencia de selección purificante (Marshall *et al*., 2009) y en cíclidos se identificó la presión selectiva sobre el gen de rodopsina (Sugawara *et al*., 2002).

La producción del genoma completo requiere una aproximación híbrida para optimizar la calidad del resultado, lo cual implica la combinación de lecturas cortas (Illumina) con otra plataforma que proporcione secuencias largas, como las que brinda la tecnología Oxford Nanopore. Las secuencias Illumina ofrecen buen precio y buena calidad, pero su corta longitud (75 a 250 bases) no garantiza la continuidad de un ensamble genómico. Las lecturas largas de Nanopore alcanzan longitudes de

más de 100 kb; sin embargo, no tienen una buena calidad asociada (Magi *et al*., 2018).

El objetivo del presente trabajo es la caracterización del genoma de *Skiffia lermae* mediante una aproximación híbrida y el análisis de la presión de selección en genes particulares con información genómica de especies cercanas.

2. ANTECEDENTES

2.1 Tecnologías de Secuenciación

En la actualidad, la búsqueda por información genómica exige la rápida generación de datos genómicos, con lecturas económicas y precisas. Este reto ha propiciado el desarrollo de tecnologías de secuenciación de la siguiente generación (conocidas como NGS, por sus siglas en inglés), que cumplen con el objetivo de generar lecturas económicas y precisas. Las tecnologías de secuenciación incluyen métodos que se agrupan ampliamente en; preparación del molde, secuenciación y el correspondiente análisis de datos. La combinación única de protocolos específicos distingue una tecnología de otra y determina el tipo de datos producidos en cada plataforma (Metzker, 2010).

La metodología de Nanopore es un ejemplo de 'secuenciación de una hebra', que involucra secuenciar al pasar una hebra de ADN a través de un nanoporo proteico embebido en una membrana de polímero sintética (Cummings *et al*., 2017); la longitud de los fragmentos secuenciados se encuentra en el orden de las decenas de kilobases, y en algunos casos en el orden de hasta megabases (Branton *et al*., 2010). Los cambios del potencial de membrana son monitoreados en tiempo real cuando el ADN pasa a través del poro y dependen de la composición de las bases del ADN, por lo que son registrados e interpretados computacionalmente. El proceso de *base-calling* (llamado de bases) genera archivos de secuencias con valores de calidad asociados (formato fastq) a partir de los archivos de flujo de voltaje intermediarios en formato .fast5, los cuales contienen los niveles de la señal eléctrica cruda medida en los nanoporos. Las lecturas provenientes de esta plataforma tienen valores de calidad Phred de 10 a 20 (entre un 90 y 99% de precisión de que la base sea predicha de forma correcta), por lo que la detección confiable de mutaciones de un sólo nucleótido requiere información redundante de dicha posición (Ebler *et al*., 2018).

La plataforma Illumina está basada en secuenciación por síntesis, registrando la adición de nucleótidos etiquetados mientras la cadena de ADN es copiada, en un procedimiento masivo en paralelo. Esta tecnología es accesible y popular por sus bajos costos y datos de alta calidad (Shin

et al., 2013); sin embargo, las lecturas producidas por la plataforma Illumina tienen una longitud máxima de 250 bases. Si bien las secuencias obtenidas son muy precisas (99.9 a 99.99%; que equivale a valores de calidad Phred de 30 a 40), en regiones altamente repetitivas generan ambigüedad por lo que no se pueden utilizar para reconstruir regiones genómicas o cromosomas eucariotas a través de su ensamble (George *et al*., 2017).

Al unificar los beneficios de las dos tecnologías mencionadas anteriormente se obtiene un ensamble híbrido; un enfoque en el cual las lecturas largas representan el molde general para un genoma contiguo y las lecturas cortas funcionan como soporte para corregir los errores asociados a las lecturas largas. Con esta aproximación se han ensamblado genomas de varias especies de peces, entre las que destacan: *Danionella translucida,* Cyprinidae (Kadobianskyi *et al*., 2019)*; Gadus morhua,* Gadidae (Kirubakaran *et al*., 2020)*; Thamnaconus septentrionalis,* Monacanthidae (Bian *et al*., 2019b)*; Maccullochella peelii,* Percichthyidae (Austin *et al*., 2019); *Neoceratodus forsteri*, Ceratodontidae (Meyer *et al*., 2021) porque incluyeron ambas tecnologías para la obtención de genomas quasi-completos.

2.2. Recursos Genómicos

La generación de recursos genómicos es un proceso exhaustivo que mayormente se ha dirigido a la caracterización del genoma mitocondrial por su facilidad; aunque es una pequeña porción de toda la información genómica, su análisis permite entender las relaciones filogenéticas y caracterizar las tendencias evolutivas. La información del genoma completo abre más posibilidades de interpretación, sin embargo, el ensamble *de novo* representa un reto bioinformático, por lo que sólo se tiene información para pocas especies.

2.2.1. Genoma Mitocondrial

La mitocondria es un organelo celular que tiene la función de formar ATP y contiene ADN mitocondrial (ADNmt), al cual también se le puede llamar genoma mitocondrial. Es heredado

generalmente de la madre y tiene una tasa evolutiva de 5 a 10 veces más rápida que el ADN nuclear (Ramon, 1998). Los patrones de variación en el ADNmt son cada vez más investigados en especies amenazadas, ya que se han identificado aplicaciones de sus análisis que se clasifican en dos áreas distintas: la primera es 'conservación de genes', a través de la identificación y el manejo de la diversidad genética; la segunda área corresponde a 'ecología molecular', la cual consiste en usar las variaciones del ADNmt para dirigir estudios demográficos de poblaciones (Moritz, 1994). Además, estudios *in vitro* e *in situ* demuestran que el ADNmt permite analizar la respuesta a ciertos compuestos genotóxicos como; pesticidas, metales, contaminantes de aire, toxinas algales, entre otros (Roubicek *et al*., 2017).

2.2.2. Genoma Completo

Desde el momento en el que se describió la estructura de doble hélice del ADN por Watson y Crick (1953), ellos definieron que "la secuencia de las bases a lo largo de la cadena es irregular" lo cual demuestra el papel del ADN en el almacenamiento de la información genética y resalta la importancia de descifrar la secuencia exacta de las bases de cada organismo (Giani *et al*., 2020). El surgimiento de los genomas completos posterior a este descubrimiento fue gradual (Fig. 1), comenzando por genomas sencillos hasta poder obtener genomas completos de más de 32 Gb. Una vez que se obtiene un genoma completo se procede a realizar una anotación genómica. Esto consiste en localizar los genes, determinar su estructura y, por lo tanto, las proteínas que producen. Lo anterior comúnmente se complementa con un transcriptoma, un conjunto de secuencias del ARN mensajero (mARN) que permiten definir cuándo y dónde se activa cada gen en las células o tejidos de un organismo (NIH, 2019). Para entender la caracterización de genomas completos es imprescindible comprender los siguientes términos: "*contig*" serie de secuencias de ADN que se traslapan para crear un mapa físico que reconstruye una secuencia consenso de ADN (Gregory, 2005); "*BUSCO completeness*" es un criterio de valoración de qué tan completos están los ensambles (es decir su completitud), utilizando como referencia la cantidad de genes ortólogos de una sola copia presentes en el genoma y su valor dependerá de la base de datos utilizada; "*scaffold*" es un enlace de bases no determinadas entre dos series de secuencias genómicas no-contiguas, separadas por brechas de longitud conocidas (Waterson, 2002).

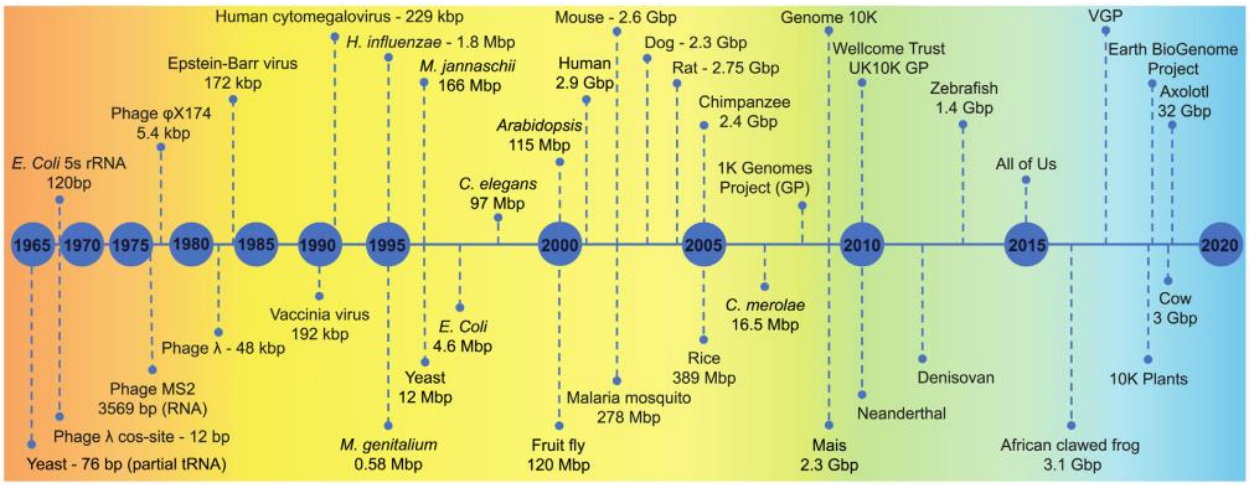


Figura 1. Progreso de la adquisición de genomas completos. Imagen tomada de Giani *et al*., 2020.

El grupo de los peces teleósteos representa más de la mitad de los organismos vertebrados, tiene más de 36,000 especies válidas, de las cuáles solo se tiene información genómica disponible para el 0.2% (Malmstrøm *et al*., 2017). En el Cuadro 1 se muestran las especies de teleósteos a los que su genoma ha sido completamente secuenciado. Es notable que el interés por obtener el genoma completo de una especie surge principalmente por propósitos comerciales y de conservación.

Cuadro 1. Genomas completos disponibles dentro del grupo de los teleósteos

Especie	Scaffold N50 (kb)	Contig N50 (kb)	BUSCO completeness (%)	Cantidad de lecturas	Tamaño obtenido (Mb)	Autor
1. *Fugu rebripes*			95	3.71 M Illumina	332.5	Aparicio *et al*., 2002
2. *Cynoglossus semilaevis*	867	26.5		205 x 10^6 lecturas Shotgun	477	Chen *et al*., 2014
3. *Larimichthys crocea*	1,030	63	95	324 Gb y 36 Gb	679	Ao *et al*., 2015
4. *Scleropages formosus*	58		93.95	297 M paired-end 290 M mate-pair	708	Austin *et al*., 2015
5. *Scophthalmus maximus*	4,300	31.2	98	580 M Illumina Genome Analyzer	544	Figueras *et al*., 2016

				Y HiSeq 2000		
6. *Mola mola*	9,000	20	74	98.22 Gb Illumina HiSeq	642	Pan *et al*., 2016
7. *Protosalanx hyalocranius*	1163	17.2	98.39	252. 1 Gb Illumina HiSeq 2000	536	Liu *et al*., 2017
8. *Maccullochella peelii*	109.9		94.20	47.2 Gb Illumina HiSeq y MiSeq	633	Austin *et al*., 2017
				804 Mb MinION		
9. *Arapaima gigas*	668	51.23	94.60	103 Gb	661	Vialle *et al*., 2018
10. *Pelteobagrus fulvidraco*	3,650	970	94.80	314.37 Gb Illumina	714	Zhang *et al*., 2018
				25.47 Gb PacBio		
11. *Bagarius yarrelli*	3,100	1,600	93.40	53.9 y 79.3 Gb Illumina		Jiang *et al*., 2019
				20.2 Gb PacBio	571	
12. *Thamnaconus septentrionalis*		22,460	94.33	45.97 Gb Illumina	474	Bian *et al*., 2019b
				50.95 Gb Nanopore		
13. *Oreochromis aureus*	1,100	53.2	97.80	239 Gb Illumina HiSeq 2500	920	Bian *et al*., 2019a
14. *Danionella translucida*	340	133.13	91.50	1.3 Billones Illumina	735	Kadobianskiy *et al*., 2019
				825k Nanopore		
15. *Epinephelus lanceolatus*	1,500	1,460	93.10	82.8 Gb Illumina	1,120	Wang *et al*., 2019
				31.15 Gb PacBio		
16. *Anguilla japonica*	1,030	11.47	83.90	268.61 Gb Illumina	1,130	Chen *et al*., 2019
17. *Poeciliopsis retropinna*	21,600		97.50	240 M HiSeq	621	
				6.75 M PacBio		

18. *Poeciliopsis turrubarensis*	4,200		95.50	344 M HiSeq X	597	Van Kruistum *et al*., 2020
19. *Gadus morhua*		10,500	94.10	71M Illumina 2.8M Nanopore	685	Kirubakaran *et al*., 2020
20. *Fundulus xenicus*		2.57	90.50	327.5 M Illumina HiSeq 38.5 Gb PromethION	1,070	Johnson *et al*., 2020
21. *Fundulus catenatus*		3.6	90.40	316.5 M Illumina HiSeq 40.3 Gb PromethION	1,160	Johnson *et al*., 2020
22. *Fundulus nottii*		3.749	94.40	197 M Illumina HiSeq 33.4 Gb PromethION	1,080	Johnson *et al*., 2020
23. *Fundulus olivaceus*		3.67	92.10	601.9 M Illumina NovaSeq 50.1 Gb PromethION	1,190	Johnson *et al*., 2020
24. *Neoceratodus forsteri*	1,750,000	1,860	91.4	1.2 Tb Nanopore 500Gb Illumina	43,000	Meyer *et al*., 2021

Los genomas de peces son muy variados en cuanto a su longitud, ya que dependen de la complejidad del ciclo de vida que se lleve, los hay desde 332 Mb hasta 1.19 Gb (Aparicio *et al*., 2012; Johnson *et al*., 2020). El genoma completo es una herramienta versátil que ha ayudado a descubrir diversos patrones evolutivos (Austin *et al*., 2015), recursos biomédicos (Bian *et al*., 2019), identificación de toxinas (Zhang *et al*., 2018), adaptaciones a distintos estilos de vida (Chen *et al*., 2014), genes peptídicos con funciones antimicrobianas (Wang *et al*., 2019), señales de selección debido a adaptación por contaminación (Osterberg *et al*., 2018), tasas de crecimiento (Pan *et al*., 2016), tiempos de cladogénesis (Vialle *et al*., 2018), estrategias reproductivas (van Kruistum *et al*., 2020), masculinización espontánea (Fraslin *et al*., 2020), adaptaciones a estrés (Ao

et al., 2015), entre otras aplicaciones.

Con respecto a la subfamilia Goodeinae, la mayor parte de la información genética se ha dirigido a entender las relaciones filogenéticas (Webb *et al*., 2004: Doadrio y Domínguez, 2004), los patrones biogeográficos y de distribución (Domínguez *et al*., 2006), así como los aspectos reproductivos (Vega *et al*., 2007). Sin embargo, no se cuenta con la caracterización del genoma completo de ninguna especie de esta subfamilia e incluso de la familia Goodeidae.

2.3. Los Goodeidos

En la familia Goodeidae se reconocen dos subfamilias bien delimitadas: Empetrichthyinae y Goodeinae. Las especies pertenecientes a la subfamilia Empetrychtyinae se distribuyen en Estados Unidos de Norteamérica y son ovíparas con fertilización externa (Foster y Piller, 2018). Mientras que las que pertenecen a la subfamilia Goodeinae sólo se encuentran en el centro de México y coloquialmente se les llama mexclapiques (Page *et al*., 2013). Presentan dimorfismo sexual, tienen una longitud menor a 15 cm y se encuentran en cuerpos de agua como lagos, lagunas y algunos ríos de flujo lento (Domínguez *et al*., 2005). En esta subfamilia los organismos tienen viviparismo matrotrófico, condición caracterizada por la presencia de un análogo placentario para la nutrición embrionaria, llamada trofotenia (Nelson, 1994). Las poblaciones de las especies de esta subfamilia van en declive desde hace décadas, debido a la progresiva y constante alteración de sus hábitats.

En especies de este grupo se han realizado estudios sobre su sensibilidad a diversos compuestos. Por ejemplo, en *Girardinichthys viviparus* se observó que al tomar organismos de acuarios y exponerlos a aguas del Lago Texcoco y el Lago Zumpango (localidad candidato para su reintroducción) se inducen patrones de estrés oxidativo y se observaron efectos de disrupción endócrina (Vega-López *et al*., 2008; Vega-López *et al*., 2007). Al exponer a *Chapalichthys pardalis* a nanopartículas de plata también se observó estrés oxidativo y un incremento de la energía consumida por los peces (Valerio-García *et al*., 2007). En otros estudios donde se expuso a *Ameca splendens* a altas concentraciones de bifenilos policlorados (PCBs) se registraron cambios en la expresión de vitelogenina, sugiriendo que este contaminante puede alterar la distribución de sexos en la naturaleza (Vega-López *et al*., 2008). En *Skiffia multipunctata y Goodea atripinnis* se

estudió la respuesta a los nitritos, mostrando efectos subletales a concentraciones por debajo de la Norma Nacional Mexicana sobre la calidad del agua (Rueda-Jasso *et al*., 2017). En hembras gestantes de *Xenotoca eiseni* se evaluó la exposición a 17α-etinilestradiol y se encontró que este compuesto a dosis altas sobre regula la vitelogenina hepática (Tingueley, 2015).

La subfamilia goodeinae está conformada por 41 especies clasificadas en 19 géneros (Domínguez y Pérez, 2007; Page *et al*., 2013). Uno de los géneros más abundantes es el género *Skiffia*, este contiene dos especies amenazadas (*S. lermae* y *S. multipunctata*), una especie críticamente amenazada (*S. bilineata*) y una especie extinta (*S. francesae*), por lo que representan un panorama relevante de la subfamilia.

Las especies del género *Skiffia* se caracterizan por presentar un esqueleto pélvico; un sistema sensorial cefálico solo de neuromastos expuestos, sin poros o canales supraorbitarios; la aleta dorsal (con 11-17 radios) detrás de la inserción pélvica; la aleta anal (con 13-17 radios) más corta en la base que los radios más largos. Los machos se distinguen por un moteado sobre la mitad dorsal del cuerpo y una media luna oscura en la base de la aleta caudal; la aleta dorsal con muescas y los primeros cuatro (o cinco) radios separados de los siguientes por una membrana interradial dentada. Las poblaciones de las especies que comprenden el género *Skiffia* se han reducido drásticamente, al grado que *S. francesae* se considera extinta en la naturaleza, sólo se conservan ejemplares en acuarios. Para iniciar la caracterización genómica de los goodeinos se seleccionaron las cuatro especies del género *Skiffia* (donde *S. lermae* destaca por ser una especie amenazada con algunas poblaciones estables) y a *Girardinichthys viviparus*, ya que en conjunto son representativas de las reducciones poblacionales de la subfamilia.

A continuación, se presenta información sobre los aspectos biológicos y estado de conservación para cada una de las especies consideradas en este estudio, tomada mayormente de Domínguez (2006) y en el Anexo 1 se muestran imágenes al respecto:

***Skiffia bilineata* (Bean, 1887)**. Encontrada históricamente en 16 localidades de la cuenca del lago de Cuitzeo en Michoacán y la cuenca adyacente del río Lerma en Guanajuato. No se han visto organismos en el Lago Cuitzeo desde la década de 1990. Actualmente sobrevive sólo en cinco sitios. Esta especie se encuentra críticamente amenazada debido a la contaminación del agua y especies invasoras que amenazan a las poblaciones restantes. Se estima que sólo dos poblaciones logren persistir dadas las tendencias de degradación ambiental y el estado de las poblacionales actuales. Los miembros de esta especie resisten temperaturas bajas y pueden producir crías incluso

a 10°C; de igual manera tiene una tolerancia a temperaturas por encima de 30°C. La temporada reproductiva abarca de marzo a mayo. Es una especie omnívora-micrófaga, prefiere comer invertebrados pequeños y detritus, la talla máxima descrita es 42 mm de longitud total.

***Skiffia francesae* Kingston, 1978**. No se han recolectado individuos de la naturaleza desde la década de 1970, por lo que se considera extinta. La fragmentación y modificación de los manantiales de Teuchitlán para convertirlos en un área de recreación posiblemente contribuyó a la desaparición de esta especie. Se conservan poblaciones cautivas en acuarios de Europa y Norteamérica. Tiene poca capacidad reproductiva (8-15 individuos) y muchas de las crías tienen deformidades y características erráticas de nado. Se requieren análisis genéticos antes de poder armar un plan de reintroducción para esta especie. Su dieta es principalmente herbívora. La talla máxima descrita es 43 mm de longitud total.

***Skiffia lermae* Meek, 1902**. Solía distribuirse en 18 sitios en las cuencas de Pátzcuaro, Zirahuén y Cuitzeo, en el lago de Zacapu y en algunas áreas del río Lerma en el Estado de Guanajuato, coloquialmente se le conoce como *Skiffia* olivo. La distribución y abundancia de esta especie ha declinado de forma constante durante los últimos 50 años con pérdidas fuertes desde el 2000. Ha desaparecido del lago de Yuriria, lago de Cuitzeo, río Grande, La Maiza, Cointzio y completamente de la cuenca del lago de Zirahuén (Lyons *et al.*, 1998; De la Vega-Salazar, 2003; Domínguez-Domínguez *et al.*, 2005; Fig. 2). Las poblaciones más estables se encuentran en el lago Zacapu y en el manantial La Mintzita en la cuenca del lago de Cuitzeo. Es una especie omnívora, pero con hábitos herbívoros fuertes, puede consumir insectos en la superficie del agua. La talla máxima descrita es de 64 mm de longitud total, a cantidad de crías varía de 3 a 15.

***Skiffia multipunctata* (Pellegrin, 1901)**. Conocida de los sitios en la parte baja de la cuenca del río Lerma, lago de Chapala y la parte superior la cuenca del río Santiago cerca de Guadalajara. Las poblaciones cerca de Guadalajara y en el lago de Chapala parecen haber desaparecido y aquellas en el río Lerma están fuertemente reducidas. Actualmente se distribuye en seis sitios y ha desaparecido de al menos otros ocho sitios. Aún se distribuye en algunos sitios en el drenaje del río Duero, incluyendo el reservorio Orandino, el arroyo Chilchota y el reservorio La Luz. Estos sitios están amenazados por la urbanización y escorrentías de agricultura que es complementado por el uso insostenible del agua de la cuenca. Su dieta es principalmente herbívora y la talla máxima descrita es de 72 mm de longitud total.

***Girardinichthys viviparus* (Bustamante, 1837).** Históricamente esta especie se distribuía

ampliamente por el valle de México, pero esta área ha sufrido degradación ambiental extrema debido al crecimiento de la Ciudad de México. Ahora sólo se mantienen tres poblaciones aisladas y su viabilidad está en duda. En muestreos anuales recientes sólo se encontraron tres especímenes. *G. viviparus* se encuentra extinto en el lago de Texcoco. La talla máxima observada es de 65 mm de longitud total, se reproducen casi todo el año excepto en los meses fríos. Son carnívoros, se alimentan de larvas de mosquitos e insectos pequeños.

Particularmente se seleccionó a *S. lermae* debido a su condición amenazada, pero con algunas poblaciones estables, ya que representa un buen candidato como modelo de la subfamilia para caracterizar su genoma. Respecto a la reducción de sus poblaciones, en 2003 se evaluaron los 14 sitios probables de distribución y se encontró sólo en tres, sin embargo, la densidad de las poblaciones era estable por lo que se le determinó como una especie de riesgo bajo (De la Vega-Sálazar *et al*., 2003). En el año 2008 se realizó un seguimiento sobre el estado de los goodeidos y se encontró para esta especie un valor de presencia de 0.45. Este valor se obtiene de la razón del número de localidades con registros actuales entre el número de localidades con registros previos de presencia, implica que se distribuye en menos de la mitad de sus localidades con registros previos y es una especie que corre peligro de extinguirse en los próximos años. En el 2019 fue integrada a la lista roja de especies amenazadas de la IUCN (International Union for Conservation of Nature) (Domínguez *et al*., 2008; Koeck, 2019).

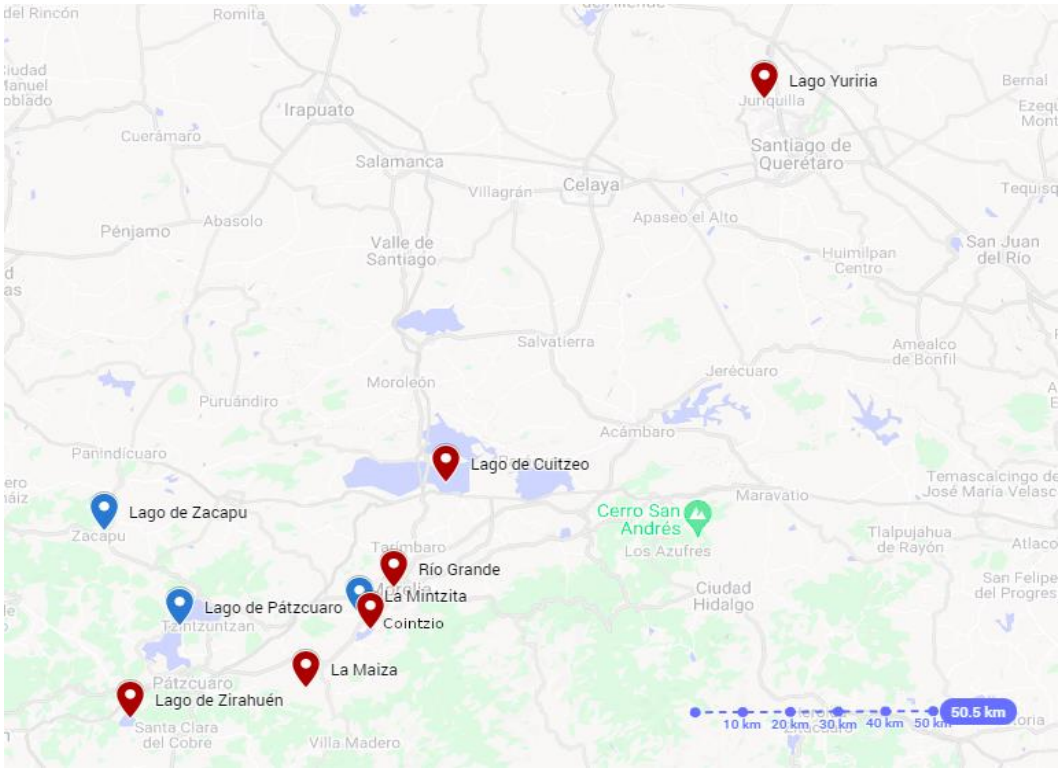


Figura 2. Distribución de *Skiffia lermae.* Los sitios azules son sus sitios de distribución hoy en día y los sitios rojos son sus sitios de distribución de hace 50 años, donde no se encuentra en la actualidad.

2.6 Información Genómica y Evaluación de Patrones de Selección

La información genómica ha cambiado paradigmas y ha permitido derivar los principios evolutivos fundamentales (Lewin *et al*., 2018). Este conocimiento ha sido crucial en la conservación a través de la identificación de alelos adaptativos y ha mejorado el rescate evolutivo basado en patrones genómicos de endogamia (Supple y Shapiro, 2018). Los principios del rescate evolutivo proponen que los seres vivos pueden recuperarse de las presiones ambientales a través de una modificación genética ventajosa, contrario al flujo génico, la migración de individuos o la dispersión (Carlson *et al*., 2014).

Para identificar estas modificaciones genéticas existen diversos estudios que han utilizado la proporción de las sustituciones no sinónimas por sitio no sinónimo (d_N) sobre las sustituciones sinónimas por sitio sinónimo (d_S) para hacer inferencias sobre la presión de selección en especies de peces simpátricas. Considerando que dicha proporción (d_N/d_S = ω) genera tres escenarios

posibles: a) si un aminoácido tiene un cambio neutral, será fijado a la misma tasa que una mutación sinónima por lo que se obtendrá ω = 1; b) si el cambio en el aminoácido es deletéreo, la selección purificadora reducirá su tasa de fijación, por lo tanto se obtendrá ω <1; y c) si el cambio en el aminoácido ofrece una ventaja selectiva éste se fijará a una tasa mayor que una mutación sinónima, lo que resultará en ω >1 (Yang y Bielawski, 2000). En la práctica, estimar directamente la tasa d_N/d_S no es una tarea sencilla, pero las distancias no sinónimas y sinónimas entre genes ortólogos en una filogenia sí se pueden estimar de un alineamiento de secuencias, llegando finalmente a la estimación de ω (Jeffares *et al.,* 2015). Existen dos clases de métodos para realizar esta estimación, la primera clase son métodos intuitivos; se basan en contar los sitios sinónimos (*S*) y no sinónimos (*N*) en dos secuencias, calcular la diferencia entre estos conteos y después hacer correcciones para sustituciones múltiples en el mismo sitio. La mayoría de los aminoácidos en una proteína funcional están bajo limitaciones estructurales y funcionales, por lo que la evolución adaptativa probablemente incide sólo en algunos sitios y en ciertos puntos temporales, por lo tanto, este enfoque de promediar tasas de todos los sitios a través del tiempo tiene poco poder (Yang y Nielsen, 2002). La segunda clase son métodos basados en máxima verosimilitud apoyados en modelos explícitos de sustitución de codones y sus parámetros como la divergencia de las secuencias, la tasa de transición/transversión y la tasa de mutaciones, que son estimadas a partir de los mismos datos. Este enfoque permite realizar pruebas estadísticas para evaluar si d_N es significativamente mayor a d_S, estableciendo un modelo nulo con ω fijado a 1, mientras que el modelo alternativo estima ω como un parámetro libre. Para evaluar si ω es diferente de 1, se compara el doble de la diferencia del logaritmo de verosimilitud y se aproxima mediante una distribución chi cuadrada, con grados de libertad igual a la diferencia en el número de parámetros libres en los dos modelos anidados (Yang, 1998; Swanson *et al*., 2001).

El parámetro ω se ha utilizado en múltiples organismos, que van desde bacterias (Kryazhimskiy y Plotkin, 2008), virus (Moury y Simon, 2011; Nielsen, 1999), ratones, chimpancés hasta humanos (Biswas *et al*., 2016), para evaluar los patrones de selección. En un estudio de 2001 se comprobó que los genes codificantes para proteínas ZP2, ZP3 y OGP se encuentran bajo selección positiva en las hembras de diferentes mamíferos a través de distintas especies (Swanson *et al*., 2001), lo que demuestra su utilidad independientemente del grupo de organismos. Particularmente en peces, se identificó la presión selectiva (ω >1) en genes ortólogos correspondientes al metabolismo y procesos biosintéticos, en dos especies simpátricas de cíclidos que difieren en sus conductas

reproductivas y en sus hábitos alimenticios (Elmer *et al*., 2010). En otras especies de cíclidos se identificaron cambios asociados con la forma del cuerpo, estructuras tróficas, coloración y determinación sexual debido a la fuerte selección aplicada en sus genes (Fan *et al*., 2012). En tilapia se encontró que los genes relacionados al sistema inmune han sido seleccionados positivamente mostrando que están sometidos a presión ambiental (Xiao *et al*., 2015).

3. HIPÓTESIS

Dado que las poblaciones de la mayoría de las especies de goodeinos están sujetas a procesos de reducción de distribución, con los recursos genómicos generados se espera identificar tasas de evolución acelerada ($\omega>1$) en regiones codificantes de genes asociados al metabolismo de xenobióticos y esto evidencie presión de selección divergente

4. OBJETIVOS

4.1 Objetivo General

Realizar la caracterización genómica de *Skiffia lermae* mediante una aproximación híbrida y analizar la presión de selección en genes particulares con información genómica de especies cercanas

4.2 Objetivos Específicos

1.- Caracterizar el genoma mitocondrial y nuclear de *S. lermae*

2.- Ensamblar y anotar el transcriptoma de *S. lermae*

3.- Ensamblar y anotar el transcriptoma de hígado de *S. bilineata*, *S. francesae, S. multipunctata* y *Girardinichthys viviparus*

4.- Identificar tasas de evolución acelerada en regiones codificantes de genes asociados al metabolismo de xenobióticos a partir del transcriptoma de hígado en *S. bilineata, S. francesae, S. lermae*, *S. multipunctata* y *G. viviparus*

5. MATERIALES Y MÉTODOS

5.1 Adquisición de Tejidos y Ácidos Nucleicos

5.1.1 Obtención de organismos

Se obtuvieron siete organismos de *Skiffia lermae* del Laboratorio de Biología Acuática de la Facultad de Biología, de la Universidad Michoacana de San Nicolás de Hidalgo, proporcionados por el Dr. Omar Domínguez y la estudiante de doctorado M. en C. Ivette Villa. Se transportaron a Mazatlán donde se mantienen en cultivo a 26°C para establecer una colonia. Por otra parte, se realizaron disecciones de machos de otras especies en el mismo laboratorio de origen y se les extrajo el hígado. Las especies seleccionadas fueron: *S. bilineata, S. francesae, S. multipunctata* y *Girardinichthys viviparus.*

5.1.2 Adquisición de Tejidos de *S. lermae*

Se sacrificó un macho proveniente de Zacapu infestado por parásitos y el músculo se conservó en ultracongelación (-60°C) para su posterior extracción de ADN. Siete tejidos fueron colocados en 400 µl de trizol para realizar una posterior extracción de ARN: bazo, corazón, cerebro, hígado, intestino, branquia, gónada y parásitos. Adicional a esto, una larva recién nacida de la misma localidad fue sacrificada y colocada en trizol para realizar secuenciación de su ARN.

Durante la disección del organismo adulto macho de *S. lermae* se observaron trematodos parásitos que fueron identificados como *Tylodelphys sp.* con base en búsquedas en la literatura (Navarrete, 2013; Nava *et al*., 2004; Martínez-Aquino *et al.*, 2012) y asesoría especializada (*com. pers.* Dr. Neptalí Morales); una muestra de estos fue homogenizada en trizol para secuenciación transcriptómica que no forma parte de los objetivos, pero que se decidió incluir en el procesamiento de análisis moleculares.

5.1.3. Extracción de ADN y ARN

Se utilizó el protocolo de extracción de ADN basado en una lisis y digestión con proteinasa K. La precipitación de los ácidos nucleicos se realizó con una solución saturada de NaCl (Miller *et al*., 1988). El procedimiento detallado, conocido como *salting-out*, se indica en el Anexo 2. Se probaron también extracciones con protocolos basados en fenol-cloroformo, así como con el kit de Zymo Research Quick-DNA/RNA^{TM} Magbead (catalog # R2130). Se verificó la integridad del ADN mediante electroforesis y se cuantificó mediante espectrofotometría UV empleando un espectrofotómetro Nanodrop Lite.

La extracción de ARN de cada tejido se realizó siguiendo el protocolo de trizol, añadiendo cloroformo y precipitación con isopropanol (Rio *et al*., 2010). Los pasos detallados se encuentran en el Anexo 3. Se extrajo ARN de los tejidos tomados del macho de y una larva de *S. lermae* descritos en la sección 5.1.2. Posteriormente se extrajo ARN de hígado de: *S. bilineata*, *S. francesae*, *S. multipunctata* y *G. viviparus* y se cuantificó con el espectrofotómetro Nanodrop Lite. Todas las extracciones de ARN fueron procesadas para envío a Genewiz de acuerdo a lo explicado en la sección 5.4.1.

5.2. Secuenciación en Nanopore

5.2.1. Preparación de Librerías para MinION

El ADN fue cuantificado mediante fluorometría con Qubit. A partir de 2.7 µg de ADN totales, se preparó una biblioteca con el kit de Genomic DNA by Ligation (SQK-LSK109). Se siguió el protocolo indicado por el proveedor, con ligeras modificaciones (Anexo 4). En resumen, consiste en reparar el ADN de posibles daños por almacenamiento y manipulación, adicionar un residuo de adenina en el extremo 3’ para optimizar la ligación de adaptadores. Luego se prepara el Flowcell para el cargado, en este caso, se agregó 1 µg de librería. Se utilizó un dispositivo MinION (Oxford

Nanopore Technologies, Oxford, UK) y se usaron Flowcells R9.4.1 hasta lograr un mínimo de 20 Gb de lecturas.

5.2.2. Experimentos de Secuenciación

Para generar la cantidad de lecturas necesaria se tuvieron que realizar dos experimentos de secuenciación. El primer experimento tuvo una duración de 72 h y generó 12 Gb de lecturas mientras que el segundo duró 54 h y generó 6.2 Gb. Después de obtener los archivos crudos se procedió a utilizar *Guppy_basecaller* para producir las secuencias con su calidad correspondiente, con la configuración de GPU (Anexo 5, sección 1).

5.2.3. Verificación de la Calidad de las Lecturas Obtenidas por MinION

El trabajo bioinformático indicado a continuación fue llevado a cabo en el servidor Chihuil-ICMYL (http://www.icmyl.unam.mx/mazatlan/), el cual posee 192 núcleos y 512 Gb de RAM. Para verificar la calidad de las lecturas producidas se utilizó *pycoQC* (Leger y Leonardi, 2019). Este programa produce un archivo html con gráficas interactivas y recibe como argumento un resumen de la corrida que se genera en automático por *MinKNOW* al terminar de secuenciar. Debido a que se interrumpió el proceso de basecalling el archivo sequencing.summary.txt era parcial (no incluía la totalidad de las lecturas generadas) por lo cual se tuvo que generar otro con todas las lecturas una vez recuperadas. El programa utilizado para esto forma parte del ambiente de *pycoQC*. Los detalles de estos comandos se encuentran en el Anexo 5, sección 2.

5.3 Genoma Mitocondrial y Nuclear de *S. lermae*

5.3.1 Preparación de ADN para Secuenciación en GeneWiz

Partiendo del ADN extraído por *salting-out*, se llevó a cabo un filtrado con RNAsa para evitar impurezas y se siguió el protocolo que se encuentra en el Anexo 6. El ADN extraído fue purificado con perlas AMPure para mandarlo al servicio de secuenciación GeneWiz. Se utilizaron 20 µl de ADN y 8 µl de perlas (proporción 0.4X) y se solicitaron 350 M de lecturas Illumina pareadas.

5.3.2 Preparación de las Lecturas

Para revisar la calidad de las lecturas cortas provenientes de Illumina se utilizó *FastQC* (Andrews, 2010) y para las lecturas largas se usó *PycoQC* (Anexo 5, sección 2) (Leger *et al*., 2019). Para limpiar las lecturas cortas de adaptadores se utilizó *trimmomatic* (Bolger *et al*., 2014) y para las lecturas largas se utilizaron dos programas: *porechop* para retirar adaptadores y *canu* (Koren *et al*., 2017) para corregir y cortar los extremos. Los comandos utilizados están detallados en el Anexo 5, sección 2 y 4.

5.3.3 Ensamble del Genoma Mitocondrial de *S. lermae*

A partir de las lecturas producidas de Illumina transformadas a formato FASTA, se utilizó *Mirabait* (Chevreux *et al*., 1999) para seleccionar mediante empalmes parciales en ciclos iterativos aquellas secuencias que correspondan con un genoma mitocondrial de referencia (*Xenotoca eiseni*, acceso GenBank AP006777.1). Para ensamblar el genoma mitocondrial se utilizó *Megahit* (Li *et al*., 2015). Por último, para anotar el genoma mitocondrial se usó *MitoZ* desde su repositorio en docker

(Meng *et al*., 2018); los parámetros específicos se reportan en el Anexo 5, sección 3. El mitogenoma fue reorientado manualmente, fijando la región codificante para el tRNA-Phe (ARN de transferencia para la fenilalanina) como sitio de inicio, basado en lo observado en mitogenomas de; *Xenotoca eiseni* (AP006777.1), *Fundulus heteroclitus* (NC_012312.1) y *Poecilia reticulata* (NC_024238).

5.3.4 Estimación del Tamaño del Genoma Completo

Para estimar el genoma se utilizó la aproximación de conteo de k-meros, sugerida en el Vertebrate Genome Project (VGP). El programa usado fue *jellyfish* para el conteo y para los cálculos se utilizó R. Brevemente, se realizó un conteo de k-meros de longitud 31, recomendado para estimar el tamaño del genoma completo de vertebrados por el VGP (Rhie *et al*., 2020). Este conteo después se transformó en un histograma, se graficó y visualizó en R para encontrar la región de una sola copia. Una vez identificada esta zona se realizó la suma de k-meros totales y se aplicó una ecuación para calcular el tamaño del genoma aproximado. Los detalles de los comandos utilizados se encuentran en el Anexo 5, sección 5.

5.3.5 Ensamble Híbrido del Genoma de *S. lermae*

El ensamble de las lecturas largas fue llevado a cabo con *Smartdenovo* (Liu *et al*., 2021) ya que este se ha usado para producir genomas a escala cromosoma (Descchamps *et al*., 2018), después se hizo el alineamiento de las lecturas cortas al ensamble obtenido de las lecturas largas, se utilizó *Minimap2* (Li, 2018). Este alineamiento fue ordenado e indexado con *samtools* (Danecek *et al*., 2021). Posteriormente se utilizó el alineamiento y el índice para pulir el ensamble con *Pilon* (Walker *et al*., 2014), se usaron todas las opciones de corrección. La descripción de los comandos utilizados se encuentra en el Anexo 5, sección 6.

5.3.6 Evaluación y Anotación del Ensamble

Se utilizó *quast* para obtener algunas estadísticas generales del ensamble. Para evaluar qué tan completo está el genoma se realizó una búsqueda con *BUSCO* (Manni *et al*., 2021) utilizando la opción de *Augustus*, para también realizar la anotación del genoma. Adicional a esto se utilizaron funciones en *R* para determinar el N50 y N25 (Hamm, 2014). Estos comandos se encuentran detallados en el Anexo 5, sección 7 y 8.

5.4 Obtención y Procesamiento del Transcriptoma

5.4.1 Preparación de ARN para Secuenciación en Genewiz

Se utilizaron tubos de estabilización por desecación (RNA Stabilization; catálogo GTR5025-GW; Genewiz) y se calculó el volumen de muestra y agua a agregar, hasta colocar 2 µg de ARN de cada tejido por tubo y un mínimo de 20 µl de líquido. El protocolo seguido para secar las muestras fue el siguiente; se realizaron las diluciones para suspender 1 y 2 µg de ARN en 20 µl de agua en tubos eppendorf de 1.5 ml, de ahí se pasaron a los tubos de estabilización, primero al fondo del tubo y se dejaron incubar 5 minutos a temperatura ambiente. Posteriormente se mezclaron por pipeteo al menos diez veces. Estos tubos se colocaron en una campana de flujo laminar y se dejaron secar por 25 h y se prepararon para envío. Debido a que algunas de estas muestras llegaron en mal estado a su laboratorio destino como se observa en la Fig. 3, se realizó un segundo envío con diferentes concentraciones que llegó en buen estado, la calidad se muestra en la Fig. 4.

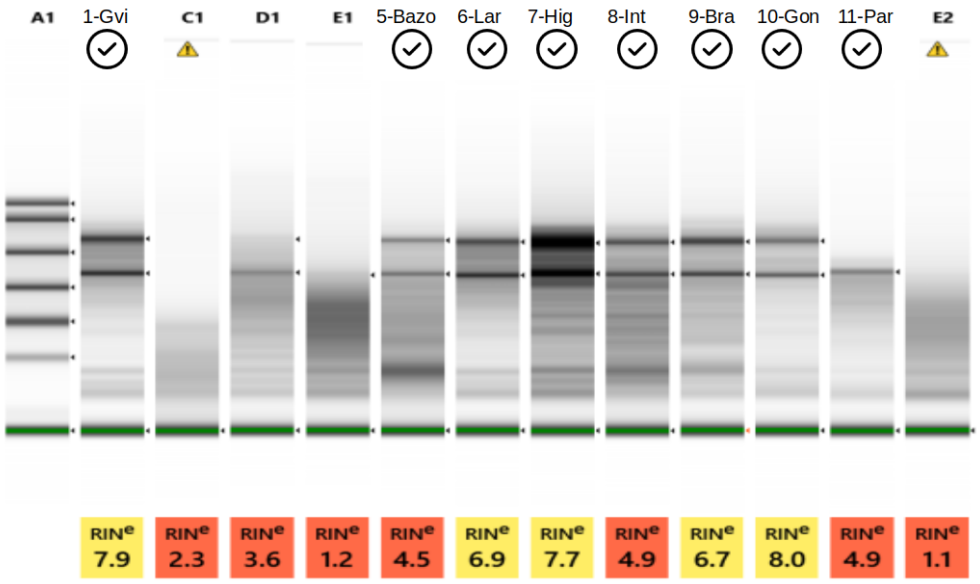


Figura 3. QC por parte de GeneWiz del primer envío realizado. Se indica cuáles muestras procedieron a la preparación de librerías con una marca de verificación sobre la barra. Los tejidos aprobados corresponden a *Girardinichthys viviparus* y a *Skiffia lermae*

Figura 4. QC por parte de GeneWiz de las muestras re sometidas debido a baja calidad. Se indica cuáles muestras procedieron a la preparación de librerías con una marca de verificación sobre la barra. Los tejidos aprobados corresponden a *S. bilineata*, *S. francesae*, *S. lermae* y *Skiffia multipunctata*,

5.4.2 Ensamble y Anotación de Transcriptomas

Para el caso de *S. lermae* las lecturas de ARN provienen de los siete tejidos y una larva completa, mencionados anteriormente en la sección 5.1.2. Para el resto de las especies el único tejido utilizado fue el hígado. La siguiente línea de trabajo fue seguida en todo el conjunto de lecturas, con la única diferencia de que para *S. lermae* se incluyeron más archivos. Primero se revisó la calidad de las lecturas con el programa *fastqc* (Andrews, 2010). Después se hicieron los ensambles *de novo* con *Trinity* (Grabherr *et al.*, 2011), utilizando la opción –trimmomatic que pasa un filtro a las lecturas antes de proceder al ensamble. Para ver las características del ensamble se utilizó *TrinityStats* y para observar la representatividad de las lecturas en el ensamble se utilizó el programa *bowtie2* (Langmead y Salzberg, 2012). Para analizar qué tan completos se encontraban los segmentos genómicos ensamblados se utilizó *BUSCO* (Benchmarking Universal Single Copy Orthologs; Manni *et al*., 2021) con dos bases de datos para el caso de los peces: *Actinopterygii* con 3640 grupos de BUSCO y *Cyprinodontiformes* con 15,213. Para el caso de los parásitos se utilizó la base de datos *Metazoa* con 954 representantes de ortólogos.

Se utilizó *TransDecoder* para identificar regiones codificantes en el ensamble, conservando los marcos de lectura abiertos (ORF) más largos, después se ejecutaron análisis con *BLAST* contra la base de datos de UniProt para describir las regiones codificantes. Se utilizó *HMMER* (Eddy, 1992) para buscar secuencias contra la base de datos de pFAM en el archivo saliente de *TransDecoder* que contiene las secuencias peptídicas. Se usó *RNAMMER* (Lagesen *et al*., 2007) para identificar las regiones ribosomales 16S/18S, 5S/8S y 23S/28S. Luego se utilizó la herramienta *TransDecoder.predict* para confirmar aquellos marcos de lectura con anotaciones basadas en similitud a pFAM y UniProt. Con el propósito de generar un reporte con toda la información sobre la anotación del transcriptoma se utilizó *Trinotate* (Bryant *et al*., 2017). Los detalles de todos estos comandos se encuentran en el Anexo 5, sección 4.

5.5 Estimación de Tasas de Mutación

5.5.1 Selección de Ortólogos

Los ortólogos utilizados para este análisis se enfocaron en genes donde se han observado respuestas diferenciales conforme a la literatura, para poder contextualizar los valores ω. Una vez construida la base de ortólogos se utilizó *BLAST* para encontrarlos en las secuencias de hígado de todas las especies: *S. lermae, S. bilineata, S. francesae, S. multipunctata* y *G. viviparus.* Los detalles de estos comandos se encuentran en el Anexo 5, sección 10. Se documentaron alrededor de 70 ortólogos, pero sólo se utilizaron aquellos encontrados en las cinco especies evaluadas. En el Anexo 7 se muestran los ortólogos descartados. Cabe destacar que este análisis se puede extender incluso en genes sin función definida, pero en este trabajo se limitó a los indicados en el Cuadro 2.

Cuadro 2. Ortólogos seleccionados para la determinación del parámetro ω

N°	Nombre	Abreviación	Longitud*	Referencia
1	Citocromo P450	CYP450	505	Andersson y Förlin, 1992
2	Citocromo P450 2N2	CYP450_2	498	Andersson y Förlin, 1992
3	Citocromo P450, familia 51	CYP51	500	Široká y Drastichova, 2004
4	Citocromo P450 2K1	CYP2K	522	Široká y Drastichova, 2004
5	Proteína de choque térmico 90-alfa	HSP90A	724	Iwama *et al*., 1998
6	Proteína de choque térmico beta-8	HSPB8	222	Iwama *et al*., 1998
7	Transportador de Zinc 5	ZNT5	769	Wang *et al*., 2017
8	Transportador de cobre	CTR1	530	Wang et al., 2019
9	ATPasa B transportadora de cobre	ATP7B	891	Wang *et al*., 2017
10	Transportador de serotonina sodio-dependiente	SERT	674	Burkina *et al*., 2015
11	Reductasa glutatión-disulfuro	GSR	512	Awasthi *et al*., 2018
12	NADPH oxidasa 1	NOX1	566	Awasthi *et al*., 2018
13	DNA metiltransferasa 1	DNMT1	1500	Laing *et al*., 2017
14	6-fosfofructo-2-quinasa/fructosa-2,6-bifosfatasa 3	PIK3R1	719	Nourizadeh *et al*., 2009

15	MDM4 regulador de p53	MDM4	364	Nourizadeh *et al*., 2009
16	Factor nuclear (eritroide-derivado) 2	NRF2	621	Burkina *et al*., 2015
17	Peroxisoma proliferador-activado receptor alfa	PPAR	492	Burkina *et al*., 2015
18	Proteina desacoplante 2	UCP2	313	Shaw *et al*., 2019
19	Pregnane X-receptor	PXR	570	Burkina *et al*., 2015
20	Receptor de estrógenos beta 1	ERB1	673	Burkina *et al*., 2015

* Longitud en bases nucleotídicas

5.5.2 Modelos de Mutación

Los ortólogos se alinearon con *clustal omega* (Larkin *et al*., 2007) el cual también realizó un árbol filogenético por cada ortólogo. Los detalles de este proceso se encuentran en el Anexo 5, sección 11. Posteriormente a cada ortólogo se le aplicaron los modelos de sustitución de codones M0 (una tasa), M3 (discreto), M7 (beta) y M8 (beta y ω) (Cuadro 3) con el programa *EasyCodeML* (Gao *et al*., 2019), en su interfaz gráfica. Con los resultados de estos modelos se corrieron pruebas de verosimilitud para verificar la significancia de la existencia de sitios seleccionados positivamente. La fórmula para el cálculo de la tasa de mutaciones se describe en la ecuación (1).

$$q_{ij}^{(h)} = \begin{cases} 0, & \text{si } i \text{ y } j \text{ difieren en dos o tres posiciones nucleotídicas} \\ \pi_j, & \text{si } i \text{ y } j \text{ difieren en una transversión sinónima} \\ \mathrm{K}\pi_j, & \text{si } i \text{ y } j \text{ difieren por una transición sinónima} \\ \omega^{(h)}\pi_j, & \text{si } i \text{ y } j \text{ difieren por una transversión no sinónima} \\ \omega^{(h)}\mathrm{K}\pi_j, & \text{si } i \text{ y } j \text{ difieren por una transición no sinónima} \end{cases} \quad \ldots.(1)$$

donde π_j es la frecuencia de equilibrio del codón j, K es la tasa de transición/transversión, y $\omega^{(h)}$ es la tasa de mutaciones d_N/d_S en el sitio h

Cuadro 3. Descripción de los modelos para determinar la variable ω entre sitios

Modelo	Descripción	Parámetros libres	Núm. De parámetros libres
M0 (una tasa)	Una tasa ω para todos los sitios	ω	1
M3 (discreto)	K = 3 clases de sitios	$\omega_0, \omega_1, \omega_2. p_0, p_1$	5
M7 (beta)	$\omega \sim B(p, q)$	p, q	2
M8 (beta y ω)	Proporción p_0 de sitios $\sim B(p, q)$, p_1 de sitios con clase discreta ω	$p, 1, p_0, \omega$	4

6. RESULTADOS Y DISCUSIÓN

6.1 Calidad del Material Inicial

6.1.1 Calidad del ADN para Secuenciación Nanopore

Del tejido muscular se realizó una serie de extracciones de manera sucesiva. El Cuadro 4 indica los parámetros de calidad obtenidos de Nanodrop lite, de cada una de las extracciones realizadas.

Cuadro 4. Calidad de las extracciones de ADN de *Skiffia lermae* realizadas

Extracción	Concentración (ng/µl)	Relación de absorbancia A_{260}/A_{280}
1.- Magbeads Zymo Research	29.4	1.81
2.- *Salting-out*	91	2.0
3.- Fenol cloroformo	190	2.01
4.- Fenol cloroformo	466	2.02
5.- *Salting-out*	108	1.72
6.- *Salting-out*	147	2.08

Se corrieron geles de agarosa al 0.8% en una cámara de electroforesis para verificar la integridad de las extracciones y se observó que las extracciones por *salting-out* resultaron las más consistentes (Fig. 5).

De los métodos utilizados para extraer ADN se puede destacar que el kit de Magbeads de Zymo no muestra una banda en la Fig. 5 (1). Esto pudo deberse a que el kit no estaba en el mejor estado o hubo algún error al interpretar el protocolo, aunque estos kits son reconocidos por arrojar ADN libre de impurezas (Zymo Research, 2020). Por otro lado, *salting-out* es un método estandarizado para la extracción de ADN, incluso de muestras consideradas “complicadas” por ejemplo, tejidos embebidos en parafina (Rivero *et al*., 2006). Este método arrojó los mejores resultados (Fig. 5:

(2)(5)(6)). Adicional a esto, se intentó la extracción por fenol cloroformo ya que es altamente recomendada para extraer ADN de alto peso molecular (Green y Sambrook, 2017). En este proyecto no fue apropiado, quizás por el estado original de la muestra, la cual probablemente ya estaba degradada para el momento en el que se intentó esta extracción.

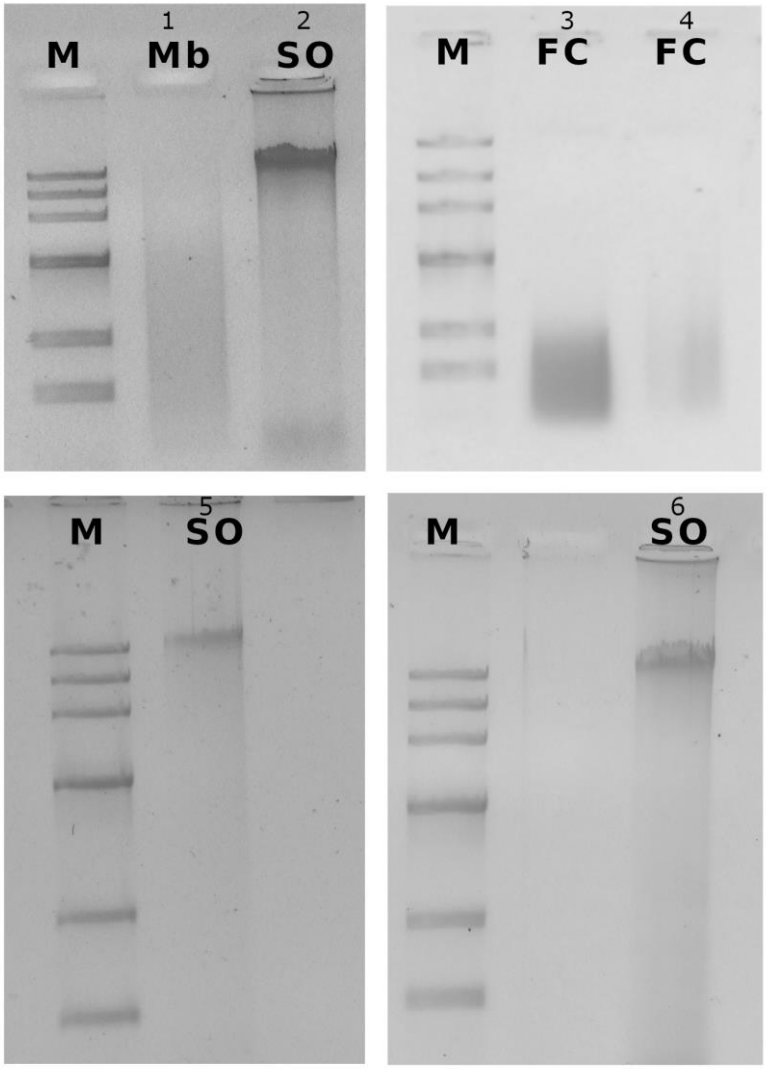


Figura 5. Electroforesis de las extracciones de ADN de músculo. M: escalera de peso molecular (Marker 4, Dongsheng Biotech), Mb: extracción con Magbeads, FC: extracción por Fenol-Cloroformo y SO: extracción por *salting-out.*

Dado lo anterior el ADN que se utilizó para la preparación de librerías fue proveniente de las extracciones por *salting-out*. Sin embargo, la librería producida con el ADN de la extracción (5)

no fue secuenciada debido a que el producto final contenía muy poca librería (50 ng) y el dispositivo necesita entre 500 ng - 1 µg de material inicial para funcionar apropiadamente.

6.1.2 Calidad del ARN para Transcriptoma

La calidad del ARN extraído de los tejidos extraídos de *S. lermae* y su concentración calculada en Nanodrop Lite se muestra en el Cuadro 5.

Cuadro 5. Evaluación de ARN de muestras de *Skiffia lermae*

Tejido	Cuantificación (ng/µl)	A260/A280
1.- Bazo y corazón	118	1.95
2.- Branquia	352	1.98
3.- Cerebro	390	2.01
4.- Gónada	469	1.92
5.- Hígado	1560	1.96
6.- Intestino	587	2.02
7.- Músculo	170	2.04
8.- Larva	1779	1.92
9.- Parásitos	100	1.65

6.1.3 Calidad de las Lecturas de MinION Producidas para Genoma

La primera corrida proveniente de la extracción (2) de *salting-out* generó 12 Gb de lecturas con un N50 de 4.7 kb con una Phred score promedio de 11. La segunda corrida proveniente de la extracción (6) mediante *salting-out* arrojó 6.16 Gb con una N50 de 5.3 kb y una Phred score promedio de 9. La calidad obtenida de las secuenciaciones Nanopore se encuentra dentro de los rangos normales (Austin *et al*., 2015) y el N50 de las lecturas está también dentro de los intervalos normales para

secuencias de peces que van desde 1 kb a 46 kb (Meyer *et al*., 2021; Kadobianskyi *et al*., 2019).

6.1.4 Calidad de las Lecturas de ADN de Illumina Obtenidas para Genoma

Se recibieron dos archivos, uno con las secuencias forward y otro con las secuencias reverse. Cada uno de estos archivos contiene 530,935,503 lecturas, otorgando un total de 99.8 Gigabases. Esto implica una cobertura de 69X para la estimación preliminar del genoma completo (1.43 Gb). La calidad promedio por lectura fue de 36. Las lecturas que se obtuvieron tienen suficiente calidad ya que la literatura indica que una phred score por debajo de 30 en secuencias Illumina es baja y por encima de 35 es de buena calidad para un genoma (Pawlowski *et al*., 2014).

6.2 Genoma Mitocondrial y Nuclear de *S. lermae*

6.2.1 Ensamble del Genoma Mitocondrial de *S. lermae*

El genoma mitocondrial producido fue de 16,551 bases. La anotación arrojó 37 genes, de los cuales 13 codifican para proteínas, 22 son correspondientes a tRNA y 2 a rRNA (Cuadro 6). La visualización espacial del mitogenoma se ilustra en la Fig. 6. Al observar la anotación y circularidad se demuestra que éste está completo, la longitud es casi idéntica al mitogenoma de otro goodeino, *Xenotoca eiseni* con 16,735 bases, y la anotación también tiene los mismos componentes (Setiamarga *et al*., 2008).

Cuadro 6. Descripción de los genes anotados en el mitogenoma de *Skiffia lermae*

Inicio	Final	Longitud (bp)	Tipo	Nombre	Producto	Frecuencia*
1	69	69	tRNA	trnF(gaa)	tRNA-Phe	1
69	1016	948	rRNA	s-rRNA	12S ARN ribosomal	1
1016	1088	73	tRNA	trnV(uac)	tRNA-Val	1
1089	2791	1703	rRNA	l-rRNA	16S ARN ribosomal	1
2791	2865	75	tRNA	trnL(uaa)	tRNA-Leu	2
2865	3840	976	CDS	ND1	NADH deshidrogenasa subunidad 1	1
3846	3916	71	tRNA	trnI(gau)	tRNA-Ile	1
3915	3986	72	tRNA	trnQ(uug)	tRNA-Gln	1
3985	4054	70	tRNA	trnM(cau)	tRNA-Met	1
4054	5101	1048	CDS	ND2	NADH deshidrogrenasa subunidad 2	1
5099	5172	74	tRNA	trnW(uca)	tRNA-Trp	1
5175	5244	70	tRNA	trnA(ugc)	tRNA-Ala	1
5246	5319	74	tRNA	trnN(guu)	tRNA-Asn	1
5356	5421	66	tRNA	trnC(gca)	tRNA-Cys	1
5424	5494	71	tRNA	trnY(gua)	tRNA-Tyr	1
5495	7046	1552	CDS	COX1	Citocromo c oxidasa subunidad 1	1
7060	7132	73	tRNA	trnS(uga)	tRNA-Ser	2
7135	7206	72	tRNA	trnD(guc)	tRNA-Asp	1
7212	7911	700	CDS	COX2	Citocromo c oxidasa subunidad II	1
7903	7877	75	tRNA	trnK(uuu)	tRNA-Lys	1
7978	8146	169	CDS	ATP8	ATP sintasa F0 subunidad 8	1
8136	8820	685	CDS	ATP6	ATP sintasa F0 subunidad 6	1
8819	9605	787	CDS	COX3	Citocromo c oxidasa subunidad III	1
9604	9676	73	tRNA	trnG(ucc)	tRNA-Gly	1
9676	10027	352	CDS	ND3	NADH deshidrogenasa subunidad 3	1
10025	10094	70	tRNA	trnR(ucg)	tRNA-Arg	1
10094	10391	298	CDS	ND4L	NADH deshidrogenasa subunidad 4L	1
10384	11770	1387	CDS	ND4	NADH deshidrogenasa subunidad 4	1
11765	11834	70	tRNA	trnH(gug)	tRNA-His	1
11834	11902	69	tRNA	trnS(gcu)	tRNA-Ser	2

11912	11985	74	tRNA	trnL(uag)	tRNA-Leu	2
11985	13824	1840	CDS	ND5	NADH deshidrogenasa subunidad 5	1
13820	14342	523	CDS	ND6	NADH deshidrogenasa subunidad 6	1
14342	14410	69	tRNA	trnE(uuc)	tRNA-Glu	1
14414	15555	1142	CDS	CYTB	Citocromo b	1
15555	15626	72	tRNA	trnT(ugu)	tRNA-Thr	1
15625	15695	71	tRNA	trnP(ugg)	tRNA-Pro	1

* Frecuencia total de ocurrencia

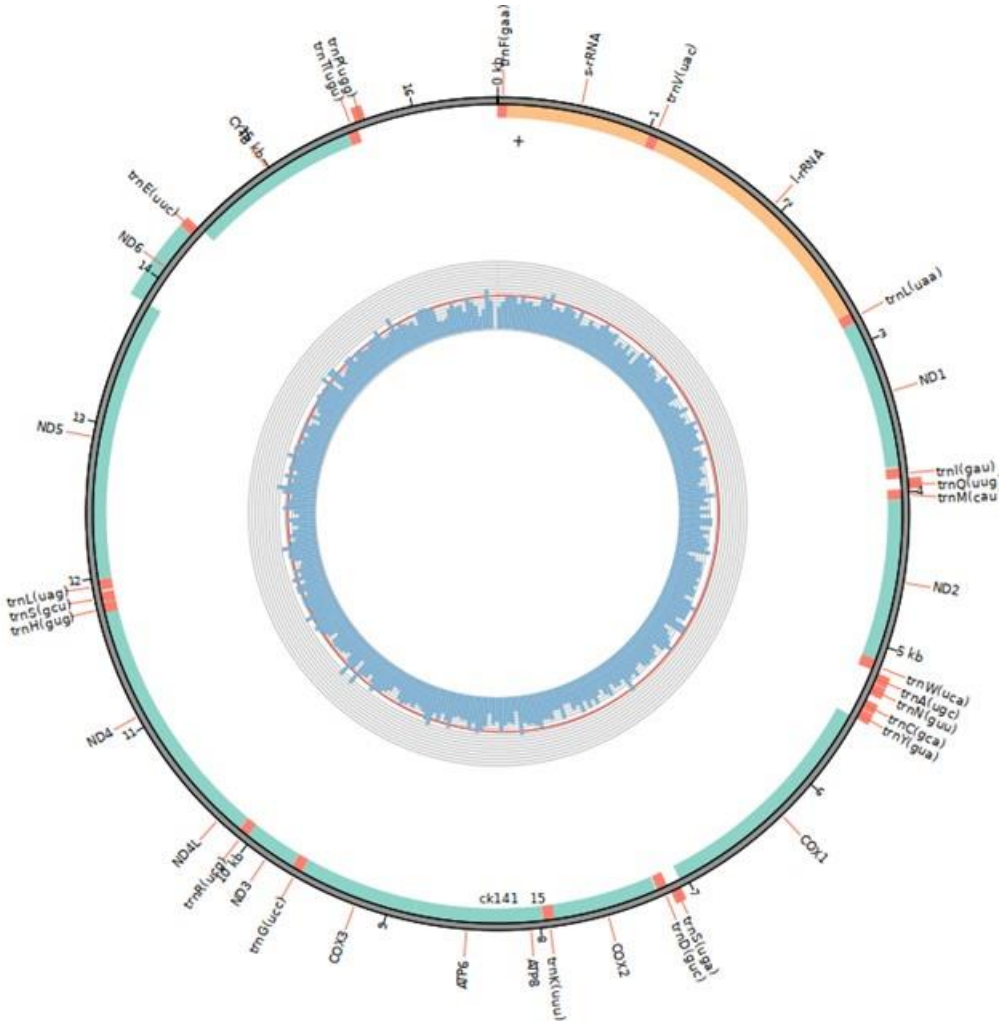


Figura 6. Organización circular del genoma mitocondrial de *Skiffia lermae*

6.2.3 Ensamble y Anotación del Genoma Completo

Las estadísticas generales del ensamble se encuentran en el Cuadro 7. Estos valores fueron obtenidos con *Quast*, después de realizar un pulido con *Pilon*. El valor de BUSCO al correr con *Augustus* arrojó que el genoma tiene un 35.2% de completitud, por lo que todavía se considera un ensamble *draft*. De los 841,715 transcritos obtenidos del ensamble de *S. lermae*, se encontraron 449,522 en el genoma completo usando *BLAST*.

Cuadro 7. Estadísticas del ensamble del genoma completo de *Skiffia lermae*

Estadística	Valor
Contigs totales	8150
Contig más largo (kb)	264.25
Contig más corto (kb)	8.5
Contig N25 (kb)	78.7
Contig N50 (kb)	54.7
Genes codificantes	14, 680
Longitud total (Mb)	400.69
Lecturas mapeadas al genoma	79.91%

Por otro lado, la longitud de 400 Mb resulta corta con respecto a otros peces cyprinodontiformes, ya que los genomas completos conocidos son de 597 Mb para *Poeciliopsis turrubarensis*, 621 Mb para *P. retropinna* (van Kruistum *et al*., 2020), 1.07 Gb para *Fundulus xenicus*, 1.08 Gb para *F. nottii*, 1.16 Gb para *F. catenatus* y 1.19 Gb para *F. olivaceus* (Johnson *et al*., 2020). Adicional a esto, los 14,680 genes que codifican para proteínas también están por debajo de lo registrado en la literatura, los cuales comienzan en 19,884 en *Protosalanx hyalocranius* (Liu *et al*., 2017), 21,562 en *Pelteobagrus fulvidraco* (Zhang *et al*., 2018), 22,067 en *Thamnaconus septentrionalis* (Bian *et al*., 2019b), 25 mil en *P. retropinna* (van Kruistum *et al*., 2020) y 31,120 en *Neoceratodus forsteri* (Meyer *et al*., 2021). En conjunto todos los valores son consistentes e indican que el genoma no está completo, ya que sólo se cuenta con alrededor de un tercio de la longitud total estimada.

Sin embargo, el valor de 54.7 kb de contig N50 se encuentra dentro de los valores reportados para genomas de peces con ensamble híbrido, los cuales van desde 20 kb para *Mola mola* (Pan *et al*., 2016), 26.5 kb para *Cynoglossus semilaevis*, 64 kb para *Larimichthys crocea* (Ao *et al*., 2015) y un valor elevado de 4.3 Mb para *Scophthalmus maximus* (Figueras *et al*., 2016). Esto indica que,

pese a no tener un genoma completo como tal, es un *draft* de buena calidad. Además, el número de 8,150 contigs es reducido para el tamaño del ensamble. Los valores reportados se encuentran comúnmente en el intervalo de 36 mil a 106 mil (Kadobianskyi *et al*., 2019; Bian *et al*., 2019), por lo que se construyó un genoma altamente contiguo gracias al método de ensamble utilizado.

6.3 Ensamble de Transcriptomas

La cantidad de lecturas de ARN obtenidas y otros datos relevantes proporcionados por *fastqc* se encuentran descritos en el Cuadro 8.

Cuadro 8. Datos de las lecturas crudas de tejidos de goodeinos obtenidos por fastqc

Organismo/tejido	Cantidad de lecturas pareadas	Porcentaje GC	Calidad promedio de las lecturas
S. lermae – bazo y corazón	27,575,846	44	36
S. lermae - branquia	44,011,256	47	40
S. lermae - cerebro	28,039,568	45	36
S. lermae - intestino	43,126,102	47	40
S. lermae - gónada	33,240,227	48	40
S. lermae - hígado	31,177,864	47	36
S. lermae - larva	30,076,159	47	36
S. bilineata - hígado	44,531,868	47	40
S. francesae - hígado	46,440,668	49	40
S. multipunctata - hígado	38,506,261	47	40
G. viviparus - hígado	25,463,998	48	36
Tylodelphys sp. (parásito)	34,773,614	44	40

Tras realizar el ensamble con *Trinity*, los resultados producidos y los valores arrojados por *BUSCO* sobre los genes ortólogos de una sola copia se encuentran registrados en el Cuadro 7. La longitud

de los ensambles concuerda con lo reportado en transcriptomas de hígado de peces, de 25 Mb a 507 Mb (Machado *et al*., 2018). La cantidad de transcritos y genes está ligeramente por debajo de los valores reportados en peces. Para genes van desde 20 mil hasta 238 mil y los transcritos desde 66 mil hasta 392 mil (Liu *et al*., 2020; Richards *et al*., 2018; Liu *et al*., 2018). Sobre el porcentaje de lecturas que mapean a los ensambles, calculado con *bowtie2*, en todos los casos hubo un valor mayor al 80% de mapeo. Los valores arrojados por BUSCO sobre los genes ortólogos de una sola copia son consistentes a lo encontrado de acuerdo a la longitud, cantidad de transcritos y de contigs en los ensambles. El ensamble más completo es el de *S. lermae* con 742 Mb de longitud y 841 mil transcritos. El resto de los ensambles no superan los 200 mil transcritos, por lo que se consideran en una categoría de ensambles *draft*.

Cuadro 9. Características generales de los ensambles iniciales de hígado de goodeinos

Especie	Genes	Transcritos	Porcentaje GC	Contig N50 (kb)	Longitud (Mb)	BUSCO completeness (%) Actinopterygii	Cyprinodontiformes
S. lermae	654,728	841,705	42.31	1.73	742.4	96.6	90.1
S. bilineata	152,175	197,463	43.69	2.09	178.4	74.5	56.5
S. francesae	116,616	150,312	44	2.14	141.3	71.4	52.4
S. multipunctata	108,564	142,608	43.96	2.21	141.65	73.8	55
G. viviparus	65,958	77,326	45.73	1.40	60	52.8	36.4
Tylodelphys sp. (parásito)	96,181	117,410	43.21	1.24	86.3	82.3*	

*Metazoa

6.4 Cálculo de Tasas de Sustitución

Los resultados de los modelos y el valor de ω se encuentran en el Cuadro 10. Se encontraron valores tanto de procesos de presión selectiva diversificante (ω>1), como valores neutrales (ω=1). Los primeros elementos analizados pertenecen a la superfamilia citocromo P450: citocromo P450 (CYP450), citocromo P450 2N2 (CYP2N2), citocromo P450, familia 51 (CYP51) y citocromo P450 2K1 (CYP2K). Un gran número de sus enzimas metaboliza contaminantes ambientales, interviniendo en la detoxificación; su actividad se ve afectada por estrés ambiental por lo que se han utilizado como biomarcadores de contaminación acuática (Andersson y Förlin, 1992; Uno *et*

al., 2012; Široká y Drastichova, 2014). Se puede observar que hay dos elementos con selección positiva en este grupo; P450(ω=7.6) y CYP51(ω=11.13). Estas moléculas se consideran inestables y codifican para enzimas que detoxifican compuestos xenobióticos (Thomas, 2007). También se han identificado previamente con selección positiva en humanos y ratones (Voight *et al*., 2006; Büntge, 2010).

En segunda instancia se muestran las Heat Shock Proteins: proteína de choque térmico 90-alfa (HSP90A) y proteína de choque térmico beta-8 (HSPB8). Éstas son reconocidas por expresarse ante una gran variedad de estresores bióticos y abióticos. Se consideran también un mecanismo de defensa que se activa posterior al daño celular (Iwama *et al*., 1998; Beere, 2005). Ambas secuencias analizadas en este trabajo, HSP90A(ω=1.3) y HSPB8(ω=11), mostraron tener sitios seleccionados positivamente, la mayoría encontrados en sitios codificantes para serina. Estos resultados ya se han observado en nemátodos (Him *et al*., 2009), en áfidos (Fares *et al*., 2002) y en cabras (Gade *et al*., 2010).

Asimismo, aparecen elementos asociados al metabolismo de minerales: transportador de Zinc 5 (ZNT5), transportador de cobre (CTR1), ATPasa B transportadora de cobre (ATP7B), transportador de serotonina sodio-dependiente (SERT) y elementos que responden ante la exposición a metales a concentraciones nocivas: factor nuclear (eritroide-derivado) 2 (NRF2) y proteína de desacoplamiento 2 (UCP2) (Wang *et al*., 2017; Burkina *et al*., 2015; Shaw *et al*., 2019). En esta categoría hubo dos secuencias que mostraron selección positiva en sus sitios; ATP7B(ω=12.7) y SERT(ω=24.3). Las ATPasas transportadoras de cobre se consideran evolutivamente estables de acuerdo a Gupta y Lutsenko, (2012) y no existen estudios recientes que demuestren presión selectiva en estas secuencias, por otro lado, se documentó divergencia genética en SERT en un estudio realizado en aves *Turdus merula* (Mueller *et al*., 2013).

A continuación, se encuentra la secuencia de ADN metiltransferasa 1 (DNMT1), asociada a cambios epigenéticos y a baja expresión en condiciones ambientales con compuestos xenobióticos (Laing *et al*., 2017). La DNMT1(ω=1) se encuentra en un estado de selección neutral. En la literatura se ha demostrado que las DNMTs tienen capacidad de cambiar por evolución en vertebrados, para adaptar la maquinaria de metilación (Álvarez-Ponce *et al*., 2018), pero en este caso no se observaron cambios significativos.

Por otro lado, se muestran las enzimas 6-fosfofructo-2-quinasa/fructosa-2,6-bifosfatasa 3 (PIK3R1) y la reguladora de p53 (MDM4). Ambas asociadas a cambios en el desarrollo y la

reproducción, además de responder ante la exposición a Contaminantes Orgánicos Persistentes (COP) Nourizadeh *et al*., 2009). PIK3R1(ω=5.7) mostró señales de selección positiva. No existen estudios que hayan observado selección en este gen, no obstante, se ha documentado la presencia de mutaciones en humanos en este gen están asociadas a incrementar la propensión a cáncer de tiroides (Murugan *et al*., 2011) y al síndrome SHORT (Bárcena *et al.*, 2014).

Se presentan genes fuertemente relacionados a estrés oxidativo, daño al ADN y apoptosis en el hígado de peces: reductasa glutatión-disulfuro (GSR) y NADPH oxidasa 1 (NOX1) (Awasthi *et al*., 2018). De estas, NOX1(ω=7.6) mostró evidencia de selección positiva en su secuencia. No hay otros estudios que hayan observado esto; sólo se ha documentado que la familia de las NADPH oxidasas ha evolucionado muy poco con el paso del tiempo. Estos genes son regulados por subunidades y las pequeñas mutaciones caracterizadas están correlacionadas a cambios en regiones clave con funciones catalíticas o regulatorias (Sumimoto, 2008; Kawahara *et al*., 2007).

Por último, se muestran secuencias correspondientes a proteínas relacionadas con respuesta a farmacéuticos que funcionan de la mano con la superfamilia Citocromo P450: el receptor activado por proliferador peroxisomal alfa (PPAR), el receptor X de pregnano (PXR) y el receptor de estrógenos beta 1 (ERB1) (Burkina *et al.*, 2015). En esta categoría se encontró a PPAR(ω=14.8) y PXR(ω=18) con signos de selección positiva. Para el caso de PPAR existe literatura que indica que responde a presión selectiva y ésta es esencial para diversificar sus funciones (Zhou *et al*., 2015). Además, la ruta de PPAR se encuentra sujeta a selección positiva en ganado (Zhao *et al*., 2015). Por otro lado, se ha demostrado que PXR está sometido a presión selectiva positiva en pruebas realizadas con nueve especies, desde peces y animales domésticos hasta humanos. Asimismo, se demostró que la selección natural ha favorecido la divergencia en sus secuencias, incrementando las diferencias a dos escalas; entre especies y entre ligandos (Krasowski *et al*., 2005; Reschly y Krasowski, 2006).

Cuadro 10. Resultados del análisis de tasas de mutaciones de ortólogos de goodeinos

Ortólogo	ω	Longitud (bases)	*p* M0 vs M3	*p* M7 vs M8
CYP450	7.6	505	0.0029	0.0098
CYP450_2	12.2	498	0.0270	0.3621
CYP51	11.13	500	0.0258	0.0344
CYP2K	1	522	1	0.9870

HSP90A	1.3	724	5E-09	1E-04
HSPB8	11	222	0.0001	0.0001
ZNT5	1	769	1	0.9999
CTR1	1	530	1	0.99
ATP7B	12.7	891	0	1.3E-06
SERT	24.3	674	3.7E-06	8.5E-06
NRF2	1	512	1	0.9999
UCP2	1	566	1	0.9999
DNMT1	1	1500	1	0.9977
PIK3R1	5.7	719	5E-05	6E-04
MDM4	3.4	364	0.37445	0.33557
GSR	1.9	621	0.5732	0.5904
NOX1	7.6	492	0.0029	0.0098
PPAR	14.8	313	0.0002	0.0046
PXR	18	570	0	0
ERB1	6	673	0.2965	0.0168

En resumen, los procesos que mostraron genes con selección positiva se relacionan con detoxificación de xenobióticos (CYP450, CYP51), respuesta a estrés (HSP90A, HSPB8), metabolismo de minerales (ATP7B, SERT), efectos de desarrollo y reproductivos (PIK3R1), respuesta a estrés oxidativo (NOX1) y a farmacéuticos (PPAR, PXR, ERB1). Esto indica que hay presión de selección actuando sobre ellas, por lo que estos cambios favorecen la adaptación fenotípica. Es muy probable que esta presión de selección se asocie a la prolongada exposición a compuestos nocivos en los cuerpos de agua y el resultado divergente en definitiva favorece ciertas mutaciones que permiten su permanencia a pesar de las amenazas derivadas de actividades antropogénicas.

7. CONCLUSIONES

En cuanto a los recursos genómicos de *Skiffia lermae* se obtuvo el mitogenoma completo y con circularidad, el primero para este género de goodeidos. Por otro lado, se logró ensamblar un genoma de buena calidad, buena contigüidad, pero con una longitud corta y un valor de completitud bajo (<40%), mientras que el transcriptoma con siete tejidos prácticamente está completo (>90%). En adición, se produjeron ensambles de transcriptoma de hígado de cuatro especies de goodeidos, una contribución parcial ya que al ser sólo de un tejido los transcriptomas están incompletos; no obstante, esta información es crucial para basar investigaciones próximas. Por último, el análisis de la tasa de mutaciones d_N/d_s nos ayudó a revelar ciertas regiones codificantes con evolución acelerada ($\omega>1$), pero con implicaciones documentadas sólo para algunos genes, ya que en otros no hay referentes. De cualquier manera, se evidenció que hay una presión selectiva diversificante afectando procesos relacionados con la interacción de los organismos y su entorno, que fomenta la adaptación fenotípica.

8. RECOMENDACIONES

Es recomendable incluir una mayor porción de secuencias largas de Nanopore o PacBio para verdaderamente enriquecer el ensamble y es necesario incrementar la cobertura por parte de esas lecturas para lograr obtener un ensamble lo más cercano posible a completo. Se recomienda agregar a este estudio otros géneros de goodeidos para poder profundizar en las adaptaciones evolutivas que puedan haber ocurrido a nivel familia.

9. REFERENCIAS

Álvarez-Ponce, D., Torres-Sánchez, M., Feyertag, F., Kulkarni, A., & Nappi, T. (2018). Molecular evolution of DNMT1 in vertebrates: duplications in marsupials followed by positive selection. PLoS One. 13(4).

Andrews S. (2010). FastQC: a quality control tool for high throughput sequence data. Disponible en: http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Andersson, T., & Förlin, L. (1992). Regulation of the cytochrome P450 enzyme system in fish. Aquatic Toxicology. 24(1-2):1-19.

Ao, J., Mu, Y., Xiang, L. X., Fan, D., Feng, M., Zhang, S., ... & Nie, L. (2015). Genome sequencing of the perciform fish *Larimichthys crocea* provides insights into molecular and genetic mechanisms of stress adaptation. PLoS Genet. 11(4).

Aparicio, S., Chapman, J., Stupka, E., Putnam, N., Chia, J. M., Dehal, P., ... & Gelpke, M. D. S. (2002). Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. Science. 297(5585):1301-1310.

Austin, C. M., Tan, M. H., Croft, L. J., Hammer, M. P., & Gan, H. M. (2015). Whole genome sequencing of the Asian arowana (*Scleropages formosus*) provides insights into the evolution of ray-finned fishes. Genome biology and evolution. 7(10):2885-2895.

Austin, C. M., Tan, M. H., Harrisson, K. A., Lee, Y. P., Croft, L. J., Sunnucks, P., ... & Gan, H. M. (2017). De novo genome assembly and annotation of Australia's largest freshwater fish, the Murray cod (*Maccullochella peelii*), from Illumina and Nanopore sequencing read. GigaScience. 6(8).

Awasthi, Y., Ratn, A., Prasad, R., Kumar, M., & Trivedi, S. P. (2018). An in vivo analysis of Cr6+ induced biochemical, genotoxicological and transcriptional profiling of genes related to oxidative stress, DNA damage and apoptosis in liver of fish, *Channa punctatus* (Bloch, 1793). Aquatic Toxicology. 200:158-167.

Ayuntamiento de Morelia. (2009). Manantial “La Mintzita”. Michoacán, México: Morelia, gobierno municipal. Recuperado de http://archivohistorico.morelia.gob.mx/index.php/micrositio-areas-naturales-protegidas/areas-naturales-protegidas-manantial-la-mintzita

Babraham bioinformatics. FastQC: A Quality Control Tool for High Throughput Sequence Data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (2015), "FastQC," https://qubeshub.org/resources/fastqc

Bárcena, C., Quesada, V., De Sandre-Giovannoli, A., Puente, D. A., Fernández-Toral, J., Sigaudy, S., ... & López-Otín, C. (2014). Exome sequencing identifies a novel mutation in PIK3R1 as the cause of SHORT syndrome. BMC medical genetics. 15(1):1-6.

Basu, N., Todgham, A. E., Ackerman, P. A., Bibeau, M. R., Nakano, K., Schulte, P. M., & Iwama, G. K. (2002). Heat shock protein genes and their functional significance in fish. Gene. 295(2):173-183.

Beere, H. M. (2005). Death versus survival: functional interaction between the apoptotic and stress-inducible heat shock protein pathways. The Journal of Clinical Investigation.115(10):2633-2639.

Bemanian, V., Male, R., & Goksøyr, A. (2004). The aryl hydrocarbon receptor-mediated disruption of vitellogenin synthesis in the fish liver: Cross-talk between AHR-and ERα-signalling pathways. Comparative Hepatology. 3(1):1-14.

Bian, C., Li, J., Lin, X., Chen, X., Yi, Y., You, X., ... & Shi, Q. (2019a). Whole genome sequencing of the blue tilapia (*Oreochromis aureus*) provides a valuable genetic resource for biomedical research on tilapias. Marine Drugs. 17(7):386.

Bian, L., Li, F., Ge, J., Wang, P., Chang, Q., Zhang, S., ... & Li, X. (2019b). Chromosome-level genome assembly of the greenfin horse-faced filefish (*Thamnaconus septentrionalis*) using Oxford Nanopore PromethION sequencing and Hi-C technology. Molecular Ecology Resources.

Biswas, K., Chakraborty, S., Podder, S., & Ghosh, T. C. (2016). Insights into the dN/dS ratio heterogeneity between brain specific genes and widely expressed genes in species of different complexity. Genomics. 108(1):11-17.

Branton, D., Deamer, D. W., Marziali, A., Bayley, H., Benner, S. A., Butler, T., ... & Jovanovich, S. B. (2010). The potential and challenges of nanopore sequencing. In Nanoscience and technology: A collection of reviews from Nature Journals. 261-268 pp.

Brandies, P., Peel, E., Hogg, C. J., & Belov, K. (2019). The value of reference genomes in the conservation of threatened species. Genes. 10(11):846.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina Sequence Data. Bioinformatics.

Booker, T. R., Jackson, B. C., & Keightley, P. D. (2017). Detecting positive selection in the genome. BMC biology. 15(1):98.

Büntge, A. (2010). Tracing signatures of positive selection in natural populations of the house mouse (Tesis doctoral, Christian-Albrechts-Universität Kiel).

Burkina, V., Zlabek, V., & Zamaratskaia, G. (2015). Effects of pharmaceuticals present in aquatic environment on Phase I metabolism in fish. Environmental Toxicology and Pharmacology. 40(2): 430-444.

Caizergues, A. E., Grégoire, A., & Charmantier, A. (2018). Urban versus forest ecotypes are not explained by divergent reproductive selection. Proceedings of the Royal Society B: Biological Sciences. 285(1882).

Carlson, S. M., Cunningham, C. J., & Westley, P. A. (2014). Evolutionary rescue in a changing world. Trends in Ecology & Evolution. 29(9):521-530.

Chevreux, B., Wetter, T. and Suhai, S. (1999): Genome Sequence Assembly Using Trace Signals and Additional Sequence Information. Computer Science and Biology: Proceedings of the German Conference on Bioinformatics (GCB). 1:45-56.

Chen, S., Zhang, G., Shao, C., Huang, Q., Liu, G., Zhang, P., ... & Hong, Y. (2014). Whole-genome

sequence of a flatfish provides insights into ZW sex chromosome evolution and adaptation to a benthic lifestyle. Nature Genetics. 46(3):253-260.

Chen, W., Bian, C., You, X., Li, J., Ye, L., Wen, Z., ... & Gu, R. (2019). Genome Sequencing of the Japanese Eel (*Anguilla japonica*) for Comparative Genomic Studies on tbx4 and a tbx4 Gene Cluster in Teleost Fishes. Marine Drugs. 17(7):426.

CONACYT. (2021). Áreas naturales protegidas del estado de Michoacán. Michoacán, México: gob.mx. Recuperado de www.conacyt.gob.mx/cibiogem/index.php/anpl/michoacan

Cummings, P. J., Olszewicz, J., & Obom, K. M. (2017). Nanopore DNA sequencing for metagenomic soil analysis. Journal of Visualized Experiments. 130.

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., ... & Li, H. (2021). Twelve years of SAMtools and BCFtools. Gigascience. 10(2).

Deschamps, S., Zhang, Y., Llaca, V., Ye, L., Sanyal, A., King, M., ... & Lin, H. (2018). A chromosome-scale assembly of the sorghum genome using nanopore sequencing and optical mapping. Nature Communications. 9(1):1-10.

De La Vega-Salazar, M. Y., Avila-Luna, E., & Macías-Garcia, C. (2003). Ecological evaluation of local extinction: the case of two genera of endemic Mexican fish, Zoogoneticus and Skiffia. Biodiversity & Conservation. 12(10):2043-2056.

Doadrio, I., & Domínguez, O. (2004). Phylogenetic relationships within the fish family Goodeidae based on cytochrome b sequence data. Molecular Phylogenetics and Evolution. 31(2):416-430.

Domínguez-Domínguez O. (2006). Viviparous Fishes. The Goodeids. The Viviparous Goodeid Fishes. 558-565 pp.

Domínguez-Domínguez, O., Doadrio, I., & Pérez-Ponce de León, G. (2006). Historical biogeography of some river basins in central Mexico evidenced by their goodeine freshwater fishes: a preliminary hypothesis using secondary Brooks parsimony analysis. Journal of Biogeography. 33(8):1437-1447.

Domínguez-Domínguez, O., Mercado-Silva, N., Lyons, J., & Grier, H. J. (2005). The viviparous goodeid fishes. Viviparous fishes, MC Uribe and HJ Grier (eds.). New Life Publications, Homestead, Florida. 525-569 pp.

Domínguez-Domínguez, O., Pedraza-Lara, C., Gurrola-Sánchez, N., Pérez-Rodríguez, R., Israde-Alcántara, I., Garduño-Monroy, V. H., ... & Brooks, D. R. (2010). Historical biogeography of the Goodeinae (Cyprinodontiforms). Viviparous Fishes. 2:13-30.

Domínguez-Domínguez, O., & Pérez-Ponce de León, G. (2007). Los goodeidos, peces endémicos del centro de México. Biodiversitas. 75:12-15.

Domínguez-Domínguez, O., Zambrano, L., Escalera-Vázquez, L. H., Pérez-Rodríguez, R., & Pérez-Ponce de León, G. (2008). Cambio en la distribución de goodeidos (Osteichthyes: Cyprinodontiformes: Goodeidae) en cuencas hidrológicas del centro de México. Revista Mexicana de Biodiversidad. 79(2):501-512.

Ebler, J., Haukness, M., Pesout, T., Marschall, T., & Paten, B. (2018). Haplotype-aware genotyping from noisy long reads. bioRxiv

Eddy, S. (1992). HMMER user's guide. Department of Genetics, Washington University School of Medicine. 2(1):13.

Elmer, K. R., Fan, S., Gunter, H. M., Jones, J. C., Boekhoff, S., Kuraku, S., & Meyer, A. (2010). Rapid evolution and selection inferred from the transcriptomes of sympatric crater lake cichlid fishes. Molecular Ecology. 19:197-211.

Fan, S., Elmer, K. R., & Meyer, A. (2012). Genomics of adaptation and speciation in cichlid fishes: recent advances and analyses in African and Neotropical lineages. Philosophical Transactions of the Royal Society B: Biological Sciences. 367(1587):385-394.

Fares, M. A., Barrio, E., Sabater-Munoz, B., & Moya, A. (2002). The evolution of the heat-shock protein GroEL from Buchnera, the primary endosymbiont of aphids, is governed by positive selection. Molecular Biology and Evolution. 19(7):1162-1170.

Figueras, A., Robledo, D., Corvelo, A., Hermida, M., Pereiro, P., Rubiolo, J. A., ... & Gut, I. G. (2016). Whole genome sequencing of turbot (*Scophthalmus maximus*; Pleuronectiformes): a fish adapted to demersal life. DNA Research. 23(3):181-192.

Ford, M. J. (2001). Molecular evolution of transferrin: evidence for positive selection in salmonids. Molecular Biology and Evolution. 18(4):639-647.

Foster, K. L., & Piller, K. R. (2018). Disentangling the drivers of diversification in an imperiled group of freshwater fishes (Cyprinodontiformes: Goodeidae). BMC evolutionary biology. 18(1):116.

Gade, N., Mahapatra, R. K., Sonawane, A., Singh, V. K., Doreswamy, R., & Saini, M. (2010). Molecular characterization of heat shock protein 70-1 gene of goat (*Capra hircus*). Molecular Biology International.

Gao, F., Chen, C., Arab, D.A., Du, Z., He, Y., Ho, S.Y.W., 2019. EasyCodeML: A visual tool for analysis of selection using CodeML.Ecology and Evolution. 9:3891-3898.

George, S., Pankhurst, L., Hubbard, A., Votintseva, A., Stoesser, N., Sheppard, A. E., ... & Phan, H. T. (2017). Resolving plasmid structures in Enterobacteriaceae using the MinION nanopore sequencer: assessment of MinION and MinION/Illumina hybrid data assembly approaches. Microbial Genomics. 3(8).

Giani, A. M., Gallo, G. R., Gianfranceschi, L., & Formenti, G. (2020). Long walk to genomics: History and current approaches to genome sequencing and assembly. Computational and Structural Biotechnology Journal. 18:9-19.

Giordano, F., Aigrain, L., Quail, M. A., Coupland, P., Bonfield, J. K., Davies, R. M., ... & Ning, Z. (2017). De novo yeast genome assemblies from MinION, PacBio and MiSeq platforms. Scientific Reports. 7(1):1-10.

Gómez Navarrete, R. A. (2013). Carga parasitaria y su posible efecto en algunos parámetros morfométricos de *Girardinichthys multiradiatus* (Cyprinidontiformes: goodeidae) (Tesis de maestría). Universidad Autónoma del Estado de México.

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., ... & Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature Biotechnology. 29(7):644-652.

Green, M. R., & Sambrook, J. (2017). Isolation of high-molecular-weight DNA using organic solvents. Cold Spring Harbor Protocols. 2017(4).

Gregory, S. (2005). Contig Assembly. Encyclopedia of Life Sciences.

Gupta, A., & Lutsenko, S. (2012). Evolution of copper transporting ATPases in eukaryotic organisms. Current Genomics. 13(2):124-133.

Hamm, C. (2014). Github.com: N50. Recuperado de https://github.com/butterflyology/N50

Hernández, M. A. H., Quispe, S. T. S., Campos, J. A. A., & Solera, A. S. (2013). Modelación integral de la gestión del sistema Zacapu y Pastor Ortiz en la Cuenca del Río Angulo (México).

Him, N. A., Gillan, V., Emes, R. D., Maitland, K., & Devaney, E. (2009). Hsp-90 and the biology of nematodes. BMC evolutionary biology. 9(1):1-13.

Hinchliff, C. E., Smith, S. A., Allman, J. F., Burleigh, J. G., Chaudhary, R., Coghill, L. M., ... & Cranston, K. A. (2015). Synthesis of phylogeny and taxonomy into a comprehensive tree of life. Proceedings of the National Academy of Sciences. 112(41):12764-12769.

Hudson, R. R., Bailey, K., Skarecky, D., Kwiatowski, J., & Ayala, F. J. (1994). Evidence for positive selection in the superoxide dismutase (Sod) region of *Drosophila melanogaster*. Genetics. 136(4):1329-1340.

Iwama, G. K., Thomas, P. T., Forsyth, R. B., & Vijayan, M. M. (1998). Heat shock protein expression in fish. Reviews in Fish Biology and Fisheries. 8(1):35-56.

Jeffares, D. C., Tomiczek, B., Sojo, V., & dos Reis, M. (2015). A beginners guide to estimating the non-synonymous to synonymous rate ratio of all protein-coding genes in a genome. In Parasite Genomics Protocols. Humana Press, New York, NY. 65-90 pp.

Jiang, W., Lv, Y., Cheng, L., Yang, K., Bian, C., Wang, X., ... & Yang, J. (2019). Whole-Genome Sequencing of the Giant Devil Catfish, *Bagarius yarrelli*. Genome Biology and Evolution. 11(8):2071-2077.

Johnson, L. K., Sahasrabudhe, R., Gill, J. A., Roach, J. L., Froenicke, L., Brown, C. T., & Whitehead, A. (2020). Draft genome assemblies using sequencing reads from Oxford Nanopore Technology and Illumina platforms for four species of North American *Fundulus killifish*. GigaScience. 9(6).

Kadobianskyi, M., Schulze, L., Schuelke, M., & Judkewitz, B. (2019). Hybrid genome assembly and annotation of Danionella translucida. Scientific Data. 6(1):1-7.

Kawahara, T., Quinn, M. T., & Lambeth, J. D. (2007). Molecular evolution of the reactive oxygen-generating NADPH oxidase (Nox/Duox) family of enzymes. BMC evolutionary biology. 7(1):1-21.

Kayhan, F. E., & Duman, B. S. (2010). Heat shock protein genes in fish. Turkish Journal of Fisheries and Aquatic Sciences. 10(2).

Kirubakaran, T. G., Andersen, Ø., Moser, M., Arnyasi, M., McGinnity, P., Lien, S., & Kent, M. (2020). A nanopore based chromosome-level assembly representing Atlantic cod from the Celtic Sea. G3: Genes, Genomes, Genetics.

Koeck, M. 2019. Skiffia lermae. The IUCN Red List of Threatened Species 2019:

e.T191714A2000019. https://dx.doi.org/10.2305/IUCN.UK.2019-2.RLTS.T191714A2000019. En. Downloaded on 29 January 2021.

Koren S, Walenz BP, Berlin K, Miller JR, Phillippy AM. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Research.

Krasowski, M. D., Yasuda, K., Hagey, L. R., & Schuetz, E. G. (2005). Evolutionary selection across the nuclear hormone receptor superfamily with a focus on the NR1I subfamily (vitamin D, pregnane X, and constitutive androstane receptors). Nuclear Receptor. 3(1):1-20.

van Kruistum, H., Guernsey, M. W., Baker, J. C., Kloet, S. L., Groenen, M. A., Pollux, B. J., & Megens, H. J. (2020). The genomes of the livebearing fish species *Poeciliopsis retropinna* and *Poeciliopsis turrubarensis* reflect their different reproductive strategies. Molecular Biology and Evolution. 37(5):1376-1386.

Kryazhimskiy, S., & Plotkin, J. B. (2008). The population genetics of dN/dS. PLoS Genetics. 4(12).

Lagesen, K., Hallin, P., Rødland, E. A., Stærfeldt, H. H., Rognes, T., & Ussery, D. W. (2007). RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Research. 35(9):3100-3108.

Laing, L. V., Viana, J., Dempster, E. L., Trznadel, M., Trunkfield, L. A., Uren Webster, T. M., ... & Santos, E. M. (2016). Bisphenol A causes reproductive toxicity, decreases dnmt1 transcription, and reduces global DNA methylation in breeding zebrafish (*Danio rerio*). Epigenetics. 11(7):526-538.

Langmead B, Salzberg S. (2012). Fast gapped-read alignment with Bowtie 2. Nature Methods. 9:357-359.

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., ... & Higgins, D. G. (2007). Clustal W and Clustal X version 2.0. Bioinformatics. 23(21):2947-2948.

Larsen, P. A., Harris, R. A., Liu, Y., Murali, S. C., Campbell, C. R., Brown, A. D., ... & Worley, K. C. (2017). Hybrid de novo genome assembly and centromere characterization of the gray mouse lemur (*Microcebus murinus*). BMC biology. 15(1):1-17.

Leger, A., & Leonardi, T. (2019). pycoQC, interactive quality control for Oxford Nanopore Sequencing. Journal of Open Source Software. 4(34):1236.

Lewin, H. A., Robinson, G. E., Kress, W. J., Baker, W. J., Coddington, J., Crandall, K. A., ... & Zhang, G. (2018). Earth BioGenome Project: Sequencing life for the future of life. Proceedings of the National Academy of Sciences. 115(17):4325-4333.

Li, D., Liu, C. M., Luo, R., Sadakane, K., & Lam, T. W. (2015). MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics. 31(10):1674-1676.

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

Liu, H., Wu, S., Li, A., & Ruan, J. (2021). SMARTdenovo: A de novo assembler using long noisy reads. Gigabyte. 1:1-9.

Liu, K., Xu, D., Li, J., Bian, C., Duan, J., Zhou, Y., ... & Yu, H. (2017). Whole genome sequencing

of Chinese clearhead icefish, *Protosalanx hyalocranius*. GigaScience. 6(4).

Liu, L., Zhang, R., Wang, X., Zhu, H., & Tian, Z. (2020). Transcriptome analysis reveals molecular mechanisms responsive to acute cold stress in the tropical stenothermal fish tiger barb (*Puntius tetrazona*). BMC genomics. 21(1):1-14.

Liu, Q. N., Xin, Z. Z., Liu, Y., Zhang, D. Z., Jiang, S. H., Chai, X. Y., ... & Tang, B. P. (2018). De novo transcriptome assembly and analysis of differential gene expression following lipopolysaccharide challenge in *Pelteobagrus fulvidraco*. Fish & Shellfish Immunology. 73:84-91.

Llorente-Bousquets, J., y S. Ocegueda. (2008). Estado del conocimiento de la biota, en Capital natural de México, vol. I: Conocimiento actual de la biodiversidad. CONABIO, México, 283-322 pp.

Lyons, J., González-Hernández, G., Soto-Galera, E., & Guzmán-Arroyo, M. (1998). Decline of freshwater fishes and fisheries in selected drainages of West-Central Mexico. Fisheries. 23(4):10-18.

Lyons, J., Piller, K. R., Artigas-Azas, J. M., Dominguez-Dominguez, O., Gesundheit, P., Köck, M., ... & Findley, K. M. (2019). Distribution and current conservation status of the Mexican Goodeidae (Actinopterygii, Cyprinodontiformes). ZooKeys. 885:115.

Machado, A. M., Tørresen, O. K., Kabeya, N., Couto, A., Petersen, B., Felício, M., ... & C Castro, L. F. (2018). "Out of the Can": A draft genome assembly, liver transcriptome, and nutrigenomics of the European sardine, *Sardina pilchardus*. Genes. 9(10):485.

Magi, A., Semeraro, R., Mingrino, A., Giusti, B., & D'Aurizio, R. (2018). Nanopore sequencing data analysis: state of the art, applications and challenges. Briefings in Bioinformatics. 19(6):1256-1272.

Malmstrøm, M., Matschiner, M., Tørresen, O. K., Jakobsen, K. S., & Jentoft, S. (2017). Whole genome sequencing data and de novo draft assemblies for 66 teleost species. Scientific data. 4(1):1-13.

Manni, M., Berkeley, M. R., Seppey, M., Simao, F. A., & Zdobnov, E. M. (2021). BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. arXiv preprint.

Marshall, H. D., Coulson, M. W., & Carr, S. M. (2009). Near neutrality, rate heterogeneity, and linkage govern mitochondrial genome evolution in Atlantic cod (*Gadus morhua*) and other gadine fish. Molecular Biology and Evolution. 26(3):579-589.

Martínez-Aquino, A., Pérez-Rodríguez, R., Hernández-Mena, D. I., Garrido-Olvera, L., Aguilar-Aguilar, R., & Pérez-Ponce de León, G. (2012). Endohelminth parasites of seven goodein species (Cyprinodontiformes: Goodeidae) from Lake Zacapu, Michoacán, Central Mexico Plateau. Hidrobiológica. 22(1):89-93.

Meng, G., Li, Y., Yang, C., & Liu, S. (2018). MitoZ: A toolkit for mitochondrial genome assembly, annotation and visualization. bioRxiv. 489955.

Metzker, M. L. (2010). Sequencing technologies—the next generation. Nature Reviews Genetics, 11(1):31-46.

Meyer, A., Schloissnig, S., Franchini, P., Du, K., Woltering, J., Irisarri, I., ... & Schartl, M. (2021).

Giant lungfish genome elucidates the conquest of land by vertebrates. Nature. 1-9.

Miller, S. A., Dykes, D. D., & Polesky, H. F. R. N. (1988). A simple salting out procedure for extracting DNA from human nucleated cells. Nucleic Acids Research. 16(3):1215.

Moritz, C. (1994). Applications of mitochondrial DNA analysis in conservation: a critical review. Molecular Ecology. 3(4):401-411.

Moury, B., & Simon, V. (2011). dN/dS-based methods detect positive selection linked to trade-offs between different fitness traits in the coat protein of potato virus Y. Molecular Biology and Evolution. 28(9):2707-2717.

Mukhopadhyay, D., & Chattopadhyay, A. (2014). Induction of oxidative stress and related transcriptional effects of sodium fluoride in female zebrafish liver. Bulletin of Environmental Contamination and Toxicology. 93(1):64-70.

Murugan, A. K., Dong, J., Xie, J., & Xing, M. (2011). Uncommon GNAQ, MMP8, AKT3, EGFR, and PIK3R1 mutations in thyroid cancers. Endocrine Pathology. 22(2):97-102.

National Human Genome Research Institute (NIH). (2019). Transcriptoma. Recuperado de https://www.genome.gov/es/about-genomics/fact-sheets/Transcriptoma#

Nelson, J.S., (1994). Fishes of the world. Third edition. John Wiley & Sons, Inc., New York. 600 pp.

Nielsen, R. (1999). Changes in ds/dn in the HIV-1 env gene. Molecular Biology and Evolution, 16(5):711-714.

Nourizadeh-Lillabadi, R., Lyche, J. L., Almaas, C., Stavik, B., Moe, S. J., Aleksandersen, M., ... & Ropstad, E. (2009). Transcriptional regulation in liver and testis associated with developmental and reproductive effects in male zebrafish exposed to natural mixtures of persistent organic pollutants (POP). Journal of Toxicology and Environmental Health, Part A. 72(3-4):112-130.

Osterberg, J. S., Cammen, K. M., Schultz, T. F., Clark, B. W., & Di Giulio, R. T. (2018). Genome-wide scan reveals signatures of selection related to pollution adaptation in non-model estuarine Atlantic killifish (*Fundulus heteroclitus*). Aquatic Toxicology. 200:73-82.

Ownby, D. R., Newman, M. C., Mulvey, M., Vogelbein, W. K., Unger, M. A., & Arzayus, L. F. (2002). Fish (*Fundulus heteroclitus*) populations with different exposure histories differ in tolerance of creosote-contaminated sediments. Environmental Toxicology and Chemistry: An International Journal. 21(9):1897-1902.

Page, L. M., Espinosa-Pérez, H., Findley, L. T., Gilbert, C. R., Lea, R. N., Mandrak, N. E., & Mayden, R. L. (2013). New Seventh Edition of Common and Scientific Names of Fishes: Changes include capitalization of common names. Fisheries. 38(4):188-189.

Pan, H., Yu, H., Ravi, V., Li, C., Lee, A. P., Lian, M. M., ... & Zhang, G. (2016). The genome of the largest bony fish, ocean sunfish (*Mola mola*), provides insights into its fast growth rate. GigaScience. 5(1).

Pawlowski, J., Esling, P., Lejzerowicz, F., Cedhagen, T., & Wilding, T. A. (2014). Environmental monitoring through protist next-generation sequencing metabarcoding: Assessing the impact of fish farming on benthic foraminifera communities. Molecular Ecology Resources. 14(6):1129-1140.

Pierron, F., Bourret, V., St-Cyr, J., Campbell, P. G., Bernatchez, L., & Couture, P. (2009). Transcriptional responses to environmental metal exposure in wild yellow perch (*Perca flavescens*) collected in lakes with differing environmental metal concentrations (Cd, Cu, Ni). Ecotoxicology. 18(5):620-631.

Piola, R. F., & Johnston, E. L. (2006). Differential tolerance to metals among populations of the introduced bryozoan *Bugula neritina*. Marine Biology. 148(5):997-1010.

Ramon, M. (1998). Mitochondrial DNA: a tool for populational genetics studies.

RAMSAR. 2009. Ramsar.org: Ficha Informativa de los Humedales de Ramsar (FIR) - versión 2009-2012. Recuperado de https://rsis.ramsar.org/RISapp/files/RISrep/MX1919RIS.pdf

Rhie, A., McCarthy, S. A., Fedrigo, O., Damas, J., Formenti, G., Koren, S., ... & Jarvis, E. D. (2021). Towards complete and error-free genome assemblies of all vertebrate species. Nature. 592(7856):737-746.

Richards, D. J., Renaud, L., Agarwal, N., Starr Hazard, E., Hyde, J., & Hardiman, G. (2018). De Novo Hepatic Transcriptome Assembly and Systems Level Analysis of Three Species of Dietary Fish, *Sardinops sagax*, *Scomber japonicus*, and *Pleuronichthys verticalis*. Genes. 9(11):521.

Rivero, E. R., Neves, A. C., Silva-Valenzuela, M. G., Sousa, S. O., & Nunes, F. D. (2006). Simple salting-out method for DNA extraction from formalin-fixed, paraffin-embedded tissues. Pathology-Research and Practice. 202(7):523-529.

Rocha, E. P., Smith, J. M., Hurst, L. D., Holden, M. T., Cooper, J. E., Smith, N. H., & Feil, E. J. (2006). Comparisons of dN/dS are time dependent for closely related bacterial genomes. Journal of Theoretical Biology. 239(2):226-235.

Roubicek, D. A., & de Souza-Pinto, N. C. (2017). Mitochondria and mitochondrial DNA as relevant targets for environmental contaminants. Toxicology. 391:100-108.

Rueda-Jasso, R. A., De los Santos-Bailón, A., & Campos-Mendoza, A. (2017). Nitrite toxicity in juvenile Goodeinae fishes *Skiffia multipunctata* (Pellegrin, 1901) and *Goodea atripinnis* (Jordan, 1880). Journal of Applied Ichthyology. 33(2):300-305.

Sánchez-Nava, P., Salgado-Maldonado, G., Soto-Galera, E., & Cruz, B. J. (2004). Helminth parasites of *Girardinichthys multiradiatus* (Pisces: Goodeidae) in the upper Lerma River sub-basin, Mexico. Parasitology Research. 93(5):396-402.

Schraiber, J. G., Evans, S. N., & Slatkin, M. (2016). Bayesian inference of natural selection from allele frequency time series. Genetics. 203(1):493-511.

Secretaría de Medio Ambiente, Recursos Naturales y Pesca (SEMARNAP). 2000. paot.org.mx: Áreas naturales protegidas de México con decretos federales (1899-2000). Recuperado de http://www.paot.org.mx/centro/ine-semarnat/anp/AN01.pdf

Secretaría de Medio Ambiente y Recursos Naturales (SEMARNAT). 2019. Modificación del anexo normativo III, Lista de especies en riesgo de la Norma Oficial Mexicana NOM-059-SEMARNAT–2010, "Protección ambiental-Especies nativas de México de flora y fauna silvestres-Categorías de riesgo y especificaciones para su inclusión, exclusión o cambio-lista de especies en riesgo". Diario Oficial de la Federación (DOF), publicado el jueves 14 de noviembre de 2019, México, D.F.

Setiamarga, D. H., Miya, M., Yamanoue, Y., Mabuchi, K., Satoh, T. P., Inoue, J. G., & Nishida, M. (2008). Interrelationships of Atherinomorpha (medakas, flyingfishes, killifishes, silversides, and their relatives): the first evidence based on whole mitogenome sequences. Molecular Phylogenetics and Evolution. 49(2):598-605.

Shaw, P., Mondal, P., Bandyopadhyay, A., & Chattopadhyay, A. (2019). Environmentally relevant concentration of chromium activates Nrf2 and alters transcription of related XME genes in liver of zebrafish. Chemosphere. 214:35-46.

Shin, S. C., Ahn, D. H., Kim, S. J., Lee, H., Oh, T. J., Lee, J. E., & Park, H. (2013). Advantages of single-molecule real-time sequencing in high-GC content genomes. PloS one. 8(7).

Široká, Z., & Drastichova, J. (2004). Biochemical marker of aquatic environment contamination-cytochrome P450 in fish. A review. Acta Veterinaria Brno. 73(1):123-132.

Sugawara, T., Terai, Y., & Okada, N. (2002). Natural selection of the rhodopsin gene during the adaptive radiation of East African Great Lakes cichlid fishes. Molecular Biology and Evolution. 19(10):1807-1811.

Sumimoto, H. (2008). Structure, regulation and evolution of Nox-family NADPH oxidases that produce reactive oxygen species. The FEBS Journal. 275(13):3249-3277.

Supple, M. A., & Shapiro, B. (2018). Conservation of biodiversity in the genomics era. Genome Biology. 19(1):1-12.

Swanson, W. J., Yang, Z., Wolfner, M. F., & Aquadro, C. F. (2001). Positive Darwinian selection drives the evolution of several female reproductive proteins in mammals. Proceedings of the National Academy of Sciences. 98(5):2509-2514.

Thomas JH (2007) Rapid Birth–Death Evolution Specific to Xenobiotic Cytochrome P450 Genes in Vertebrates. PLoS Genet. 3(5):67.

Tinguely, S. M. (2015). *Xenotoca eiseni* (Cyprinodontiformes, Goodeidae) as a Potential New Model for Studies on Maternal Transfer of Environmental Contaminants.

Tyson, J. R., O'Neil, N. J., Jain, M., Olsen, H. E., Hieter, P., & Snutch, T. P. (2017). Whole genome sequencing and assembly of a *Caenorhabditis elegans* genome with complex genomic rearrangements using the MinION sequencing device. BioRxiv. 099143.

Uno, T., Ishizuka, M., & Itakura, T. (2012). Cytochrome P450 (CYP) in fish. Environmental Toxicology and Pharmacology. 34(1):1-13.

Valerio-García, R. C., Carbajal-Hernández, A. L., Martínez-Ruíz, E. B., Jarquín-Díaz, V. H., Haro-Pérez, C., & Martínez-Jerónimo, F. (2017). Exposure to silver nanoparticles produces oxidative stress and affects macromolecular and metabolic biomarkers in the goodeid fish *Chapalichthys pardalis*. Science of the Total Environment. 583:308-318.

Vega-López, A., Jiménez-Orozco, F. A., García-Latorre, E., & Domínguez-López, M. L. (2008). Oxidative stress response in an endangered goodeid fish (*Girardinichthys viviparus*) by exposure to water from its extant localities. Ecotoxicology and Environmental Safety. 71(1):94-103.

Vega-López, A., Jiménez-Orozco, F. A., Ramón-Gallegos, E., García-Latorre, E., & Domínguez-López, M. L. (2008). Estrogenic effects of polychlorinated biphenyls and relation to cytochrome P4501A activity in the endangered goodeid fish *Ameca splendens*.

Environmental Toxicology and Chemistry: An International Journal. 27(4):963-969.

Vega-López, A., Ortiz-Ordóñez, E., Uría-Galicia, E., Mendoza-Santana, E. L., Hernández-Cornejo, R., Atondo-Mexia, R., ... & Domínguez-López, M. L. (2007). The role of vitellogenin during gestation of Girardinichthys viviparus and *Ameca splendens*; two goodeid fish with matrotrophic viviparity. Comparative Biochemistry and Physiology Part A: Molecular & Integrative Physiology, 147(3), 731-742.

Vega-López, A., Ramón-Gallegos, E., Galar-Martínez, M., Jiménez-Orozco, F. A., García-Latorre, E., & Domínguez-López, M. L. (2007). Estrogenic, anti-estrogenic and cytotoxic effects elicited by water from the type localities of the endangered goodeid fish *Girardinichthys viviparus*. Comparative Biochemistry and Physiology Part C: Toxicology & Pharmacology. 145(3):394-403.

Vialle, R. A., de Souza, J. E. S., Lopes, K. D. P., Teixeira, D. G., Alves Sobrinho, P. D. A., Ribeiro-dos-Santos, A. M., ... & Hamoy, I. G. (2018). Whole genome sequencing of the pirarucu (*Arapaima gigas*) supports independent emergence of major teleost clades. Genome Biology and Evolution. 10(9):2366-2379.

Voight, B. F., Kudaravalli, S., Wen, X., & Pritchard, J. K. (2006). A map of recent positive selection in the human genome. PLoS biology. 4(3):72.

Walker, B. H. (1992). Biodiversity and ecological redundancy. Conservation Biology. 6(1):18-23.

Wang, C. C., Si, L. F., Guo, S. N., & Zheng, J. L. (2019). Negative effects of acute cadmium on stress defense, immunity, and metal homeostasis in liver of zebrafish: the protective role of environmental zinc dpre-exposure. Chemosphere. 22:91-97.

Wang, D., Chen, X., Zhang, X., Li, J., Yi, Y., Bian, C., ... & You, X. (2019). Whole genome sequencing of the giant grouper (*Epinephelus lanceolatus*) and high-throughput screening of putative antimicrobial peptide genes. Marine Drugs. 17(9):503.

Waterston, Robert (2002). "On the Sequencing of the Human Genome". Proceedings of the National Academy of Sciences of the United States of America. 99(6):3712–3716.

Watson, J. D., & Crick, F. H. (1953). Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. Nature. 171(4356):737-738.

Webb, S. A., Graves, J. A., Macias-Garcia, C., Magurran, A. E., Foighil, D. O., & Ritchie, M. G. (2004). Molecular phylogeny of the livebearing Goodeidae (Cyprinodontiformes). Molecular Phylogenetics and Evolution. 30(3):527-544.

Wilson, D. J., Hernandez, R. D., Andolfatto, P., & Przeworski, M. (2011). A population genetics-phylogenetics approach to inferring natural selection in coding sequences. PLoS genetics. 7(12).

Xiao, J., Zhong, H., Liu, Z., Yu, F., Luo, Y., Gan, X., & Zhou, Y. (2015). Transcriptome analysis revealed positive selection of immune-related genes in tilapia. Fish & Shellfish Immunology. 44(1):60-65.

Yang, Z. (1998). Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Molecular Biology and Evolution. 15(5):568-573.

Yang, Z., & Bielawski, J. P. (2000). Statistical methods for detecting molecular adaptation. Trends in Ecology & Evolution. 15(12):496-503.

Yang, Z., & Nielsen, R. (2002). Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Molecular Biology and Evolution. 19(6):908-917.

Zhang, S., Li, J., Qin, Q., Liu, W., Bian, C., Yi, Y., ... & Liu, Y. (2018). Whole-genome sequencing of Chinese yellow catfish provides a valuable genetic resource for high-throughput identification of toxin genes. Toxins. 10(12):488.

Zhao, F., McParland, S., Kearney, F., Du, L., & Berry, D. P. (2015). Detection of selection signatures in dairy and beef cattle using high-density genomic information. Genetics Selection Evolution, 47(1):1-12.

Zhou, T., Yan, X., Wang, G., Liu, H., Gan, X., Zhang, T., ... & Li, L. (2015). Evolutionary pattern and regulation analysis to support why diversity functions existed within PPAR gene family members. BioMed Research International.

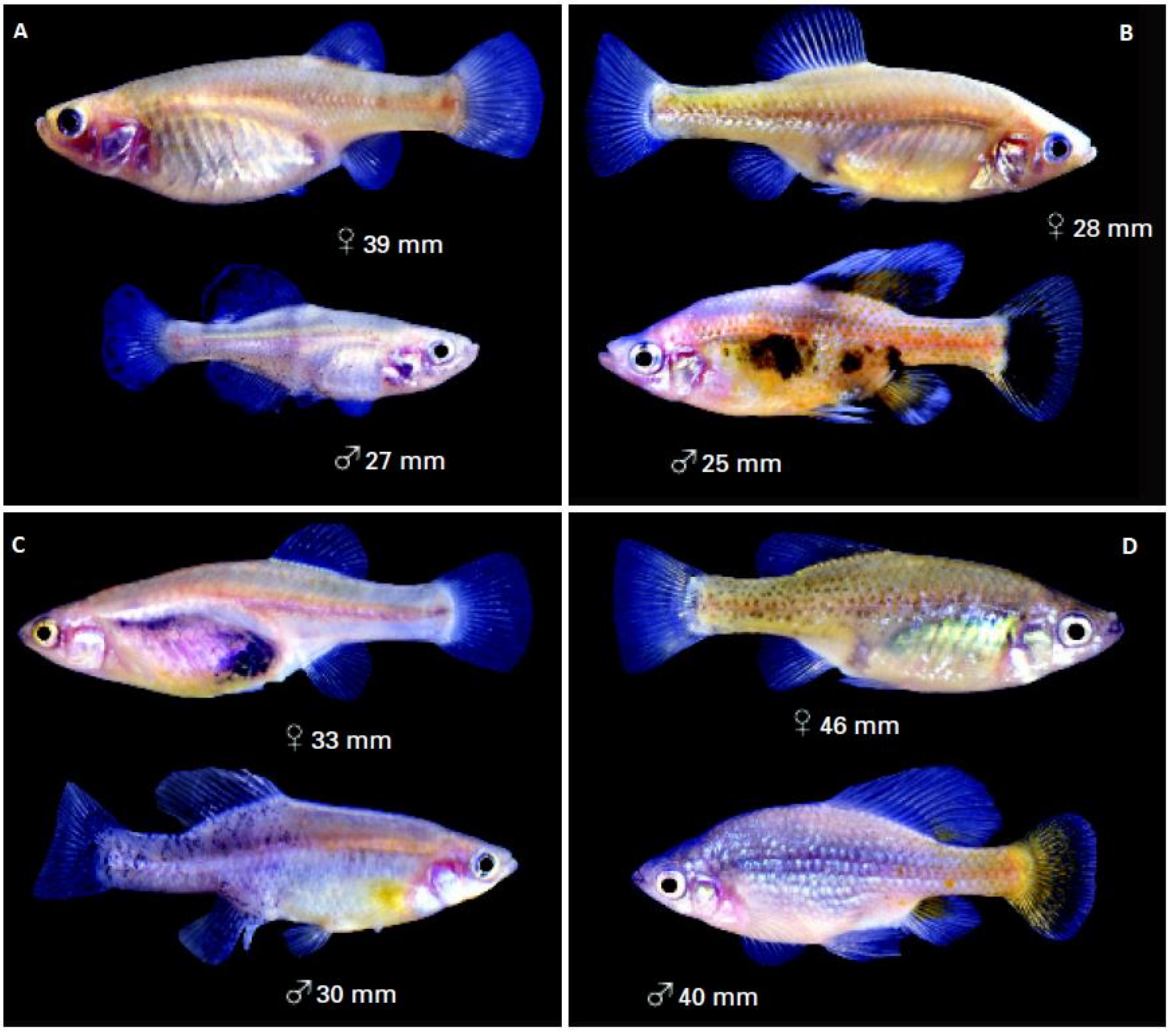
Zhu, T., Feng, S., Liu, X., & Li, Q. (2017). Next-generation sequencing yields the complete mitochondrial genome of the mummichog, *Fundulus heteroclitus*. Mitochondrial DNA Part A. 28(1):121-122.

Zimin, A. V., Puiu, D., Luo, M. C., Zhu, T., Koren, S., Marçais, G., ... & Salzberg, S. L. (2017). Hybrid assembly of the large and highly repetitive genome of *Aegilops tauschii*, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. Genome Research. 27(5):787-792.

Zymo Research. (2020). Zymoresearch.com: Quick-DNA/RNA™ MagBead. https://files.zymoresearch.com/protocols/_r2130_r2131_quick-dna_rna_magbead.pdf

10. ANEXOS

10.1. Imágenes de los peces goodeidos y localidades de *Skiffia lermae*



Fotos de los peces goodeidos utilizados en este estudio. A: *G. viviparus*. B: *S. multipunctata*. C: *S. bilineata*. D: *S. francesae*

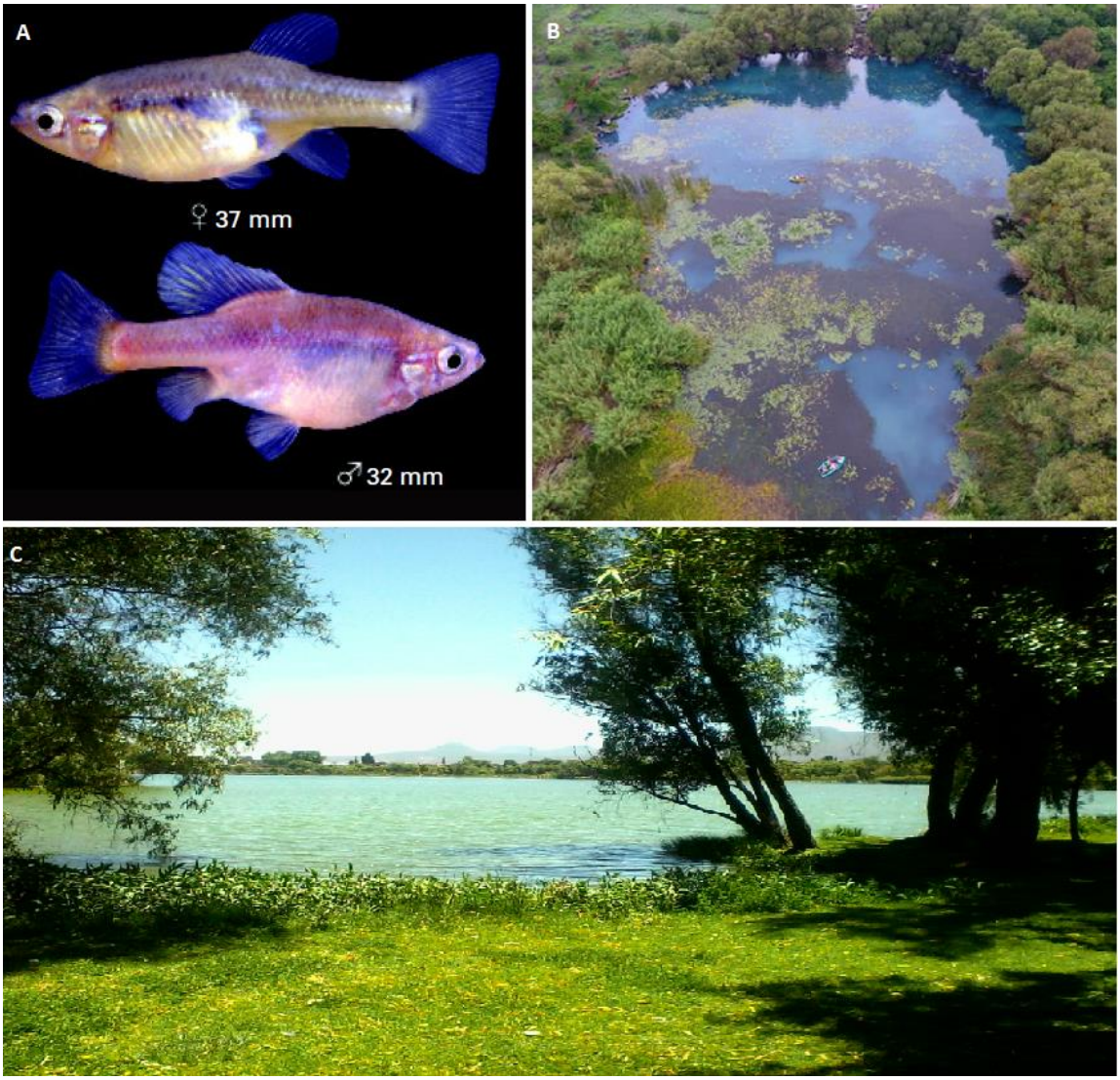


Foto de *S. lermae* (A), Manantial La Mintzita (B) y el Lago de Zacapu (C). Fotos de SEMACCDET

10.2. Extracción de ADN

1.- Incubar el tejido en 400 µl de buffer de lisis y 20 µl de proteinasa K a 55°C durante la noche

2.- Añadir 200µl de NaCl saturado y agitar vigorosamente

3.- Incubar en hielo durante 10 minutos

4.- Centrifugar a 10,000 rpm durante 10 minutos

5.- Transferir sobrenadante (500-600 µl) a un tubo con 1 ml de etanol absoluto

6.- Invertir el tubo varias veces para precipitar el ADN. El ADN debe ser visible en etanol, transferir el ADN a un tubo con 100 µl de TE

7.- Correr un gel de agarosa al 0.7% a 100V por 45 minutos con extractos crudos de ADN para verificar la extracción

Buffer de lisis (50 ml):

5M NaCl	1 ml	(concentración final 100mM)
1M Tris pH 8	2.5 ml	(concentración final 50mM)
0.5 M EDTA pH 8	10 ml	(concentración final 100mM)
10% SDS	5 ml	(concentración final 1%)

Aforar a 50 ml con agua desionizada

10.3. Extracción de ARN

1.- Homogenizar tejido con 1 ml de Trizol.

2.- Incubar por 5 min a temperatura ambiente.

3.- Añadir 200 µl de cloroformo por 1 ml de Trizol. (respectivamente)

4.- Agitar vigorosamente e incubar 2-3 min a temperatura ambiente.

5.- Centrifugar a 12,000 g durante 15 min a 4°C.

6.- Transferir sobrenadante a un tubo limpio.

7.- Añadir 500 µl de isopropanol por 1 ml de Trizol (respectivamente)

8.- Incubar a temperatura ambiente durante 15 min.

9.- Centrifugar a 12,000 g durante 10 min a 4°C.

10.- Descartar sobrenadante y lavar con 1 ml de etanol 70%.

11.- Centrifugar a 12,000 g 5 min a 4°C.

12.- Descartar sobrenadante y exceso de etanol, resuspender en 50 µl de agua DEPC.

Tratamiento con DNAsa I:

13.- Preparar mezcla maestra, hacer el cálculo dependiendo del número de muestras:

a. Agua DEPC	4.5 µl por muestra
b. 10X buffer para DNAsa I	6.5 µl por muestra
c. RNAsin	0.5 µl por muestra
d. DNAsa I (Promega)	3.5 µl por muestra

Mezclar bien con la pipeta y agregar 15µl de mezcla maestra a cada muestra de 50 µl de RNA (volumen final de 65 µl)

14.- Incubar durante 30 min a 37°C

15.- Añadir 6.5 µl de LiCl 4M.

16.- Añadir 65 µl de fenol/cloroformo y vórtex.

17.- Centrifugar a máxima velocidad (~ 13,000 rpm) durante 10 min.

18.- Transferir sobrenadante a un tubo limpio y añadir 2.5 volúmenes (~165 µl) de etanol absoluto.

19.- Incubar por lo menos 60 min a –20°C, puede permanecer toda la noche.

20.- Centrifugar 20 min a máxima velocidad a 4°C.

21.- Lavar con 0.5 ml de etanol 70% y centrifugar durante 10 min a 4°C a máxima velocidad.

22.- Eliminar etanol y resuspender en 25-50 µl de agua DEPC (dependiendo del pellet).

23.- Almacenar a –20°C o a –70°C.

10.4. Preparación de Librerías Nanopore, Modificado del Protocolo SQK-LSK009

Reparación de ADN y End-prep

1.- Colocar 3-5 µg de ADN en un tubo de PCR de 0.2 ml y subir el volumen hasta 47 µl utilizando agua libre de nucleasas

2.- A ese tubo agregar lo siguiente (todos los reactivos deben estar descongelados y haber pasado por vórtex):

* 1 µl de DNA CS, proveniente del kit SQK-LSK109
* 3.5 µl de NEBNext FFPE DNA Repair Buffer
* 2 µl de NEBNext FFPE DNA Repair Mix
* 3.5 µl de Ultra II End-prep reaction buffer
* 3 µl de Ultra II End-prep enzyme mix

3.- Mezclar bien agitando el tubo (sin vórtex) y centrifugar brevemente

4.- Usando un termociclador, incubar a 20°C por 10 minutos y a 65°C por 5 minutos

5.- Resuspender perlas AMPure con vórtex

6.- Transferir la muestra de ADN a un tubo Eppendorf Lobind

7.- Añadir 30 µl (o 60 µl para 1X) de perlas magnéticas a la reacción y mezclar agitando el tubo

8.- Incubar con agitación leve por 10 minutos a temperatura ambiente

9.- Preparar 500 µl de etanol 70% (350 µl de etanol absoluto + 150 µl de agua)

10.- Centrifugar la muestra y colocar en un rack magnético hasta que el líquido sea translúcido, desechar sobrenadante

11.- Aun estando en el rack magnético enjuagar con 200 µl de etanol 70% sin deshacer el pellet y descartar el sobrenadante. Repetir este paso

12.- Centrifugar y poner en el rack magnético, sacar cualquier resto de etanol y dejar secar por ~30 segundos

13.- Retirar el tubo del rack magnético y resuspender en 61 µl de agua libre de nucleasas. Incubar por 2 minutos a temperatura ambiente

14.- Colocar el tubo en el rack magnético y esperar a que el líquido sea translúcido. Recuperar el ADN en un tubo LoBind diferente.

15.- Cuantificar 1 µl en Qubit

Ligación de adaptadores y limpieza

Todos los reactivos de esta sección necesitan sacarse a descongelar en hielo, ponerse en vórtex y centrifugar antes de utilizarse. El buffer de ligación se debe resuspender con pipeta dada su

viscosidad

Sacar los reactivos LFB o SFB y EB a descongelar y colocar en hielo

1.- En el mismo tubo Lobind en el que se encuentran los 60 µl de ADN agregar lo siguiente:

* 25 µl de Ligation buffer (LNB) (kit SQK-LSK109)
* 5 µl de Quick T4 ligase (E6057A)
* 5 µl de Quick ligase (M2200L)
* 5 µl de Adapter Mix (AMX kit SQK-LSK109)

2.- Incubar la reacción por 20 minutos a temperatura ambiente

3.- Añadir 40 µl de perlas AMPure a la reacción y mezclar agitando el tubo

4.- Incubar con agitación leve por 10 minutos a temperatura ambiente

5.- Centrifugar la muestra y colocar en el rack magnético hasta que el líquido se vea transparente

6.- Lavar las perlas con 125 µl de Long fragment buffer (LFB), quitar del rack magnético y re suspender, re colocar en el rack magnético y desechar sobrenadante. Repetir este paso

7.- Centrifugar y colocar la muestra en el rack magnético. Retirar cualquier resto de sobrenadante. Dejar secar por ~30 segundos

8.- Retirar el tubo del rack magnético y añadir 15 µl de Elution Buffer (EB)

9.- Centrifugar e incubar por 10 minutos a 37°C

10.- Colocar el tubo en el rack magnético y recuperar los 15 µl de librería

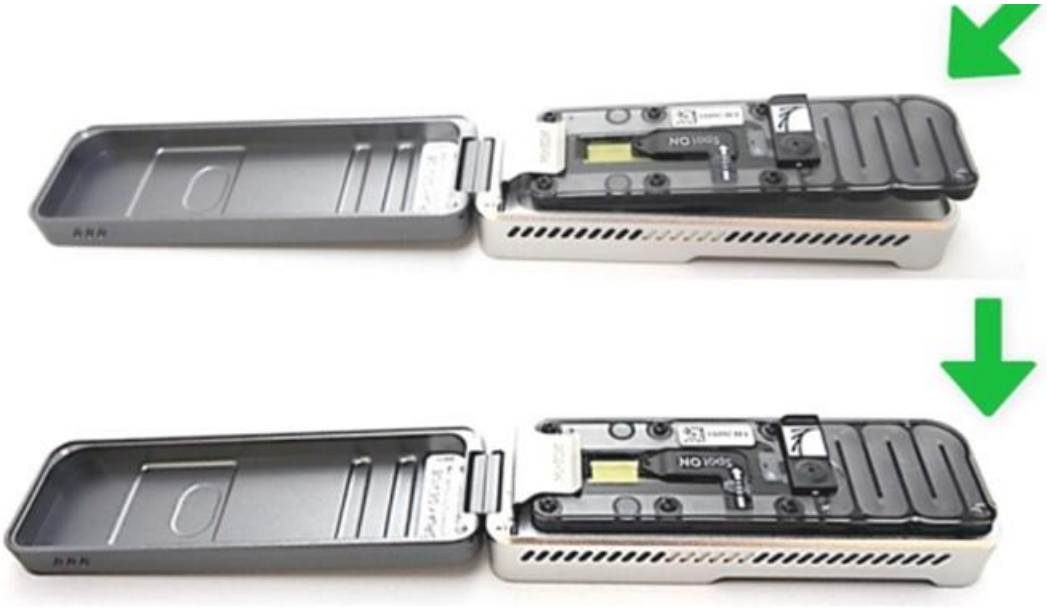
11.- Cuantificar 1 µl en Qubit

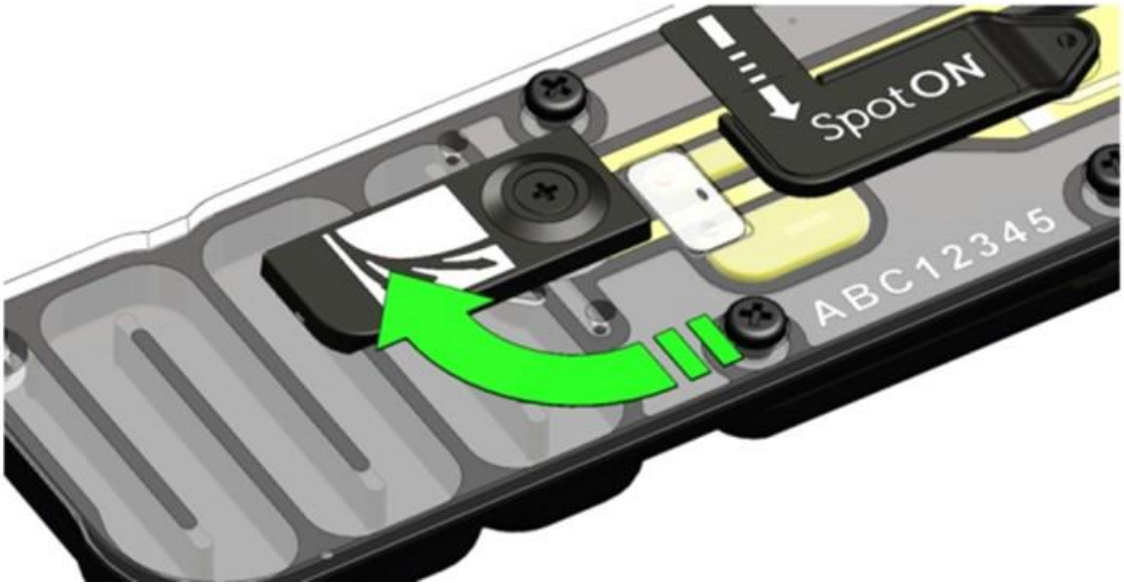
Priming y cargado de la Flowcell

Se utilizan reactivos del Flow cell priming kit (EXP-FLP002) y del Ligation Sequencing kit (SQK-LSK109)

1.- Colocar el flowcell dentro del dispositivo MinION

2.- Abrir el priming port girando 90° en sentido de las manecillas del reloj





3.- Colocar una pipeta de 1000 µl en 200 µl, insertar la punta en el priming port, gira el volumen de la pipeta hasta que esté en 230 µl para retirar un pequeño volumen de buffer

4.- Añadir 30 µl de Flush Tether (FLT) directamente a un tubo de Flush buffer (FB)

5.- Cargar 800 µl del priming mix (realizado en el paso anterior) en el priming port con cuidado de no insertar burbujas de aire

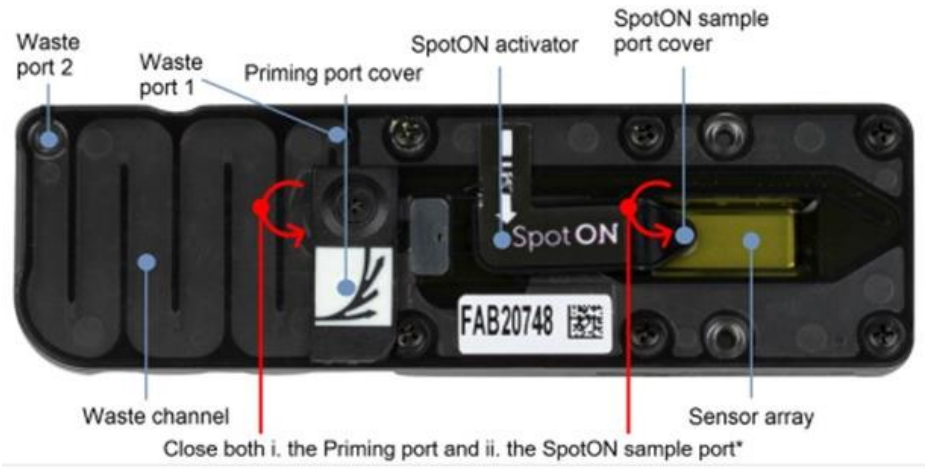
6.- En un tubo nuevo añadir lo siguiente:

* 37.5 µl de Sequencing buffer (SQB)
* 25.5 µl Loading beads (LB) resuspendidas con pipeta inmediatamente antes de agregar
* 14 µl de librería de ADN

7.- Abrir el sample port y cargar 200 µl de priming mix dentro del priming port, se verán burbujas saliendo por el sample port

8.- Resuspender por pipeteo la librería y cargar inmediatamente después en el sample port (76 µl) gota por gota

9.- Cerrar el priming port y el sample port, conectar el dispositivo a la computadora y acceder a la interfaz MinKNOW



Se recomienda cargar al menos 50 fmoles o alrededor de 1 µg de ADN

10.5. Detalles Bioinformáticos

Sección 1. Lecturas Crudas y Basecalling

Comando para realizar el basecalling:

> guppy_basecaller -i prueba/ -s recovered_basecalled/ -c dna_r9.4.1_450bps_hac.cfg -x cuda:0 --min_qscore 8

Parámetros:

-i → indica el directorio donde se encuentran los archivos .fast5

-s → indica el directorio donde se guardarán los archivos .fastq producidos

-c → indica el archivo de configuración dictado por el kit y el tipo de flowcell usado

--min_qscore → determina la Q score mínima para aprobar una base

-x → permite la lectura de CUDA, se le señala con un número qué dispositivo de memoria gráfica utilizar

Sección 2. Verificación de Calidad de las Lecturas de Oxford Nanopore

Comando de verificación de la calidad por pycoQC:

> pycoQC -f sequencing_summary.txt -o todos_fast5.html –quiet

Parámetros:

-f → ruta del archivo sequencing.summary.txt generado por el basecaller

-o → ruta donde se alojará el archivo de salida en formato html

--quiet → para que no emita el progreso al stdout (esto vuelve el proceso más ágil)

Comando de escritura del archivo summary:

> Fast5_to_seq_summary -f all_fast5/ -s sequencing_summary.txt

Parámetros:

-f → corresponde a la ruta de los archivos .fast5

-s → ruta donde se guardará el sequencing_summary.txt.

Sección 3. Ensamble del Mitogenoma de *S. lermae*

Capturar secuencias similares al mitogenoma de *Xenotoca eiseni:*

> nohup mirabait -b NC_011381.1.fasta -p gDNA_R1_001.fastq.gz gDNA_R2_001.fastq.gz -t 8 -o mira &

Parámetros:

-b → indica el archivo de referencia

-p → indica que la referencia son 2 archivos, un forward y un reverse

-t → es la cantidad de threads a utilizar (puede no especificarse)

-o → indica el archivo de salida, será un archivo multi fasta (el programa lo crea)

Ensamble del mitogenoma:

> nohup megahit -r mira -t 12 -m 0.8 -o megahit &

Parámetros:

-r → indica un archivo fasta no pareado

-t → cantidad de threads a asignar a la tarea

-m → fracción de la memoria que puede utilizar el programa

-o → directorio de salida

Anotación del mitogenoma:

> docker run -v megahit/megahit -w megahit --rm guanliangmeng/mitoz:2.3 python3 MitoZ.py annotate --fastafile sler_mitocondrial.fa --outprefix sler_anotacion --clade Chordata --annotation

Parámetros:

-v → indica donde montar el programa

-w → indica el directorio de trabajo (las lecturas a procesar deben estar en este directorio)

--rm → se le indica la versión a correr

python3 → se le da la ruta del ejecutable

annotate → es la opción para sólo anotar, tiene otras opciones para ensamble

--fastafile → indica el archivo input

--outprefix → el prefijo que estará en los archivos de salida

--clade → clado al que pertenece el organismo, solo acepta Chordata o Arthropoda

--annotation → señalar que sí va a anotar (podría omitirse si solo se usa la opción annotate)

Sección 4. Preparación de las Lecturas

Para visualizar la calidad de las lecturas cortas se usó FastQC

fastqc gDNA_R1_001.fastq

fastqc gDNA_R2_001.fastq

Filtrado de lecturas cortas con *trimmomatic*

> java -jar /opt/Trimmomatic-0.39/trimmomatic-0.39.jar PE -threads 20 ~/Slermae_illumina/gDNA_R1_001.fastq ~/Slermae_illumina/gDNA_R2_001.fastq sler.fastq.P.qtrim sler.fastq.U.qtrim sler2.fastq.P.qtrim sler2.fastq.U.qtrim ILLUMINACLIP:/opt/trinityrnaseq-v2.11.0/trinity-plugins/Trimmomatic/adapters/TruSeq3-PE.fa:2:30:10 SLIDINGWINDOW:4:25 LEADING:5 TRAILING:5 MINLEN:25 HEADCROP:2

Parámetros:

PE → indica que las lecturas vienen pareadas

-threads → la cantidad de núcleos que se asignan al proceso

Posicionalmente se le dan los 2 input y se les dan 4 claves a los archivos de salida (2 archivos por cada input, uno de lecturas pareadas y otro con lecturas que quedaron sin par)

ILLUMINACLIP → es el trimmer a utilizar, recibe un archivo con adaptadores y otros atributos

SLIDINGWINDOW → establece el tamaño de las bases que se van a promediar y la calidad específica promedio requerida

LEADING → remueve bases de baja calidad al comienzo

TRAILING →remueve bases de baja calidad al final de la lectura

MINLEN →longitud mínima de las lecturas para conservarlas

HEADCROP → corta bases al inicio de la lectura, sin importar su calidad

Retiro de adaptadores de las lecturas largas con porechop

> porechop -i nanopore.fastq.gz --format auto -t 12 -o chopped_reads.fastq.gz

Parámetros:

–format → en qué formato está el archivo, se puede dejar en ‘auto’ para indicar que lo detecte automático

-t → threads, núcleos para asignarle al proceso

-o → nombre del output generado

Corrección y recorte de las lecturas largas con canu

> canu -p canu -d nanopore_canu genomeSize=1000m -correct -maxThreads=35 -nanopore-raw chopped_nanopore_reads.fastq.gz

> canu -p canu -d canu_trimmed genomeSize=1000m -trim -maxThreads=35 -nanopore-corrected canu_corrected/canu.correctedReads.fasta.gz

Parámetros:

-p → prefijo para los archivos creados

-d → carpeta creada para los archivos de salida

genomeSize → tamaño aproximado del genoma

-correct → llevar a cabo el paso de corrección

-nanopore-raw → el archivo de entrada son lecturas crudas de nanopore

-maxThreads → limitar los threads disponibles para canu

-trim → llevar a cabo el paso de recorte

-nanopore-corrected → el archivo de entrada son lecturas corregidas de nanopore

Sección 5. Estimación del Tamaño del Genoma

Conteo de k-meros con jellyfish

> jellyfish count -t 12 -C -m 31 -s 20G -o 31mer_out --min-qual-char=? sler.fastq.P.qtrim sler2.fastq.P.qtrim

Parámetros:

-t → cantidad de núcleos a asignar al proceso

-C → contar ambas lecturas pareadas, representación canónica

-m → longitud del k-mero a contar

-s → tamaño inicial del proceso

-o → nombre del archivo de salida

--min-qual-char=? → indica que el mínimo de calidad aceptable en codificación Phred para que una secuencia sea procesada es “?”

Generación del histograma con jellyfish

> jellyfish histo -o 31mer_out.histo 31mer_out

-o → indica el nombre del archivo de salida

El archivo de entrada es un argumento posicional

Importación del archivo a R

> dataframe31 <- read.table("Slermae_illumina/trimmomatic2/31mer_out.histo")

Gráfica de los primeros 200 valores

> grafica31 <- plot(dataframe31[3:200,], type="l")

Se descartan los valores iniciales que visualmente se muestren mucho más altos que el resto

Gráfica con puntos para localizar la región de una sola copia

> plot(dataframe31[3:100,],type="l")

> points(dataframe31[3:100,]) # Se grafican los puntos

Estimación de los k-meros totales

> sum(as.numeric(dataframe31[3:9993,1]*dataframe31[3:9993,2])) # El resultado es 20096341310

Determinar la posición del punto más alto de la gráfica

> dataframe31[5:25,] # El pico es el valor 14

Cálculo del tamaño del genoma en Mb

> sum(as.numeric(dataframe31[3:9993,1]*dataframe31[3:9993,2]))/14 # Resultado de 1.43 Gb

Sección 6. Ensamble Híbrido del Genoma

Ensamble con smartdenovo

> nohup /opt/smartdenovo/smartdenovo.pl -p wtasm -e zmo -t 40 -c 1 canu.trimmedReads.fasta.gz > wtasm.mak

> make -f wtasm.mak

Parámetros:

El archivo de entrada es un argumento por sí solo

-p → prefijo para los archivos de salida

-e → el ensamblador a usar

-t → threads a asignar para el proceso

-c → generar el consenso

Alinear lecturas cortas al ensamble de lecturas largas con Minimap2

> minimap2 -ax sr -t 40 ensamble_consenso.cns sler.fastq.P.qtrim sler2.fastq.P.qtrim -o alineamiento.sam

Parámetros:

-ax → la a indica que el output sea en sam, la x indica que se usará un pre-set

sr → modo pre-set para alinear un ensamble a lecturas illumina

-t → cantidad de threads asignados al proceso

Ordenado e indexado del alineamiento con samtools

> samtools sort -m 1000M -@ 30 illumina.bam -o illumina_sorted.bam

> samtools index illumina_sorted.bam

Parámetros:

El archivo de entrada es un argumento posicional

-m → cantidad de memoria máxima por núcleo

-@ → cantidad de núcleos asignados al proceso

-o → nombre del archivo de salida

Pulido del ensamble con pilon

> java -Xmx420G -jar /opt/pilon/pilon-1.23.jar --genome ensamble_consenso.fa --frags illumina_sorted2.bam --output pilon_completo --vcf --threads 45

Parámetros:

-Xmx420G → indica cuánta memoria darle a la máquina virtual de Java

-jar → ejecutar el archivo .jar

–genome → genoma ensamblado de referencia, debe estar en fasta

–frags → alineamiento realizado con lecturas paired-end

–output → prefijo de los archivos de salida

–threads → cantidad de procesos paralelos permitido

Sección 7. Evaluación del Ensamble

Obtención de estadísticas generales con quast

> /opt/quast/quast.py pilon.fasta -1 sler.fastq.P.qtrim -2 sler2.fastq.P.qtrim -o quast_post_pilon

Parámetros:

El archivo de entrada es un argumento posicional

-1 → archivo pareado izquierdo con lecturas illumina

-2 → archivo pareado derecho con lecturas illumina

-o → nombre del archivo de salida

Evaluación y anotación del ensamble con BUSCO

> busco -i pilon_completo.fasta --augustus -m geno -l /botete/bases_de_datos/BUSCO/cyprinodontiformes_odb10/ -c 45 --long -o genoma_busco -f

-i → es el archivo de entrada

–augustus → le indica que usará augustus para buscar genes eucariotas

-m → es el modo de anotación, puede ser geno(genoma), tran(transcriptoma) o prot(proteoma)

-l → es la ubicación de la carpeta con la base de datos del linaje

-c → cantidad de núcleos a asignar al proceso

–long → incrementa la precisión de la predicción de genes, toma más tiempo en finalizar el comando

-o → etiqueta para los archivos de salida

-f → forzar sobreescritura de archivos

Sección 8. Cálculo de N50, N25 en R

Creación de variables

> genome <- readDNAStringSet("R/pilon_completo.fasta", format="fasta", use.names = FALSE)

N1 <- list(assembly1 = width(genome))

reflength1 <- sapply(N1, sum)

Función utilizada

contigStats <- function(N, reflength, style = "data", pch = 19,

```
        xlab="Percentage of Assembly Covered by Contigs of Size >= Y",
        ylab = "Contig Size [bp]", main = "Cumulative Length of Contigs",
        sizetitle = 14, sizex = 12, sizey = 12, sizelegend = 9,
        xlim, ylim) {
    Nl <- lapply(names(N), function(x) rev(sort(N[[x]])))
    names(Nl) <- names(N)
    Nlcum <- lapply(names(Nl), function(x) cumsum(Nl[[x]]))
    names(Nlcum) <- names(Nl)
    N50 <- sapply(seq(along = N), function(x) Nl[[x]]
    [which(Nlcum[[x]] - reflength[x]/2 >= 0)[1]])
    names(N50) <- names(N)
  if(style == "data") {
    N75 <- sapply(seq(along = N), function(x) Nl[[x]]
      [which(Nlcum[[x]] - reflength[x] * 0.75 >= 0)[1]])
    names(N50) <- names(N)
    N25 <- sapply(seq(along = N), function(x) Nl[[x]]
      [which(Nlcum[[x]] - reflength[x] * 0.25 >= 0)[1]])
    names(N50) <- names(N)
  stats <- cbind(N25, N50, N75, Longest = sapply(N, max),
    Mean = sapply(N, mean), Median = sapply(N, median), Shortest = sapply(N, min),
N_Contigs = sapply(N, length))
          return(c(Nlcum, Contig_Stats = list(stats)))
    }
      if(style == "plot") {
      if(missing(xlim)) xlim <- c(0, 100)
      if(missing(ylim)) ylim <- c(0, max(unlist(N)))
      split.screen(c(1,1))
      for(i in seq(along = Nl)) {
            if(i == 1) {
      plot(Nlcum[[i]]/reflength[[i]] * 100, Nl[[i]], col = i,
      pch = pch, xlim = xlim, ylim = ylim, xlab = xlab,
```

```
        ylab = ylab, main = main)
            }
        screen(1, new = FALSE)
        plot(Nlcum[[i]]/reflength[[i]] * 100, Nl[[i]], col = i,
        pch = pch, xlim = xlim, ylim = ylim, xaxt = "n",
        yaxt = "n", ylab = "", xlab = "", main = "", bty = "n")
        }
        legend("topright", legend = paste(names(N50), ": N50 = ",
        N50, sep = ""), cex = 1.2, bty = "n", pch = 19,
        pt.cex = 1.2, col = seq(along = Nl))
        close.screen(all = TRUE)
    }
}
```

Obtención de los resultados

```
stats1 <- contigStats(N = N1, reflength = reflength1, style = "data")
stats1[["Contig_Stats"]]

stats2 <- contigStats(N = N1, reflength = reflength1, style = "plot")
```

Sección 9. Ensamble y Anotación de Transcriptomas

Revisar la calidad de las lecturas:

```
> fastqc genewiz/*.fastq.gz
```

Ensamble *de novo* con Trinity:

```
> nohup Trinity --seqType fq --left 1-Gvi_R1_001.fastq.gz --right 1-Gvi_R2_001.fastq.gz --CPU
12 --max_memory 80G --trimmomatic --output trinity_gviv --full_cleanup &
```

Parámetros:

--seqType → qué tipo de secuencias va a recibir, fasta o fastq

--left y –right → archivos pareados de entrada

--CPU → núcleos que puede utilizar el programa

–max_memory → máxima memoria asignada para este proceso

--trimmomatic → realizar cortes de calidad antes de ensamblar

--output → nombre que llevarán los archivos de salida

--full_cleanup → limpia todos los archivos temporales después de terminar el ensamble

Observar las características del ensamble con TrinityStats:

> TrinityStats.pl trinity_gviv.Trinity.fasta

Representatividad del ensamble con bowtie2, primero se construye un índice:

> bowtie2-build trinity_gviv.Trinity.fasta trinity_gviv.Trinity.fasta

> nohup bowtie2 -p 10 -q --no-unal -k 20 -x trinity_gviv.Trinity.fasta -1 1-Gvi_R1_001.fastq.gz -2 1-Gvi_R2_001.fastq.gz 2>align_stats.txt | samtools view -@10 -Sb -o bowtie2.bam &

Parámetros:

-p → número de threads a permitir

-q → indica que los archivos de entrada son fastq

--no-unal → que no lance registros SAM para las lecturas que no se alinearon

-k → cantidad máxima de alineamientos por lectura que puede reportar

-x → archivo de entrada (que ya tiene sus índices)

-1 y -2 → indican los archivos fastq de los que provino el ensamble

> → indica que la salida será a align_stats.txt

BUSCO para evaluar qué tan completos están los ensambles:

> busco -i trinity_smul.Trinity.fasta -o ensamble -l actinopterygii_odb10/ -m tran -c 20 -f

Parámetros:

-i → archivo de entrada, el ensamble producido por trinity

-o → nombre del folder con los archivos de salida

-l → ubicación de la base de datos correspondiente al linaje

-m → modo de uso, en este caso tran para transcriptoma

-c → núcleos a utilizar

-f → forzar la sobre escritura en caso de tener algunos archivos generados anteriormente

Predicción de regiones codificantes:

> nohup TransDecoder.LongOrfs -t trinity_gviv.Trinity.fasta --gene_trans_map trinity_gviv.Trinity.fasta.gene_trans_map -m 200 --output_dir longorfs/ &

Parámetros:

-t → el archivo de salida con el ensamble de trinity

--gene_trans_map → mapa arrojado también por trinity

-m → a partir de qué longitud considerar cada proteína como válida

--output_dir → nombre de la carpeta que contendrá los archivos producidos

Descripción de las regiones codificantes con BLAST:

> nohup blastp -query trinity_gviv.Trinity.fasta.transdecoder.pep -db uniprot_sprot.pep -num_threads 10 -max_target_seqs 1 -evalue 1e-5 -outfmt 6 > blastp.gviv &

> nohup blastx -query trinity_gviv.Trinity.fasta -db uniprot_sprot.pep -num_threads 10 -max_target_seqs 1 -evalue 1e-5 -outfmt 6 > blastx.gviv &

Parámetros:

-query → archivo. pep de transdecoder

-db → ubicación de la base de datos

-max_target_seqs → cantidad máxima de hits por secuencia

-outfmt → se especifica el formato del archivo de salida

-evalue → establece el evalue mínimo a considerar

-num_threads → número de hilos permitido para el proceso

Búsqueda con la base de datos de pfam:

> nohup hmmscan --cpu 8 --domtblout pfam.domtblout Pfam-A.hmm trinity_smul.Trinity.fasta.transdecoder.pep &

Parámetros:

--cpu → cantidad de cpus a utilizar en el proceso

--domtblout → el nombre del archivo de salida

Rnammer para identificar las regiones ribosomales:

> RnammerTranscriptome.pl --transcriptome trinity_gviv.Trinity.fasta --path_to_rnammer /opt/RNAMMER/rnammer

Parámetros:

--transcriptome → le especifica el ensamble

--path_to_rnammer → la ruta donde el programa está instalado

Transdecoder para encontrar marcos de lectura considerando los hits con pfam y blast:

> TransDecoder.Predict -t trinity_sfran.Trinity.fasta --retain_pfam_hits pfam.domtblout --retain_blastp_hits blastp.sfran --output_dir transdecoder/longorfs/

Parámetros:

-t → es el archivo de salida de trinity

--single_best_only → opción para retener sólo los mejores ORFs

--output_dir → se le otorga el directorio de salida del paso anterior

Anotación del transcriptoma:

> Trinotate Trinotate_3.2.0.sqlite init --gene_trans_map trinity_smul.Trinity.fasta.gene_trans_map --transcript_trinity_smul.Trinity.fasta --transdecoder_pep trinity_smul.Trinity.fasta.transdecoder.pep &&
Trinotate Trinotate_3.2.0.sqlite LOAD_swissprot_blastp blastp.gviv
Trinotate Trinotate_3.2.0.sqlite LOAD_pfam pfam.domtblout
Trinotate Trinotate_3.2.0.sqlite LOAD_swissprot_blastx blastx.gviv
Trinotate Trinotate_3.2.0.sqlite report > trinotate_annotation_report.xls

Parámetros:

Recibe un archivo .sqlite que es una base de datos

--gene_trans_map → se le otorga el archivo .gene_trans_map saliente de trinity

--transdecoder_pep → se le da el archivo saliente de transdecoder con extensión .pep

El comando debe correrse varias veces con las opciones LOAD_swissprot_blastp (para cargar la búsqueda de blastp), LOAD_pfam (para cargar el archivo pfam.domtblout), LOAD_swissprot_blastx (para cargar la búsqueda blastx)

Por último, se debe correr el comando report y otorgarle un nombre al archivo de salida con >

Sección 10. Selección de Ortólogos

Realizar bases de datos a partir de los transcriptomas ensamblados

> makeblastdb -in transdecoder_sler.cds -dbtype nucl

Parámetros:

-in → archivo del que se hará una base de datos blasteable

-dbtype → tipo de base de datos a generar

Búsqueda individual de ortólogos

> blastn -query cyp450.fa -db transdecoder_sler.cds -outfmt 7

Parámetros:

-query → el archivo con la secuencia a buscar

-db → la base de datos donde se buscará

-outfmt → formato de salida de los resultados de blast

Recuperación manual de cada ortólogo en archivos separados

> grep TRINITY_DN455_c0_g1_i4.p1 transdecoder_sler.cds -A22

Parámetros:

El primer argumento es qué va a buscar y el segundo argumento es dónde, en qué archivo

-A22 → que imprima las 22 líneas siguientes (depende de la longitud del ortólogo este número cambió)

Sección 11. Análisis de Mutaciones

Alineamiento con clustalo

> clustalo -i all_p450.nucl -t DNA --threads 40 --guidetree-out guide_tree_p450 -o alineamiento_p450

Parámetros:

-i → archivo de entrada con secuencias a alinear

-t → tipo de archivo de entrada

--threads → cantidad de núcleos a asignar al proceso

--guidetree-out → pedir un árbol filogenético

-o → etiqueta de los archivos de salida

10.6. Digestión con RNAsa

1.- Añadir al ADN extraído 140 µl de agua miliQ y 10 µl de RNAsa A

2.- Agregar 1/10 volúmenes de acetato de sodio 3M y 2.5 volúmenes de etanol 100%

3.- Colocar a -80°C por 30 minutos

4.- Centrifugar a 12,000rpm durante 20 minutos

5.- Decantar cuidadosamente el sobrenadante

6.- Lavar el pellet con alcohol 70% frío

7.- Centrifugar por 3-5 minutos

8.- Sacar la mayor cantidad posible de etanol con una pipeta

9.- Dejar secar el pellet

10.- Re suspender con TE

10.7. Ortólogos Dotrocumentados

Estos ortólogos formaron parte de la base de datos inicial pero no fueron encontrados en las secuencias de goodeidos.

N°	Nombre	Abreviación	Referencia
1	Citocromo P450 2D6	CYP2D6	Uno *et al*., 2012
2	Citocromo P450 3A4 isoforma 2	CYP3A4	Uno *et al*., 2012
3	Citocromo P450 2C8 isoforma b	CYP2C8	Uno *et al*., 2012
4	Proteína de choque térmico beta-7	HSPB7	Basu *et al*., 2002
5	Proteína de choque térmico beta-1	HSPB1	Basu *et al*., 2002
6	Proteína de choque térmico B6	HSPB6	Basu *et al*., 2002
7	ATPasa activadora de 90 kDa proteína de choque térmico	HSPATP	Kayhan y Duman, 2010
8	Proteína de choque térmico 105 kDa	HSP105	Kayhan y Duman, 2010
9	Receptor de arilo-hidrocarburo	AHR	Bemanian *et al*., 2004
10	Represor del receptor arilo-hidrocarburo	AHRR	Bemanian *et al*., 2004
11	Super oxido dismutasa	SOD	Awasthi *et al*., 2018
12	Catalasa	CAT	Awasthi *et al*., 2018
13	Proteína tumoral p53	P53	Awasthi *et al*., 2018
14	Regulador de apoptosis BCL2 asociada X	BAX	Awasthi *et al*., 2018

15	Regulador de apoptosis BCL2	BCL2	Awasthi *et al*., 2018	
16	Peptidasa apoptótica factor 1 de activación	APAF1	Awasthi *et al*., 2018	
17	Caspasa 3, relacionada a apoptosis cisteína peptidasa	CASP3A	Awasthi *et al*., 2018	
18	Glutatión S-transferasa	GST	Mukhopadhyay y Chattopadhyay, 2014	
19	Glutatión peroxidasa 1a	GPX	Mukhopadhyay y Chattopadhyay, 2014	
20	Hormonas glycoproteicas, polipéptido alfa	CGA	Nourizadeh *et al.*, 2009	
21	Coactivador 3, receptor nuclear	NCOA3	Nourizadeh *et al.*, 2009	
22	6-fosfofructo-2-quinasa/fructosa-2,6-bifosfatasa 3	PFKFB3	Nourizadeh *et al.*, 2009	
23	Citocromo oxidasa subunidad I	PTB1	Nourizadeh *et al.*, 2009	
24	NAD(P)H deshidrogenasa, quinona 1	NQO1	Shaw *et al*., 2019	
25	Hemo oxigenasa 1	HO1	Shaw *et al*., 2019	
26	Metaloteioneina-1	MTS	Pierron *et al.*, 2009	
27	Metalotioneina-A	MTSA	Pierron *et al.*, 2009	
28	Metalotioneina 2	MT2	Wang *et al*., 2019	
29	Factor de crecimiento insulínico tipo 1	IGF1	Burkina *et al*., 2015	
30	Receptor de arilo-hidrocarburo 2	AHR2	Burkina *et al*., 2015	
31	Receptor de estrogenos	ER	Burkina *et al*., 2015	
32	Receptor 5-hidroxitriptamina 1A	HT1A	Burkina *et al*., 2015	
33	Citocromo P450 familia 11 subfamilia B miembro 1	CYP11B1	Burkina *et al*., 2015	
34	Citocromo P450 familia 11 subfamilia A miembro 1	CYP11A	Burkina *et al*., 2015	
35	Citocromo P450 familia 11 subfamilia B miembro 12	CYP11B2	Burkina *et al*., 2015	
36	Citocromo P450 17alfa-hidroxilasa	CYP17	Zhu *et al*., 2019	
37	Citocromo P450, familia 19, subfamilia A	CYP19	Zhu *et al*., 2019	
38	21-hidroxilasa esteroidea	CYP21	Zhu *et al*., 2019	
39	Vitelogenina	VTG	Zhu *et al*., 2019	
40	Hormona receptora liberadora de gonadotropina	GNRH	Zhu *et al*., 2019	
41	Factor nuclear de transcripción κB	P65	Wang *et al*., 2019	
42	Factor de transcripción regulatorio de metales	MTF1	Wang *et al*., 2019	
43	Factor de necrosis tumoral α	TNFA	Wang *et al*., 2019	
44	Interleucina-6	IL6	Wang *et al*., 2019	
45	Sintasa inducible de óxido nítrico	INOS	Wang *et al*., 2019	
46	Factor de transcripción de choque térmico 1	HSF1	Wang *et al*., 2019	
47	Factor de transcripción de choque térmico 2	HSF2	Wang *et al*., 2019	