UNIVERSIDAD NACIONAL AUTÓNOMA DE MEXICO
MAESTRÍA EN CIENCIAS (NEUROBIOLOGÍA)

Análisis de la carga mutacional en el genoma mitocondrial en el cerebro del ratón

TESIS
Que para optar por el grado de
Maestra en Ciencias

PRESENTA:
I.B.Q. Kenya Lizbeth Contreras Ramírez

Tutor:
Dr. Alfredo Varela Echavarría, INB. UNAM

Comité Tutor:
Dra. Carla Daniela Robles Espinoza, LIIGH, UNAM
Dra. Teresa Edith Garay Rojas, INB, UNAM

Juriquilla, Qro., Abril, 2021

“Sería inútil referirme a las [aventuras] de ayer, porque yo entonces era una persona distinta. “

<Alicia en el país de las maravillas: Lewis Carroll>

I. AGRADECIMIENTOS

Este trabajo se llevó a cabo con el apoyo de una beca CONACYT (No. **858018**) como requisito de graduación del Programa de Maestría en Ciencias (Neurobiología) en el Instituto de Neurobiología, UNAM.

FINANCIAMIENTO:

Este trabajo se realizó con fondos de los proyectos CONACYT QRO-2018-01-01-88344, 238566 y 299041 y PAPIIT IN229620.

Instituto de Neurobiología, UNAM

Laboratorio de diferenciación neural y axogénesis

- Carlos Lozano Flores

Laboratorio de mapeo de función cerebral

- Leopoldo González Santos

Unidad de proteogenómica

- Michael Conrad Jeziorski
- Adriana González Gallardo
- Anaid Antaramian Salas

Unidad Cómputo

- Alberto Lara Ruvalcaba
- Omar González Hernández
- Ramón Martínez Olvera

Laboratorio Universitario de Bioterio, UNAM

- Alejandra Castilla León
- José Martín García Servín

Laboratorio Nacional de visualización científica y avanzada, UNAM

- Luis Alberto Aguilar Bautista
- Alejandro de León Cuevas
- Carlos Sair Flores Bautista

Instituto de Mátematicas, UNAM

- Maribel Hernández Rosales
- Janeth de Anda Gil

II. DEDICATORIA

A mis padres Zorahida y Benjamín que fueron mis pilares emocionales en los momentos más duros del camino y que siempre me han apoyado para ir hasta donde yo quiera ir, yo no llegaría a los lugares que quiero sin el soporte, los consejos, el apoyo y el amor que me brindan y sé que celebran conmigo cada paso.

A mis hermanos Benja y Katya que han apoyado mi camino siempre y por todas las risas que hacen siempre más ligero el viaje.

Al Dr. Alfredo, gracias por la oportunidad de aprender de usted, de sus ideas y proyectos y por abrirme las puertas de su laboratorio y de su increíble equipo de trabajo.

A Daniela Robles, por toda la paciencia, el apoyo y el tiempo tomado para enseñarme, para ayudarme cuando lo necesité y todos los ánimos y consejos. Apoyaste mis ideas, las impulsaste y me ayudaste a dirigirlas. Gracias por dejarme aprender tanto de ti.

A mis amigos del laboratorio Carlos, Katia y Emilio fue un tiempo tan divertido el poder compartir con ustedes y aprender de ustedes. Carlos gracias por toda la dedicación para que yo aprendiera técnicas nuevas y todo el apoyo. Katia gracias por darme tu amistad, tus ideas, tu empuje, he aprendido tanto de ti y te admiro muchísimo, espero todos nuestros proyectos sigan avanzando. Emilio no sé si leas esto algún día, pero estoy muy contenta de que hayas pasado por el laboratorio y haber compartido contigo eres brillante amigo.

A mis amigas Ali y Miri, gracias por todos los momentos juntas en la maestría y por escuchar todas mis frustraciones, por las cenas, las salidas, las pijamadas, las noches de Friends, las tardes de estudio, todo, hicieron cada momento memorable.

A Gus, que siempre está impulsándome a que alcance mis metas en la ciencia y en la literatura hasta a ganar mi primer concurso. Gracias por los consejos, los ánimos y todo.

A mi amiga Sir, gracias por siempre estar ahí, escuchar con atención mis pláticas científicas y siempre recordarme lo que quiero llegar a ser aún cuando a mí se me olvida. Gracias por decirme que soy brillante y talentosa.

A todas la personas que me apoyaron. Tal vez no mencioné a todos, pero gracias a cada uno.

Finalmente, a mí misma, por la paciencia, el empeño y la resiliencia para salir adelante pese a todas eventualidades que presentó el desarrollo de esta tesis y por no darme por vencida.

INDICE DE CONTENIDO

Índice de Tablas

Índice de figuras

1. RESUMEN

La mitocondria es un organelo citoplasmático cuya principal función consiste en la producción de energía para la célula, que contiene un genoma de DNA circular de doble cadena y codifica proteínas que forman parte de los complejos de la fosforilación oxidativa además de genes para RNAs de trasferencia y ribosomales de distribución restringida a la matriz mitocondrial. La tasa de mutación en el genoma mitocondrial es mayor que la del genoma nuclear, lo que genera variantes de diversos tipos que incluyen deleciones, sustituciones e inserciones de bases. En el humano variantes en el genoma mitocondrial están ligadas al desarrollo de distintas enfermedades neurodegenerativas y neuromusculares que con frecuencia cursan con la disfunción mitocondrial.

El desarrollo de la secuenciación de nueva generación (NGS) ha permitido mejorar el análisis del genoma mitocondrial tanto en la clínica como en la investigación. Sin embargo, para su uso se requiere de herramientas bioinformáticas específicas o la adaptación de las utilizadas para el análisis del DNA nuclear tanto para el alineamiento con el genoma de referencia como para la identificación y cuantificación de variantes. Estas herramientas deben tomar en cuenta las características particulares del genoma mitocondrial como la presencia de múltiples copias por célula y su estructura circular.

En este trabajo se llevó a cabo un análisis de la diversidad y carga mutacional en el genoma mitocondrial de cerebros de ratón en distintas etapas de la vida (embrionario, joven y adulto envejecido). Para ello se utilizaron distintas herramientas bioinformáticas incluyendo un prototipo desarrollado como parte de este proyecto y se comparó con otras previamente publicadas. El hallazgo principal de este estudio es la alta diversidad de variantes de alta y baja frecuencia que en conjunto revelan una alta abundancia de variantes genéticas en el genoma mitocondrial en todas las etapas de la vida estudiadas.

2. ABSTRACT

Mitochondria are cytoplasmic organelles whose main function is the production of energy for the cell. Mitochondria contain a double-stranded circular DNA genome which encodes proteins that are part of oxidative phosphorylation complexes, as well as transfer and ribosomal RNAs restricted to the mitochondrial matrix. The mutation rate of the mitochondrial genome is higher than that of the nuclear genome and contains variants of various types such as deletions, substitutions and insertions of bases. In humans, variants of the mitochondrial genome cause different diseases including neurodegeneration and neuromuscular alterations that in many cases have been related to mitochondrial dysfunction.

The development of next generation sequencing (NGS) technologies has allowed in depth analyses of the mitochondrial genome both in clinical and in research settings. However, they require specific bioinformatics tools or the adaptation of those used for the analysis of nuclear DNA both for for alignment to the reference genome and the identification and quantification of variants taking into account the unique features of the mitochondrial genome such as the presence of multiple copies or its circular structure. In this work, we carried out an analysis of the diversity and mutational load in the mitochondrial genome of mouse brains at different stages of life (embryonic, young and old adult). For this, different bioinformatics tools were used, including a prototype developed as part of this project and compared with others previously published. The main finding of this study is the high diversity of high and low frequency variants that together reveal a high abundance of genetic variants of the mitochondrial genome at all stages of life studied.

3. INTRODUCCIÓN

3.1 Aspectos Generales

Las mitocondrias son organelos multifuncionales heredados por la madre a través del citoplasma del huevo los cuales forman redes dinámicas que responden al estado metabólico de las células. El mantenimiento de tales redes depende de un delicado equilibrio entre la biogénesis mitocondrial y su degradación por mitofagia. Las mitocondrias juegan un papel preponderante en la producción y el almacenamiento de energía por la oxidación de sustratos orgánicos en condiciones aeróbicas a través de la respiración, aunque también tienen muchas funciones anabólicas (Alston, Rocha, Lax, Turnbull, & Taylor, 2017; Friedman & Nunnari, 2014; Kowaltowski, 2000).

Las mitocondrias son organelos citoplasmáticos de doble membrana. La membrana externa separa la mitocondria del citoplasma y la membrana interna forma pliegues o invaginaciones hacia la luz mitocondrial formando crestas lo que a su vez define la matriz mitocondrial que contiene enzimas que catalizan las reacciones asociadas con la producción de energía (Taanman, 1999). En la membrana interna también se encuentran alojados los cinco complejos enzimáticos que forman el sistema de fosforilación oxidativa (Taanman, 1999).

En 1988 se describieron por primera vez enfermedades causadas por anomalías en la función mitocondrial o mutaciones en su genoma. A la fecha, distintas patologías se han asociado con estas alteraciones, como mitopatías, enfermedades neurodegenerativas, cardiovasculares, desórdenes metabólicos, autoinmunes, musculoesqueléticos, fatigantes, psiquiátricos y cáncer (Nicolson, 2014). Además, muchas de las variantes se trasmiten por línea materna, lo que hace que el diagnóstico en un individuo pueda tener implicaciones en varias generaciones de una familia (Montoya, Playán, Solano, & Alcaine, 2000).

3.1.1 Morfología y dinámica

Las mitocondrias son organelos muy dinámicos y sus cambios morfológicos se regulan a través de los mecanismos de fusión (las membranas externas mitocondriales se unen) y fisión en el cual se experimenta un fenómeno de fragmentación de la red mitocondrial (Escobar-Henriques & Anton, 2013). Esto

genera mitocondrias de diferentes formas y tamaños lo cual tiene implicaciones funcionales ya que la morfología mitocondrial tiende a variar en función del tipo de célula, la etapa del desarrollo y el ambiente (Meyer et al., 2013).

Los mecanismos de fusión y fisión están asociados con la segregación del ADN mitocondrial (ADNmt) en las células eucariotas superiores. Los eventos de fusión facilitan el transporte de los productos generados por la expresión del ADNmt (Escobar-Henriques & Anton, 2013). La fusión también se ha asociado con el transporte mitocondrial en el citoesqueleto y recientemente se ha planteado la hipótesis de que podría estar involucrada en la capacidad de las células humanas para tolerar altos niveles de ADNmt patógeno (Friedman & Nunnari, 2014).

El equilibrio entre los mecanismos de fusión y fisión determina el mantenimiento y funcionalidad de la red mitocondrial (Picard, Taivassalo, Gouspillou, & Hepple, 2011). Se ha demostrado que la disminución o pérdida de alguno de estos mecanismos, genera daños severos tanto a nivel estructural como funcional en la red, ya sea por fragmentación de segmentos o por la generación de mitocondrias alargadas que aumentan las interconexiones de la red (Friedman & Nunnari, 2014).

Otras observaciones han revelado que factores de estrés pueden propiciar cambios en la morfología mitocondrial causando aumentos de producción de especies reactivas de oxígeno y la inducción de la apoptosis (Roberts, 2016).

3.1.2 Funciones

Las mitocondrias desempeñan distintas funciones altamente interconectadas y van más allá de los límites de la célula como unidad, ya que influyen en la fisiología de los organismos que expresan cierto fenotipo mediante la regulación de diversos procesos fisiológicos (Nunnari & Suomalainen, 2012).

Entre las funciones más conocidas de las mitocondrias se encuentran la producción de energía mediante la fosforilación oxidativa, la regulación de calcio, la apoptosis, la modulación de la actividad sináptica (Roberts, 2016), la autofagia y la mitofagía, además de estar implicadas en los mecanismos de homeostasis celular y biosíntesis de diversos metabolitos (Baker et al., 2017).

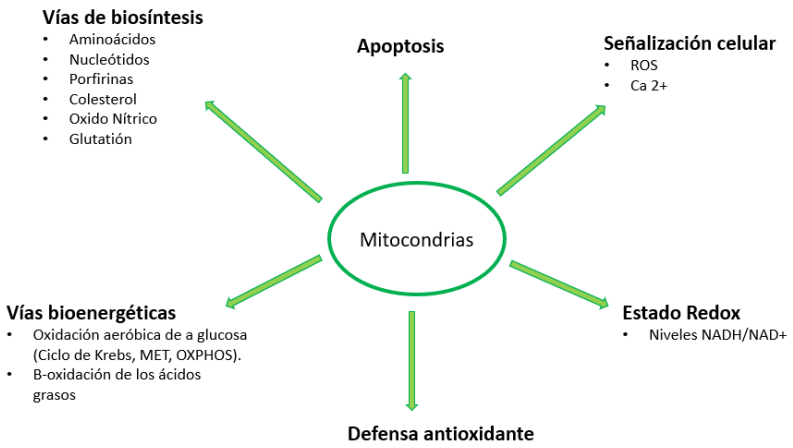

Figura 1. Funciones mitocondriales. *Principales funciones mitocondriales involucradas en procesos y vías fundamentales para la célula (tomado de Herst, Rowe, Carson, & Berridge, 2017).*

La fosforilación oxidativa (FOSFOX) es uno de los procesos más importantes en los que la mitocondria está involucrada, pues cubre alrededor del 90% de los requerimientos de ATP de la célula para llevar a cabo sus funciones endergónicas. Este proceso se lleva a cabo mediante la cadena respiratoria, la cual clásicamente se define como cuatro complejos que regulan el transporte de electrones que derivan del metabolismo intermedio. Esta transferencia de electrones se acopla con un gradiente de protones que luego se disipa por la acción de la ATPasa que funge como un quinto complejo que forma el ATP (Vafai & Mootha, 2012).

Debido a los procesos metabólicos que ocurren en las mitocondrias, son los sitos donde se generan la mayor cantidad de especies reactivas de oxígeno (ERO) las cuales son productos secundarios de la FOSFOX. Estos subproductos son dañinos ya que generan radicales como el peróxido de hidrógeno, el superóxido y el hidroxilo; que son perjudiciales para la célula por generar daños en sus componentes macromoleculares (proteínas, lípidos y ADN) (Lax, Turnbull, & Reeve, 2011).

3.2 El genoma mitocondrial

El ADNmt de los mamíferos es un genoma circular de doble cadena compuesto por alrededor de 16,000 pares de bases con dos orígenes de replicación (O_H y O_L). Este genoma codifica para 13 polipéptidos que son parte de los 5 complejos de membrana que actúan en la cadena de transporte para la producción de ATP. De éstos, siete son subunidades del complejo I (ND1, 2, 3, 4L, 4, 5, 6), una subunidad del complejo III (cytb), tres subunidades del complejo IV (COXI, II, III) y dos subunidades del complejo V (ATP 6 & 8). Los 24 genes restantes codifican para 22 ARNs de transferencia (ARNt) y dos ARNs ribosomales (ARNr) (12S, 16S) los cuales son esenciales para la síntesis de proteínas en la mitocondria. Además de estas regiones, en el genoma mitocondrial hay una región de control principal conocida como *D-loop* que contiene el origen de la replicación (O_H) y los promotores (P_H y P_L) para la transcripción del ARN mitocondrial en dos cistrones que posteriormente se fragmentan (Gorman et al., 2015; Stewart & Chinnery, 2015).

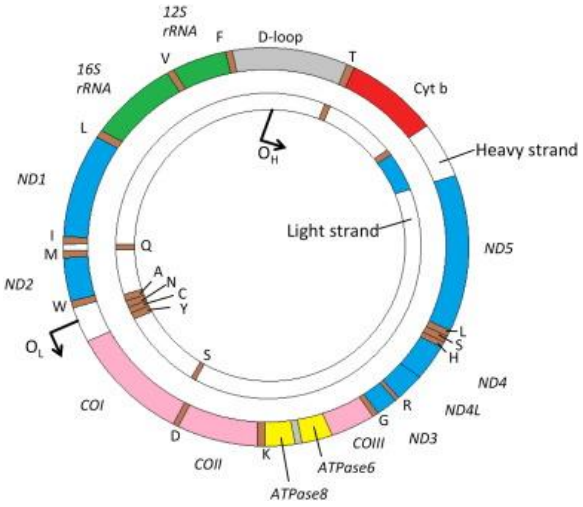

Figura 2. Genoma mitocondrial*. En verde se ilustran los dos genes que codifican ARN ribosomal (ARNr), en rosa los que codifican la citocromo oxidasa c, en azul los genes que codifican las subunidades del complejo I, en rojo se puede observar el citocromo b del complejo III. Las subunidades de ATP sintasa están en amarillo. La región de control no codificante (D-loop) es gris, y los orígenes de la replicación de la cadena pesada y ligera se denominan O_H y O_L, respectivamente (tomado de Keogh & Chinnery, 2015).*

Las cadenas que conforman el ADNmt reciben los nombres de pesada (H) y ligera (L) debido a la diferencia en su contenido de GC. La mayor parte de la información genética se codifica en la cadena pesada que contiene los genes para dos ARNr, 14 ARNt y 12 polipéptidos, mientras que la cadena ligera codifica ocho

ARNt y un polipéptido. El ADNmt de mamíferos es una estructura compacta y organizada, sus genes carecen de intrones y con excepción de la región reguladora *D-loop*, las secuencias intergénicas están limitadas a unas pocas bases (Stewart & Chinnery, 2015; Taanman, 1999).

Pese a que el genoma mitocondrial sólo codifica para 13 péptidos, se ha estimado que el mitoproteoma contiene alrededor de 1500 proteínas. Las proteínas restantes se producen por los ribosomas en el citoplasma mediante ARN mensajero que proviene del núcleo; éstas se transportan de manera activa para su ensamblaje y forman parte de los componentes mitocondriales (Nunnari & Suomalainen, 2012).

Se cree que en la mayoría de los organismos eucariotes que presentan reproducción sexual, el ADNmt proviene predominantemente de vía materna, ya que el número de copias en el ovocito no fertilizado asciende a unos cientos de miles, mientras que en los espermatozoides es alrededor de 100 (Stewart & Chinnery, 2015). Se ha planteado que las mitocondrias de los espermatozoides son eliminadas durante la embriogenésis temprana mediante destrucción selectiva, inactivación o dilución por el vasto contenido de mitocondrias ovocitarias. Sin embargo, algunos estudios han demostrado que si bien, el dogma central de la transferencia del ADNmt sigue siendo válido, existen algunos casos donde el ADNmt paterno puede transmitirse a la descendencia, aunque el mecanismo de este modo de herencia es aún desconocido (Luo et al., 2018; Sambasivarao, 2014; Stewart & Chinnery, 2015).

Dado que hay entre 100 y 10,000 mitocondrias por célula en mamíferos y estas mitocondrias contienen de dos a diez copias del genoma mitocondrial (Guo et al., 2012), es posible diferenciar dos condiciones en las células respecto a tales copias, la heteroplasmia y la homoplasmia. Si la mayoría de las moléculas son idénticas prevalece una condición de homoplasmia; sin embargo, la presencia de variantes de secuencia en distintas copias del ADNmt se define como heteroplasmia, la cual puede variar entre células, tejidos, órganos e individuos. Los niveles de heteroplasmia se han asociado con la predisposición y el desarrollo

de enfermedades (Stewart & Chinnery, 2015; Tuppen, Blakely, Turnbull, & Taylor, 2010).

La mitocondrias cuentan con su propio sistema de replicación en el cual están involucradas diferentes proteínas nucleares como la DNA polimerasa γ (Polγ). La eficiencia de la replicación del genoma mitocondrial está determinada por la expresión de las proteínas involucradas y la concentración de los precursores de desoxiribonucleótidos (dNTP), ya que durante este proceso se ha sugerido que las mitocondrias pueden tener un desequilibrio que conduce a una disminución de la fidelidad de Polγ y mayores tasas de mutación (Doudican, Song, Shadel, & Doetsch, 2005).

La preservación de la integridad del genoma mitocondrial durante toda la vida de un organismo se enfrenta con varios desafíos y depende principalmente de su replicación y reparación (Copeland, 2016, Kazak, Reyes, & Holt, 2012). Algunos de los mecanismos conocidos de reparación del ADNmt son la reparación por escisión de bases (BER), la reparación de desajustes (*mismatch repair*), la recombinación homóloga y unión final no homóloga, además de la degradación selectiva, la cual juega también un papel importante en el mantenimiento de su integridad (Shokolenko et al., 2009; Liu et al., 2008).

De este último mecanismo se sabe poco. Sin embargo, se han sugerido dos procesos alternativos, uno que involucra nucleasas que degradan el ADNmt mientras que el otro se asocia con los mecanismos de autofagia y mitofagia (Moretton et al, 2017).

3.2.1 Mutaciones en el ADNmt

La combinación de la exposición continua a las especies reactivas de oxígeno, la falta de histonas protectoras y la baja fidelidad de reparación del genoma mitocondrial se han asociado con la acumulación y generación de mutaciones somáticas (mutaciones puntuales y grandes reordenamientos o deleciones) aunque aún no está claro si existen procesos mutacionales específicos en células somáticas (Seok et al., 2019).

Además de las mutaciones somáticas, las mutaciones del ADNmt de línea germinal pueden heredarse de los ovocitos y dispersarse durante el desarrollo en

el cuerpo en todos o la mayoría de los órganos y tejidos. Las variantes genéticas también pueden surgir durante el desarrollo embrionario temprano que dependiendo de los linajes progenitores afectados, limitan su distribución a órganos o tejidos específicos del cuerpo. En contraste, las mutaciones somáticas del ADNmt de aparición tardía en el desarrollo pueden persistir en un patrón de mosaico en algunas pocas células (Ma et al., 2018).

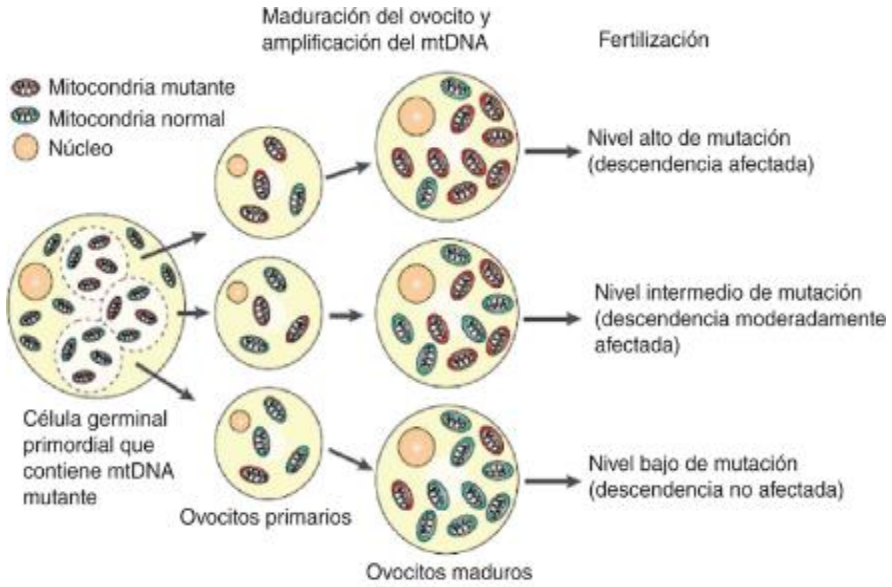

Figura 3. Maduración de los ovocitos y segregación del ADNmt *Durante el proceso de ovogénesis puede existir una distribución sesgada de variantes que se transmiten al embrión (tomado de Skorecki & Behar, 2019).*

La transmisión transgeneracional de variantes del genoma mitocondrial está relacionada con el proceso de ovogénesis. El ovocito primario contiene un número bajo de copias de ADNmt con una distribución distinta a partir de la célula germinal. Posteriormente, durante la maduración de los ovocitos hay un incremento de estas copias mediante un proceso de replicación rápida que puede modificar la carga inicial de variantes genéticas del mismo en cada uno de ellos. Esto genera ovocitos procedentes de una misma hembra que difieren entre ellos en el grado de heteroplasmia. Esta carga puede estar sesgada hacia variantes específicas que serán heredadas por el individuo (Skorecki & Behar, 2019).

Los avances recientes en la tecnología de secuenciación han hecho evidente que la mayoría de los individuos, si no es que todos, tienen niveles bajos de variantes heteroplásmicas (Kappulia et al, 2016). Originalmente se creía que la

heteroplasmia estaba únicamente asociada con estados patológicos, pero en los últimos años se ha demostrado que la heteroplasmia se presenta también en individuos sanos y en estados asintomáticos (Naue et al, 2014). A la fecha, se han descrito cientos de variantes en el ADNmt tanto patogénicas como neutras que en su mayoría son polimórficas (Tuppen et al, 2010).

Se cree que las deleciones o rearreglos grandes del genoma en la mayoría de los casos ocurren de forma espontánea por errores de replicación del ADNmt en el ovocito o en el embrión temprano. También hay condiciones en las que se presentan deleciones en el genoma mitocondrial causadas por mutaciones en los genes nucleares necesarios para la replicación del ADNmt y la síntesis o transporte de nucleótidos. En el caso de ratones, por ejemplo, se ha reportado que la inactivación de mitofusina aumenta los niveles de deleciones en el ADNmt. La cantidad de tales deleciones varía según el tejido o incluso entre las células individuales debido a la deriva aleatoria de moléculas que son eliminadas (Damas et al, 2012; Kappulia et al, 2016).

Las deleciones en el genoma mitocondrial se clasifican en dos categorías, las conocidas como tipo I, que se encuentran flanqueadas por secuencias repetidas homólogas o casi homólogas y las que están flanqueadas por secuencias no repetidas que son conocidas como deleciones tipo II. En ambos casos se cree que su causa es el deslizamiento de la cadena creciente de DNA durante su replicación, aunque los mecanismos exactos de su producción durante el desarrollo no se conocen completamente (Kappulia et al., 2016, Saneto & Russell, 2017). En el ratón y la mosca de la fruta se ha observado la presencia de ADNmt lineal con deleciones que es considerado una subclase distinta de ADNmt en la que los puntos de ruptura generalmente se encuentran en los orígenes de replicación. Se ha sugerido que esta forma alternativa de ADNmt con eliminaciones representa productos fallidos de la replicación que normalmente se degradan (Kappulia et al, 2017).

La presencia de variantes en el genoma mitocondrial puede tener un impacto importante en el fenotipo expresado por el individuo. Las mutaciones en genes codificantes de proteínas afectan la función o la expresión de ese producto

solamente, mientras que las mutaciones encontradas en los ARNt o en los ARNr pueden tener un impacto negativo sobre la traducción de los 13 polipéptidos contenidos en el ADNmt (Duun et al, 2011). En la región de control conocida como *D-loop* también se han detectado mutaciones, se sabe que es una región susceptible a alteraciones y que además contiene dos regiones hipervariables, HV1 en las posiciones 16024-16383 y HV2 en las posiciones 57-372, en las que se han detectado mutaciones asociadas a cánceres humanos (Sharma et al, 2005).

Una de las preguntas más relevantes en cuanto a la biología del genoma mitocondrial y la heteroplasmia es cómo puede una variante acumularse hasta alcanzar el nivel suficiente para causar un defecto funcional. Este proceso se conoce como expansión clonal, el cual puede ocurrir de forma selectiva (la expansión del ADNmt mutado a expensas del silvestre) o de forma neutral (aumento de su abundancia relativa por deriva aleatoria).

Típicamente el ADNmt mutante debe alcanzar un abundancia relativa umbral que varía en diferentes tejidos para causar un efecto deletéreo que oscila entre el 60% y el 70%. Esto se debe a que las variantes del ADNmt presentan un fenómeno de complementación en el que algunas copias de un genotipo variante no provocan una enfermedad en presencia de suficientes copias silvestres que cumplen con la función normal (Stewart & Chinnery, 2015; Herst et al, 2017).

Esto ha aumentado el interés en la caracterización de las mutaciones tempranas y en el proceso de expansión clonal que aún no está claramente definido (Payne & Chinnery, 2015).

3.2.2 Enfermedades asociadas a mutaciones en el genoma mitocondrial

Pese a que el ADNmt sólo codifica aproximadamente el 1% del mitoproteoma, las 13 proteínas que constituyen esa fracción son esenciales para la función mitocondrial. Por ello, sus alteraciones pueden conducir a un deterioro grave de la conversión de energía celular y a la disfunción celular o tisular. Además de las mutaciones en los genes asociados a la FOSFOX, se sabe que cambios en otras

regiones del ADNmt tienen implicaciones en la salud (Kauppila et al, 2017; Chinnery, 2015; Keogh y Chinnery, 2015).

Las alteraciones en el DNA mitocondrial han sido implicadas principalmente en enfermedades relacionadas a tejidos con altas demandas de energía. Estas mutaciones son una causa importante de enfermedades hereditarias que incluyen las mitopatías como la neuropatía óptica de Leber (LHON), el síndrome de epilepsia mioclónica asociado a fibras rojas rasgadas (MERF), la neuropatía atáxica y la retinitis pigmentosa, el síndrome de Kearns-Sayre, el síndrome de Leigh y la encefalomiopatía mitocondrial, acidosis láctica y episodios parecidos a un accidente cerebrovascular (MELAS) en la cual se ha implicado una variante en uno de los genes de ARNt por sustitución de adenina a guanina en la posición 3243 (Piotroswka et al, 2018). Se ha estimado que la prevalencia de este tipo de desórdenes es de aproximadamente de 1 en 5000 individuos, aunque el espectro fenotípico y la gama de enfermedades asociadas con la disfunción mitocondrial se ha expandido considerablemente (Gorman et al, 2015). Por otro lado, se ha demostrado que las alteraciones en el genoma mitocondrial no sólo están relacionadas con enfermedades hereditarias, sino que también pueden alterar el riesgo de padecer otras patologías (Husdon et al, 2014). Las mutaciones puntuales y deleciones en el genoma mitocondrial también han sido implicadas en enfermedades neurodegenerativas, neuromusculares, cardiovasculares, diabetes, trastornos gastrointestinales, de la piel, envejecimiento y cáncer (Piotroswka et al,2018).

En el caso de cáncer se ha reportado que las variantes en la línea germinal pueden desempeñar un papel asociado al crecimiento de tumores hematopoyéticos, de próstata, de mama y el cáncer renal. Asimismo se ha propuesto que mutaciones somáticas están posiblemente involucradas en distintos tipos de cáncer como el cáncer de mama, colorrectal, de vejiga, esofágico, de cabeza y cuello, de ovario, renal, pulmonar, de tiroides y leucemia. El efecto de las mutaciones somáticas de ADNmt en la tumorigénesis depende de los efectos funcionales y de umbral de la variante. Respecto a las variantes asociadas a cáncer se ha observado que más del 50% de las mutaciones se encuentran en

genes que codifican para ARNt (Herst et al, 2017). En cambio, en cánceres como el carcinoma hepatocelular y el de pulmón las mutaciones se encuentran con frecuencia en la región reguladora *D-loop* (Vivian et al, 2018).

En el caso de las enfermedades neurodegenerativas, se sabe que el cerebro tiene altas demandas energéticas para su funcionamiento normal por lo cual es un órgano muy susceptible a la disfunción mitocondrial. Esto se debe a que las variantes en el ADNmt pueden conducir a cambios en la señalización celular y a déficits de vías bioenergéticas que pueden conducir a desarrollar ciertas patologías. En, particular, en individuos afectados por la enfermedad de Alzheimer (AD), Parkinson y esclerosis lateral amiotrófica (ALS), se ha observado un patrón de herencia materna del ADNmt que se asocia con la susceptibilidad a padecerlas (Wilkins et al, 2017).

La principal problemática a la que se enfrenta la investigación de las enfermedades asociadas a mutaciones en el genoma mitocondrial es determinar el papel exacto de las variantes, debido a la variabilidad de sus manifestaciones clínicas, la frecuencia variable con la que se presentan, el mosaicismo somático, así como los desafíos que representa el análisis de múltiples copias del genoma mitocondrial por célula (Bris et al, 2018). Actualmente, además de la caracterización de las variantes en el ADNmt que por sí solas determinan el riesgo de padecer una enfermedad compleja, también se considera de importancia el efecto acumulativo de múltiples variantes en el mismo individuo (Piotrowska et al, 2018).

3.2.3 Envejecimiento

El envejecimiento es un proceso normal, complejo y progresivo que se caracteriza por un deterioro en las funciones de un organismo que eventualmente lleva a la enfermedad y la muerte. Se ha propuesto que es el resultado del daño molecular acumulado asociado con la actividad metabólica regulada por factores genéticos y ambientales, aunque existen diversas teorías al respecto. La teoría clásica involucra el daño que genera la acumulación de mutaciones en el genoma mitocondrial ocasionadas por las EROs en células somáticas lo cual rompe la cadena de electrones creando un ciclo de daño que culmina en la apoptosis. Esto

se ha relacionado con el desarrollo enfermedades de inicio tardío en el humano (Ma et al, 2018; DeBalsi et al, 2016).

Otra teoría es que las mutaciones en ADNmt se acumulan por expansión clonal. Según esta hipótesis el ADNmt contiene mutaciones iniciales ya sean heredadas o *de novo* en algunas células portadoras, las cuales experimentan una expansión de modo que el umbral del genoma mitocondrial alcanza un nivel patogénico con la edad que eventualmente afecta a todo el organismo (Payne & Chinnery, 2015; TA, 2014; Pinto & Moraes,2015).

En humanos, se han reportado variantes somáticas del ADNmt asociadas con el envejecimiento. Tales variantes se han detectado en tejidos posmitóticos como el músculo esquelético y las neuronas, así como en tejidos replicativos. En este proceso también se ha encontrado disminución de la función mitocondrial. Sin embargo, no se ha confirmado que exista una relación causal entre la disfunción mitocondrial y el envejecimiento (Payne & Chinnery, 2015). Otros estudios apoyan la teoría de tejidos tipo mosaico, en los que también se ha demostrado que con la edad hay una disminución de la función de la citocromo oxidasa (Müller-Höcker, 1990; Pinto & Moraes, 2015). Esto sugiere una relación entre la función mitocondrial y el envejecimiento por el aumento progresivo a lo largo de la vida del individuo de un grupo limitado de variantes iniciales causando los cambios fisiológicos asociados con la edad (Payne & Chinnery, 2015).

3.2.4 Detección de variantes

El desarrollo de la secuenciación de nueva generación (NGS) ha aumentado la capacidad de detectar mutaciones en el genoma mitocondrial mediante el análisis sistemático de todo el genoma mitocondrial con un consecuente aumento en la sensibilidad de detección. Sin embargo, pese a que se ha mejorado mucho el diagnóstico de trastornos mitocondriales y la identificación de variantes, se requieren herramientas bioinformáticas específicas, que tomen en cuenta la biología y las características estructurales del genoma mitocondrial (Bris et al, 2018).

Previo al desarrollo de la tecnología de NGS, el diagnóstico e investigación de trastornos mitocondriales se basaba en el uso y combinación de otras técnicas

y herramientas para la detección y cuantificación de los niveles de heteroplasmia. Entre estos métodos destacan la secuenciación dirigida Sanger para la detección de variantes y la transferencia Southern, para la detección de rearreglos o depleción del ADNmt (Bris et al, 2018). Otras técnicas utilizadas incluyen el MitoChip v.2.0 de Affymetrix que contiene sondas complementarias a la secuencia de referencia del genoma mitocondrial y el análisis PCR-RFLP (Ye et al, 2014).

Todos estos métodos tienen desventajas dado que dependen de la búsqueda dirigida de mutaciones específicas en regiones definidas a priori, el nivel de heteroplasmia no se puede cuantificar con precisión y los métodos son demasiado complejos para ser aplicable a un gran número de muestras. Además, la eficacia de la detección puede variar sustancialmente de un laboratorio a otro, incluso cuando se aplica el mismo método, como resultado del uso de diferentes instrumentos, reactivos o estándares para determinar heteroplasmia (Li et al, 2013).

Por otro lado, la mayoría de los estudios del genoma mitocondrial se han dirigido a las regiones codificantes. Sin embargo, en la región de control y los segmentos hipervariables es difícil determinar la heteroplasmia pese a que es probable que estas regiones sean de mayor importancia como causa de algunos padecimientos. Por lo tanto, se necesitan métodos más precisos y eficientes para el análisis y cuantificación de la heteroplasmia en todo el genoma mitocondrial sin conocimiento previo de las regiones afectadas por mutaciones (Li et al, 2013).

Además, distinguir la heteroplasmia en una muestra de ADNmt puede presentar otros desafíos que complican la interpretación de los resultados como la contaminación por secuencias mitocondriales nucleares (NUMTs) y errores de secuenciación (Just et al, 2015). Otra consideración en el análisis de la heteroplasmia es que todos los métodos presentan diferentes niveles de detección por lo cual aún no se ha definido un umbral consenso. Estudios previos han establecido que el límite de detección oscila entre 1 y 10% según la tecnología de NGS utilizada. Sin embargo, debido a algunas de las características intrínsecas de estos estudios existen algunos errores sistemáticos que aún no se pueden evitar (Li et al, 2013).

Una vez obtenidos los datos de secuenciación es necesario analizar y procesar la información de forma adecuada. El procesamiento de los datos obtenidos se puede dividir en tres etapas: a) evaluación y control de calidad de los datos crudos, b) alineamiento de secuencias y c) llamado o identificación de variantes. En cada una de estas etapas es necesario el monitoreo de las métricas de control de calidad (Guo et al, 2013). La primera etapa consiste en una evaluación de los datos crudos para analizar la calidad de la secuenciación y las lecturas; el siguiente paso, el alineamiento, es la yuxtaposición de las secuencias de la muestra contra el genoma de referencia para identificar cambios en la primera, mientras que el último paso se refiere al conteo e identificación de variantes en base a los resultados del alineamiento (Li et al, 2013).

Uno de los enfoques clásicos para el estudio del genoma mitocondrial consiste en la secuenciación por NGS de extractos de DNA total seguida de alineamiento con la secuencia de referencia. En esta técnica es necesario implementar un paso para detectar NUMTS que pueden alterar los resultados como el número de variantes o la profundidad de lectura. Se han propuesto para ello distintas alternativas y la más aceptada consiste en realizar un primer alineamiento con el genoma de referencia nuclear y filtrar aquellas lecturas que se alinean para luego hacerlo con el genoma mitocondrial (Ye et al, 2014). Este proceso, sin embargo, resulta en la pérdida de lecturas de las regiones del ADNmt colineares con los NUMTs con la consecuente pérdida de capacidad de detección de variantes en tales regiones. El potencial de un método para detectar la heteroplasmia depende en gran medida de la profundidad de la cobertura de la secuenciación, es decir, el número de veces que una región es secuenciada. Los dos requisitos para detectar niveles de heteroplasmia por debajo del 1% son una profundidad relativamente alta y tasas de error de secuenciación reducidas. Al respecto de esto, se ha determinado que el nivel más bajo de heteroplasmia detectable se define como 1/D (en el que D corresponde al valor de la profundidad de secuenciación). Sin embargo, a medida que disminuye el umbral detectable de heteroplasmia, se vuelve cada vez más difícil distinguir entre la heteroplasmia verdadera y los errores de secuenciación (Guo et al, 2012; Ye et al, 2014).

Adicionalmente, un tipo de error que puede ser inducido por el método de análisis es el de PCR que ocurre por una amplificación sesgada. Este tipo es de los más difíciles de identificar y prevenir. Otra fuente de errores es por la plataforma de secuenciación que puede ser reducida aplicando un filtro de calidad estricto en el llamado de bases (Yiru Guo et al., 2012).

4. ANTECEDENTES

El genoma mitocondrial de ratón está compuesto por 16299 pb y al igual que el del humano codifica 13 proteínas, dos ARNr (16S y 12S) y 22 ARNt. La secuenciación de distintas cepas de ratón ha demostrado que ratones de líneas consanguíneas presentan polimorfismos de un solo nucleótido (SNP) que son únicos entre cepas. Esto ha permitido identificar haplotipos diferentes entre 16 cepas distintas de ratones (Wallace & Chalkia,2013; Goios et al, 2007). Por otro lado, la distribución de los marcos de lectura abiertos en el genoma mitocondrial de ratón es similar a la del humano, ya que, de las 13 proteínas codificadas, 12 están codificadas en la cadena pesada y una en la cadena ligera. La región control del *D-loop* tiene un tamaño aproximado de 800 nucleótidos y contiene los promotores para la transcripción tanto de la cadena H como L y el origen de replicación de la cadena H (Pogozelski et al, 2008).

Pese a que el ADNmt del resto de los mamíferos tiene similitudes con el del humano, sus variantes difieren fenotípica y genotípicamente. Sin embargo, el análisis del ADNmt en modelos animales es muy útil para comprender su biología y los mecanismos de patogenicidad de sus variantes (Dunn et al, 2012). Por ello, el ratón es un modelo conveniente para investigar los mecanismos moleculares subyacentes al envejecimiento y a la biología del genoma mitocondrial de los mamíferos, aunque existan diferencias importantes con el del humano en cuanto a propiedades fisiológicas, a los mecanismos de patogenia y la historia de vida a nivel poblacional (Demetrius, 2006; Kazachkova et al, 2013).

Como antecedente directo de este trabajo, en nuestro laboratorio se desarrolló el proyecto de investigación de Rueda (2017) cuyo objetivo fue detectar deleciones de baja frecuencia en el genoma mitocondrial humano empleando un método de extracción de ADN circular por lisis alcalina y digestión enzimática del ADN lineal remanente. En este trabajo se identificaron y caracterizaron distintos tipos de deleciones en distintas regiones del genoma mitocondrial siendo las más comunes deleciones de un solo nucleótido.

Además, se determinó que al menos un 41.3% del genoma mitocondrial contenía alguna deleción. Otra observación importante fue que las deleciones se distribuyen en todo el genoma mitocondrial, aunque se encontraron deleciones específicas de mayor frecuencia en la región del *D-loop* en las posiciones 521-524, 5750-5752 y 2462 en distintos niveles de heteroplasmia. Estos experimentos sugieren que a todo lo largo del genoma mitocondrial de individuos sanos existe una gran variabilidad y abundancia de múltiples deleciones de baja frecuencia.

En el estudio reciente (J. Ma, Purcell, Showalter, & Aagaard, 2015), se analizó la variación del genoma mitocondrial en humanos en comparaciones madre-hijo. El objetivo de dicho trabajo fue identificar si las variantes en la progenie se heredan por vía materna, ocurren en el útero durante la gestación o en etapas tempranas de la vida y si se acumulan o se generan a lo largo de la vida. Por ello, el propósito fue examinar la tasa de mutación heredable de la madre y las mutaciones de *novo* en el DNA fetal a través de secuenciación de alta profundidad. Para realizar el estudio se tomaron muestras de 90 pares materno-fetales de sangre y placenta respectivamente. En este estudio se localizaron 665 SNP´s y 82 variantes de inserción-deleción (indels), con una alta profundidad de secuenciación. En este caso el menor grado de heteroplasmia que se pudo identificar fue mayor a 1%. Además, una de las particularidades de este estudio fue que la mayor cantidad de mutaciones que se encontraron fueron heredables por vía materna, sin embargo, se encontraron también dos variaciones en el *D-loop* y en CO2.

Las variantes encontradas se filtraron para análisis posteriores para garantizar la alta calidad de los SNP´s y filtrar posibles errores de secuenciación, por lo que sólo las variantes con una frecuencia superior al 1% se incluyeron. En este estudio la mayor cantidad de mutaciones de novo que se encontraron fueron mutaciones sinónimas, es decir que no causan un cambio en la secuencia de la proteína codificada. Sin embargo, se propuso que algunas de las variantes identificadas no sinónimas pueden estar relacionadas con algunos padecimientos.

En los dos casos anteriores las variantes se identificaron mediante el uso de distintas herramientas diseñadas por cada grupo. Sin embargo, existen

herramientas especificas con este propósito, aunque generalmente están diseñadas para genoma nuclear como SomaticSniper (Larson *et al*, 2012), gatk (Mckenna *et al*, 2010), JoinSNVmix (Roth *et al,* 2012) y Strelka (Saunders *et al*, 2012). En contraste, para el análisis del genoma mitocondrial son pocas las herramientas específicas como lo es Mitoseek (Yan Guo, Li, Li, Shyr, & Samuels, 2013). Existen además otros programas como MuTect2 (Cibulskis et al., 2013) que fue creado con la intención de detectar variantes en cáncer. Recientemente, sin embargo, se añadió a este programa un modo mitocondrial, que toma en cuenta la características inherentes del genoma mitocondrial para su análisis. Con este trabajo aportamos una descripción y caracterización de la distribución de variantes a lo largo de todo el genoma mitocondrial empleando un método integral que incluyó el desarrollo de una técnica experimental para su extracción directa y el uso de herramientas bioinformáticas adecuadas para su análisis con la finalidad de disminuir posibles fuentes de error.

5. HIPÓTESIS

El genoma mitocondrial contiene una alta variabilidad de variantes de alta y baja frecuencia que en conjunto generan una alta abundancia relativa en individuos sanos.

6. OBJETIVOS

Objetivo general:

- Determinar la abundancia de diversos tipos de variantes en el genoma mitocondrial de ratón en estadios embrionarios y adultos.

Objetivos específicos:

- Determinar el estado inicial de las secuencias de genoma mitocondrial de cerebro embrionario, de adulto joven, de adulto envejecido y de secuencias control para su análisis posterior.

- Obtener un catálogo de diversos tipos de variantes y su frecuencia en el genoma mitocondrial de ratón en diferentes etapas de la vida.

- Establecer comparaciones de las variantes encontradas y su frecuencia en las muestras de ratón en distintas etapas de la vida con diferentes herramientas.

7. JUSTIFICACIÓN

El genoma mitocondrial ha sido ampliamente estudiado. A finales de 1980 se hizo el primer hallazgo de una variante asociada con una enfermedad y la fecha se han descubierto cientos de variantes, de las cuales se especula que podrían estar involucradas en una gran variedad de enfermedades como mitopatías, neurodegeración, cáncer, diabetes y enfermedades neuromusculares (Bereiter-Hahn, 2014; Keogh & Chinnery, 2015; Wilkins, Weidling, Ji, & Swerdlow, 2017).

El impacto de las variantes en el genoma mitocondrial y su proporción en la salud humana aún continúa en estudio. Se ha observado que las variantes en condición de homoplasmia se relacionan principalmente con las mitopatías y daños acumulados en un solo tejido. En cambio, la heteroplasmia de distintas variantes en diferentes niveles se ha correlacionado con disfunción mitocondrial, el desarrollo de enfermedades y la predisposición a ciertos padecimientos (Bereiter-Hahn, 2014; Ma, Purcell, Showalter, & Aagaard, 2015; Stewart & Chinnery, 2015; Tuppen, Blakely, Turnbull, & Taylor, 2010).

Dado lo anterior, se ha sugerido que el nivel de la carga mutacional en el genoma mitocondrial se asocia directamente con el grado de severidad del fenotipo clínico presentado y defecto bioquímico que afecta al órgano o tejido involucrado. Esto es generalmente severo en tejidos como el cerebro, el corazón, los órganos endocrinos, los nervios periféricos y el músculo. Por ello, la detección de los niveles de heteroplasmia se ha vuelto muy importante para comprender la biología del ADNmt tanto en salud como en enfermedad (Chinnery & Turnbull, 2001; Stewart & Chinnery, 2015).

Pese a la gran cantidad de variantes reportadas sólo 84 tienen un estado confirmado de patogenicidad en humanos, mientras que se desconoce si el resto de un total de 595 tiene un efecto en la salud (Bris et al., 2018). Una de las principales limitaciones para estos estudios son las características intrínsecas del ADNmt, el gran número de copias y la variabilidad que presenta dificultan la detección, cuantificación y caracterización de los niveles bajos de heteroplasmia.

Estas dificultades técnicas han limitado la capacidad de comprender cómo ocurren estas mutaciones, cómo los niveles de heteroplasmia aumentan, cómo afectan el organismo y en qué frecuencia estas variantes generan un fenotipo específico (Elliott, Samuels, Eden, Relton, & Chinnery, 2008; Ma et al., 2015; Stewart & Chinnery, 2015). Además, aunque se ha propuesto que la generación y acumulación de variantes en el genoma mitocondrial se asocia al envejecimiento y el estrés oxidativo, existen estudios que contradicen esta idea.

Por lo anterior, el propósito de este trabajo fue analizar la carga mutacional en el genoma mitocondrial en cerebros de ratón de distintas edades para identificar variantes y sus frecuencias en los estadios de vida de embrión, adulto joven y adulto envejecido.

Para ello fue necesario desarrollar un procedimiento integral con un enfoque experimental y bioinformático que permitiera la detección y cuantificación de variantes en cualquier región del genoma mitocondrial. Tal procedimiento incluyó distintas herramientas diseñadas para el alineamiento de secuencias y el llamado de variantes incluyendo un prototipo desarrollado para este estudio.

8. MATERIALES Y MÉTODOS

8.1 Muestras biológicas

Se seleccionaron individuos de distintas edades de la cepa de ratones C57BL/6J (Tabla1). En el caso de los embriones se obtuvieron individuos de 15.5 días de gestación (E15.5), los adultos jóvenes fueron de 2 a 4 meses y por último los adultos envejecidos fueron de 21 a 24 meses. Las muestras control consistieron en una mezcla equimolar de seis plásmidos (A-F), cada uno conteniendo un fragmento de ADNmt de ratón de la cepa de ratón CD1 de aproximadamente 3000 pb (Tabla 2).

Tabla 1. Muestras utilizadas para el análisis del genoma mitocondrial del ratón de tres grupos de edad.

Etapa	Muestra	Edad
Adulto Viejo	B1	23 meses
	B2	22 meses
	B3	22 meses
	E1	24 mese
	E2	21 meses
Adulto Joven	B4	4 meses
	E4	2 meses
	E5	2 meses
	E6	2 meses
	E7	2 meses
Embrión	D6	15.5 días
	D9	15.5 días
	D10	15.5 días
	D11	15.5 días
	D12	15.5 días
Clonado	C1	Mezcla A-F
	C2	Mezcla A-F
	C3	Mezcla A-F
	C4	Mezcla A-F
	C5	Mezcla A-F

El número de individuos se seleccionó de forma empírica de acuerdo a la cantidad de DNA necesaria para el procedimiento experimental. Para embriones se utilizaron 35 individuos por muestra, mientras que para jóvenes y adultos envejecidos se utilizó un individuo por muestra.

Tabla 2. Fragmentos clonados del genoma mitocondrial de la cepa CD1 empleados como control. *En la tabla se muestran los fragmentos de ADN mitocondrial utilizados en el método A para el análisis y la comparación con las muestras.*

Fragmento	Coordenada de inicio	Coordenada de término
A	87	3363
B	2875	6034
C	5572	9000
D	8287	11427
E	10834	14098
F	13862	904

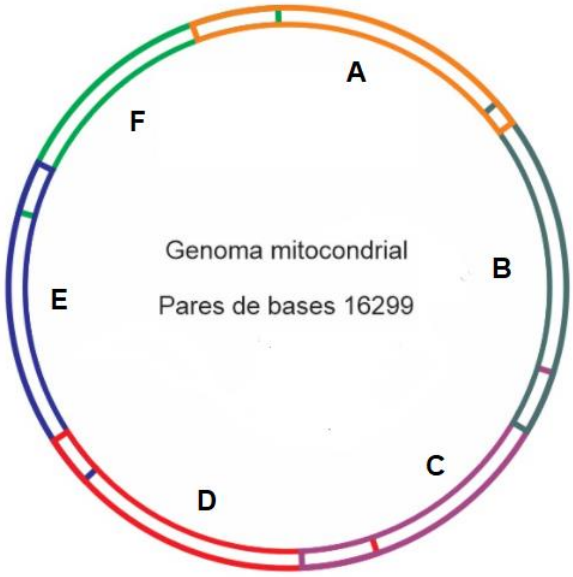

Figura 4. Diagrama de la ubicación de los fragmentos del ADN mitocondrial*. En la figura se muestran los fragmentos del genoma mitocondrial mostrados en la tabla 2. Cada fragmento está representado con un color distinto.*

8.2 Métodos

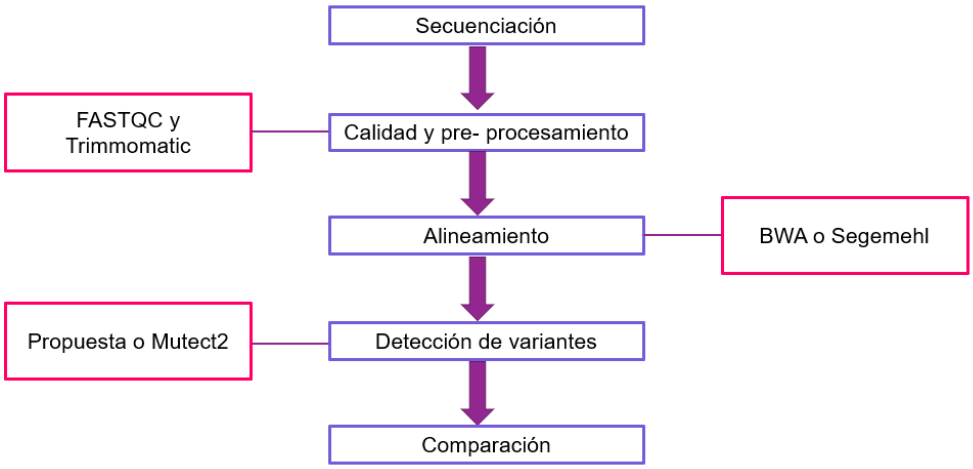

Figura 5. Diagrama de flujo general de trabajo para la detección de variantes en el genoma mitocondrial.

8.2.1 Obtención de muestras

Las muestras se obtuvieron de cerebros de ratón de distintas edades que se procesaron para extraer ADN enriquecido en moléculas circulares. Para ello se utilizó una adaptación al método de lisis basado usando el "Quiafliter Plasmid Purification Midi Kit" de Qiagen* (Rueda, 2017). Posteriormente, para eliminar remanentes de DNA linear provenientes del genoma nuclear se utilizó una exonucleasa dependiente de ATP (Plasmid Safe, Epicentre. Madison, WI) con la que se llevó a cabo la digestión del eluido obtenido de la columna de extracción. Una reacción de digestión típica contenía 25 µg de ADN y 600 unidades de enzima en un volumen total de 900 µl de buffer de reacción 1X. Esta mezcla se incubó por 14 horas a 37ºC seguido de una segunda digestión por 6 horas después de añadir 170 µl de buffer de reacción 1X conteniendo 100 unidades de enzima. Después de la digestión, el ADN circular fue limpiado por una extracción de fenol-cloroformo, precipitado con acetato de sodio y etanol, resuspendido en agua conteniendo 0.5% SDS, extraído nuevamente con fenol-cloroformo y precipitado con acetato de sodio y etanol, para ser resuspendido finalmente en agua.

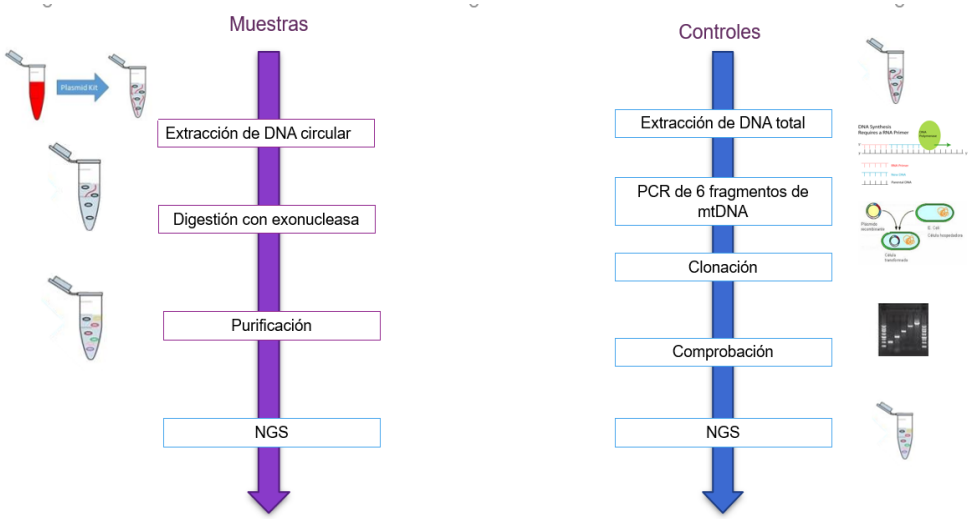

Figura 6. Procesamiento de muestras biológicas y controles clonados.

8.2.2 Secuenciación masiva

La secuenciación del ADNmt de cada una de las muestras se realizó mediante la plataforma Illumina Miseq en modo de lecturas pareadas para la generación de la biblioteca de secuenciación. Se generaron dos lecturas (R1 y R2) por cada fragmento el cual se seleccionó de entre 500 – 1000 nucleótidos, utilizando 300 ciclos con la finalidad de generar lecturas de un tamaño aproximado de 300 nucleótidos.

8.2.3 Análisis bioinformático de secuencias

8.2.3.1 Calidad de datos crudos

La calidad de los datos obtenidos de la secuenciación se comprobó con el uso del paquete FastQC (High Throughput Sequence QC Report, Version 0.11.5, Andrews, Lindenbaum, Howard, Ewels 2011-15) que permite evaluar la calidad de secuenciación mediante el puntaje Phread que corresponde a la probabilidad de de llamado correcto de base (Bacterial Pangenomics, 2012; Erwin 1998). En este paso se identifican problemas que se generan ya sea en el secuenciador o en el material de la biblioteca de inicio y se genera un reporte con las estadísticas y gráficas de los datos las cuales incluyen la calidad secuencia por base, por contenido de GC, por contenido de Ns y adaptadores, entre otras que

proporcionan una visión general del estado de los datos y si es necesario hacer ajustes.

Para realizar el ajuste de las secuencias obtenidas se utilizó Trimmomatic, una herramienta que permite el recorte, filtrado de lecturas, identificación y eliminación de adaptadores en una ruta de trabajo. Entre sus algoritmos cuenta con el filtrado de calidad, el cual se puede llevar a cabo mediante 2 métodos; uno realiza el escaneo de la secuencia y recurre a la eliminación del extremo 3´ si la calidad de la lectura cae por debajo del promedio del grupo y otra mediante la retención de bases adicionales en el extremo 5'. Estos enfoques toman en cuenta el puntaje de calidad para cada posición, acorde a los datos típicos generados por la plataforma Illumina los cuales generalmente tienen una calidad inferior en el extremo 3' (Bolger, Lohse, & Usadel, 2014). En este trabajo se aplicaron distintos parámetros da calidad y pre-procesamiento a los datos con la finalidad de identificar las implicaciones sobre el número de variantes cuantificadas.

8.1.1.1 Alineamiento de secuencias

Otro paso para el procesamiento de los datos de NGS es el alineamiento, que en parte determina si un experimento ha tenido éxito. La selección de un algoritmo para el alineamiento de secuencias está influenciada tanto por el experimento como por la tecnología de secuenciación utilizada. Para la selección de algoritmos también debe tomarse en cuenta el tamaño de las lecturas (Flicked & Birney, 2009).

En este proyecto las secuencias se mapearon al genoma de referencia mitocondrial de ratón C57BL/6J (GRCm38/mm10) obtenido del Genome Browser de la Universidad de California Santa Cruz (https://genome.ucsc.edu/cgi-bin/hgGateway?hgsid=804062049_AhXteqD9D7nyoptbA4SASO1i35Fa) con Segemehl (Hoffman et al, 2009). Una de las características principales de este software de alineamiento de secuencias es que su algoritmo toma en consideración parámetros para el mapeo de moléculas circulares de DNA. Se

realizaron mapeos pareados y sencillos utilizando distintas condiciones de filtrado y ajuste de calidad de las secuencias (Hoffman et al, 2009).

Con la finalidad de comparar los resultados de esta herramienta se utilizó el software BWA (Li & Durbin, 2010) que es un paquete utilizado en múltiples estudios tanto de ADN nuclear como de ADNmt para mapear secuencias contra el genoma de referencia. Para este estudio se utilizó el algoritmo BWA-MEM el cual se recomienda para análisis de alta calidad, ya que es más rápido y preciso. BWA-MEM también tiene mejor rendimiento para lecturas de Illumina de 70-100bp.

Para evaluar los resultados de ambos alineamientos se utilizó SAMtools/1.9, que es una biblioteca de utilidades para el análisis de archivos de alineamiento que permite visualizar estadísticas y gráficos de las secuencias alineadas, así como para ordenar, indizar y calcular la profundidad de lecturas por posición (Li et al, 2009). Además, se empleó bamstats con la cual se calcularon y se mostraron de forma gráfica diversas métricas derivadas de archivos de SAM / BAM del alineamiento. Con esto se obtuvieron gráficos de contenido de GC y del tamaño del inserto, entre otras que proporcionan información sobre el mapeo.

En esta parte del análisis también se utilizó el Integrative Genomics Viewer (IGV) (Robinson et al, 2011), un software de visualización de grandes volúmenes de información, que es utilizado para la exploración de genomas de los conjuntos de datos de secuenciación de nueva generación (NGS). IGV varía el nivel de detalle mostrado según la escala de resolución y fue útil para evaluar la calidad general y diagnosticar problemas técnicos en las ejecuciones de secuencia. Además, la visualización de los alineamientos con esta herramienta permitió observar cambios en la secuencia como SNPs o rearreglos del genoma tales como deleciones y el tamaño de las lecturas.

8.1.1.2 Detección y cuantificación de variantes

La identificación de variantes se realizó mediante distintos programas con el objetivo de evaluar la sensibilidad de la detección. Debido a que la finalidad es la

identificación de la heteroplasmia, la sensibilidad y la precisión son factores fundamentales.

Uno de los métodos utilizados (**Método A**) consiste en un programa prototipo diseñado por Janeth de Anda Gil como una colaboración con el grupo de la Dra. Maribel Hernández (Instituto de Matemáticas, UNAM) con el objetivo de cuantificar e identificar variantes en el genoma mitocondrial. Este programa parte de los archivos SAM resultantes del mapeo con Segemehl e identifica las variantes por comparación con el genoma de referencia. Posteriormente normaliza la frecuencia de cada variante tomando en cuenta la profundidad en cada posición del genoma, lo que de forma indirecta genera que aquellas variantes que se encuentran en una sola lectura tiendan a una frecuencia aproximada a 0, ya que es probable que éstas sean un error. El programa toma en cuenta los controles clonados de ADNmt estableciendo la comparación entre muestras biológicas y controles. Para esto utiliza una prueba estadística pareada no paramétrica Wilcoxon y verifica la significancia de los datos con un valor de p (p-value) menor a 0.05 y determina si existen diferencias significativas entre grupos.

El Método A utiliza distintos filtros para descartar falsos positivos de las variantes cuantificadas inicialmente, que son: que existan diferencias estadísticamente significativas entre la muestra y el control y que la frecuencia del control no exceda la frecuencia de la muestra. En general, el análisis consiste en detectar las variantes utilizando los controles clonados en plásmidos como una población de moléculas semejantes las cuales son sometidas al mismo procedimiento que las muestras biológicas con fines comparativos. El propósito de esta comparación consiste en descartar variantes o ruido introducido por la técnica utilizada por lo que el programa toma en cuenta si las variantes encontradas son de mayor frecuencia en la muestra o en el control. Este método, favorece la detección de variantes presentes en varias muestras de una edad y no detecta variación individual.

El conteo de las variantes se realizó utilizando el archivo procedente del mapeo de secuencias en formato SAM que contiene información sobre el mapeo

de lecturas frente a secuencias de referencia. El archivo SAM se divide en dos secciones: una sección de encabezado y una sección de alineación. Uno de los campos de información que contiene este formato es la cadena CIGAR que describe cómo la lectura se alinea con la referencia y se estructura de uno o más componentes. Cada componente comprende un operador y el número de bases a las que se aplica el operador. Utilizando esta cadena se contabiliza el número de cambios.

Ya que el Método A consiste en un programa de prueba, se requiere comparar los resultados obtenidos utilizado herramientas ya establecidas para el llamado de variantes presentes en el ADNmt. Para esto se utilizó Mutect2 (https://github.com/broadinstitute/gatk/blob/master/scripts/mitochondria_m2_wdl/MitochondriaPipeline.wdl) que para los fines del presente proyecto corresponde al **Método B**. Este método está especializado en la identificación de alelos de baja frecuencia ya que se centra en la suposición de que las mitocondrias presentan el fenómeno de heteroplasmia o fracción de alelo variable. El programa considera la estructura circular del genoma mitocondrial tomando en cuenta que el genoma de referencia está generalmente linealizado de forma artificial ubicando el punto de ruptura en la región de control *D-loop* altamente variable por lo que aumenta la sensibilidad a la variación en esta región con dicha consideración. Además de lo anterior, Mutect2 considera la presencia de las inserciones del ADNmt en el genoma autosómico (NUMTs) que pueden interferir en el análisis. Para esto utiliza un método de realineación y marcaje que aumenta la precisión e identifica este tipo de secuencias para excluirlas.

El **Método B** utiliza un modelo bayesiano como algoritmo para identificar variantes. El teorema de Bayes expresa la probabilidad condicional de un evento aleatorio A dado B en términos de la distribución de probabilidad condicional del evento B dado A. Este método toma en cuenta la bases que cubren un sitio, la tasa de error de secuenciación y calcula un puntaje logarítmico en donde los sitios con puntaje más alto tienen más probabilidades de expresar una variante. Además

de las consideraciones anteriores, el Método B cuenta con una gama amplia de opciones y filtros de ajuste.

Predicción de estructura secundaria de ARN de transferencia

Para predecir la estructura secundaria de ARNt se empleó la herramienta tRNAscan-SE (http://lowelab.ucsc.edu/tRNAscan-SE/).

9. RESULTADOS

9.1 Secuenciación

Mediante el método de secuenciación utilizado se obtuvieron lecturas pareadas de cada muestra con un tamaño de máximo de 301 pb.

Tabla 3. Longitud de lectura.

Etapa	Muestra	Edad	Longitud Phread >20	
			R1	**R2**
Adulto envejecido	B1	23 meses	285	210
	B2	22 meses	285	210
	B3	22 meses	285	185
	E1m	24 meses	289	210
	E2m	21 meses	285	210
Adulto Joven	B4	4 meses	285	214
	E4	2 meses	285	210
	E5m	2 meses	289	200
	E6m	2 meses	285	210
	E7m	2 meses	285	200
Embrión	D6	15.5 días	285	185
	D9	15.5 días	285	185
	D10	15.5 días	285	210
	D11	15.5 días	285	189
	D12	15.5 días	285	214
Clonado	C1	A-F cloned mix	285	185
	C2	A-F cloned mix	285	185
	C3	A-F cloned mix	285	185
	C4	A-F cloned mix	285	164
	C5	A-F cloned mix	285	185

9.1.1 Calidad de datos crudos

La evaluación de la calidad de los datos crudos se realizó utilizando la herramienta FastQC. Esta herramienta produce una serie de gráficos con la finalidad de identificar problemas que se generan ya sea durante la secuenciación o en el material de la biblioteca de inicio. Una de las características de este programa es que produce un archivo html que proporciona una visión general del estado de los datos, así como una descripción de la muestra.

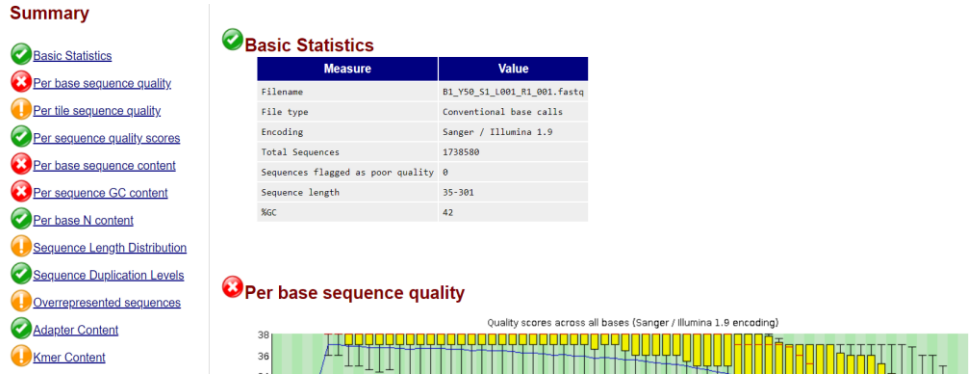

Figura 7. Estadísticas básicas y resumen gráfico de calidad. *Como parte del reporte de la muestra B1, se presenta un resumen indicando si la muestra cumple o no con los parámetros de calidad, con la finalidad de visualizar si es necesario realizar un ajuste.*

En el caso de las muestras de este estudio, la mayoría no pasó el filtro de calidad de secuencia por base, que se refiere a la probabilidad de que la base reportada sea incorrecta en cualquiera de las posiciones. Ésta generalmente falla si la mediana del Phread de las lecturas en esa posición para cualquier base es menor a 20, lo que indicaría que 1 de cada 100 bases es incorrecta en las lecturas teniendo una precisión del 99%.

Con respecto a el contenido de GC, se muestra un ejemplo y se observa que este módulo no cumple con los requerimientos necesarios para pasar los estándares de calidad aplicados (Figura 7). Este módulo falla si la suma de las desviaciones de la distribución normal de GC representa más del 30% de las lecturas.

Por último, el contenido de bases de la secuencia y otros módulos presentaron advertencias como lo muestra la Figura 6. Con base en estos puntos se analizaron los gráficos de los módulos fallidos para realizar ajustes.

El primer gráfico que se genera por FastQC es el de calidad de secuencia (eje Y) por cada posición (eje X) del conjunto de todas las lecturas de una muestra. Esta gráfica de cajas y bigotes representa los puntajes de calidad Phread. Ésta contiene además otros elementos descriptivos de calidad: la línea roja representa el valor de la mediana, mientras que la línea azul representa la media. El fondo del gráfico a su vez se presenta en tres colores (verde, naranja y

rojo) que ilustran la calidad de la base desde muy buena, buena y baja respectivamente dependiendo del valor de Phread de las lecturas en esa posición.

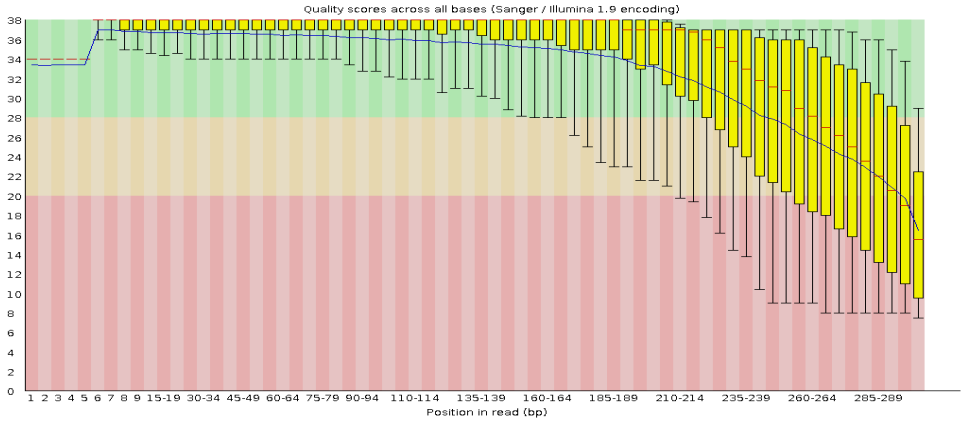

Figura 8. Calidad de secuencia por base B1R1*. Se presenta la gráfica de cajas y bigotes (Boxwhisker) de la muestra B1 para las lecturas R1. En el grafico se observa que las lecturas tienen una longitud aproximada de 289, la calidad decae en la sección roja a partir aproximadamente 301 bases. Esto es normal para la mayoría de las plataformas de secuenciación a medida que aumenta el número de ciclos, tal es el caso de Illumina.*

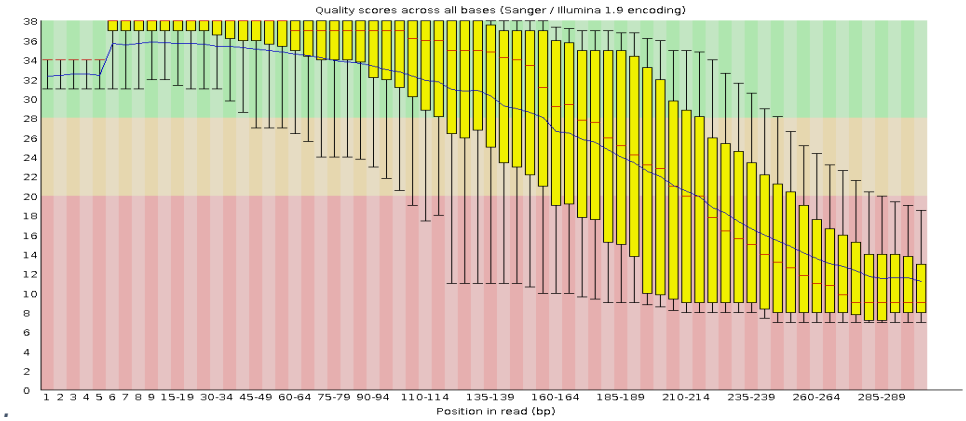

Figura 9. Calidad de secuencia por base B1R2*. Se observa que pese a que su longitud máxima igualmente 301 bases, los puntajes de calidad son menores para la muestra B1. Se observa que a partir de la posición 189 la mediana de la calidad cae a la sección roja correspondiente a baja calidad.*

Otro de los rubros que presentó un fallo fue el contenido de bases por secuencia. Esto determina la proporción de cada una de las cuatro bases de DNA normales en cada posición del conjunto de lecturas. En una biblioteca aleatoria, se espera poca o ninguna diferencia entre las bases de una ejecución de secuencia, por lo que las líneas de esta gráfica deberían correr paralelas entre sí. Las advertencias o errores en este módulo se presentan si la diferencia entre A y T, o G y C excede al 10- 20%, respectivamente. Las muestras presentaron un fallo en este módulo ya que se reportó un sesgo positivo hacia ciertos nucleótidos (T y A)

con respecto a los nucleótidos restantes, en las primeras 15 posiciones y en las últimas como se muestra en la Figura 10.

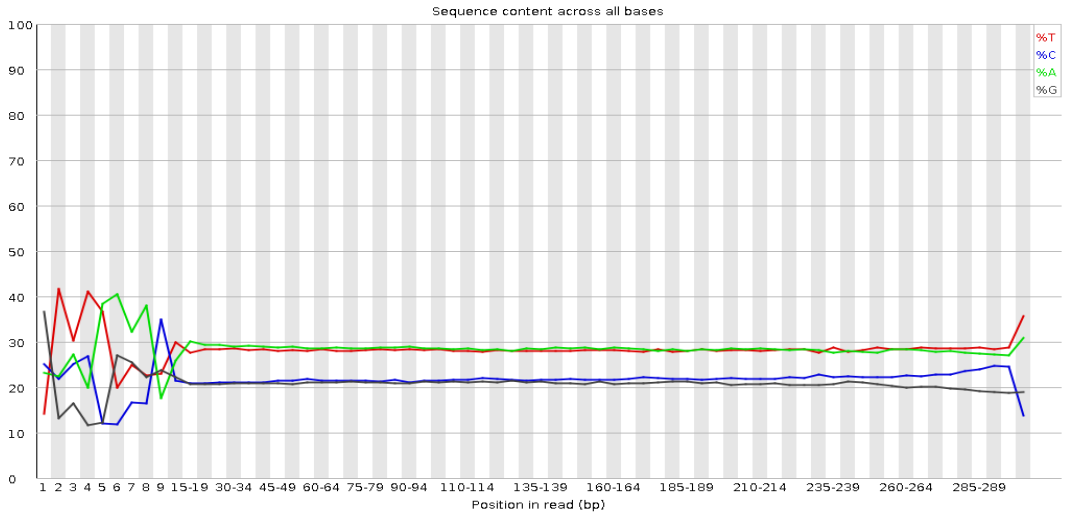

Figura 10. Contenido de bases por secuencia B1R1. *Se ilustra el contenido de bases por posición para la muestra B1. Idealmente se esperan líneas paralelas si no hay un sesgo hacia una o más bases específicas. Se observa un desequilibrio en las primeras 15 bases y en el extremo final de las lecturas. La línea roja representa Timina, la azul citosina, la verde adenina y la gris guanina.*

Respecto al contenido de GC de las secuencias, generalmente se espera una distribución normal. Los datos obtenidos presentaron una distribución inusual ligeramente desplazada de la estimación teórica propuesta lo cual indica una biblioteca contaminada o algún otro tipo de subconjunto sesgado. Es decir, un sesgo sistemático que es independiente de la posición base. Regularmente el fallo de este módulo está asociado con contaminaciones ya sea por adaptadores, secuencias sobrerrepresentadas o contaminación de una especie diferente que se calcula como la sumatoria de las desviaciones estándar de la distribución normal.

Como parte de la información obtenida mediante el reporte de FastQC se encontró que, los datos de todas las muestras no mostraron presencia de adaptadores, así como un bajo contenido de N, sin embargo, el módulo de longitud de la secuencia en todas las muestras presentó un tamaño de fragmento variable.

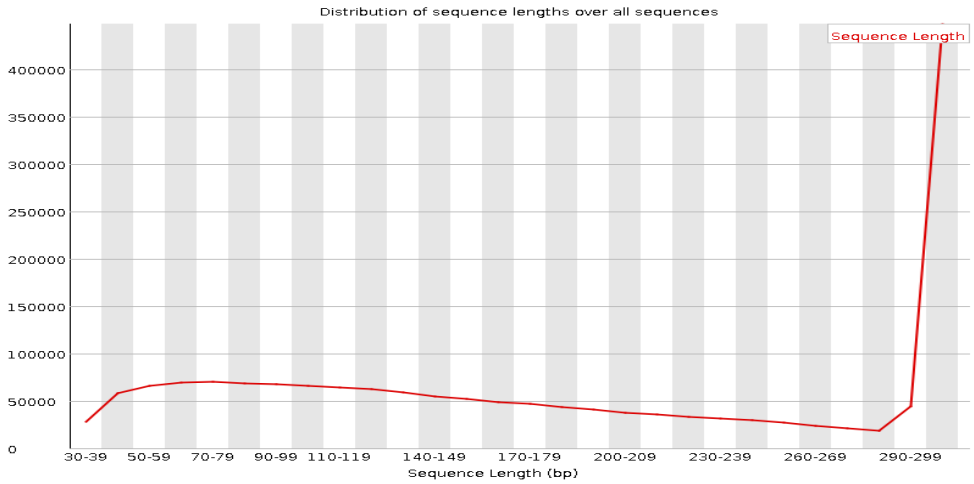

Figura 11. Longitud de secuencia. *Se observa que en esta muestra (B1) hay lecturas que van desde los 30 nucleótidos hasta 300 lo cual muestra que hay una longitud variable de un rango muy amplio.*

9.1.2 Ajuste y calidad de los datos

Posterior al análisis de calidad de los datos crudos, se utilizó el programa Trimmomatic para realizar recortes de secuencia y filtrar calidades en base a los resultados presentados en la sección anterior. Se decidió aplicar las siguientes correcciones:

- Recorte de las primeras 15 bases de cada lectura
- Recorte para eliminar bases al final de la secuencia con calidad menor a 20

Con estos ajustes, se analizó nuevamente la calidad de las secuencias post-procesamiento con FastQC. Los resultados de las correcciones aplicadas no mostraron módulos fallidos. Cabe mencionar que los ajustes se realizaron por separado para R1 y R2 (lecturas pareadas) de acuerdo con las características de cada set de lecturas y sus calidades con la finalidad de conservar la mayor cantidad de información posible de buena calidad.

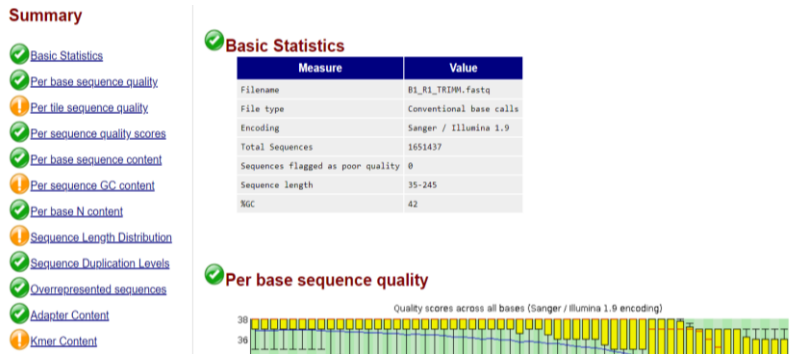

Figura 12. Reporte de calidad post-procesamiento. *Aplicando las correcciones previamente mencionadas a la muestra, se observa que se eliminan los módulos fallidos y mejora la calidad de las muestras.*

Al comparar la Figura 8 con la Figura 13 se observa que el tamaño de las lecturas disminuye a 240 pb por los recortes realizados. Sin embargo, las medianas de todos los cuartiles se encuentran en la región verde que indica muy buena calidad con valores de Phread mayores a 28. De la misma manera, al comparar la Figura 9 correspondiente a la calidad por base en la secuencia de las lecturas R2 con la Figura 14 que corresponde al mismo análisis después de las correcciones aplicadas, se observa el mismo comportamiento ya que disminuye el tamaño de las lecturas hasta 180 bases, así mismo la mediana de los puntajes de calidad (Phread) muestra solamente valores mayores a 28 puntos.

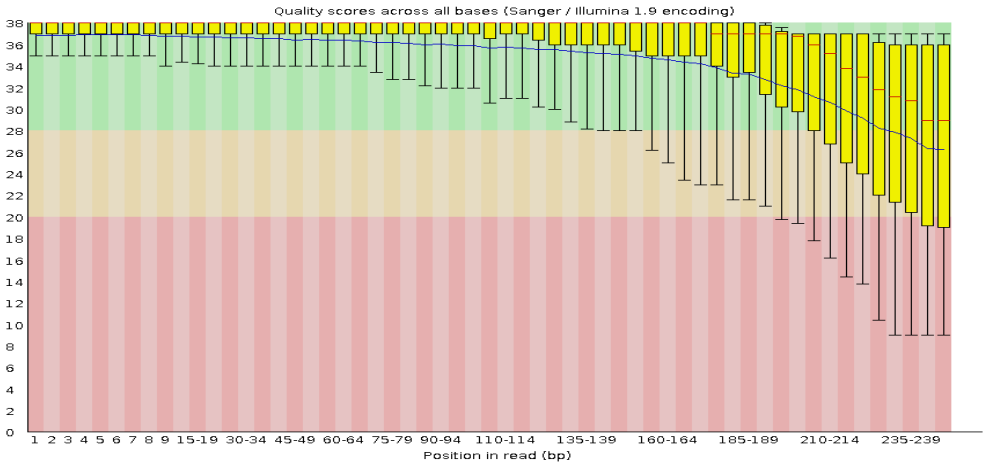

Figura 13. Calidad de secuencia por base B1R1. *Calidad de los datos después del procesamiento con Trimmomatic para la muestra B1 lecturas R1 (forward).*

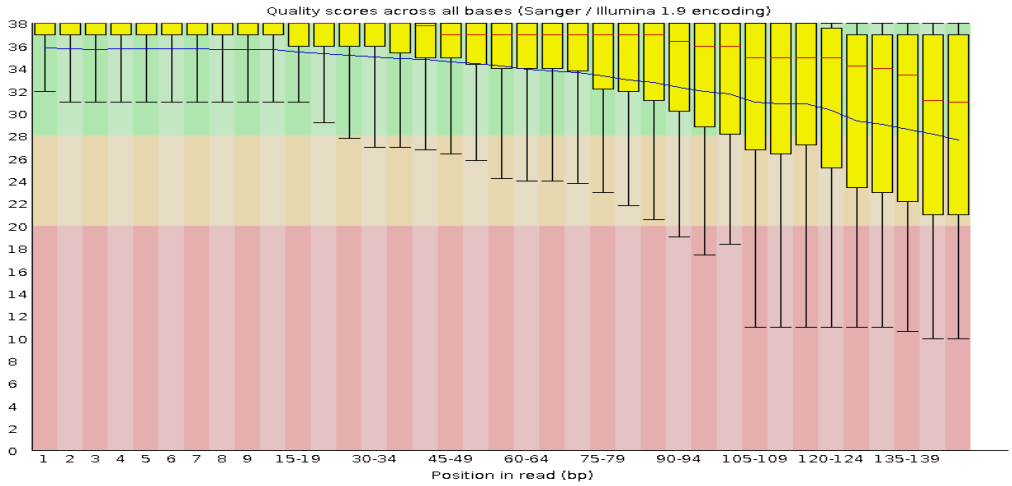

Figura 14. Calidad de secuencia por base B1R2*. Calidad de los datos correspondientes a la lectura R2 (reverse) de la muestra B1 después del procesamiento con Trimmomatic****.*

Como parte de la comparación y la comprobación de los cambios realizados mediante los ajustes aplicados a las muestras, se observó que se eliminaron las regiones problemáticas al inicio y al final de las lecturas obtenidas como se muestra en la Figura 15 que muestra un contenido de bases uniforme a lo largo de las lecturas.

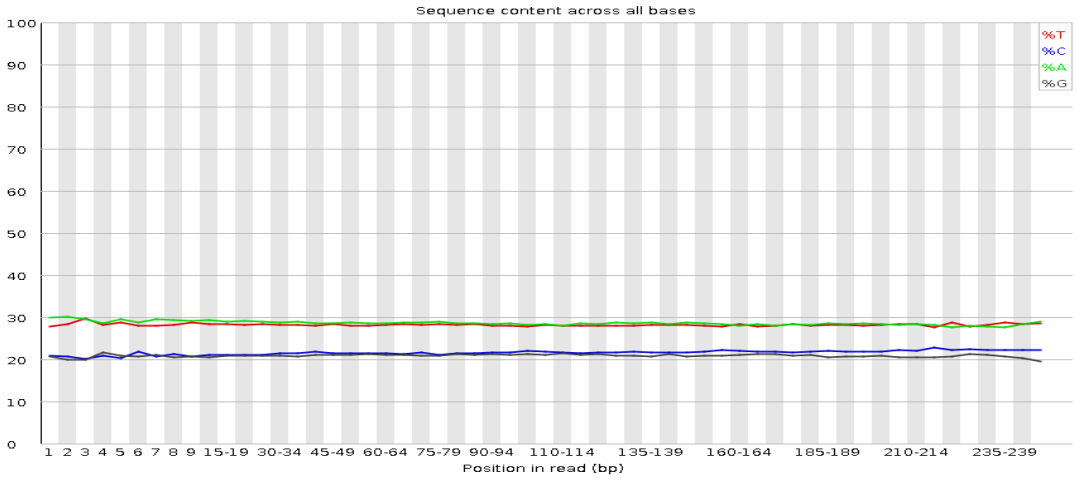

Figura 15. Contenido de base post-procesamiento de muestra B1. *Luego de los ajustes realizados con Trimmomatic se observa que se han eliminado las zonas de las lecturas donde había desequilibrio de bases. Las zonas en las que había sesgos en a la lectura hacia alguna de las bases, se han eliminado. La línea roja representa timina, la verde adenina, la azul citosina y la negra guanina.*

El mapeo a un genoma de referencia es un requisito previo para la mayoría de los flujos de trabajo de secuenciación NGS, ya que a partir del alineamiento de secuencias se obtiene el formato de archivo SAM/BAM el cual contiene la información de todas las lecturas de la muestra. Se realizaron alineamientos con

lecturas pareadas y alineamientos con las lecturas utilizadas como “single-end”. Al visualizar los resultados de ambos mapeos con IGV se observaron tamaños de lectura variables y lecturas sobrelapadas en las muestras pareadas (Figura 16).

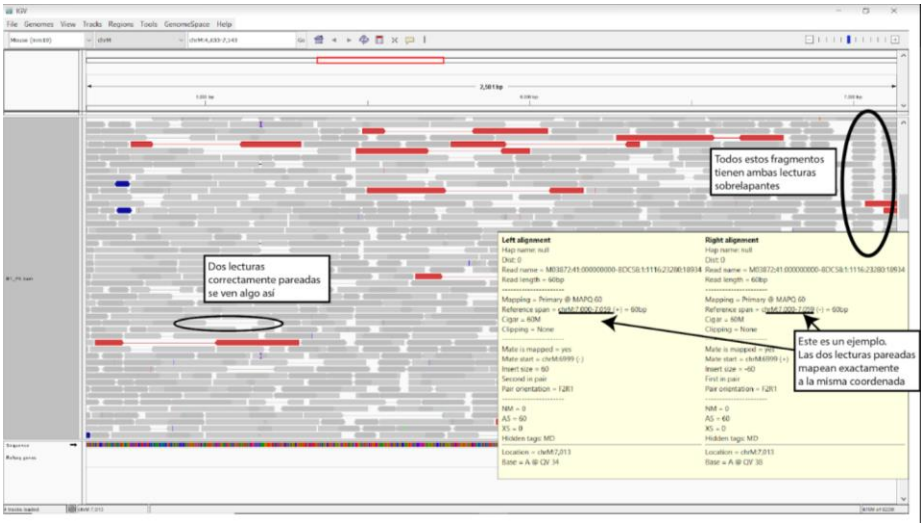

Figura 16. Visualización de lecturas pareadas con IGV. *Se muestra una parte del mapeo resultado de la muestra B1 alineada como pair-end. Las lecturas en rojo representan pares muy alejados, las gris ocuro y azules representan lecturas 100% sobrelapadas.*

Esto se corroboró haciendo un análisis del porcentaje de sobrelape de las lecturas R1 y R2 y se encontró que la mayoría de las lecturas analizadas tenían de un 50 -100% de sobrelape en la misma muestra, es decir que mapean exactamente en la misma zona. También se utilizó BAMstats que es en una herramienta de software que permite la evaluación gráfica de varias métricas de calidad de los archivos SAM/BAM. Con este programa se detectó que había un tamaño de inserto variable en un rango de 40pb a 500pb, en la mayoría de las muestras. En la Figura 17 se encontró nuevamente representado el sobrelape ya que se muestra que hay una gran cantidad de mapeos “outwards”.

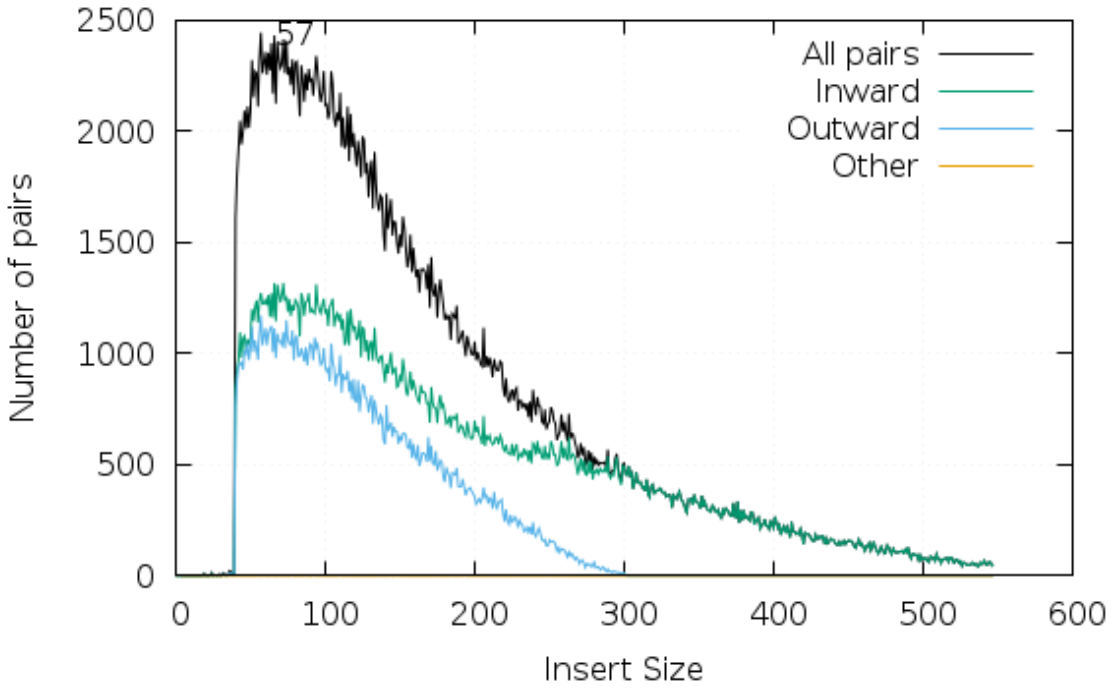

Figura 17. Talla del inserto de la muestra B1 pareada. *Mediante BAMstats se obtuvo una gráfica del tamaño del inserto, esta grafica se obtuvo haciendo una prueba pareada, en la que se aprecia que la mayoría de las lecturas tienen un tamaño de 57 pares de bases.*

Con este programa, al igual que con el uso de Samtools, se determinó la profundidad por posición en el genoma mitocondrial y se encontró una profundidad heterogénea, que variaba en el rango desde 300X hasta 12000X para todas las muestras.

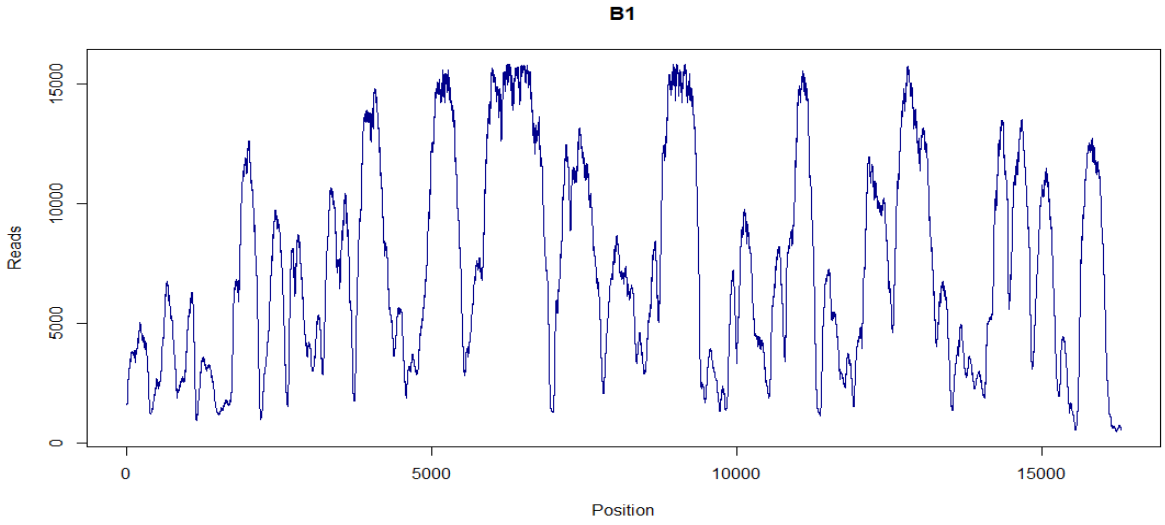

Figura 18. Profundidad de la muestra B1*. Se observa una profundidad variable, este comportamiento con patrones similares se observó en todas las muestras.*

9.2 Cuantificación e identificación de variantes

La identificación de variantes se realizó mediante dos métodos (Método A y Método B) con la finalidad de evaluar el Método A para el análisis del genoma mitocondrial y determinar si realmente era efectiva en la detección de variantes de baja heteroplasmia. Además, se buscó determinar su alcance y definir las consideraciones necesarias para el análisis de este tipo de muestras.

9.2.2 Método A

9.2.2.1 Distribución de variantes

Con este método se detectaron alrededor de 1628 variantes distintas entre deleciones, inserciones y sustituciones de base. Como se menciona en la sección anterior. éstas corresponden únicamente a cambios de pocos nucleótidos sin tomar en cuenta rearreglos grandes del genoma por las condiciones de los datos de secuenciación.

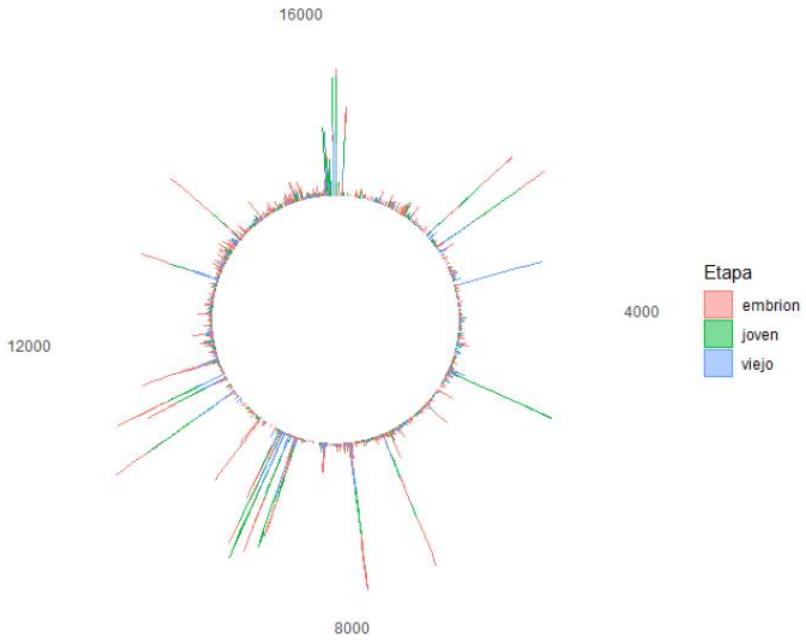

***Figura 19. Distribución y frecuencia**. Distribución de las variantes localizadas en los tres grupos de muestras analizadas. El color rojo representa a la muestra de embrión, el verde al joven y el azul al cerebro de ratón envejecido.*

Las variantes encontradas con este método se encuentran distribuías a lo largo de todo el genoma mitocondrial (Figura 19). Sin embargo, se aprecian zonas de mayor acumulación y recurrencia de mutaciones. Las variantes identificadas mediante este método se encuentran en un rango de frecuencias de entre 1 a 0.0000029. Además, se observó que una gran cantidad de ellas correspondían a variantes únicas de frecuencias muy bajas que oscilaban en rangos menores a 0.00001 en cada grupo de muestras.

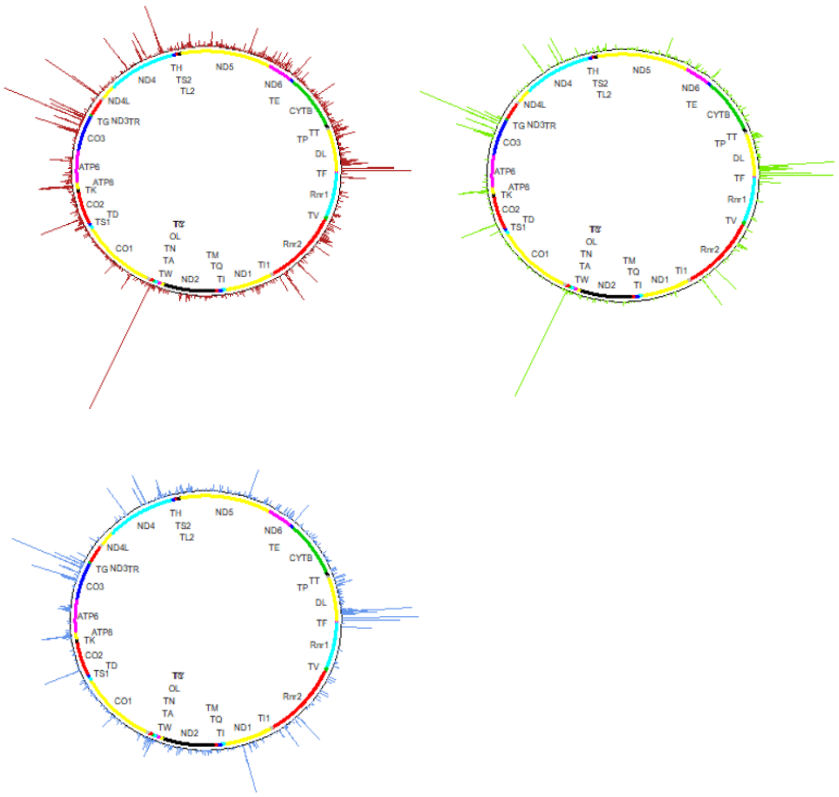

Figura 20. Distribución de variantes por región. *Los gráficos circulares presentan las variantes por región en el genoma mitocondrial. El circulo corresponde a los 16299 pb de longitud del genoma, las líneas salientes representan las variantes por posición y la altura la frecuencia de la variante, siendo la más alta 0.016. Donde el rojo representa los embriones, el verde los jóvenes y el azul los adultos envejecidos.*

En la Figura 20 se aprecia que existe un sesgo hacia ciertas regiones funcionales de genoma mitocondrial. Pese a que se distribuyen a todo lo largo de su extensión, se encontró que las regiones donde se acumula un mayor número de variantes corresponden a COIII, *D-loop*, ND4, Rnr2, ND5, ATPase8 y en algunos casos al O_L.

9.2.2.2 Carga total de variantes

Con los resultados anteriores se calculó la sumatoria total de las frecuencias de todas las variantes normalizadas encontradas por grupo (embrión, adulto joven y adulto envejecido), es decir el conjunto de variantes con sustituciones de bases, deleciones e inserciones (X,D,I)(Tabla 4A). Para el cálculo de la carga total, se tomaron en cuenta en la primera columna todas las variantes incluyendo SNPs y variantes de alta frecuencia, mientras que la columna de baja frecuencia considera únicamente todas aquellas que están por debajo de 0.05. Esto se hizo para separar las variantes de alta heteroplasmia que pueden incluir diferencias entre el genoma mitocondrial de la colonia empleada con el genoma de referencia.

Tabla 4. Carga mutacional total por muestra.

Muestra	Frecuencia acumulada total	**A:** Sin aplicar filtro 1/D		B: Aplicando el filtro 1/D	
		Número de mutaciones acumuladas	Frecuencia acumulada de variantes de Baja frecuencia	Número de mutaciones acumuladas	Frecuencia acumulada de variantes de Baja frecuencia
Embrión	1.94977	979	0.552624	793	0.52496
Adulto Joven	1.28353	380	0.271672	268	0.26185
Adulto viejo	134411	564	0.340263	389	0.30931

En la tabla se representa la media de la sumatoria de la frecuencia de las variantes identificadas en el genoma mitocondrial de cada muestra, así como la media del número de variantes encontradas en cada condición. La columna de "Todas las variantes" representa la carga mutacional previa al filtrado de las variantes de alta frecuencia (>0.05) presentadas en la sección anterior tomando las consideraciones estadísticas pertinentes y el filtro 1/D para eliminar variantes cuya frecuencia es más baja que el límite de detección confiable comúnmente utilizado en este tipo de estudios en función de la profundidad de secuenciación.

Cuando se aplicó el filtro para eliminar variantes con frecuencia menor al límite de detección 1/D (Tabla 4B), se encontró que del 26 al 51% de los genomas secuenciados por cada grupo de muestras presentaron una mutación de baja frecuencia siendo el grupo de embriones el que tuvo el mayor número de variantes y de mayor frecuencia acumulada, mientras que los adultos jóvenes y envejecidos presentaron un número menor de variantes y carga acumulada similar entre ellos. En los análisis subsecuentes de los resultados de este método se consideraron solamente las variantes de frecuencia mayor al valor de ajuste 1/D.

Es importante notar que, para el cálculo de la sumatoria de frecuencias por muestra, se asumió que cada mutación se presenta de manera independiente a otras en regiones distantes del genoma mitocondrial dado que el tipo de secuenciación utilizado no permite determinar la coexistencia de más de una mutación por genoma secuenciado.

Cuando se analizaron las variantes de acuerdo a su tipo, observamos que las sustituciones corresponden al más abundante con un rango de70-80% de la carga total de variantes en cada grupo de edad.

Tabla 5. Carga parcial de variantes de baja frecuencia.

Muestra	Tipo de variante	Frecuencia acumulada	No. de mutaciones acumuladas	No. De mutaciones acumulas C/filtro 1/D	Frecuencia acumulada C/filtro 1/D
Embrión	X	0.482911	767	675	0.46553
	I	0.0311646	90	48	0.027004
	D	0.0385489	122	70	0.032428
Adulto Joven	X	0.230007	253	177	0.17267
	I	0.014834	52	42	0.04802
	D	0.026831	75	49	0.04115
Adulto envejecido	X	0.304515	421	338	0.28405
	I	0.0229878	75	23	0.01723
	D	0.0126172	67	28	0.00802

Se enlista las carga y número de variantes identificadas por muestra para cada tipo de variante respecto al genoma de referencia, en la que X= Sustituciones, I= Inserciones y D= Deleciones.

Con lo anterior se determinó que los tres grupos de edad estudiados tienen una alta carga de variantes de baja frecuencia siendo el embrionario el que tiene la mayor carga y que las variantes más comunes son sustituciones de base.

9.2.2.3 Identificación y filtrado de variantes de alta frecuencia

En otro análisis se identificaron únicamente variantes con una frecuencia mayor al 0.4 (es decir al 40%), tomando en cuenta que una de las variantes se encontró en una frecuencia del 43% mientras que otras dos se encontraron en una frecuencia mayor al 90% siendo estas en conjunto las variantes con más alta incidencia identificadas con este método. Las cuales se identificaron y las tres corresponden a diferencias entre la colonia estudiada y el genoma de referencia.

Tabla 6. Variantes de alta frecuencia.

Posición	Tipo	Secuencia / Referencia	Variante	Frecuencia en Muestras	Frecuencia en Control
9348	X	G	A	0.98 (Todas)	0
9829	I	*	A	0.43 (Embrión)	0.039
9461	X	T	C	0.99 (Joven)	0.98

Se representan las variantes que se identificaron en las muestras con una frecuencia mayor 0.4. El tipo representa la clase de variante localizada en el que X= Sustitución, I= Inserción y D= Deleción (En este caso la frecuencia se presenta como una mediana obtenida por el programa A que analiza cada muestra y como resultado se obtiene la mediana de las variantes en esa posición).

La variante localizada en la posición 9348 se identificó en las tres muestras analizadas con una frecuencia aproximada de 0.98, es decir en el 98% de los genomas secuenciados, para los tres casos, mientras que en el control no se encontró esta variante. En la posición 9829 se encontró una inserción de una adenina, que se identificó en heteroplasmia en el 43% de los casos únicamente en las muestras del grupo de embriones. Por último, en la posición 9461 se halló una variante que corresponde a una sustitución de timina (T) por citosina (C) con frecuencia de 0.993 detectada solamente en adultos jóvenes. Estas variantes se analizaron de forma separada al resto de las encontradas que se hallaron por debajo del 2% de frecuencia (0.02).

Las mutaciones previamente descritas se buscaron en la base de datos de la UCSC en el genoma de referencia de ratón con la finalidad de identificar las regiones afectadas y las implicaciones de los cambios detectados. También se verificó si las variantes de alta frecuencia habían sido reportadas previamente.

Tabla 7. Localización e implicación de las variantes de alta frecuencia

Posición	Tipo	Referencia	Secuencia	Localización	Original	Mutado
9348	X	G	A	COIII	GTC-Val	ATC-Ile
9829	I	*	A	mt-Trna(Arg) adyacente a poliA		
9461	X	T	C	Ultima base primer codón de ND3	ATT- Ile	ATC-Ile

En la tabla se enlistan el consenso de las variantes de alta frecuencia localizadas en uno o más grupos de muestras como se indica en Tabla 6.

El cambio de guanina a adenina en la posición 9348 ha sido previamente reportada en distintos artículos como un polimorfismo encontrado en la cepa C57BL/6J (Ishikawa et al., 2019). La versión G9348 contiene un sitio de reconocimiento para la enzima de restricción AspI que se interrumpe cuando el polimorfismo A9348 está presente. Sin embargo, se ha demostrado que por sí sola no tiene un efecto patógeno (Bayona-Bafaluy et al., 2003; Vivian, Hagedorn, Jensen, Brinker, & Welch, 2018).

La variante identificada en la posición 9829 se ha reportado previamente en el genoma mitocondrial de cerebro de ratón y está ubicada en una región adyacente a una cadena de poliA en el gen del ARNt de Arginina. Se ha observado que la inserción de una adenina en esta zona altamente variable no altera la estructura de este ARNt. Con la finalidad de corroborarlo se utilizó una herramienta de predicción de la estructura secundaria que confirmó que este cambio no tiene implicaciones estructurales (Bayona-Bafaluy et al., 2003; Kiebish & Seyfried, 2005).

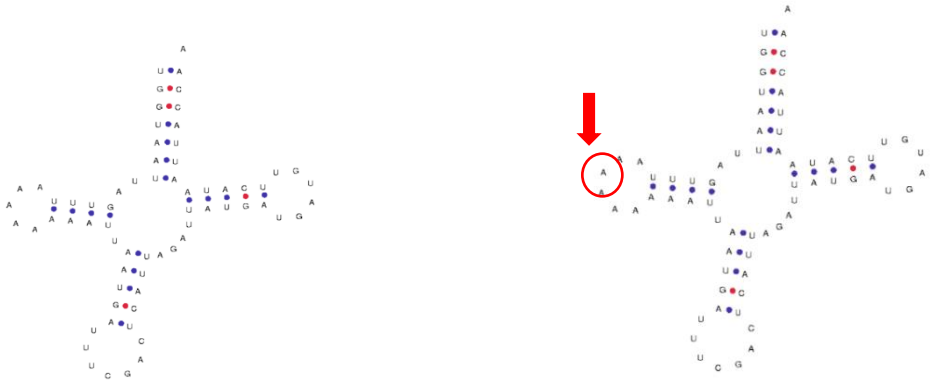

Figura 21. Estructura secundaria de mt-TR. . *La figura de la izquierda representa la estructura sin variante, mientras que la de la derecha representa la estructura con una inserción en la posición 9829.*

Por último, la variante en la posición 9461 que sólo se encuentra en adulto joven es una variante sinónima en el primer codón ND3, que codifica para Isoleucina. Esta variante también se ha reportado en la cepa C57BL/6J y se sabe que al igual que G9348A, interrumpe el sitio de reconocimiento para BclI (Bayona-Bafaluy et al., 2003; Moreno-Loshuertos et al., 2013).

9.2.2.4 Variantes de baja frecuencia: cuantificación y caracterización

La mayor parte de las variantes identificadas con el Método A corresponden a variantes de baja frecuencia. En esta sección se describe primero la carga total de mutaciones en cada grupo de muestras y posteriormente se analizan por diferentes características para clasificarlas con la finalidad de detectar sesgos hacia mutaciones especificas o en regiones puntuales.

9.2.2.5 Variantes de baja frecuencia: Localización y efecto

Al ser el presente un análisis de prueba de la herramienta computacional desarrollada en este trabajo se optó por identificar cuáles eran las variantes de baja frecuencia con mayor recurrencia en cada tipo de muestra y en qué regiones del genoma mitocondrial se encontraron. De entre ellas, en las Tablas 8, 9 y 10 se muestran las diez variantes de mayor frecuencia en cada uno de los grupos de edad estudiadas. Las regiones afectadas por las variantes de estos grupos fueron las regiones regulatorias O_L y el *D-loop,* las regiones codificantes para la citocromo oxidasa subinidad 3 (COIII), NADH deshidrogenasa subunidad 4 (ND4) y NADH deshidrogenasa subunidad 1 (ND1) y los ARNt para fenilalanina y arginina. Las regiones que tuvieron el mayor número de variantes fueron el *D-loop* y COIII. De las variantes en regiones codificantes, la mayoría implica cambios en codones.

Tabla 8. Variantes de baja frecuencia en el embrión.

Posición	Tipo	Referencia	Variante	Frecuencia	Región	Codón en referencia	Codón en variante
5182	D(1)	*	*	0.0163	OL		
9237	X	G	T	0.0094	COIII	GGA-GLY	TGA-End
1	I	*	ACAA	0.0076	RNAt (Phe)		
9821	X	*	A	0.0073	mt-tRNA de Arginina poli A		
11371	X	T	G	0.0058	ND4		
16254	X	G	T	0.0057	D-loop		
16252	X	G	T	0.0057	D-loop		
9238	X	G	T	0.0050	COIII	GGA-GLY	GTA-VAL
8995	X	C	A	0.0048	COIII	CCA -Phe	CAA-Gln
7129	X	C	A	0.0048	COII	CTC- Leu	CTA-Leu

Variantes de baja frecuencia más comunes en la muestra de embrión, la tabla indica las características de la variante, la frecuencia y la región en la que se ubica.

***Tabla 9. Variantes identificadas con el método en muestras de ratón joven**.*

Posición	Tipo	Referencia	Variante	Frecuencia	Región	Codón en referencia	Codón en variante
5182	D	*	*	0.0136	OL		
9237	X	G	T	0.0094	COIII	GGA-Gly	TGA- End
16115	X	G	T	0.0082	D-loop		
1	I	*	ACAA	0.0080	RNAt (Phe)		
16254	X	G	T	0.0056	D-loop		
8995	X	C	A	0.0043	COIII	CCA -Phe	CAA- Gln
7129	X	C	A	0.0041	COII	CTC- Leu	CTA- Leu
10624	X	C	A	0.0039	ND4	ACC- Thr	AAC- Asn
9299	X	T	G	0.0039	COIII	CAT- His	CAG- Gln
9239	X	A	C	0.0038	COIII	GGA- GLY	GGC - GLY

Caracterización y localización de las diez variantes de baja frecuencia más comunes en la muestra de cerebro de ratón joven.

Tabla 10. Variantes de baja frecuencia más recurrentes.

Posición	Tipo	Referencia	Secuencia	Frecuencia	Localización	Codón en referencia	Codón en variante
16254	X	G	T	0.0084	D-loop		
9237	X	G	T	0.0082	COIII	GGA- GLY	TGA- End
16132	X	CT	AC	0.0076	D-loop		
1	I	*	ACAA	0.0063	mt-tRNA (Phe)		
3365	X	A	G	0.0055	ND1	TCA- Ser	TCG- Ser
8995	X	C	A	0.051	COIII	CCA- Phe	CAA- Gln
7129	X	C	A	0.0041	COII	CTC- Leu	CTA- Leu
9238	X	G	T	0.0039	COIII	GGA- Gly	GTA- VAL
16159	X	A	C	0.0038	D-loop		
10624	X	C	A	0.0037	ND4	ACC- Thr	AAC- Asn

En la tabla se muestran las diez variantes más recurrentes encontradas con el programa de prueba en las muestras de ratón adulto viejo.

Del total de 1628 variantes de baja frecuencia, se detectaron 89 que se comparten por los tres grupos, 63 que se comparten entre embriones y adultos envejecidos, 36 entre adultos jóvenes y embriones y 18 entre adultos jóvenes y envejecidos (Tabla 11). Esto indica que la mayoría de las variantes localizadas corresponden a variantes presentes en una sola edad.

Tabla 11. Frecuencia acumulada de las variantes encontradas por grupo de comparación.

Número de mutaciones	Embriones	Adultos jóvenes	Adultos viejos
89	0.168455	0.15415602	0.15125546
36	0.03338334	0.02678387	-
63	0.03914066	-	0.03121627
18	-	0.00656084	0.00570536
791	0.31164492	-	-
237	-	0.08417064	-
394	-	-	0.1520856
	0.5526	0.2716	0.3402

La tabla indica la media de la sumatoria de las frecuencias de las variantes encontradas por cada comparación por grupo de muestras

9.2.3 Método B para la identificación de variantes

9.2.3.1 Distribución de mutaciones en genoma mitocondrial

Mediante el Método B se detectaron 41 sitios del genoma mitocondrial con 54 tipos de cambios distintos en las muestras de los tres grupos de edad en estudio. Estas variantes se presentan a lo largo de todo el genoma mitocondrial, sin embargo, existen sitios de mayor recurrencia. En la Figura 22 se observa que en la región correspondiente a las coordenadas de 9000-10000 hay recurrencia de distintas variantes en diferente frecuencia algunas de las cuales tuvieron frecuencias cercanas a la homoplasmia lo que corresponde a diferencias en secuencia con el genoma de referencia. En esta región también se observan variantes con frecuencia de alelo variable que van desde 0.001 hasta las ya mencionadas de 1.00.

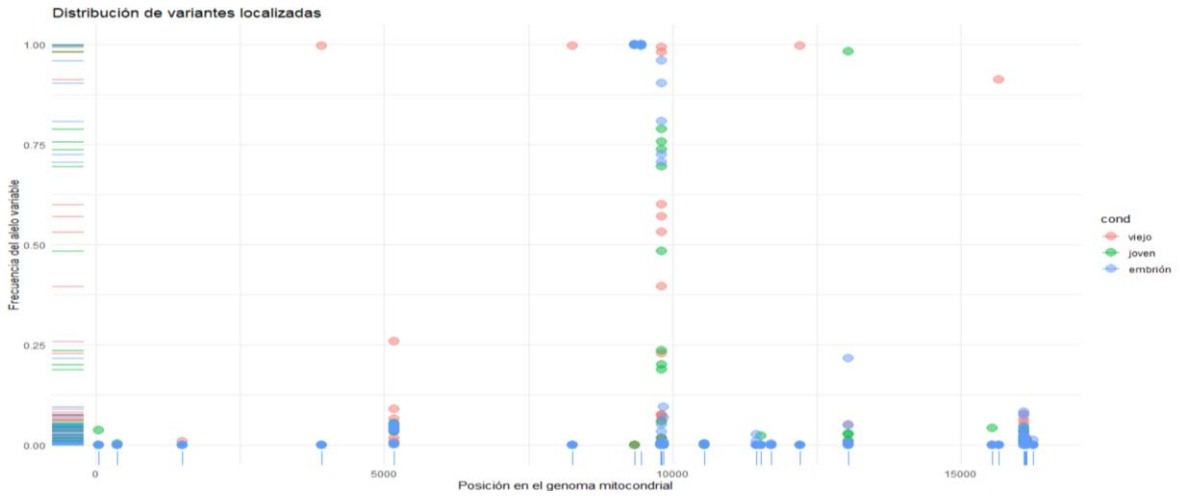

Figura 22. Distribución de variantes en el genoma mitocondrial*. La longitud del genoma mitocondrial de ratón (16299 pb) se encuentra representada en el eje x, mientras que el eje y representa la frecuencia del alelo variable para cada mutación encontrada en al menos una de las tres condiciones estudiadas.*

Figura 23. Distribución de variantes en el genoma mitocondrial. *En las gráficas circulares se presenta la longitud del ADN mitocondrial y la frecuencia de variantes localizadas en cada grupo.*

En cambio, en otras regiones no se encontraron mutaciones o presentaron variantes únicas en una sola condición. Las regiones como el *D-loop*, ND3, y ND5 presentaron variantes en todas las condiciones, pero de baja frecuencia.

9.2.3.2 Carga total de mutaciones

En la sección de la Carga total de variantes obtenidas con el Método A se indica que en el proyecto se asume que cada variante provine de una copia del mtDNA distinta. Para estimar la carga total de variantes se realizó una sumatoria de las frecuencia de alelo variable para cada grupo (embrión, joven y viejo), con la finalidad de detectar la fracción de copias del genoma afectado por variantes en cada muestra con respecto al número total de genomas secuenciados (Tabla 12).

Tabla 12. Promedio de la carga mutacional en cada grupo.

Condición	Todas las variantes	Alta heteroplasmia sin SNP	Baja heteroplasmia (Variantes <0.4)
Embrión	3.1179174	1.1189192	0.2977192
Joven	3.00416089	1.20576086	0.31616086
Viejo	3.575392	1.977092	0.4600494

En la primera columna se presenta cada grupo por edad, la segunda columna representa la sumatoria de frecuencia de todas las variantes incluyendo polimorfismos identificados en la cepa, la columna de alta heteroplasmia corresponde a las variantes con frecuencia mayor a 0.5 pero que no han sido reportadas como polimorfismos. La última columna corresponde a la sumatoria de la frecuencia de todas las variantes con una baja incidencia (0.5>x).

En la Tabla 13 se muestran la variantes de alta y baja frecuencia para cada muestra de los distintos grupos de edad. Se denominan como variantes de alta frecuencia a aquellas que se encuentran al menos en el 40% de los genomas analizados, sin contar los polimorfismos propios de la cepa, mientras que el resto son variantes de baja frecuencia En la segunda columna se muestran todas las variantes identificadas en todas las condiciones incluyendo los polimorfismos detectados previamente en la cepa C57BL/6J. En las siguientes columnas se muestra la carga de variantes de alta y baja frecuencia que son mutaciones nuevas y que no corresponden a polimorfismos propios la cepa. Estos resultados revelan una alta abundancia de variantes de alta y baja frecuencia. Separando las de alta frecuencia que corresponden a variantes polimórficas en homoplasmia, la carga acumulada de variantes de baja frecuencia se observó en un rango de 30 a 46%. Aunque no se encontraron diferencias estadísticamente significativas en la carga de variantes entre grupos de diferentes edades, se aprecia que hay muestras con una alta carga de variantes de alta frecuencia que no corresponden a polimorfismos como B4 (adulto joven, 2.24785) y B3 (adulto envejecido, 5.457961).

Tabla 13.Sumatoria de frecuencias por muestras.

Condición	Muestra	Todas las variantes	Alta heteroplasmia sin SNP	Baja heteroplasmia (Variantes <0.4)
Embrión	D10	3.156493	1.157493	0.253493
Embrión	D11	3.11379	1.113799	0.406799
Embrión	D12	3.216105	1.218105	0.257105
Embrión	D6	2.936383	0.937383	0.211383
Embrión	D9	3.166816	1.167816	0.359816
Joven	B4	3.24285	2.24785	0.47485
Joven	E4	3.0989503	1.0999503	0.3619503
Joven	E5	3.191703	1.191703	0.434703
Joven	E6	2.620873	0.621873	0.137873
Joven	E7	2.866428	0.867428	0.171428
Viejo	B1	2.2	1.2045	0.60351
Viejo	B2	2.221109	1.222109	0.691109
Viejo	B3	7.455961	5.457961	0.570961
Viejo	E1	2.853486	0.853486	0.283486
Viejo	E2	3.146404	1.147404	0.153404

En la tabla se presentan las sumatorias de frecuencias tomado los parámetros anteriores y desglosados por muestra.

9.2.3.3 Identificación de variantes de alta frecuencia

Con nuestros análisis se identificaron ocho variantes con frecuencias mayores a 0.8 en los diferentes grupos de edad. En este grupo se identificaron variantes correspondientes a polimorfismos reportados previamente para la cepa C57BL/6J y variantes únicas en las muestras B3 (adulto envejecido) y B4 (adulto joven).

Tabla 14. Variantes con frecuencias mayores a 0.8.

		Embrión					Joven					Viejo				
Posición	**Cambio**	**D6**	**D9**	**D10**	**D11**	**D12**	**B4**	**E4**	**E5**	**E6**	**E7**	**B1**	**B2**	**B3**	**E1**	**E2**
9461	**T-C**	**0.99**	**0.99**	**0.99**	**1.0**	**0.99**	**0.99**	**1.0**	**1.0**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**1.0**	**0.99**
9348	**G-A**	**1.0**	**1.0**	**1.0**	**1.0**	**0.99**		**0.99**	**1.0**	**1.0**	**1.0**			**0.99**	**1.0**	**1.0**
9820*	**T-TAA/TA**	**0.732**	**0.859**	**0.9110**	**0.71**	**0.993**	**0.995**	**0.996**	**0.99885**	**0.484**	**0.696**	**0.998**	**0.9992**	**0.997**	**0.646**	**0.998**
3919	**T-C**													**0.998**		
8260	**G-A**													**0.998**		
12217	**A-G**													**0.998**		
13052	**T-C**		**0.00695**	**0.049**	**0.216**		**0.983**				**0.028**	**0.00785**			**0.05**	
15672	**C-T**													**0.912**		

El (*) en la posición 9820 significa que esta posición presenta más de un genotipo alternativo, el gris oscuro muestra aquellos individuos donde uno de los genotipos alternativos supera el 80% de frecuencia, mientras que en lo lila???? es la suma de todos lo genotipos lo que da una alta frecuencia acumulada. En la posición 13052 sólo en la muestra B4 (adulto joven) se identifica esta variante, sin embargo, en otros individuos se encuentra esta variante a menor frecuencia.

De las variantes identificadas con alta frecuencia en las muestras, tres corresponden a polimorfismos previamente descritos y detectados con el Método A (9348, 9829 en la región de poli A y 9461). Sin embargo, también se identificaron variante únicas con frecuencias mayores a 0.8 en sólo en una de las muestras. Éstas se encontraron en las posiciones 3919, 8260, 13052 y 15672 para la muestra B3 correspondiente a un adulto envejecido y en la posición 12217 para la muestra B4 correspondiente a un adulto joven.

Tabla 15. Variantes de alta frecuencia identificadas con el Método B.

Posición	Tipo	Secuencia referencia	Secuencia variante	Localización	Original	Mutado
9348	X	G	A	COIII	GTC-Val	ATC- Ile
9820	I	*	A	mt-tRNA (Arg) poliA		
9461	X	T	C	ND3	ATT- Ile	ATC- Ile
3919	X	T	C	ND2	TCC- 0. Ser	CCC-Pro
8260	X	G	A	ATPase6	GCC-Ala	AGG-Thr
12217	X	A	G	ND5	ACG-Thr	GCG-Ala
13052	X	T	C	ND5	TTC-Phe	TCC-Ser
15672	X	C	T	D-loop		

Las tres primeras variantes reportadas corresponden a polimorfismos. El tipo representa la clase de variante localizada en el que X=Sustitución, I=Inserción y D=Deleción.

En la posición 9461, la sustitución de una base correspondiente a timina por una citosina, así como el cambio de guanina por adenina en la posición 9848 son considerados polimorfismos específicos de la cepa C57BL/6J. Las inserciones en la posición 9820 equivalen a otras de manera ambigua por corresponder a un conjunto de adeninas de longitud variable en el que no es posible determinar el sitio exacto de inserción.

Las variantes se encuentran localizadas en distintos loci. La variante T3919C se localiza en el gen correspondiente a ND2, G8260A en ATP6, A12217G en ND5, 15672 en el *D-loop* y por último la variante de muy alta frecuencia localizada en la muestra B4 se localiza en gen ND5. Mutaciones en estas regiones se han reportado con una relación en distintas enfermedades en humanos. Sin embargo, de las anteriores sólo una mutación se ha reportado directamente asociada a una patología en el MITOMAP, la variante T3919C la cual se ha relacionado con el padecimiento LHON.

9.2.3.4 Identificación de variantes con distintos genotipos

Entre las variables con frecuencia mayor a 0.5 se encontraron también dos variantes en las posiciones 9820 y 5171. Ambos cambios se encuentran en regiones de poliA que han sido previamente reportados como hipervariables asociadas al número de copias de ADNmt y a la disminución de la vida media del

individuo (Sachadyn, Zhang, Clark, Naviaux, & Heber-katz, 2009). El número de repeticiones de adenina en estos loci varía entre cepas e incluso entre los individuos de la misma edad.

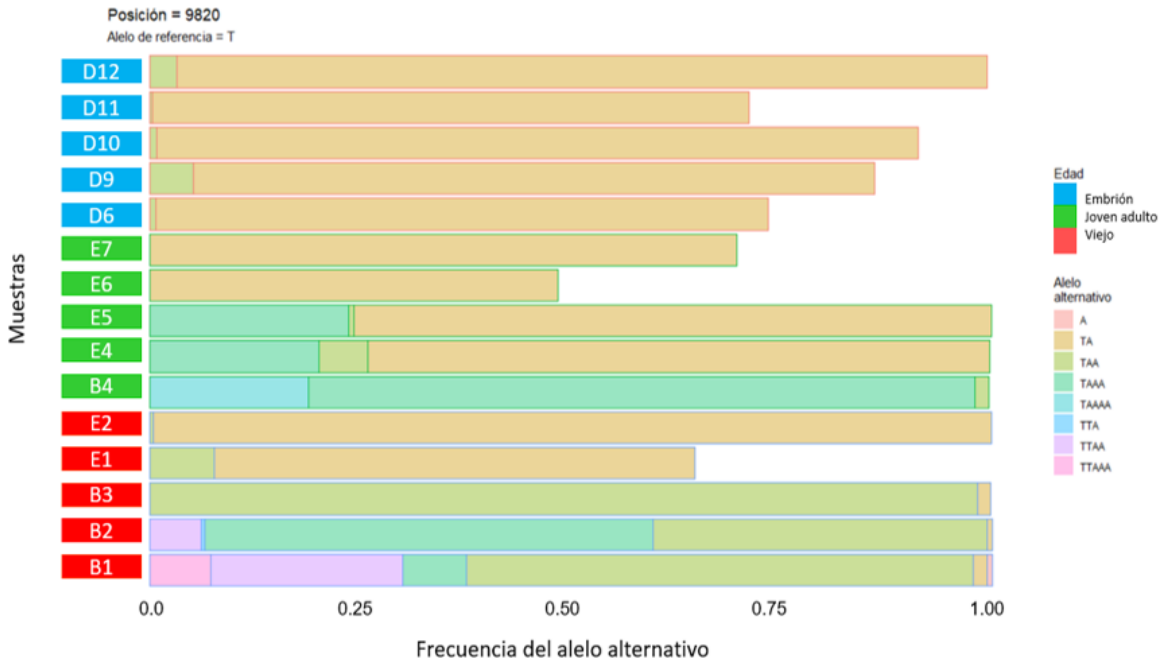

Figura 24. Frecuencia de alelos alternativos variables en la posición 9820. *En esta posición se presentan ocho genotipos diferentes, con distinta frecuencia en las muestras de los distintos grupos de edad. Todas las barras representan variantes con respecto al genoma de referencia y su longitud es proporcional a la frecuencia.*

La posición 9820 corresponde al inicio de una cadena de poliA en el ARNt de arginina. Previamente se han reportado inserciones de una a seis adeninas en este locus en algunas cepas de ratón como MRL/MpJ (Sachadyn, Zhang, Clark, Naviaux, & Heber-katz, 2009) en la que se aprecia que la longitud de este tracto aumenta con la edad. En la Figura 24 se observa también que la longitud de esta región se ve afectada presentado copias en donde es ligeramente más larga y variable en los ratones jóvenes y viejos que en el caso de los embriones.

En la Figura 25 se muestran los alelos encontrados en la posición 5171. Esta posición en el genoma de referencia presenta una guanina (G) y como se observa en la figura se detectaron distintos tipos de cambios. Este locus corresponde al origen de replicación de la cadena ligera O_L y es adyacente a una de las regiones polimórficas del genoma mitocondrial que contiene una cadena de

poli A de longitud variable detectada también con el Método A (Bayona-Bafaluy et al., 2003; Wanrooij & Falkenberg, 2010).

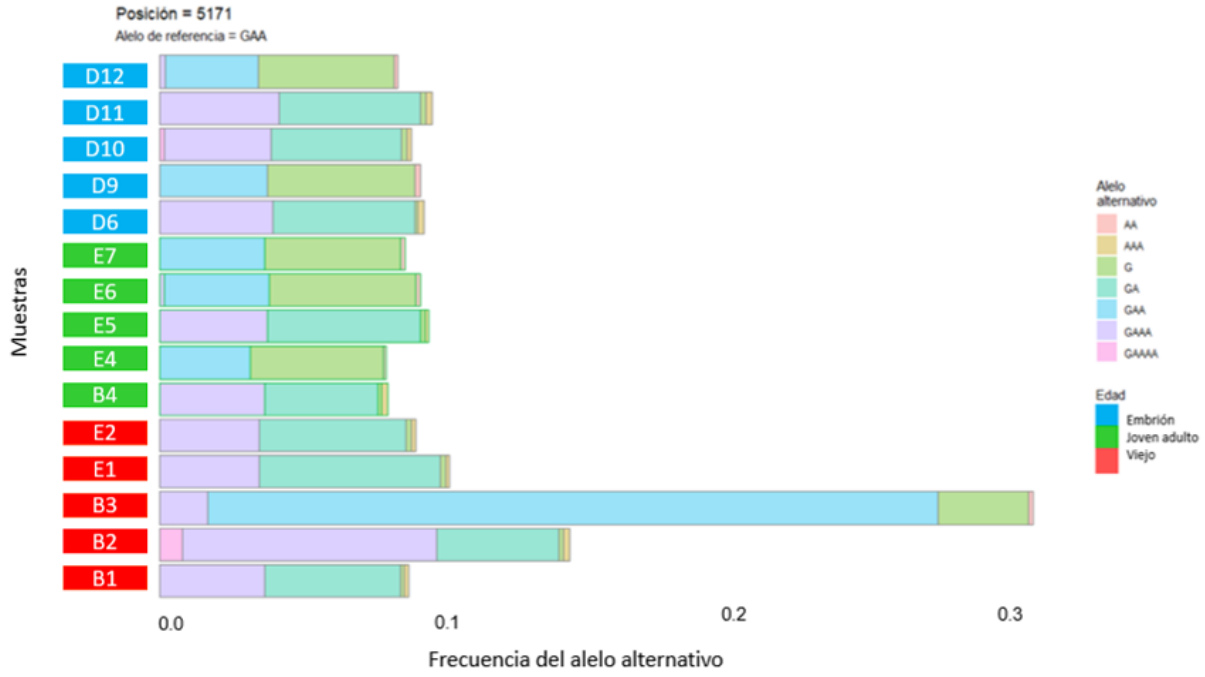

Figura 25. Alelo de frecuencia variable en la posición 5171. En esta posición se presentan distintos genotipos en frecuencias bajas. *Todas las barras representan variantes con respecto al genoma de referencia y su longitud es proporcional a la frecuencia*

Consistentemente con lo previamente reportado, las variantes con mayor frecuencia en esta posición fueron 10A y 9A en la mayoría de los casos. También se encontró que 12 de las 15 muestras estudiados presentaron la variante 12A en un rango de frecuencias de 0.048 a 0.001. Otras de las variantes que se identificaron en esta región incluyen una de longitud 13A, la deleción la base G5171 y una sustitución G5171A, todas frecuencias por abajo de 0.001.

9.2.3.5 Identificación de variantes de baja frecuencia

Entre las 52 variantes diferentes identificadas mediante el Método B, se observó que, además de las ocho previamente mencionadas y las dos de alelo variable, la mayoría de las mutaciones son de frecuencias menores a 0.5. En el caso de los embriones, se identificaron 32 variantes, en su mayoría presentes en todos las muestras del grupo (Tabla 16). Se encontró además que el 53% de ellas se encuentran en la región del D-loop, como el cambio T-C en la posición 16105 que

está en todos las muestras con frecuencia variable en el rango de 0.083 a 0.023. Otros cambios en el D-loop corresponden a variantes únicas en algunas muestras, pero de más baja frecuencia que los cambios que se comparten entre dos o más de ellas.

Tabla 16. Variantes de baja frecuencia detectadas en embriones.

Muestra	Posición	Secuencia en Referencia	Secuencia en Variante	Frecuencia	Región
D10	9849	G	A	0.005693	ARNtArg
D12	9849	G	A	0.094	
D6	9849	G	A	0.006172	
D9	9849	G	A	0.068	
D10	10554	CTTA	C	0.002646	ND4
D11	10554	CTTA	C	0.002709	
D12	10554	CTTA	C	0.003066	
D6	10554	CTTA	C	0.002128	
D10	11459	A	G	0.026	
D9	11459	A	G	0.008809	
D11	11710	CA	C	0.003619	ARNt-Leu
D11	13051	T	TC	0.003123	ND5
D10	13052	T	C	0.049	
D11	13052	T	C	0.216	
D9	13052	T	C	0.006959	
D10	16099	A	C	0.004685	*D-loop*
D10	16104	C	A	0.005336	
D10	16105	T	C	0.031	
D11	16105	T	C	0.046	
D12	16105	T	C	0.023	
D6	16105	T	C	0.073	
D9	16105	T	C	0.083	
D10	16108	T	C	0.015	
D11	16108	T	C	0.023	
D12	16108	T	C	0.021	
D6	16108	T	C	0.024	
D9	16108	T	C	0.041	
D9	16110	T	C	0.008993	
D10	16112	A	T	0.005768	
D11	16115	G	T	0.013	
D6	16115	G	T	0.006004	
D10	16267	A	T	0.012	

En el caso de adultos jóvenes se encontraron 29 variantes totales en las cinco muestras analizadas (Tabla 17). Del mismo modo que en el caso de los embriones, la mayoría de estas variantes (65%) se encuentran en el *D-loop*.

Tabla 17. Variantes de baja frecuencia detectadas en adultos jóvenes

Muestra	Posición	Secuencia en Referencia	Secuencia en Variante	Frecuencia	Región
E5	48	G	A	0.036	ARNt-Phe
E5	370	TATAA	T	0.002356	mt-Rnr1
E4	10554	CTTA	C	0.001715	ND4
E5	10554	CTTA	C	0.003133	
E5	11545	G	C	0.023	
B4	13048	GT	G	0.01	ND5
B4	13051	T	TC	0.025	
B4	13052	T	C	0.983	
E7	13052	T	C	0.028	
B4	15558	T	C	0.042	*D-loop*
B4	16099	A	C	0.00427	
E6	16099	A	C	0.003223	
E7	16099	A	C	0.01	
B4	16105	T	C	0.043	
E4	16105	T	C	0.013	
E5	16105	T	C	0.023	
E6	16105	T	C	0.033	
E7	16105	T	C	0.033	
B4	16108	T	C	0.032	
E4	16108	T	C	0.009275	
E5	16108	T	C	0.011	
E6	16108	T	C	0.009757	
E7	16108	T	C	0.014	
B4	16112	A	T	0.007775	
B4	16115	G	T	0.005746	
B4	16121	C	A	0.003951	
B4	16130	C	A	0.005623	
B4	16146	G	A	0.009997	

El resto corresponden a variantes en regiones codificantes en los genes ND4 y ND5. Sin embargo, la muestra E5 presenta dos variantes únicas, una en el ARNt de fenilalanina y la otra en el gen de ARN ribosomal mt-Rnr1. También se aprecia que hay variantes compartidas con el grupo de los embriones en las posiciones 13052, 16105, 10554 y 16108 con frecuencias similares.

Para el caso de los adultos envejecidos se identificaron 35 variantes entre todas las muestras analizadas (Tabla 18). Al igual que en los grupos previos, la mayor cantidad de variantes (75%), corresponden a la región del *D-loop*. Otras regiones afectadas fueron también ND3, ND4 y ND5, además de variantes en el ARNt de arginina y el correspondiente al ARNr de la subunidad 16S.

Tabla 18. Total, de variantes de baja frecuencia en adultos envejecidos.

Muestra	Posición	Referencia	Cambio	Frecuencia	Región
B3	1487	TA	T	0.007644	mt-Rnr2
B3	9800	A	T	0.004073	ND3
B1	9819	T	TAAA	0.004418	mt-TR
B3	10554	CTTA	C	0.003227	ND4
E1	10554	CTTA	C	0.003097	
E2	10554	CTTA	C	0.002205	
B1	13048	GT	G	0.00197	ND5
E1	13052	T	C	0.05	
B1	13052	T	TC	0.007581	
E2	16097	C	A	0.002598	D-loop
B3	16105	T	C	0.077	
B2	16105	T	C	0.063	
B1	16105	T	C	0.054	
E2	16105	T	C	0.036	
E1	16105	T	C	0.033	
B3	16108	T	C	0.023	
E2	16108	T	C	0.018	
B1	16108	T	C	0.016	
B2	16108	T	C	0.015	
E1	16108	T	C	0.014	
B3	16112	A	T	0.007811	

Como se observa en las tablas 16, 17 y 18 la mayoría de las variantes son compartidas entre grupos, aunque existen algunas de alta y baja frecuencia presentes en sólo una de las muestras.

9.3 Comparación entre métodos

De los métodos utilizados para la identificación y la cuantificación de variantes, el Método A realiza un análisis comparativo entre los valores de frecuencia de las muestras de un grupo con respecto a las muestras control permitiendo la identificación de variantes comunes a las muestras y ausentes de las muestras control. En cambio, el Método B permite un análisis de cada muestra y sus valores de calidad y calcula los valores de frecuencia de los genotipos variantes en comparación con los que corresponden a la secuencia de referencia. Esto facilita la identificación de variantes únicas en un individuo y posteriormente permite detectar patrones específicos en cada grupo.

Tabla 19. Comparación de los resultados de cada método de análisis de la carga de variantes en el genoma mitocondrial*. En la taba se presentan algunos de los resultados identificados con cada programa y las diferencias entre ellos, las cuales tienen inferencia en la cantidad de variantes identificadas y el tiempo de análisis.*

Método A	Método B (MuTect2, modo mitocondrial)
Alto número de variantes identificado: **1450 las cuáles incluyen polimorfismo y variantes de baja frecuencia tomando en cuenta el filtro 1/D**	Bajo número de variantes identificado **60 inlcuyendo polimorfimos y descartando variantes repetidas entre grupos.**
Las variantes reportadas en alta frecuencia son: *9461 T C (99%, Jóvenes) *9829 * A (43%, Embriones) *9348 G A (98%, todos) Son excepción de las tres variantes previas el resto de las variantes localizadas por este método se encontraron por debajo del 20%.	Las variantes reportadas en alta frecuencia son: ***9461 T C (99-100%, todos)** ***9820 * A (40-60%, todos)** ***9348 G A (98-100%, todos)** Se localizaron 5 variantes únicas de alta frecuencia (mayor a 0.9) en muestras individuales (B4 y B3) las cuales no fueron identificadas por el programa A (Figura 18). Se localizaron variantes por debajo del 0.4 hasta 0.0009603 de frecuencia, muchas de las variantes se encuentran en más de un individuo. De las variantes detectadas con el método A se encontraron los polimorfismos y algunas de las variantes en las regiones de alelo variable 9820 y 5172 en frecuencias similares en algunos casos.
Analiza todas las lecturas	Alrededor del 30% de lecturas se filtran por distintos parámetros de calidad.
Análisis por grupo	Análisis individual
Detecta mayor acumulación de variantes en las regiones correspondientes a COIII**, D-loop, ND4, ND5 y Rnr2.**	Detecta mayor acumulación de variantes en las regiones correspondientes al **D-loop, ND4 y ND5.**

En el comparativo de resultados observamos que las variantes de alta frecuencia (>0.4) fueron detectadas por los dos métodos (Tabla 19). El Método A, sin embargo, no las detectó en todos los grupos de edad a pesar de que al inspeccionar los valores de frecuencia se detectaron en todos los grupos al igual que con el Método B. Esto se debió a que el Método A desechó algunos de los grupos por tener niveles de frecuencia ligeramente menores a los presentados por

el control durante la comparativa realizada. Por otro lado, el Método B detectó 60 totales de las cuales algunas de las mismas fueron consistentes en todas las muestras o en subgrupos de ellas. Además, con este método se identificaron variantes que estaban presentes sólo en algunas muestras que no fueron detectadas por el Método A dado que éste requiere que las variantes estén presentes en todos los miembros de un grupo para ser válidas. El Método A, en cambio, detectó 1450 variantes de frecuencias mucho más bajas que las detectadas por el Método B. Esta comparación revela que cada método tiene sus fortalezas y limitaciones que en conjunto que pueden permitir obtener una visión global de la carga de variantes en el genoma mitocondrial del ratón en diferentes etapas de la vida, sin embargo, en algunos casos es necesario trabajar a detalle el método de análisis para evitar la presencia de falsos positivos. Por lo tanto, considerando que pueden ser necesarios ajustes, el método A podría resultar útil para análisis grandes en poblaciones completas mientras que el método B es un método más particular que permitió identificar variantes individuales que no se logran apreciar con el Método A.

Otra de las consideraciones es que en ambos casos la distribución de variantes es consistente relativamente mostrado por los dos métodos que hay regiones de mayor recurrencia donde se acumulan las variantes como el D-loop. ND4, ND5 o incluso COIII, lo cual pareciera indicar una distribución en zonas discretas en el genoma más que una acumulación en todas las regiones del mismo.

El método B permitió identificar regiones interesantes como las regiones de alelo variable, las cuales en algunos casos han sido ya reportadas como en la posición 9820. Sin embargo, resulta interesante conocer la proporción de estas variantes en cada individuo.

10. DISCUSIÓN

Este trabajo plantea un método integral para el análisis del genoma mitocondrial que incluye una técnica experimental modificada para favorecer la extracción de moléculas circulares, eliminando residuos de DNA lineal por un método enzimático y seguido de secuenciación masiva. Asimismo, el procedimiento experimental se complementa con técnicas bioinformáticas para realizar el análisis de las secuencias obtenidas.

Uno de los obstáculos para el análisis del genoma mitocondrial es la dificultad para diferenciar entre variantes reales y errores introducidos por los métodos de análisis. Esto se debe a la gran cantidad de copias por célula del genoma mitocondrial y a las variantes que contiene, aunando al mosaicismo genético de un mismo individuo. Además, las variantes en el ADNmt pueden ser o no patogénicas y su efecto puede depender de su combinación con otras variantes de frecuencias variables. Esto ocasiona que mutaciones similares puedan tener expresiones clínicas variables y que síndromes similares sean causados por mutaciones diferentes como en el caso de distintos tipos de cáncer, enfermedades neurodegenerativas y neuromusculares.

En este estudio se utilizó una herramienta informática generada por nuestro grupo de trabajo y se comparó con una herramienta disponible públicamente. Por lo tanto, uno de los objetivos del mismo fue determinar los parámetros necesarios para la cuantificación e identificación de variantes de alta y baja heteroplasmia de forma eficiente con ambos métodos. El primer método corresponde a un flujo de trabajo que incluye la herramienta diseñada por nuestro grupo como identificador de variantes (Método A) y el segundo incluye a MuTect2 con el mismo propósito (Método B). Los procedimientos experimentales, el control de calidad y el alineamiento de secuencias al genoma de referencia fueron iguales en los dos métodos.

El Método A fue diseñado para identificar variantes en las muestras con respecto al genoma de referencia, este método hace uso de muestras control que consisten en seis fragmentos de ADNmt clonados en plásmidos y de análisis estadísticos con la finalidad de determinar la diversidad y abundancia de variantes

comunes a un grupo de muestras. El segundo método empleado, fue originalmente desarrollado para identificar mutaciones asociadas a cáncer con la premisa de que en tejido canceroso existen células sanas y otras portadoras de mutaciones en frecuencias bajas (Cibulskis *et al*., 2013). Además, este programa tiene un modo mitocondrial que toma en cuenta los niveles bajos de heteroplasmia, la circularidad del ADNmt y la presencia de regiones hipervariables, con lo que optimiza la búsqueda de variantes.

Mediante la búsqueda de variantes con el Método A se identificaron 1631 variantes, mientras que con el método B se detectaron únicamente 60. Estas están distribuidas a lo largo de todo el genoma mitocondrial e incluyen polimorfismos de la cepa C57BL/6J y variantes de alta y baja frecuencia encontradas en los tres grupos experimentales. La ventaja del Método A es, entonces, que permite detectar patrones y comportamientos generales de un grupo de muestras pertenecientes a la misma condición. Sin embargo, por diseño, no detecta variación individual de cada muestra. Además, a pesar de la baja abundancia de cada una de las variantes detectadas en comparación con las variantes válidas identificadas por el Método B, la significancia de su hallazgo reside en que se detectaron en todas las muestras de un grupo, que estuvieron ausentes de las secuencias control y que pasaron el filtro del límite empírico de detección (1/D).

El Método B, por otra parte, constituye un método más estricto, pero a su vez más preciso. Es por esto que detecta una menor cantidad de variantes, pero detecta consistentemente polimorfismos que con el Método A no se detectan. Esto se debe a que el Método A lleva a cabo una comparación control-muestra conservando sólo las variantes cuya frecuencia (mediana) en las muestras es mayor que la del control. Por ejemplo, elimina una variante si su mediana en las muestras del grupo es ligeramente menor que la del control, aunque ambas estén cerca el la homoplasmia. Tal es el caso de algunos polimorfismos de la cepa como el cambio en la posición 9461 que con el Método A sólo se detecta en jóvenes al 99% y con MuTect2 se detecta en todas las muestras con abundancias relativas

de cerca del 100%. Este hallazgo revela un aspecto en el que el Método A puede ser ajustado para fortalecerlo.

Otra diferencia importante radica en la variante localizada en la posición 9829 ya que con el Método A se identifica al 43% de heteroplasmia únicamente en los embriones. Sin embargo, los resultados con el Método B mostraron que esta variante se encontraba en la posición 9820 en todos los grupos e incluso en combinaciones de distintos genotipos que no se detectaron con el primer método. Esta diferencia de localización se debe a que el Método B cuenta con realineamiento de *indels* y reporta la diferencia al inicio de la secuencia de poli A que hay en esa región. Por ello se determinó que el Método B resulta útil para identificar variación individual y detectar cambios que no se logran detectar en el Método A diseñado para la comparación entre grupos.

Debido a lo anterior se determinó que si bien el Método A puede llegar a ser un herramienta bastante útil para el análisis de grupos como en este estudio que se analizan tres edades diferentes, aun requiere algunas modificaciones para hacerlo más preciso como una de las pocas herramientas que existen para el análisis de ADNmt. Se ha propuesto también que para abordar la problemática de que no detecta variación individual se debe incluir un modo de extraer información individual previo a la comparación entre muestras de un mismo grupo. Por lo anterior para el análisis de las variantes se emplearon los hallazgos obtenidos con MuTect2.

Variantes de alta abundancia relativa

Entre las variantes de alta frecuencia se detectaron cambios individuales y polimorfismos correspondientes a la cepa C57BL/6J. Los polimorfismos identificados corresponden a los previamente reportados en la literatura para esta cepa en las regiones de ND3, el ARNt de arginina y COIII. Con el Método B se identificaron 8 variantes que superan el 80% de heteroplasmia en al menos una muestra.

La variante T9461C se identificó en todos los individuos de los tres grupos analizados con una frecuencia aproximada del 100%, presentándose en menor frecuencia la secuencia de referencia. Esta variante corresponde a un

polimorfismo previamente reportado en la cepa utilizada que tiene como consecuencia la pérdida de un sitio de restricción y corresponde a un cambio sinónimo en el primer codón de ND3, que codifica para isoleucina. Otra de las variantes identificadas como polimorfismo es la variante G9348A, la cual causa también la pérdida de un sitio de restricción y corresponde a un cambio valina-isoleucina. Los resultados revelan consistentemente estas dos variantes en todas las muestras por lo cual concluimos que son diferencias entre el genoma de nuestra colonia con el genoma de referencia.

Dentro de las variantes de mayor abundancia relativa también se encontraron variantes en la posición 9820 en todos los individuos en frecuencia variable y distintos genotipos en una misma posición. Esto se debe a que esta región corresponde a una cadena de poliA y es reconocida por ser un *hot-spot* de variantes (Jandova et al., 2012; Stoneking, 2000). En el caso de la cepa C57BL/6J se ha reportado en la posición 9820 un polimorfismo de longitud 9A, con respecto al genoma de referencia (C57BL/6J, mm10) que tiene con una longitud de 8A. Estudios previos han reportado que polimorfismos de longitud variable de adeninas en la posición 9821 que corresponden al bucle DHU del ARNmt de Arginina no tienen implicaciones en los niveles de especies reactivas de oxígeno y en el de número de copias de ADNmt con excepción de la longitud 10A que causa aumentos significativos en estos parámetros (Moreno-Loshuertos, 2006).

Algo relevante es que se observa una tendencia al aumento en la variabilidad y longitud de esta región con la edad. En esta posición algunas muestras presentaron la variante 9A con frecuencia por encima del 90% de como en el caso de las muestras D10 y D12 de embriones y, B3 y E2 de adultos envejecidos, mientras que en las demás muestras la carga mutacional de esta región es alta por la combinación de distintos genotipos que implican la inserción de dos o tres adeninas.

Anteriormente se creía que esta variante no tenía implicaciones funcionales sobre los organismos (Bayona-Bafaluy et al., 2003). Sin embargo, estudios más recientes han reportado variabilidad en la longitud de esta región en el rango de 5 a 8 nucleótidos. Algunas de estas variantes se han asociado con un deterioro leve

en la síntesis de proteínas mitocondriales y el aumento de especies reactivas de oxígeno (ROS). Además, se sabe que las modificaciones en esta región son responsables de fenotipos profundos y altamente pleomórficos en otros mt-ARNt (Moreno-Loshuertos, Pérez-Martos, Fernández-Silva, & Enríquez, 2013).

Un caso particular por su patrón de variantes fue la muestra B3 de un individuo envejecido la cual contiene cuatro variantes únicas de alta frecuencia en las posiciones T3919C (98%), G8260A (98%), A12217G (98%) y C15672T (91%), mismas que no fueron detectadas ni en baja frecuencia en ningún otro individuo analizado. Esto sugiere que algún mecanismo particular causó la generación y fijación de estas variantes en el linaje de este individuo y apoya la idea de la existencia de procesos repentinos que generan variantes con altas frecuencias en una generación (Hudson, Gomez-Duran, Wilson, & Chinnery, 2014; Meirelles & Smith, 1997). Por la presencia de estas cuatro variantes en alta abundancia relativa en un solo individuo es improbable que se trate de eventos separados asociados al envejecimiento. Es más probable, en cambio, que su origen haya sido la imprecisión de la maquinaria de replicación o reparación del genoma mitocondrial en la línea germinal, seguido de un mecanismo complementario que causó el incremento en abundancia relativa a cerca del 100%. Es notable, que todas estas variantes corresponden a transiciones que generan cambios no sinónimos en las regiones codificantes indicadas que sorprendentemente no son letales pues se detectaron en un individuo envejecido aparentemente sano.

En el caso de la variante en posición 13052 hay un cambio no sinónimo de fenilalanina a serina. Esta variante corresponde a la última posición de un bloque de cuatro T consecutivas que se detectó en el individuo joven B4 con una abundancia de 98%. Sin embargo, esta variante fue identificada en frecuencias variables (0.007-0.49) en otros individuos de todos los grupos, a diferencia de otras ragiones no presentaron otros genotipos alternativos. Esto sugiere que existen en nuestra colonia diferentes linajes portadores de esta variante y, por extensión lógica, probablemente de otras más.

Variantes con más de un alelo alternativo

La posición 5171, ubicada en el origen de replicación de la cadena ligera (O_L), también presentó múltiples alelos alternativos. En el genoma de referencia esta posición corresponde a una guanina (G) y es adyacente a una cadena de poliA de longitud 11A. En nuestro estudio detectamos que la longitud predominante es 10A y también detectamos variantes de 11A a 13A en un rango de frecuencias individuales bajas. Adicionalmente se detectaron en este locus la deleción de la base G5171 y la sustitución G5171A, también en bajas frecuencias. En cepas de laboratorio y en ratones domésticos de vida libre se han encontrado polimorfismos en esta cadena de poliA que en la mayoría de las copias del ADNmt (70% ó más) tiene longitud de 11A, mientras que el porcentaje restante puede exhibir desde 9 a más de 11 (Bayona-Bafaluy et al., 2003; Wanrooij & Falkenberg, 2010). Esto indica que tal patrón de variación se encuentra de forma natural y puede representar una consecuencia de la presión de mutación y posiblemente estar sujetos a una selección natural (Hirose et al., 2018). Adicionalmente, la abundancia relativa de la variante 12A se ha correlacionado negativamente con la vida media de los individuos portadores, ya que su aumento se ha asociado con la disminución del número de copias del ADNmt que se compensa con un incremento en la expresión de los genes que codifican proteínas que participan en la FOSFOX alterando el metabolismo de la glucosa (Hirose et al., 2018).

Variantes de baja abundancia relativa

El resto de las variantes identificadas corresponden a variantes de baja frecuencia que se encuentran distribuidas en todo el ADNmt. Sin embargo, existen zonas de mayor recurrencia de variantes, las cuales principalmente se encontraron en las zonas del *D-loop*, ND4 y ND5. La mayoría de las variantes de baja frecuencia se comparten en al menos dos individuos del mismo grupo o de grupos distintos. Adicionalmente, hay variantes en distintas regiones que se presentan sólo en un individuo de los estudiados.

De las variantes de baja frecuencia, la mayoría están ubicadas en el *D-loop* en todos los grupos estudiados. Tal es el caso de las variantes localizadas en las

posiciones 16105 y 16108 que se presentan en la mayoría de los individuos de los tres grupos, en muchos casos con frecuencias similares. La presencia de estas variantes en los tres grupos de edad y baja abundancia relativa sugiere que cambios en este sitio tienen consecuencias deletéreas en funciones mitocondriales por lo que pueden estar sujetas selección funcional negativa.

Alta abundancia por la acumulación de variantes

Nuestro estudio reveló variantes de abundancia cercana a la homoplasmia en sitios en los que la secuencia difiere de la del genoma de referencia en todos los individuos analizados por lo que se pueden definir como polimórficas en nuestra colonia de ratones de la cepa C57BL/6J. Este estudio también reveló variantes de alta abundancia en algunos individuos y baja en otros, variantes de baja frecuencia y sitios con múltiples variantes de diversas frecuencias que parecen aumentar con la edad, sitios con variantes adyacentes a otros de alta variabilidad y hasta un individuo con cuatro variantes de alta abundancia en sitios distintos. Sin tomar en cuenta las variantes polimórficas, todo esto indica que hay una alta diversidad de variantes en todas las muestras, apoyando la idea de la generación frecuente de variantes en este genoma. Sin embargo, dado que en este primer estudio con este método tan sensible no se tomaron en cuenta las relaciones de linaje entre los individuos estudiados, se requiere de estudios adicionales con diseños experimentales que incorporen este aspecto para confirmar y cuantificar la tasa de generación de variantes como las descritas.

Las variantes polimórficas de alta abundancia podrían estar restringidas a sublinajes de nuestra colonia y no parecen estar asociadas a disfunción mitocondrial pues en este estudio se emplearon animales sanos. Las de baja abundancia, en cambio, representan una carga basal que podría actuar como sustrato para la generación de enfermedades por cambios en su frecuencia en línea germinal, por mosaicismo somático durante el desarrollo embrionario o durante el envejecimiento.

Existen controversias acerca del origen de las variantes en el genoma mitocondrial y la relevancia de la edad en el aumento de variantes o su frecuencia (Chen et al., 1995; Kazachkova, 2013; Payne & Chinnery, 2015; Reddy et al.,

2015). En este trabajo se observa que existe un aumento con la edad de algunas variantes y sólo en sitios discretos sin un aumento generalizado de la carga total de variantes. El aumento selectivo aparentemente asociado a la edad se observa en la región del *D-loop* y del ARNt-Arginina, lo que podrá ser corroborado en estudios subsecuentes con el método desarrollado en este trabajo.

No obstante, la visión final de este estudio es que aun dejando de lado las diferencias polimórficas entre cepas, un alto porcentaje de las copias del genoma mitocondrial de animales sanos de etapas embrionarias o adultas parecen ser portadores de variantes de alta y baja frecuencia. Si asumimos como se indicó previamente, que cada variante ocurre independientemente de cualquier otra en un locus distinto del genoma mitocondrial, una estimación global con los dos métodos empleados indica que entre el 30 y el 50% de los genomas secuenciados contienen al menos una variante. Si bien se ha propuesto que el genoma mitocondrial es portador de diversas variantes de secuencia en heteroplasmia (Ju et al., 2014; H. Ma et al., 2018; Rebolledo-Jaramillo et al., 2014; Wai, Teoli, & Shoubridge, 2008), hasta esta fecha su diversidad, abundancia relativa y carga acumulada no se había podido cuantificar de manera tan precisa como en este estudio.

Este trabajo proporciona un flujo de trabajo experimental y bioinformático para la identificación y cuantificación de variantes de secuencia en el genoma mitocondrial que reveló información novedosa y sin duda permitirá abordar en diferentes estudios las controversias mencionadas arriba. Además, ofrece la posibilidad de desarrollar alternativas para el análisis del ADNmt en muestras clínicas ya que detecta variantes de alta y baja frecuencia en cualquier región de la molécula circular del DNA mitocondrial sin sospecha previa de su localización o abundancia. Al ser un trabajo inicial de desarrollo del método analítico, se identificaron algunos aspectos de cada fase que es necesario abordar para mejorar en el análisis del DNA mitocondrial de muestras clínicas como sangre y biopsias de músculo y optimizar la sensibilidad de detección. Esto representará un avance considerable en el diagnóstico de enfermedades asociadas a mutaciones en el genoma mitocondrial.

11. CONCLUSIONES

- En este estudio se desarrolló un flujo de trabajo experimental y bioinformático para localizar con precisión variantes en el genoma mitocondrial que permite detectar variantes de alta y baja frecuencia sin conocimiento previo de su abundancia y localización. Para el análisis bioinformático se probaron dos métodos distintos, uno generado en nuestro grupo de investigación y otro previamente publicado por otro grupo y se demostró su efectividad para el análisis de datos de genoma mitocondrial. De estos métodos se seleccionó el previamente publicado para los análisis más detallados.
- Se detectaron variantes polimórficas en la cepa C57BL/6J que ya estaban reportadas en otras cepas y otras que se detectaron por vez primera.
- Se detectaron variantes de alta abundancia presentes sólo en algunas de las muestras estudiadas lo que sugiere que existen linajes con diferencias en secuencia en la colonia de ratones empleada apoyando la idea de la generación frecuente de variantes.
- Se detectaron variantes de baja abundancia relativa que afectan a distintas regiones del genoma mitocondrial, pero con más frecuencia a las regiones correspondientes a ND4, ND5 y la región regulatoria *D-loop*, además del mt-ARNt de Arginina, el cual junto con el *D-loop* son regiones hipervariables en las que se encontró la mayor recurrencia.
- Se encontraron dos sitios en los que se detectaron diversos alelos de frecuencia variable, uno en el mt-TR cuya variabilidad y longitud parece estar ligada al envejecimiento y otra en la región del O_L) que presenta una frecuencia baja y constante en todos los individuos.
- Se encontraron patrones mutacionales complejos en dos muestras como la presencia de variantes de alta frecuencia en regiones únicas o patrones inusuales en regiones de alelo variable. Estos hallazgos sugieren que existen eventos de generación repentina de variantes en un mismo individuo que podrían establecerse en una o pocas generaciones por transmisión materna.

- La abundancia combinada de variantes de baja frecuencia indica que entre el 30 y el 50% de las copias del genoma mitocondrial del ratón es portador de diferencias en su secuencia.
- Este trabajo ofrece la posibilidad de desarrollar un método de aplicación clínica a futuro para el análisis de enfermedades complejas asociadas con el genoma mitocondrial.

12. PRESPECTIVAS

Este trabajo plantea un método integral tanto experimental como bioinformático para el análisis del genoma mitocondrial en su totalidad. Este enfoque amplía la posibilidad del estudio del mosaicismo en tejidos individuales, la dinámica de la acumulación de variantes de forma discreta en regiones definidas, y la comparación entre distintos individuos o condiciones experimentales. Hasta ahora, esto ha representado un gran reto debido a la variabilidad de ciertas regiones, a la gran cantidad de copias presentes por célula, a la heteroplasmia y a la mayor frecuencia de generación de variantes que en el genoma nuclear.

Los resultados obtenidos hasta ahora se muestran congruentes con algunos estudios previos. Sin embargo, nuestro estudio arroja resultados nuevos e interesantes. Una de la perspectivas de este trabajo a futuro es que nuestro enfoque pueda tener aplicación clínica para el diagnóstico en humanos de enfermedades complejas asociadas a mutaciones en el genoma mitocondrial. Para ello es necesario realizar estudios adicionales en los que se consideren aspectos no analizadas inicialmente por el carácter exploratorio de este estudio, entre las cuales se tome en consideración el linaje de los sujetos analizados, el análisis y comparación de distintas muestras procedentes de otros tejidos o en condiciones patológicas asociadas con algún desorden que posteriormente permitan escalar el mismo a humanos.

La posibilidad de escalar este estudio a humanos requiere también analizar el tipo y el manejo de las muestras a utilizar. Es decir, es necesario adaptar los procedimientos al uso de ADN derivado de muestras de sangre y biopsias de músculo, que serían las muestras disponibles para el estudio de enfermedades humanas.

Otra área de oportunidad para este proyecto es la revisión y mejoras del método de llamado de variantes desarrollado por nuestro grupo para mejorar su precisión.

13. BIBLIOGRAFÍA

Alston, C. L., Rocha, M. C., Lax, N. Z., Turnbull, D. M., & Taylor, R. W. (2017). The genetics and pathology of mitochondrial disease. *Journal of Pathology*, *241*(2), 236–250. https://doi.org/10.1002/path.4809

Bayona-Bafaluy, M. P., Acín-Pérez, R., Mullikin, J. C., Park, J. S., Moreno-Loshuertos, R., Hu, P., … Enríquez, J. A. (2003). Revisiting the mouse mitochondrial DNA sequence. *Nucleic Acids Research*, *31*(18), 5349–5355. https://doi.org/10.1093/nar/gkg739

Chen, X., Prosser, R., Simonetti, S., Sadlock, J., Jagiello, G., & Schon, E. A. (1995). Rearranged mitochondrial genomes are present in human oocytes. *American Journal of Human Genetics*, *57*(2), 239–247. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/7668249%5Cnhttp://www.ncbi.nlm.nih.gov/pmc/articles/PMC1801549/pdf/ajhg00034-0045.pdf

Cibulskis, K., Lawrence, M. S., Carter, S. L., Sivachenko, A., Jaffe, D., Sougnez, C., … Getz, G. (2013). A n a ly s i s Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature Biotechnology*, *31*(3), 213–219. https://doi.org/10.1038/nbt.2514

Copeland, W. C. (2016). DNA polymerases in the mitochondria A critical review of the evidence. *Frontiers in Bioscience*, *22*(4), 692–709. https://doi.org/10.2741/4510

Doudican, N. A., Song, B., Shadel, G. S., & Doetsch, P. W. (2005). Oxidative DNA Damage Causes Mitochondrial Genomic Instability in. *Society*, *25*(12), 5196–5204. https://doi.org/10.1128/MCB.25.12.5196

Escobar-Henriques, M., & Anton, F. (2013). Mechanistic perspective of mitochondrial fusion: Tubulation vs. fragmentation. *Biochimica et Biophysica Acta - Molecular Cell Research*, *1833*(1), 162–175. https://doi.org/10.1016/j.bbamcr.2012.07.016

Friedman, J. R., & Nunnari, J. (2014). Mitochondrial form and function. *Nature*,

505(7483), 335–343. https://doi.org/10.1038/nature12985

Gorman, G. S., Schaefer, A. M., Ng, Y., Gomez, N., Blakely, E. L., Alston, C. L., … McFarland, R. (2015). Prevalence of nuclear and mitochondrial DNA mutations related to adult mitochondrial disease. *Annals of Neurology*, *77*(5), 753–759. https://doi.org/10.1002/ana.24362

Guo, Yan, Li, J., Li, C., Shyr, Y., & Samuels, D. C. (2013). Sequence analysis MitoSeek : extracting mitochondria information and performing high-throughput mitochondria sequencing analysis, *29*(9), 1210–1211. https://doi.org/10.1093/bioinformatics/btt118

Guo, Yiru, Flaherty, M. P., Wu, W.-J., Tan, W., Zhu, X., Li, Q., & Bolli, R. (2012). Genetic background, gender, age, body temperature, and arterial blood pH have a major impact on myocardial infarct size in the mouse and need to be carefully measured and/or taken into account: results of a comprehensive analysis of determinants of infarc. *Basic Research in Cardiology*, *107*(5), 288. https://doi.org/10.1007/s00395-012-0288-y

Hirose, M., Schilf, P., Gupta, Y., Zarse, K., Künstner, A., Fähnrich, A., … Johann, K. (2018). Low-level mitochondrial heteroplasmy modulates DNA replication , glucose metabolism and lifespan in mice, (October 2017), 1–15. https://doi.org/10.1038/s41598-018-24290-6

Hudson, G., Gomez-Duran, A., Wilson, I. J., & Chinnery, P. F. (2014). Recent Mitochondrial DNA Mutations Increase the Risk of Developing Common Late-Onset Human Diseases. *PLoS Genetics*, *10*(5). https://doi.org/10.1371/journal.pgen.1004369

Ishikawa, K., Kobayashi, K., Yamada, A., Umehara, M., Oka, T., & Nakada, K. (2019). Concentration of mitochondrial DNA mutations by cytoplasmic transfer from platelets to cultured mouse cells. *PLoS ONE*, *14*(3), 1–20. https://doi.org/10.1371/journal.pone.0213283

Jandova, J., Eshaghian, A., Shi, M., Li, M., King, L. E., Janda, J., & Sligh, J. E. (2012). Identification of an mtDNA mutation hot spot in UV-induced mouse skin tumors producing altered cellular biochemistry. *Journal of Investigative Dermatology*, *132*(2), 421–428. https://doi.org/10.1038/jid.2011.320

Ju, Y. S. eo., Alexandrov, L. B., Gerstung, M., Martincorena, I., Nik-Zainal, S., Ramakrishna, M., … Campbell, P. J. (2014). Origins and functional consequences of somatic mitochondrial DNA mutations in human cancer. *ELife*, *3*, 1–28. https://doi.org/10.7554/eLife.02935

Kazachkova, N. (2013). Mitochondrial DNA Damage Patterns and Aging: Revising the Evidences for Humans and Mice. *Aging and Disease*, *4*(6), 337–350. https://doi.org/10.14336/ad.2013.0400337

Kazak, L., Reyes, A., & Holt, I. J. (2012). Minimizing the damage: Repair pathways keep mitochondrial DNA intact. *Nature Reviews Molecular Cell Biology*, *13*(10), 659–671. https://doi.org/10.1038/nrm3439

Kiebish, M. A., & Seyfried, T. N. (2005). Absence of pathogenic mitochondrial DNA mutations in mouse brain tumors. *BMC Cancer*, *5*, 1–8. https://doi.org/10.1186/1471-2407-5-102

Kowaltowski, A. J. (2000). Alternative mitochondrial functions in cell physiopathology: Beyond ATP production. *Brazilian Journal of Medical and Biological Research*, *33*(2), 241–250. https://doi.org/10.1590/S0100-879X2000000200014

Larson, D. E., Harris, C. C., Chen, K., Koboldt, D. C., Abbott, T. E., Dooling, D. J., … Ding, L. (2012). SomaticSniper : identification of somatic point mutations in whole genome sequencing data, *28*(3), 311–317. https://doi.org/10.1093/bioinformatics/btr665

Lax, N. Z., Turnbull, D. M., & Reeve, A. K. (2011). Mitochondrial Mutations: Newly Discovered Players in Neuronal Degeneration. *The Neuroscientist : A Review*

Journal Bringing Neurobiology, Neurology and Psychiatry, *17*(6), 645–658. https://doi.org/10.1177/1073858411385469

Ludwig, L. S., Lareau, C. A., Ulirsch, J. C., Christian, E., Muus, C., Li, L. H., … Sankaran, V. G. (2019). Lineage Tracing in Humans Enabled by Mitochondrial Mutations and Single-Cell Genomics. *Cell*, *176*(6), 1325-1339.e22. https://doi.org/10.1016/j.cell.2019.01.022

Luo, S., Valencia, C. A., Zhang, J., Lee, N.-C., Slone, J., Gui, B., … Huang, T. (2018). Biparental Inheritance of Mitochondrial DNA in Humans. *Proceedings of the National Academy of Sciences*, *115*(51), 13039–13044. https://doi.org/10.1073/pnas.1810946115

Ma, H., Lee, Y., Hayama, T., Van Dyken, C., Marti-Gutierrez, N., Li, Y., … Mitalipov, S. (2018). Germline and somatic mtDNA mutations in mouse aging. *PLoS ONE*, *13*(7). https://doi.org/10.1371/journal.pone.0201304

Ma, J., Purcell, H., Showalter, L., & Aagaard, K. M. (2015). Mitochondrial DNA sequence variation is largely conserved at birth with rare de novo mutations in neonates. *American Journal of Obstetrics and Gynecology*, *212*(4), 530.e1-530.e8. https://doi.org/10.1016/j.ajog.2015.02.009

Mckenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., … Depristo, M. A. (2010). The Genome Analysis Toolkit : A MapReduce framework for analyzing next-generation DNA sequencing data, 1297–1303. https://doi.org/10.1101/gr.107524.110.20

Meirelles, F. V., & Smith, L. C. (1997). Mitochondrial genotype segregation in a mouse heteroplasmic lineage produced by embryonic karyoplast transplantation. *Genetics*, *145*(2), 445–451.

Meyer, J. N., Leung, M. C. K., Rooney, J. P., Sendoel, A., Hengartner, M. O., Kisby, G. E., & Bess, A. S. (2013). Mitochondria as a target of environmental toxicants. *Toxicological Sciences*, *134*(1), 1–17.

https://doi.org/10.1093/toxsci/kft102

Montoya, J., Playán, A., Solano, A., & Alcaine, M. (2000). Enfermedades del ADN mitocondrial, *31*(4), 324–333.

Moreno-Loshuertos, R., Pérez-Martos, A., Fernández-Silva, P., & Enríquez, J. A. (2013). Length variation in the mouse mitochondrial tRNAArg DHU loop size promotes oxidative phosphorylation functional differences. *FEBS Journal*, *280*(20), 4983–4998. https://doi.org/10.1111/febs.12466

Müller-Höcker, J. (1990). Cytochrome c oxidase deficient fibres in the limb muscle and diaphragm of man without muscular disease: An age-related alteration. *Journal of the Neurological Sciences*, *100*(1–2), 14–21. https://doi.org/10.1016/0022-510X(90)90006-9

Nicolson, G. L. (2014). Mitochondrial Dysfunction and Chronic Disease: Treatment With Natural Supplements. *Integrative Medicine (Encinitas, Calif.)*, *13*(4), 35–43. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/26770107%0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4566449

Nunnari, J., & Suomalainen, A. (2012). Mitochondria: In sickness and in health. *Cell*, *148*(6), 1145–1159. https://doi.org/10.1016/j.cell.2012.02.035

Payne, B. A. I., & Chinnery, P. F. (2015). Mitochondrial dysfunction in aging: Much progress but many unresolved questions. *Biochimica et Biophysica Acta - Bioenergetics*, *1847*(11), 1347–1353. https://doi.org/10.1016/j.bbabio.2015.05.022

Picard, M., Taivassalo, T., Gouspillou, G., & Hepple, R. T. (2011). Mitochondria: isolation, structure and function. *The Journal of Physiology*, *589*(18), 4413–4421. https://doi.org/10.1113/jphysiol.2011.212712

Rebolledo-Jaramillo, B., Su, M. S.-W., Stoler, N., McElhoe, J. A., Dickins, B., Blankenberg, D., … Makova, K. D. (2014). Maternal age effect and severe

germ-line bottleneck in the inheritance of human mitochondrial DNA. *Proceedings of the National Academy of Sciences*, *111*(43), 15474–15479. https://doi.org/10.1073/pnas.1409328111

Reddy, P., Ocampo, A., Suzuki, K., Luo, J., Bacman, S. R., Williams, S. L., … Izpisua Belmonte, J. C. (2015). Selective elimination of mitochondrial mutations in the germline by genome editing. *Cell*, *161*(3), 459–469. https://doi.org/10.1016/j.cell.2015.03.051

Roberts, R. C. (2016). Postmortem studies on mitochondria in schizophrenia. *Schizophrenia Research*, *187*, 17–25. https://doi.org/10.1016/j.schres.2017.01.056

Roth, A., Ding, J., Morin, R., Crisan, A., Ha, G., Giuliany, R., … Shah, S. P. (2012). JointSNVMix : a probabilistic model for accurate detection of somatic mutations in normal / tumour paired next-generation sequencing data, *28*(7), 907–913. https://doi.org/10.1093/bioinformatics/bts053

Sachadyn, P., Zhang, X., Clark, L. D., Naviaux, R. K., & Heber-katz, E. (2009). Mouse, *8*, 358–366. https://doi.org/10.1016/j.mito.2008.07.007.Naturally-Occurring

Sambasivarao, S. V. (2014). Mitochondrial DNA: impacting central and peripheral nervous systems. *Neuron*, *18*(9), 1–39. https://doi.org/10.1016/j.micinf.2011.07.011.Innate

Saunders, C. T., Wong, W. S. W., Swamy, S., Becq, J., Murray, L. J., & Cheetham, R. K. (2012). Strelka : accurate somatic small-variant calling from sequenced tumor – normal sample pairs, *28*(14), 1811–1817. https://doi.org/10.1093/bioinformatics/bts271

Shtolz, N., & Mishmar, D. (2019). The Mitochondrial Genome–on Selective Constraints and Signatures at the Organism, Cell, and Single Mitochondrion Levels. *Frontiers in Ecology and Evolution*, *7*(September), 1–9.

https://doi.org/10.3389/fevo.2019.00342

Stewart, J. B., & Chinnery, P. F. (2015). The dynamics of mitochondrial DNA heteroplasmy: Implications for human health and disease. *Nature Reviews. Genetics*, *16*(9), 530–542. https://doi.org/10.1038/nrg3966

Stoneking, M. (2000). Hypervariable sites in the mtDNA control region are mutational hotspots. *American Journal of Human Genetics*, *67*(4), 1029–1032. https://doi.org/10.1086/303092

TA, W. (2014). Cellular and molecular mechanisms of fibrosis. *Journal of Pathology*, *214*(November 2008), 199–210. https://doi.org/10.1002/path

Taanman, J. W. (1999). The mitochondrial genome: Structure, transcription, translation and replication. *Biochimica et Biophysica Acta - Bioenergetics*, *1410*(2), 103–123. https://doi.org/10.1016/S0005-2728(98)00161-3

Tuppen, H. A. L., Blakely, E. L., Turnbull, D. M., & Taylor, R. W. (2010). Mitochondrial DNA mutations and human disease. *Biochimica et Biophysica Acta*, *1797*(2), 113–128. https://doi.org/10.1016/j.bbabio.2009.09.005

Vafai, S. B., & Mootha, V. K. (2012). Mitochondrial disorders as windows into an ancient organelle. *Nature*, *491*(7424), 374–383. https://doi.org/10.1038/nature11707

Vivian, C. J., Hagedorn, T. M., Jensen, R. A., Brinker, A. E., & Welch, D. R. (2018). Mitochondrial polymorphisms contribute to aging phenotypes in MNX mouse models. *Cancer and Metastasis Reviews*, *37*(4), 633–642. https://doi.org/10.1007/s10555-018-9773-6

Wai, T., Teoli, D., & Shoubridge, E. A. (2008). The mitochondrial DNA genetic bottleneck results from replication of a subpopulation of genomes. *Nature Genetics*, *40*(12), 1484–1488. https://doi.org/10.1038/ng.258

Wanrooij, S., & Falkenberg, M. (2010). The human mitochondrial replication fork in

health and disease. *Biochimica et Biophysica Acta - Bioenergetics*, *1797*(8), 1378–1388. https://doi.org/10.1016/j.bbabio.2010.04.015

Wilson, I. J., Carling, P. J., Alston, C. L., Floros, V. I., Pyle, A., Hudson, G., … Chinnery, P. F. (2016). Mitochondrial DNA sequence characteristics modulate the size of the genetic bottleneck. *Human Molecular Genetics*, *25*(5), 1031–1041. https://doi.org/10.1093/hmg/ddv626