Facultad de Ingeniería
Escuela de Ingeniería Civil en Bioinformática

Análisis loci-específico de elementos transponibles en la progresión de esclerosis lateral amiotrófica en el modelo de ratón transgénico SOD1^{G93A}

Memoria para optar al título de Ingeniero Civil en Bioinformática

Alumno: Esteban Andrés Arancibia G.
Profesor Tutor: Dr. Braulio Valdebenito
Profesor Informante: Dr. Gonzalo Riadi
Talca, Chile.
2021

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, 2021

Indice

Resumen

Esclerosis Lateral Amiotrófica (ELA), es una enfermedad neurodegenerativa, en la cual las neuronas motoras mueren y ya no pueden enviar mensajes a los músculos. Esto afecta una serie de regiones a lo largo del cuerpo, como los brazos y las piernas, y también la capacidad de hablar y deglutir. Con el tiempo, más grupos musculares desarrollan problemas, generando dificultades en el movimiento. Finalmente, cuando los músculos en la zona torácica dejan de trabajar, esto resulta en un fallo respiratorio, y la muerte del paciente. Aproximadamente el 10% de los casos de ELA son hereditarios y 90% son esporádicos. Dentro del primer grupo, varios genes se han asociado causalmente a la enfermedad, siendo la mutación $SOD1^{G93A}$ una de las más comunes. Ratones transgénicos que llevan la mutante $SOD1^{G93A}$ de humano recapitulan la progresión de la enfermedad. Se ha especulado que como los casos hereditarios y esporádicos son fenotípicamente indistinguibles, hallazgos en el modelo de ratón $SOD1^{G93A}$ pueden ser informativos para ambos tipos de ELA.

Los Elementos Transponibles (TEs) tienen la capacidad de movilizarse en un genoma, y con el paso del tiempo aumentar su número de copias. Debido a esta naturaleza repetitiva, confunden estudios de expresión génica con datos de RNA-seq. Trabajos recientes han mostrado que TEs están expresados en enfermedades neurodegenerativas, incluida ELA, pero no han utilizado herramientas en las que se mantiene la ubicación genómica del TE expresado. Por esto, el impacto loci-específico que pueden tener en regulación génica en ELA no ha sido estudiado. Recientemente, se han hecho públicas herramientas que permiten incluir TEs de manera loci-específica en estos análisis, pero no han sido utilizadas ampliamente.

Exclusivamente desde el modelo de ratón $SOD1^{G93A}$ hay datos públicos de RNA-seq en las distintas etapas de ELA. En este estudio se analizaron esos datos utilizando las nuevas herramientas para estudio de TEs a fin de entender el rol de los TEs en la progresión de la enfermedad, y para evaluar el potencial impacto loci-específico que estos tienen en la expresión de algunos genes. Como principal resultado de este trabajo, se encontró un grupo de genes (*Slc15a2, Ube3a, Snhg14, Chd9, Serpina3n, Cep85*), cuya expresión parece estar modulada por TEs dentro de ellos. Finalmente, mediante búsqueda bibliográfica, se confirmó una relación entre dichos genes y la neurodegeneración, a partir de lo cual se puede especular que los TEs podrían de manera indirecta vincularse a la progresión de ELA en el modelo estudiado.

Introducción

1. Esclerosis Lateral Amiotrófica (ELA)

Según el *National Institute of Neurological Disorders and Stroke*[1] existen más de 600 enfermedades neurodegenerativas entre las que destacan, la Enfermedad de Alzheimer, la Enfermedad de Parkinson, la Enfermedad de Huntington y la Esclerosis Lateral Amiotrófica. Estas enfermedades se caracterizan por una progresiva disfunción del Sistema Nervioso Central.[2] En particular, resultan en una degeneración y/o muerte de las células nerviosas, lo cual causa problemas con el movimiento del cuerpo (llamado ataxia) o con el funcionamiento mental (llamado demencia) de las personas que las padecen.[3] Las causas de estas patologías aún no están completamente dilucidadas, ya que hay casos hereditarios, y también hay casos adquiridos (también denominados esporádicos).

Este estudio se concentrará en la enfermedad llamada Esclerosis Lateral Amiotrófica (ELA). El nombre de esta enfermedad tiene dos componentes: (i) “Esclerosis lateral”, que se refiere al endurecimiento de las columnas laterales de la médula espinal; y (ii) “Amiotrófica”, que se refiere a la atrofia muscular producto de que las neuronas motoras inferiores están siendo afectadas, lo que causa debilidad y contracciones musculares involuntarias.[4]

Esta enfermedad fue descrita por primera vez por Charcot en el año 1874, quien es considerado el padre de la neurología (Kumar et al., 2011), En su descripción basada en una observación y documentación, Charcot identificó pacientes débiles sin dificultades sensoriales, epilepsia u otros movimientos involuntarios. Mientras que todos eran débiles, algunos eran espásticos con contracturas y otros eran con atrofia. En ese entonces, Charcot no tenía ideas firmes sobre la etiología de ELA. Reconoció que aproximadamente un tercio de los casos tenían exposición previa a la humedad o al frío, pero no atribuyó la enfermedad a estos problemas

1 https://www.ninds.nih.gov/
2 https://www.neuronup.com/
3 https://www.neurodegenerationresearch.eu/
4 http://www.alsa.org/

ambientales. El sesgo primario de Charcot era que todos los trastornos neurológicos primarios se debían a influencias hereditarias, y que las familias transmitían de generación en generación una propensión a las enfermedades neurológicas. Dependiendo de las influencias ambientales, las manifestaciones del trastorno neurológico cambiaban de un individuo a otro, de modo que un árbol genealógico podría incluir varios trastornos neurológicos fenomenológicamente diferentes. Entre sus pacientes, Charcot no encontró casos de otros familiares con ELA (Goetz, 2000).

Se estima que la prevalencia (cantidad total de individuos afectados) de ELA es 5 de cada 100.000 personas en todo el mundo. En Europa, la incidencia (cantidad de casos nuevos en un período de tiempo determinado) varía de 2 a 3 casos por cada 100.000 individuos. La incidencia es menor en el este de Asia (~0,8 casos por 100.000 individuos) y el sur de Asia (~0,7 casos por 100.000 individuos). En algunas regiones como Guam y la península de Kii de Japón la incidencia reportada fue muy alta, pero se ha reducido sustancialmente en los últimos 30 años por razones que aún no están claras. En áreas donde diferentes poblaciones pre-coloniales viven muy cerca (como en América del Norte), la incidencia de la ELA en poblaciones indígenas es baja (0,63 casos por 100.000 individuos), mientras que las incidencias reportadas en regiones de poblaciones relativamente homogéneas (como Irlanda, Escocia y las Islas Feroe) son altas (2,6 casos por 100.000 individuos) (Hardiman et al., 2017).

1.1 Etapas de la Enfermedad[5]

A grandes rasgos, en ELA, las neuronas motoras inferiores se desgastan o mueren y ya no pueden enviar mensajes a los músculos. Esto afecta una serie de regiones a lo largo del cuerpo, como se muestra en la ***Figura 1***. La debilidad puede afectar primero los brazos y las piernas mostrando espasticidad, atrofia e incluso calambres, o la capacidad de hablar y deglutir (disartria y disfagia, respectivamente). A medida que la enfermedad empeora, más grupos musculares desarrollan problemas. Con el tiempo, esto lleva a debilitamiento muscular, espasmos e incapacidad para mover los brazos, las piernas y el cuerpo. Finalmente, cuando los

5 https://alstreatment.com/

músculos en la zona torácica dejan de trabajar, esto resulta en un fallo respiratorio, y la muerte del paciente.

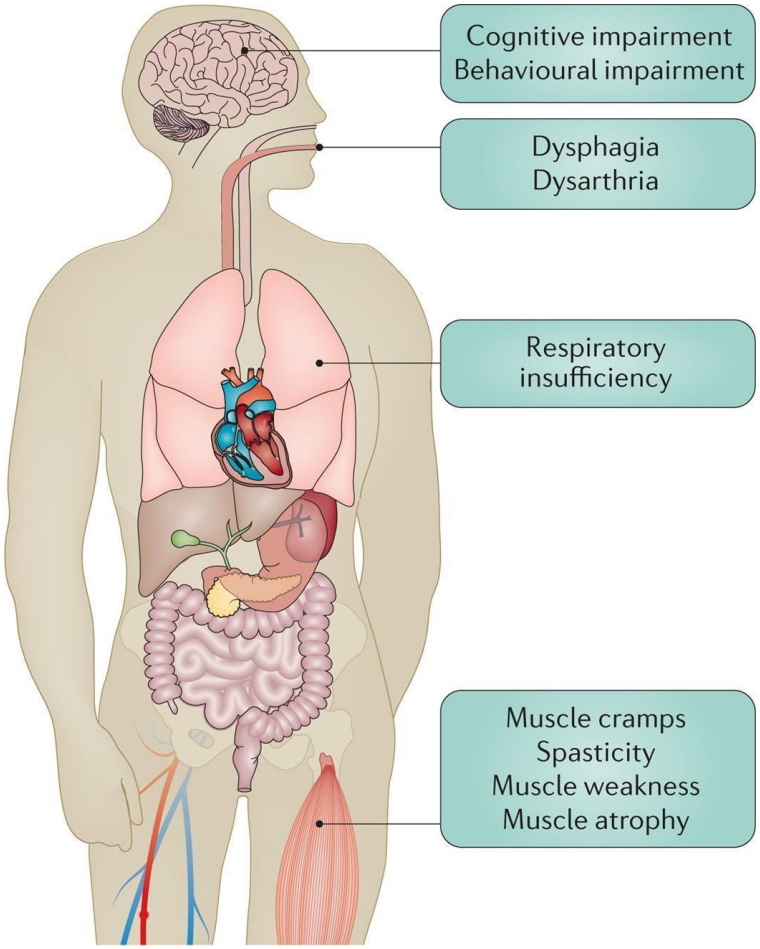


*Figura 1**: Manifestaciones Clínicas de ELA (figura tomada de Hardiman et al., 2017)***

Las manifestaciones motoras son los principales síntomas de ELA como la debilidad muscular, la disfagia (dificultad para tragar) y disartria (dificultad con el habla). Pero, además, cerca de la mitad de los pacientes presentan síntomas no motores, como deterioro cognitivo .

ELA no afecta los sentidos, y la mayoría de las personas es capaz de pensar normalmente. Sin embargo, una pequeña cantidad presenta demencia, lo que provoca problemas de memoria, discapacidad cognitiva o desorden del comportamiento.[6] De hecho, existe evidencia que vincula ELA con Demencia Frontotemporal (Pattamatta et al., 2018), (Balendra & Isaacs, 2018).

6 https://medlineplus.gov/

Recapitulando lo expuesto anteriormente, ELA es una enfermedad neurológica progresiva caracterizada por la destrucción de las células nerviosas que son responsables de controlar el movimiento muscular voluntario. Algunos ejemplos de movimiento muscular voluntario incluyen masticar, caminar, hablar y respirar. Cabe mencionar que algunos pacientes pueden no experimentar los mismos síntomas y, para otros, la enfermedad progresa más lentamente.

En general, la progresión de ELA puede dividirse en tres etapas distintas: temprana, media y tardía.

1.1.1 Etapa Temprana

En las primeras etapas de la progresión de ELA, los pacientes tienden a tener músculos débiles. Es común experimentar espasmos y calambres musculares, y pérdida de masa muscular. Estos síntomas pueden ocurrir en una sola región del cuerpo, o síntomas leves podrían afectar a más de una parte del cuerpo. La persona puede estar fatigada, faltarle equilibrio, decir mal sus palabras, tener un agarre débil o tropezar al caminar.

1.1.2 Etapa Media

En las etapas medias de ELA, los síntomas en los músculos se generalizan. Algunos músculos pueden paralizarse, mientras que otros no se ven afectados o simplemente se debilitan. Los músculos no utilizados pueden provocar contracturas, donde las articulaciones se vuelven dolorosas, rígidas e incluso deformadas. Si una persona se cae, es posible que no pueda volver a levantarse sola. Ya no pueden conducir y experimentan debilidad al tragar, así como una mayor dificultad para controlar la saliva y comer. La debilidad en los músculos asociados a la respiración puede provocar insuficiencia respiratoria, especialmente al acostarse. Algunas personas experimentan episodios de llanto o risa incontrolables.

1.1.3 Etapa Tardía

En las etapas finales de ELA, la mayoría de los músculos voluntarios se han paralizado. Los músculos que ayudan a mover el aire dentro y fuera de los pulmones de la persona están gravemente comprometidos. La movilidad, en este punto, está severamente limitada. La persona necesita de ayuda en la mayoría de sus funciones diarias personales. Puede que no sea posible hablar, beber y comer. La mala respiración puede provocar fatiga, dolores de cabeza, pensamiento confuso y una mayor susceptibilidad a la neumonía. La insuficiencia respiratoria es la causa principal de muerte para las personas con ELA.

1.2 Componentes Genéticos

A pesar de que se han realizado varios estudios sobre ELA, no se conoce por completo las causas posibles a nivel genético. Se ha estimado que el 90% de los casos son esporádicos y tan solo el 10% son hereditarios/familiares, de los cuales aproximadamente el 20% están relacionados con mutaciones en el gen SOD1 (Phatnani et al., 2013). Otros genes que se han asociado a ELA familiar son TARDBP, FUS, VCP, C9orf72 y PFN1, los cuales representan aproximadamente el 60%–70% de esos casos (***Figura 2***, Renton et al., 2014).

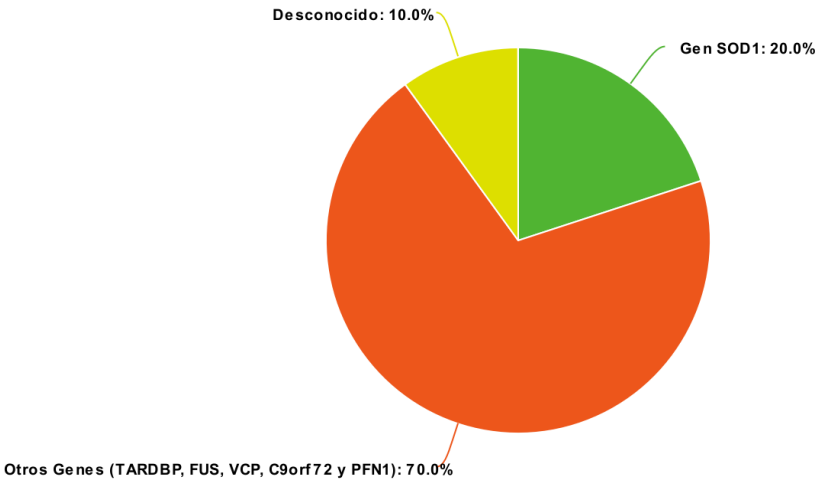


*Figura 2****: Proporciones en los casos hereditarios/familiares de la enfermedad.***

Actualmente se estima que más de 25 genes están asociados a ELA (ya sea familiar o esporádica), siendo las mutaciones en SOD1 y C9orf72 las más comunes (Nguyen et al., 2018). Gracias a avances en secuenciación masiva en paralelo, como la secuenciación del genoma completo y la secuenciación del exoma completo, se han ido identificando nuevos genes asociados a la enfermedad. Con estos estudios se han identificado variantes raras que confieren riesgo de enfermedad. Así, recientemente se han identificado 9 genes que portan tales variantes causales raras, incluyendo TBK1, CHCHD10, TUBA4A, MATR3, CCNF, NEK1, C21orf2, ANXA11 y TIA1 (Nguyen et al., 2018).

Como en términos de fenotipo, ELA familiar y ELA esporádica no se pueden diferenciar, se cree que hallazgos en modelos de ELA pueden ser útiles para ganar conocimiento transversal sobre diversos aspectos de la enfermedad. Un modelo de la enfermedad establecido, y altamente utilizado es el de ratones transgénicos que llevan la versión mutante de SOD1 de humanos (Phatnani et al., 2013).

Este trabajo estará exclusivamente enfocado en el modelo de ratón con la mutante $SOD1^{G93A}$. El motivo de esto es que de toda la información de expresión a partir de RNA sequencing (explicado en la siguiente sección) disponible públicamente en la base de datos Sequence Read Archive (SRA), únicamente de este modelo hay datos a lo largo de la progresión de la enfermedad. Adicionalmente, en el trabajo de Phatnani et al. (2013), que es donde se generaron estos datos, no se hizo estudio de TEs. Se hace particular énfasis en estos puntos, ya que está directamente asociado con la hipótesis de este trabajo.

Estos estudios han llevado a hacer una correlación entre la duración de las etapas de la enfermedad en humanos y en ratones[7], como se muestra en la *Tabla 1* (Aziza, 2018).

Modelo	Etapa Temprana	Etapa Media	Etapa Tardía
Humano	~12 meses	~30 meses	~61 meses
Ratón	~8 semanas	~13 semanas	~17 semanas

Tabla 1: Correlación de las etapas de la enfermedad entre humanos y ratones.

7 https://emedicine.medscape.com/

2. RNA-seq

La secuenciación de ARN (RNA sequencing en inglés, abreviado, y referido de aquí en adelante como RNA-seq) es la aplicación de cualquier variedad de técnicas de secuenciación para estudiar el ARN. A grandes rasgos, lo que se hace para secuenciar el ARN es convertirlo a cDNA, ya que este último es más estable. Luego de eso, se somete a los protocolos estándar de secuenciación, que incluyen adición de adaptadores y la posterior secuenciación (entendiéndose como secuenciación en este punto al proceso específico de generar los reads, y no a la etapa en general) *(**Figura 3**, parte superior)*. Con el análisis de datos de RNA-seq se puede realizar un ensamble de los transcritos, identificar eventos de *splicing* alternativo o incluso descubrir nuevos transcritos, además de hacer la cuantificación de transcripción por gen. Esto último es lo que comúnmente se conoce hoy como análisis de expresión de genes desde RNA-seq, el cual se realiza mediante un mapeo de los reads generados en la etapa de secuenciación contra un genoma de referencia para saber a qué genes corresponde cada uno. Un área de estudio derivada de esto es el análisis de expresión diferencial de genes, lo que es útil cuando se requiere comparar dos condiciones diferentes, por ejemplo, un paciente enfermo contra uno sano (Chu & Corey, 2012).

Aunque RNA-seq sigue siendo una tecnología en desarrollo, ofrece varias ventajas clave sobre las tecnologías existentes. Una de las más importantes es que, a diferencia de los enfoques basados en hibridación, como microarreglos, RNA-seq no se limita a detectar transcripciones que corresponden a secuencias ya conocidas. Esto hace que RNA-seq sea particularmente atractivo para organismos no modelo con secuencias genómicas que aún no se han determinado. Además, RNA-seq puede revelar la ubicación precisa de los límites de transcripción, con nivel de resolución a nivel de base *(**Figura 3**, parte inferior)*. Estos factores hacen que RNA-seq sea útil para estudiar transcriptomas complejos. Finalmente, RNA-seq también puede revelar variaciones de secuencia (por ejemplo, *single-nucleotide polymorphisms,* conocidos comúnmente como SNPs) en las regiones transcritas (Z. Wang et al., 2009).

A pesar de que RNA-seq es utilizada ampliamente para estudios de expresión génica, presenta un problema al tratar de estudiar reads originados en elementos repetitivos, como los elementos transponibles.

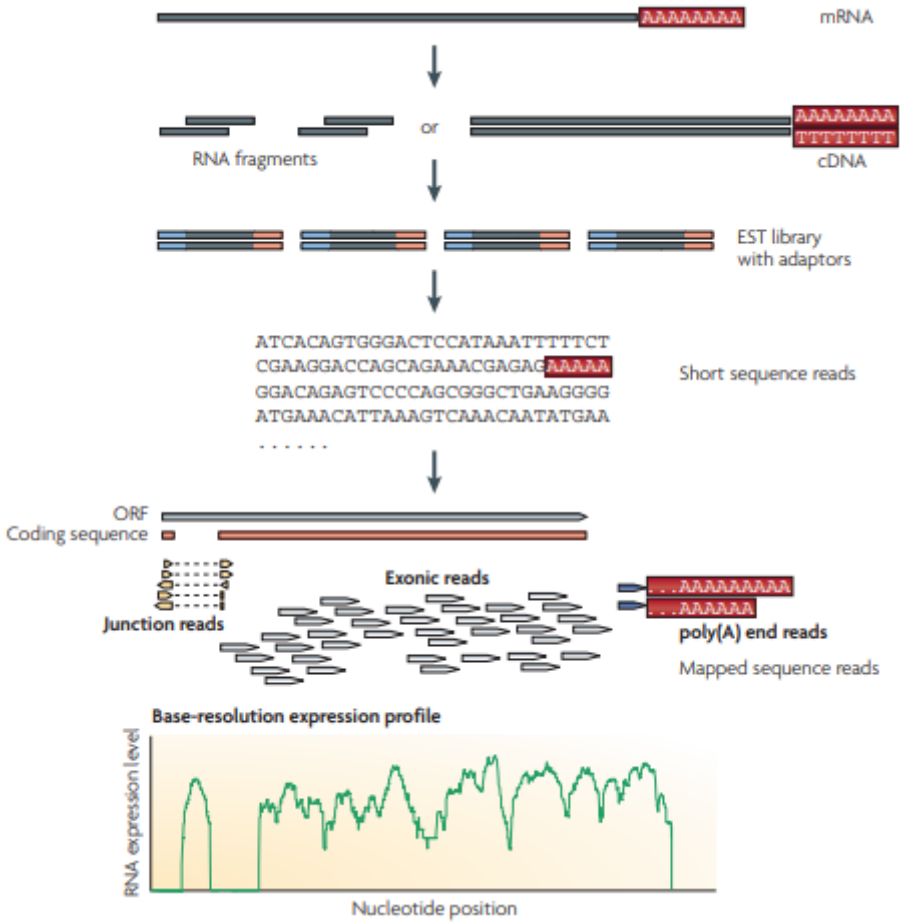


Figura 3: ***Procedimiento RNA-seq*** *(tomada de Z. Wang et al., 2009).*

En la figura se muestra que los ARN largos primero se convierten en una biblioteca de fragmentos de ADNc mediante fragmentación de ARN o posteriormente se añaden adaptadores de fragmentación de ADN (azul) a cada fragmento de cDNA y se obtiene una secuencia corta de cada ADNc. Los reads resultantes son alineados contra el genoma o transcriptoma de referencia, y se clasifican en tres tipos: reads exónicos, reads de unión y reads finales poli-A. Estos tres tipos se utilizan para generar un perfil de expresión para cada gen, como se ilustra en la parte inferior.

3. Elementos Transponibles

Los Elementos Transponibles (en inglés Transposable Elements, abreviados de aquí en adelante como TEs) son secuencias genéticas móviles presentes repetidamente en todos los genomas eucariontes. Los TEs usan diferentes

estrategias replicativas, que involucran intermediarios de ARN (retrotransposones) o intermediarios de ADN (transposones de ADN) (Siomi et al., 2011).

A. Los transposones de ADN (***Figura 4****, parte a, sección verde*) se mueven alrededor del genoma principalmente por un mecanismo de "cortar y pegar". Los transposones de ADN codifican una enzima transposasa que cataliza el movimiento de dichos elementos. Las transposasas ingresan al núcleo y se unen a los extremos de los transposones, en repeticiones invertidas/directas para escindirlos del genoma. Posteriormente, este elemento escindido se integra en un nuevo sitio en el genoma.
B. Los retrotransposones de repetición terminal larga (LTR) (***Figura 4***, *parte b, sección rosa*) se mueven alrededor del genoma mediante un mecanismo de "copiar y pegar" Los retrotransposones LTR contienen secuencias con repeticiones largas en ambos extremos y contienen la información para codificar integrasas y proteínas glicosaminoglicanas (GAG). Las proteínas GAG forman partículas similares a los virus, en las que se capturan los ARNm de los retrotransposones. Luego, en esta partícula tipo virus ocurre la transcripción inversa de los ARNm, generando intermediarios de ADN. Esta reacción tiene lugar en el citoplasma. Los intermediarios de ADN asociados con las integrasas emergen de las partículas, se importan al núcleo y se integran en el genoma posteriormente.
C. Los retrotransposones no LTR (***Figura 4****, parte c, sección azul*) también mueven alrededor del genoma mediante un mecanismo de "copiar y pegar", pero no tienen la repetición terminal larga y se pueden dividir en dos subtipos: elementos nucleares largos intercalados (*Long Interspersed Nuclear Elements*, LINE) y elementos nucleares cortos intercalados (*Short Interspersed Nuclear Elements*, SINE). Los LINE codifican dos proteínas, ORF1 y ORF2, que tienen un dominio de transcriptasa inversa. Las ORF1 y ORF2 forman complejos de ribonucleoproteína (RNP) con sus propios ARNm en el citoplasma. Los complejos RNP se transportan al núcleo y los ARNm se integran en nuevos sitios en el genoma mediante la transcripción inversa con los promotores del objetivo. Los SINE no codifican una transcriptasa inversa funcional y, por lo tanto, usan la enzima de LINE u otros elementos transponibles para su transposición.

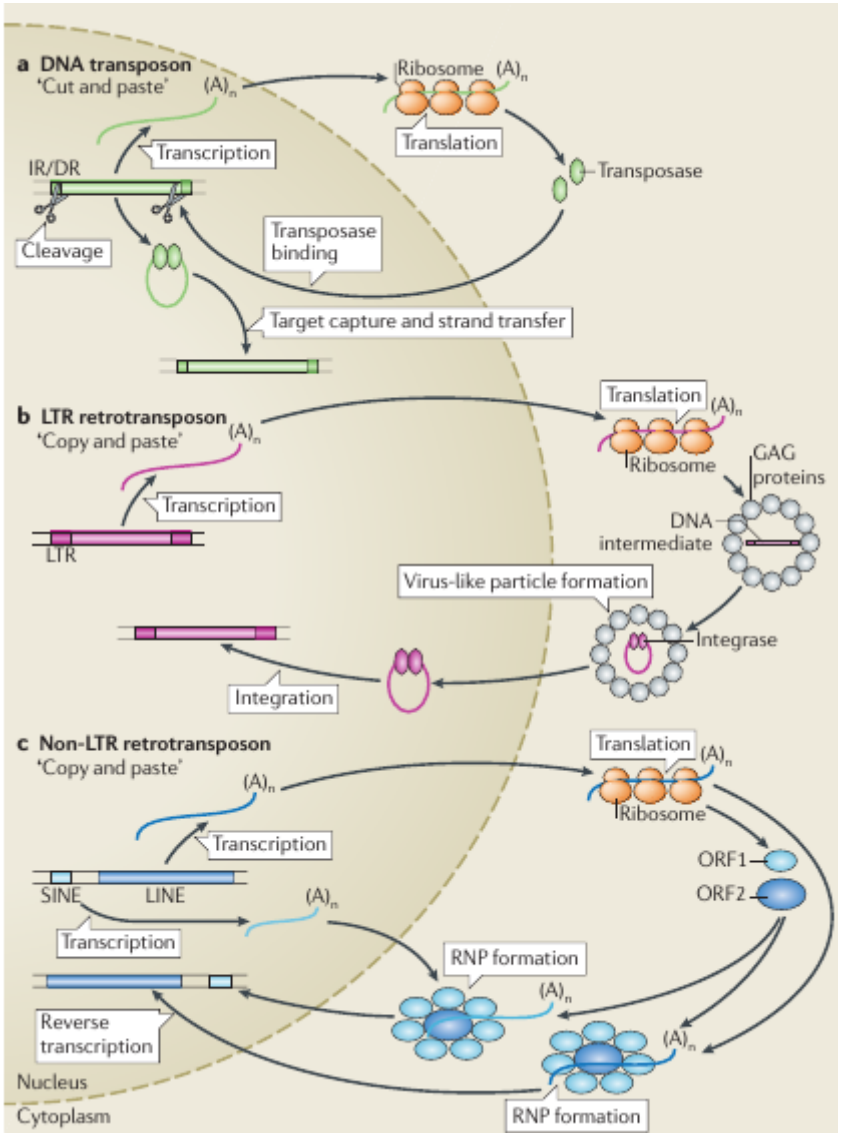


*Figura 4**: Mecanismo de Acción TEs (Siomi et al., 2011)***

En la figura se muestran los diferentes mecanismos de acción de los tres tipos de de TEs. Los transposones de ADN (barra verde) se mueven alrededor del genoma principalmente por un mecanismo de "cortar y pegar" catalizado por una enzima transposasa (óvalos verdes). Los retrotransposones de repetición terminal larga (LTR) (barra rosa) se mueven alrededor del genoma mediante un mecanismo de "copiar y pegar", los cuales codifican integrasas (óvalos rosados) y proteínas de glucosaminoglucano (GAG) (círculos grises) para su movimiento. Por último, los retrotransposones no LTR (barra azul) también se mueven mediante un mecanismo de "copiar y pegar", y se pueden dividir en dos subtipos, elementos nucleares largos intercalados (LINE; azul oscuro) y elementos nucleares cortos intercalados (SINE; azul claro). Los LINE codifican dos proteínas que forman un complejo para su transposición, ORF1 (óvalos azules claros) y ORF2 (óvalos azules oscuros), que tienen un dominio de transcriptasa inversa. Los SINE no codifican una transcriptasa inversa funcional y, por lo tanto, usan la enzima de LINE u otros elementos transponibles para su transposición. Los óvalos naranjas representan los ribosomas y "(A)n" indica la cola poli (A) formada durante la traducción.

Independiente del mecanismo de movilización, los TEs pueden tener varios impactos a nivel genético, según donde se inserten respecto a un gen. A nivel transcripcional, un TE insertado río arriba de un gen puede tener varias consecuencias, como: insertar secuencias promotoras e introducir un sitio de inicio de transcripción alternativo, interrumpir los elementos reguladores en cis ya existentes, introducir un nuevo elemento en cis, tal como un sitio de unión de factores de transcripción. Además, un TE insertado dentro de un intrón puede impulsar la transcripción antisentido e interferir potencialmente con la transcripción normal del gen. Finalmente, un TE puede servir como centro de nucleación para la formación de heterocromatina que potencialmente silencia la transcripción de genes adyacentes (Feschotte, 2008a).

A nivel post-transcripcional, un TE insertado en el extremo UTR 3' de un gen puede introducir un sitio de poliadenilación alternativo, un sitio de unión para un micro ARN o para una proteína de unión a ARN. Un TE insertado dentro de un intrón puede interferir con el patrón de *splicing* normal de un pre-ARNm, provocando varias formas de *splicing* alternativo (retención de intrón, omisión de exón, etc.) (Feschotte, 2008a).

Un TE que contenga sitios de *splicing* críptico, insertado dentro de un intrón, se puede incorporar ("*exonizar*") como un exón alternativo. Esto puede resultar en la traducción de una nueva isoforma proteica, o en la desestabilización o degradación del ARNm a través de la vía de descomposición mediada sin sentido (NMD), especialmente si el TE exonizado introduce un codón de stop prematuro (Feschotte, 2008b).

Adicionalmente al impacto que puedan tener en genes, los transposones han ido a lo largo del tiempo poblando los genomas. En humanos se estima que alrededor de un 50% del genoma está ocupado por transposones, y alrededor de un 80% en maíz (Hoen et al., 2015). Por este motivo, estudiar la expresión de los TE a través de RNA-seq se hace difícil, ya que por su naturaleza repetitiva, están presentes en muchas copias dentro de un genoma y, por ende, los reads originados de TEs mapearán en múltiples lugares. Los enfoques actuales que abordan los reads con

multi-mapeo se centran en la cuantificación de la expresión y no en encontrar el origen de la expresión.

Abordar el origen genómico de los TE expresados podría ayudar aún más a comprender el papel que los TE podrían tener en la célula (Valdebenito-Maturana & Riadi, 2018). Mediante el uso de parámetros predeterminados, los alineadores más comunes de RNA-seq, como Bowtie 2 o STAR, entregan una ubicación seleccionada al azar entre todas las asignaciones posibles de un read que mapea en múltiples ubicaciones.

Para resolver el problema del estudio de los TEs sin perder su ubicación real en el genoma con datos de RNA-seq, existen dos software desarrollados en los últimos años: TEcandidates que estima el origen de expresión de TEs al hacer un ensamble *de novo* de los reads de RNA-seq, y entrega TEs candidatos para ser considerados en el análisis de expresión posterior, junto con una versión modificada de un genoma de referencia adecuado para mapear los reads de RNA-seq. De este modo, se evita la ambigüedad en el mapeo de reads provenientes de TEs (Valdebenito-Maturana & Riadi, 2018). Por otro lado, SQuIRE cuantifica la expresión de TEs y realiza análisis de expresión diferencial en TEs y genes, distribuyendo los reads en todas las instancias de cada TE (Yang et al., 2019). Así entonces, se puede argumentar que tanto TEcandidates y SQuIRE permiten la estimación de expresión de TEs de manera loci-específica. Con esto, se pueden establecer asociaciones entre TEs y genes en base a su ubicación genómica.

3.1 Rol de los TEs en la Neurodegeneración

Existe evidencia que muestra que TEs están expresados en distintas enfermedades neurodegenerativas, como Enfermedad de Alzheimer, Demencia Frontotemporal y ELA (Ochoa Thomas et al., 2020)**.** En particular, hay evidencia que sugiere que los TEs pueden promover directamente la disfunción y/o pérdida neuronal. Esto debido a que su activación puede ser dañina, potencialmente interrumpiendo el proceso transcripcional y desencadenando una respuesta inmunológica. Con la movilización

de TEs, también pueden ocurrir mutagénesis de inserción somática y reordenamientos genómicos. Por lo tanto, numerosos sistemas celulares han evolucionado para suprimir la actividad de los TEs, pero estos mecanismos se superponen con los que regulan la estructura de la cromatina y la reparación del ADN. Sin embargo, el mecanismo de vigilancia de los TEs puede deteriorarse con el envejecimiento cerebral, lo que lleva a la activación de retrotransposones (Guo et al., 2018).

Hasta la fecha, ningún trabajo ha presentado evidencia que permita sugerir que la actividad de TEs es una causa o consecuencia de las enfermedades neurodegenerativas, y **es una pregunta que aún no está resuelta**. Más aún, en los trabajos que han llegado a hallazgos de TEs expresados en ELA, se han utilizado datos de RNA-seq con metodologías que tienen como consecuencia la pérdida del origen de expresión del TE. Por esto, tampoco se ha podido establecer un vínculo entre TEs expresados en la enfermedad y potenciales genes afectados.

Por un lado, entonces, tenemos que hay sets de datos de RNA-seq a lo largo de la progresión de la enfermedad en el modelo de ratón (en particular, para las etapas asintomática, temprana, media y tardía), en los cuales no hicieron estudio de TEs (mencionado anteriormente), y por otro lado, tenemos que sólo recientemente las herramientas para un análisis loci-específico de TEs se han publicado. Esto representa una nueva oportunidad para explotar esos datos y recopilar evidencia que permita elucidar si los TEs son causa o consecuencia de ELA. Considerando esto, se puede plantear la hipótesis descrita en la siguiente página.

Hipótesis

Los elementos transponibles tienen un efecto causal en la progresión de ELA en el modelo de ratón SOD1^{G93A}, mediante la interrupción loci-específica de la expresión de algunos genes.

Para testear esta hipótesis, se plantea lo siguiente en términos de objetivos:

Objetivo general

Analizar la expresión loci-específica de TEs y genes, a lo largo de la progresión de ELA en el modelo de ratón SOD1^{G93A}.

Objetivos específicos

1. Analizar la expresión loci-específica de TEs en las etapas asintomática, temprana, media y tardía, del modelo de ratón transgénico SOD1^{G93A}.
2. Asociar la expresión loci-específica de TEs con la expresión genes en las etapas asintomática, temprana, media y tardía, del modelo de ratón transgénico SOD1^{G93A}.
3. Analizar el potencial impacto de los genes afectados por TEs en la progresión de ELA en el modelo de ratón transgénico SOD1^{G93A}.

Materiales y Métodos

Materiales

Software:

- **BEDtools 2.27.0** (Quinlan & Hall, 2010). Software que permite la comparación, manipulación y anotación de características genómicas en formato BED y GFF. Además, soporta la comparación de alineamientos de secuencias genómicas en formato BAM.

- **BioPerl** (Stajich et al., 2002). Módulo de Perl que permite el manejo y manipulación de información biológica.

- **Bowtie2 v2.3** (Langmead & Salzberg, 2012). Software que permite el alineamiento de reads de secuenciación contra un genoma de referencia.

- **SAMtools v1.4.1** (Li et al., 2009). Librería y paquete de software que permite el parseo y manipulación de alineamientos en formato SAM/BAM.

- **Trinity v2.4.0** (Haas et al., 2013). Combinación de tres softwares independientes (*Inchworm, Chrysalis and Butterfly*) para el procesamiento de grandes cantidades de datos de RNA-seq y, además, provee de un método eficiente y robusto para la reconstrucción *de novo* de transcriptomas con datos de RNA-seq.

- **SRAtoolkit** (Leinonen et al., 2011). Permite la lectura de archivos de secuenciación de la base de datos SRA y, además, escribir archivos en formato SRA.

- **STAR 2.5.3a** (Dobin et al., 2013). Software de alineamiento rápido de datos de RNA-seq.

- **Stringtie 1.3.3b** (Pertea et al., 2015). Software que permite el ensamble de sets de datos para formar transcritos.

- **R 3.4.1** (R Development Core Team, 2006). Lenguaje utilizado para la programación estadística y gráficos.

- **Python 2.7**. Lenguaje de programación interpretado.

- **DESeq2 1.16.1** (Love et al., 2014). Módulo de BioConductor que permite un análisis diferencial de conteos de datos.

- **HTseq 0.11.1 (*Count*)** (Anders et al., 2015). Librería de Python que permite el procesamiento de datos de RNA-seq para análisis de expresión diferencial mediante el conteo de traslapes de reads con genes.

- **TEcandidates** (Valdebenito-Maturana & Riadi, 2018). Pipeline que permite incluir TEs en análisis de expresión sin perder su ubicación en el genoma (locus-específico).

- **SQuIRE** (Yang et al., 2019). Pipeline que permite un análisis de datos de RNA-seq cuantitativo y locus-específico de la expresión de TEs.

Hardware:

- 2 CPUs Intel Xeon E7-8867 v3 @ 2.50GHz, con 64 cores cada una, 128 procesadores en total.

- 256GB RAM.

- 40TB HD en 6 discos SATA de 8TB.

- Sistema operativo: GNU/Linux Debian para una máquina x86_64 con kernel 4.19.0-6-amd64, versión #1 SMP Debian 4.19.67-2+deb10u2 (2019-11-11).

Datos:

- Set de Datos de RNA-seq del modelo de ratón $SOD1^{G93A}$ (Identificador en Base de Datos GEO GSE43879) (Phatnani *et al.*, 2013), los cuales contienen perfiles de:

    - 16 muestras de médula espinal en las 4 diferentes etapas de la enfermedad (asintomática, temprana, media y tardía), *Wild Type (WT)* y $SOD1^{G93A}$ mutante.

- Genoma de Referencia *Mus musculus GRCm38.p6.*

Métodos

La metodología que se utilizó en este trabajo se resume en la ***Figura 5***, mostrada a en la siguiente pagina. Brevemente, en este proyecto se utilizaron principalmente dos softwares: TEcandidates (Valdebenito-Maturana & Riadi, 2018) y SQuIRE (Yang et al., 2019) con datos de RNA-seq de las distintas fases de la enfermedad de ELA del modelo de ratón $SOD1^{G93A}$.

El objetivo de TEcandidates era predecir la ubicación de los TEs que se están expresando a partir de datos de RNA-seq. Con esto, TEs se pueden incluir

posteriormente en análisis de expresión, y se puede estudiar su asociación con otros componentes en el genoma en base a su ubicación. TEcandidates tiene como dependencias: BEDtools (Quinlan & Hall, 2010), BioPerl (Stajich et al., 2002), Bowtie2 (Langmead & Salzberg, 2012) y Trinity (Haas et al., 2013). Este pipeline recibe tres inputs principales: archivos de reads de RNA-seq, genoma de referencia y anotación de TEs. Primero, TEcandidates realiza un alineamiento de los reads contra el genoma de referencia con el fin de restringir los reads a TEs exclusivamente, lo cual ayuda a reducir la cantidad de reads que se utilizan en los pasos siguientes. Posteriormente, estos reads se le entregan a Trinity para realizar un ensamble de transcriptoma *de novo*. Con los transcritos reconstruidos se hace un mapeo con Bowtie 2 contra el genoma de referencia, y a continuación se evalúan las intersecciones entre posiciones de alineamiento de los transcritos *de novo* y los TEs anotados utilizando BEDtools. Finalmente, los TEs candidatos se escriben en un archivo con formato GFF3 y las posiciones de los TEs anotados que no fueron seleccionados son enmascarados en la secuencia del genoma con letras ‘X’. Este pipeline entrega como output tres archivos: el archivo que contiene los TEs candidatos, otro que contiene los TEs que no fueron seleccionados y uno que contiene el genoma con los TEs que no fueron seleccionados enmascarados.

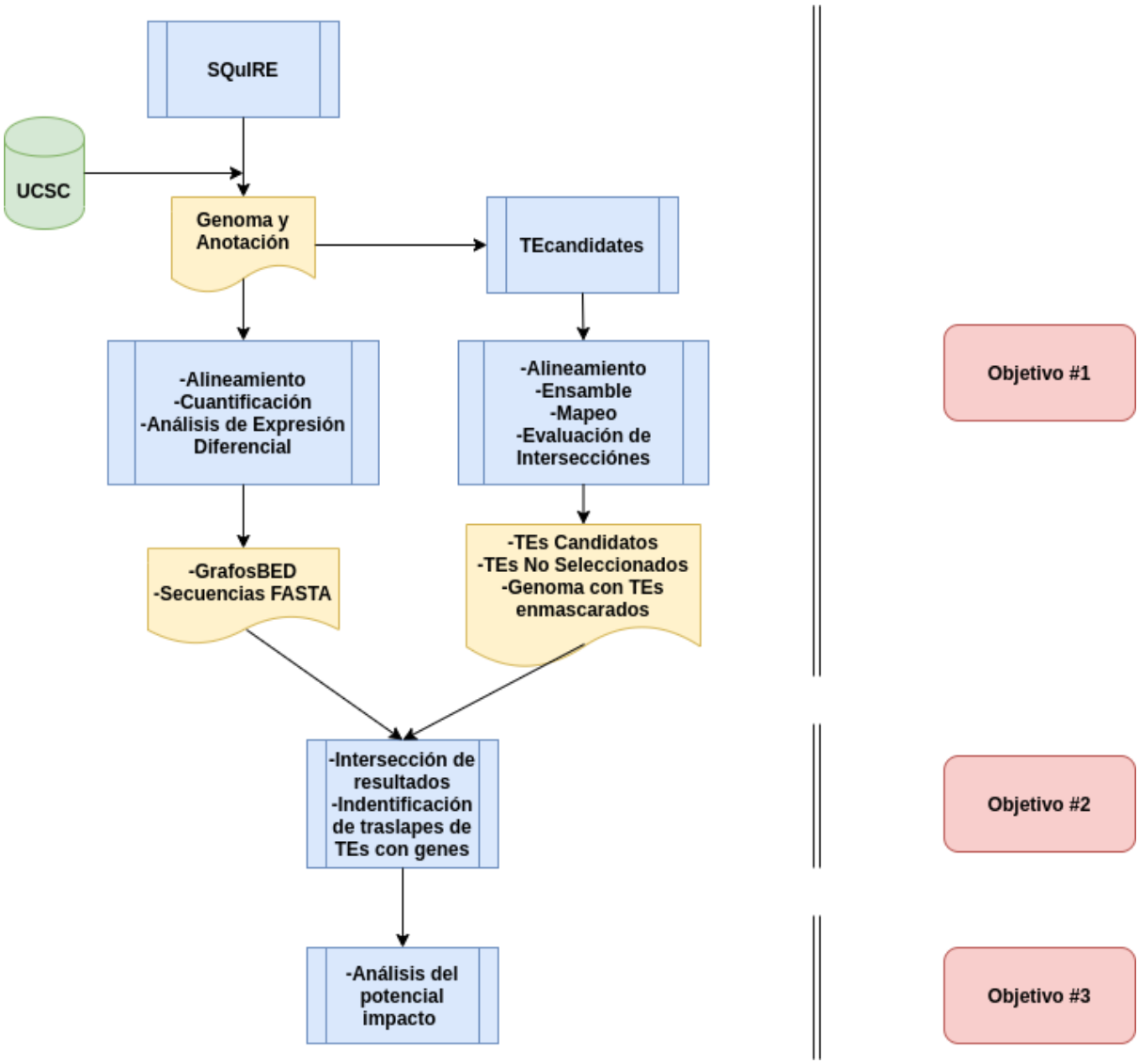


Figura 5: Diagrama de flujo general de la metodología.

Con el fin de obtener resultados que potencialmente correspondan a TEs realmente expresados, también se utilizó SQuIRE, y se evaluó la intersección de TEs predichos por esta herramienta y TEcandidates. SQuIRE es set de herramientas con el cual se estiman los niveles de expresión de TEs a partir de datos de RNA-seq, sin perder su ubicación en el genoma. Similar a TEcandidates, SQuIRE tiene como dependencias: STAR (Dobin et al., 2013), BEDtools (Quinlan & Hall, 2010), SAMtools (Li et al., 2009), StringTie (Pertea et al., 2015), DESeq2 (Love et al., 2014), R (R Development Core Team, 2006) y Python 2.7.

Objetivo #1

SQuIRE posee diversas herramientas para el análisis de TEs que son utilizadas a lo largo del pipeline, el cual se organiza en cuatro etapas: Preparación, Cuantificación, Análisis y Seguimiento (***Figura 6***).

En la etapa de preparación, la herramienta *Fetch* descarga los archivos de anotación requeridos de las especies con genomas ensamblados disponibles en *University of California Santa Cruz (UCSC) Genome Browser.* Estos archivos de anotación incluyen información de RefSeq de los genes en formatos BED y GTF, además de la información de los TE de RepeatMasker en un formato estándar. Esta herramienta también realiza la indexación de los archivos de cromosomas en formato FASTA para el software de alineamiento STAR (a diferencia de TEcandidates que utiliza Bowtie). *Clean* hace un reformateo de la información de anotación de los TEs de RepeatMasker y lo transforma en formato BED.

La etapa de Cuantificación incluye dos pasos: el paso de alineamiento (*Map*) y el paso de cuantificación de RNA-seq (*Count*). *Map* hace un alineamiento de los datos de RNA-seq usando STAR, dando como resultado un archivo BAM. Luego, *Count* cuantifica la expresión de los TEs usando un algoritmo específico de SQuIRE que incorpora los reads con mapeo único y con multimapeo. Adicionalmente, *Count* cuantifica la expresión de los genes anotados en RefSeq con el software ensamblador de transcritos StringTie.

Posteriormente, en la etapa de Análisis, *Call* realiza un análisis de expresión diferencial para los TEs y genes de RefSeq con el paquete de Bioconductor DESeq2.

Para poder tener predicciones comparables, TEcandidates se utilizó con los archivos de genoma y respectiva anotación descargados por “SQuIRE *Fetch*” en la etapa de *Preparación*. En el caso de TEcandidates, la cuantificación de reads por gen y TEs se realiza con HTseq-count. Los conteos se utilizan como entrada en DESeq2, para hacer análisis de expresión diferencial.

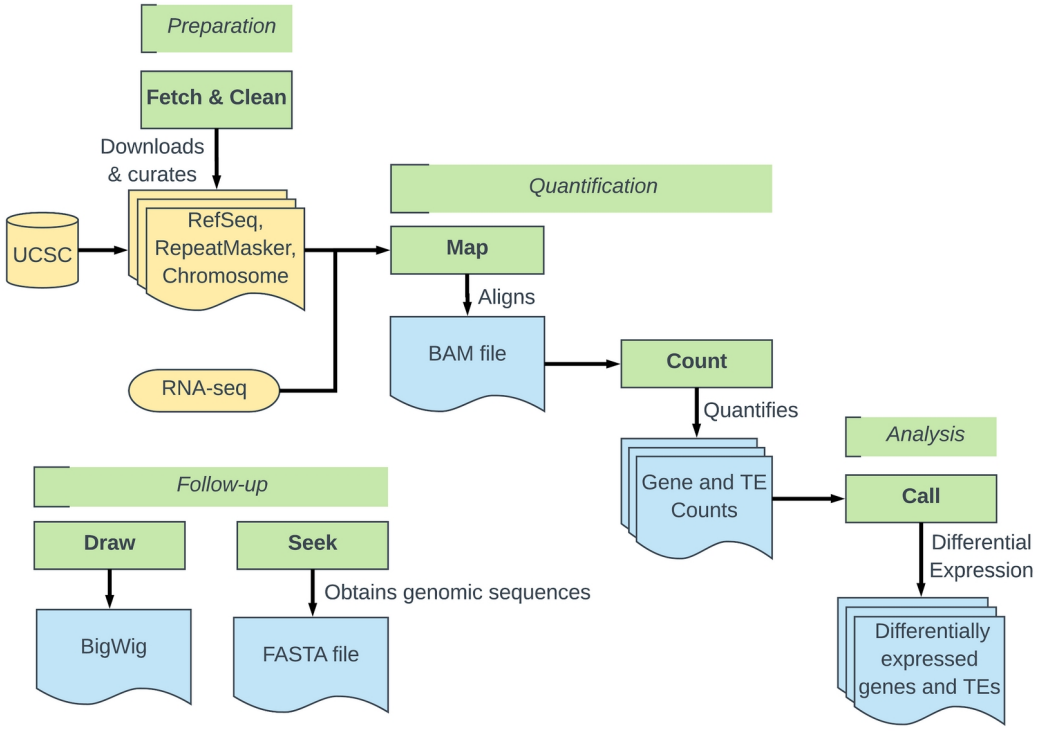


*Figura 6**: Esquema del pipeline de SQuIRE (Yang et al., 2019)***

Los recuadros verdes con texto en negrita representan las herramientas de SQuIRE con la etapa del pipeline (Preparación, Cuantificación, Análisis y Seguimiento) arriba en recuadros verdes con texto en cursiva. Las figuras amarillas representan los inputs requeridos y las azules, los outputs generados.

Para los análisis de expresión diferencial, se compararon las muestras en cada punto de tiempo entre la condición mutante y *wild type*. Como criterio para definir TEs con expresión diferencial se utilizó el Fold Change (FC) en escala logarítmica en base 2 ($\log_2$(FC)), y el valor P-ajustado (también conocido como *False Discovery Rate*, FDR). *Fold Change* (FC) es una medida que describe cuánto una cantidad en términos relativos cambia desde un valor inicial hasta un valor final, el cual se calcula como la relación entre el valor final y el valor inicial. En términos de análisis de expresión diferencial, de manera simplificada, se calcula como cantidad de reads en condición B (usualmente "mutante") dividido por la cantidad de reads en condición A (usualmente "*wild type*"):

$$FC = reads\ en\ condición / ❑ reads\ en\ condición\ wild\ type$$

Posterior al cálculo del FC, éste se pasa a escala logarítmica en base 2, ya que los genes que están reprimidos tendrán valores entre 0 y 1, y por otro lado, los que están sobre-expresados estarán entre 1 e infinito, entonces, con esta nueva escala, los que estén sobre-expresados estarán entre 0 e infinito positivo y por el contrario, los que estén reprimidos tendrán valores entre 0 e infinito negativo.

Así, se utilizará como punto de corte $\log_2$(FC) mayor a 2, o menor a -2, ya que esto sería indicativo de genes que están al menos 2 veces más expresados en una condición que en otra. De esta manera, se espera que los resultados no sean falsos positivos. En relación a esto, adicionalmente se utiliza el P-value ajustado, también conocido como FDR (Tasa de Descubrimiento Falso, en inglés *False Discovery Rate*), menor a 0.05, el estándar usado en RNA-seq actualmente. Con este valor de FDR se estimaría que dentro de los genes seleccionados con el $\log_2$(FC) recién mencionado, habrá a lo más 5% de falsos positivos, ya que este parámetro da una idea del número de falsos positivos que se puede esperar si el experimento se realizará un número infinito de veces (Aubert et al., 2004).

Por último, en la etapa de Seguimiento, con el fin de permitir al usuario visualizar de mejor manera los alineamientos de los TEs de interés, *Draw* crea grafos BED para cada muestra y *Seek* recupera las secuencias genómicas proporcionadas por el usuario en formato FASTA.

Objetivo #2

La siguiente etapa se realizó con los TEs identificados en las distintas fases de la enfermedad de ELA del modelo de ratón $SOD1^{G93A}$. Se hizo una asociación loci-específica de TEs identificados anteriormente con genes, luego de realizar un análisis de expresión diferencial de genes y TEs por separado. Los TEs con valores de $\log_2$(FC) mayor o igual a 2 y FDR menor o igual a 0.05 se utilizan para identificar traslapes con genes por medio del software BEDtools, el cual recibirá como input un archivo que contenga las coordenadas de las posiciones de TEs seleccionados en la etapa anterior y otro con las coordenadas de genes.

Objetivo #3

Finalmente, para analizar el potencial impacto de los genes afectados por TEs en la progresión de ELA en el modelo de ratón $SOD1^{G93A}$, se toman los TEs sobre-expresados en la condición mutante que tuvieron traslapes con genes, se comparan sus valores de $\log_2$(FC) y se especula sobre posibles relaciones entre ellos a lo largo del desarrollo de la enfermedad. Además, se hizo una revisión bibliográfica, con el fin de realizar un análisis funcional de genes, es decir, identificar función del gen y las vías metabólicas en las que está implicado descritas en literatura, mediante NCBI-Gene con el módulo *GeneRIF*, el cual vincula la información que existe dentro de la literatura con funciones de genes.

Resultados

Análisis de Calidad de Reads

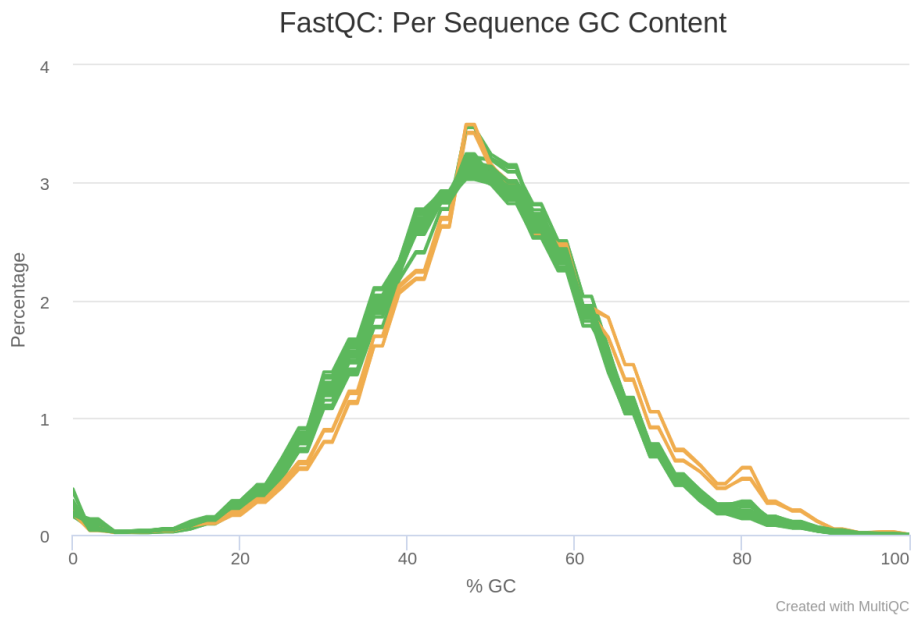
Un proceso importante que se debe realizar previo a la utilización de cualquier set de reads obtenidos por medio de una secuenciación, es el análisis de calidad de los mismos. Los reads fueron descargados mediante la herramienta de SRA-toolkit llamada "*fastq-dump*" especificando la opción "*--split-files*" ya que se tratan de archivos de *reads* del tipo *paired-end.* Posterior a la descarga de los archivos, se realizó un análisis de la calidad de los *reads* mediante la herramienta FastQC y, posteriormente, MultiQC (Ewels et al., 2016) para obtener una sumarización de todos los análisis realizados.

Las principales métricas que se considerarán para evaluar la calidad de los reads son: Contenido GC, Contenido de adaptadores, Contenido de Ns, Calidad por base, y Calidad promedio.

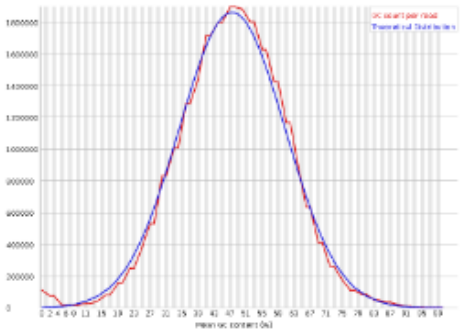
Contenido GC

La importancia del análisis de contenido de GC que se tiene a lo largo de los *reads* radica en que es un indicador de la cobertura que se tiene, en este caso, a lo largo del transcriptoma estudiado, es decir, un alto contenido GC indicaría una alta profundidad de secuenciación (Y.-C. Chen et al., 2013). En ***Figura 7.a*** se ve que el promedio del contenido de GC de los reads sigue una distribución normal, lo cual es lo típico en los análisis de calidad de este tipo En el gráfico se pueden ver líneas con dos colores, por un lado las líneas de color verde son los archivos que el *software* MultiQC definió como correctas **(*Figura 7.b*)**, y por el contrario, las de color naranjo son las que se definen con algún tipo de *warning* ***(Figura 7.c)***.

a.



b.



c.

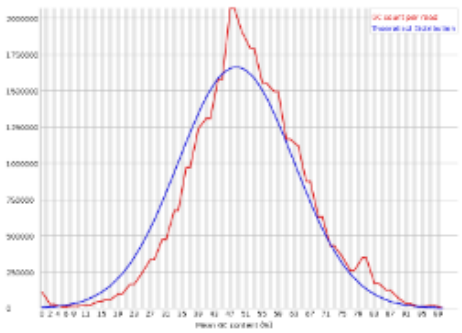


Figura 7: ***Contenido GC - a.*** *Contenido de GC promedio en el conjunto de reads analizados mediante el software MultiQC.* ***b.*** *Ejemplo de distribución de contenido GC, en donde la del set de datos es parecida a la esperada (línea azul) (parte* ***a,*** *líneas verdes.).* ***c.*** *Ejemplo de distribución distinta a la esperada por el software FastQC (parte* ***a,*** *líneas naranjas) al momento de compararla con las distribuciones de los demás reads.*

Se puede ver, además, que el *peak* está cerca del 50% de contenido GC de los reads analizados, lo cual se asemeja a lo que presentan los exones del genoma de *Mus musculus*, ya que éste presenta cerca de un 50,7% luego de un análisis realizado por medio de las herramientas Bedtools y AWK.

Contenido de Ns

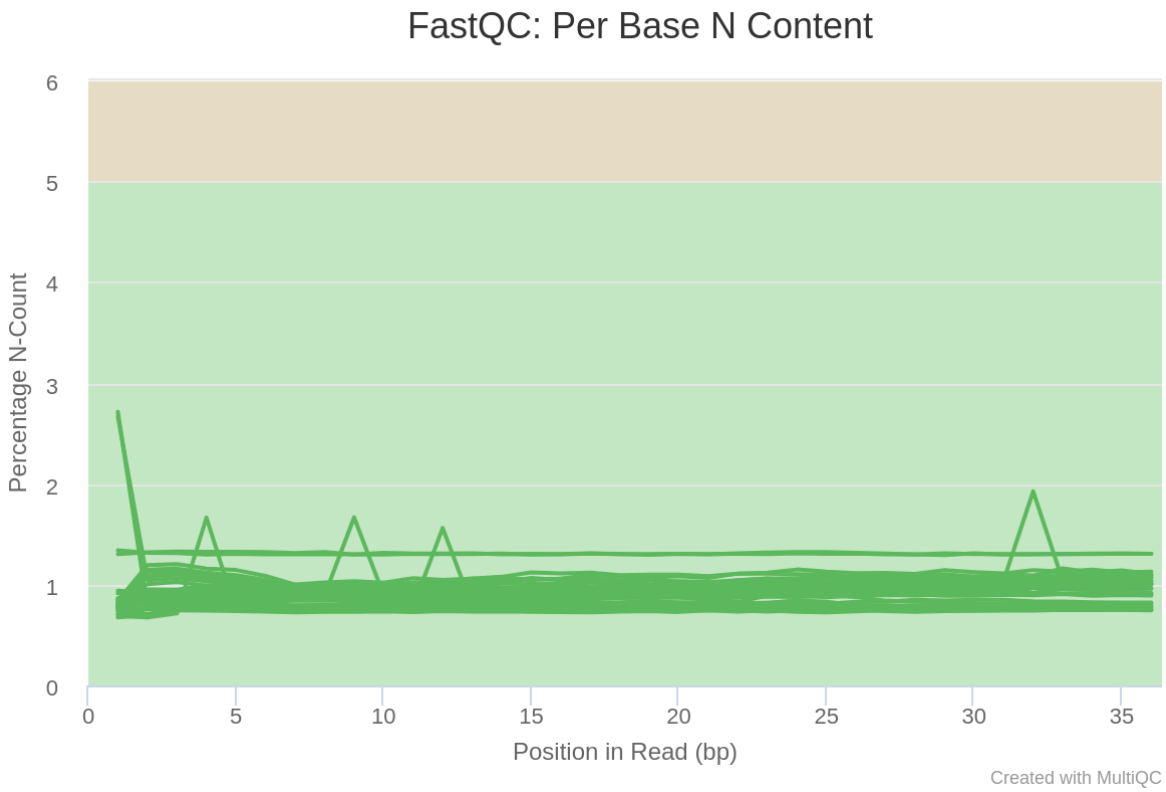


Figura 8: ***Contenido de Ns por posición -*** *En el eje Y se muestra el porcentaje de Ns que se encontró en cada posición de los reads estudiados, y en el eje X la posición de cada base a lo largo del read. Cada línea corresponde a cada uno de los archivos de reads.*

El gráfico de la ***Figura 8*** muestra el porcentaje de *base calls* en cada posición que fueron asignadas con una N. Se puede ver que prácticamente a lo largo de todos los reads el porcentaje rodea el 1 %, lo cual sugiere que, por cada archivo de *reads*, habrían cerca de 300.000 bases desconocidas al tratarse de librerías de ~25.000.000 de *reads*.

Contenido de adaptadores

El *software* MultiQC reporta que no se encontró contaminación por adaptadores en las librerías de *reads* entregadas, lo cual hace que en el proceso de *trimming* no sea necesario enfocarse en este tipo de secuencias.

Calidad Principal de Secuencia

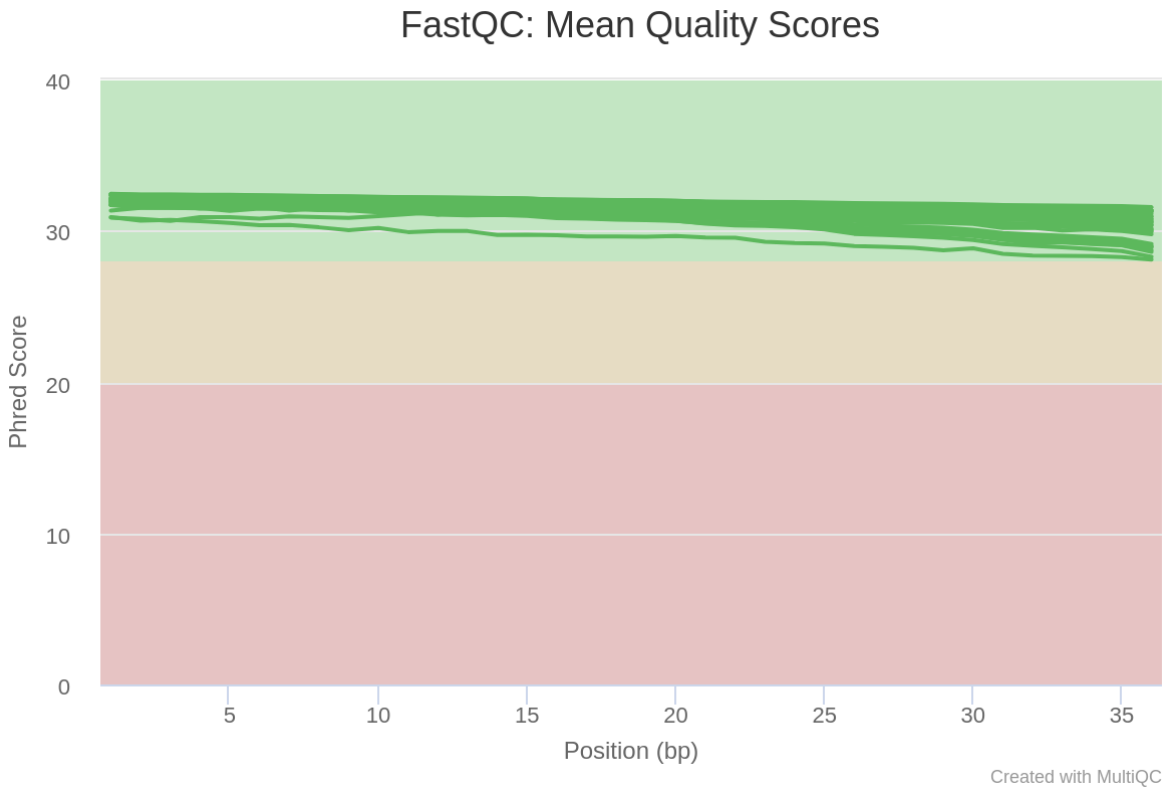


Figura 9: ***Calidad por posición -*** *En el eje X se muestra la posición a lo largo de los reads (36 pb en total), y en el eje Y se pueden ver los puntajes de calidad promedios por base. Cada línea corresponde a cada uno de los archivos de reads. .*

Al igual que en el gráfico siguiente ***(Figura 10)***, se muestra la calidad de la secuencia ***(Figura 9)***, pero en este caso se ve el promedio por cada posición del read. Cabe mencionar que todos los reads tienen un largo de 36 pb y la calidad promedio por posición es cerca de 32, lo que sugiere que hay ~0.1% de que el *base-call* esté incorrecto.

Puntajes de Calidad por Secuencia

a. **b.**

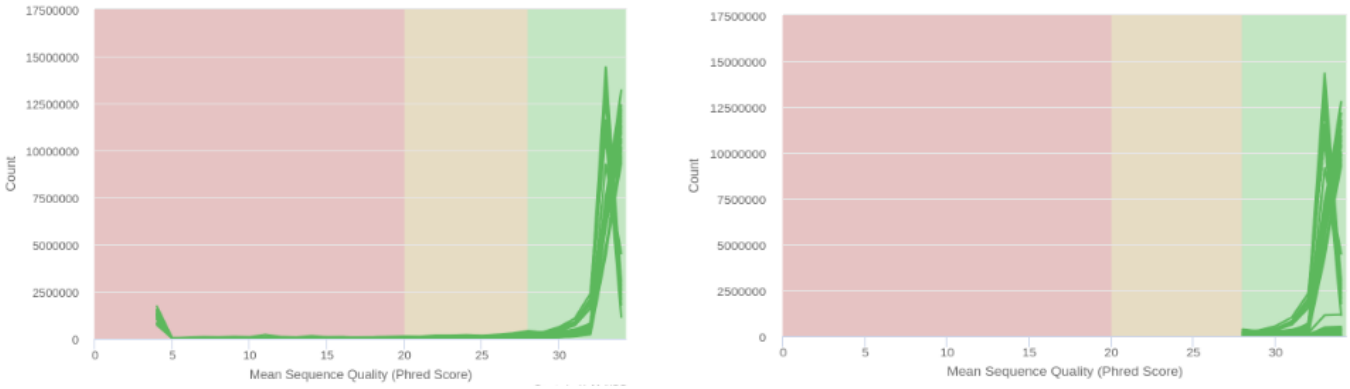


Figura 10: ***Calidad promedio secuencia -*** *En en eje X se muestra la calidad promedio por secuencia en una escala de puntajes Phred 33, y en el eje Y se muestra la cantidad total de secuencias analizadas. -* ***a.*** *Distribución de calidades antes del proceso de trimming.* ***b.*** *Distribución de calidades después del proceso de trimming.*

El gráfico de la ***Figura 10.a*** muestra el número de reads junto con los puntajes promedio de calidad, además se puede ver si existe algún subset de reads que tengan una baja calidad. En este caso se tiene un primer *peak* cercano a una calidad 5, lo que podría estar relacionado con el punto anterior, tratándose de las secuencias con un alto contenido de Ns. Por otro lado, se tiene un segundo *peak* cercano al valor de calidad 32, lo cual es bueno, ya que en este tipo de análisis se considera un buen valor de calidad valores sobre 28 aproximadamente. Esto debido a que desde este valor se estima que la probabilidad de *base-call* incorrecto es de ~0.1%. A raíz del *peak* cercano a 5, se decide realizar un proceso de *trimming* para una posterior comparación entre ambos sets de datos y decidir si es factible perder toda esa información ***(Figura 10.b)***. Esta etapa se realizó mediante el *software* Trimmomatic (versión 0.39) (Bolger et al., 2014) con la opción “Average Quality”, la cual corta todos los *reads* que estén bajo una calidad dada, que en este caso fue 28. Posterior a esto, se realizó un alineamiento de ambos sets de datos contra el

genoma de *Mus musculus* (versión GRCm38) con el *software* HISAT2 (Kim et al., 2019) ***(Figura 11)***.

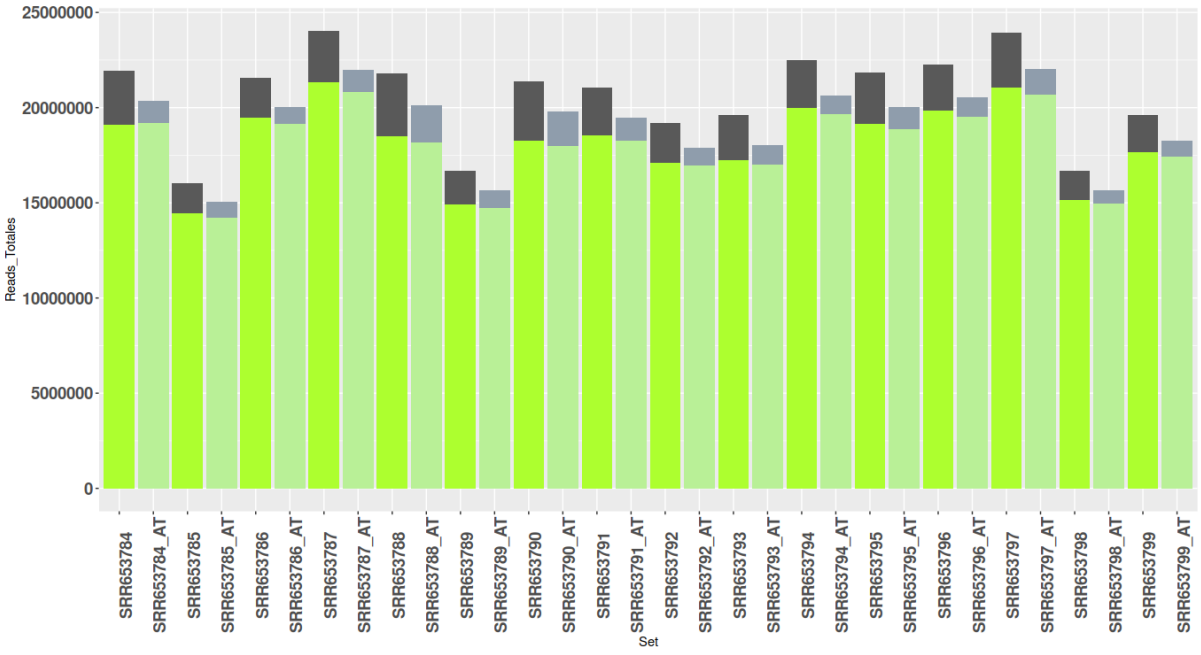


Figura 11: ***Comparación sets de datos -*** *El gráfico muestra en el eje x el nombre del set de reads analizado, siendo los sets filtrados los que en sus nombres contienen "_AT" (After Trimming). En el eje Y se muestra la cantidad de reads que contiene cada set. En color verde se muestra la cantidad de reads que alinearon contra el genoma de Mus musculus. En color gris se muestra la diferencia que existe entre cantidad total de reads de cada set y los que alinearon.*

Si bien, en cuanto a proporción no es mucha la información que se está perdiendo, en el set de datos filtrado se estarían cortando o eliminando *reads* que sí alinearon contra el genoma de referencia, lo que significa que se estaría eliminando información que posteriormente podría ser valiosa para los análisis. Por otro lado, también se está corriendo el riesgo de mantener información "basura" que podrían influir en los siguientes análisis pudiendo llevar a conclusiones equivocadas. Considerando que al hacer la cuantificación de reads por gen o TE, las herramientas pueden eliminar *reads* que si alinearon se decidió seguir trabajando con el set de datos sin filtrar.

Objetivo #1: *Análisis de expresión loci-específica de TEs en las diferentes etapas de la enfermedad.*

Análisis de expresión de TEs

SQuIRE

Luego de la ejecución del *pipeline* completo de SQuIRE, se han categorizado los TEs encontrados en las diferentes etapas de la enfermedad según sus valores de FDR y Fold Change como Reprimidos, No Significantes o Sobre-Expresados (ver Metodología) en la condición de *Mutante* con respecto a su condición *Wild Type*, tal como se presenta en la siguiente tabla:

Semanas	Reprimidos	No significantes	Sobre-expresados
4	23	22.670	28
8	52	26.156	281
12	43	22.685	77
17	52	21.197	111

Tabla 2: ***Resumen de resultados SQuIRE:*** *Cantidad de TEs encontrados en cada categoría luego del análisis de expresión diferencial mediante edgeR.*

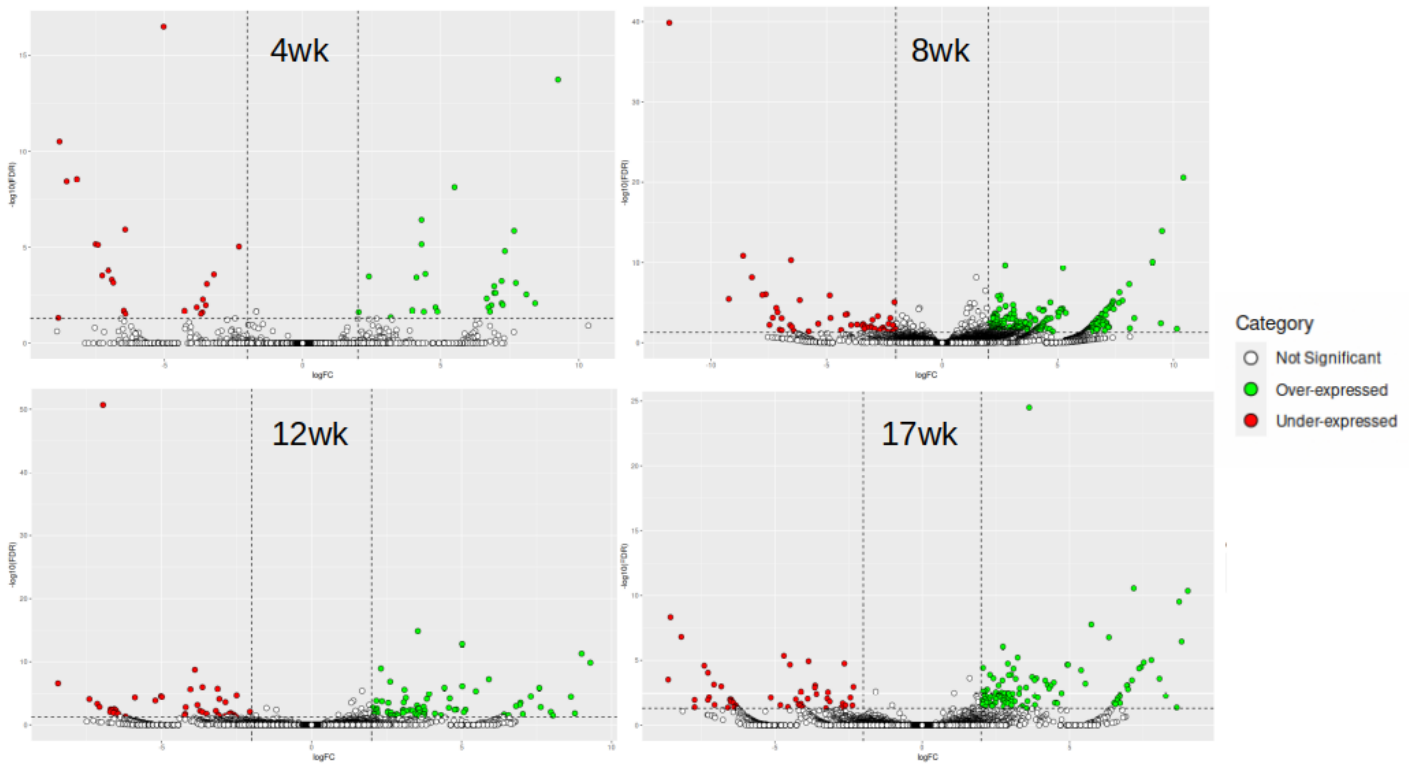


Figura 12: ***Resultados Pipeline SQuIRE:*** *Se muestran los resultados del análisis de expresión diferencial con los resultados del pipeline SQuIRE representados en cuatro volcano plots separados por etapas de la enfermedad. En verde se representan los TEs que están sobre-expresados en la condición mutante con respecto a su condición Wild Type; en rojo se representan los TEs que están siendo reprimidos y en blanco los que tienen valores no significativos para este tipo de análisis. Para la categorización de dichos TEs se utilizaron como puntos de corte un FDR menor a 0.05 (eje Y) y un* log_2 *(Fold Change) menor o igual a -2 y mayor o igual a 2 (eje X).*

TEcandidates

TEcandidates es un *pipeline* que entrega información de manera cualitativa acerca de la expresión de los TEs a partir de datos de RNA-seq, por lo tanto los resultados obtenidos se resumieron en un gráfico de barras **(Figura 13).** Cabe mencionar que en esta etapa se utilizó la última versión de TEcandidates (2.0) aún no publicada.

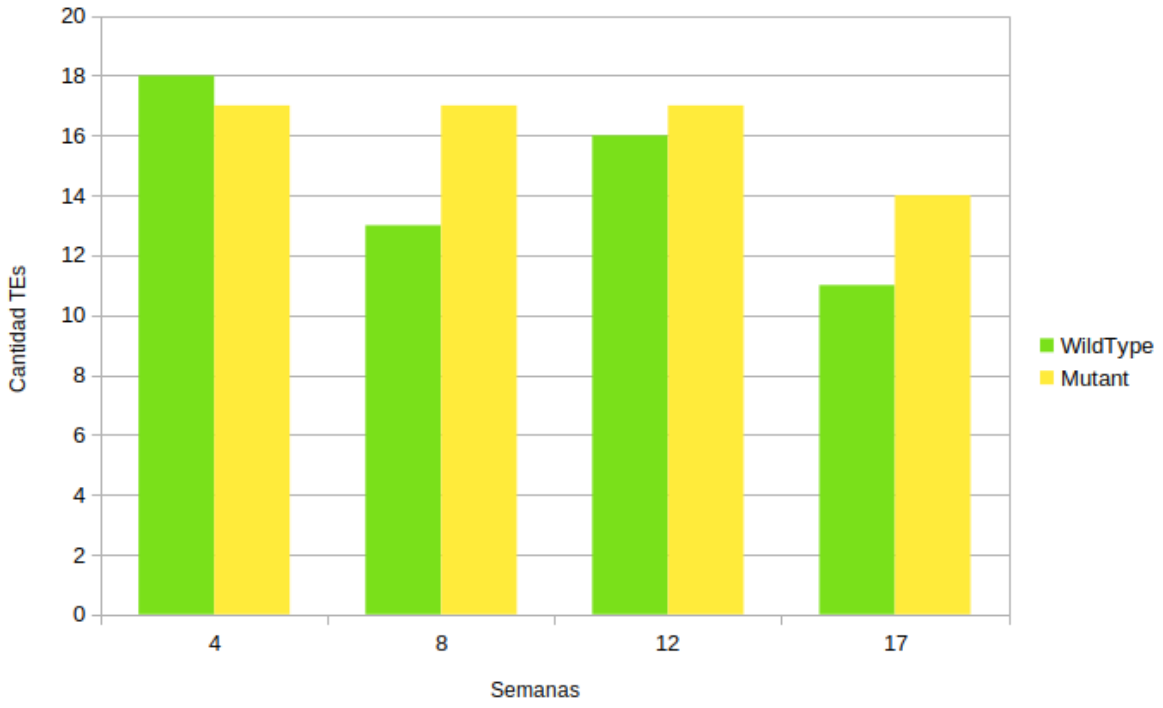


Figura 13: ***Resultados Pipeline TEcandidates:*** *En el eje X se muestran los set de datos utilizados separados por semanas de la enfermedad. En el eje Y se muestra la cantidad de TEs que fueron predichos como origen de expresión por TEcandidates. En verde se presentan los TEs predichos en la condición Wild Type, y en amarillo se presentan los TEs predichoss en la condición mutante.*

La cantidad de TEs encontrados por TEcandidates fue mucho menor a los encontrados por el *pipeline* de SQuIRE por lo que posterior a la ejecución de ambos pipelines (SQuIRE **(*Figura 12*)** y TEcandidates **(*Figura 13*)**) se realizó una intersección de resultados con el fin de tener mayor seguridad en los resultados obtenidos y con estos continuar con los análisis posteriores ***(Figuras 14 - 17)***.

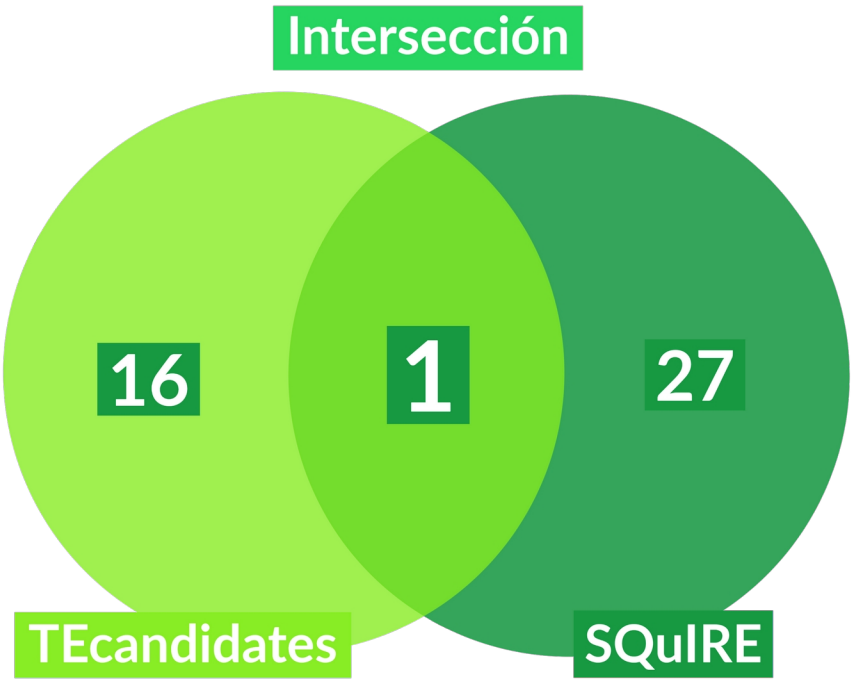


Figura 14: ***Cantidad TEs encontrados por cada pipeline sobre-expresándose en la condición mutante en la semana 4.***

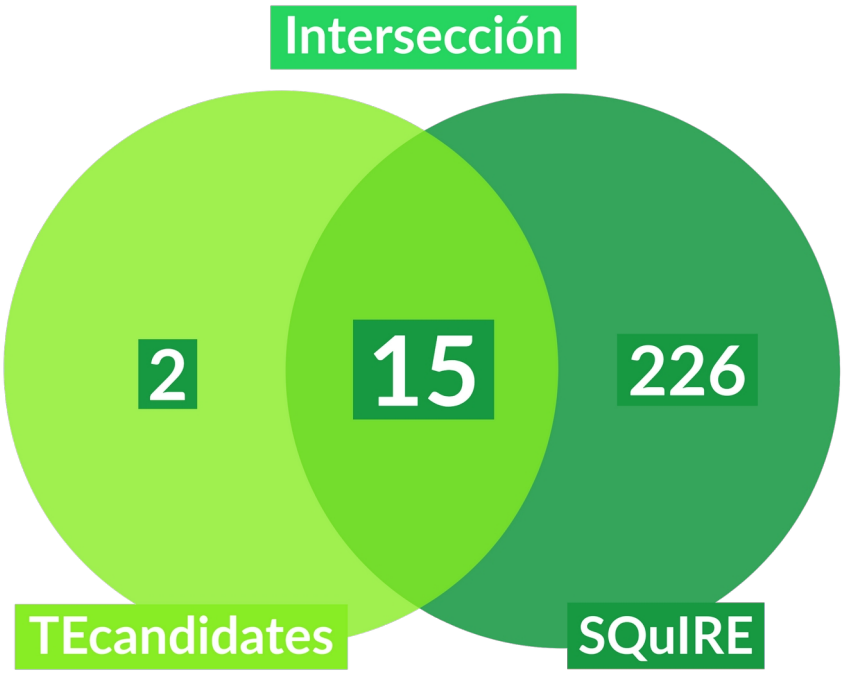


Figura 15: ***Cantidad TEs encontrados por cada pipeline sobre-expresándose en la condición mutante en la semana 8.***

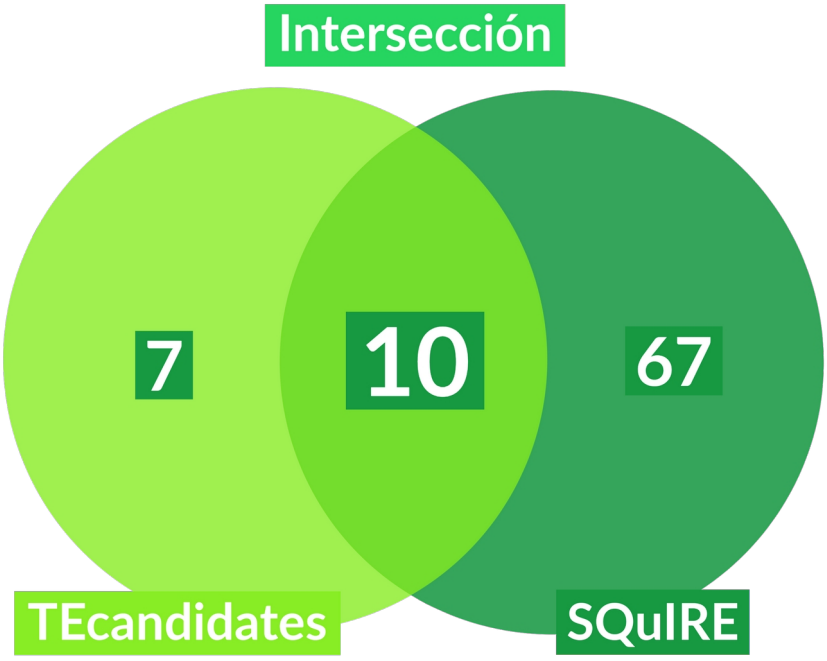


Figura 16: ***Cantidad TEs encontrados por cada pipeline sobre-expresándose en la condición mutante en la semana 12.***

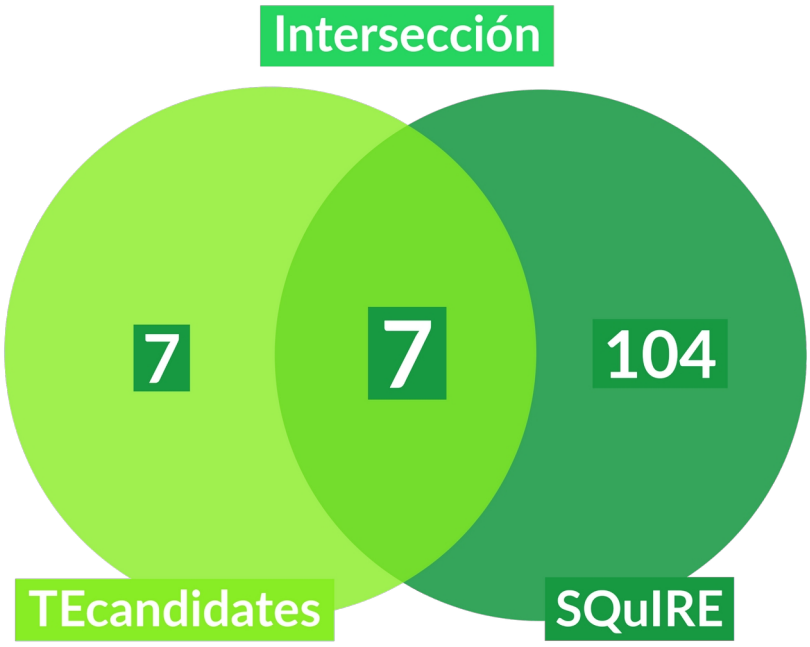


Figura 17: ***Cantidad TEs encontrados por cada pipeline sobre-expresándose en la condición mutante en la semana 17.***

Objetivo #2: *Correlación de expresión loci-específica de TEs con genes.*

Intersección TEs - Genes

Con los TEs que fueron encontrados sobre-expresados en la condición mutante por ambas herramientas se realizó un análisis de intersecciones entre genes y TEs mediante el *software* Bedtools ***(Figura 18)***.

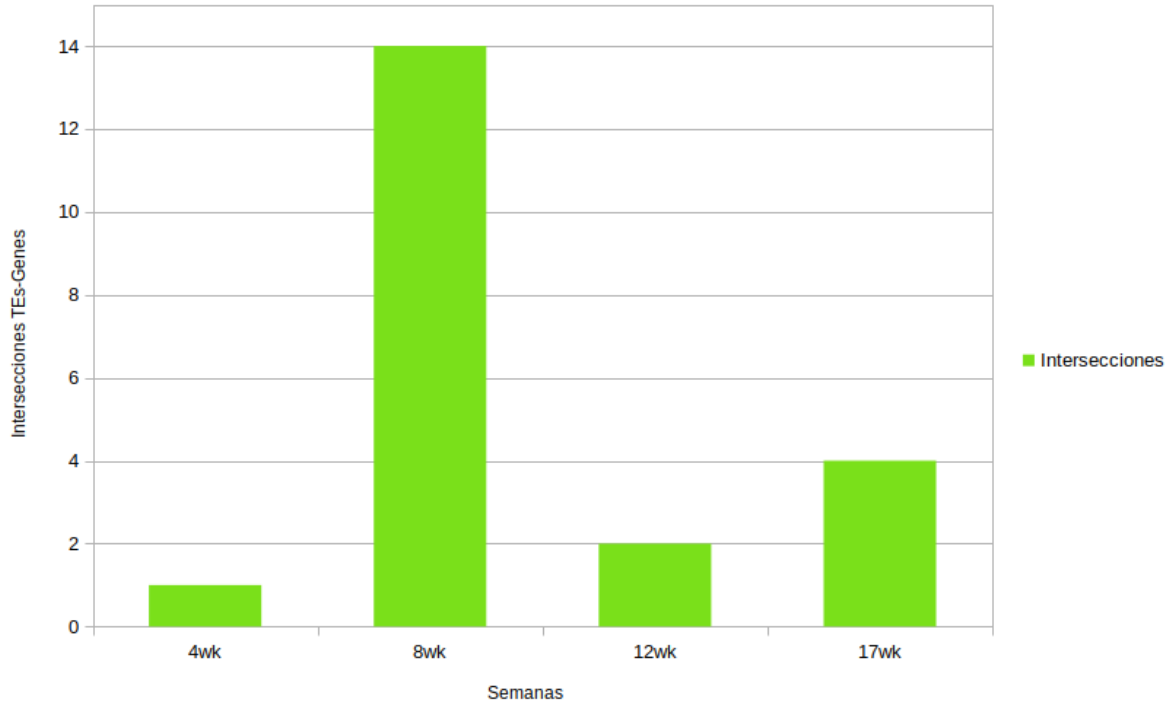


Figura 18: ***Intersecciones encontradas entre TEs y Genes:*** *Luego de la intersección de resultados de ambos pipelines, con los TEs encontrados por ambas herramientas se hizo una búsqueda de intersecciones con genes, en donde se encontró 1 TEs dentro de un Gen en la Semana 4, 14 en la Semana 8, 2 en la Semana 12 y 4 en la Semana 17.*

Gen	TE
Slc15a2	IAPEz-int:ERVK:LTR
Ube3a	PB1D10:Alu:SINE
Ube3a	L1MB2:L1:LINE
Ube3a	L1_Mm:L1:LINE
Snhg14	PB1D10:Alu:SINE
Snhg14	B3:B2:SINE
Snhg14	L1M4:L1:LINE
Chd9	RLTR4_Mm:ERV1:LTR
Serpina3n	L3:CR1:LINE
Cep85	RLTR4_MM-int:ERV1:LTR

*Tabla 3: **Intersección Gen-TE:** En la tabla se muestran las intersecciones encontradas entre genes de Mus musculus y TEs encontrados sobre-expresados en la condición mutante por ambos pipelines mediante el software Bedtools.*

Para la realización de la etapa de evaluación de intersecciones entre genes de *Mus musculus* y los TEs encontrados sobre-expresados por los pipelines de SQuIRE y TEcandidates se utilizó el archivo BED entregado por la herramienta *squire-fetch.* Por otro lado, los archivos de TEs se filtraron y formatearon mediante *awk* en la consola de Linux. Además, las intersecciones anteriormente encontradas (*Tabla 3*) se verificaron mediante el *software* Integrative Genomics Viewer (IGV) (Thorvaldsdóttir et al., 2013).

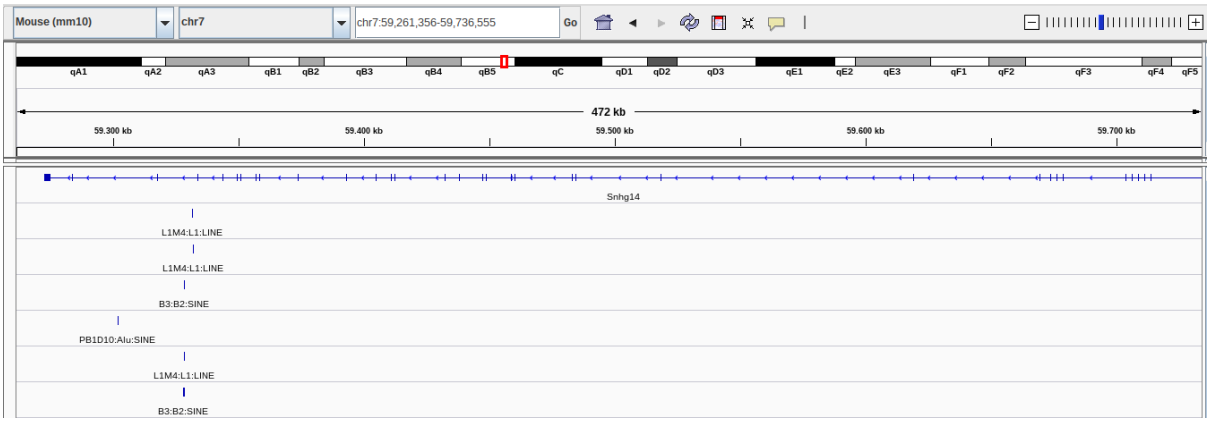


*Figura 19: **Visualización intersección gen Snhg14 - TEs:** Se encontraron seis intersecciones de TEs con el gen Snhg14: PB1D10:Alu:SINE (1 intersección), B3:B2:SINE (2 intersecciones) y L1M4:L1:LINE (3 intersecciones). Todas aquellas intersecciones que se encontraron con TEs se ubican en regiones intrónicas.*

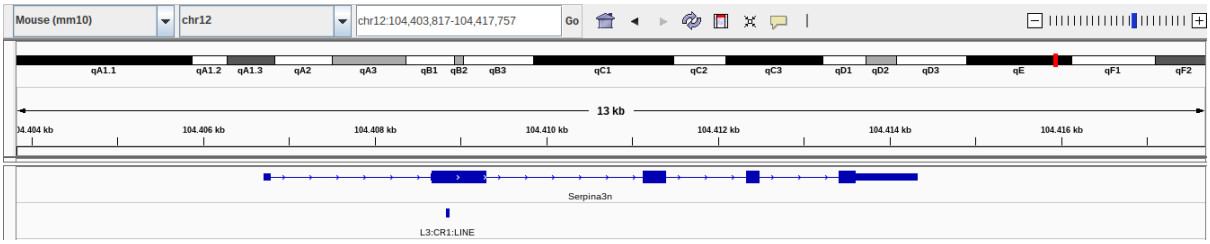


Figura 20: ***Visualización intersección gen Serpina3 - TE:*** *Se encontró una intersección de un TE (L3:CR1:LINE) con el gen Serpina3, el cual se encuentra dentro de un exón.*

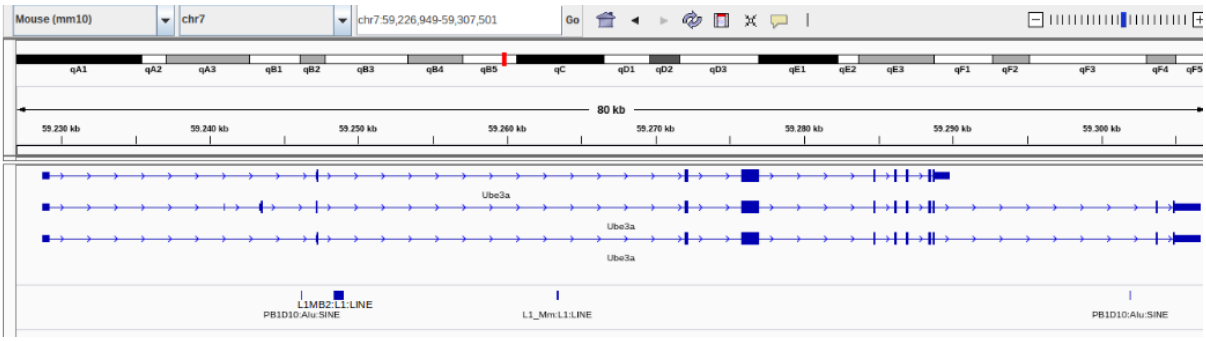


Figura 21: ***Visualización intersección gen Ube3a - TEs:*** *Se encontraron cuatro intersecciones de TEs con el gen Ube3a: PB1D10:Alu:SINE (dos veces), L1MB2:L1:LINE (una vez) y L1_Mm:L1:LINE (una vez). Todos los TEs encontrados intersectan en regiones intrónicas del gen.*

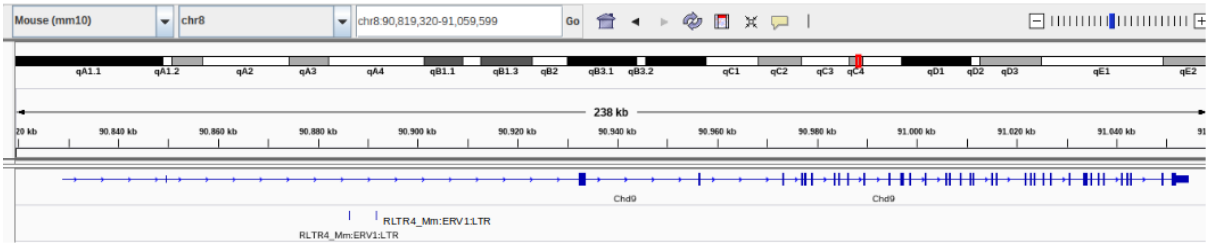


Figura 22: ***Visualización intersección gen Chd9 - TEs:*** *Se encontraron dos intersecciones del TE RLTR4_Mm:ERV1:LTR con el gen Chd9, las cuales se encuentran ambas en regiones intrónicas del mismo.*

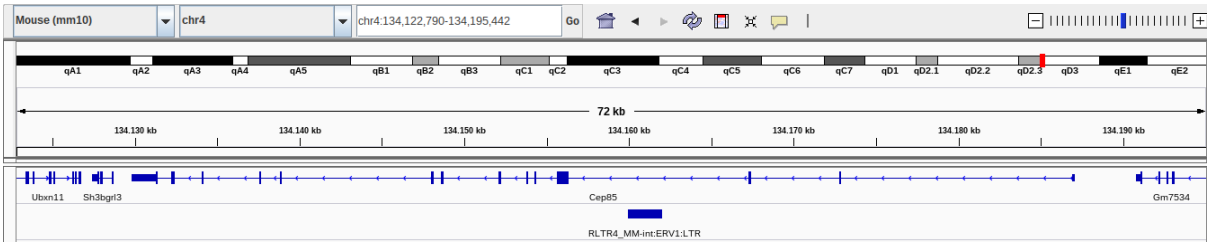


Figura 23: ***Visualización intersección gen Cep85 - TEs:*** *Se encontró una intersección del TE RLTR4_MM-int:ERV1:LTR con el gen Cep85, el cual se encuentra dentro de una región intrónica.*

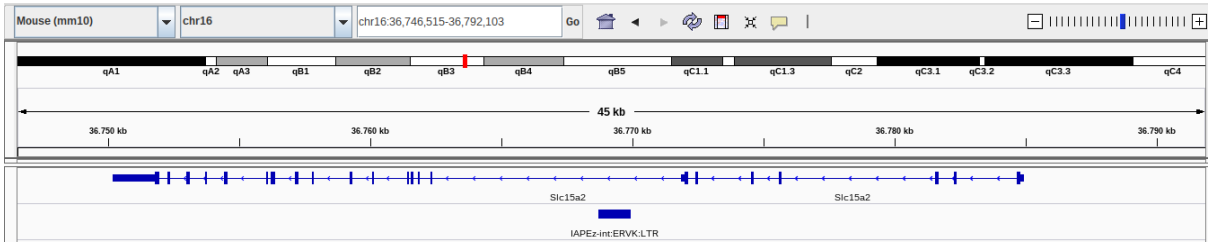


Figura 24: ***Visualización intersección gen Slc15a2 - TE:*** *Se encontró una intersección del TE IAPEz-int:ERVK:LTR con el gen Slc15a2, el cual se encuentra dentro de una región intrónica.*

Posterior a la visualización de la posición en donde se encuentran las intersecciones de los TEs con genes **(*Figuras 19 - 24*)** se procedió a analizar el aumento o disminución de la transcripción de ambos elementos a lo largo de la enfermedad mediante la comparación de sus TPM (*Transcritos por millón*) **(*Figuras 25 - 33*)**. Los gráficos resultantes se dividieron en tres grupos: Co-transcripción, Modulado Positivo y No Relacionados.

Objetivo #3: *Análisis del potencial impacto de los genes afectados por TEs en la progresión de la ELA.*

Co-transcripción:

La co-transcripción de un TE y Gen es un fenómeno en donde se puede observar una tendencia de transcripción similar para tanto el TE, como para su Gen hospedero. Con los datos utilizados en este trabajo no se puede distinguir si el TE impulsa la transcripción del Gen, o vice-versa. Por esto, se categoriza como “co-transcripción”. Para pareja de TE-Gen que se encuentra en este grupo, se puede especular que ocurre una co-transcripción hasta la semana 12, ya que el TE está dentro de un exón del gen **(*Figura 20*)**. Posterior a esto es probable que comience a transcribirse otra versión del gen que no incluya el exón en el que está inmerso dicho TE.

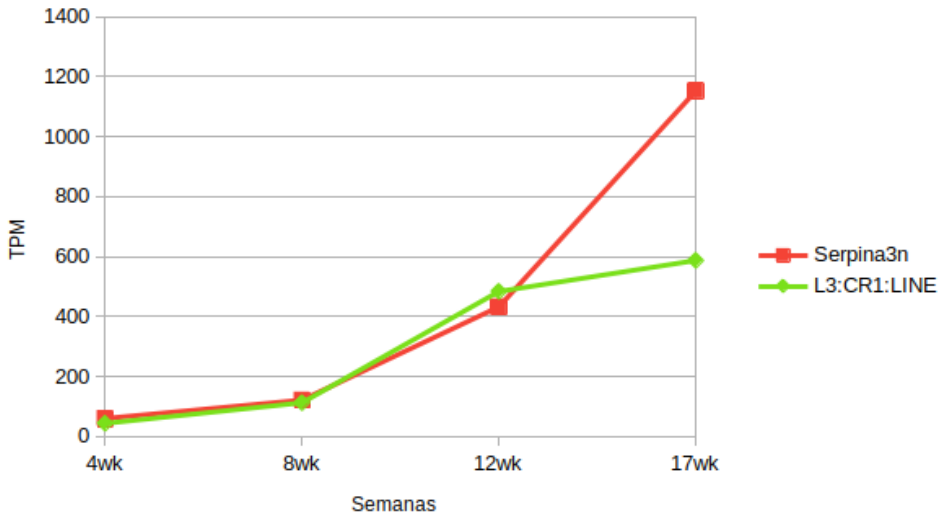


Figura 25: ***Análisis TPM a lo largo de la progresión de la enfermedad del gen Serpina3 y el TE L3:CR1:LINE:*** *Se presume que ambos elementos, gen y TE, se están co-transcribiendo hasta la semana 12 de la enfermedad y, posterior a eso, existe una sobre-expresión del gen.*

Modulado Positivo:

Para las parejas de TE-Genes que se encuentran en este grupo, se especula que el TE podría estar teniendo un rol de modular la transcripción del gen, al menos hasta la semana 12 de la enfermedad, ya que en la mayoría de ellos, posterior a esta semana comienza una transcripción diferente de TE y gen, es decir, uno de los dos elementos comienza a expresarse más y el otro a reprimirse. Similar al caso anterior, una hipótesis alternativa es que el Gen sea el que está impulsando la actividad transcripcional del TE.

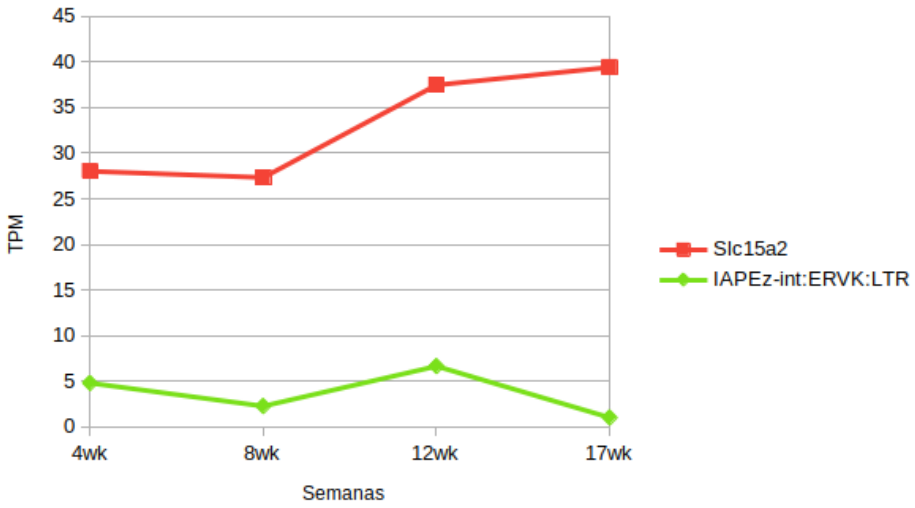


Figura 26: ***Análisis TPM a lo largo de la progresión de la enfermedad del gen Slc15a2 y el TE IAPEz-int:ERVK:LTR:*** *Se presume que hasta la semana 12 existe un modulado del gen por parte del TE que luego deja de suceder a partir de dicha semana, por lo que el gen comienza a transcribirse mayormente y el TE a reprimirse.*

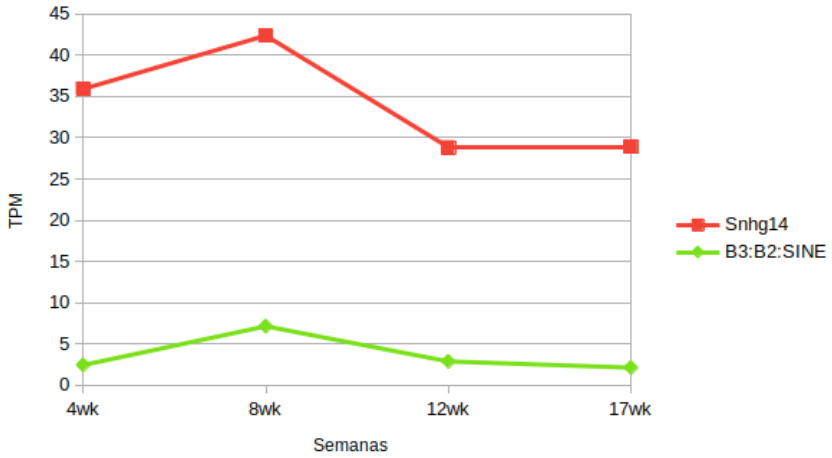


Figura 27: ***Análisis TPM a lo largo de la progresión de la enfermedad del gen Snhg14 y el TE B3:B2:SINE:*** *Se presume que hasta la semana 12 existe un modulado del gen por parte del TE que luego deja de suceder a partir de dicha semana, por lo que el gen comienza a transcribirse al mismo nivel pero el TE a reprimirse.*

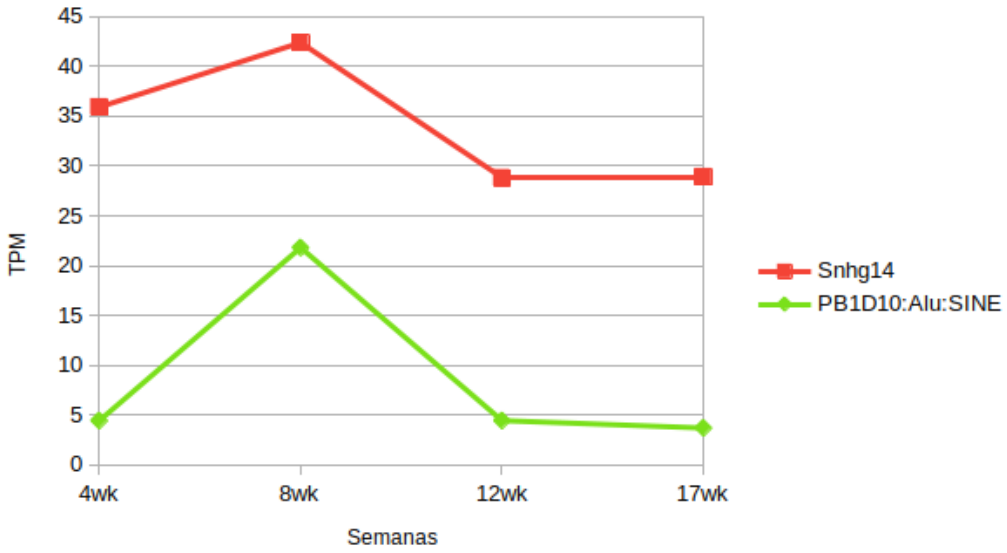


Figura 28: ***Análisis TPM a lo largo de la progresión de la enfermedad del gen Snhg14 y el TE PB1D10:Alu:SINE:*** *Se presume que hasta la semana 12 existe un modulado del gen por parte del TE que luego deja de suceder a partir de dicha semana, por lo que el gen comienza a transcribirse al mismo nivel pero el TE a reprimirse.*

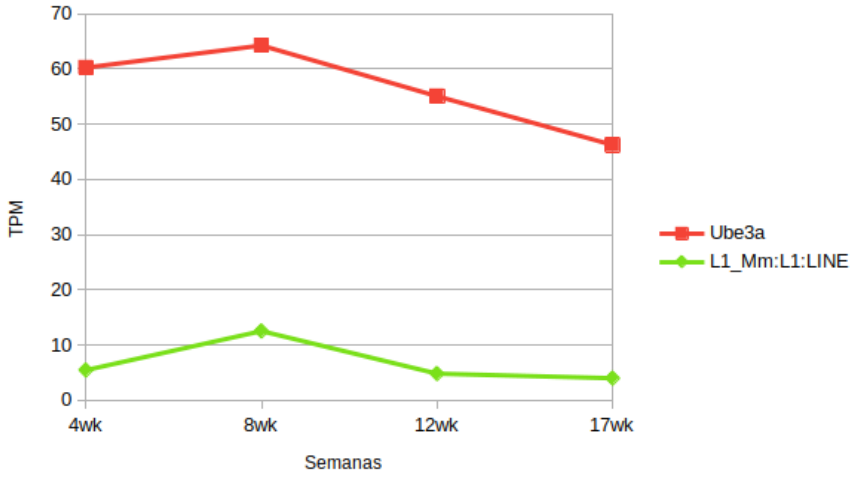


Figura 29: ***Análisis TPM a lo largo de la progresión de la enfermedad del gen Ube3a y el TE L1_Mm:L1:LINE:*** *Se presume que hasta la semana 12 existe un modulado del gen por parte del TE que luego deja de suceder a partir de dicha semana, por lo que el gen comienza a reprimirse y el TE se sigue transcribiendo al mismo nivel.*

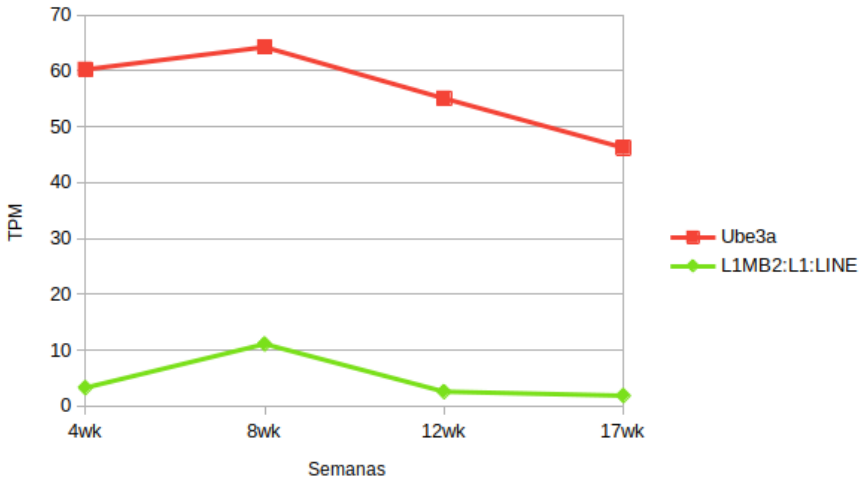


Figura 30: ***Análisis TPM a lo largo de la progresión de la enfermedad del gen Ube3a y el TE L1MB2:L1:LINE:*** *Se presume que hasta la semana 12 existe un modulado del gen por parte del TE que luego deja de suceder a partir de dicha semana, por lo que el gen comienza a reprimirse y el TE se sigue transcribiendo al mismo nivel.*

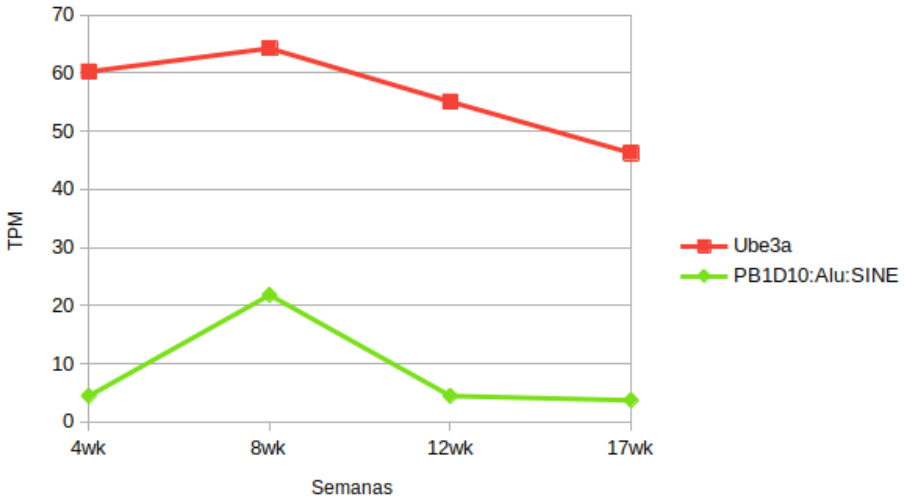


Figura 31: ***Análisis TPM a lo largo de la progresión de la enfermedad del gen Ube3a y el TE PB1D10:Alu:SINE:*** *Se presume que hasta la semana 12 existe un modulado del gen por parte del TE que luego deja de suceder a partir de dicha semana, por lo que el gen comienza a reprimirse y el TE se sigue transcribiendo al mismo nivel.*

No Relacionados:

Dentro de las parejas de TE-Gen que se encuentran en este grupo no se encontró una relación clara en términos de transcripción.

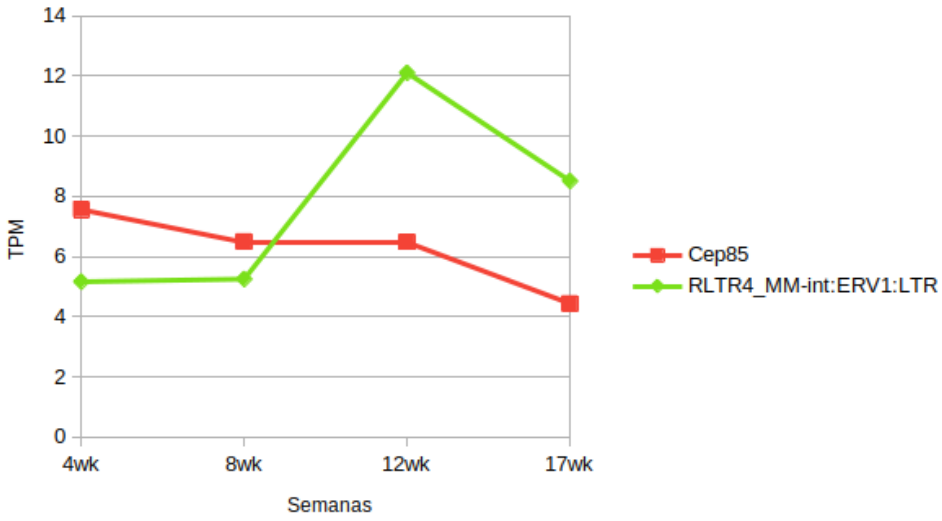


Figura 32: ***Análisis TPM a lo largo de la progresión de la enfermedad del gen Cep85 y el TE RLTR4_MM-int:ERV1:LTR.***

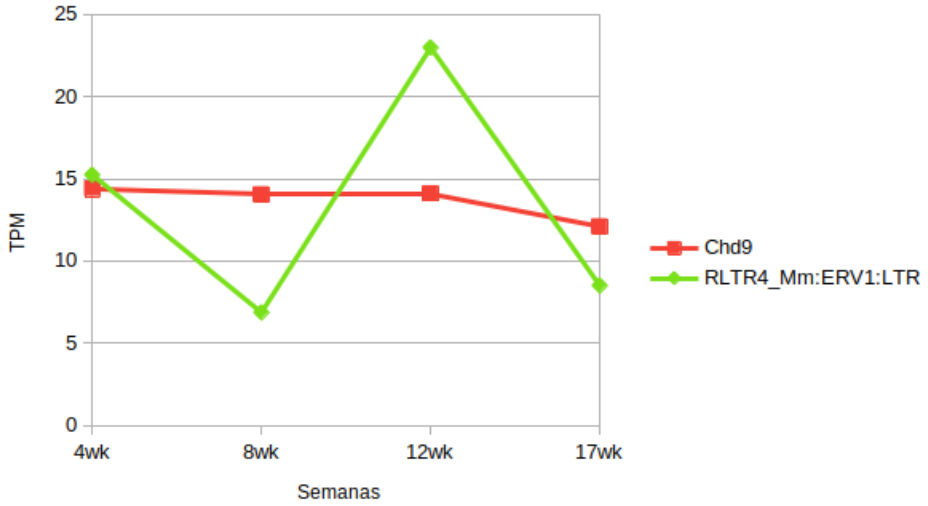


Figura 33: ***Análisis TPM a lo largo de la progresión de la enfermedad del gen Chd9 y el TE RLTR4_Mm:ERV1:LTR.***

Relación genes encontrados con la neurodegeneración

Para los genes anteriores, se hizo una búsqueda bibliográfica para evaluar cuáles de ellos podrían estar involucrados con neurodegeneración. Así, se puede establecer un link putativo entre TE->Gen, y luego entre Gen->Neurodegeneración, que en este último caso correspondería a implicaciones en la progresión de ELA.

Slc15a2

La superfamilia de transportadores de solutos (SLC) es uno de los principales subgrupos de proteínas de membrana en células en mamíferos. Las proteínas transportadoras de solutos incluyen más de 400 transportadores de solutos diferentes que atraviesan la membrana, organizados con 65 familias en el ser humano. En la familia de los transportadores de solutos de las neuronas, el neurotransmisor se considera un *target* de los fármacos neuropsiquiátricos debido a su importante papel en la recuperación de neurotransmisores como GABA, glutamato, serotonina, dopamina y noradrenalina y, también, en la regulación de su concentración en regiones sinápticas. Por lo tanto, los transportadores de solutos desempeñan funciones vitales y diferentes en los trastornos neurodegenerativos (Ayka & Şehirli, 2020).

El gen Slc15a2 tiene principalmente como función inhibir la enzima aminopeptidasa que se encarga de degradar proteínas en oligopéptidos más pequeños[8], por lo que su sobre-expresión podría implicar acumulación de proteínas que deberían ser degradadas. Además, se piensa que, junto a la expresión otras proteínas, pueden ser biomarcadores en enfermedades neurodegenerativas, tales como el Alzheimer. Por otro lado, el transportador Slc15a2 (también conocido como Pept2) es la principal proteína involucrada en la recuperación de aminoácidos unidos a péptidos y fármacos similares a péptidos en el riñón. También se cree que Pept2 tiene un papel en la regulación de la homeostasis de los neuropéptidos y la permeabilidad de la barrera hematoencefálica. Las alteraciones de esta son un aspecto importante de las enfermedades neurodegenerativas en humanos (Arisi et al., 2011). Considerando todo lo anterior, y que en ELA uno de los defectos celulares que ocurre es la acumulación de proteínas, este es un interesante ejemplo de un

8 https://www.uniprot.org/uniprot/Q16348#function

hallazgo realizado en este trabajo que potencialmente vincularía TEs con alguna de las consecuencias fenotípicas que tiene la enfermedad.

Snhg14

Se tiene evidencia de que hay una elevada expresión de Snhg14 en tejido cerebral del modelo de ratón en la enfermedad de Párkinson y, además, que SNHG14 reprimido puede mitigar las lesiones neuronales en ratones con Párkinson (Zhang et al., 2019). También se sabe que el silenciamiento de Snhg14 provoca un fuerte aumento de la viabilidad celular, una notable represión de la apoptosis celular, así como una clara reducción de la inflamación celular (Zhou et al., 2020).

Ube3a

Ube3a cumple una función en el desarrollo del cerebro y, además, en la respuesta celular al estímulo de factores neurotróficos derivados del cerebro, los cuales son una familia de proteínas que favorecen la supervivencia de las neuronas.[9]

Por otro lado, se ha demostrado que la ubiquitin-ligasa Ube3a promueve la eliminación de varias proteínas mal plegadas ligadas a enfermedades neurodegenerativas y también está implicada en la progresión del modelo de ratón de la enfermedad de Huntington. Ube3a también juega un papel importante en la función sináptica y la plasticidad. Además, la pérdida de la función del gen Ube3a heredado por la madre causa el síndrome de Angelman, que se caracteriza por discapacidades intelectuales y del desarrollo. Estos hallazgos sugieren que la función aberrante de Ube3a podría influir en la progresión del Alzheimer y restaurar el nivel normal de Ube3a podría ser beneficioso para dicha enfermedad (Singh et al., 2017).

Serpina3

Existe evidencia que Serpina3 juega un papel central en los efectos protectores de la melatonina, la cual es un agente antiinflamatorio bien conocido con actividad neuroprotectora significativa. El cloruro de trimetiltin (TMT) es una potente

9 https://www.uniprot.org/uniprot/Q05086

neurotoxina que causa neuroinflamación y muerte de células neuronales. Además, la sobre-expresión de Serpina3 en el hipocampo de ratón puede reprimir los efectos protectores de la melatonina sobre la neuroinflamación y neurotoxicidad inducidas por TMT. La melatonina protege a las células contra la neurotoxicidad inducida por TMT al inhibir la neuroinflamación mediada por Serpina3 (Xi et al., 2019).

Por otro lado, se ha visto que Serpina3 está fuertemente sobre-expresada en el cerebro de todas las enfermedades priónicas humanas, con una leve sobre-expresión en Alzheimer. Se ha demostrado que esta llamativa sobre-expresión, tanto a nivel de ARNm como de proteínas, está presente en todos los tipos de enfermedades priónicas humanas analizadas, aunque en diferente medida para cada trastorno específico. Se sugiere que Serpina3 puede estar involucrado en la patogénesis y la progresión de las enfermedades priónicas, lo que representa una herramienta válida para distinguir diferentes formas de estos trastornos en humanos. (Vanni et al., 2017)

Por último, se tiene que la versión KO de Serpina3 puede resultar en un deterioro del aprendizaje y memoria en ratones. (Z.-M. Wang et al., 2020) Lo cual sugeriría que, tanto su sobre-expresión, como su represión podrían tener un efecto adverso en el individuo, por lo tanto necesitaría estar cuidadosamente modulada para tener un rol “normal” dentro del organismo.

Chd9

La familia de proteínas de unión a ADN de cromodominio helicasa (CHD) son remodeladores de cromatina dependientes de ATP que contribuyen a la reorganización de la estructura de la cromatina y al depósito de las variantes de histonas necesarias para regular la expresión génica. Las proteínas CHD juegan un papel importante en el neurodesarrollo, ya que las variantes patogénicas se han asociado con una variedad de fenotipos neurológicos, incluido el trastorno del espectro autista, la discapacidad intelectual y la epilepsia. (Lamar & Carvill, 2018)
Esta función de reorganización de la cromatina podría ser importante, ya que la manipulación genética de las vías de modificación de la cromatina o de reparación del ADN puede suprimir la neurotoxicidad, lo que sugiere que el mantenimiento de la

integridad genómica y la neurodegeneración en la enfermedad de Alzheimer puede estar relacionado causalmente en lugar de simplemente una consecuencia posterior de la muerte celular. (Guo et al., 2018)

Cep85

Nek2 se ha implicado en la disyunción del centrosoma al inicio de la mitosis para promover la formación del huso bipolar, y la hiperactivación de Nek2 conduce a la separación prematura del centrosoma. Su actividad, por tanto, necesita estar estrictamente regulada. Se ha reportado que Cep85, una proteína del centrosoma, actúa como un *binding partner* ("socio de unión") de Nek2A. Se co-localiza con la isoforma A de Nek2 (Nek2A) en los centrosomas y forma una red de gránulos que envuelve los extremos proximales de los centriolos. Cep85, a la vez,e coopera con PP1γ (también conocido como PPP1CC) para antagonizar la actividad de Nek2A con el fin de mantener la integridad del centrosoma en interfase en células de mamíferos. Cabe mencionar que el centrosoma es un orgánulo que actúa como el principal centro organizador de microtúbulos para promover la formación del huso bipolar y la progresión mitótica oportuna. Entonces, una desregulación en la expresión de Cep85 significaría una desestabilización de la célula y, por lo tanto, una muerte de la misma. (C. Chen et al., 2015).

Toda la información expuesta anteriormente se resume en la Tabla 4, mostrada a continuación:

Gen	**Expresión**	**Correspondencia con neurodegeneración**
Slc15a2	Sobre-expresión	Positiva
Snhg14	Represión	Positiva
Ube3a	Sobre-expresión/Represión	Negativa
Serpina3	Sobre-expresión/Represión	Negativa
Chd9	Represión	Negativa
Cep85	Sobre-expresión/Represión	Negativa

Tabla 4: ***Resumen de las consecuencias en el cambio de expresión del gen su la correlación con la neurodegeneración.*** *Para cada gen (columna "Gen"), se indican cambios en su expresión (columna "Expresión"), y cómo estos cambios se correlacionan (columna "Correspondencia con neurodegeneración") según la información obtenida en la bibliografía.*

Discusión

En términos de planificación de este proyecto, en un principio, se pensaba que mediante la correlación del log_2(Fold Change) entre genes y TEs, se tendría una idea de qué genes podrían estar siendo afectados de manera putativa por un TE. Sin embargo, esta medida sólo informa sobre cambios relativos, y no tiene sentido biológico comparar estos cambios relativos entre un gen y un TE. Por esto, se procedió a hacer un cálculo del TPM y posterior análisis mediante la visualización de los valores a lo largo de la progresión de la enfermedad en gráficos por parejas (Gen - TE).

Considerando los resultados obtenidos entre número de TEs, primero que nada, es necesario mencionar que SQuIRE parece generar muchos más resultados que TEcandidates. Así, la intersección de resultados entre programas, redujo el número de TEs a una cantidad pequeña, para la cual se pudo mostrar resultados en más detalle. Una de las principales limitaciones de todas formas, es que ninguna de estas herramientas, al menos en este trabajo, permite discernir si eventos de co-transcripción ocurren como consecuencia de la actividad transcripcional del TE impactando a su gen hospedero, o viceversa.

Si bien, como se ha mencionado anteriormente, una hipótesis alternativa que no se descarta es que en todos los casos el gen podría impulsar la expresión del TE. No obstante, en la parte final de este trabajo, la conclusión tomada es de que el TE podría estar impactando en la regulación del Gen. Así, se piensa que los TEs pueden estar teniendo una actividad regulatoria de la transcripción y expresión de los genes en lo que se encuentran inmersos, al menos hasta la semana 12 del modelo de ratón transgénico $SOD1^{G93A}$, ya que luego de esto, en la mayoría de los casos, se encontró un cambio en el patrón de transcripción de ambos elementos involucrados. Además, se confirma mediante la búsqueda bibliográfica que los genes involucrados con TEs están relacionados de alguna forma en la neurodegeneración, ya sea, con evidencia en ELA o en otras enfermedades de la misma familia, tales como el Parkinson o Alzheimer.

Conclusiones

El estudio de la expresión de los Elementos Transponibles en diferentes condiciones puede dar una idea de lo que puede estar sucediendo en dichos estados o, incluso, confirmar información que ya se tenía sobre genes que se están viendo afectados, más aún ahora con las herramientas computacionales recientemente disponibles para este tipo de análisis.

Este es uno de los primeros trabajos en explotar las herramientas actuales para análisis loci-específico de TEs, permitiendo correlacionarlos de alguna manera con Genes. Así, el principal hallazgo de este trabajo está en que es posible encontrar TEs expresados antes de la manifestación fenotípica de la enfermedad ELA. Potencialmente entonces, se podrían implicar como agentes causantes de esta enfermedad. Sin embargo, futuros trabajos serán necesarios para seguir elucidando el rol que estos elementos puedan tener en la regulación génica, y progresión de ELA.

Dicho esto, en relación a la hipótesis planteada inicialmente antes de la realización de este estudio haciendo alusión a que los TEs tienen un efecto causal en la enfermedad de ELA, a raíz de los resultados obtenidos se puede concluir que dichos elementos podrían tener un rol más regulatorio de la expresión de ciertos genes involucrados que un rol causante de la enfermedad.

Bibliografía

1. Anders, S., Pyl, P. T., & Huber, W. (2015). HTSeq—A Python framework to work with high-throughput sequencing data. *Bioinformatics*, *31*(2), 166–169. https://doi.org/10.1093/bioinformatics/btu638
2. Arisi, I., D'Onofrio, M., Brandi, R., Felsani, A., Capsoni, S., Drovandi, G., Felici, G., Weitschek, E., Bertolazzi, P., & Cattaneo, A. (2011). Gene expression biomarkers in the brain of a mouse model for Alzheimer's disease: Mining of microarray data by logic classification and feature selection. *Journal of Alzheimer's Disease: JAD*, *24*(4), 721–738. https://doi.org/10.3233/JAD-2011-101881
3. Aubert, J., Bar-Hen, A., Daudin, J.-J., & Robin, S. (2004). Determination of the differentially expressed genes in microarray experiments using local FDR. *BMC Bioinformatics*, *5*, 125. https://doi.org/10.1186/1471-2105-5-125
4. Ayka, A., & Şehirli, A. Ö. (2020). The Role of the SLC Transporters Protein in the Neurodegenerative Disorders. *Clinical Psychopharmacology and Neuroscience*, *18*(2), 174–187. https://doi.org/10.9758/cpn.2020.18.2.174
5. Aziza, R. (2018). From Mouse Models to Human Disease: An Approach for Amyotrophic Lateral Sclerosis. *In Vivo*, *32*(5), 983–998. https://doi.org/10.21873/invivo.11339
6. Balendra, R., & Isaacs, A. M. (2018). C9orf72-mediated ALS and FTD: Multiple pathways to disease. *Nature Reviews. Neurology*, *14*(9), 544–558. https://doi.org/10.1038/s41582-018-0047-2
7. Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics (Oxford, England)*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170
8. Chen, C., Tian, F., Lu, L., Wang, Y., Xiao, Z., Yu, C., & Yu, X. (2015).

Characterization of Cep85—A new antagonist of Nek2A that is involved in the regulation of centrosome disjunction. *Journal of Cell Science*, *128*(17), 3290–3303. https://doi.org/10.1242/jcs.171637

9. Chen, Y.-C., Liu, T., Yu, C.-H., Chiang, T.-Y., & Hwang, C.-C. (2013). Effects of GC bias in next-generation-sequencing data on de novo genome assembly. *PloS One*, *8*(4), e62856. https://doi.org/10.1371/journal.pone.0062856
10. Chu, Y., & Corey, D. R. (2012). RNA Sequencing: Platform Selection, Experimental Design, and Data Interpretation. *Nucleic Acid Therapeutics*, *22*(4), 271–274. https://doi.org/10.1089/nat.2012.0367
11. Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England)*, *29*(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635
12. 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
13. Feschotte, C. (2008a). The contribution of transposable elements to the evolution of regulatory networks. *Nature Reviews. Genetics*, *9*(5), 397–405. https://doi.org/10.1038/nrg2337
14. Feschotte, C. (2008b). Transposable elements and the evolution of regulatory networks. *Nature Reviews Genetics*, *9*(5), 397–405. https://doi.org/10.1038/nrg2337
15. Goetz, C. G. (2000). *Amyotrophic lateral sclerosis: Early contributions of Jean-Martin Charcot*. 8.
16. Guo, C., Jeong, H.-H., Hsieh, Y.-C., Klein, H.-U., Bennett, D. A., De Jager, P. L., Liu, Z., & Shulman, J. M. (2018). Tau Activates Transposable Elements in Alzheimer's Disease. *Cell Reports*, *23*(10), 2874–2880. https://doi.org/10.1016/j.celrep.2018.05.004
17. Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J.,

Couger, M. B., Eccles, D., Li, B., Lieber, M., MacManes, M. D., Ott, M., Orvis, J., Pochet, N., Strozzi, F., Weeks, N., Westerman, R., William, T., Dewey, C. N., … Regev, A. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols*, *8*(8), 1494–1512. https://doi.org/10.1038/nprot.2013.084

18. Hardiman, O., Al-Chalabi, A., Chio, A., Corr, E. M., Logroscino, G., Robberecht, W., Shaw, P. J., Simmons, Z., & van den Berg, L. H. (2017). Amyotrophic lateral sclerosis. *Nature Reviews Disease Primers*, *3*(1), 1–19. https://doi.org/10.1038/nrdp.2017.71
19. Hoen, D. R., Hickey, G., Bourque, G., Casacuberta, J., Cordaux, R., Feschotte, C., Fiston-Lavier, A.-S., Hua-Van, A., Hubley, R., Kapusta, A., Lerat, E., Maumus, F., Pollock, D. D., Quesneville, H., Smit, A., Wheeler, T. J., Bureau, T. E., & Blanchette, M. (2015). A call for benchmarking transposable element annotation methods. *Mobile DNA*, *6*(1), 13. https://doi.org/10.1186/s13100-015-0044-6
20. Kim, D., Paggi, J. M., Park, C., Bennett, C., & Salzberg, S. L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology*, *37*(8), 907–915. https://doi.org/10.1038/s41587-019-0201-4
21. Kumar, D. R., Aslinia, F., Yale, S. H., & Mazza, J. J. (2011). Jean-Martin Charcot: The Father of Neurology. *Clinical Medicine & Research*, *9*(1), 46–49. https://doi.org/10.3121/cmr.2009.883
22. Lamar, K.-M. J., & Carvill, G. L. (2018). Chromatin Remodeling Proteins in Epilepsy: Lessons From CHD2-Associated Epilepsy. *Frontiers in Molecular Neuroscience*, *11*. https://doi.org/10.3389/fnmol.2018.00208
23. 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
24. Leinonen, R., Sugawara, H., & Shumway, M. (2011). The Sequence Read Archive. *Nucleic Acids Research*, *39*(Database issue), D19–D21. https://doi.org/10.1093/nar/gkq1019

25. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352
26. Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 550. https://doi.org/10.1186/s13059-014-0550-8
27. Nguyen, H. P., Broeckhoven, C. V., & Zee, J. van der. (2018). ALS Genes in the Genomic Era and their Implications for FTD. *Trends in Genetics*, *34*(6), 404–423. https://doi.org/10.1016/j.tig.2018.03.001
28. Ochoa Thomas, E., Zuniga, G., Sun, W., & Frost, B. (2020). Awakening the dark side: Retrotransposon activation in neurodegenerative disorders. *Current Opinion in Neurobiology*, *61*, 65–72. https://doi.org/10.1016/j.conb.2020.01.012
29. Pattamatta, A., Cleary, J. D., & Ranum, L. P. W. (2018). All in the Family: Repeats and ALS/FTD. *Trends in Neurosciences*, *41*(5), 247–250. https://doi.org/10.1016/j.tins.2018.03.010
30. Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T.-C., Mendell, J. T., & Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology*, *33*(3), 290–295. https://doi.org/10.1038/nbt.3122
31. Phatnani, H. P., Guarnieri, P., Friedman, B. A., Carrasco, M. A., Muratet, M., O'Keeffe, S., Nwakeze, C., Pauli-Behn, F., Newberry, K. M., Meadows, S. K., Tapia, J. C., Myers, R. M., & Maniatis, T. (2013). Intricate interplay between astrocytes and motor neurons in ALS. *Proceedings of the National Academy of Sciences*, *110*(8), E756–E765. https://doi.org/10.1073/pnas.1222361110
32. Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841–842. https://doi.org/10.1093/bioinformatics/btq033

33. R Development Core Team, R. (2006). A language and environment for statistical computing. *Computing*, *1*. https://doi.org/10.1890/0012-9658(2002)083[3097:CFHIWS]2.0.CO;2
34. Renton, A. E., Chiò, A., & Traynor, B. J. (2014). State of play in amyotrophic lateral sclerosis genetics. *Nature Neuroscience*, *17*(1), 17–23. https://doi.org/10.1038/nn.3584
35. Singh, B. K., Vatsa, N., Kumar, V., Shekhar, S., Sharma, A., & Jana, N. R. (2017). Ube3a deficiency inhibits amyloid plaque formation in APPswe/PS1δE9 mouse model of Alzheimer's disease. *Human Molecular Genetics*, *26*(20), 4042–4054. https://doi.org/10.1093/hmg/ddx295
36. Siomi, M. C., Sato, K., Pezic, D., & Aravin, A. A. (2011). PIWI-interacting small RNAs: The vanguard of genome defence. *Nature Reviews Molecular Cell Biology*, *12*(4), 246–258. https://doi.org/10.1038/nrm3089
37. Stajich, J. E., Block, D., Boulez, K., Brenner, S. E., Chervitz, S. A., Dagdigian, C., Fuellen, G., Gilbert, J. G. R., Korf, I., Lapp, H., Lehväslaiho, H., Matsalla, C., Mungall, C. J., Osborne, B. I., Pocock, M. R., Schattner, P., Senger, M., Stein, L. D., Stupka, E., … Birney, E. (2002). The Bioperl Toolkit: Perl Modules for the Life Sciences. *Genome Research*, *12*(10), 1611–1618. https://doi.org/10.1101/gr.361602
38. Thorvaldsdóttir, H., Robinson, J. T., & Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. *Briefings in Bioinformatics*, *14*(2), 178–192. https://doi.org/10.1093/bib/bbs017
39. Valdebenito-Maturana, B., & Riadi, G. (2018). TEcandidates: Prediction of genomic origin of expressed transposable elements using RNA-seq data. *Bioinformatics*, *34*(22), 3915–3916. https://doi.org/10.1093/bioinformatics/bty423
40. Vanni, S., Moda, F., Zattoni, M., Bistaffa, E., De Cecco, E., Rossi, M., Giaccone, G., Tagliavini, F., Haïk, S., Deslys, J. P., Zanusso, G., Ironside, J. W., Ferrer, I., Kovacs, G. G., & Legname, G. (2017). Differential overexpression of SERPINA3 in human prion diseases. *Scientific Reports*, *7*(1), 15637. https://doi.org/10.1038/s41598-017-

15778-8

41. Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: A revolutionary tool for transcriptomics. *Nature Reviews Genetics*, *10*(1), 57–63. https://doi.org/10.1038/nrg2484
42. Wang, Z.-M., Liu, C., Wang, Y.-Y., Deng, Y.-S., He, X.-C., Du, H.-Z., Liu, C.-M., & Teng, Z.-Q. (2020). SerpinA3N deficiency deteriorates impairments of learning and memory in mice following hippocampal stab injury. *Cell Death Discovery*, *6*(1), 1–11. https://doi.org/10.1038/s41420-020-00325-8
43. Xi, Y., Liu, M., Xu, S., Hong, H., Chen, M., Tian, L., Xie, J., Deng, P., Zhou, C., Zhang, L., HE, M., Chen, C., Lu, Y., Reiter, R., Yu, Z., Pi, H., & Zhou, Z. (2019). Inhibition of SERPINA3N-dependent neuroinflammation is essential for melatonin to ameliorate trimethyltin chloride-induced neurotoxicity. *Journal of Pineal Research*, *67*. https://doi.org/10.1111/jpi.12596
44. Yang, W. R., Ardeljan, D., Pacyna, C. N., Payer, L. M., & Burns, K. H. (2019). SQuIRE reveals locus-specific regulation of interspersed repeat expression. *Nucleic Acids Research*, *47*(5), e27–e27. https://doi.org/10.1093/nar/gky1301
45. Zhang, L.-M., Wang, M.-H., Yang, H.-C., Tian, T., Sun, G.-F., Ji, Y.-F., Hu, W.-T., Liu, X., Wang, J.-P., & Lu, H. (2019). Dopaminergic neuron injury in Parkinson's disease is mitigated by interfering lncRNA SNHG14 expression to regulate the miR-133b/ α-synuclein pathway. *Aging (Albany NY)*, *11*(21), 9264–9279. https://doi.org/10.18632/aging.102330
46. Zhou, S., Zhang, D., Guo, J., Zhang, J., & Chen, Y. (2020). Knockdown of SNHG14 Alleviates MPP+-Induced Injury in the Cell Model of Parkinson's Disease by Targeting the miR-214-3p/KLF4 Axis. *Frontiers in Neuroscience*, *14*. https://doi.org/10.3389/fnins.2020.00930