Facultad de Ingeniería
Escuela de Ingeniería en Bioinformática

Relación entre número de exonizaciones de elementos transponibles y metilaciones CpG del Reloj epigenético de Horvath

Loreto Farías Pavez

Profesor Tutor: Gonzalo Riadi

Co-Tutor: Bairon Hernández

Profesor Informante: Janin Riedelsberger

Memoria para optar al título de Ingeniera Civil en Bioinformática

Talca-Chile
Fecha de Examen de Título: 13 de julio 2022

CONSTANCIA

La Dirección del Sistema de Bibliotecas a través de su unidad de procesos técnicos certifica que el autor del siguiente trabajo de titulación ha firmado su autorización para la reproducción en forma total o parcial e ilimitada del mismo.

Talca, 2022

Agradecimientos

El presente trabajo de memoria de título lo dedico a:

A mi profesor tutor, PhD Gonzalo Riadi, por su comprensión en tiempos difíciles. Su paciencia y habilidades me motivaron a la superación propia cada semana.

A los docentes de carrera, quienes siempre tuvieron sus puertas abiertas para dudas y consultas, incluso cuando ya habían pasado años desde la última vez que tuvimos clases.

A mis padres, Myriam y Juvenal; a mis hermanos, Paula y Rodolfo; y a mi novio, Marcelo. Porque a pesar de que no siempre entendían lo que yo hacía, me apoyaron de manera incondicional durante el transcurso de toda mi carrera.

Finalmente, a mi pequeño/a garbancito, que a pesar de que aún no naces, eres un gran motivo para seguir adelante cada día. Te quiero mucho ♥.

Facultad de Ingeniería
Escuela de Ingeniería en Bioinformática

Relación entre número de exonizaciones de elementos transponibles y metilaciones CpG del Reloj epigenético de Horvath

Loreto Farías Pavez

Nombre : Gonzalo Riadi
Profesor Tutor

Nombre : Bairon Hernández
Co-Tutor

Nombre : Janin Riedelsberger
Profesor Informante

Talca-Chile.
Fecha de Examen de Título: 13 de julio de 2022

ÍNDICE DE CONTENIDO

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

RESUMEN

El envejecimiento es un conjunto de cambios fisiológicos y morfológicos en un organismo, que se acumulan con el paso del tiempo, y generan un deterioro progresivo de las funciones biológicas, por ejemplo, la reducción de fuerza muscular y otros aspectos cognitivos.

Existen al menos una decena de teorías que tratan de explicar el envejecimiento, las cuales se clasifican en tres grupos: teorías programadas, teorías de acumulación de daño y teorías combinadas. Las teorías de envejecimiento se basan principalmente en alteraciones que impactan a la cromatina mediante marcas epigenéticas, como son las metilaciones de ADN (ADNm), las cuales, en su mayoría disminuyen durante el envejecimiento. Estas marcas son las únicas que afectan directamente al ADN, suprimiendo la expresión y movimiento de elementos transponibles (TEs), y ocurren preferentemente en posiciones específicas del genoma llamadas islas CpG.

Las ADNm fueron utilizadas por Horvath para el desarrollo de su reloj epigenético, basado en teorías combinadas del envejecimiento, especialmente en la de pérdida de heterocromatina, el cual es un modelo predictor de la edad cronológica, que analiza el grado de metilación de islas CpG específicas en múltiples tejidos humanos.

Los sectores del genoma que son ricos islas CpG, susceptibles a la ADNm son, entre otras, las regiones de TEs adyacentes a genes, los cuales son importantes debido a que los TEs se encuentran extensamente repetidos en el genoma, y a pesar de que son insertados mayoritariamente en zonas no codificantes, pueden formar parte de nuevos exones (exonizaciones) gracias al splicing alternativo. Las exonizaciones de TEs pueden aumentar en número si las marcas del tipo ADNm disminuyen.

Aún no se ha demostrado que pueda existir una relación entre las exonizaciones de TEs y las ADNm que son analizadas por Horvath. Pero, si se logra demostrar que existe un aumento de exonizaciones de TEs durante el envejecimiento, y éstas se ven afectadas por ADNm de islas CpG de Horvath, se podrían relacionar a las exonizaciones de TEs con las islas CpG de Horvath. Debido a esto, nace la pregunta: “¿Existe relación entre el número de exonizaciones de TEs y las metilaciones de Horvath?”.

El presente trabajo, pretende recoger evidencia que apoye la idea de que existe un número creciente de exonizaciones de TEs durante el envejecimiento, y que dicho número podría estar relacionado con el grado de metilación de las islas CpG analizadas en el reloj de Horvath.

Se utilizan muestras sanas, de fastq y ADNm, de tejidos de rombencéfalo (cerebelo, bulbo raquídeo y puente de Varolio) humano en distintas edades, desde las 4 semanas de gestación a los 58 años. El número de exonizaciones de TEs de cada muestra (fastq) se obtiene a partir de herramientas bioinformáticas, principalmente Tophat. Las islas CpG se obtienen desde la investigación de Horvath. La información es analizada con respecto a la edad real de las muestras, y se espera encontrar correlación entre el número de exonizaciones de TEs y las islas CpG usadas por Horvath.

ABSTRACT

Aging is a set of physiological and morphological changes in an organism, which accumulate with the lenght of time and generate a progressive deterioration of biological functions, for example, the reduction of muscle strength and other cognitive aspects.

There are at least a dozen theories that try to explain aging, which are classified into three groups: programmed theories, damage accumulation theories and combined theories. Aging theories are mainly based on alterations that impact chromatin through epigenetic marks, such as DNA methylations (mDNA), which mostly decrease during aging. These marks are the only ones that directly affect DNA, suppressing the expression and movement of transposable elements (TEs), and occur preferentially in specific positions of the genome called CpG islands.

mDNAs were used by Horvath for the development of his epigenetic clock, based on combinatorial theories of aging, especially heterochromatin loss, which is a model predictor of chronological age that analyzes the degree of methylation of specific CpG islands in multiple human tissues. Sectors of the genome that are rich in CpG islands, susceptible to mRNA are, among others, TE regions adjacent to genes, which are important because TEs are found extensively repeated in the genome, and although they are mostly inserted in non-coding regions, they can form part of new exons (exonizations) by means of alternative splicing. Exonizations of TEs can increase in number if mRNA-like marks decrease.

It has not yet been demonstrated that there could be a relationship between TE exonizations and the mRNAs that are analyzed by Horvath. But, if it can be shown that there is an increase in TE exonizations during aging, and these are affected by Horvath CpG island mRNAs, then TE exonizations could be related to Horvath CpG islands. Because of this, the question arises, "Is there a relationship between the number of TE exonizations and Horvath methylations?".

The present work aims to gather evidence to support the idea that there is an increasing number of TE exonizations during aging, and that this number could be related to the degree of methylation of the CpG islands analyzed in the Horvath clock.

Healthy fastq and mDNA samples from human hindbrain tissues (cerebellum, medulla oblongata and pons) at different ages, from 4 weeks of gestation to 58 years old, are used. The number of TEs exonizations of each sample (fastq) is obtained from bioinformatics tools, mainly Tophat. The CpG islands are obtained from Horvath research. The information is analyzed with respect to the actual age of the samples, and it is expected to find correlation between the number of TEs exonizations and the CpG islands used by Horvath.

INTRODUCCIÓN

El envejecimiento

El concepto de vejez es tan usual como desafiante la tarea de encasillarlo en una única y simple definición. Debido a esto, se han generado diferentes definiciones para el término envejecimiento. Entre ellas destacan: La Enciclopedia Británica, que define al envejecimiento como cambios fisiológicos progresivos de un organismo, que conducen a la senescencia o deterioro de funciones biológicas (Rogers et al., 2020); el diccionario de Oxford lo define como un proceso de los seres vivos que comporta una serie de cambios que aparecen gracias al paso del tiempo y no por enfermedades ni accidentes (Oxford, s. f.); una definición biológica se refiere al cambio gradual en un organismo que conduce a un mayor riesgo de debilidad, enfermedad y muerte (Alvarado G & Salazar M, 2014); y como definición médica es, el conjunto de modificaciones morfológicas y fisiológicas que tienen lugar como consecuencia de la acción del tiempo sobre los seres vivos (Rico-Rosillo et al., 2018).

Una manera de unificar todas las definiciones anteriores es: El envejecimiento es un conjunto de cambios fisiológicos y morfológicos de un organismo, que se acumulan con el paso del tiempo, y generan un deterioro progresivo de las funciones biológicas, como es la reducción de fuerza muscular y otros aspectos cognitivos.

El envejecimiento afecta a todas las células del cuerpo, es más notable de forma física en tejidos como la piel y los músculos, pero los tejidos internos también se ven afectados por el envejecimiento, causando enfermedades. Los tejidos de interés para esta investigación son los del rombencéfalo (cerebelo, puente de Varolio y bulbo raquídeo), ya que se ha demostrado que existe una neurodegeneración en el cerebelo humano, lo que puede producir enfermedades de gran impacto como son la demencia senil, Alzheimer(Gellersen et al., 2021).

Como la población a nivel mundial envejece debido a que la esperanza de vida ha aumentado en los últimos años, se espera que para el 2030 una de cada seis personas en el mundo tendrá 60 años o más (OMS, 2021), la aparición de enfermedades relacionadas al envejecimiento han aumentado dramáticamente, es por esto, el envejecimiento ha tomado gran importancia para investigadores y científicos, que han

creado decenas de teorías generalizadas sobre por qué ocurre el envejecimiento, las cuales se dividen en tres grupos: (1) Teoría de envejecimiento Programado; (2) Teoría de acumulación de daño; (3) Teorías combinadas que unen teorías de los grupos anteriores (Pinto da Costa et al., 2016).

(1) Teorías de envejecimiento programado.

Las teorías del envejecimiento programado, o también llamadas teorías del envejecimiento activo, sugieren que el envejecimiento es en última instancia resultado de un mecanismo o programa biológico que intencionalmente causa o permite el deterioro y la muerte. Según estas teorías, el envejecimiento sigue un calendario biológico, que podría depender de la expresión génica, afectando los mecanismos de mantención, reparación y defensa del organismo (Adav & Wang, 2021; Goldsmith, 2012). Dentro de las teorías del envejecimiento programado se destaca la teoría de acortamiento de telómeros:

Teoría del Acortamiento de Telómeros

Esta teoría sugiere que el envejecimiento es causado por el acortamiento de los telómeros. Los telómeros son secuencias de ADN repetidas que se encuentran en los extremos de los cromosomas lineales, que no pueden ser replicadas completamente por las ADN polimerasas en las hebras retardadas durante el proceso de división celular, lo que genera su acortamiento con cada división de manera natural. Esto ocurre debido a la ausencia de la expresión de la telomerasa, enzima ribonucleoproteica, no presente en la mayoría de los mamíferos (Pinto da Costa et al., 2016; West et al., 2019), daños oxidativos y modificaciones epigenéticas como son las metilaciones de ADN (ADNm)

Esta teoría sostiene tres principios específicos: (1) que el envejecimiento está programado; (2) la detención irreversible del ciclo celular ocurre en respuesta al acortamiento de telómeros; y, (3) que el número total de divisiones celulares en ausencia de telomerasa no pueden exceder el límite Hayflick, límite que denota que las células somáticas se dividen un número fijo de veces, y las células humanas, como los fibroblastos, se dividen entre cuarenta y sesenta veces, antes de la senescencia celular (Bernadotte et al., 2016; Liu et al., 2019).

(2) Teorías de acumulación de daño.

Estas teorías sugieren que el envejecimiento están relacionadas a interacciones medioambientales que inducen daños acumulativos al ADN en varios niveles de los organismos, los cuales no son reparados, como por ejemplo los patógenos que generan enfermedades dañando al organismo (Jin, 2010; Maklakov & Immler, 2016).

Dentro de las teorías de acumulación de daño se destaca la teoría de inestabilidad del genoma.

La teoría de inestabilidad del genoma.

Esta teoría sugiere que la estabilidad e integridad del ADN se ve afectada por agentes físicos, químicos y biológicos exógenos (externos al organismo) y endógenos (internos del organismo) como son los errores de replicación del ADN, reacciones hidrolíticas espontáneas y especias reactivas de oxígeno (ROS) (López-Otín et al., 2013).

Las lesiones genéticas ocasionadas por inestabilidad genética incluyen mutaciones puntuales, translocaciones, ganancia y pérdidas cromosómicas, acortamiento de telómeros y disrupción genética a causa de la integración de virus o transposones.

La inestabilidad del genoma puede producir acumulación de daños genéticos además de numerosas enfermedades de envejecimiento prematuro, como el síndrome de Huschinson-Gilford (progeria) y el síndrome de Werner (López-Otín et al., 2013; Swahari & Nakamura, 2016).

La teoría de inestabilidad del genoma apunta a la maquinaria de reparación del ADN presenta defectos como errores en la replicación del ADN y cambios espontáneos (mutaciones y cambios epigenéticos) que podrían estar relacionados a la inestabilidad genómica y a la senescencia (López-Otín et al., 2013; Pinto da Costa et al., 2016).

(3) Teorías combinadas.

Las teorías combinadas se desarrollaron con la intención de unificar teorías del envejecimiento programado con las teorías de acumulación de daño. El proceso de envejecimiento se considera en un grado más completo y global, pero aun así, los

factores que afectan al envejecimiento no pueden explicarse completamente (Pinto da Costa et al., 2016). Entre ellas, se destaca la teoría de pérdida de heterocromatina:

La teoría de pérdida de heterocromatina.

La heterocromatina corresponde a regiones de la cromatina que se encuentran compactadas y son en su mayoría transcripcionalmente silenciosas en comparación con regiones no compactadas (eucromatina). La heterocromatina se divide en constitutiva (segmentos de centrómeros y telómeros que adoptan la formación silenciosa) y facultativa (encontrado en el cromosoma X inactivado) (Tsurumi & Li, 2012). La teoría de pérdida de heterocromatina propone que los dominios de heterocromatina establecidos en la embriogénesis se descomponen durante el proceso de envejecimiento, contribuyendo a la desrepresión de genes silenciados y conduciendo a patrones de expresión génica aberrante, transcribiéndose elementos del genoma que son generalmente silenciados, como es el caso de los elementos transponibles. Durante el proceso de envejecimiento, se produce una disminución global de heterocromatina constitutiva lo que conduce a un aumento de heterocromatina facultativa (Andersen et al., 2017; Tsurumi & Li, 2012). La regulación de la heterocromatina ocurre mediante marcas epigenéticas de modificación de histonas, como es la acetilación (y desatelación), metilación (y desmetilación) y fosforilación (Bannister & Kouzarides, 2011).

Las mutaciones que producen pérdida de heterocromatina en la lámina nuclear de la línea germinal, es decir, en el entramado proteico encontrado en el núcleo de las células que separa la membrana interna del núcleo de la cromatina, en el proceso de desarrollo embrionario, pueden producir enfermedades que imitan al envejecimiento, como es el caso de los síndromes de progeria y de Werner (Guo & Fang, 2014; Swahari & Nakamura, 2016; Tsurumi & Li, 2012).

A modo de resumen, se puede decir que, durante el envejecimiento, el estado de la cromatina de un individuo se ve modificado (**¡Error! La autoreferencia al marcador no es válida.**), debido a la incorporación de diferentes variantes de histonas, patrones de metilación de ADN de histonas alterados, lo que da como resultado modificaciones en la compactación de la cromatina (Pal & Tyler, 2016).

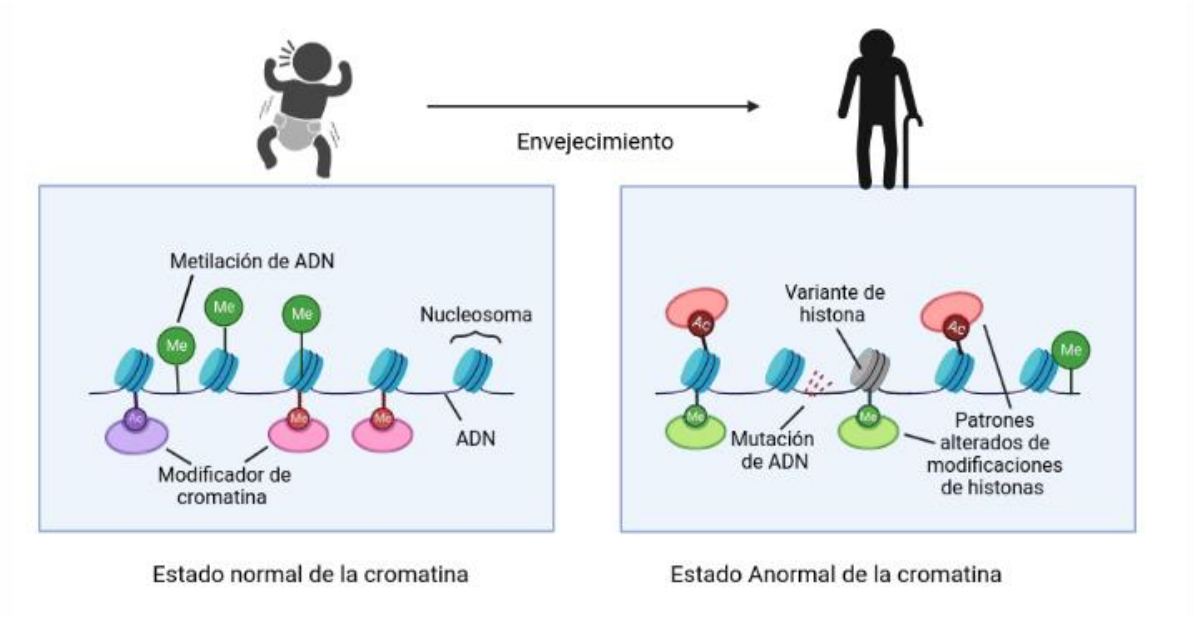


Fig. 1 Estado de la cromatina durante el envejecimiento.

Durante el envejecimiento, la información epigenética cambia en respuesta a factores exógenos y endógenos. El estado anormal de la cromatina resultante se caracteriza por la incorporación de diferentes variantes de histonas, patrones de metilación del ADN y modificadores de histonas alterados, lo que da como resultado el reclutamiento de diferentes transformadores de cromatina. El estado anormal de la cromatina en las células senescentes incluye patrones de transcripción modificados y deriva transcripcional dentro de la población. Este estado también conduce a la inserción de nuevos elementos transponibles en el genoma y a la inestabilidad genómica, incluidas las mutaciones del ADN.

Figura creada con Biorender[1] , basada en investigación de Pal & Tylor (Pal & Tyler, 2016).

Las alteraciones de eucromatina (cromatina ligeramente compactada) y heterocromatina producen patrones más permisivos o más restrictivos de la expresión génica, variando la tasa de envejecimiento de un individuo (West et al., 2019).

La expresión de genes puede ser regulada a distintos niveles: transcripcional (a nivel de síntesis de ARN), postranscripcional (corte y empalme), traduccional (síntesis de proteínas a partir de ARN mensajero), y postraduccional (modificaciones postraduccionales) (Cavagnari, 2012).

La forma de regular la expresión génica, que toma en cuenta la relación de los genes con el medio ambiente, está a cargo de los mecanismos epigenéticos. Entre los mecanismos más estudiados, se encuentran las modificaciones de histonas, las

metilaciones de ADN, y el silenciamiento genético mediado por ARN no codificantes (Cavagnari, 2012).

Marcas epigenéticas

La epigenética es el estudio de cambios hereditarios en la expresión génica que no implican cambios en las secuencias de ADN, causados por la edad y expresión de factores ambientales (alimentación, ejercicio y sustancias químicas) (Cavagnari, 2012; NIH, 2011; Teng et al., 2020). Es capaz de modular y regular la expresión de genes a través de modificaciones químicas que cambian la conformación espacial de la cromatina: compactándola, evitando la unión de factores de transcripción al ADN; o des compactándola, permitiendo la unión de factores de transcripción (Tiffon, 2018).

En la actualidad, se sabe que los cambios epigenéticos ocurren mediante marcas químicas que son detectables en el genoma. Estás marcas pueden traspasarse a las siguientes generaciones o perderse a medida que el individuo envejece (García Robles & Ayala Ramírez, 2012; Teng et al., 2020).

Los mecanismos epigenéticos más estudiados que intervienen en la regulación de la transcripción (expresión génica) son: el silenciamiento genético mediado por ARN no codificante (micro ARN) (Cavagnari, 2012; Pal & Tyler, 2016), las modificaciones postraduccionales de las histonas(Plunk & Richards, 2020) y la metilaciones de ADN (Cavagnari, 2012).

De estas marcas epigenéticas se destacan las metilaciones de ADN (ADNm), las que son fundamentales para el desarrollo de esta investigación.

Metilación del ADN

La metilación de ADN (ADNm) desempeña un rol importante en la regulación de la expresión génica, alterando el fenotipo sin perturbar el genotipo (Ecker & Beck, 2019). Las enzimas que metilan el ADN se dirigen a los residuos de citosina ubicados en dinucleótidos de citosina-fosfato-guanina (CpG), también llamadas germline differentially methylated regions (regiones germinales metiladas diferencialmente o gDMRs). Añadiendo un metilo en la posición 5 del anillo de la citosina (5-metilcitosina), conduciendo a la represión epigenética de la expresión génica y la cual disminuye

notablemente a medida que el organismo envejece (Benoit, 2014; Plunk & Richards, 2020; Sturm et al., 2015).

La metilación de la citosina en los sitios CpG es la única modificación epigenética conocida que afecta directamente la molécula de ADN, y es necesaria para el desarrollo embrionario en mamíferos, los cuales están marcados por una alta frecuencia de sitios CpG (Gabory et al., 2011; Plunk & Richards, 2020).

Durante el envejecimiento, existe una reducción de la actividad de la ADN metiltransferasa 1, que contribuye a la disminución de la metilación global del ADN. Este evento es asociado con el riesgo de desarrollar enfermedades relacionadas con el envejecimiento como son la ateroesclerosis, cáncer y enfermedades autoinmunes (Rico-Rosillo et al., 2018).

La disminución progresiva en la metilación de ADN tiene lugar preferentemente en los elementos transponibles dispersos a lo largo de todo el genoma, ya que estos elementos son ricos en islas CpG y generalmente se encuentran altamente metilados, debido a que la ADNm reprimen su expresión y actúa limitando su potencial genotóxico, es decir, su capacidad para afectar la replicación del ADN, impactando negativamente la integridad del material genético.(Jansz, 2019; Plunk & Richards, 2020). El potencial genotóxico se puede medir utilizando técnicas para determinar la reparación del ADN, las aberraciones cromosómicas, niveles de apoptosis y necrosis, entre otros (Nai et al., 2015).

El uso del marcador epigenético del tipo ADNm no se limita únicamente a la expresión y represión de genes. Esto fue comprobado el investigador genetista Steve Horvath, que demostró que el grado de metilación en las islas CpG está correlacionada con la edad cronológica de las muestras, desarrollando uno de los primeros mecanismos de estimación de edad biológica, el Reloj Epigenético de Horvarth (Horvath, 2013).

Reloj Epigenético de Horvath

A partir de una investigación previa (Hannum et al., 2013), Steve Horvath corroboró la existencia de la relación entre las metilaciones de ADN en islas CpG específicas y la edad cronológica de tejidos humanos. Hannum, con solo 71 CpG (de las cuales solo

7 de ellas se encuentran en la plataforma Illumina 27K), logró predecir la edad de muestras del tejido de sangre total humana con una mediana de error absoluto (MAE) de 8 años y una correlación de 0.82 (alta)(Hannum et al., 2013). Sin embargo, fue Horvath quien logró mejorar la predicción de Hammun, aumentando el número de islas CpG específicas utilizadas, mencionadas en el Archivo suplementario 2 (Horvath, 2013), que contiene el proceso de la creación del reloj multitejido de Horvath, detallando las fórmulas que permiten calcular el grado de metilación de las muestras con edades conocidas y creando un modelo de predicción de edad a partir de muestras de metilación con edades desconocidas.

A la fecha, Horvath ha publicado 4 diferentes relojes epigenéticos (multitejido (Horvath, 2013), Skin and Blood (Ecker & Beck, 2019; Horvath et al., 2018), PhenoAge (Levine et al., 2018) y GrimAge (Lu et al., 2019)), los cuales utilizan información de ADNm de muestras de uno o varios tejidos a diferentes edades cronológicas, para predecir sus edades biológicas a partir del grado de metilación que presentan. La principal diferencia entre estos relojes es que su diseño se basa en el análisis del grado de metilación de distintas islas CpG para cada reloj, con el objetivo de generar predictores con una MAE más baja para los tejidos o muestras, por lo tanto, el uso de distintas islas CpG para el diseño de los relojes epigenéticos de Horvath genera una mayor especificidad para ciertos tejidos.

Por propósito de esta investigación, solo se menciona el reloj multitejido, ya que es éste el que fue utilizado posteriormente.

Reloj multitejido

Corresponde al primer método de estimación de la edad epigenética que funciona con precisión (MAE de 3.6 años) en casi todos los tejidos y tipos de células humanas, como son: sangre total, células mononucleares de sangre periférica, muestras cerebelosas, corteza occipital, epitelio bucal, colon, tejido adiposo, hígado, pulmón, saliva y cuello uterino (Ecker & Beck, 2019; Horvath, 2013).

El reloj multitejido está diseñado a partir de 353 islas CpG específicas (Archivo suplementario 3 Horvath, 2013) y analiza sets datos de ADNm pre-existentes de muestras de diferentes tejidos y edades (Horvath, 2013), los que se obtienen a partir

de la tecnología BeadChip de Illumina con montaje infinium, técnica de detección de islas CpG (Pidsley et al., 2016), la información se encuentra depositada en la plataforma de libre acceso GEO(Clough & Barrett, 2016).

El reloj Multitejido cuenta con una plataforma web de acceso rápido, en donde se depositan un archivo con formato de columnas separadas por coma (.csv) que contiene los IDs de las islas CpG y los valores de metilación de cada una de ellas.

Hoy en día, se han identificado tres sectores del genoma que son ricos en islas CpG, y por lo tanto son más susceptibles a cambios epigenéticos del tipo ADNm: regiones promotoras de housekeeping genes, o genes domésticos, expresados en casi todas las células de un organismo y son capaces de proporcionar funciones para mantener todos los tipos de células vivas; elementos reguladores de imprinted genes, o genes impresos, se expresa solo en el cromosoma heredado de uno de los padres, poseen marcas que permiten diferenciar los alelos; y elementos transponibles adyacentes a genes (Plunk & Richards, 2020). Estos últimos son importantes ya que se ha demostrado que estos elementos se encuentran altamente metilados, para poder suprimir su función de modificación del ADN y así mantener la estabilidad e integridad del genoma(Alexeeff et al., 2013).

Elementos Transponibles

Los elementos transponibles (TEs) son secuencias de ADN que tienen la capacidad de movilizarse por el genoma y se encuentran altamente repetidos. El genoma humano está compuesto por entre 45% y cerca del 60% de elementos repetidos, de los cuales aproximadamente el 80%, corresponde a TEs (**¡Error! No se encuentra el origen de la referencia.**) (Benoit, 2014; Bourque et al., 2018; LaRocca et al., 2020).

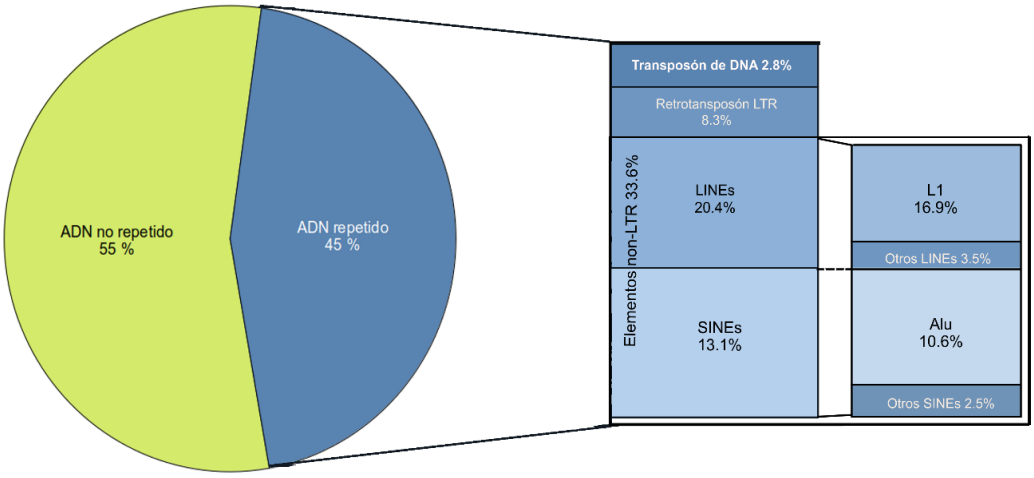


Fig. 2 Proporción de elementos transponibles en el genoma humano.

La figura muestra que el porcentaje de elementos repetidos en el genoma humano corresponde al 45%, del cual 2.8% corresponden a Transposones de ADN y el resto corresponde a retrotransposones Figura extraída extraía desde investigación de Benoit (Benoit, 2014).

Movilización y Regulación de TEs

Los TEs tienen un impacto importante en la regulación génica de todo el genoma, ya que, al tener la capacidad de movilizarse por el genoma, pueden afectar directamente la secuencia de ADN. Debido a ello, generalmente los TEs se encuentran silenciados o reprimidos mediante marcas epigenéticas de ADNm, las cuales a su vez pueden alterar la expresión de un gen (Bourque et al., 2018).

Los TEs se mueven en el genoma, y como consecuencia un alto número de copias, pueden inducir diferentes tipos de reordenamientos, los cuales incluyen deleciones, duplicaciones, inversiones y translocaciones (Warren et al., 2015). La movilización de TEs puede conducir a inestabilidad genómica, por ejemplo, si un TE salta a una región

funcional (codificadora o reguladora) del genoma, la inserción puede dar como resultado la pérdida de función, lo que podría inducir la muerte de la célula afectada. La ocurrencia masiva de eventos de transposición puede conducir a varios procesos degenerativos (O'Donnell & Burns, 2010). Sin embargo, La transposición de TEs se lleva a cabo generalmente en regiones intrónicas (no codificantes) del genoma, y raramente en regiones exónicas (codificantes), pero de ocurrir una inserción en una región codificante, la secuencia del TE podría pasar a ser parte del transcrito del ARN mensajero (ARNm) y generar posibles proteínas aberrantes.

Los TEs han enriquecido el transcriptoma humano mediante intronizaciones (creación de nuevos intrones) y exonizaciones (creación de nuevos exones), estas últimas se llevan a cabo principalmente a partir de splicing alternativo de TEs transpuestos en el genoma (Sela et al., 2010). El splicing alternativo es el proceso por el cual un ARNm inmaduro puede codificar para diferentes proteínas. En humanos, el 100% de los genes producen al menos dos variantes de ARNm alternativas (Lee & Rio, 2015).

Los TEs se dividen en dos clases principales según su mecanismo de transposición (clase I, Retrotransposones; y Clase II, Transposones de ADN) *(Bourque et al., 2018)* (Fig. 3).

En humanos la mayoría de las exonizaciones de TEs son originadas a partir del transposón específico de primates llamado Alu, que corresponden a la clase I de TEs (Sela et al., 2010).

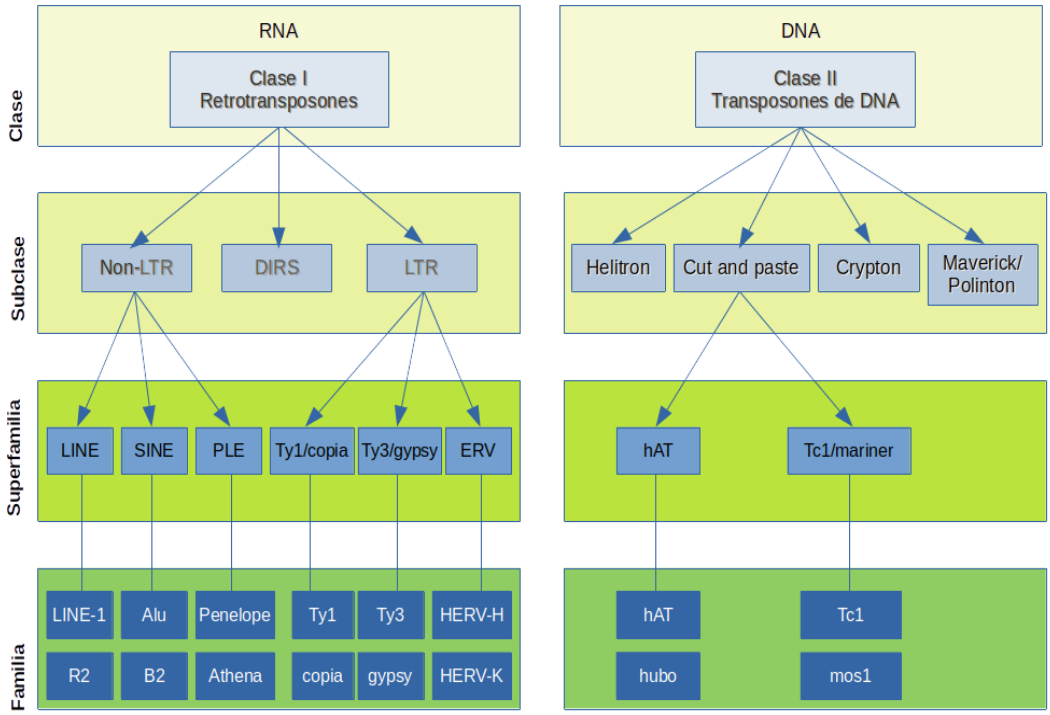


Fig. 3 Resumen de las clases de elementos transponibles en eucariotas.

Se pueden observar las relaciones entre clases, subclases, superfamilias y familias de los TEs. Cada clase se subdivide en subclases según el mecanismo de integración cromosómica. Cada subclase de TE se divide en subgrupos (o superfamilias) que normalmente se encuentran en una amplia gama de organismos, pero comparten una organización genética común y un origen monofilético (con un mismo antepasado en común). En el nivel más detallado de la clasificación de TEs, los elementos se agrupan en familias o subfamilias, que son un grupo de elementos estrechamente relacionados que se pueden rastrear como descendientes de una sola unidad ancestral (Andrenacci et al., 2020).

Figura extraída y modificada desde el artículo de Bourque (Bourque et al., 2018)

Clases de TEs

Clase I

La clase I de TEs, también conocidos como retrotransposones, se movilizan a través de un mecanismo de "copiar y pegar" mediante el cual un intermedio de ARN se transcribe de forma inversa (gracias a la proteína transcriptasa reversa), en una copia de ADNc que se integra en otra posición del genoma Fig. 4a). Para los retrotransposones de repetición terminal larga (LTR), la integración se produce mediante una reacción de escisión y transferencia de cadena catalizada por una

integrasa muy parecida a los retrovirus. Para los retrotransposones no LTR, que incluyen elementos nucleares intercalados tanto largos como cortos (LINE y SINE), la integración cromosómica se acopla a la transcripción inversa a través de un proceso denominado transcripción inversa cebada por objetivo (Bourque et al., 2018).

Clase II

Los elementos de clase II, también conocidos como transposones de ADN, se movilizan de manera completa a través del ADN, puede movilizarse con el mecanismo de "cortar y pegar". La proteína ADN transposasa, corta la secuencia donde se encuentra el TE (escisión) y lo vuelve a insertar en otro sector del genoma (Fig. 4b) (Bourque et al., 2018).

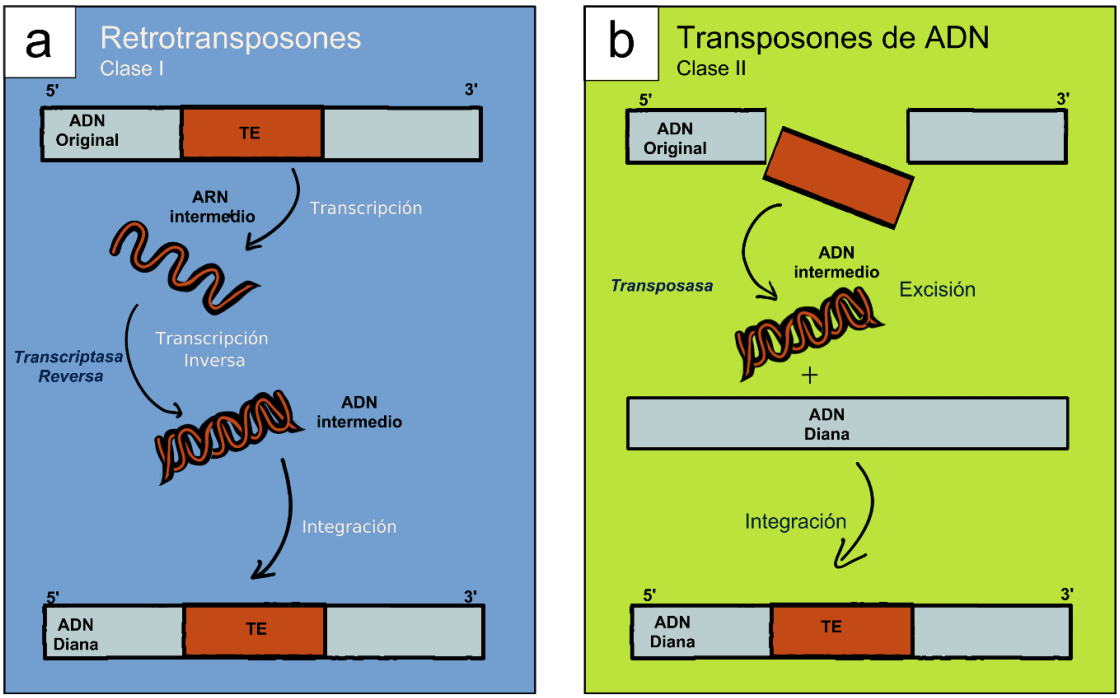


Fig. 4 Mecanismo de movilización de elementos transponibles.

(a) Clase I: Retrotransposones, ocurre transcripción inversa con ayuda de la enzima transcriptasa reversa. (b) Clase II. Transposones de ADN, mediante el mecanismo "Cortar y Pegar", el desplazamiento ocurre directamente desde un lugar del genoma a otro, mediante la enzima transposasa (Bourque et al., 2018).

Los TEs pueden verse afectados comúnmente por las marcas epigenéticas del tipo ADNm, y la presencia o ausencia de estas marcas puede conducir a importantes daños en el genoma. Por ejemplo, la hipermetilación derivada de elementos Alu humanos

puede silenciar los genes supresores de tumores, siendo de gran importancia para enfermedades como el cáncer (Saleh et al., 2019). Además, la hipometilación del promotor LINE-1 puede activar una transcripción alternativa de la MET oncogén en tumores de vejiga; induciendo la expresión ectópica de genes (expresión “fuera de lugar” de genes ejerciendo su función donde generalmente no lo hacen) y posiblemente alterar la susceptibilidad a enfermedades (Saleh et al., 2019; Wolff et al., 2010).

TEs y Envejecimiento

A lo largo de la vida del humano, se ha detectado una disminución sign*ificativa en la metil*ación promedio de Alu (el SINE más común de primates), por lo que se puede presumir, que la edad se asocia negativamente con los niveles de metilación de las secuencias Alu (Jintaridth & Mutirangura, 2010; Sela et al., 2010).

La activación de los retrotroelementos se ha observado e implicado en una variedad e transtornos neurológicos, además de patologías relacionadas con el envejecimiento que incluyen cáncer, diabetes, enfermedades cardiovasculares y neurodegenerativas (Brunet & Berger, 2014; Saleh et al., 2019).

Se ha demostrado que los mecanismos que reprimen la actividad de LINE-1 son menos eficientes durante el proceso de envejecimiento. Estudios comprueban que se produce un aumento de los niveles de transcripción de LINE-1 en células senescentes. La acumulación de ADNc de LINE-1 citoplasmático impulsa la expresión de fenotipo secretor asociado a la senescencia (De Cecco et al., 2019; Saleh et al., 2019).

Exonizaciones y Exonizaciones de TEs

La exonización es un proceso mediante el cual los genes obtienen nuevos exones, producida a partir de mutaciones en los intrones, es decir, de secuencias de ADN que no codifican proteínas, a menudo dicha secuencia corresponde a un TE debido a la presencia de estructuras internas similares a sitios de empalme dentro de sus secuencias (Huda & Bushel, 2013).Las exonizaciones se encuentran restringidas a limitaciones evolutivas como: que los nuevos exones generalmente pasan por el

proceso de Splicing alternativo y su tasa de inclusión es baja (5.2% de las exonizaciones pasan a ser parte de una región codificante) (Huda & Bushel, 2013; Sorek et al., 2002).

Los elementos transponibles se encuentran en regiones codificantes de proteínas de ~4% de los genes humanos, y los elementos Alu representan aproximadamente un tercio de esas inserciones. La insersión de un elemento Alu en un ARNm maduro puede causar una enfermedad genéticas, sin embargo la misma inserción puede contribuir a la variedad y versatilidad de la proteína (Sorek et al., 2002).

La inserción de la secuencia (o parte de ella) de un elemento transponible a un exón, puede ocurrir en técnicamente dos posiciones, después del exón (al lado derecho), como se muestra en (a) de la Fig. 5; antes del exón (lado izquierdo) como se muestra en (b) de la Fig. 5. Además, una inserción puede unir dos exones, insertándose al lado derecho de un exón y al lado izquierdo de un segundo exón, como en (c) de la Fig. 5 (Sela et al., 2010).

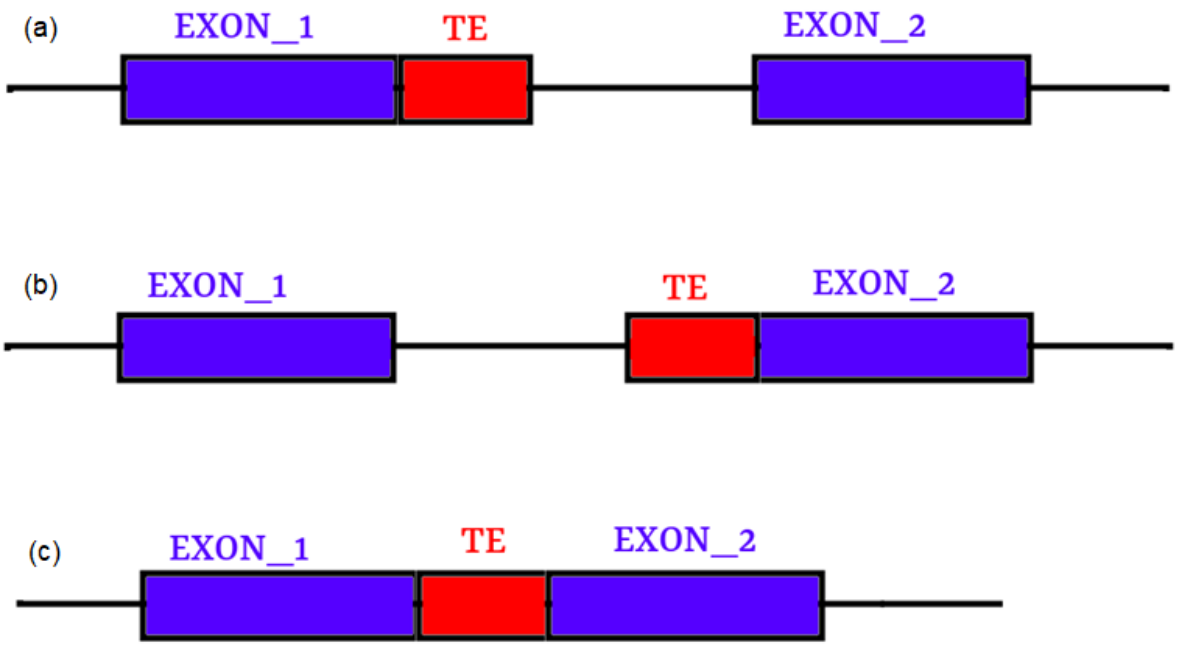


Fig. 5 Inserción de TE en el Genoma para la formación de un nuevo exón

La figura muestra las posiciones en las que se puede insertar un TE (rojo) al exón (axúl). Figura creada a partir de lo mencionado por Sela en su artículo (Sela et al., 2010).

Las exonizaciones de TEs pueden estar relacionadas al envejecimiento, ya que los TEs se encuentran altamente relacionados a las islas CpG, siendo en ocasiones casi un tercio de su secuencia, por consecuencia, son metilados, evitando que se movilicen por el genoma (Li & Zhang, 2014; Sorek et al., 2002). Debido a la presencia de islas CpG en los TEs es posible pensar que al momento de exonizarse un TE, este pueda contener alguna de las islas CpG usadas por Horvath para predecir la edad de las muestras (Horvath, 2013).

PROBLEMÁTICA

Las teorías del envejecimiento que se encuentran relacionadas con la metilación del ADN (ADNm) mencionan que la falta o exceso de ADNm genera inestabilidad génica o genómica.

Las marcas epigenéticas producidas por la ADNm son utilizadas por Horvath para predecir el envejecimiento de tejidos humanos. Además, ADNm se encuentra relacionada elementos transponibles (TEs), siendo uno de los principales mecanismos de regulación de su expresión. Si estos elementos no son regulados, es decir, si se movilizaran libremente por el genoma, pueden generar inestabilidad genómica, produciendo enfermedades y/o síndromes. Es decir, la metilación de ADN es un proceso de gran importancia para la regulación génica, es el eje central para el desarrollo del reloj de Horvath y uno de los principales represores de la expresión de elementos transponibles.

La disminución de las marcas epigenéticas durante el envejecimiento, disminuyan el control sobre la expresión de genes y TEs. Si este proceso afecte a genes que tengan funciones importantes como de splicing alternativo, podría desregular la expresión de variantes de genes, aumentando el número de exonizaciones de TEs a medida que el organismo envejece.

Aún no se ha demostrado que pueda existir una relación entre las exonizaciones de elementos transponibles y las metilaciones que son analizadas por el reloj de Horvath. Sin embargo, este es un punto de interés, ya que las ADNm se llevan a cabo en las islas CpG, las cuales se encuentran en gran medida en los elementos transponibles. Si se lograra demostrar que existen exonizaciones de TEs en los mismos sectores donde se producen las ADNm del reloj de Horvath, se podrían relacionar a las exonizaciones de TEs en el proceso de envejecimiento. Debido a esto, nace la pregunta: “¿Existe relación entre el número de exonizaciones de TEs y las metilaciones de Horvath?”.

Para poder responder a esta pregunta, se propone calcular el número de islas CpG de Horvath presentes en las muestras que se encuentren relacionadas con exonizaciones de TEs a lo largo de la vida. Esto se lleva a cabo en muestras de tejidos de

rombencéfalo de *Homo sapiens (humano)* de diferentes edades, mediante el uso de dos sets de datos, el primero de metilaciones de islas CpG (17 muestras intercaladas, desde 1 a 60 años) y el segundo mediante archivos en formato FASTQ de secuenciación RNA-seq (57 muestras intercaladas, desde 4 semanas de gestación a 58 años).

El set de metilaciones se encuentra depositada en la base de datos GEO (Clough & Barrett, 2016) de NCBI, posteriormente es procesada utilizando la plataforma del reloj multitejido de Horvath[1].

El set en formato FASTQ para la obtención de las exonizaciones de TEs relacionados al envejecimiento, se obtendrán a partir de archivos RNA-seq (desde la base de datos SRA[2]), posteriormente son procesados mediante un pipeline modificado de la investigación de Wang (Wang et al., 2016) y herramientas bioinformáticas: Bowtie (Langmead & Salzberg, 2012), Tophat(Kim & Salzberg, 2011) y Bedtools(Aaron, 2021).

La investigación está centrada en tejidos de rombencéfalo de *Homo sapiens*, ya que los tejidos que conforman el rombencéfalo, es decir, cerebelo, bulbo raquídeo y puente de Varolio, se encuentran ampliamente relacionados a enfermedades neurodegenerativas relacionadas con el envejecimiento como son el Parkinson, Alzheimer y demencia senil (Gellersen et al., 2017).

1 Plataforma de Predictor de edad, reloj epigenético de Horvath (https://ADNmage.genetics.ucla.edu/)
2 Plataforma SRA (Sequence Read Archive https://www.ncbi.nlm.nih.gov/sra)

HIPÓTESIS Y OBJETIVOS

Hipótesis.

Existe relación entre el número de exonizaciones de TEs y el número de metilaciones de Horvath en rombencéfalo a lo largo de la vida del humano.

Objetivos.

Objetivo General.

Relacionar el número y tipos de exonizaciones de TEs con el número de metilaciones de Horvath a lo largo del tiempo.

Objetivo específicos y actividades.

1 **Cuantificar las metilaciones de Horvath a lo largo de la edad.**

1.1 Búsqueda y extracción coordenadas de metilaciones CpG del reloj epigenético de Horvath en tejido rombencéfalo Humano.

1.2 Procesamiento de análisis cualitativo preliminar de información obtenida con el reloj epigenético de Horvath.

2 **Cuantificar exonizaciones de TEs.**

2.1 Descarga de genoma *Homo sapiens*, RepeatMasker y set de RNA-seq de tejidos de rombencéfalo.

2.2 Obtención de archivo de exonizaciones relacionadas a TEs.

3 **Relacionar exonizaciones de TEs, metilaciones de Horvath y edades cronológicas.**

3.1 Relacionar por separado el número de exonizaciones de TEs y metilaciones CpG de Horvath con la edad de las muestras en cada set de datos.

3.2 Extracción de coordenadas de metilaciones y exonizaciones de TEs.

3.3 Intersección de coordenadas de islas CpG usadas por de Horvath y de exonizaciones de TEs.

3.4 Visualización, análisis e interpretación de gráficos y archivos resultantes de la relación entre metilaciones de Horvath y exonizaciones de TEs.

RECURSOS

Set de Metilación de ADN de tejidos de rombencéfalo a diferentes edades de humano

Los valores de metilación utilizados para esta investigación son los mencionados por Horvath en su artículo (Horvath, 2013), el set de metilaciones cuenta con archivos separados de 17 muestras desde 1 hasta los 60 años de edad. En la Fig. 6 se observa un histograma de frecuencias con la información de las edades e las 17 muestras de metilaciones utilizada.

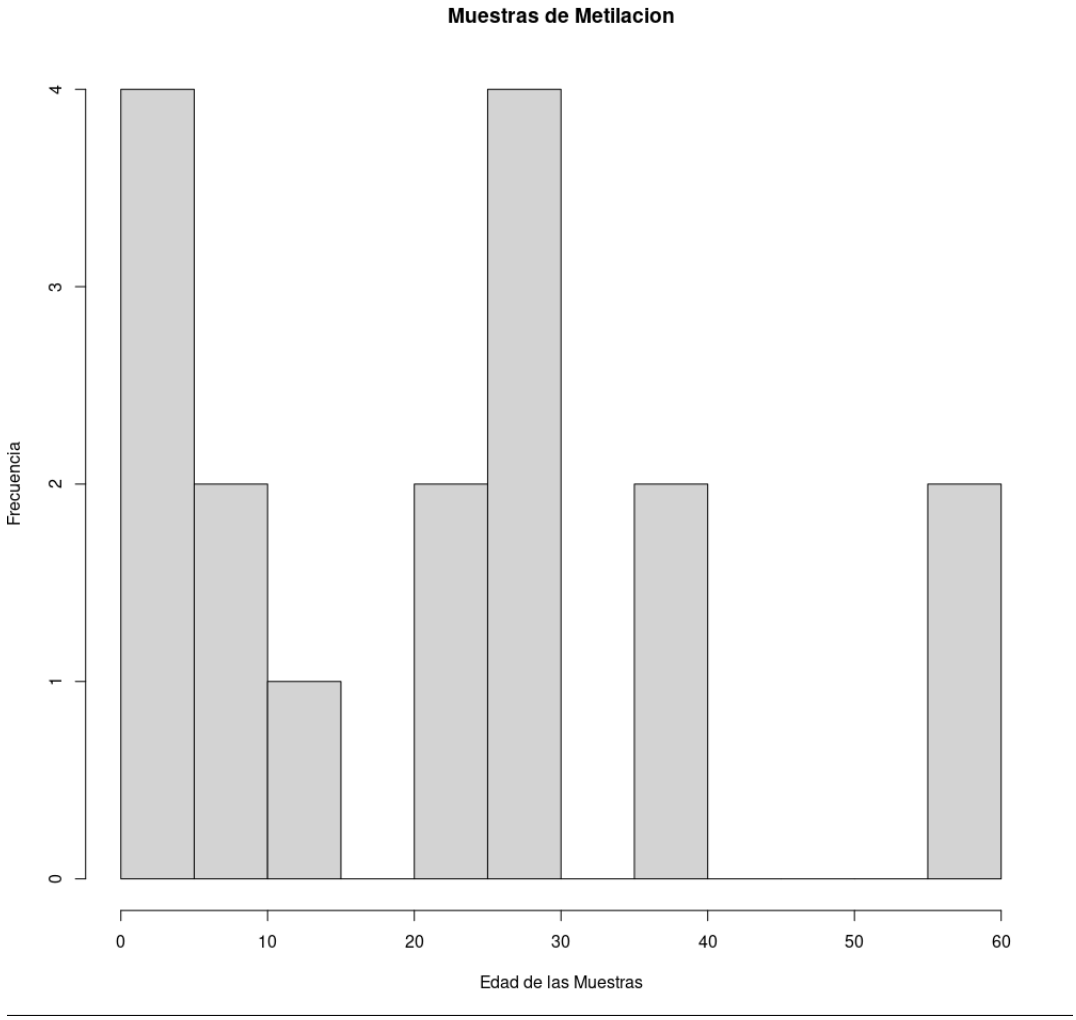


Fig. 6 Histograma de Frecuencias de edades en muestras de metilación

Las muestras de tejido rombencefálico de *Homo sapiens* mencionadas por Horvath fueron descargadas a partir de la base de datos Gene Expression Omnibus (GEO) de NCBI. GEO[3] es un repositorio público que acepta datos procesados y sin procesar para estudios de expresión génica, genómica funcional de alto rendimiento, metilación del genoma, perfiles de proteínas, entre otros (Clough & Barrett, 2016). El set de datos fue descargado en formato de matriz de series, en dos columnas, la primera con los IDs de cada isla CpG y la segunda con el grado de metilación de cada una de ellas.

Genoma humano de referencia

El genoma de *Homo sapiens* (GRCh37/hg19) fue utilizado. Además, se descargó la anotación de genes y elementos transponibles (TEs) correspondientes a dicha versión del genoma. Esta información se obtuvo a partir de las bases de datos UCSC Browser[4] (Navarro Gonzalez et al., 2020).

Set de RNA-seq tejidos de rombencéfalo a diferentes edades de humano

El set de RNA-Seq utilizado corresponde a tejido rombencefálico en formato FASTQ de la investigación de Cardoso-Moreira et al., 2019. Las muestras de tejidos analizados en la investigación de Cardoso-Moreira corresponden a rombencéfalo humano desde las 4 semanas de gestación, hasta los 58 años de edad. En total corresponden a 59 librerías, de las cuales fueron eliminadas dos que poseían calidades y número de reads inferiores al resto, quedando un total de 57 archivos FASTQ. La Tabla Anexa 2 señala las muestras que utilizadas con su respectivo identificador SRA y la edad de cada una de las muestras (Leinonen et al., 2011).

Todas las librerías relacionadas a *Homo sapiens* que se mencionan en la investigación de Cardoso-Moreira (Cardoso-Moreira et al., 2019) se encuentran depositadas en la base de datos ArrayExpress de EMBL-EBI, con el código de acceso E-MTAB-6814 (EMBL-EBI, 2021).

Software Bioinformáticos

Se utilizó SRAtoolkit (Klymenko, 2014/2021), para descargar los archivos SRA (.fastq), FastP (Chen et al., 2018) para preprocesar los archivos FASTQ, FastQC[5] analizar los

3 Plataforma GEO (Gene Expression Omnibus https://www.ncbi.nlm.nih.gov/geo/)
4 Plataforma del Instituto de Genómica de la Universidad de Santa Cruz, California (UCSC Genome Browser https://genome.ucsc.edu/)
5 FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/)

parámetros de calidades de las muestras y MultiQC (Ewels et al., 2016) para visualizarlos en conjunto, Bowtie (Langmead et al., 2009), alineador de reads contra el genoma de referencia, con el que se generan índices del genoma que posteriormente fueron utilizados para mapear los reads de los distintos archivos SRA; TopHat (Kim & Salzberg, 2011), con su función tophat-fusion se identificó exonizaciones de TEs mediante el pipeline editado de Wang (Wang et al., 2016). Bedtools (Aaron, 2021), permitió identificar la relación entre las exonizaciones de TEs y las metilaciones de Horvath a través de sus coordenadas. R y sus paquetes, ggplot2, tidyr pheatmap y scatterplot3d, con los que, a partir de la información cuantitativa de la cantidad de exonizaciones de TEs y las metilaciones de Horvath, se generaron gráficos mostrando la relación entre ambos términos y las edades cronológicas.

Finalmente, para la visualización de algunos resultados también se usó IGV o Integrative Genomics Viewer (Robinson et al., 2011), que permite la visualización de múltiples archivos en formato bed al mismo tiempo.

Los cálculos fueron realizados mediante el uso de computadora personal con sistema operativo Linux Ubuntu 22.04, 8 GB de ram, tarjeta gráfica NVIDIA GTX 1660Ti y espacio de almacenamiento de 2TB; y el cluster Exxact, compuesto por 128 procesadores modelo Intel (R) Xeon (R) CPU E78867 v3 de velocidad 2,5GHz, 256GB de memoria RAM, 48TB de capacidad de almacenamiento (40Tb disponibles para usuarios), y sistema operativo Linux.

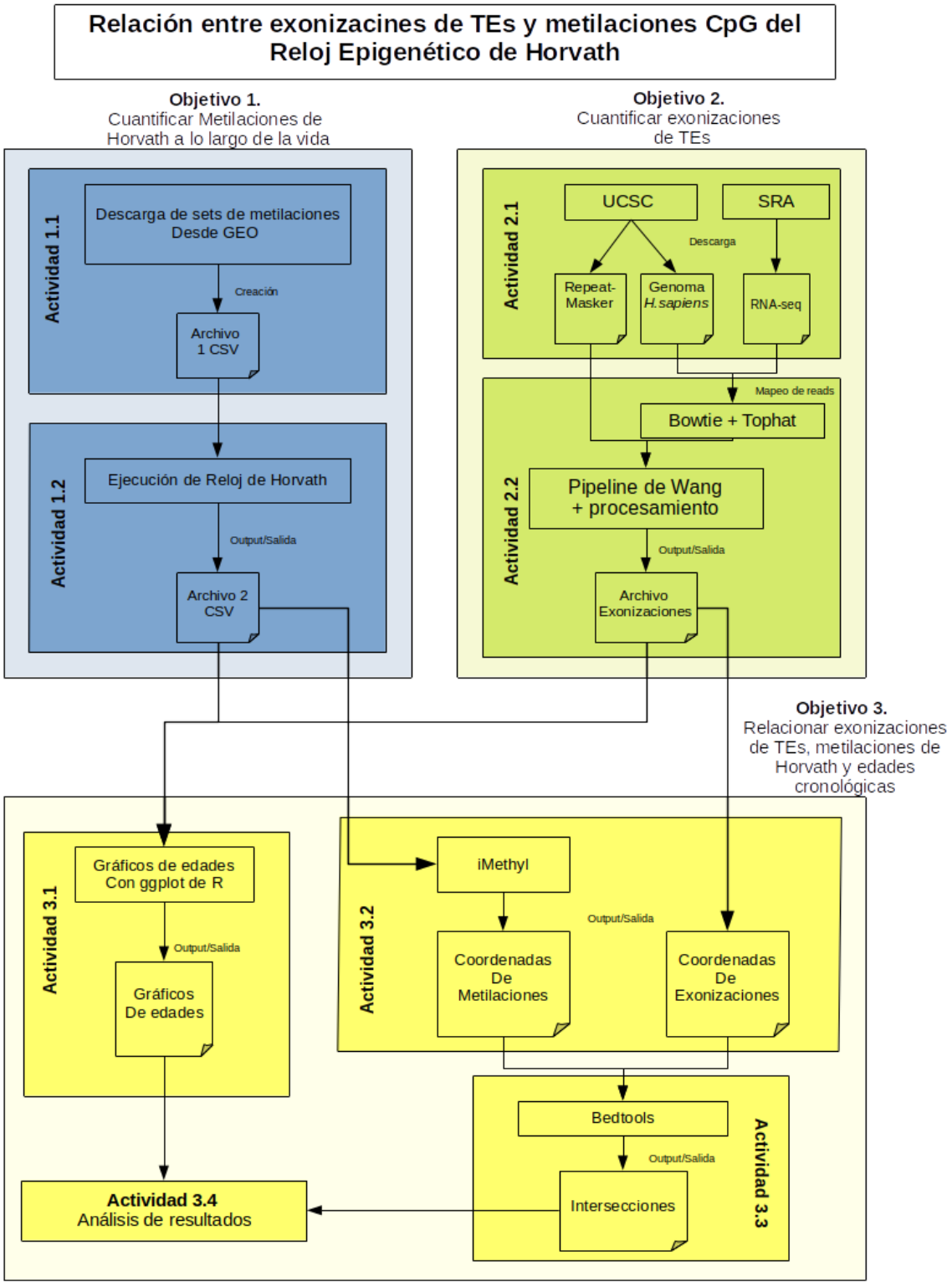


Fig. 7 Diagrama de Métodos, separado por objetivos y tareas.

Diagrama de Métodos, separado por objetivos y tareas. Las tareas son resumidas por su índice y sus actividades principales.

MÉTODOS

Los métodos son resumidos en la Fig. 1, donde se pueden observar las principales tareas de cada uno de los objetivos. La realización de los objetivos 1 y 2 puede ser en conjunto, pero ambos objetivos deben estar realizados antes de comenzar el objetivo 3.

Objetivo 1: Cuantificar Metilaciones de Horvath a lo largo de la vida.

1.1 Búsqueda y extracción de metilaciones del reloj epigenético de Horvath en tejido rombencefálico Humano.

El reloj epigenético de Horvath es un predictor que permite estimar la edad biológica de muestras a partir de la metilación del ADN (ADNm) de la mayoría de tejidos y células humanas.

Se descargaron las 17 muestras de metilaciones de tejido cerebeloso (uno de los tres tejidos de rombencéfalo) humano mencionadas por Horvath en su artículo (Horvath, 2013), desde la base de datos GEO de NCBI (Clough & Barrett, 2016), con edades de entre infantes de 1 año y adultos ancianos de 60 años. A partir de esta información se generó un archivo en formato CSV (comma separated values, .csv), que posteriormente fue utilizado por la plataforma del reloj Multitejido de Horvath[6] para predecir las edades de las muestras. Dicho archivo .csv (Fig. 7, “Archivo 1 CSV”), contiene columnas, donde la primera columna corresponde al identificador de illumina (IllmnID) de las islas CpG, las siguientes columnas corresponderán a las muestras y a los valores de metilación correspondientes a cada isla CpG. Es importante mencionar que estas 17 muestras contienen no solo las 353 islas CpG usadas por Horvath, sino más de 27.000 islas CpG presentes en el genoma humano.

Finalmente, mediante la búsqueda de los archivos suplementarios mencionados en la investigación de Horvath (Horvath, 2013), se identificaron las 353 marcas epigenéticas de metilación que Horvath utilizó en el modelo del reloj epigenético para predecir la edad de diferentes tejidos.

6 *Plataforma de Predictor de edad, reloj epigenético de Horvath (ADN Methylation Age Calculator https://ADNmage.genetics.ucla.edu/)*

1.2 Procesamiento de análisis cualitativo preliminar de información obtenida con el reloj epigenético de Horvath.

El archivo 1 (Fig. 1, “Archivo 1 CSV”), fue ingresado en la plataforma de predicción del reloj epigenético de Horvath. Y como resultado, la página entregó un segundo archivo .csv (Fig. 1, “Archivo 2 CSV”) con las predicciones de edad, IDs y otros datos de interés, como son los valores mínimos, máximos y medios de metilación en valor Beta de cada una de las muestras ingresadas.

Se compararon de manera manual los resultados obtenidos con el reloj epigenético de Horvath y los datos de metilación ingresados para su ejecución, se esperaba encontrar diferencias muy bajas con un error de 3.6 años aproximadamente entre las edades cronológicas (edades reales) y la predicción de las edades biológicas entregadas por el reloj para cada una de las muestras, ya que el reloj multitejido el cual se está utilizando en esta ocasión tiene un MAE de 3.6 años.

Descripción del protocolo de actividades Objetivo 1

Desde GEO (Clough & Barrett, 2016) se descargaron los archivos de metilación en muestras de tejido cerebeloso de *Homo sapiens*, con el código de acceso GSE38608, en donde se encuentran las 17 muestras necesarias para la ejecución del reloj epigenético de Horvath. La información descargada corresponde a 17 archivos .csv separados en dos columnas, en la primera columna se encuentra el ID de la isla CpG y en la segunda el grado de metilación en formato de valor M, números decimales entre positivos y negativos sin unidad, que indican el grado de metilación de cada una de las islas CpG, calculados mediante scanner de fluorescencia de la intensidad de la sonda en longitud de onda (nm).

Un valor M, es la relación $\log_2$ (bites) de las intensidades de la sonda metilada frente a la sonda no metilada. Sus valores son positivos y negativos. Si el valor M es cercano a 0 indica una intensidad similar entre las sondas metiladas y no metiladas, lo que significa que el sitio CpG está metilado en algún grado. Los valores M positivos significan que hay más moléculas metiladas que no metiladas, mientras que los valores M negativos significan lo contrario (Du et al., 2010).

Para poder ejecutar el reloj epigenético de Horvath, es necesario transformar el valor M a valor Beta.

El valor Beta es la relación entre la intensidad de la sonda metilada y la intensidad general (suma de las intensidades de la sonda metilada y no metilada). El valor Beta da como resultado un número de entre 0 y 1, que en condiciones ideales un valor de cero indica que todas las copias del sitio CpG en la muestra están completamente sin metilar, y un valor uno indica que cada copia del sitio se encuentra metilado (Du et al., 2010). Sin embargo, para Horvath, una isla CpG se considera completamente metilada si su valor Beta es igual o superior a 0.9, y completamente desmetilada si su valor Beta es igual o inferior a 0.1(Horvath, 2013).

La transformación de valor M a Beta se lleva a cabo mediante la ecuación señalada en la Fig. 8, donde se puede observar la equivalencia entre ambos valores, de cada isla CpG, señalada con el subíndice i.

$$Beta_i = \frac{2^{M_i}}{2^{M_i} + 1}$$

Fig. 8 Equivalencia entre valor Beta y valor M. Subíndice i representa a la isla CpG corresponidiente. Ecuación extraída de investigación de Du y Zhang (Du et al., 2010)

Luego de transformar las muestras a valor Beta, se pueden unir de forma manual (con ayuda de una hoja de cálculo) o mediante comandos de la terminal Linux, como join o paste. La manera en que se asegure la correcta unión entre los datos es mediante el comando join, pero este tiene una limitación, que puede unir solo 2 archivos a la vez, y al ser 17 archivos, se debe realizar muchas veces el proceso. Por lo tanto, se analizaron los archivos, al ser todos extraídos de la misma investigación, contienen las mismas islas CpG como ID, por lo tanto, se usó el comando paste.

```
paste muestra* > 17_con_IDrep
```

Como el comando paste solo añade las columnas hacia el lado, quedaron las columnas de “ID” repetidas en las columnas 16 veces (descontando la primera columna). Para reparar esto, se utilizó el comando awk.

```
awk 'BEGIN{OFS=FS=","}{print $1, $2, $4, $6, $8, $10, $12, $14,
$16, $18, $20, $22, $24, $26, $28, $30, $32, $34}' 17 _con_IDrep >
input_Horvath
```

Como resultado de la edición queda el archivo .csv necesario para la ejecución del Reloj epigenético de Horvath, con el ID de cada isla en la primera columna, y cada una de las muestras con el valor Beta de metilación en las columnas siguientes como se señala en la Tabla 1.

ID isla CpG	$Muestra_1$	→	$Muestra_{17}$
CpG_1	β (1,1)	---	β (1,17)
↓	↓	---	↓
CpG_m	β (m,1)	---	β (m,17)

Tabla 1 representación visual del archivo de entrada para la plataforma del Reloj Epigenético de Horvath

Los resultados entregados fueron comparados mediante gráficos con respecto a la edad real de las muestras con ayuda del paquete ggplot2 de R, lo que permitió identificar diferencias entre la información real y aquella predicha por la plataforma del reloj epigenético de Horvath , la Tabla Anexa 1 contiene esta información de los resultados.

Objetivo 2: Cuantificar exonizaciones de Tes.

2.1 Descarga de genoma *Homo sapiens,* RepeatMasker y set de RNA-seq de tejidos de rombencéfalo

Desde UCSC Genome Browser(Navarro Gonzalez et al., 2021) fue descargado el genoma humano (versión GRCh37/hg19), el archivo de anotaciones que describen

los genes (KnowGenes) y el archivo de coordenadas de RepeatMasker, en donde se encuentran las anotaciones de TEs, entre otros elementos repetidos.

Los archivos de anotaciones fueron editados para resumir mejor su información, generando archivos en formato .bed, de genes, exones y TEs (Ver Código Anexo 1).

El set de RNA-seq de tejido rombencefálico, mencionados por Cardoso-Moreira (Cardoso-Moreira et al., 2019), fueron descargados a partir de la plataforma SRA con ayuda de la herramienta SRA-tooolkit. Los identificadores SRA se obtuvieron a partir de la información depositada en la base de datos ArrayExpress[7] de EMBL-EBI (EMBL-EBI, 2021), con el código de acceso "E-MTAB-6814". Dichos identificadores se encuentran especificados en la Tabla Anexa 2.

2.2 Obtención de archivo de exonizaciones relacionadas a TEs

El genoma descargado es utilizado de referencia para realizar el mapeo de los reads del archivo .fastq con el programa Bowtie (Langmead et al., 2009), se generan índices que fueron utilizados por el programa TopHat (Kim & Salzberg, 2011), con su función tophat-fusion que identifica los puntos de fusión (fragmentos de reads que mapean en dos lugares distintos del genoma) presentes en el genoma entregado.

Se utilizó como referencia el script especificado en la investigación de Wang para crear un script editado (Código Anexo 2) que extrae las coordenadas de todas las fusiones presentes, sin pasar por una base de datos, lo que permitió acelerar el proceso de ejecución para la obtención de Exonizaciones de TEs a solo un par de minutos por archivo.

A partir del script editado, se obtiene un archivo con todas las coordenadas de las fusiones, que posteriormente pasaron por un segundo script (Código Anexo 3) que procesa los datos y realizó 5 filtros para poder para poder identificar correctamente las coordenadas de exonizaciones de TEs.

Descripción de protocolo de actividades Objetivo 2

Dentro de la investigación de Cardoso-Moreira (Cardoso-Moreira et al., 2019) de datos de RNA-seq, se encuentran mencionados en la información suplementaria, con el código de acceso E-MTAB-6814 (EMBL-EBI, 2021) para todos los tejidos humanos,

7 Plataforma ArrayExpress-Functional genomic data (https://www.ebi.ac.uk/arrayexpress/)

desde ahí, se extrajeron los sets únicamente de rombencéfalo (hindbrain), esto se llevó a cabo mediante el comando awk seleccionando solo las columnas 24, 5 y 7, cuando la columna 11 corresponda a rombencéfalo *(awk 'BEGIN{OFS=FS="\t"}($11=="hindrbrain"){print $24, $5, $7}'*), como resultado se obtuvo un archivo con el identificador SRA, edad y sexo de cada una de las muestras.

A partir de los identificadores SRA, se descargaron los archivos de reads en formato FASTQ con el comando fastq-dump de SRA-toolkit. La información de relevancia de los archivos se encuentra resumida en la Tabla 2.

Especie	Tejido	Observación	Tipo de secuenciación	Tipo de read	Largo de read
H. sapiens	Rombencéfalo	Sano	RNA-seq	Single End	101

Tabla 2 información básica de archivos SRA descargados.

Dichos archivos fueron preprocesados FastP (Chen et al., 2018), luego analizados con FastQC[5] y finalmente visualizados con MultiQC, para asegurar una calidad buena antes de ejecutar TopHat (Kim & Salzberg, 2011) con la función Tophat-fusion que, con ayuda de Bowtie, se mapean los archivos FASTQ en busca de fusiones presentes en cada una de las muestras.

El software FastP se ejecutó con los siguientes parámetos:

```
fastp -i archivoX.fastq.gz -o fastp_archivoX.fastq.gz --
json=fastp_achivoX.json --html=fastp_ERR2598080.html -f 6 -q 27 --
cut_front_window_size 5 --cut_tail_window_size 5 -5 -l 50 -U --
umi_loc=read1 --umi_len=6 -w 10
```

Donde:

- -f: se realiza un corte frontal de 6 nucleótidos (nt)
- -q: se espera que la calidad promedio de cada lectura (read) sea de 27
- --cut_front_window_size: corte de ventana frontal 5nt

- --cut_atil_window_size: corte de ventana final 5nt
- -l: largo mínimo de 50 de cada lectura
- -U –umi_loc: búsqueda de repeticiones en la lectura 1 (Single End)
- --umi_len=6: largo de la repetición en cada lectura
- -w: 10 uso de 10 hebras (uso de recursos computacionales)

Luego del preprocesamiento con FastP, los reads disminyeron de largo y cantidad, pasaron de tener reads de 101nt a tener largos de entre 50 y 80nt, y en promedio se perdieron 5.2% de los reads luego de la ejecución de FastP. Se utilizó FastQC para observar las calidades de los archivos previos y posteriores al uso de FastP, como se muestra en Fig. 9, donde se encuentran dos gráficos de “Per base sequence qualily” (calidad de secuencia por base).

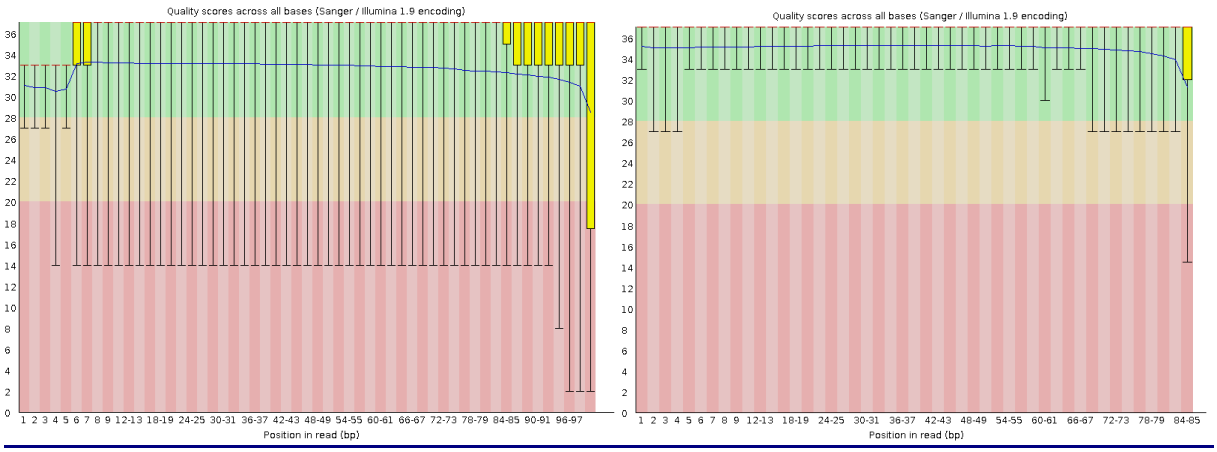


Fig. 9 Mejora de calidad post preprocesamiento de archivo fastq con FastP.

La primera imagen muestra las calidades iniciales de un archivo fastq aleatorio del set de datos, la segunda imagen corresponde al mismo archivo fastq después de preprocesar con FastP. Si bien las calidades no son perfectas en el archivo posterior al uso de FastP, son mucho mejores a lo que eran inicialmente.

Con MultiQC, se observan todos los archivos fastq pre y post FastP, de una manera más visual gráfica como se ve en la Fig. 10. En donde, ña calidad de los datos es representada con el color verde cuando son buenas, con amarillo cuando su calidad es intermedia, mientras que cuando una muestra tiene una calidad baja es representada por el color rojo.

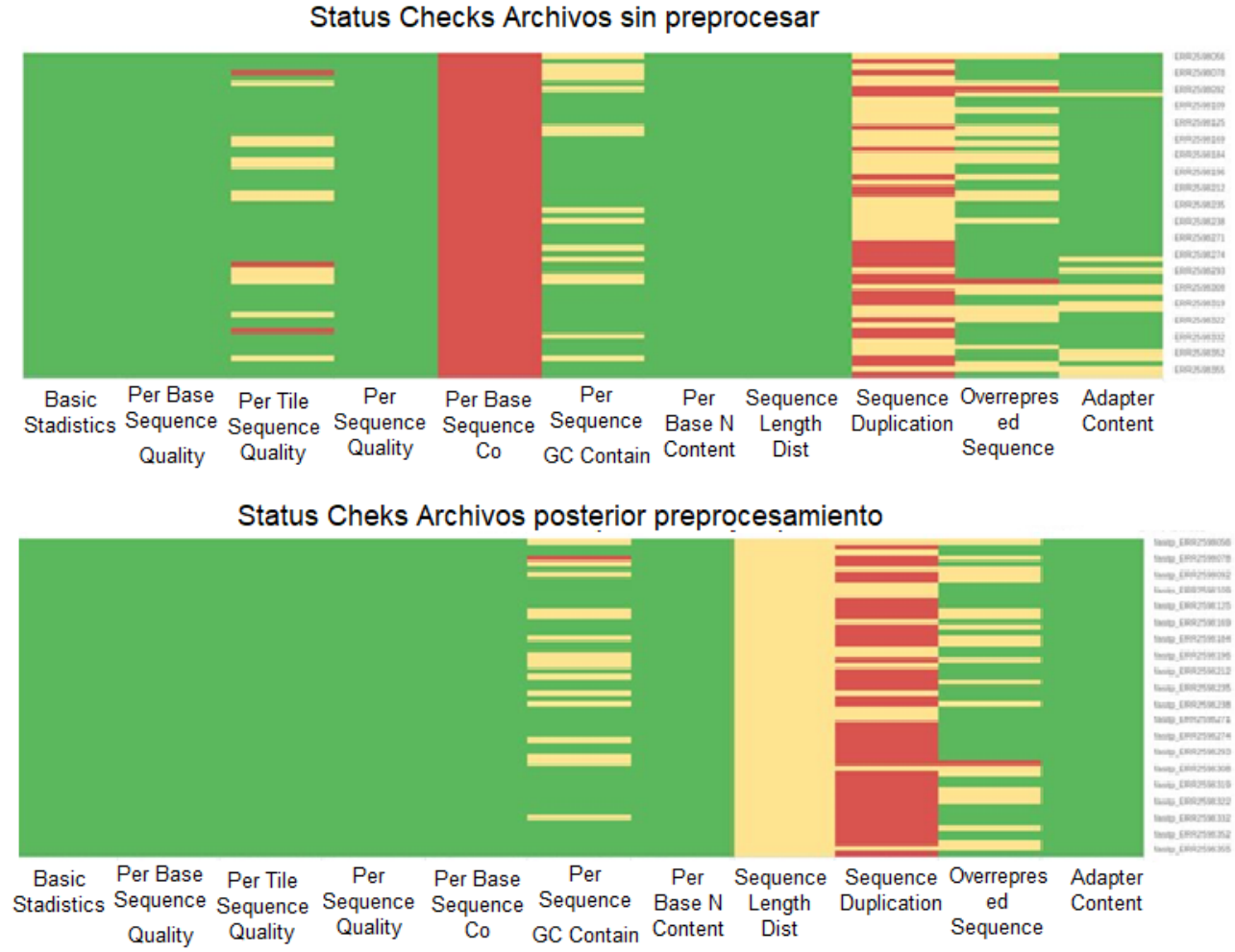


Fig. 10 Resumen Status Checks de FastQC (comprobaciones de estado) de los archivos antes (arriba) y después (abajo) de ser preprocesados con FastP.

Cada fila en las imágenes corresponden a una muestra, mientras que cada columna corresponde a un dato de interés.

La ejecución de tophat-fusions requiere la previa creación de índices mediante Bowtie, el cual requiere del genoma en formato fasta (fna). La ejecución del siguiente comando es a través de “la Terminal” de Linux.

```
bowtie-build genoma.fna índice_genoma
```

Posterior a la creación de los índices, se puede ejecutar tophat fusions, mediante el siguiente comando:

```
tophat --bowtie1 -g 1 -N 1 -o SALIDA -p 10 --max-intron-length 10 -
 -max-segment-intron 10 --fusion-search --fusion-min-dist 10 --fu-
   sion-read-mismatches 1 --fusion-multipairs 1 índice_genoma ar-
                          chivo.fastq.gz
```

Donde:

- - - bowtie1 : se decreta el uso del software Bowtie para el mapeo de los reads.
- -g 1: Si hay mas de un alineamiento con el mismo puntaje, se establece uno de forma aleatoria.
- N 1: cantidad de incopatibilidad mínima permitida
- -o SALIDA: se genera un archivo contenedor de los resultados con el nombre “SALIDA”.
- -p 10: Cantidad de procesadores que son utilizados por el software.
- - - max-intron-lenght 10: largo máximo de intrones permitidos.
- - - max-segment_intron 10: largo máximo de intrón que se puede encontrar durante la búsqueda del segmento dividido.
- - -fusion-search: habilita la función de mapeo de fusiones.
- - -fusion-min-dist 10: búsqueda de fusiones separadas por una distancia de 10 nt.
- - -fusion-read-mismatches 1: Cantidad de reads que apoyan las fusiones.
- - -fusion-multipair 1: se consideran los reads que mapean una única vez
- Índice_genoma: indexación generada con bowtie-build
- Archivo.fastq.gz: corresponde al archivo fastq a analizar.

Inicialmente se había optado por usar el script de la investigación de Wang (Wang et al., 2016) para la búsqueda de exonizaciones de TEs, sin embargo, por motivos de tiempo de ejecución y de gastos computacionales, se decidió utilizar solo una parte del script, y crear un script editado, que entrega resultados menos procesados pero en un

menor tiempo de ejecución (Código Anexo 2), lo que permite ser procesados posteriormente de manera manual (Código Anexo 3).

Para poder corroborar que los resultados del script de Wang y los del script editados, se procesaron ambos de manera paralela con la misma muestra elegida al azar. El script de Wang requiere el uso de una base de datos MySQL, que contenía las anotaciones de exones y las de elementos transponibles extraídas de los archivos TE.bed y exón.bed (ver Código Anexo 1).

Al corroborar que técnicamente se obtenían los mismos resultados, se decidió usar el script editado para la obtención de exonizaciones de TEs.

Con el uso del script editado a partir del de Wang (Wang et al., 2016), se obtuvieron las coordenadas de todas las fusiones, o candidatos a exonizaciones de TEs, presentes en cada una de las muestras (ver

La figura señala que la salida de Tophat-fusions se separa en dos partes en formato bed mediante el uso de comando “awk”, estas partes, izquierda y derecha (left.bed y right. Bed, respectivamente), con ayuda de Bedtools intersect, son procesados identificando si las coordenadas encontradas en la salida de tophat-fusions corresponden a exones (Exones.bed) o a elementos transponibles (TE.bed). Con ayuda de los comandos “join” y “awk”, pasan por los 5 filtros mencionados anteriormente.

donde se especifica el funcionamiento del script editado, y los Código Anexo 2 y Código Anexo 3). Estos resultados fueron procesados a través de 5 filtros: (1) la fusión haya aparecido en el análisis más de una vez (jCount>1); (2) que corresponda a Splicing Canónico; (3) el fragmento de read sea ≥ 25nt;(4) que ambos lados de la fusión correspondan al mismo cromosoma; y (5) que la fusión sea entre un exón y un TE (TE-exón o exón-TE).

Estos 5 filtros permitieron reducir las fusiones encontradas con tophat-fusions para identificar las posibles exonizaciones de TEs presentes en cada una de las muestras.

Objetivo 3: Relacionar exonizaciones de TEs, metilaciones de Horvath y edades cronológicas.

3.1 Relacionar por separado el número de exonizaciones de TEs y metilaciones CpG de Horvath con la edad de las muestras en cada set de datos.

Esto se llevó a cabo mediante ambos sets de datos por separado, los de metilaciones (archivos de entrada y de salida del Reloj Epigenético de Horvath) y el archivo resultante de las exonizaciones de TEs obtenido en la actividad 2.2, con la intención de encontrar relaciones entre las muestras por separado con respecto a la edad real (cronológica) de cada muestra. Con ayuda de paquetes R (ggplot2, tidyr y pheatmap), se generan gráficos que permiten un análisis visual del comportamiento de las exonizaciones de TEs versus la edad cronológica de las muestras y las metilaciones relacionadas a las islas CpG de Horvath a lo largo de la vida, sin tomar en cuenta las coordenadas de éstas en el genoma.

3.2 Extracción de coordenadas de metilaciones y exonizaciones de TEs

Cada una de las metilaciones de Horvath está enlazada a un IllmnID (identificador de illumina), por lo que será necesario encontrar la ubicación específica de cada isla CpG en el genoma humano. Esto fue llevado a cabo mediante el uso de la base de datos iMthyl[8], que permite la descarga de todas las islas CpG presentes en el genoma Humano, cada una de ellas se encuentra identificada mediante IllmnID, por lo que es posible seleccionar solo las islas que son utilizadas por Horvath para su reloj epigenético.

El uso de iMehyl entrega solo la coordenada de inicio y la cadena de cada isla CpG, por lo que con ayuda de comandos de Linux, se generan archivos en formato bed, con las primeras 6 columnas, donde la primera corresponde al cromosoma, la segunda y tercera son las coordenadas de incio y final, la cuarta al IllmnID, la quinta al puntaje (representado por un punto “.”) y la última corresponde a la cadena (+ o -).

8 iMethyl, base de datos de metilación del ADN integrativo. Contiene la ubicación de todas las islas CpG presentes en el genoma humano, versión GRCh37/hg19. (http://imethyl.iwate-megabank.org/)

Las coordenadas de las exonizaciones de TEs, fueron extraídas a partir del archivo de salida obtenido en la actividad del objetivo 2.2, y mediante comandos Linux, se generó un archivo en formato bed similar al de las islas CpG, solo se diferenciará por la cuarta columna que corresponde al TE específico de cada exonización, mencionando su clase, familia y super familia a la vez, separado por símbolo numeral (#), tal como se señala el archivo de elementos transponibles (TE.bed) de Código Anexo 1.

3.3 Intersección de coordenadas de islas CpG usadas por de Horvath y de exonizaciones de TEs

Teniendo las coordenadas de las islas CpG y el de exonizaciones de TEs, se pudo observar si existe una relación entre ambas coordenadas. Este paso se llevó a cabo mediante la función intersect de bedtools (Aaron, 2021), que identificó la presencia de islas CpG en genes que poseen exonizaciones de TEs

Se generó un archivo de 12 columnas, en donde se observarán en las columnas 4 y 8 el TEs y la isla CpG específica de la intersección.

3.4 Visualización, análisis e interpretación de gráficos y archivos resultantes de la relación entre metilaciones de Horvath y exonizaciones de TEs

Se observaron los gráficos de edades generados (3.1) y los resultados de la intersección (3.3), con el objetivo de determinar la correlación entre las exonizaciones de TEs y las metilaciones en las islas CpG del reloj epigenético de Horvath.

Los gráficos generados (3.1), esperaban mostrar la capacidad de las metilaciones y de las exonizaciones de TEs para predecir la edad biológica de una muestra. Se esperaba encontrar una correlación significativa entre el número de exonizaciones de TEs relacionados a islas CpG con la edad cronológica de las muestras.

Con la información generada en la actividad 3.3, se generan gráficos de: (1) proporción de número de Exonizaciones de TEs v/s la edad cronológica; (2) proporción de número de Exonizaciones de TEs con islas CpG v/s la edad cronológica; (3) proporción de número de Exonizaciones de TEs v/s proporción de número de Exonizaciones de TEs con islas CpG; (4) un gráfico tridimensional de proporción de número de Exonizaciones de TEs con islas CpG v/s proporción de número de Exonizaciones de TEs con islas CpG v/s edad cronológica; y (5) un análisis de genes relacionados a exonizaciones de

TEs con islas CpG, en busca de rutas metabólicas y/o enfermedades que puedan estar relacionadas al envejecimiento.

Descripción de protocolo de actividades Objetivo 3

Los datos generados en las actividades del objetivo 1 fueron evaluadas comparando las edades reales versus las edades predichas por el reloj epigenético de Horvath, se tomó en cuenta la información presente en la Tabla Anexa 1 para generar un heatmap, con la herramienta pheatmap de R, en donde las columnas correspondían a las predicciones del reloj de Horvath de las 17 muestras de metilaciones de ADN, mientras que las filas correspondían a las islas CpG (Fig. Anexa 2).

Para el desarrollo de la actividad 3.2, fue necesario identificar claramente las islas 353 CpG utilizadas por Horvath para la creación de su reloj, mediante su archivo suplementaria 3 (Horvath, 2013), se obtiene el ID de cada una de las islas y se guarda en un archivo en forma de lista con el encabezado "ID" (lista obtenida en actividad 1.1), para luego, con ayuda de la base de datos iMethyl[8], identificar sus coordenadas iniciales junto a su cadena.

La lista de islas CpG entragadas por iMethyl se encuentra en formato de tabla, separada por tabulaciones, con los encabezados ID, Inicio y Cadena:

ID	Inicio	Cadena
cgXXXXXXXX	100000	+

Con ayuda del comando "join" de la terminal Linux, se une el archivo con la lista de los ID de las islas CpG de Horvath con la información entregada por la base de datos iMethyl.

El comando utilizado fue:

```
join -1 1 -2 1 -t '\t' lista_CpG_Horvath lista CpG_iMethyl >
CpG_Horvath
```

Donde:

- -1 1 hace referencia a la unión de la primera columna del primer archivo
- -2 1 hace referencia a la unión de la primera columna del segundo archivo

- -t ‘\t’, hace referencia al tipo de separador de salida, en este caso corresponde a una tabulación.
- lista_CpG_Horvath: corresponde al primer archivo (353 CpG específicas).
- lista_CpG_iMethyl: corresponde al segundo archivo.
- > CpG_Horvath: archivo de salida, que contiene las 353 islas CpG específicas de Horvath, con las coordenadas de inicio y cadenas

Las coordenadas finales de cada isla CpG fueron calculadas de manera pareja, añadiendo 25.000 pb (25 Kb) a las coordenadas iniciales, esto debido a que la metilación de una isla CpG afecta al gen o conjunto de genes que comparten la misma función (Li & Zhang, 2014). Y debido a que el tamaño un gen humano es de entre 20 y 30 Kb, es correcto usar 25 Kb para identificar la presencia de una isla CpG en un gen.

Se creó un archivo formato bed con la información de las islas CpG mediante el comando awk de la terminal Linux:

```
awk ‘BEGIN{OFS=FS=”\t”}($2=!”Inicio”){print $1,$2,$2+25000,”.”,$3}’
CpG_Horvath > CpG
```

El comando anterior añade 25000 a la coordenada de inicio, siempre que ésta coordenada corresponda a un valor numérico, es decir, sin tomar en cuenta el encabezado “Inicio”.

Los resultados obtenidos en las actividades del objetivo 2 fueron normalizados y evaluados generando datos cuantitativos en un archivo de columnas, con los IDs de las muestras, edades cronológicas, la cantidad de exonizaciones de TEs por muestra, proporción de éstas, tipos de TE, proporción de tipos de TE, entre otros. Estos datos son capaces de ser utilizados en R, con los paquetes ggplot2 ytidyr. Fue posible generar un gráfico del número de exonizaciones de TEs (normalizados mediante la ecuación de la Fig. 11) versus la edad cronológica de las muestras, mediante un gráfico de puntos, coloreado por tipo de muestra (pre o posnatal), añadiendo también líneas de pendientes. También, con el uso de ggplot2 y tidyr en conjunto se pudieron generar gráficas de los porcentajes de las distintas familias de TEs presentes en cada una de

muestras pre y posnatales (Fig. Anexa 3 y Fig. Anexa 4). Con la intención de encontrar quizás una variación entre las Familias de TEs durante el crecimiento.

$$Normalización = \frac{num.de\ Exonizaciones\ de\ TEs\ por\ muestra}{num.de\ candidatos\ iniciales\ por\ muestra}$$

Fig. 11 Cálculo de normalización de las exonizaciones de TEs

Posteriormente, se utilizó bedtools intersect para encontrar las intersecciones entre las coordenadas de islas CpG y las exonizaciones de TEs, para añadir la información de proporción del número de exonizaciones de TEs relacionadas a islas CpG de cada una de las muestras al archivo creado anteriormente (Ver Tabla Anexa 3), y con dicha información, se pudo generar otro gráfico entre las proporción de exonizaciones de TEs relacionadas a islas CpG versus la edad cronológica de las muestras, además de un gráfico comparativo entre ambos parámetros.

El cuarto gráfico corresponde a un gráfico tridimensional, generado con el paquete scatterplot3d de R, utilizando la proporción de exonizaciones de TE v/s exonizaciones de TE con islas CpG v/s la edad cronológica, de esta manera se busca encontrar una correlación entre las 3 variables, mediante un plano de regresión, calculado con R.

Con la información resultante se calcula el coeficiente de determinación múltiple R^2 y el coeficiente correlación múltiple mediante la ecuación de la siguiente ecuación.

$$\boldsymbol{R_{123}} = \sqrt{\frac{\boldsymbol{r_{12}^2 + r_{13}^2 - 2r_{12} * r_{13} * r_{23}}}{\boldsymbol{1 - r_{23}^2}}}$$

Fig. 12 Ecuación de Correlación entre variables calculadas en gráfico 3d

Todos los códigos usados en R para la creación de gráficos en la actividad 3.4 se encuentran especificados en Código Anexo 4.

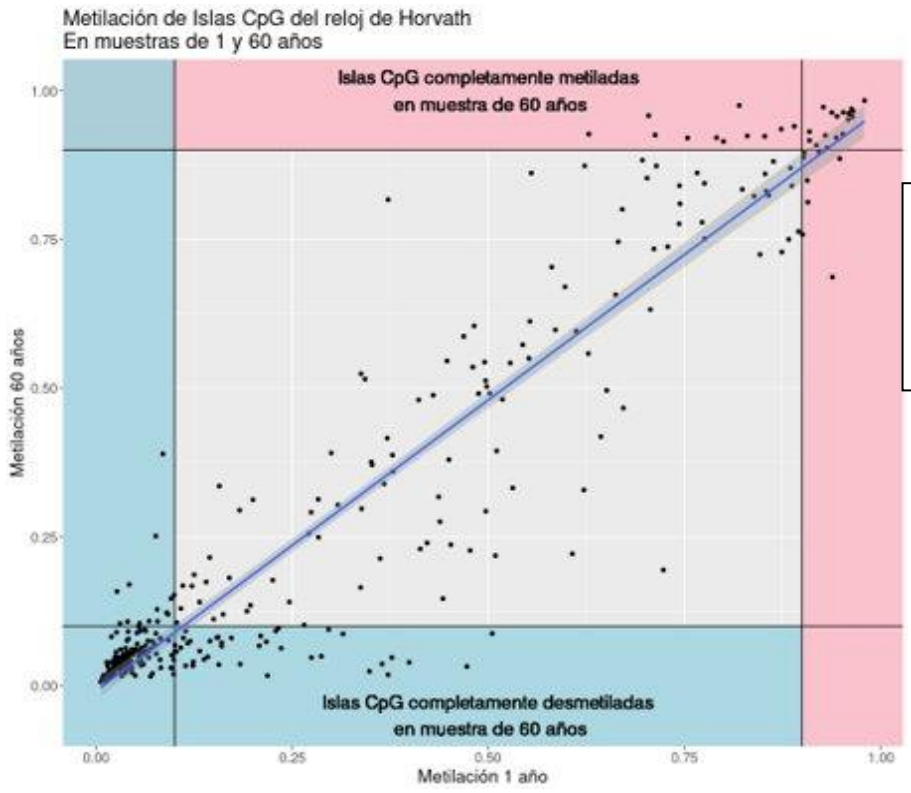
Adicionalmente, con ayuda de bedtools, se identificaron los genes que contienen exonizaciones de TEs relacionados a islas CpG, uno a uno, mediante la plataforma Keeg[9], con el fin de encontrar rutas metabólicas y enfermedades relacionados a dichos genes.

[9] KEGG PATHWAY Database https://www.genome.jp/kegg/pathway.html

RESULTADOS

Resultados Objetivo 1

Durante el proceso de análisis de las 17 muestras de metilaciones en la plataforma del reloj epigenético de Horvath, se compararon los datos sin predecir de una de estas muestras más jóvenes (1 año) y una de las más longevas (60 años), con la intención de observar una estimación del comportamiento de las islas CpG a lo largo de la vida del ser humano. La Fig. 13 muestra gráfico de puntos, donde Se puede observar que claramente ambas variables aumentan de manera casi pareja, ya que cuenta con una pendiente muy cercana a 1 (0.977) y una correlación entre los datos de 0.944. El cuadrante superior muestra que existen islas en la muestra de 1 año que aumentan su metilación en la muestra de 60 años, mientras que el cuadrante inferior señala que existen islas que se desmetilan a lo largo de la edad.



correlación R	0,944
R^2	0,891
pendiente	0,977

Fig. 13 Comparación entre muestras de Metilación de ADN más joven y más longevas

La figura muestra a cada una de las islas CpG del reloj epigenético de Horvath como un punto, los ejes x e y corresponden a muestras de 1 y 60 años respectivamente y los colores del gráfico en celeste y rosado, representan si se considera que una isla se encuentra desmetilada o metilada.

Esta comparación (entre la muestra de 1 año y la de 60 años), permite identificar un grupo de islas CpG que poseen una mayor varianza entre ambas muestras, lo que se ve reflejado en la Fig. 14, para poder obtener este conjunto de islas CpG, se calculó de promedio del valor beta de cada una de las 353 islas de Horvath en ambas edades, aquellas que superaran el promedio fueron seleccionadas, en este caso, se trata de 50. En la figura se puede observar *un leve patrón de comportamiento entre las dos muestras, si una isla a la edad de 1 año es mayor a 0.7, es más probable que aumente su metilación a que disminuya.*

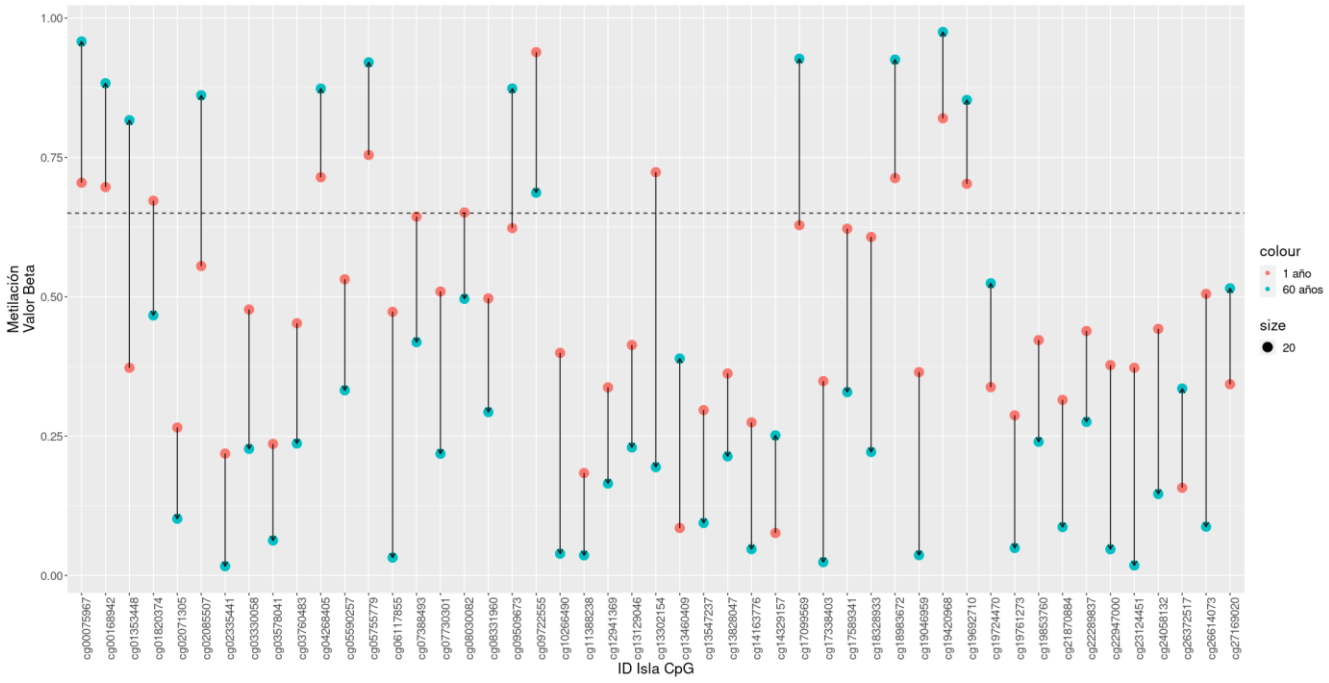


Fig. 14 Muestra de 50 islas CpG con mayor varianza en edades de entre 1 y 60 años

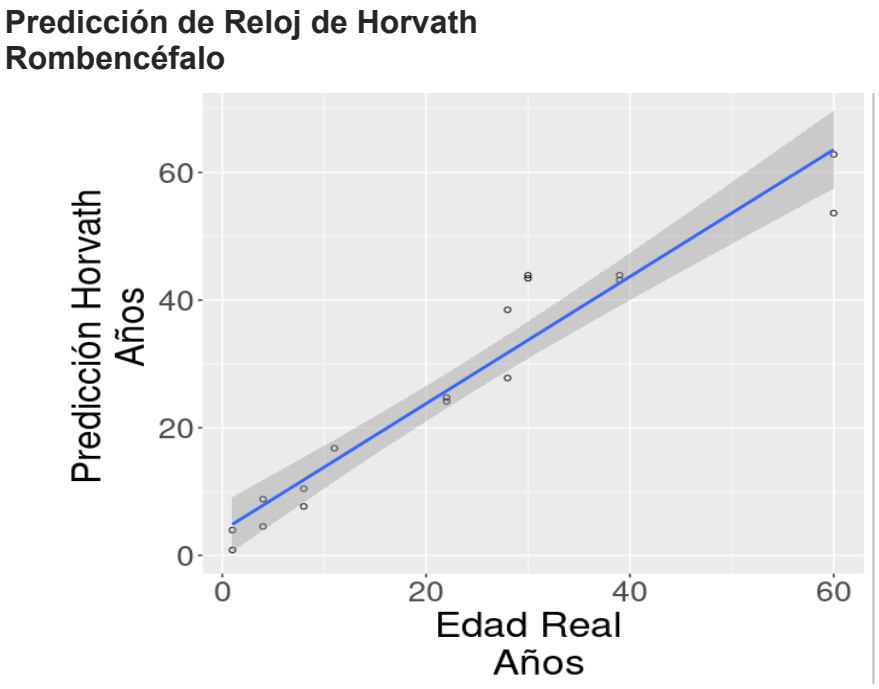
Se observa en el eje x el conjunto de 50 islas CpG con mayor varianza entre las muestras de 1 y 60 años, en el eje y el valor beta de cada una de las islas, disminuye o aumenta con la edad, los puntos rojos representan el valor beta de una isla específica a la edad de 1 año, mientras que los puntos azules representan los de 60 años. Fue añadida una flecha por cada una de las islas CpG que indican si la metilación aumenta o disminuye en cada isla. Además se inseró una línea punteada en y=0.7.

Los resultados del reloj epigenético de Horvath (ver Tabla Anexa 1 y Tabla 3), se obtiene una predicción de la edad de las muestras, y estos resultados fueron graficados para identificar la eficiencia del reloj frente a las muestras entregadas. En la Fig. 15, se observa un gráfico de las edades reales o cronológicas de cada muestra versus la predicción de Horvath. El error esperado de los resultados del reloj epigenético multitejido según Horvath era de 3.6 años, y de 4.2 años para muestras específicas de rombencéfalo, y el error (desviación estándar) de las muestras entregadas fue de 4.309, lo que permite inferir que el reloj epigenético entregó resultados esperados. En muchas ocasiones, la edad predicha es similar a la real, aunque existen excepciones, diferencias de hasta 13 años (Tabla 3).

Predicción del Reloj Multitejido	Edad real de las muestras	Predicción del Reloj Multitejido	Edad real de las muestras
0,84	1	**27,79**	28
3,96	1	**38,47**	28
4,54	4	**43,15**	39
7,67	8	**43,40**	30
8,82	4	**43,89**	30
10,46	8	**43,90**	39
16,79	11	**53,63**	60
24,13	22	**62,82**	60
24,73	22		

Tabla 3 Resultados resumidos de Reloj Epigenético de Horvath

La tabla se encuentra ordenada a partir de la predicción del reloj epigenético de Horvath, se puede observar que, la diferencia entre la predicción y la edad real de las muestras.



sd	4,309
menor	0,150
mayor	13,89
pendiente	0,995
R	0.965
R²	0.931

Fig. 15 Predicción de Reloj epigenético de Horvath

La figura muestra la comparación de la edad real o cronológica de las muestras versus su respectiva predicción de edad según Horvath. Los resultados son bastante confiables, con una desviación estandar (sd) de 4.309, una pendiente de 0.995 y una correlación entre los datos de 0.965.

Resultados Objetivo 2

Luego de la descarga de los sets de datos necesarios para obtención de exonizaciones de TEs, fue necesario realizar un preprocesamiento de las muestras con FastP (Chen et al., 2018). Inicialmente existían 59 muestras de rombencéfalo humano, de las cuales 2 fueron eliminadas debido a que su calidad y cantidad de reads era menor al resto, por lo que el análisis solo se llevó a cabo sobre las 57 muestras en formato fastq presentes en la Tabla Anexa 2.

Al ejecutar tophat-fusions, los “candidatos” a exonizaciones de TEs eran en promedio más de 140.000, y al momento de ejecutar los 5 fiiltros, que aumentan la probabilidad de que una fusión corresponda a una exonización de TEs, se obtiene un número mucho menor en de resultados (en promedio poco más de 480 exonizaciones de TEs). La Fig. 16 muestra esta disminución de manera gráfica.

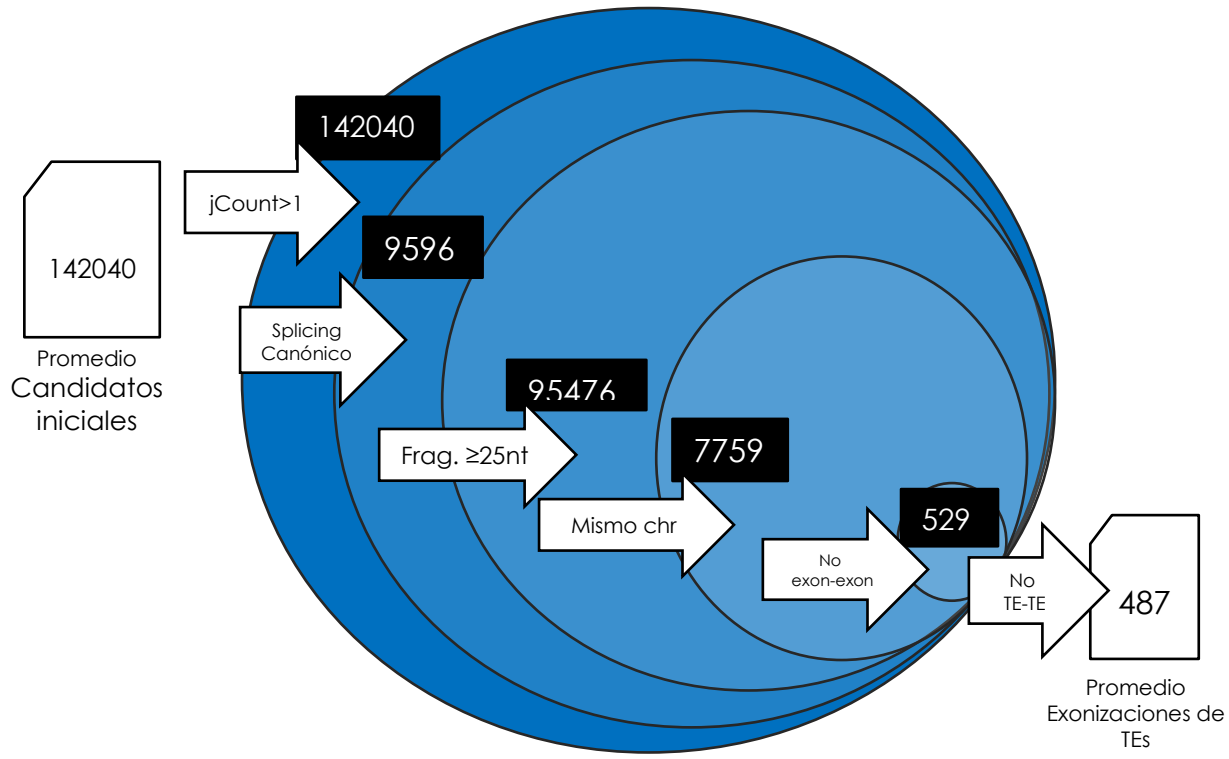


Fig. 16 Ejecución de 5 filtros para encontrar exonizaciones de TEs

La figura muestra de manera gráfica con un diagrama de Venn la disminución de resultados desde los 142040 candidatos a iniciales, hasta las 487 exonizaciones de TEs en promedio. Los símbolos de archivos representan los candidatos iniciales y el archivo de exonizaciones de TEs, cada flecha representa uno de los filtros utilizados para la obtención de exonizaciones de TEs, exceptuando las últimas dos flechas (No exón-exón y No TE-TE), que corresponden al último filtro (debe tener una interacción TE-exón o exón-TE).

Los archivos resultantes de la ejecución de los 5 filtros corresponden a archivos de 10 columnas como se puede observar en la Tabla 4, que se encuentra separada en dos partes, izquierda (L) y derecha (R). Como es mencionado en la Fig. 5, es necesario identificar la parte derecha e izquierda de una exonización para identificar cual de estos lados corresponde a un exón y cuál a un TE. Se puede observar que en cada una de las líneas se encuentra un exón y un elemento transponible independientemente del lado, ya que no importa el lado en donde se encuentre el exón o el TE, lo importante es que exista una interacción entre ambos elementos.

ID_exo	CHR_L	inicioL	FinL	NombreL	CadenaL	CHR_R	inicioR	FinR	NombreR	CadenaR
108	chr1	1164263	1164326	exon-NM_016176.6-2	-	chr1	1166887	1166921	MIRb#SINE#MIR	-
366	chr1	1608233	1608284	exon-NM_001110781.3-2	-	chr1	1675689	1675731	MLT1F2#LTR#ERVL-MaLR	-
368	chr1	1622414	1622454	MLT1F2#LTR#ERVL-MaLR	-	chr1	1671099	1671142	exon-NM_001199787.2-3	-
397	chr1	1671080	1671142	exon-NM_001199787.2-3	-	chr1	1675689	1675727	MLT1F2#LTR#ERVL-MaLR	-
588	chr1	3506445	3506489	MER103C#DNA#hAT-Charlie	-	chr1	3511900	3511940	exon-NM_001409.4-3	-

Tabla 4 archivo de salida de exonizaciones de TEs.

Presenta un extracto de un archivo de salida de una muestra de rombencéfalo humano, el archivo cuenta con 10 columnas, donde la primera columna corresponde al ID de la fusión, un identificador único que posee cada exonización; las columnas 2 y 7 corresponden a los cromosomas de cada lado de la exonización de TEs; 3 y 8 son las coordenadas iniciales; 4 y 9 son las coordenadas finales; y, 5 y 10 son las cadenas (+ o -).

Resultados Objetivo 3

Con la predicción entregada por Horvath en el objetivo 1, fue posible generar heatmaps, en busca de patrones claros de metilación o desmetilación a lo largo de la edad, es decir, se esperaba ver una degradación de izquierda a derecha o de derecha a izquierda en el Heatmap, a medida que la muestra fuera aumentando en edad, para ello, se utilizaron inicialmente las 353 islas CpG, sin embargo, no se observó un patrón claro (ver Fig. Anexa 2). Debido a esto, se decidió recurrir a la clusterización de las islas, desde k=150, 100 y 50 pero, no se observaron patrones claros.

Finalmente, se recurrió a la información suplementaria 2 del artículo de Horvath (Horvath, 2013), que menciona la existencia de 7 islas CpG específicas con las que se puede realizar una predicción bastante acertada de la edad real de las muestras (con error de hasta 8 años), pero esta predicción corresponde a una investigación previa a la de Horvath (Hannum et al., 2013).

Según Hannum, con solo estas 7 islas CpG (cg04474832, cg05442902, cg06493994, cg09809672, cg19722847, cg21296230, cg22736354) es posible predecir la edad de una muestra, pero según Horvath, es necesario añadir “un par de cientos” más, para

que la predicción tenga menos errores y pueda ser válida para todos los tejidos humanos.

Dentro del reloj epigenético de Horvath solo se encuentran 6 de estas 7 islas CpG, y fueron utilizadas para la creación de un último heatmap (Fig. 17). En esta ocasión, se observan claros patrones de desmetilación con la edad en dos islas cg05442902 y cg19722847, en donde sus colores pasan de blanco a celeste/azul mientras aumenta la edad de las muestras. Ocurre lo contrario en las islas cg22736354 y cg06493994, que pasan de un azul más intenso a uno menos intenso, por lo que se puede decir que se observa una metilación de esas islas con el paso del tiempo. Las dos islas sobrantes, no presentan un claro patrón de metilación.

Dos de las 6 islas mencionadas en la figura se encuentran relacionadas a genes, el resto se encuentra en regiones no codificantes del genoma. La isla cg04474832 (que no posee un patrón claro de metilación) se encuentra relacionada a los genes ABHD14A. ABHD14A-ACY1 y ACY1, los cuales codifican a proteínas con un posible papel en el desarrollo de neuronas granurales (neuronas pequeñas encontradas en diversas estructuras del encéfalo), además de catalizar la hidrólisis de aminoácidos N-acetilados a acetato y aminoácidos libres. Mientras que la isla cg05442902 se encuentra relacionada al gen SLC7A4, que codifica a la proteína con su mismo nombre que interviene en el transporte de los aminoácidos catiónicos (arginina, lisina y ornitina).

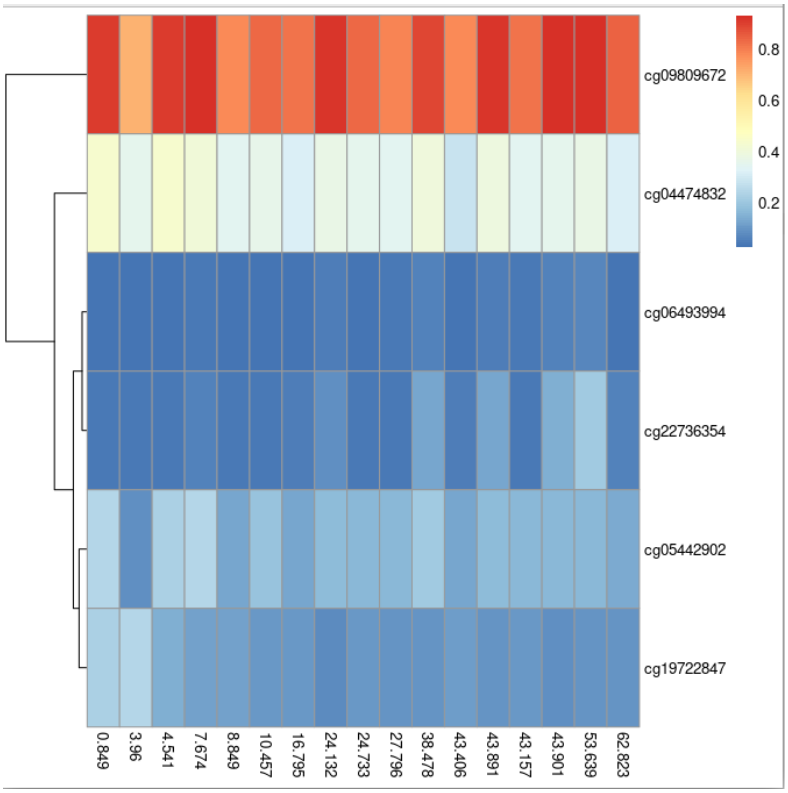


Fig. 17 Heatmap Comparación de las 6 de las 7 islas CpG específicas de Hannum. Estas islas se encuentran ubicadas en las filas, mientras que cada una de las muestras se encuentra en las columnas, ordenadas según la predicción del reloj epigenético de Horvath, de menor a mayor desde los 0.849 hasta los 62.823 años.

Mediante la obtención de las coordenadas de exonizaciones y de las islas CpG se generaron gráficos comparativos de estos parámetros por separado, los gráficos de proporciones de exonizaciones de TEs v/s edad de la edad de las muestras (Fig. 18), proporción de Exonizaciones de TEs relacionados a islas CpG de Horvath v/s la edad de las muestras (Fig. 19) y el gráfico que busca encontrar una correlación entre ambos, Proporción de Número de Exonizaciones de TEs v/s Exonizaciones de TEs con islas CpG (Fig. 20).

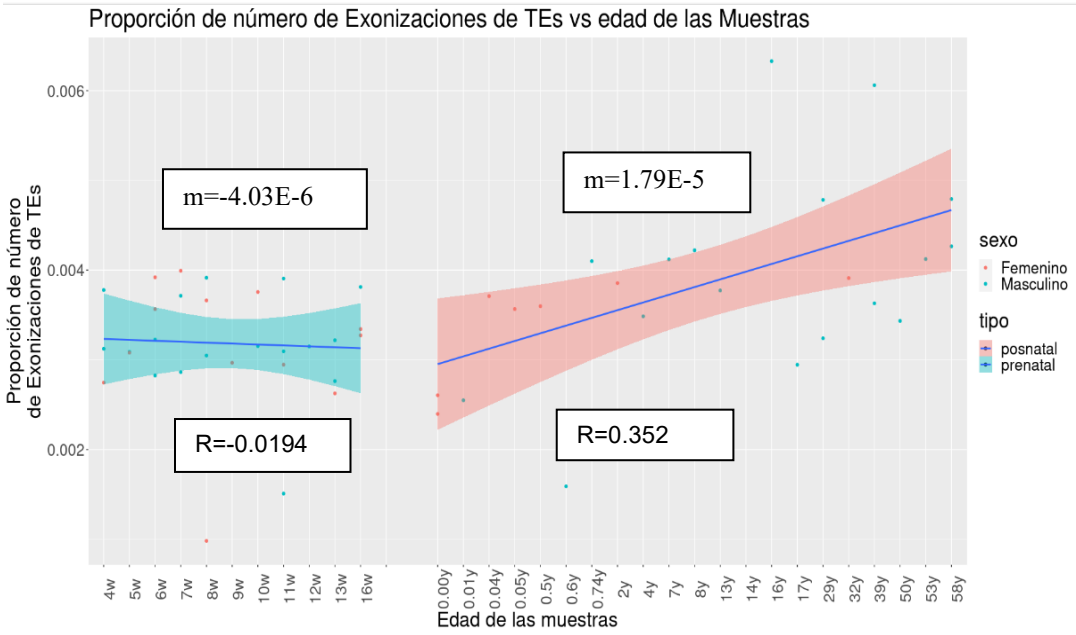


Fig. 18 Exonizaciones de TEs v/s edad cronológica de las muestras.

La figura especifica cada una de las muestras mediante puntos de distintos colores, rosa femenino y azul masculino, además presenta dos líneas de pendientes de la recta, una para muestras prenatales (sombreado celeste) y las posnatales (sombreado rosa). El eje x representa la edad de las muestras, desde las 4 semanas de gestación (w), hasta los 58 años (y), Mientras que el eje y representa la proporción de exonizaciones de TEs presentes en cada una de las muestras.

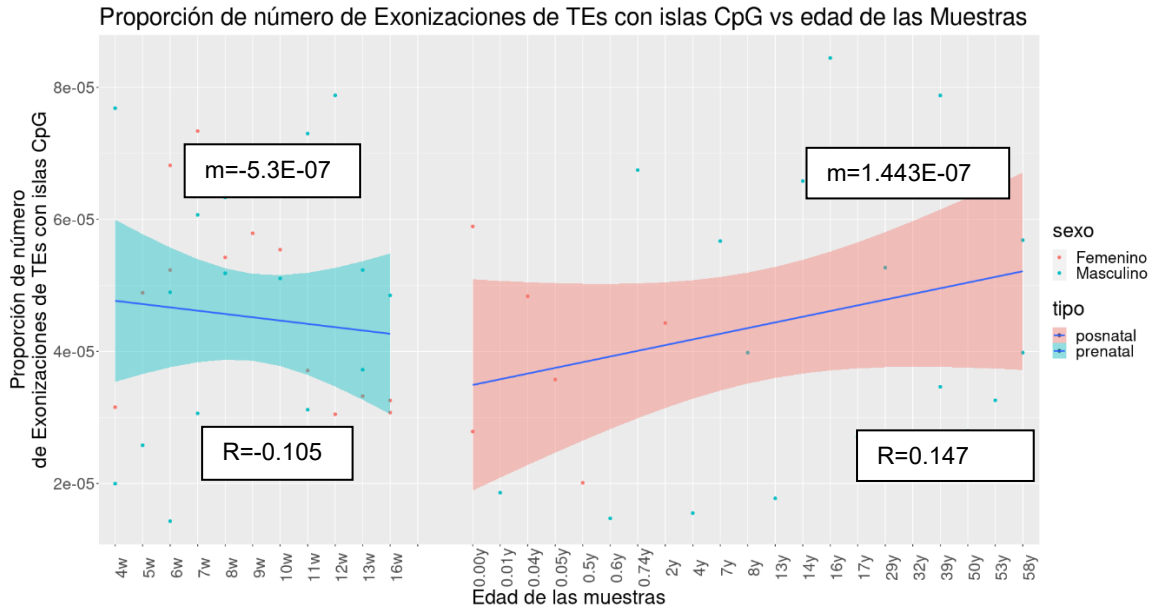


Fig. 19 Relación entre número de exonizaciones relacionadas a islas CpG v/s edad cronológica de las muestras

La figura muestra la proporción de exonizaciones de TEs con islas CpG a lo largo de la edad. Al igual que en la Fig. 18, el eje x representa la edad de las muestras, desde las 4 semanas de gestación (w), hasta los 58 años (y), Mientras que el eje y representa la proporción de exonizaciones de TEs presentes en cada una de las muestras.

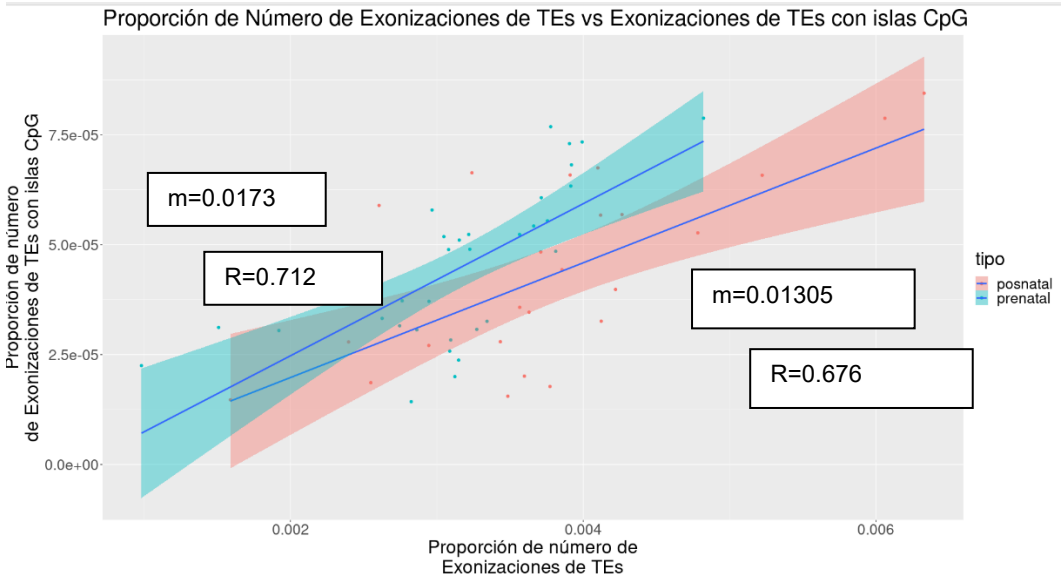


Fig. 20 Proporción de Número de Exonizaciones de TE v/s Exonizaciones de TEs con islas CpG

En el eje x se encuentra la proporción de número de exonizaciones de TEs , mientras que en el eje y se encuentra la proporción de número de exonizaciones de TEs con islas CpG. Las pendientes que poseen las muestras son bajas pero positivas, a diferencia de Fig. 18 y Fig. 19, donde las pendientes son prácticamente 0. Sin embargo, las correlaciones varían con respecto a las figuras anteriores.

Para poder identificar correctamente si existe una correlación entre el número de exonizaciones de TEs y el número de exonizaciones de TEs con islas CpG de Horvath, es necesario generar un último gráfico de 3 ejes (tridimensional) en el que se añade el parámetro de edad. La Fig. 21 muestra un plano de regresión lineal, que entregó los siguientes parámetros:

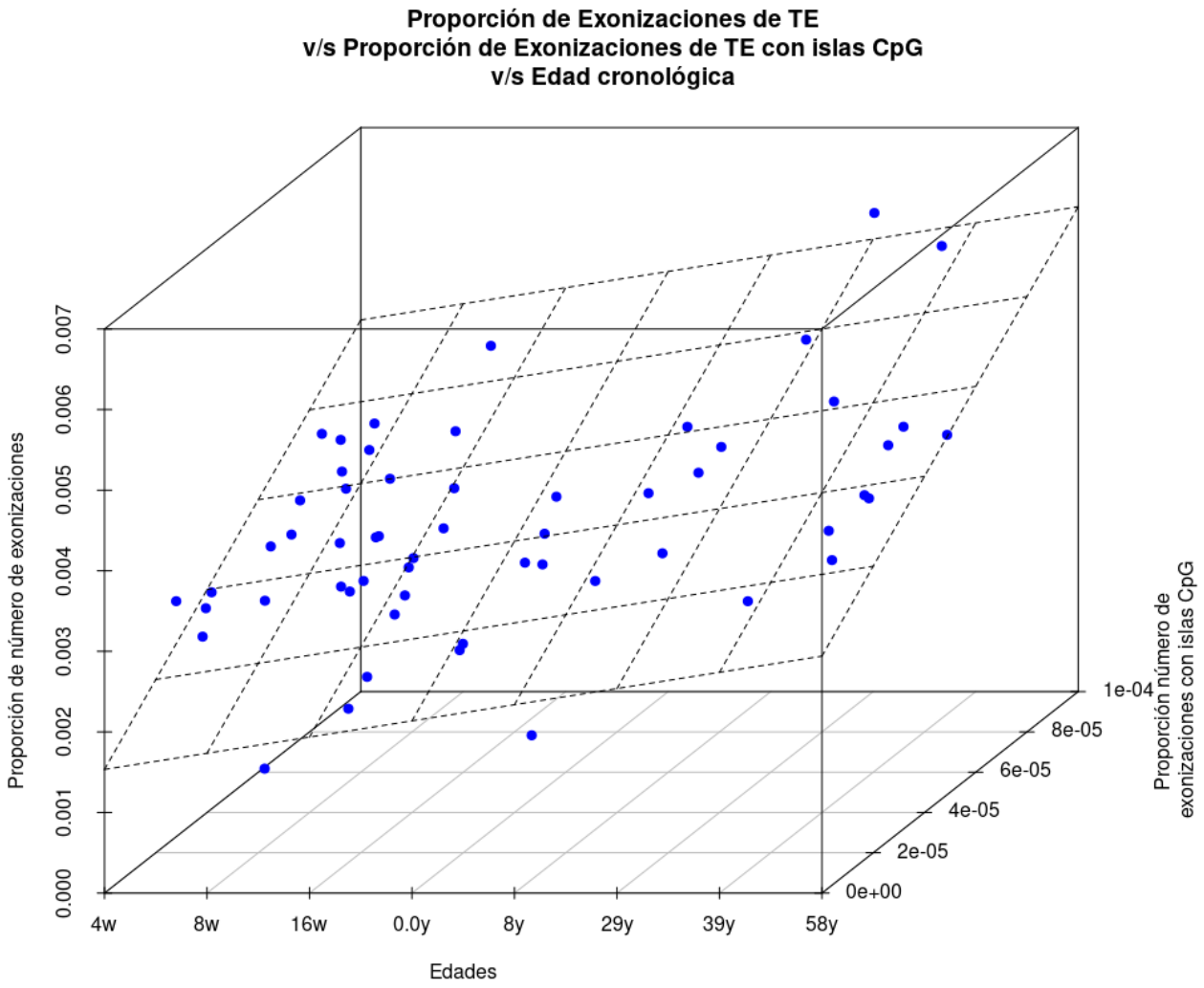


Fig. 21 Gráfico comparativo entre 3 variables

Se observa en el eje x las edades de las muestras, eje y la proporción de número exonizaciones de TEs con islas CpG y en el eje z se encuentra la proporción de número de TE. Además, se añadió un plano de regresión para calcular la correlación entre los datos.

La correlación a partir de los datos del gráfico de la Fig. 21, y el cálculo señalado en la Fig. 12 es de 0.71 (Ver Cálculo Anexo 1). Lo que indica una correlación alta entre los tres parámetros a la vez, utilizando los coeficientes de correlación mencionados en las figuras Fig. 18, Fig. 19 y Fig. 20.

Finalmente, se analizaron los genes relacionados a exonizaciones de TEs con islas CpG mediante el uso de bedtools instersect (Ver Código Anexo 3). Se encontraron 30 genes (ver Fig. Anexa 5), los cuales fueron analizados por separado, a través de la plataforma Keeg[9] en busca de rutas metabólicas y enfermedades en común entre ellos. De los 30 genes encontrados inicialmente solo 10 de ellos interactuaban dentro de las mismas rutas o están relacionados a enfermedades, dicha información es resumida en la Fig. 22. Los genes relacionados a exonizaciones de TEs con islas CpG se encuentra detallada en la Tabla Anexa 4.

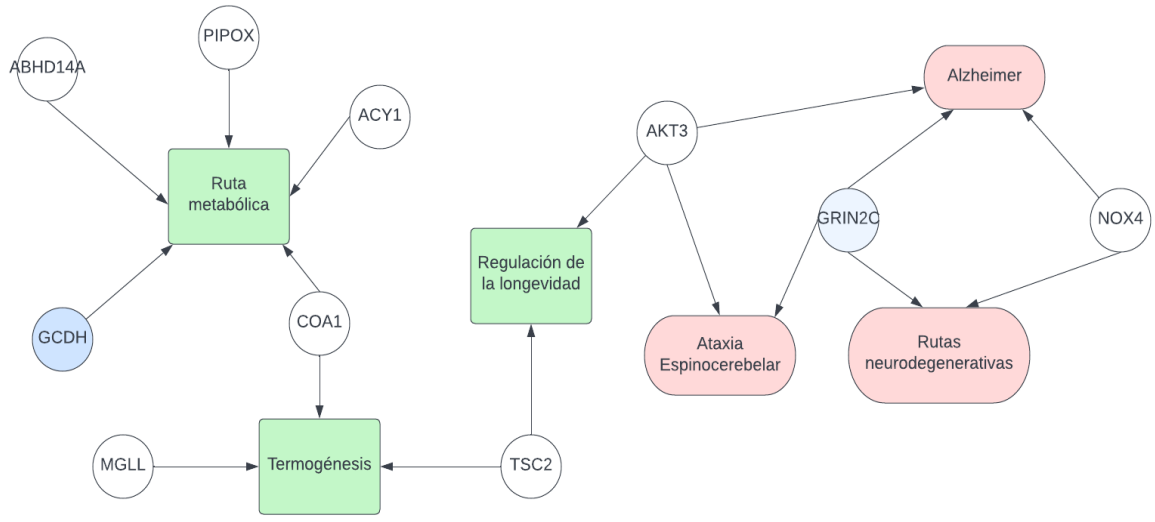


Fig. 22 Rutas metabólicas y enfermedades relacionadas a genes con exonizaciones de TEs e islas CpG

La figura muestra los genes (en círculos) que participan en rutaes metabólicas (cuadros verdes), o se relacionan con alguna enfermedad (cuadros rojos). Existen genes directamente relacionados a la regulación de la longevidad (AKT3 y TSC2) y otros directamente relacionados con enfermedades (AKT3, GRIN2C, NOX4). Se marcaron con azul y celeste los genes GCDH y GRIN2C respectivamente, que corresponden a genes que se encuentran solamente en una etapa (pre y posnatal).

DISCUSIONES

Discusiones Objetivo 1

El resultado obtenido en la Fig. 13, en donde se observa una comparación entre dos muestras una joven y una longeva, hace presumir que los grados de metilación de las islas CpG de Horvath son directamente proporcionales, ya que son pocas las ocasiones en las que existe un cambio en el grado de metilación con la edad. Lo que desafía lo mencionado por Andersen y Tsurumi (Andersen et al., 2017; Tsurumi & Li, 2012), que mencionan que durante el envejecimiento existe una disminución de ADNm a nivel global. Pero la Fig. 14, apoya prácticamente por completo lo mencionado por los autores, ya que en las islas CpG que tienen mayor varianza si se observa que la gran mayoría de las islas sufren desmetilación más que metilación durante el envejecimiento.

El reloj epigenético multitejido de Horvath ha sido probado y validado por decenas de investigadores en distintas ocasiones, debido a ello, es de esperar que sus predicciones fueran acertadas. Sin embargo, existían predicciones muy alejadas de la realidad, lo que hace pensar que el estilo de vida de cada uno de los sujetos de estudio debe inducir a que sus tejidos posean una edad biológica diferente a la real. Según investigaciones posteriores de Horvath y Lu (Lu et al., 2019), la edad de un tejido se ve afectada por la calidad de vida del sujeto (consumo de alcohol y tabaco, alimentación y factores ambientales). Produciendo que personas que llevan una mejor calidad de vida tengan una predicción de edad más cercana a su edad real (cronológica) que personas con una calidad de vida deficiente.

Discusiones Objetivo 2

Es necesario el preprocesamiento de los datos mediante una herramienta capaz de mejorar su calidad, generar recortes de los reads y reducir o aumentar el porcentaje GC a un valor adecuado al genoma humano (entre 46 y 48%), ya que los archivos fastq que contienen un porcentaje GC diferente a lo esperado pueden estar en presencia de alguna clase de contaminación por parásitos u otros organismos, o puede

existir algún error de secuenciación que influiría en la confiabilidad de los resultados que se puedan obtener a partir de dicho archivo fastq.

El tiempo de ejecución dentro del objetivo 2 fue un factor de gran importancia, ya que, el preprocesamiento de todos los archivos fastq (57 muestras) con FastP demoraron en total casi 7 horas en ejecutar, es decir, un promedio de 10 minutos por archivo. Mientras que la ejecución de Tophat-fusion demoró un estimado de 2.5 horas por archivo. Es este el motivo por el cual se optó por no utilizar el script de Wang para la búsqueda de exonizaciones de TEs, ya que el ingreso a la base de datos MySQL presente en el Script de Wang (Wang et al., 2016), produce un tiempo de ejecución de hasta 20 horas por archivo, lo que genera grandes gastos computacionales que mantendrían al servidor Exxact ocupado por cerca de 47 días, lo que es demasiado para un servidor compartido.

Además, el script de Wang entrega no entrega todos los resultados que son detectados mediante el uso del script editado (Código Anexo 2) y el script de procesamiento (Código Anexo 3), ya que, con una prueba generada con datos propios, se encuentran muchas más exonizaciones de TEs con los scripts generados para esta investigación que con el de Wang. Y aquellas exonizaciones de TEs encontradas por el script de Wang se encuentran dentro del conjunto de exonizaciones de TEs encontrados por los scripts utilizados para esta investigación, señalado en la Fig. 23. Esto quiere decir que existe la posibilidad de que el script de Wang no detecte todas las exonizaciones de TEs presentes en una muestra.

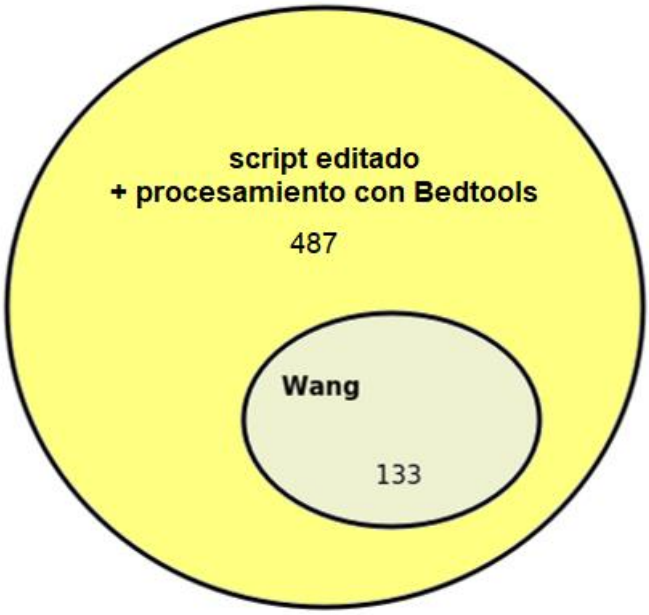


Fig. 23 Comparación de número de exonizaciones obtenidos con el Script de Wang y las obtenidas por el script editado y el de procesamiento con Bedtools

Discusiones Objetivo 3

El uso de los sets de datos por separado para la realización de las tareas de este objetivo tiene como finalidad encontrar un patrón visual que permita identificar una correlación entre los sets de metilaciones en islas CpG de Horvath y el archivo de exonizaciones de TEs versus las edades biológicas de las muestras. Es de esperar que mientras menos islas CpG se analicen a la vez, se pueda observar un patrón más claro de metilación entre las islas, sin embargo, el motivo del uso de 353 islas CpG y no menos islas como las mencionadas en la Fig. 17, es porque el uso de estas islas ayuda a proporcionar una mejor validación de la predicción en cada muestra. El hecho, de que no se puedan observar patrones claros con las 353 islas CpG infieren que el uso de 17 muestras es muy poca cantidad para poder encontrar un patrón claro en un heatmaps como el presentado en la Fig. Anexa 2. Por otro lado, el relacionar el número de exonizaciones de TEs versus la edad de las muestras no tiene mucho sentido, ya que cada una de las muestras correspondía a un archivo fastq con un tamaño (cantidad de reads) único, por lo que es posible que el número de exonizaciones de TEs se

encuentre relacionado al tamaño de la muestra inicial, o más bien, a los candidatos iniciales mencionados en Fig. 16, es decir, el número total de datos entregados por el script editado. Esto se justifica mediante el estudio de (Sorek et al., 2002), que menciona que aproximadamente el 4% de las proteínas humanas contienen secuencias de TE, pero si durante el proceso de mapeo de los reads (con Bowtie), no es posible tener una cobertura completa del genoma, es muy probable que no se logren mapear los elementos del genoma (genes, exones e intrones). Debido a esto, es normal pensar que, a mayor cantidad de candidatos iniciales, es mayor el número de exonizaciones de TEs encontradas, por lo que se recurrió a la normalización de los datos, que en este caso corresponde a una proporción entre el número de exonizaciones de TEs dividido por el número de candidatos iniciales como se observa en la Fig. 11. De esta manera, los gráficos que se generaron para el objetivo 3 tienen sentido, ya que no corresponde al número de exonizaciones de TEs, sino a la proporción de estos con respecto a la edad cronológica de cada una de las muestras.

Lo mismo ocurre al momento de querer analizar el número de exonizaciones de TEs relacionadas a islas CpG, es necesario normalizar los datos creando una proporción como en la Fig. 11, con la única diferencia de que es el número de exonizaciones relacionadas a islas CpG divididas por el número de candidatos iniciales. En ambos casos, se observa una correlación muy baja o nula (Ver Tabla Anexa 5) entre sus edades cronológicas, esto era esperable ya que las pendientes de cada gráfico (Fig. 18 y Fig. 19) son prácticamente 0.

La creación del gráfico de la Fig. 20 es el único de los gráficos bidimensionales que cuenta con unas pendientes menos cercanas a 0 y una correlación casi significativa (ver Tabla Anexa 5), demostrando que mientras más exonizaciones de TEs sean identificadas de un archivo, el número islas CpG relacionadas a exonizaciones de TEs aumentará también. Esto es esperable, ya que en el genoma humano los TEs se encuentran altamente metilados gracias a que en sus secuencias se encuentra un alto contenido de CG, siendo en ocasiones hasta un tercio de ellas(Li & Zhang, 2014; Sorek et al., 2002).

CONCLUSIONES

Inicialmente, se había pensado en rechazar la hipótesis: “Existe una relación entre el número de exonizaciones de TEs y las metilaciones de Horvath a lo largo del tiempo celular”, debido a dos grandes motivos que posteriormente fueron descartados:

El primero, se consideró que la cantidad de 17 muestras de metilaciones de islas CpG en rombencéfalo (cerebelo), eran muy pocas para poder encontrar un patrón visible de metilaciones en el heatmap de la Fig. Anexa 2, esto se justificaba debido a que para la creación del reloj epigenético de Horvath (Horvath, 2013), él utilizó 7844 muestras sanas de metilaciones de ADN de distintos tejidos, y entre más de 27000 islas CpG analizadas, y encontró las 353 con las que predice la edad biológica. Pero este motivo dejó de ser valido al momento encontrar las 7 islas de Hannum, (6 de ellas presentes en las 353 de Horvath). Como se puede observar en la Fig. 17, si existen patrones de metilación en 4 de las 6 islas, se puede pensar que al añadir más muestras si se podría encontrar un patrón más claro de metilación en dichas islas.

El segundo está relacionada a las pendientes de los gráficos del número de exonizaciones de TEs y la edad cronológica son prácticamente 0 (Fig. 18), al igual que las pendientes entre el número de exonizaciones de TEs con islas CpG con respecto a la edad cronológicas (Fig. 19), por lo que se pensaba que no existían correlaciones entre los datos. Sin embargo, en ambos casos si existen, son débiles o muy débiles en muestras pre y posnatales (muy cercanas a 0), pero, al comparar las proporciones de exonizaciones de TEs v/s exonizaciones de TEs con islas CpG (Fig. 20) se encuentran correlaciones moderadas o significativas (sobre 0.6), lo que es apoyado por el gráfico tridimensional de la Fig. 21, que tiene una correlación múltiple significativa de 0.7. Los tipos de correlación pueden observarse en la Tabla Anexa 5, donde se resumen los valores que puede tomar el coeficiente de correlación lineal.

La aparición de genes que se encuentran directamente relacionados a enfermedades neurodegenerativas y a la regulación de la longevidad (Fig. 22), hace suponer que, si existe una relación entre las exonizaciones de TEs y las islas CpG de Horvath con la

aparición de enfermedades relacionadas con el envejecimiento, lo que apoya directamente la hipótesis planteada en la investigación.

El hecho de encontrar pequeñas correlaciones entre los datos, aunque sean pequeñas, existen, por lo no es posible rechazar ni comprobar la hipótesis de manera tajante, sería necesario aumentar el número de muestras y comprobar los resultados con quizás más tejidos.

REFERENCIAS

Aaron, R. (2021). *Bedtools: A powerful toolset for genome arithmetic—Bedtools 2.30.0 documentation*. https://bedtools.readthedocs.io/en/latest/

Adav, S. S., & Wang, Y. (2021). Metabolomics Signatures of Aging: Recent Advances. *Aging and Disease*, *12*(2), 646–661. https://doi.org/10.14336/AD.2020.0909

Afanas'ev, I. (2014). New nucleophilic mechanisms of ros-dependent epigenetic modifications: Comparison of aging and cancer. *Aging and Disease*, *5*(1), 52–62. https://doi.org/10.14336/AD.2014.050052

Alexeeff, S. E., Baccarelli, A. A., Halonen, J., Coull, B. A., Wright, R. O., Tarantini, L., Bollati, V., Sparrow, D., Vokonas, P., & Schwartz, J. (2013). Association between blood pressure and DNA methylation of retrotransposons and pro-inflammatory genes. *International Journal of Epidemiology*, *42*(1), 270–280. https://doi.org/10.1093/ije/dys220

Alvarado G, A. M., & Salazar M, Á. M. (2014). Aging concept analysis. *Gerokomos*, *25*(2), 57–62. https://doi.org/10.4321/S1134-928X2014000200002

Andersen, P. R., Tirian, L., Vunjak, M., & Brennecke, J. (2017). A heterochromatin-dependent transcription machinery drives piRNA expression. *Nature*, *549*(7670), 54–59. https://doi.org/10.1038/nature23482

Andrenacci, D., Cavaliere, V., & Lattanzi, G. (2020). The role of transposable elements activity in aging and their possible involvement in laminopathic diseases. *Ageing Research Reviews*, *57*, 100995. https://doi.org/10.1016/j.arr.2019.100995

Bannister, A. J., & Kouzarides, T. (2011). Regulation of chromatin by histone modifications. *Cell Research*, *21*(3), 381–395. https://doi.org/10.1038/cr.2011.22

Benoit, C. (2014). Transposable Elements in Cancer and Other Human Diseases. *Current Cancer Drug Targets*, *15*(3), 227–242. https://doi.org/10.2174/1568009615666150317122506

Bernadotte, A., Mikhelson, V. M., & Spivak, I. M. (2016). Markers of cellular senescence. Telomere shortening as a marker of cellular senescence. *Aging*, *8*(1), 3–11. https://doi.org/10.18632/aging.100871

Bourque, G., Burns, K. H., Gehring, M., Gorbunova, V., Seluanov, A., Hammell, M., Imbeault, M., Izsvák, Z., Levin, H. L., Macfarlan, T. S., Mager, D. L., & Feschotte, C. (2018). Ten things you should know about transposable elements. *Genome Biology*, *19*(1), 199. https://doi.org/10.1186/s13059-018-1577-z

Brunet, A., & Berger, S. L. (2014). Epigenetics of Aging and Aging-related Disease. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, *69*(Suppl 1), S17–S20. https://doi.org/10.1093/gerona/glu042

Cardoso-Moreira, M., Halbert, J., Valloton, D., Velten, B., Chen, C., Shao, Y., Liechti, A., Ascenção, K., Rummel, C., Ovchinnikova, S., Mazin, P. V., Xenarios, I., Harshman, K., Mort, M., Cooper, D. N., Sandi, C., Soares, M. J., Ferreira, P. G., Afonso, S., … Kaessmann, H. (2019). Gene expression across mammalian organ development. *Nature*, *571*(7766), 505–509. https://doi.org/10.1038/s41586-019-1338-5

Cavagnari, B. (2012). Regulación de la expresión génica: Cómo operan los mecanismos epigenéticos. *Archivos Argentinos de Pediatria*, *110*(2), 132–136. https://doi.org/10.5546/aap.2012.132

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, *34*(17), i884–i890. https://doi.org/10.1093/bioinformatics/bty560

Clough, E., & Barrett, T. (2016). The Gene Expression Omnibus Database. En E. Mathé & S. Davis (Eds.), *Statistical Genomics: Methods and Protocols* (pp. 93–110). Springer. https://doi.org/10.1007/978-1-4939-3578-9_5

Crespo Santiago, D. (2011). *Capítulo 1. El envejecimiento: Definiciones y teorías.* Capítulo 1. https://ocw.unican.es/mod/page/view.php?id=700

De Cecco, M., Ito, T., Petrashen, A. P., Elias, A. E., Skvir, N. J., Criscione, S. W., Caligiana, A., Brocculi, G., Adney, E. M., Boeke, J. D., Le, O., Beauséjour, C., Ambati, J., Ambati, K., Simon, M., Seluanov, A., Gorbunova, V., Slagboom, P. E., Helfand, S. L., … Sedivy, J. M. (2019). L1 drives IFN in senescent cells and promotes age-associated inflammation. *Nature*, *566*(7742), 73–78. https://doi.org/10.1038/s41586-018-0784-9

Du, P., Zhang, X., Huang, C.-C., Jafari, N., Kibbe, W. A., Hou, L., & Lin, S. M. (2010). Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics*, *11*(1), 587. https://doi.org/10.1186/1471-2105-11-587

Ecker, S., & Beck, S. (2019). The epigenetic clock: A molecular crystal ball for human aging? *Aging (Albany NY)*, *11*(2), 833–835. https://doi.org/10.18632/aging.101712

EMBL-EBI. (2021). *E-MTAB-6814—Human RNA-seq time-series of the development of seven major organs*. https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-6814/samples/?s_sortby=col_45&s_sortorder=ascending&s_page=1&s_pagesize=100

Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, *32*(19), 3047–3048. https://doi.org/10.1093/bioinformatics/btw354

Fulop, T., Witkowski, J. M., Pawelec, G., Alan, C., & Larbi, A. (2014). On the immunological theory of aging. *Interdisciplinary Topics in Gerontology*, *39*, 163–176. https://doi.org/10.1159/000358904

Gabory, A., Attig, L., & Junien, C. (2011). Epigenetic mechanisms involved in developmental nutritional programming. *World Journal of Diabetes*, *2*(10), 164–175. https://doi.org/10.4239/wjd.v2.i10.164

García Robles, R., & Ayala Ramírez, P. (2012). *Epigenética: Definición, bases moleculares e implicaciones en la salud y en la evolución humana*. http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S1692-72732012000100006

Gellersen, H. M., Guell, X., & Sami, S. (2021). Differential vulnerability of the cerebellum in healthy ageing and Alzheimer's disease. *NeuroImage. Clinical*, *30*, 102605. https://doi.org/10.1016/j.nicl.2021.102605

Gellersen, H. M., Guo, C. C., O'Callaghan, C., Tan, R. H., Sami, S., & Hornberger, M. (2017). Cerebellar atrophy in neurodegeneration-a meta-analysis. *Journal of Neurology, Neurosurgery, and Psychiatry*, *88*(9), 780–788. https://doi.org/10.1136/jnnp-2017-315607

Goldsmith, T. C. (2012). On the programmed/non-programmed aging controversy. *Biochemistry. Biokhimiia*, *77*(7), 729–732. https://doi.org/10.1134/S000629791207005X

González-Carreró López, M. I. (2011). *Capítulo 7. Daño oxidativo y envejecimiento.* Capítulo 7. https://ocw.unican.es/mod/page/view.php?id=708

Guo, T., & Fang, Y. (2014). Functional organization and dynamics of the cell nucleus. *Frontiers in Plant Science*, *5*. https://doi.org/10.3389/fpls.2014.00378

Hannum, G., Guinney, J., Zhao, L., Zhang, L., Hughes, G., Sadda, S., Klotzle, B., Bibikova, M., Fan, J.-B., Gao, Y., Deconde, R., Chen, M., Rajapakse, I., Friend, S., Ideker, T., & Zhang, K. (2013). Genome-wide methylation profiles reveal quantitative views of human aging rates. *Molecular Cell*, *49*(2), 359–367. https://doi.org/10.1016/j.molcel.2012.10.016

Horvath, S. (2013). DNA methylation age of human tissues and cell types. *Genome Biology*, *14*(10), 3156. https://doi.org/10.1186/gb-2013-14-10-r115

Horvath, S., Oshima, J., Martin, G. M., Lu, A. T., Quach, A., Cohen, H., Felton, S., Matsuyama, M., Lowe, D., Kabacik, S., Wilson, J. G., Reiner, A. P., Maierhofer, A., Flunkert, J., Aviv, A., Hou, L., Baccarelli, A. A., Li, Y., Stewart, J. D., … Raj, K. (2018). Epigenetic clock for skin and blood cells applied to Hutchinson Gilford Progeria Syndrome and ex vivo studies. *Aging (Albany NY)*, *10*(7), 1758–1775. https://doi.org/10.18632/aging.101508

Huda, A., & Bushel, P. R. (2013). Widespread Exonization of Transposable Elements in Human Coding Sequences is Associated with Epigenetic Regulation of Transcription.

Transcriptomics: open access, *1*(1). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4028971/

Jansz, N. (2019). DNA methylation dynamics at transposable elements in mammals. *Essays in Biochemistry*, *63*(6), 677–689. https://doi.org/10.1042/EBC20190039

Jin, K. (2010). Modern Biological Theories of Aging. *Aging and Disease*, *1*(2), 72–74.

Jintaridth, P., & Mutirangura, A. (2010). Distinctive patterns of age-dependent hypomethylation in interspersed repetitive sequences. *Physiological Genomics*, *41*(2), 194–200. https://doi.org/10.1152/physiolgenomics.00146.2009

Kim, D., & Salzberg, S. L. (2011). TopHat-Fusion: An algorithm for discovery of novel fusion transcripts. *Genome Biology*, *12*(8), R72. https://doi.org/10.1186/gb-2011-12-8-r72

Klymenko, A. (2021). *Ncbi/sra-tools* [C]. NCBI - National Center for Biotechnology Information/NLM/NIH. https://github.com/ncbi/sra-tools (Original work published 2014)

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359. https://doi.org/10.1038/nmeth.1923

Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, *10*(3), R25. https://doi.org/10.1186/gb-2009-10-3-r25

LaRocca, T. J., Cavalier, A. N., & Wahl, D. (2020). Repetitive elements as a transcriptomic marker of aging: Evidence in multiple datasets and models. *Aging Cell*, *19*(7), e13167. https://doi.org/10.1111/acel.13167

Lee, Y., & Rio, D. C. (2015). Mechanisms and Regulation of Alternative Pre-mRNA Splicing. *Annual Review of Biochemistry*, *84*, 291–323. https://doi.org/10.1146/annurev-biochem-060614-034316

Leinonen, R., Sugawara, H., Shumway, M., & on behalf of the International Nucleotide Sequence Database Collaboration. (2011). The Sequence Read Archive. *Nucleic Acids Research*, *39*(suppl_1), D19–D21. https://doi.org/10.1093/nar/gkq1019

Levine, M. E., Lu, A. T., Quach, A., Chen, B. H., Assimes, T. L., Bandinelli, S., Hou, L., Baccarelli, A. A., Stewart, J. D., Li, Y., Whitsel, E. A., Wilson, J. G., Reiner, A. P., Aviv, A., Lohman, K., Liu, Y., Ferrucci, L., & Horvath, S. (2018). An epigenetic biomarker of aging for lifespan and healthspan. *Aging (Albany NY)*, *10*(4), 573–591. https://doi.org/10.18632/aging.101414

Li, E., & Zhang, Y. (2014). DNA methylation in mammals. *Cold Spring Harbor Perspectives in Biology*, *6*(5), a019133. https://doi.org/10.1101/cshperspect.a019133

Liu, J., Wang, L., Wang, Z., & Liu, J.-P. (2019). Roles of Telomere Biology in Cell Senescence, Replicative and Chronological Ageing. *Cells*, *8*(1), 54. https://doi.org/10.3390/cells8010054

López-Otín, C., Blasco, M. A., Partridge, L., Serrano, M., & Kroemer, G. (2013). The hallmarks of aging. *Cell*, *153*(6), 1194–1217. https://doi.org/10.1016/j.cell.2013.05.039

Lu, A. T., Quach, A., Wilson, J. G., Reiner, A. P., Aviv, A., Raj, K., Hou, L., Baccarelli, A. A., Li, Y., Stewart, J. D., Whitsel, E. A., Assimes, T. L., Ferrucci, L., & Horvath, S. (2019). DNA methylation GrimAge strongly predicts lifespan and healthspan. *Aging*, *11*(2), 303–327. https://doi.org/10.18632/aging.101684

Maklakov, A. A., & Immler, S. (2016). The Expensive Germline and the Evolution of Ageing. *Current Biology*, *26*(13), R577–R586. https://doi.org/10.1016/j.cub.2016.04.012

McGinty, R. K., & Tan, S. (2015). Nucleosome Structure and Function. *Chemical Reviews*, *115*(6), 2255–2273. https://doi.org/10.1021/cr500373h

Nai, G. A., Oliveira, M. C. de, Tavares, G. de O., Pereira, L. F. F., Soares, N. D. S. L., & Silva, P. G. (2015). Evaluación de la genotoxicidad inducida por la administración repetida de anestésicos locales: Un estudio experimental en ratones. *Brazilian Journal of Anesthesiology (Edicion en Espanol)*, *65*(1), 21–26. https://doi.org/10.1016/j.bjanes.2013.07.008

Navarro Gonzalez, J., Zweig, A. S., Speir, M. L., Schmelter, D., Rosenbloom, K. R., Raney, B. J., Powell, C. C., Nassar, L. R., Maulding, N. D., Lee, C. M., Lee, B. T., Hinrichs, A. S.,

Fyfe, A. C., Fernandes, J. D., Diekhans, M., Clawson, H., Casper, J., Benet-Pagès, A., Barber, G. P., … Kent, W. J. (2021). The UCSC Genome Browser database: 2021 update. *Nucleic Acids Research*, *49*(D1), D1046–D1057. https://doi.org/10.1093/nar/gkaa1070

NIH. (2011). *Impronta genética | NHGRI*. Genome.gov. https://www.genome.gov/es/genetics-glossary/Impronta-genetica

OMS. (2021). *Envejecimiento y salud*. Envejecimiento y Salud. https://www.who.int/es/news-room/fact-sheets/detail/ageing-and-health

Oxford, D. (s. f.). *ENVEJECIMIENTO | Definición de ENVEJECIMIENTO por Oxford Dictionary en Lexico.com y también el significado de ENVEJECIMIENTO*. Lexico Dictionaries | Español. Recuperado 6 de abril de 2021, de https://www.lexico.com/es/definicion/envejecimiento

Pal, S., & Tyler, J. K. (2016). Epigenetics and aging. *Science Advances*, *2*(7), e1600584. https://doi.org/10.1126/sciadv.1600584

Pidsley, R., Zotenko, E., Peters, T. J., Lawrence, M. G., Risbridger, G. P., Molloy, P., Van Djik, S., Muhlhausler, B., Stirzaker, C., & Clark, S. J. (2016). Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. *Genome Biology*, *17*(1), 208. https://doi.org/10.1186/s13059-016-1066-1

Pinto da Costa, J., Vitorino, R., Silva, G. M., Vogel, C., Duarte, A. C., & Rocha-Santos, T. (2016). A synopsis on aging—Theories, mechanisms and future prospects. *Ageing Research Reviews*, *29*, 90–112. https://doi.org/10.1016/j.arr.2016.06.005

Plunk, E. C., & Richards, S. M. (2020). Epigenetic Modifications due to Environment, Ageing, Nutrition, and Endocrine Disrupting Chemicals and Their Effects on the Endocrine System. *International Journal of Endocrinology*, *2020*, e9251980. https://doi.org/10.1155/2020/9251980

Rico-Rosillo, M. G., Oliva-Rico, D., & Vega-Robledo, G. B. (2018). Envejecimiento: Algunas teorías y consideraciones genéticas, epigenéticas y ambientales. *Revista Médica del Instituto Mexicano del Seguro Social*, *56*(3), 287–294.

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., & Mesirov, J. P. (2011). Integrative Genomics Viewer. *Nature biotechnology*, *29*(1), 24–26. https://doi.org/10.1038/nbt.1754

Rogers, K., Simic, P., & Guarente, L. P. (2020, enero 30). *Aging | Definition, Process, & Effects*. Encyclopedia Britannica. https://www.britannica.com/science/aging-life-process

Saleh, A., Macia, A., & Muotri, A. R. (2019). Transposable Elements, Inflammation, and Neurological Disease. *Frontiers in Neurology*, *10*. https://doi.org/10.3389/fneur.2019.00894

Sato, F., Tsuchiya, S., Meltzer, S. J., & Shimizu, K. (2011). MicroRNAs and epigenetics. *The FEBS Journal*, *278*(10), 1598–1609. https://doi.org/10.1111/j.1742-4658.2011.08089.x

Sela, N., Mersch, B., Hotz-Wagenblatt, A., & Ast, G. (2010). Characteristics of transposable element exonization within human and mouse. *PloS One*, *5*(6), e10907. https://doi.org/10.1371/journal.pone.0010907

Sorek, R., Ast, G., & Graur, D. (2002). Alu-containing exons are alternatively spliced. *Genome Research*, *12*(7), 1060–1067. https://doi.org/10.1101/gr.229302

Sturm, Á., Ivics, Z., & Vellai, T. (2015). The mechanism of ageing: Primary role of transposable elements in genome disintegration. *Cellular and Molecular Life Sciences*, *72*(10), 1839–1847. https://doi.org/10.1007/s00018-015-1896-0

Swahari, V., & Nakamura, A. (2016). Speeding up the clock: The past, present and future of progeria. *Development, Growth & Differentiation*, *58*(1), 116–130. https://doi.org/10.1111/dgd.12251

Teng, C.-S., Wu, B.-H., Yen, M.-R., & Chen, P.-Y. (2020). MethGET: Web-based bioinformatics software for correlating genome-wide DNA methylation and gene expression. *BMC Genomics*, *21*. https://doi.org/10.1186/s12864-020-6722-x

Tsurumi, A., & Li, W. (2012). Global heterochromatin loss. *Epigenetics*, *7*(7), 680–688. https://doi.org/10.4161/epi.20540

Vogel, C., Silva, G. M., & Marcotte, E. M. (2011). Protein Expression Regulation under Oxidative Stress. *Molecular & Cellular Proteomics : MCP*, *10*(12). https://doi.org/10.1074/mcp.M111.009217

Wang, T., Santos, J. H., Feng, J., Fargo, D. C., Shen, L., Riadi, G., Keeley, E., Rosh, Z. S., Nestler, E. J., & Woychik, R. P. (2016). A Novel Analytical Strategy to Identify Fusion Transcripts between Repetitive Elements and Protein Coding-Exons Using RNA-Seq. *PLOS ONE*, *11*(7), e0159028. https://doi.org/10.1371/journal.pone.0159028

West, M. D., Sternberg, H., Labat, I., Janus, J., Chapman, K. B., Malik, N. N., de Grey, A. D., & Larocca, D. (2019). Toward a unified theory of aging and regeneration. *Regenerative Medicine*, *14*(9), 867–886. https://doi.org/10.2217/rme-2019-0062

ANEXO

Tablas

Sam-pleID	DNA-mAge	mean-MethBy-Sample	min-MethBy-Sample	maxMethBySample	corSam-pleVS-goldstan-dard	meanAbs-Difference-SampleVS-goldstan-dard	predic-tedGen-der
M_1y	0,8499	0,26827	0,00462	0,99192	0,90762	0,08242	male
M_1y	3,9695	0,24596	0,00475	0,99315	0,84268	0,09926	male
M_4y	4,5414	0,26725	0,00433	0,99280	0,91123	0,08020	male
M_8y	7,6745	0,26976	0,00419	0,99507	0,90025	0,08512	male
M_4y	8,8190	0,25279	0,00439	0,99425	0,84492	0,10023	male
M_8y	10,4572	0,26187	0,00469	0,99375	0,84799	0,10128	male
M_11	16,7954	0,25706	0,00449	0,99586	0,83900	0,10217	male
M_22y	24,1321	0,27251	0,00371	0,99416	0,89174	0,08850	male
M_22y	24,7331	0,25964	0,00374	0,99391	0,83950	0,10331	male
M_28y	27,7969	0,26573	0,00457	0,99457	0,84217	0,10281	male
M_28y	38,4781	0,27591	0,00413	0,99506	0,90877	0,08258	male
M_39y	43,1577	0,25835	0,00429	0,99427	0,83875	0,10237	male
M_30y	43,4061	0,26190	0,00486	0,99560	0,83477	0,10413	male
M_30y	43,8923	0,27050	0,00448	0,99437	0,89259	0,08753	male
M_39y	43,9012	0,27165	0,00486	0,99381	0,88860	0,08926	male
M_60y	53,6398	0,25825	0,00349	0,99553	0,83957	0,10207	male
M_60y	62,8239	0,27675	0,00379	0,99511	0,88883	0,09021	male

Tabla Anexa 1 Resultados de Reloj Epigenético de Horvath

La tabla se encuentra ordenada a partir de la predicción del reloj epigenético de Horvath (DNAmAge), La primera columna corresponde a cada una de las muestras con su respectiva edad, desde la columna 3 a la 7 corresponden a datos estadísticos relacionados al grado de metilación de las islas CpG (promedio, mínimo, máximo correlación y diferencia de promedio absoluta)

Muestras Prenatales			Muestras Posnatales			
ID SRA	**Edad de la Muestra**	**Sexo**	**ID SRA**	**Rango de Edad de la Muestra**	**Edad**	**Sexo**
ERR2598056	10 semanas de gestación	Masculino	**ERR2598271**	adolescente	13 años	Masculino
ERR2598057	10 semanas de gestación	Femenino	**ERR2598272**	adolescente	16 años	Masculino
ERR2598077	11 semanas de gestación	Masculino	**ERR2598273**	adolescente	14 años	Masculino

ERR2598078	11 semanas de gestación	Masculino	**ERR2598274**	adolescente	17 años	Masculino
ERR2598079	11 semanas de gestación	Masculino	**ERR2598282**	anciano	58 años	Masculino
ERR2598080	11 semanas de gestación	Femenino	**ERR2598283**	anciano	58 años	Masculino
ERR2598092	12 semanas de gestación	Masculino	**ERR2598293**	infante	221 días	Masculino
ERR2598093	12 semanas de gestación	Femenino	**ERR2598294**	infante	270 días	Masculino
ERR2598094	12 semanas de gestación	Masculino	**ERR2598295**	infante	182,63	Femenino
ERR2598109	13 semanas de gestación	Masculino	**ERR2598308**	adulto medio	53 años	Masculino
ERR2598110	13 semanas de gestación	Masculino	**ERR2598309**	adulto medio	50 años	Masculino
ERR2598111	13 semanas de gestación	Femenino	**ERR2598318**	recién nacido	4 días	Masculino
ERR2598125	16 semanas de gestación	Femenino	**ERR2598319**	recién nacido	15 días	Femenino
ERR2598126	16 semanas de gestación	Masculino	**ERR2598320**	recién nacido	19 días	Femenino
ERR2598127	16 semanas de gestación	Femenino	**ERR2598321**	recién nacido	0 días	Femenino
ERR2598169	4 semanas de gestación	Femenino	**ERR2598322**	recién nacido	0 días	Femenino
ERR2598170	4 semanas de gestación	Masculino	**ERR2598330**	edad escolar	8 años	Masculino
ERR2598171	4 semanas de gestación	Masculino	**ERR2598331**	edad escolar	7 años	Masculino
ERR2598184	5 semanas de gestación	Masculino	**ERR2598340**	bebé	2 años	Femenino
ERR2598185	5 semanas de gestación	Femenino	**ERR2598341**	bebé	4 años	Masculino
ERR2598195	6 semanas de gestación	Masculino	**ERR2598352**	adulto joven	29 años	Masculino
ERR2598196	6 semanas de gestación	Masculino	**ERR2598353**	adulto joven	29 años	Masculino
ERR2598197	6 semanas de gestación	Femenino	**ERR2598354**	adulto joven	32 años	Femenino
ERR2598198	6 semanas de gestación	Femenino	**ERR2598355**	adulto joven	39 años	Masculino
ERR2598212	7 semanas de gestación	Masculino	**ERR2598356**	adulto joven	39 años	Masculino
ERR2598213	7 semanas de gestación	Femenino				
ERR2598214	7 semanas de gestación	Masculino				
ERR2598235	8 semanas de gestación	Masculino				
ERR2598236	8 semanas de gestación	Femenino				
ERR2598237	8 semanas de gestación	Femenino				
ERR2598238	8 semanas de gestación	Masculino				
ERR2598257	9 semanas de gestación	Femenino				

Tabla Anexa 2 Muestras de RNA-seq en Pre y Pos-natales de rombencéfalo humano sano.

ID_orden	ID	candidatos	edad	sexo	pDNA	pLINE	pLTR	pSINE	pOtro	exo_number	prop_exo	tipo	CpG_number	CpG_prop
1	4wf	126731	4	Feme-nino	11,49	37,93	26,7241	23,2758	0,57	348	0,002746	prenatal	4	0,00003
1	4wm1	150157	4	Mascu-lino	14,28	36,46	25,5863	23,2409	0,42643	469	0,003123	prenatal	3	0,00001
1	4wm2	156156	4	Mascu-lino	15,42	37,96	26,1016	19,6610	0,84745	590	0,003778	prenatal	12	0,00007

Tabla Anexa 3 Extracto de archivo usado para generar los gráficos en R

Gen asociado	TE asociado (Subclase – Superfamilia – Familia)	isla CpG
THUMPD3-AS1	Tigger1 DNA TcMar-Tigger	cg25781123
CRTAP	L3 LINE CR1	cg23180365
COA1	L1ME2 LINE L1	cg17408647
HSBP1L1	MLT1A LTR ERVL-MaLR	cg09441152
COA1	Tigger1 DNA TcMar-Tigger	cg17408647
DNAJC18	AluSx SINE Alu	cg19569684
TSC2	L2a LINE L2	cg08331960
COA1	MLT1K LTR ERVL-MaLR	cg17408647
BASP1-AS1	MSTA LTR ERVL-MaLR	cg08771731
ZNF497	MLT1D LTR ERVL-MaLR	cg22568540
THUMPD3-AS1	AluSz SINE Alu	cg25781123
GRIN2C	MIR3 SINE MIR	cg09722397
COA1	L4 LINE RTE	cg17408647
AKT3	L3b LINE CR1	cg11314684
GCDH	L2b LINE L2	cg15341340
CLOCK	HAL1 LINE L1	cg05960024
MGLL	L1MD2 LINE L1	cg03330058
ELP6	MSTA LTR ERVL-MaLR	cg26614073
ELP6	MSTA LTR ERVL-MaLR	cg20524216
ABHD14A-ACY1	MIR SINE MIR	cg04474832
THUMPD3-AS1	Charlie1 DNA hAT-Charlie	cg25781123
SLC20A2	L1MC4 LINE L1	cg14408969
ABHD14A	MIR SINE MIR	cg04474832
TRIOBP	MIR3 SINE MIR	cg19853760
RFC3	LTR16C LTR ERVL	cg24254120

PIPOX	MIR3 SINE MIR	cg06144905
PAWR	MLT1F LTR ERVL-MaLR	cg09418283
NOX4	L2c LINE L2	cg17063929
NOL12	MIR SINE MIR	cg19853760
NCAN	L1MB4 LINE L1	cg06952310
LZTFL1	LTR87 LTR ERVL?	cg08186124
HEXA	Charlie10b DNA hAT-Charlie	cg21801378
GRIN2C	L2c LINE L2	cg09722397
GPR68	MER112 DNA hAT-Charlie	cg03588357
ELP6	AluSx SINE Alu	cg26614073
ELP6	AluSx SINE Alu	cg20524216
CRAMP1	MIRb SINE MIR	cg06810647
CCSER1	THE1D LTR ERVL-MaLR	cg06044899
CCSER1	L1PBa LINE L1	cg06044899
ACY1	MIR SINE MIR	cg04474832

Tabla Anexa 4 Conjunto de genes relacionados a Exonizaciones con islas CpG

Se destaca la única de las 6 islas usadas para la creación del Heatmap de Fig. 17.

Coeficiente de Correlación Lineal		
Rango de valores (-1,1)		**Tipo de Correlación**
±0.96	±1.0	Perfecta
±0.85	±0.95	Fuerte
±0.70	±0.84	Significativa
±0.50	±0.69	Moderada
±0.20	±0.49	Débil
±0.10	±0.19	Muy débil
±0.09	±0.0	Nula

Tabla Anexa 5 Tabla resumen de valores de Coficiente de Correlación Lineal

El coeficiente de correlación puede tomar valores desde -1 hasta 1 pasando por 0. Cuando el valor del coeficiente de correlación lineal es -1 quiere decir que se trata de una asociación lineal perfecta negativa entre dos variables. Mientras cuando el coef. de correlación lineal tiene

un valor de 1, quiere decir que se trata de una asociación perfecta positiva entre dos variables. Pero si el valor es de 0, quiere decir que la asociación es nula o no existe una asociación entre ambas variables.

Figuras

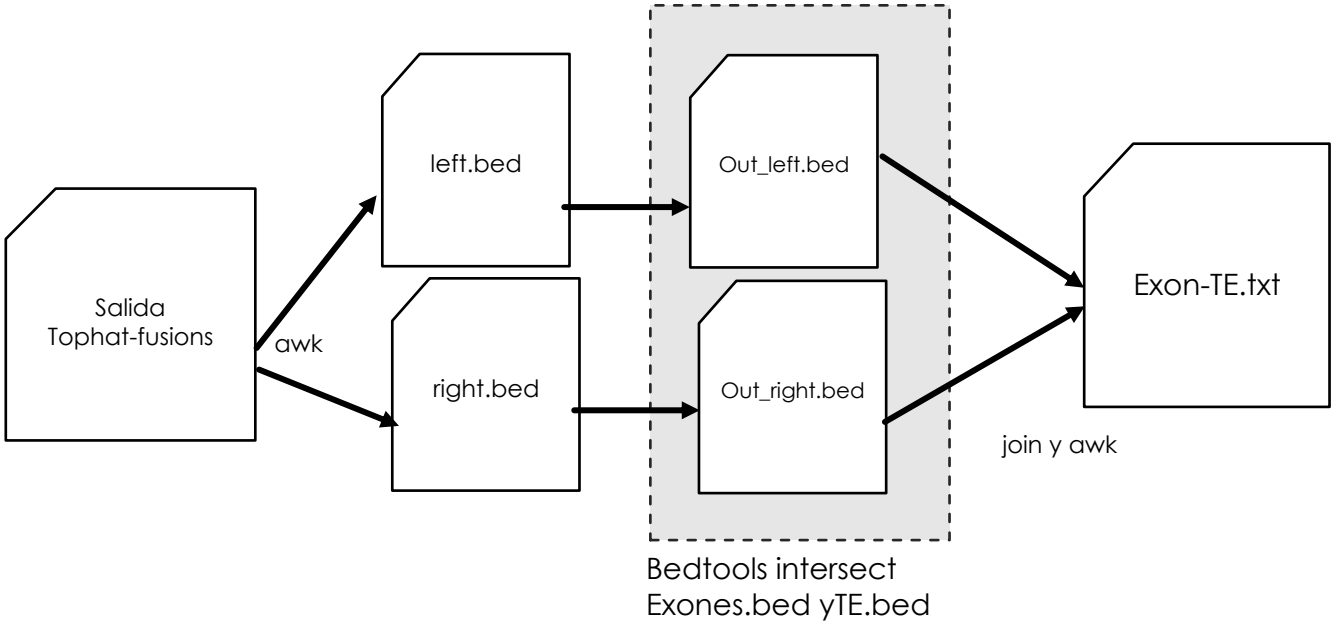


Fig. Anexa 1 Pipeline de funcionamiento de scripts.

La figura señala que la salida de Tophat-fusions se separa en dos partes en formato bed mediante el uso de comando "awk", estas partes, izquierda y derecha (left.bed y right. Bed, respectivamente), con ayuda de Bedtools intersect, son procesados identificando si las coordenadas encontradas en la salida de tophat-fusions corresponden a exones (Exones.bed) o a elementos transponibles (TE.bed). Con ayuda de los comandos "join" y "awk", pasan por los 5 filtros mencionados anteriormente.

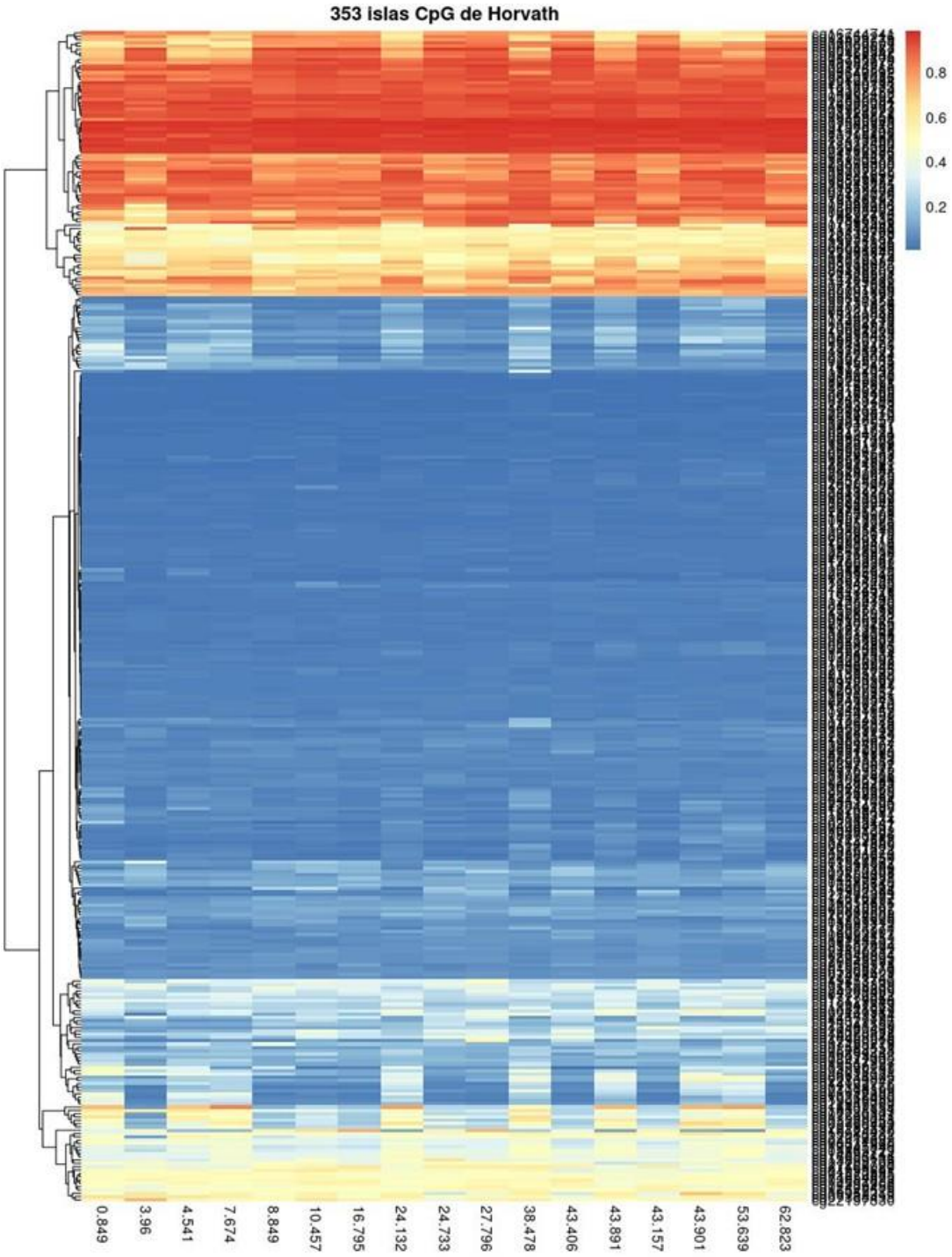


Fig. Anexa 2 Heatmap de predicción de edad cronológica v/s islas CpG

La figura muestra la predicción de cada una de las 17 muestras de metilaciones generadas por el reloj epigenético de Horvath (en sus columnas), mientras que en las filas se encuentra cada una de sus islas 353 CpG, identificadas por su ID. Los colores presentes en la imagen representan el grado de metilación de las islas CpG en cada una de las muestras, mientras

más rojo, es mayor el grado de metilación, y mientras más azul, su grado de metilación es menor.

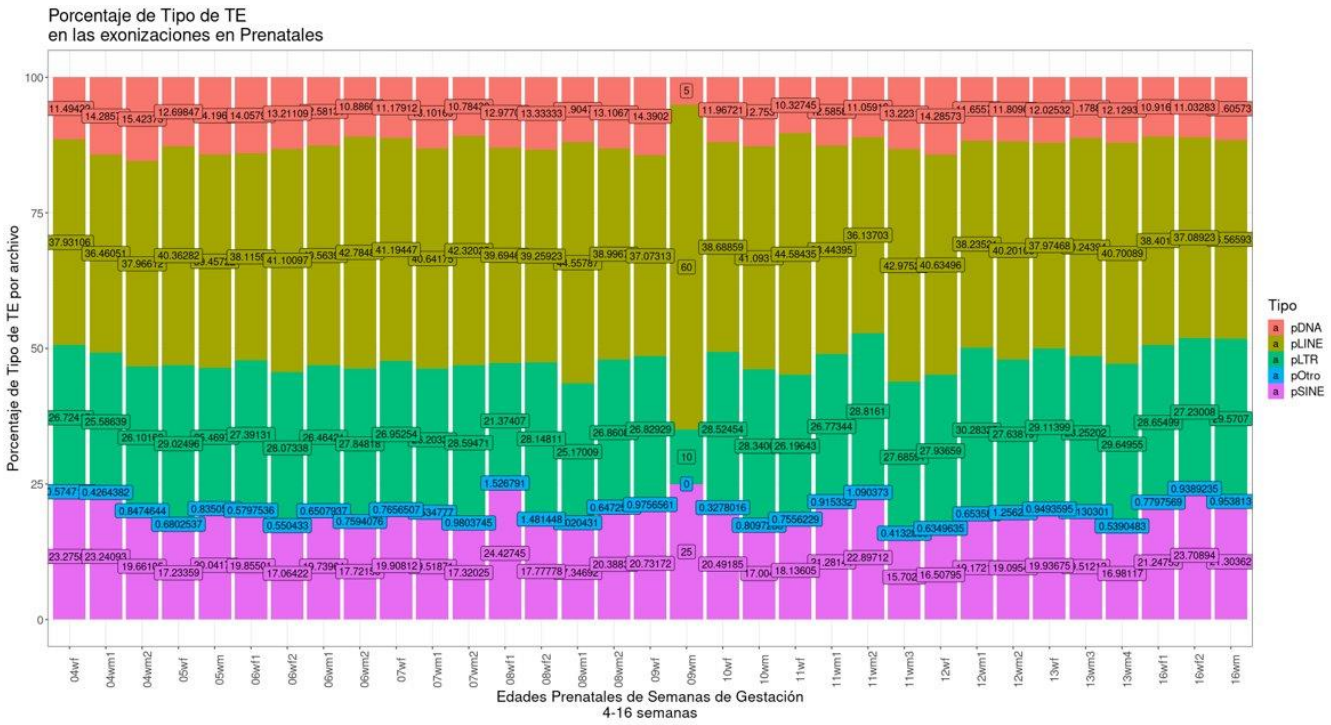


Fig. Anexa 3 Porcentaje de Tipo de Elementos Transponibles en muestras prenatales

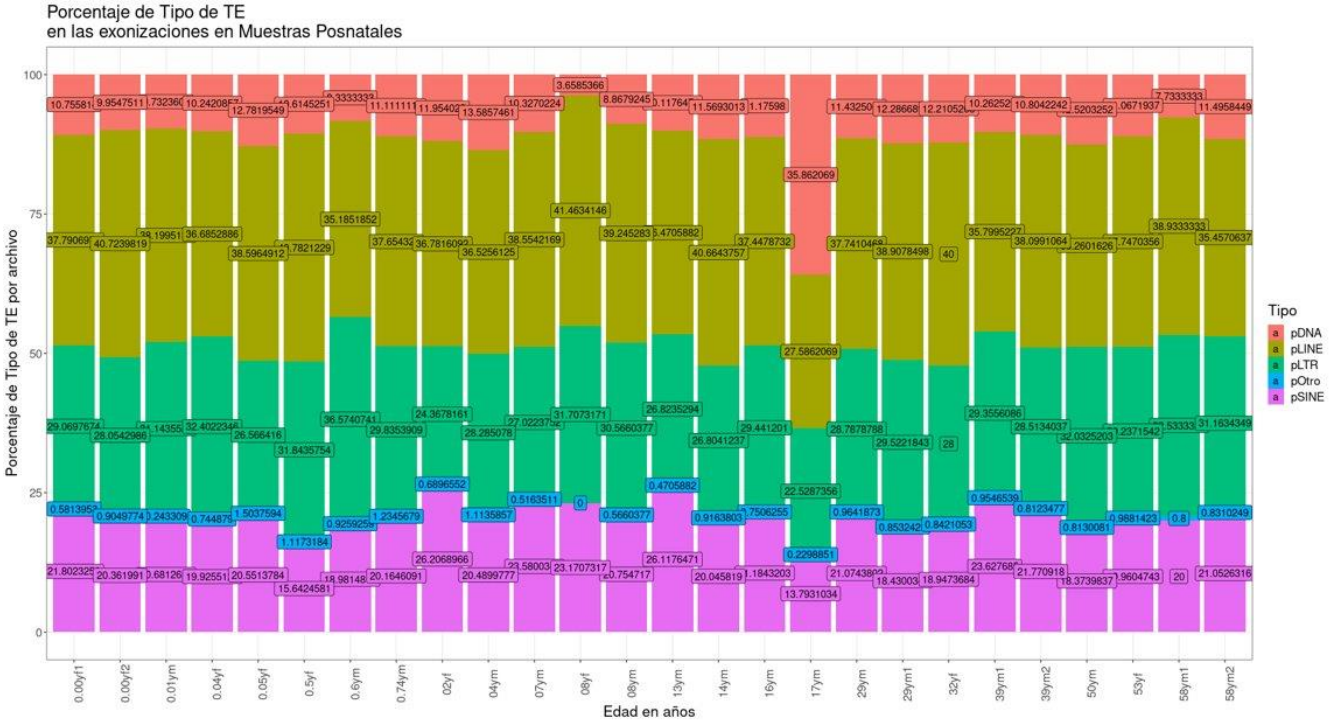


Fig. Anexa 4 Porcentaje de Tipo de Elementos Transponibles en muestras posnatales

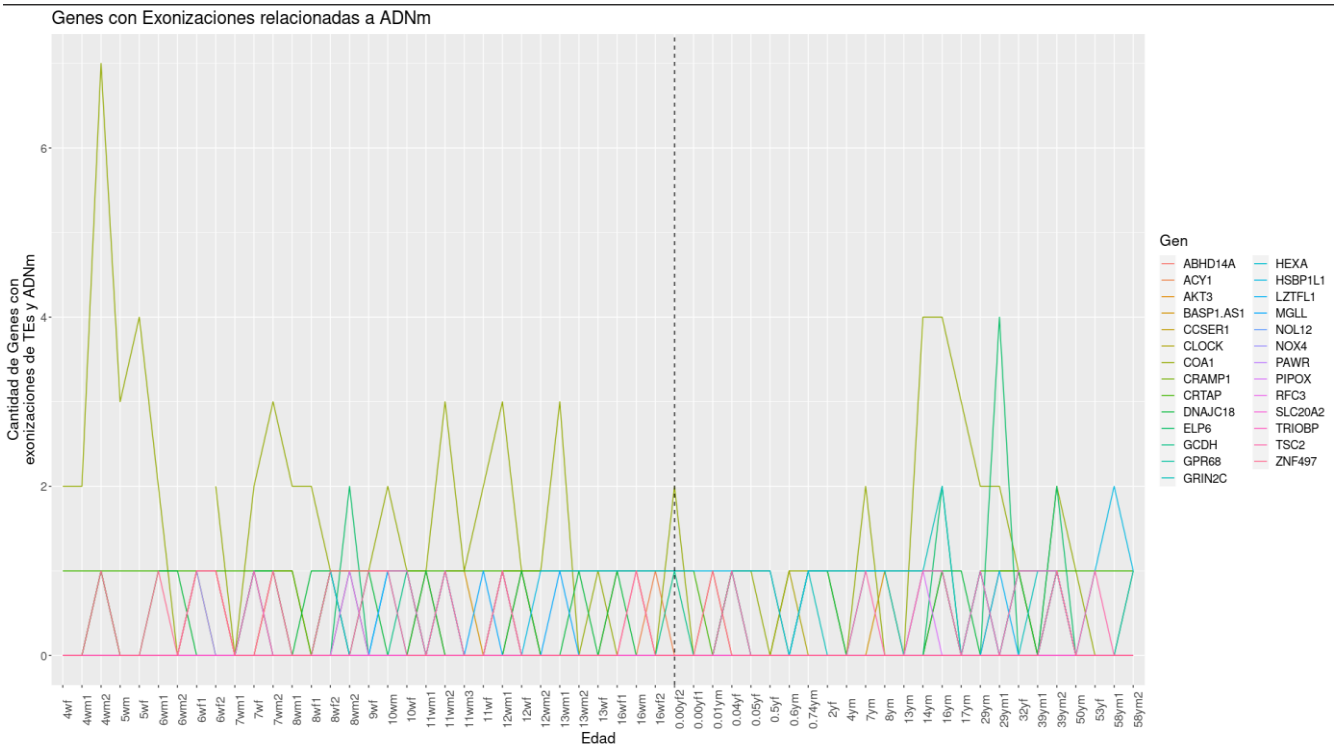


Fig. Anexa 5 Genes presentes en distintas etapas del crecimiento.

Las muestras pre y posnatales se encuentran en el eje x, mientras que en el eje y se encuentra el número de veces que se repite el gen en una misma muestra.

Cálculos

Correlación múltiple:

$$\boldsymbol{R_{123}} = \sqrt{\frac{\boldsymbol{r_{12}^2 + r_{13}^2 - 2r_{12} * r_{13} * r_{23}}}{\boldsymbol{1 - r_{23}^2}}}$$

$Coeficiente\ de\ correlación\ lineal\ prop.exonizaciones\ de\ TE\ y\ edades\ r_{12} = 0.3681$

$Coef.cor.lineal\ prop.exo.de\ TE\ y\ prop.exo.de\ TE\ con\ islas\ CpG\ r_{13} = 0.6311617$

$Coef.cor.lineal\ edades\ y\ prop.exo.de\ TE\ con\ islas\ CpG\ r_{23} = 0.06963$

$$\boldsymbol{R_{123}} = \sqrt{\frac{0.3681^2 + 0.631617^2 - 2 * 0.3681 * 0.631617 * 0.06963}{1 - 0.06963^2}}$$

$$R_{123} = \sqrt{\frac{0.1355 + 0.3989 - 2 * 0.01619}{1 - 0.0048}}$$

$$R_{123} = \sqrt{\frac{0.5344 - 0.03238}{0.995}}$$

$$R_{123} = \sqrt{\frac{0.502}{0.995}}$$

$$R_{123} = \sqrt{0.5045}$$

$$R_{123} = 0.71$$

Cálculo Anexo 1 Correlación múltiple

Códigos

```
#Comandos para editar archivos de anotaciones.

#   Archivo de anotación genoma humano (genoma.gff)
#      Obtención de los datos junto con el id de cada línea, para
evitar exceso de información (Se elimina gran parte de la informa-
ción de la última columna).

awk 'BEGIN{OFS=FS=";"}{print $1}' genoma.gff > genome.gff

#      obtención de las coordenadas de X (gene/exon) en formato
bed, se conserva el cromosoma ($1), las coordenadas de inicio y fi-
nal ($4 y $5), el id del gen o exón ($9), y la cadena ($7)

awk 'BEGIN{OFS=FS="\t"}($3=="X"){print $1,$4,$5,$9,".", "7"}' ge-
nome.gff > X.bed

#   Archivo de anotaciones RepeatMasker (repeatMasker.rmsk)
#      Se los elementos repetidos que no corresponden a TEs
awk 'BEGIN{OFS=FS="\t"}($12=="SINE" || $12=="LINE" || $12=="LTR" ||
$12=="DNA"){print $6,$7,$8,$11"#"$12"#"$13,".", "10"}' repeatMas-
ker.rmsk > TE.bed
```

Código Anexo 1 Obtención de archivos en formato bed, para la ejecución de scripts posteriores.

Uso de comando awk para la edición de archivos.

```
#!/usr/bin/perl -w
#     Creado por Wang
#     Adaptado por Loreto Farías 12/02/2022
#     Propósito:
#     Encontrar puntos de fusión y tipo de splicing para facilitar
el uso del segundo script.

use strict;
use warnings;
use feature 'say';
use diagnostics;

use List::Util qw[min max];

my $filename =$ARGV[0];

open (FUSION, '<',$filename) or die "Can't open '$filename': $!";


my $eachline;
my $linea=0;
my @datas_1;

my $FUSIONconteo=0;


while(defined($eachline=<FUSION>)){
      my @eachFusion=();
      my @chr=();
      my @exon=();
      my @repeat=();
      my @tempL=();
      my @tempR=();

      my $repeatL = 0;
      my $repeatR = 0;
      my $ExonTable = "";
      my $RMSKtable = "";
      my $StartL = 0;
      my $EndL = 0;
      my $StartR = 0;
      my $EndR = 0;
      my $Splice = 0;

      @eachFusion = split(/\s+/, $eachline);# space
      @chr = split(/-/, $eachFusion[0]);


#     Gather fusion information
      my $sizeL = min($eachFusion[8], 100);
      my $sizeR = min($eachFusion[9], 100);
```

```
     my $seqL = substr($eachFusion[18], 30); # Last 20bp
     my $spliceL = substr($eachFusion[18], 48) . substr($each-
Fusion[19], 0, 4); # Last 2bp + First 4bp
     my $spliceR = substr($eachFusion[21], 46) . substr($each-
Fusion[22], 0, 2); # Last 4bp + First 2bp
     my $seqR = substr($eachFusion[22], 0, 20); # First 20bp

     my $StrandL;
     my $StrandR;


     if (($spliceL =~ /GT/) && ($spliceR =~ /AG/)){
          my $l = 'GT';
          my $r = 'AG';

          if (index($spliceL,$l) == index($spliceR,$r)){
          $Splice = 4;
          }
          else{
          $Splice = 3;
          }
     }

     elsif (($spliceL =~ /GC/) && ($spliceR =~ /AG/)){
          my $l = 'GC';
          my $r = 'AG';

          if (index($spliceL,$l) == index($spliceR,$r)){
               $Splice = 4;
          }
          else{
               $Splice = 3;
          }
     }

     elsif (($spliceL =~ /CT/) && ($spliceR =~ /AC/)){
          my $l = 'CT';
          my $r = 'AC';

          if (index($spliceL,$l) == index($spliceR,$r))    {
               $Splice = 4;
          }
          else{
               $Splice = 3;
          }
     }

     elsif (($spliceL =~ /CT/) && ($spliceR =~ /GC/)){
          my $l = 'CT';
          my $r = 'GC';

          if (index($spliceL,$l) == index($spliceR,$r)){
```

```
                $Splice = 4;
          }
          else{
                $Splice = 3;
          }
     }

     else{
          if (($spliceL =~ /GT/) || ($spliceL =~ /GC/) || ($spli-
ceL =~ /CT/)){
                $Splice++;
          }
          if (($spliceR =~ /AG/) || ($spliceR =~ /AC/) ||
($spliceR =~ /GC/))   {
                $Splice++;
          }
     }

     my $number=($eachFusion[4]+$eachFusion[5]+$eachFusion[6]);
     if ($number > 1){
          if ($eachFusion[3] eq "ff"){
                $StartL = $eachFusion[1]-$sizeL+1;
                $EndL = $eachFusion[1];
                $StartR = $eachFusion[2];
                $EndR = $eachFusion[2]+$sizeR-1;
          }

          elsif ($eachFusion[3] eq "fr"){
                $StartL = $eachFusion[1]-$sizeL+1;
                $EndL = $eachFusion[1];
                $StartR = $eachFusion[2]-$sizeR+1;
                $EndR = $eachFusion[2];
          }

          elsif ($eachFusion[3] eq "rf"){
                $StartL = $eachFusion[1];
                $EndL = $eachFusion[1]+$sizeL-1;
                $StartR = $eachFusion[2];
                $EndR = $eachFusion[2]+$sizeR-1;
          }

          elsif ($eachFusion[3] eq "rr"){
                $StartL = $eachFusion[1];
                $EndL = $eachFusion[1]+$sizeL-1;
                $StartR = $eachFusion[2]-$sizeR+1;
                $EndR = $eachFusion[2];
          }
          $FUSIONconteo++;

          #     Write to fusions+Repeat.txt (with repeat)
```

```
           push @datas_1,[@eachFusion[0..10], $StartL, $EndL,
$StartR, $EndR, $seqL, $spliceL, $spliceR, $seqR, $Splice,$FUSION-
conteo];

      }
}


close FUSION;

# Salida del resultado
open my $DATAS , '> fusions_analysis.txt';
say     $DATAS  join "\t",
           qw(chr junctionL junctionR direction jCount jPair
jPairRead contradict  ReadsL ReadsR jScore startL endL startR endR
seqL siteL siteR seqR      Splicing ID_fusion);

for my $j (0 .. $#datas_1){
      # sacamos la información, por l{ineas
      say $DATAS join "\t",@{$datas_1[$j]};
}
close $DATAS;
```

Código Anexo 2 Script editado a partir del de Wang.

Este código se encuentra en formato perl (.pl), identifica el tipo de splicing de las fusiones presentes en el archivo fusions.out, les otorga un número, siendo el 4 el splicing canónico o esperado.

Como salida, este código entrega un archivo que contiene la ubicación los cromosomas que interactúan por lado de cada fusión, el inicio y final de cada lado, la dirección (ff, fr, rf o rr), fragmentos de read, el tipo de splicing id de la fusión, entre otros. Esta información es procesada con un segundo script (Código Anexo 3). La principal diferencia que tiene el script de Wang y el editado es que el script de Wang pide el ingreso a una base de datos MySQL, que contiene la información de las anotaciones del genoma humano, específicamente de los exones, además de las anotaciones de RepeatMasker. Esto produce un excesivo gasto de tiempo y recursos computacionales, lo que fue mejorado en el script editado, donde no se ingresa a una base de datos, por lo que los archivos resultantes se encuentran menos procesados y mediante el segundo script se genera un procesamiento de los datos, con ayuda principalmente de bedtools.

```
#!/bin/bash
#para ejecutar este archivo se requiere de:
      ## tener instalado bedtools
      ## Una lista de los archivos resultantes de tophat_fusion
      ## carpeta con los archivos resultantes de tophat_fusion
      ## las coordenadas de los exones y rpmk editados en formato
BED (Homo_sapiens hg19)
      ## el script editado de Wang (script.pl)
```

```
     ## archivo de coordenadas CpG del reloj de Horvath en formato
BED (Homo_sapiens hg19)

ls tophat/ > Tophat_list

my_file="Tophat_list"
f=0.9
for var in $(cat -t $my_file); do
     mkdir -p $var
     #se eliminan los resultados sin lado izquierdo o derecho
     awk 'BEGIN{OFS=FS="\t"}($17~/[ATCG]/ ){print $0}'
tophat/$var/fusions.out > fusions_2.out
     #ejecución de script en busca de candidatos
     perl script.pl fusions_2.out
     # 'fusions_analysis' salida de script
     # separación de lados Derecho e izquierdo R y L
     # jCount >1, número de reads que coinciden
     awk 'BEGIN{OFS=FS="\t"}($5>1){print $0}' fusions_analy-
sis.txt > L1
     # Splicing == 4 (canónico)
     awk 'BEGIN{OFS=FS="\t"}($20==4){print $0}' L1 > L2
     # Largo de los reads >=25
     awk 'BEGIN{OFS=FS="\t"}($9>=25){print $1,$12,$13,$NF,"."}'
L2 > L3
     # Separación de lados chr-chr → chr   chr
     awk 'BEGIN{OFS=FS="-"}{print $1"\t"$2}' L3 > L4
     # Coordenadas lado izquierdo
     awk 'BEGIN{OFS=FS="\t"}{print $1,$3,$4,$5,$6}' L4 >
$var/left_coordinates.bed
     awk 'BEGIN{OFS=FS="\t"}($5>1){print $0}' fusions_analy-
sis.txt > R1
     awk 'BEGIN{OFS=FS="\t"}($20==4){print $0}' R1 > R2
     awk 'BEGIN{OFS=FS="\t"}($10>=25){print $1,$14,$15,$NF,"."}'
R2 > R3
     awk 'BEGIN{OFS=FS="-"}{print $2}' R3 > $var/right_coordi-
nates.bed
     # Bedtools, uso de f, porcentaje mínimo de intersección entre
A y B
     bedtools intersect -a $var/left_coordinates.bed -b exon_coor-
dinates.bed rpmk_coordinates.bed -wa -wb -f $f > $var/out_left.bed
     bedtools intersect -a $var/right_coordinates.bed -b exon_co-
ordinates.bed rpmk_coordinates.bed -wa -wb -f $f >
$var/out_right.bed
     # Orden de salidas derecha e izquierda
     sort -t $'\t' -k4,4 -n
$var/out_left.bed >$var/sort_out_left.bed
     sort -t $'\t' -k4,4 -n $var/out_right.bed >
$var/sort_out_right.bed
     # Se inserta el título a los archivos ordenados, para poder
usar join
     sed -i
'1s/^/chr_left\tbegin_left\tend_left\tID\tscore_1\tnum\tchr_exon-
```

```
rpmk\t\begin_exon-rpmk\tend_exon-rpmk\tname_exon-
rpmk\tscore_2\tstrand\n/' $var/sort_out_left.bed
     sed -i
'1s/^/chr_rigth\tbegin_rigth\tend_rigth\tID\tscore_1\tnum\tchr_exon
-rpmk\t\begin_exon-rpmk\tend_exon-rpmk\tname_exon-
rpmk\tscore_2\tstrand\n/' $var/sort_out_right.bed
     join --nocheck-order -t $'\t' -1 4 -2 4
$var/sort_out_left.bed $var/sort_out_right.bed > $var/join_out.txt
     # Selección de los cromosomas iguales por lado ADEMÁS DE LA
MISMA CADENA
     awk 'BEGIN{OFS=FS="\t"}($2==$13){print $0}'
$var/join_out.txt > $var/precandidatos.txt
     awk 'BEGIN{OFS=FS="\t"}($12==$23){print $0}' $var/precandida-
tos.txt > $var/candidatos.txt
     # Se eliminan los resultados que corresponden a exon-exon
     awk
'BEGIN{OFS=FS="\t"}(substr($10,1,4)!=substr($21,1,4)){print $0}'
$var/candidatos.txt > $var/TE-exon_TE_TE.txt
     # Se eliminan los resultados que corresponden a TE-TE
     awk 'BEGIN{OFS=FS="#"}($5~/^$/){print $0}' $var/TE-
exon_TE_TE.txt > $var/candidatos_TE_exon.txt
     # Se muestran solo las coordenandas de ambos lados junto a su
cadena, y el tipo de TE o exón
     awk 'BEGIN{OFS=FS="\t"}($2==$13){print
$1,$2,$3,$4,$10,$12,$13,$14,$15,$21,$23}' $var/candida-
tos_TE_exon.txt > $var/TE_exon.txt
     ## COORDENADAS de TODAS LAS EXONIZACIONES :p la BEGIN=menor
coordenada inciial END= mayor coordenada final
     #awk 'BEGIN{OFS=FS="\t"}{if($8<=$19){if($9>=$20){print
$7,$8,$9,$1"#"$10"#"$21,$11,$12} else {print
$7,$8,$20,$1"#"$10"#"$21,$11,$12}} else {if($9>=$20) {print
$7,$19,$9,$1"#"$10"#"$21,$11,$12} else {print
$7,$19,$20,$1"#"$10"#"$21,$11,$12}}}' $var/candidatos_TE_exon.txt >
$var/coordenada_exonizacion.bed
     awk 'BEGIN{OFS=FS="\t"}{if(substr($10,1,4)=="exon"){print
$7,$8,$9,$1"#"$21,$11,$12} else {print
$7,$19,$20,$1"#"$10,$22,$23}}' $var/candidatos_TE_exon.txt >
$var/coordenada_exonizacion.bed
     # Se eliminan las posibles duplicaciones
     ## sort_TE-exon.txt es el archivo que contiene todas las exo-
nizaciones de TEs encontradas en el análisis.
     sort -u $var/TE_exon.txt > $var/exonizaciones_de_TEs.txt
     #### búsqueda de genes que contienen exonizaciones de TEs y
CpG de Horvath
     bedtools intersect -a genes.bed -b $var/coordenada_exoniza-
cion.bed -s -wa -wb | sort -u > $var/genes_exonizaciones_$var.bed
     bedtools intersect -a $var/genes_exonizaciones_$var.bed -b
CpG_islands/coordenadas_cpg_25k.bed -s -wa -wb | sort -u > $var/ge-
nes_ex-te_DNAM_25K_$var.bed
     awk 'BEGIN{OFS=FS="\t"}{print $4}' $var/genes_ex-
te_DNAM_25K_$var.bed | sort | uniq -c | sort -nr > $var/num_ge-
nes_ex-te_DNAM_25K
```

```
      sed -i '1i '$var'' $var/num_genes_ex-te_DNAM_25K
      wc -l L1 L2 L3 R1 R2 R3 $var/candidatos.txt $var/TE-
exon_TE_TE.txt $var/candidatos_TE_exon.txt > $var/5_filtros
      rm -r L1 L2 L3 L4 R1 R2 R3
      ## Guarda los resultados del script en la carpeta
      mv fusions_analysis.txt $var
      ## Cuenta la cantidad de familias de TEs (CAMBIAR NOMBRE)
      awk 'BEGIN{OFS=FS="#"}{print $2}' $var/exonizacio-
nes_de_TEs.txt | sort | uniq -c > $var/$var\_TE-exon
      ## Se añade un título, con el nombre del archivo, esto para
poder ordenarlas
      sed -i '1i '$var'' $var/$var\_TE-exon
done
mkdir f_$f
mv Tophat_Fusion_* f_$f
cp f_$f/Tophat_Fusion_*/genes_exonizaciones_*.bed f_$f/
cp f_$f/Tophat_Fusion_*/genes_ex-te_DNAM_*.bed f_$f/
cat f_$f/Tophat_Fusion_*/num_genes_ex-te_DNAM_25K > num_genes_ex-
te_DNAM_25K
cat f_$f/Tophat_Fusion_*/Tophat_Fusion_* > TEs
mv num_genes* TEs f_$f
```

Código Anexo 3 Script de procesamiento

Este script utiliza el resultado del Código Anexo 2 para analiza uno a uno los archivos de salida de tophat fusión, ejecutando los 5 filtros para encontrar las exonizaciones de TEs. Este Scrip entrega los resultados del objetivo 2 y parte del objetivo 3

```
#llamado a librerías
library(ggplot2)
library(tidyr)
prenatal <- read.delim("muestras_prenatal.txt", sep="\t",
header=TRUE)

#Grafico proporciones de Familias de Elementos Transponibles en las
distintas muestras
prenatal %>% gather("Tipo","Value", -ID, -edad_2, -SampleName, -
edad_1, -sexo,-exo_number, -LTR, -DNA, -SINE, -LINE, -Otro, -fu-
sion_analysis, -prop_exo) %>%
ggplot(aes(ID, Value, fill=Tipo)) +
geom_bar(position="stack", stat = "identity")+geom_label(aes(la-
bel=Value), position = position_stack(.5)) +
theme_bw()+ theme(text=element_text(size=16), axis.text.x = ele-
ment_text(angle=90)) +
labs(x='Semanas de Gestación', y='Porcentaje de Tipo de TE', ti-
tle='Porcentaje de Tipo de TE\nen las exonizaciones en Muestras
Prenatales')+ facet_wrap(~ sexo)

#leer el archivo con las muestras
muestras <- read.delim("muestras.txt", sep="\t", header=TRUE)

#GRAFICO Exonizaciones de TE v/s Edad cronológica
```

```
ggplot(muestras3, aes(ID_orden2, prop_exo, fill=tipo))+
 geom_point(aes(color=sexo))+
geom_smooth(method=lm)+labs(y='Proporción de número\nde Exonizacio-
nes de TEs', x='Edad de las muestras', title='Proporción de número
de Exonizaciones de TEs vs edad de las Muestras') +
 theme(text=element_text(size=25), axis.text.x = element_text(an-
gle=90))+
scale_x_discrete(limits=c("4w", "5w", "6w", "7w", "8w", "9w",
"10w", "11w", "12w", "13w", "16w","","","0.00y", "0.01y", "0.04y",
"0.05y", "0.5y", "0.6y", "0.74y", "2y", "4y", "7y", "8y", "13y",
"14y", "16y", "17y", "29y", "32y", "39y", "50y", "53y", "58y"))+
geom_point(aes(ID_orden2, CpG_prop))+geom_smooth(method=lm)

#GRAFICO Exonizaciones de TE con islas CpG v/s Edad cronológica
ggplot(muestras3, aes(ID_orden2, CpG_prop, fill=tipo))+
geom_point(aes(color=sexo))+
geom_smooth(method=lm)+
labs(y='Proporción de número\nde Exonizaciones de TEs con islas
CpG', x='Edad de las muestras', title='Proporción de número de Exo-
nizaciones de TEs con islas CpG vs edad de las Muestras') +
theme(text=element_text(size=25), axis.text.x = element_text(an-
gle=90))+
scale_x_discrete(limits=c("4w", "5w", "6w", "7w", "8w", "9w",
"10w", "11w", "12w", "13w", "16w","","","0.00y", "0.01y", "0.04y",
"0.05y", "0.5y", "0.6y", "0.74y", "2y", "4y", "7y", "8y", "13y",
"14y", "16y", "17y", "29y", "32y", "39y", "50y", "53y", "58y"))

#Grafico Exonizaciones de TEs v/s Exonizaciones de TEs con islas
CpG
ggplot(muestras3, aes(prop_exo, CpG_prop, fill=tipo))+
geom_point(aes(color=tipo))+
geom_smooth(method=lm)+
labs(y='Proporción de número\nde Exonizaciones de TEs con islas
CpG', x='Proporción de número de\nExonizaciones de TEs', ti-
tle='Proporción de Número de Exonizaciones de TEs vs Exonizaciones
de TEs con islas CpG', legend='Tipo de Muestra') +
theme(text=element_text(size=25))

#Grafico Tridimensional
> scatterplot3d(x=c(orden), prop_cpg, prop_exo, pch = 19, color =
"blue", xlab="Edades", ylab="\n\nProporción número de\n exonizacio-
nes con islas CpG", zlab="Proporción de número de exonizaciones",
main="Proporción de Exonizaciones de TE\nv/s Proporción de Exoniza-
ciones de TE con islas CpG\nv/s Edad cronológica",
```

```
x.ticklabs=c("4w", "8w", "16w","0.0y", "8y", "29y","39y","58y"),
grid=TRUE, box=TRUE, cex.axis=1, angle=45)$plane3d(my.lm)
> my.lm

Call:
lm(formula = prop_exo ~ orden + prop_cpg)

Coefficients:
(Intercept)        orden     prop_cpg
  1.534e-03    4.017e-05    3.078e+01
```

Código Anexo 4 Código de creación de gráfico en Rs de la actividad 3.4