Facultad de ingeniería

Escuela de Ingeniería Civil en Bioinformática

Expresión de Elementos Transponibles en la formación de memoria y el recuerdo en el ratón a partir de un enfoque multiómico

Memoria para optar al grado de Ingeniero Civil en Bioinformática
Licenciado en Ciencias de la Ingeniería

Alumno: Javier Diaz Rojas.
Profesor Tutor: Braulio Valdebenito Maturana.
Profesor informante: Gonzalo Riadi Mahias

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, 2022

Tabla de Contenidos

Índice Tablas

Índice Figuras

Abstract

Memory formation occurs mainly in the engram, a subset of neuronal cells. In order to understand the molecular mechanisms involved in this process, a recent work was performed integrating RNA-Seq and ATAC-Seq data. As a result, a model of gene expression regulation based on chromatin accessibility was proposed. Despite this breakthrough in understanding this phenomenon, Transposable Elements were not analyzed.

Transposable elements (TEs) have the potential to provide regulatory sequences, or act directly as gene regulators. An individual TE can disrupt gene expression, directly or indirectly create an advantageous modification to gene expression, or have no immediate consequence. Because their role in gene regulation is well accepted, and there is evidence that TEs are activated in the brain, the hypothesis of this work is that TEs are involved in memory formation.

Therefore, in the present work, the RNA-Seq and ATAC-Seq data from the aforementioned work were reanalyzed to detect TEs. The aim of this thesis was to evaluate whether and at what stage they are involved in memory formation and recall in the mouse. Estimating the origin of TE expression using these methodologies (RNA-Seq and ATAC-Seq) in combination with TE prediction software (TEcandidates and SQuIRE) would reliably reveal TE expression, which may help to understand the role of TEs in memory formation.

The implementation of a multi-omics approach to evaluate TE locus-specific expression in this study resulted in 3 TEs potentially related to processes involving the brain.

Resumen

La formación de memoria ocurre principalmente en el engrama, un subconjunto de células neuronales. A fin de entender los mecanismos moleculares involucrados en este proceso, recientemente se realizó un trabajo en donde integraron datos de RNA-Seq y ATAC-Seq. Producto de esto, se propuso un modelo de regulación de expresión de genes basado en la accesibilidad de la cromatina. A pesar de este gran avance en comprender este fenómeno, no se analizaron Elementos Transponibles.

Los Elementos Transponibles (TEs), tienen el potencial de proporcionar secuencias reguladoras, o actuar directamente como reguladores de genes. Un TE individual puede interrumpir la expresión de un gen, crear directa o indirectamente una modificación ventajosa para la expresión de un gen o no tener ninguna consecuencia inmediata. Debido a que es bien aceptado su rol en regulación génica, y a que existe evidencia que indica que los TEs se activan en el cerebro, la hipótesis de este trabajo es que los TEs están involucrados en la formación de memoria.

Por lo anterior, en el presente trabajo se reanalizaron los datos de RNA-Seq y ATAC-Seq del trabajo de Marco y colaboradores del año 2020, a fin de detectar TEs. El objetivo de esta tesis era evaluar si es que, y en qué fase (Basal, Early, Late y Recall) estos se involucran en formación de memoria y el recuerdo en el ratón. Estimar el origen de expresión de TEs utilizando estas metodologías (RNA-Seq y ATAC-Seq) en combinación con software de predicción de TEs (TEcandidates y SQuIRE), nos revelaría expresión de TEs de manera confiable, lo que pudo ayudar a entender el rol que estos tienen en la formación de memoria.

La implementación de un enfoque multiómico para evaluar la expresión especifica de locus de TE en este estudio, dio como resultado 3 TEs potencialmente relacionados a procesos que involucran al cerebro.

Introducción

Originalmente se creía que el engrama era el lugar físico donde se forman las memorias (Josselyn et al., 2015). Hoy en día es bien aceptado que hay un subgrupo de células neuronales llamadas “células del engrama”, en donde ocurren los procesos moleculares que llevan a la formación de memoria a largo plazo. Dentro de los procesos moleculares involucrados en este fenómeno se destacan cambios epigenéticos. Estos cambios, denominados colectivamente como el epigenoma, corresponden a modificaciones químicas en nucleótidos específicos, o a las proteínas unidas al DNA, los que a su vez impactan la accesibilidad al DNA. La accesibilidad al DNA afecta directamente la expresión de genes. Considerando todo esto, recientemente se realizó un trabajo en donde se aplicaron técnicas de secuenciación a gran escala para entender en detalle la accesibilidad al DNA (mediante ATAC-Seq) y el transcriptoma (mediante RNA-Seq) de estas células, en un modelo de ratón (Marco et al., 2020).

El diseño experimental del trabajo de Marco y colaboradores (2020) se muestra en la Figura 1. En el trabajo, se sometió un grupo de ratones (de 8 a 10 animales en triplicado) a un estímulo. Este se aplicó al inicio del estudio y luego al final de este mismo; con una ventana de 5 días entre el inicio y final de este (fueron sometidos al paradigma clásico de condicionamiento contextual del miedo pavloviano). Este estímulo consta de un leve choque eléctrico en las patas, además de la emisión de un sonido a través de un parlante en la jaula. Luego, se utilizan biomarcadores (NeuN/eYFP) para marcar neuronas asociadas al desarrollo de la memoria y el recuerdo. En base a esto, ellos pudieron establecer las fases de la formación de la memoria y el recuerdo (Basal, Early, Late y Reactivated) a evaluar en la investigación. Una vez extraídas las diferentes muestras, se procedió a preparar las bibliotecas para RNA-Seq, ATAC-Seq y pc-HiC (Marco et al., 2020) (Figura 1). Se destaca en este punto que en este proyecto sólo se utilizaron los primeros dos tipos de datos, debido al tiempo en el que se debe realizar el trabajo de tesis. A continuación se procede a explicar cada una de las fases de la formación de la memoria y el recuerdo:

1) Las neuronas de estado basal (Basal) no presentan activación las proteínas de marcaje (NeuN+/eYFP-), ya que no se han expuesto al estímulo.
2) 1,5-2 horas después de la exposición al estímulo, se colectó cerebros para identificar neuronas marcadas (NeuN+/eYFP+), que se activaron durante la exposición inicial (Early).
3) Después de 5 días, se extrajeron neuronas que fueron marcadas (NeuN+/eYFP+) en el día de entrenamiento (Late).
4) 1,5-2 horas después de la fase de Late, se realizó una reexposición (NeuN+/eYFP+) del estímulo (Reactivate).

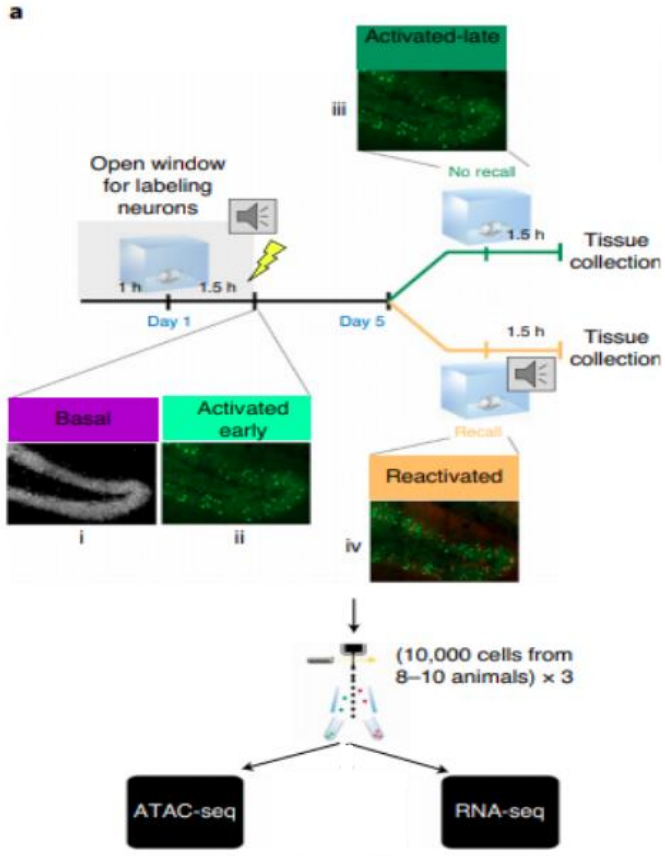

Figura 1. Representación esquemática del diseño experimental. Cuatro poblaciones neuronales diferentes (Basal, Early, Late y Reactivate) se clasificaron mediante citómetro de flujo y se sometieron a la preparación de bibliotecas para RNA-Seq y ATAC-Seq. Figura modificada de (Marco et al., 2020).

Como principal resultado del trabajo de Marco y colaboradores (2020), se propuso un modelo de la dinámica de la cromatina y la expresión de genes a lo largo de cada una de las etapas mencionadas anteriormente (Figura 2).

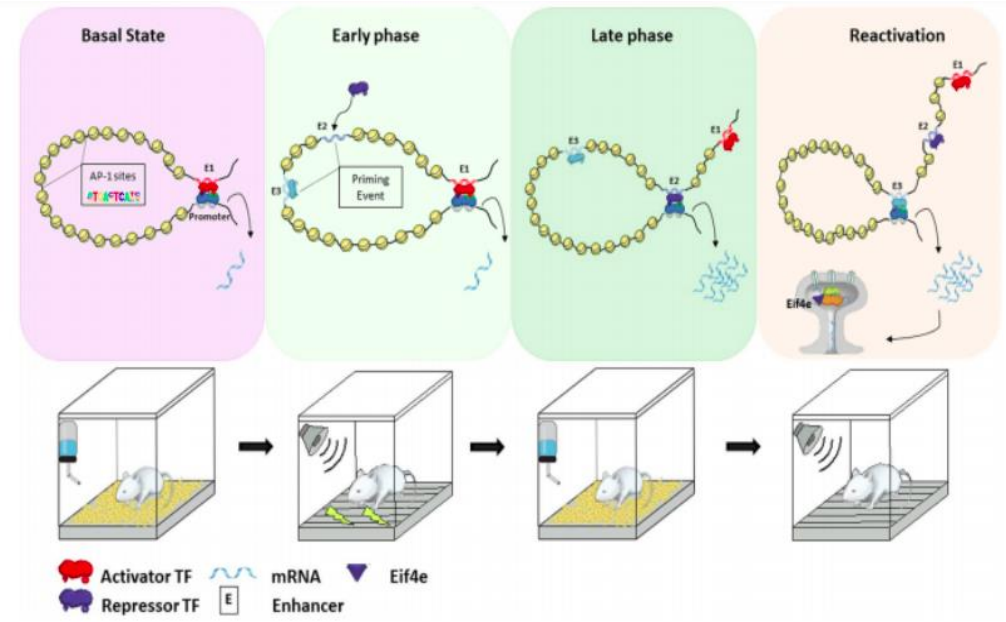

Figura 2. Propuesta de modelo sobre la accesibilidad de la cromatina y la transcripción de genes. Para cada fase (Basal, Early, Late y Reactivate) en la parte superior se esquematiza el DNA (línea negra) y los nucleosomas (círculos amarillos), los TFs activadores (figuras purpura y celeste) y represores (figura roja), los mRNAs (líneas celestes), mientras que en la parte inferior, el estado de los ratones estudiados: sin estímulo, estímulo sonoro + shock elétrico, sin estúmulo y estímulo sonoro respectivamente. Figura tomada de (Marco et al., 2020).

Para cada estado, se proponen los siguientes eventos a nivel molecular:

1. Estado Basal: los promotores de genes interactúan con enhancer que llevan carga transcripcional represora y expresan bajos niveles de mRNA.
2. Fase Early: conduce a un evento de activación, en el que los enhancer que albergan carga de activadores transcripcionales se hacen más accesibles, pero al permanecer aislados no forman interacciones con los promotores genéticos respectivos.
3. Fase Late: los promotores de genes tienen su interacción a las regiones de cebado (enhancer), que albergan motivos de activadores transcripcionales. Esta reprogramación de los promotor-enhancer da lugar a un aumento de la

expresión genética que se presume que permite la consolidación de la memoria.

4. Recall o Fase Reactivate: Las neuronas reactivadas del engrama utilizan un subconjunto de interacciones de promotores-enhancer cebados, que se asocia con cambios transcripcionales implicados en el transporte de mRNAa los compartimentos sinápticos y la traducción de proteínas.

A pesar del gran avance mostrado anteriormente, en el trabajo original no se realizó un análisis de Elementos Transponibles. La importancia de esto es que dichos elementos tienen el potencial de proporcionar secuencias reguladoras y/o de codificación de proteínas en un nuevo sitio de integración (Seberg & Petersen, 2009). Dependiendo de su secuencia de nucleótidos y del sitio de inserción genómica, un TE individual puede interrumpir la expresión de un gen, crear directa o indirectamente una modificación ventajosa para la expresión de un gen o no tener ninguna consecuencia inmediata (Elbarbary et al., 2016). Un ejemplo de ello serían los LTR que flanquean genomas ERV con frecuencia actúan como promotores y enhancer en la expresión de ERV en el cerebro de mamíferos, esta actividad podría implicar diferentes trastornos neuroinflamatorios, neurodegenerativos y neuropsiquiátricos, como la esclerosis múltiple, la esclerosis lateral amiotrófica y la esquizofrenia (Ferrari et al., 2021).

Antes de profundizar en los Elementos Transponibles, se procede a explicar las metodologías de RNA-Seq y ATAC-Seq, debido a que estas serán las utilizadas en este trabajo, como se indicó anteriormente.

Para explicar la técnica de RNA-Seq, primero debemos entender qué es el transcriptoma. Este se define como el conjunto completo de transcripciones en una célula. Entender el transcriptoma es indispensable para descifrar los elementos funcionales del genoma y descubrir los componentes moleculares de las células, así como para entender el desarrollo y las enfermedades. La finalidad de la transcriptómica es: clasificar todas las especies de transcritos, incluidos los mRNA y los RNA no codificantes (Wang et al., 2009).

En la Figura 3, se esquematiza un experimento de RNA-Seq. Este experimento parte desde una población de RNA (total o fraccionado), la que se transforma en una biblioteca de fragmentos de cDNA. Luego, se agregan adaptadores de secuenciación (azul) a cada fragmento de cDNA y se adquiere una secuencia corta ("read") de cada cDNA por medio de la tecnología de secuenciación de alto rendimiento. Las lecturas resultantes se alinean con el genoma o transcriptoma de referencia y se clasifican en tres tipos: lecturas exónicas, lecturas de unión y lecturas finales de poli(A). Estos tres tipos se utilizan para generar un perfil de expresión con resolución de bases para cada gen (Wang et al., 2009).

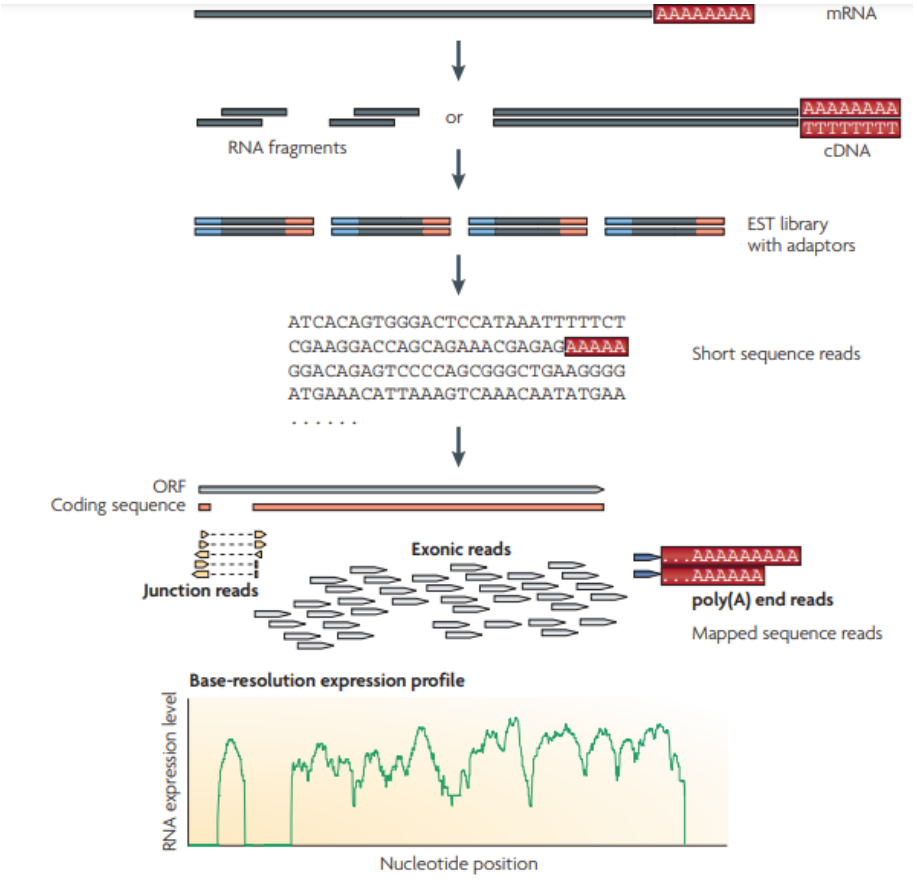

Figura 3. Esquema general de experimento de RNA-Seq. Los RNAs se convierten primero en una biblioteca de fragmentos de cDNA. Luego, se agregan adaptadores de secuenciación (azul y rojo) a cada fragmento de cDNA. Las lecturas de secuencias resultantes se alinean con el genoma o transcriptoma de referencia, lo que finalmente permite obtener un perfil de expresión con resolución a nivel de base. Figura tomada (Wang et al., 2009).

A pesar de todas las ventajas que entrega RNA-Seq, no permite dilucidar directamente elementos reguladores de la expresión génica. Los elementos reguladores se sitúan selectivamente en la cromatina accesible, que es importante para la regulación transcripcional. El mantenimiento de las configuraciones de cromatina accesibles requiere la unión de factores de transcripción para activar los genes objetivo. Por otro lado, la cromatina condensada restringe la unión de factores de transcripción y reguladores transcripcionales al promotor y / o enhancer, lo que resulta en el silenciamiento de genes. Además, la accesibilidad de la cromatina es una parte sustancial de la regulación epigenética, que se caracteriza por la metilación del DNA y la modificación de las histonas, como se mencionó anteriormente (Sun et al., 2019).

La accesibilidad de la cromatina se puede estudiar a nivel de genoma completo con ATAC-Seq *(Assay for Transposase-Accessible Chromatin using sequencing)*. La construcción de la biblioteca ATAC-Seq consta de tres pasos: preparación de núcleos, transposición y amplificación (Figura 4). En primer lugar, los tejidos o células para examen se suspenden en células individuales homogéneas, intactas, que posteriormente se incuban en el tampón de lisis para generar núcleos en bruto (Figura 4A). En segundo lugar, los núcleos resuspendidos se incuban con una versión mutada de la transposasa Tn5, la cual es hiperactiva y tiene como función cortar el DNA en regiones de cromatina abierta, y donde no haya proteínas unidas. Así, se producen fragmentos de DNA que pueden estar libres de nucleosomas y/o corresponder a regiones con uno o más nucleosomas (Figura 4B). Finalmente, estos fragmentos se amplifican y se secuencian (Figura 4C) (Sun et al., 2019).

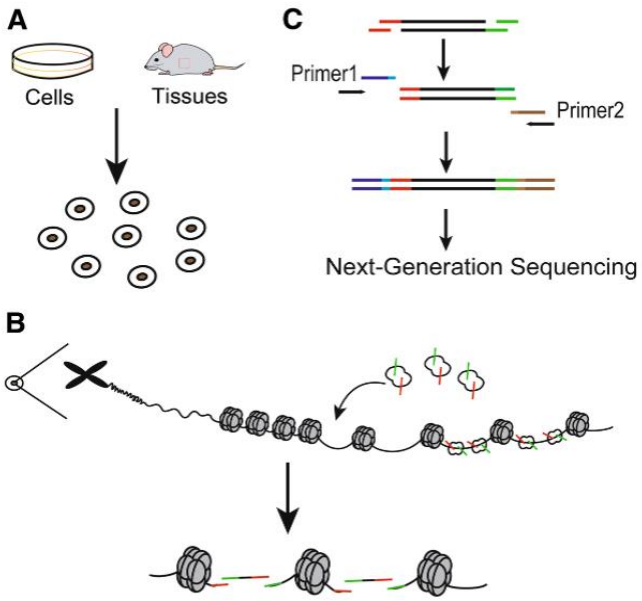

Figura 4: Procedimiento de ATAC-Seq. A) Preparación de los núcleos. B) Reacción de la transposasa: El símbolo verde representa el adaptador 1 de la transposasa Tn5, mientras que el símbolo rojo representa el adaptador 2 de la transposasa Tn5. C) Amplificación por PCR. Figura tomada de (Sun et al., 2019).

ATAC-Seq, entonces, nos puede ayudar a comprender cómo ocurre la regulación de la expresión génica en función de la accesibilidad de la cromatina, posiciones de nucleosomas y sitios de unión de factores de transcripción en todo el genoma. Esta información puede ayudar a entender la red de factores de transcripción relevantes y los mecanismos de regulación estructural de la cromatina que conducen los programas de expresión génica (Sun et al., 2019).

Hoy en día, es aceptado que los genomas eucariotas se componen en su mayoría de secuencias de DNA repetitivo intercalado. Por ejemplo, Britten & Kohne (1968) menciona: "*Un estudio de varias especies indica que las secuencias repetidas se dan amplia y probablemente de forma universal en el DNA de los organismos superiores*". La secuenciación de DNA a gran escala demuestra que del DNA repetitivo (en su mayoría) proviene de la actividad de los Elementos Transponibles (*Transposable Elements*, TEs), las cuales son secuencias que puede moverse de manera autosuficiente dentro del genoma (Feschotte, 2008). Los TEs utilizan diversos métodos replicativos, están los que involucran intermediarios de RNA,

llamados retrotransposones, o los que implican intermediarios de DNA. A estos últimos se les denomina transposones de DNA (Seberg & Petersen, 2009). El mecanismo de transposición de la clase I se denomina comúnmente "copiar y pegar", ya que, una vez transcrito un TE de esta clase, puede ocurrir una retrotranscripción mediada por una transcriptasa inversa (RT). El producto de este proceso es una copia nueva del TE, y así el retrotransposon original continua *in situ* en el lugar que se transcribe, mientras que la copia nueva se incorpora en una nueva localización genómica. Por su parte, el mecanismo de la clase II se denomina como "cortar y pegar", ya que una vez expresadas las transposasas, estas cortan del genoma el segmento de DNA original que codificó para ellas, y lo insertan en otra ubicación (Seberg & Petersen, 2009). A su vez, los retrotransposones se pueden clasificar en 3 categorías principales, las cuales son los retrotransposones de repetición terminal larga (LTR, Long Terminal Repeat), los elementos intercalados largos (LINE, Long Interspersed Nuclear Element) y los elementos intercalados cortos (SINE, Short Interspersed Nuclear Element)(Elbarbary et al., 2016). Debido a la movilización de "copiar y pegar" de los retrotransposones, estos se han acumulado en mayor proporción en los genomas eucariontes (Elbarbary et al., 2016). Los estudios empíricos y teóricos manifiestan que los TEs son denominados parásitos genómicos, debido a su capacidad de replicarse más rápido que el huésped que los porta. Brookfield (2005) menciona: "*Debido a que las secuencias de DNA móviles se replican durante la transposición y pueden dañar a sus anfitriones, pueden verse como secuencias de DNA egoístas*". Los TEs son una de las principales fuentes de mutación en el genoma, y también son responsables de provocar cambios en la expresión de los genes (Britten, 1996). El impacto a largo plazo de estos procesos es que los TEs se han insertado en regiones de numerosos genes eucariontes, impactando el control transcripcional de estos genes específicos (Britten, 1996). Una gran cantidad de estudios han demostrado que los TE pueden contribuir en la regulación de la expresión de genes cercanos, tanto a nivel transcripcional como post-transcripcional (Feschotte, 2008).

Los TEs proporcionan secuencias reguladoras que controlan la expresión de los genes del huésped. Una comparación de las secuencias del genoma humano y de

ratón con respecto a la ubicación de las secuencias reguladoras derivadas de TEs sugiere que este es realmente el caso. En consecuencia, la acción de los TEs puede proveer una maquinaria específica que fomente la diversificación reguladora de los linajes evolutivos del genoma del huésped (Mariño-Ramírez et al., 2005).

Se ha reportado actividad de TEs en varias especies, incluido los humanos (Huang et al., 2012). Las inserciones provocadas por los TEs dan como resultado impactos a niveles estructurales y funcionales en los genes y genomas. Esto se traduce en un valioso rol en la evolución del genoma (Huang et al., 2012). El efecto de las inserciones a nivel funcional de los genes puede ocurrir incluso cuando estos TEs se insertan fuera de las regiones codificantes, un ejemplo de esto sería cuando hay una inserción intrónica, esta causa transcripciones truncadas y no funcionales. (Huang et al., 2012).

Prácticamente la mitad del genoma humano y hasta el 40% del genoma del ratón están constituidos por secuencias de DNA repetitivas. En su mayoría son retrotransposones pertenecientes a las familias SINE y LINE, y estos elementos normalmente son reprimidos por mecanismos epigenéticos (Estécio et al., 2012). Los LINEs y SINEs son capaces de definir el límite entre la heterocromatina y la eucromatina. Por ejemplo, un elemento B2 de ratón actúa como elemento de restricción para impedir que la heterocromatina situada en *cis* silencie la expresión del desarrollo de los cinco genes, esta está situada en el locus de la hormona del crecimiento del ratón (Elbarbary et al., 2016). Respecto a los SINE, Ichiyanagi (2013) dice: "*Es evidente que las SINEs y otros retrotransposones han moldeado los genomas de los huéspedes cambiando las secuencias genómicas, los patrones de empalme, las redes reguladoras, las redes de cromatina y el programa de desarrollo*".

Los retrotransposones ubicados antes de los genes codificadores de proteínas son capaces de actuar como promotores ya que se han predicho sitios de unión para una gran cantidad de factores de transcripción en los SINEs (Figura 5). En cambio, no es evidente si la mayoría de los sitios de unión a factores de transcripción insertos en SINEs intervienen para modular la transcripción de genes (Elbarbary et al.,

2016). Los LINEs y SINEs igualmente tienen la capacidad de actuar como un nuevo sitio de inicio de la transcripción (TSS) (Figura 5). Se estima que entre el 6 y el 30% de los transcritos humanos y de ratón emplean TSSs relacionados a secuencias repetitivas (Elbarbary et al., 2016).

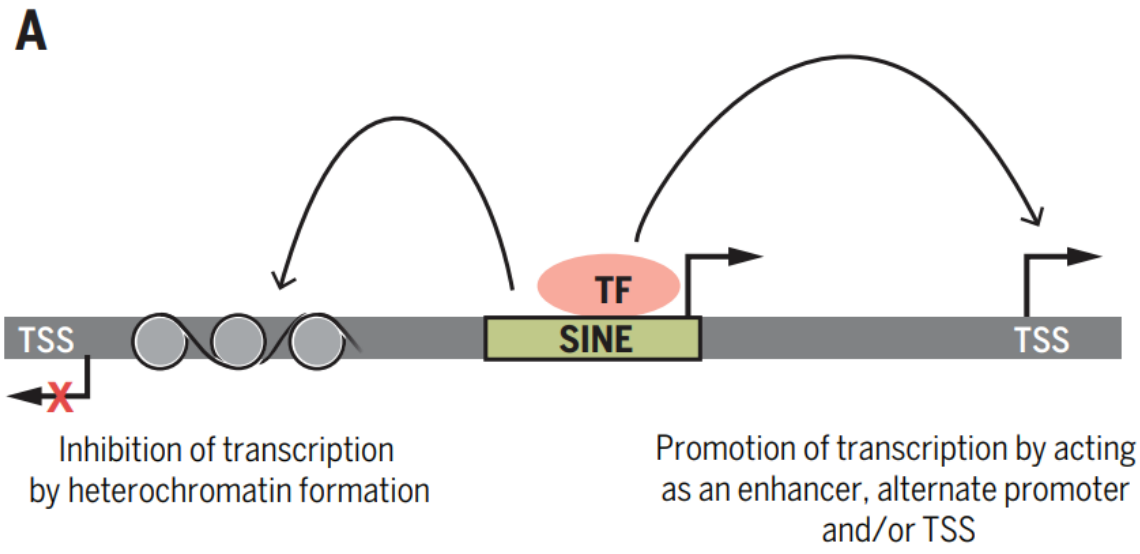

Figura 5: Regulación génica mediada por LINE y SINE. (A) Los SINEs (y LINEs) pueden promover o inhibir la transcripción de genes cercanos. Figura tomada de (Elbarbary et al., 2016)

Los elementos repetitivos Alu tienen la capacidad de insertarse en los RNA mensajeros maduros por medio de un proceso inducido por el splicing denominado exonización (creación de un nuevo exón como resultado de una inserción dentro de un intrón) (Lev-Maor, 2003). A partir de esto, a los TEs ya no se les considera como elementos "egoístas", ya que se incorporan a la maquinaria genómica del hospedador (Carmona, 2013). En base a esto Sela (2007) menciona: "*Los exones de Alu empalmados alternativamente enriquecen así el transcriptoma, la capacidad de codificación y la versatilidad reguladora de los genomas de primates con nuevas isoformas, sin comprometer la integridad y el repertorio original del transcriptoma y su proteoma resultante*".

A pesar de todo lo que se conoce actualmente sobre TEs, estos son rutinariamente desechados en estudios que utilizan datos de secuenciación de gran escala, como

RNA-Seq y ATAC-Seq. Estimar el origen de expresión de TEs utilizando solo RNA-Seq es complicado por las lecturas de mapeo múltiple provenientes de sus secuencias repetitivas. Abordar el origen genómico de los TE expresados podría ayudar aún más a comprender el papel que podrían tener los TE en la célula. Software como TEcandidates, basado en el ensamblaje de transcriptomas de novo para evaluar las instancias de TE que se expresan, junto con su ubicación (Valdebenito-Maturana & Riadi, 2018), y SQuIRE (Yang et al., 2019) son 2 avances recientes para resolver este problema. Así, nos proporcionan información de la expresión de TEs específica a nivel de locus. Datos recientemente publicados por el grupo de mi tutor indican que TEcandidates posee una alta selectividad (mayor tasa de verdaderos positivos en sus resultados) pero una baja sensibilidad (pocos resultados generados), mientras que SQuIRE tiene una alta sensibilidad y baja selectividad. La utilización del umbral de puntuación máximo (100) en SQuIRE produce una precisión del 56,1%, esto cambia cuando se utiliza el enfoque SQuIRE+TEcandidatos, en donde se observa una mejora considerable en la precisión de la predicción de los TEs (Valdebenito-Maturana et al., 2021). Por otro lado, ATAC-Seq nos puede ayudar a entender el entorno de cromatina, y así identificar regiones permisivas para la transcripción (eucromatina) y regiones no permisivas para la transcripción (heterocromatina). Con esta información, se puede estimar de mejor manera los falsos positivos de cada programa. Esto se puede evaluar específicamente confirmando el locus de TE expresado con el estado de cromatina en esa región. Por ejemplo, si un programa indica un locus como expresado, pero la cromatina está cerrada, eso sería un error de tipo falso positivo del programa

Así entonces, el enfoque multiómico de, trabajar con los datos de RNA-Seq y ATAC-Seq nos proporcionaría una forma de validar los TEs candidatos (generados a partir de los software SQuIRE y TEcandidates). Con lo ya mencionado anteriormente se plantea la pregunta de ***si los TEs se expresan diferencialmente en la formación de memoria y el recuerdo en el ratón***. En la siguiente página se expone la hipóteiss y objetivos de este trabajo.

Hipótesis

Los Elementos Transponibles participan como agentes reguladores en la formación de la memoria y el recuerdo en ratón.

Objetivo general

Identificar qué elementos transponibles se expresan en la formación de la memoria y el recuerdo en ratón con un enfoque multiómico.

Objetivos específicos

1. Analizar datos de RNA-Seq para predecir TEs expresados.
2. Analizar y procesar datos de ATAC-Seq para identificar zonas enriquecidas (heterocromatina y eucromatina).
3. Confirmar la expresión de TEs mediante la correlación de TEs predichos con las zonas enriquecidas de ATAC-Seq para identificar TEs expresados en regiones de eucromatina.
4. Analizar expresión de TEs a lo largo del proceso de formación de memoria.

Para poder llevar a cabo estos objetivos, a continuación se procede a describir los aspectos metodológicos de este trabajo.

Materiales y Métodos

Materiales

Software

- **BEDTools 2.30** (Quinlan & Hall, 2010)**:** permite intersectar, combinar, contar y complementar intervalos genómicos de varios archivos en formatos de archivos genómicos ampliamente utilizados como BAM, BED, GFF / GTF y VCF.
- **Bowtie 2 2.4.2** (Langmead & Salzberg, 2012)**:** herramienta para alinear las lecturas (reads) de secuenciación con secuencias de referencia (genoma).

- **SAMtools 1.12** (Li et al., 2009)**:** software para el parseo y manipulación de alineamientos en formato SAM/BAM.
- **Trinity 2.12** (Grabherr et al., 2011)**:** Combinación de tres software independientes (Inchworm, Chrysalis and Butterfly) para la reconstrucción de transcriptomas *de novo* con datos de RNA-seq.
- **STAR 2.8** (Dobin et al., 2013)**:** software de alineamiento de datos de RNA-seq.
- **R** (R Core Team, 2021)**:** lenguaje de programación de análisis estadístico.
- **Python 2.7:** lenguaje de programación interpretado.
- **DESeq2** (Love et al., 2014)**:** Paquete de R para realizar análisis de expresión diferencial a partir de una matriz de conteos.
- **TEcandidates** (Valdebenito-Maturana & Riadi, 2018)**:** pipeline que permite estima el origen de expresión de TEs.
- **SQuIRE** (Yang et al., 2019): pipeline que permite un análisis cuantitativo y locus-específico de la expresión de TEs a partir de datos de RNA-seq.
- **MACS3:** software para identificar regiones enriquecidas desde datos de ChIP-Seq y/o ATAC-Seq.
- **HMMRATAC** (Evan D. Tarbell and Tao Liu, 2019): herramienta de análisis ATAC-seq con un enfoque de aprendizaje automático semi-supervisado.

Hardware

Este proyecto se llevará a cabo en el computador Exxact del Dr. Gonzalo Riadi. Este computador cuenta con 2 CPUs Intel Xeon E7-8867 v3 @ 2.50GHz, con 64 cores cada una, 128 procesadores en total, y 256GB RAM. Además, posee una capacidad de almacenamiento de 40 TB.

Datos

El conjunto de datos a utilizar en este trabajo está públicamente disponible en la base de datos GEO, número de acceso GSE152956. Los datos disponibles son tanto de RNA-Seq, como de ATAC-Seq, provenientes del modelo de ratón descrito en la introducción. Una descripción detalla de las muestras se indica en las tablas 1

e 2. Para cada tipo de secuenciación, se generaron datos en las condiciones Basal, Early, Late y Reactivate.

Experiment Title	Sample Accession	Número de Reads	Largo de Read
RNA_Reactivated_rep3	SRS6866328	51.140.582	40
RNA_Reactivated_rep2	SRS6866327	52.002.939	40
RNA_Reactivated_rep1	SRS6866326	53.855.115	40
RNA_Late_rep3	SRS6866325	38.375.401	40
RNA_Late_rep2	SRS6866324	31.533.072	40
RNA_Late_rep1	SRS6866323	34.753.413	40
RNA_Early_rep3	SRS6866322	37.793.778	40
RNA_Early_rep1	SRS6866320	32.895.831	40
RNA_Basal_rep3	SRS6866319	52.082.682	40
RNA_Basal_rep2	SRS6866318	29.853.125	40
RNA_Early_rep2	SRS6866321	49.675.366	40
RNA_Basal_rep1	SRS6866317	26.052.676	40

Tabla 1. Se exponen los 12 conjuntos de datos que contiene cada muestra de RNA-Seq, como título, accession, cantidad de reads, y. largo de reads.

Experiment Title	Sample Accession	Número de Reads	Largo de Read
ATAC_Reactivated_rep3	SRS6885320	250.312.476	40
ATAC_Reactivated_rep2	SRS6885319	37.301.871	40
ATAC_Reactivated_rep1	SRS6885318	42.477.967	40
ATAC_Late_rep4	SRS6885317	63.682.914	40
ATAC_Late_rep3	SRS6885316	57.453.232	40
ATAC_Late_rep2	SRS6885315	69.096.408	40
ATAC_Late_rep1	SRS6885314	60.385.356	40
ATAC_Early_rep3	SRS6885313	51.072.530	40
ATAC_Early_rep2	SRS6885312	58.758.559	40
ATAC_Early_rep1	SRS6885311	72.971.590	40
ATAC_Basal_rep4	SRS6885310	42.863.552	40
ATAC_Basal_rep3	SRS6885309	43.928.770	40
ATAC_Basal_rep2	SRS6885308	55.950.657	40
ATAC_Basal_rep1	SRS6885307	41.436.445	40

Tabla 2. Se exponen los 14 datos que contiene cada muestra de ATAC-Seq, como título, accession, cantidad de reads, y. largo de reads.

El genoma de referencia a utilizar es *Mus musculus GRCm38.p6*, tomado desde University of California Santa Cruz (UCSC) Genome Browser (mm10).

Metodología

La metodología que se utilizó en este trabajo se resume en la Figura 6, exhibida en la siguiente página. A lo largo del texto a continuación, se describe específicamente la metodología asociada a cada objetivo.

Objetivo #1

En este proyecto se utilizaron 2 software para el análisis de TEs a partir de datos de RNA-Seq: TEcandidates (Valdebenito-Maturana & Riadi, 2018) y SQuIRE (Yang et al., 2019).

TEcandidates tiene por función predecir la posición en el genoma de los TEs que se están expresando a partir de datos de RNA-Seq. A partir del archivo de anotación (el cual contiene los TEs predichos) se puede proseguir con el análisis de expresión diferencial, pudiendo evaluar la participación que tienen los TEs en la regulación del genoma. TEcandidates tiene como dependencias BEDtools (Quinlan & Hall, 2010), Bowtie2 (Langmead & Salzberg, 2012), Bioperl (Stajich, 2002) y Trinity (Grabherr et al., 2011). Este software requiere 3 archivos de entrada: los reads de RNA-Seq, genoma de referencia y anotación de TEs. En la primera fase TEcandidates realiza un alineamiento de los reads de RNA-Seq contra el genoma de referencia con la finalidad de limitar los reads a TEs solamente. Luego Trinity utiliza los reads del

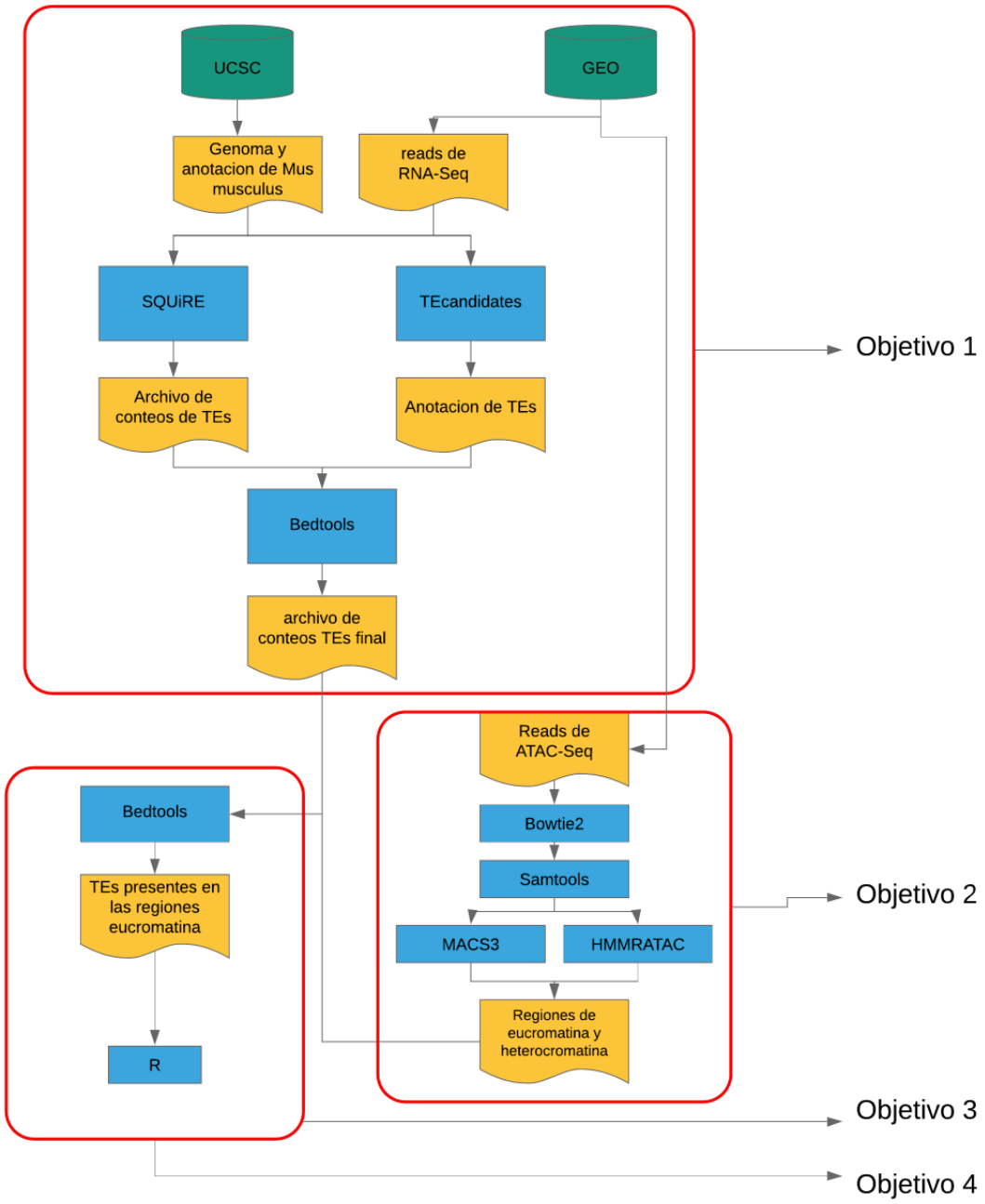

Figura 6: Diagrama de flujo general de la metodología de trabajo en base a la norma ANSI para la elaboración de diagramas de flujos. La base de datos de color verde hace referencia a University of California Santa Cruz (UCSC) Genome Browser y Gene Expression Omnibus (GEO), los símbolos de color amarillo indican los archivos utilizados en cada actividad (input/output). Las actividades/procesos denotados de color azul hace referencia a la utilización de los diversos software. Finalmente, los rectángulos de color rojo encapsulan a que objetivo corresponde esa parte de la metodología.

La principal limitación de TEcandidates es que no estima niveles de expresión de los TEs. Por esto, se utilizó adicionalmente SQuIRE. SQuIRE se compone de un set de herramientas con las cuales se evalúan los niveles de expresión de TEs a partir de datos de RNA-Seq. Este software tiene como dependencias BEDtools (Quinlan & Hall, 2010), STAR (Dobin et al., 2013), SAMtools (Li et al., 2009), StringTie (Pertea et al., 2015), DESeq2 (Love et al., 2014), R (Team, 2006) y Python 2.7.

Para este proyecto, se utilizaron las etapas de Preparación, Cuantificación y Análisis de SQuIRE. La etapa de Preparación se divide en 2 herramientas: Fetch y Clean. Fetch constó en la descarga de los archivos del genoma (mm10), y sus respectivas anotaciones de genes y TEs, desde el sitio web UCSC Genome Browser. Luego de esto, se realizó el indexado del genoma con STAR. Por su parte, Clean realizó un formateo como archivo BED de la información de anotación de los TEs de RepeatMasker. Para la etapa de Cuantificación se realizó el alineamiento de los datos de RNA-Seq usando STAR, generando como resultado un archivo BAM. A continuación, mediante un algoritmo propio de SQuIRE, se estimó la expresión de TEs. Al culminar esta etapa, se entregó una tabla de conteos de reads por genes y por TEs. Finalmente, en la etapa de análisis se realizó el análisis de expresión diferencial para los TEs y genes con el paquete DESeq2 del software R.

Ambos software (TEcandidates y SQuIRE) utilizaron los archivos de genoma de referencia y anotación descargados por la herramienta SQuIRE Fetch. Con esto se asegura la posibilidad de poder realizar análisis posteriores de intersección entre los resultados de ambas herramientas. Esto se realizó con el software BEDtools, utilizando como entrada el archivo que contiene los candidatos de TEs (uno de los archivos generados por TEcandidates) con el archivo de conteos generado con la herramienta de SQuIRE. Con esto se obtuvo un archivo de conteos que contemplan los TEs que se encuentran en ambos programas.

Objetivo #2

Para el análisis de los datos de ATAC-Seq, se procedió a mapear los reads en el genoma de referencia de *Mus musculus* con el software Bowtie 2. Una vez realizado el mapeo, se generaron los archivos SAM, los cuales fueron formateados con la herramienta SAMtools. Esta herramienta contiene la opción la opción *view* la cual permitió convertir de archivo SAM a BAM. Luego, se utilizó *sort*, la cual ordena según coordenadas el archivo BAM. Por último, con la opción *index*, generó un índice de dichos archivos BAM, lo que permite procesarlos más rápido. Estos archivos BAM se utilizaron en el software MACS3 y HMMRATAC. Con esto se generaron los archivos que contienen las regiones de eucromatina. Finalmente, con el software BEDtools, en específico la opción groupby e intersect, se procedió agrupar por replicas para así obtener 1 archivo por condición, luego estos se intersecan entre el resultados agrupado con MACS3 y HMMRATAC.

Objetivo #3

Con el archivo de conteos que almacena los datos de TEs presentes tanto en TEcandidates como SQuIRE (Objetivo 1) y el archivo que contiene las regiones eucromatina y heterocromatina (Objetivo 2), se utilizó BEDtools para identificar el entorno de cromatina de los TEs. Con esto se filtraron aún más los resultados, ya que, los TEs en regiones de heterocromatina, se consideraron como Falsos Positivos.

Objetivo #4

Con el archivo final generado (Objetivo 3) se procedió a realizar el análisis de expresión diferencial de los datos de RNA-Seq y ATAC-Seq con el software R, en particular la librería DESeq2 con las métricas de que el log2foldchange fuera mayor a 1 y menor a -1. además, que FDR o log10(pvalue) fuera menor a 0.05, luego se procedió a validar la información mediante una búsqueda bibliográfica de la función reguladora de las regiones obtenidas del objetivo 3. Además, se procedió a confirmar resultados obtenidos con la investigación de Marco y colaboradores del año 2020.

Resultados

1.Analisis de los datos de RNA-Seq para la predicción de los TEs expresados.

A continuación, se exponen los resultados obtenidos por los software TEcandidates y SQuIRE de la cantidad de TEs predichos para cada condición.

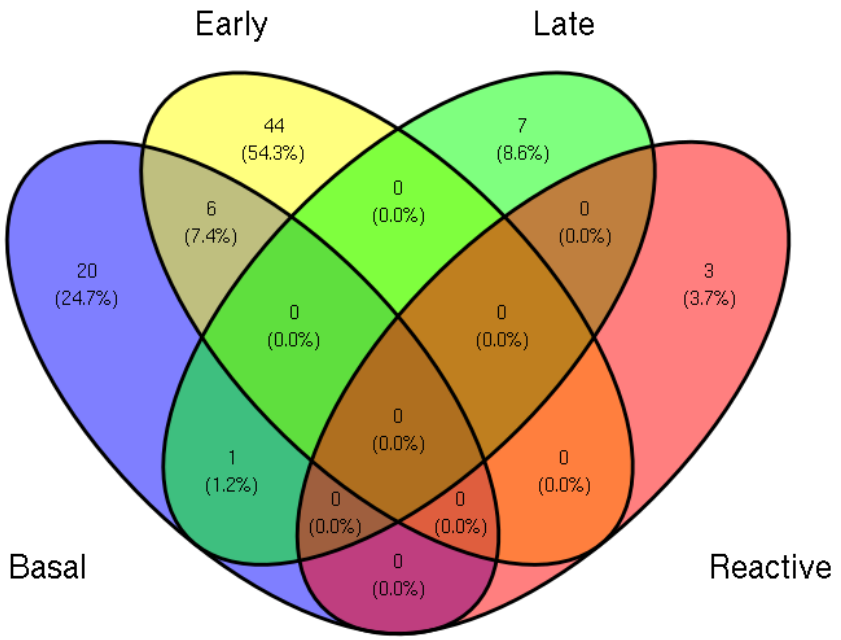

Figura 7: Diagrama de Venn con la cantidad de TEs predichos por el software TEcandidates. En el diagrama se aprecia la cantidad de TEs predichos por el software TEcandidates para la condición Basal (azul), Early (amarillo), Late (verde), recall o reactive (rojo) y sus respectivas intersecciones.

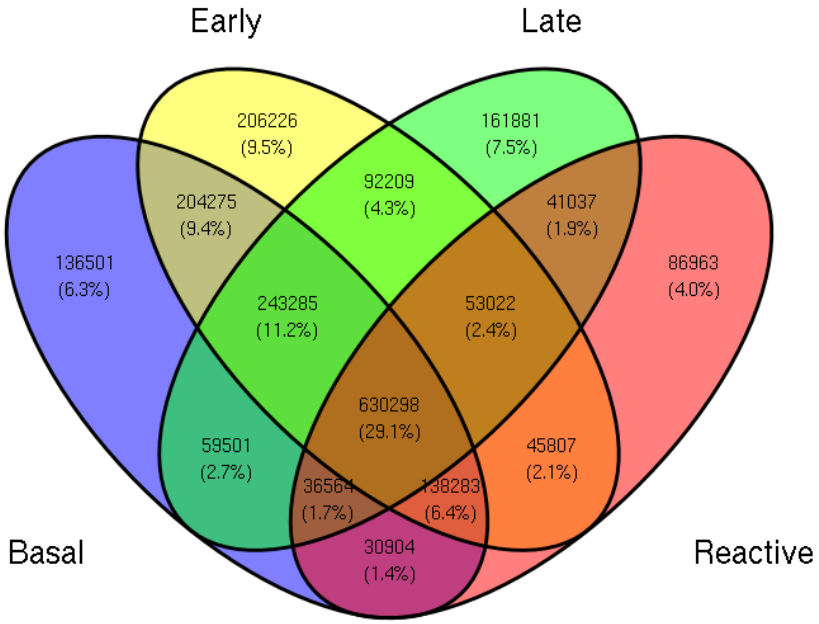

Figura 8: Diagrama de Venn con la cantidad de TEs predichos por el software SQuIRE. En el diagrama se aprecia la cantidad de TEs predichos por el software SQuIRE para la condición Basal (azul), Early (amarillo), Late (verde), recall o reactive (rojo) y sus respectivas intersecciones.

Luego, se realizó la intersección de los resultados obtenidos por los software TEcandidates y SQuIRE para cada condición.

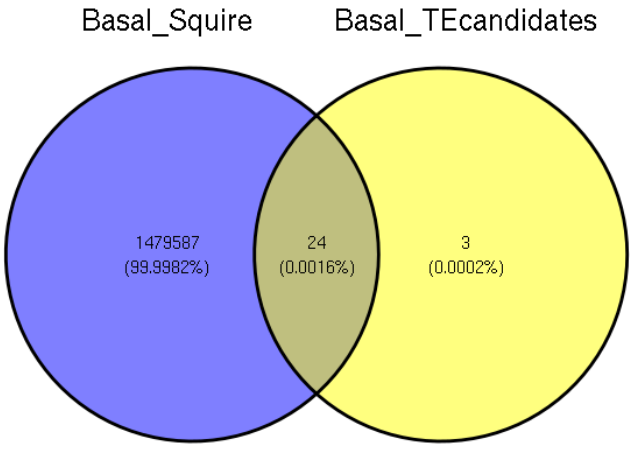

Figura 9: Diagrama de Venn con la cantidad de TEs predichos por el software SQuIRE y TEcandidates para la condición Basal. En el diagrama se aprecia la cantidad de TEs predichos por el software SQuIRE (azul) y TEcandidates (amarillo) para la condición Basal y la cantidad presente en ambos software.

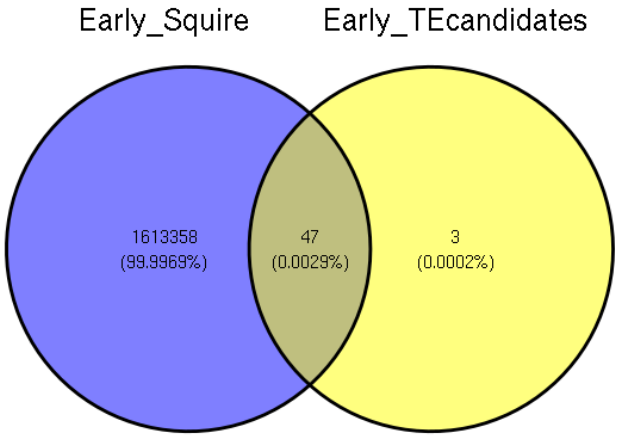

Figura 10: Diagrama de Venn con la cantidad de TEs predichos por el software SQuIRE y TEcandidates para la condición Early. En el diagrama se aprecia la cantidad de TEs predichos por el software SQuIRE (azul) y TEcandidates (amarillo) para la condición Early y la cantidad presente en ambos software.

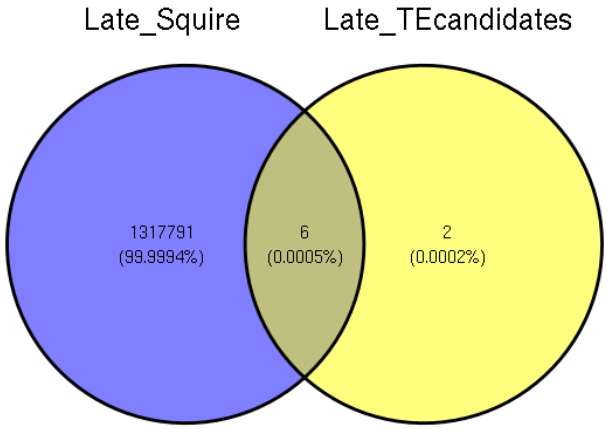

Figura 11: Diagrama de Venn con la cantidad de TEs predichos por el software SQuIRE y TEcandidates para la condición Late. En el diagrama se aprecia la cantidad de TEs predichos por el software SQuIRE (azul) y TEcandidates (amarillo) para la condición Late y la cantidad presente en ambos software.

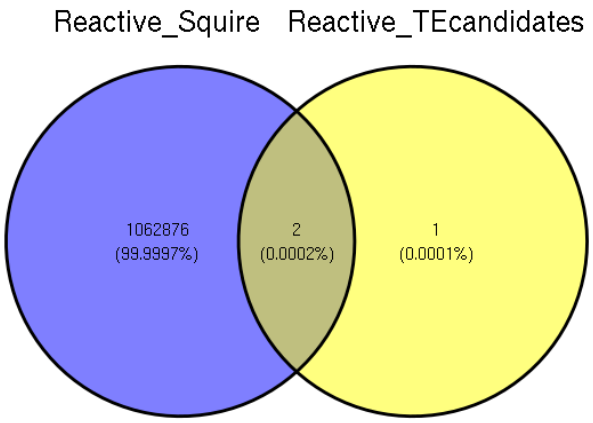

Figura 12: Diagrama de Venn con la cantidad de TEs predichos por el software SQuIRE y TEcandidates para la condición Reactive. En el diagrama se aprecia la cantidad de TEs predichos por el software SQuIRE (azul) y TEcandidates (amarillo) para la condición Reactive y la cantidad presente en ambos software.

Posteriormente, se realizó la intersección de los resultados obtenidos por los software TEcandidates y SQuIRE a lo largo de todas las condiciones.

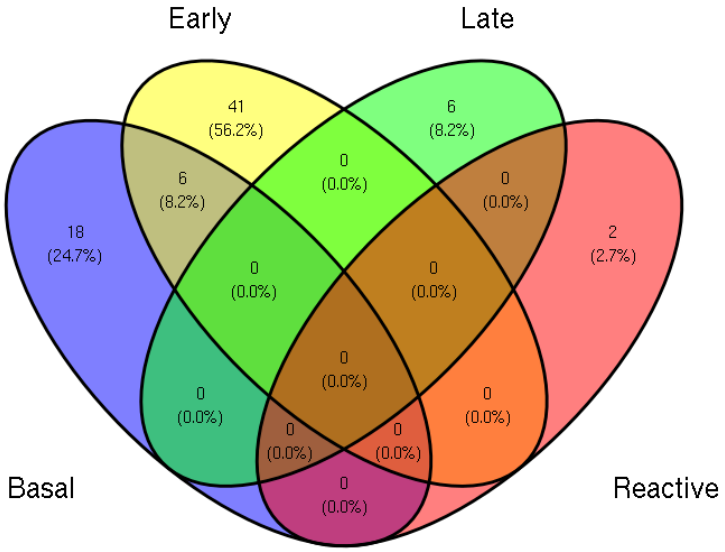

Figura 13: Diagrama de Venn con la cantidad de TEs predichos por el software SQuIRE y TEcandidates. En el diagrama se aprecia la cantidad de TEs predichos por el software SQuIRE y TEcandidates para la condición Basal (azul), Early (amarillo), Late (verde), recall o reactive (rojo) y sus respectivas intersecciones.

Finalmente se obtienen set de TEs filtrados (TEcandidate+SQuIRE) a partir de los datos de RNA-Seq para cada condición (Basal=24, Early=47, Late=6 y Reactivated=2), con este análisis se logra una mayor certeza en la exactitud de la predicción de TEs gracias al benchmark de estas herramientas de prediccion.

2. Análisis de los datos de ATAC-Seq para la identificación de las zonas enriquecidas

Luego de alinear los datos de ATAC-Seq al genoma y procesar los archivos resultantes con SAMtools, se realizó el peak calling con MACS3 y HMMRATAC. A fin de entender cuánto del genoma corresponde a regiones accesible (eucromatina) en términos generales, a continuación, se exponen los resultados obtenidos por los software MACS3 y HMMRATAC para cada condición:

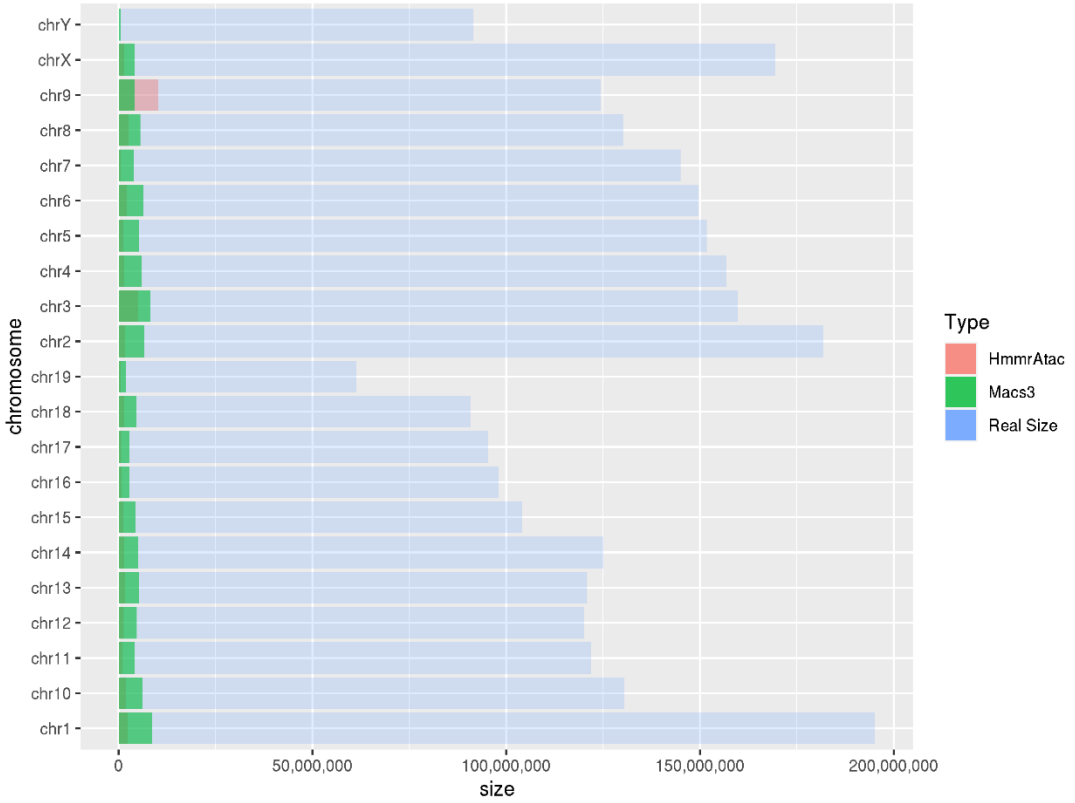

Figura 14: Grafica de barras con la cantidad total perteneciente a zonas enriquecidas de ATAC-Seq estimada por los software MACS3 y HMMRATAC para la condición Basal. El eje X hace referencia al tamaño, mientras que el eje Y representa a cada cromosoma de *Mus musculus*.

En la gráfica se aprecia la cantidad del genoma (dividido por cromosoma) perteneciente a zonas enriquecidas de ATAC-Seq estimadas por los software Macs3 (verde) y HmmrAtac (rojo), además se contrasta con el tamaño de referencia por cada cromosoma (azul).

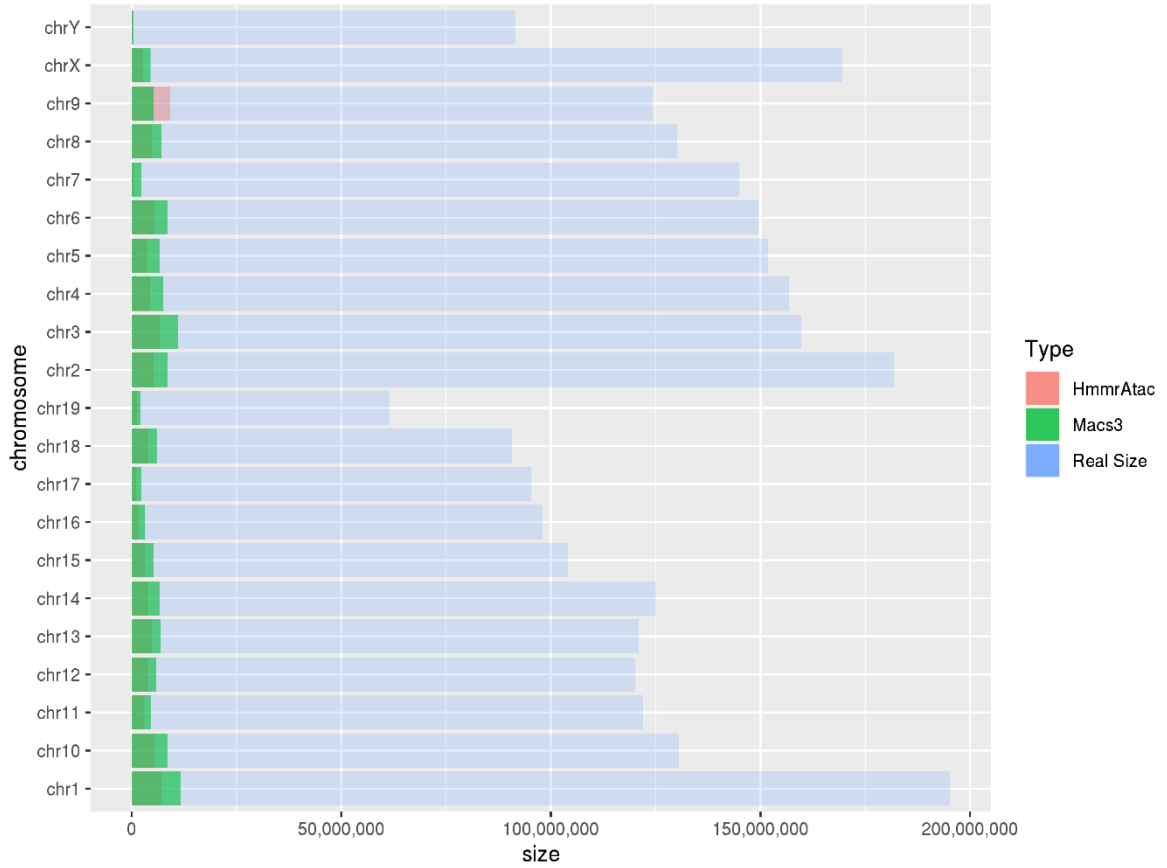

Figura 15: Grafica de barras con la cantidad total perteneciente a zonas enriquecidas de ATAC-Seq estimada por los software MACS3 y HMMRATAC para la condición Early. El eje X hace referencia al tamaño, mientras que el eje Y representa a cada cromosoma de mus musculus. En la gráfica se aprecia la cantidad del genoma (dividido por cromosoma) perteneciente a zonas enriquecidas de ATAC-Seq estimadas por los software Macs3 (verde) y HmmrAtac (rojo), además se contrasta con el tamaño de referencia por cada cromosoma (azul).

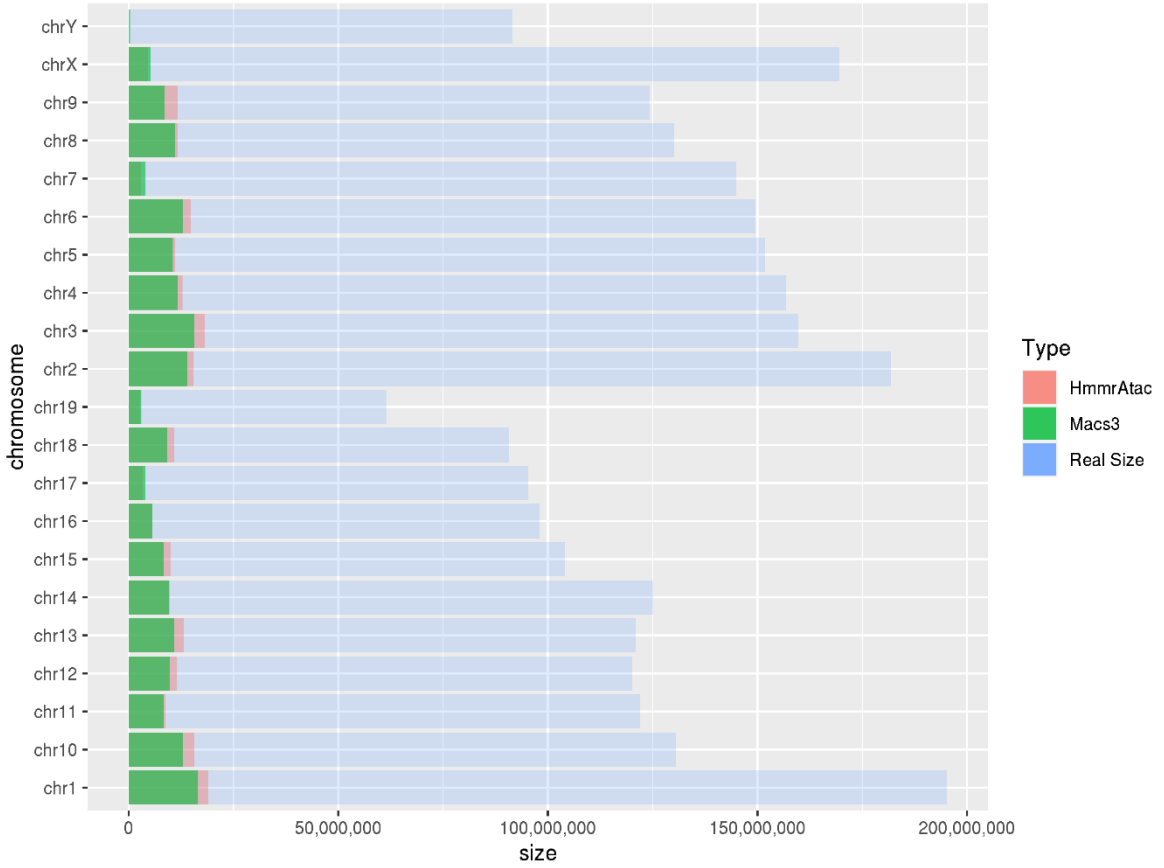

Figura 16: Grafica de barras con la cantidad total perteneciente a zonas enriquecidas de ATAC-Seq estimada por los software MACS3 y HMMRATAC para la condición Late. El eje X hace referencia al tamaño, mientras que el eje Y representa a cada cromosoma de mus musculus. En la gráfica se aprecia la cantidad del genoma (dividido por cromosoma) perteneciente a zonas enriquecidas de ATAC-Seq estimadas por los software Macs3 (verde) y HmmrAtac (rojo), además se contrasta con el tamaño de referencia por cada cromosoma (azul).

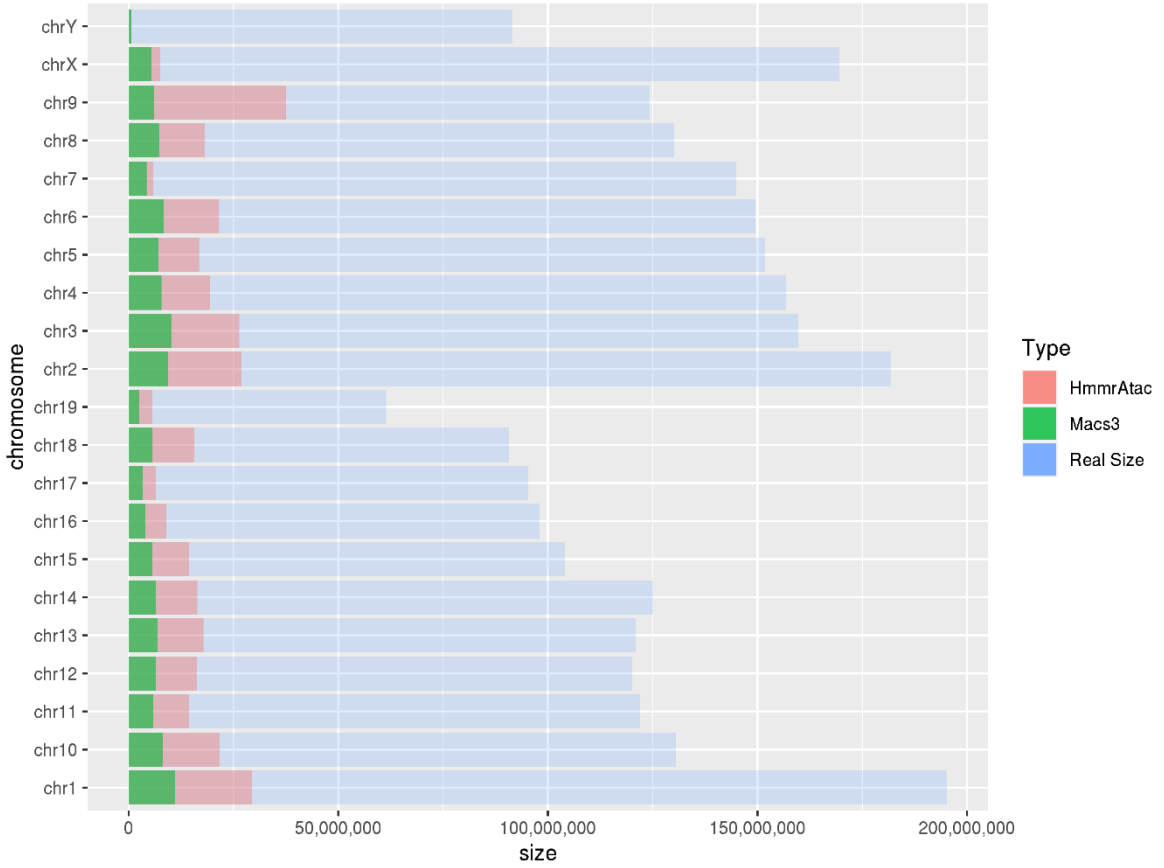

Figura 17: Grafica de barras con la cantidad total perteneciente a zonas enriquecidas de ATAC-Seq estimada por los software MACS3 y HMMRATAC para la condición Reactive. El eje X hace referencia al tamaño, mientras que el eje Y representa a cada cromosoma de mus musculus. En la gráfica se aprecia la cantidad del genoma (dividido por cromosoma) perteneciente a zonas enriquecidas de ATAC-Seq estimadas por los software Macs3 (verde) y HmmrAtac (rojo), además se contrasta con el tamaño de referencia por cada cromosoma (azul).

3. Confirmación de la expresión de TEs mediante la relación de TEs predichos a partir de los datos de RNA-Seq con las regiones de eucromatina identificadas mediante ATAC-Seq.

A continuación, se exponen los TEs predichos por el software SQuIRE, que se encuentran en las regiones de eucromatina (MACS3 y HMMRATAC). Esto se realizó con los software BEDtools y R.

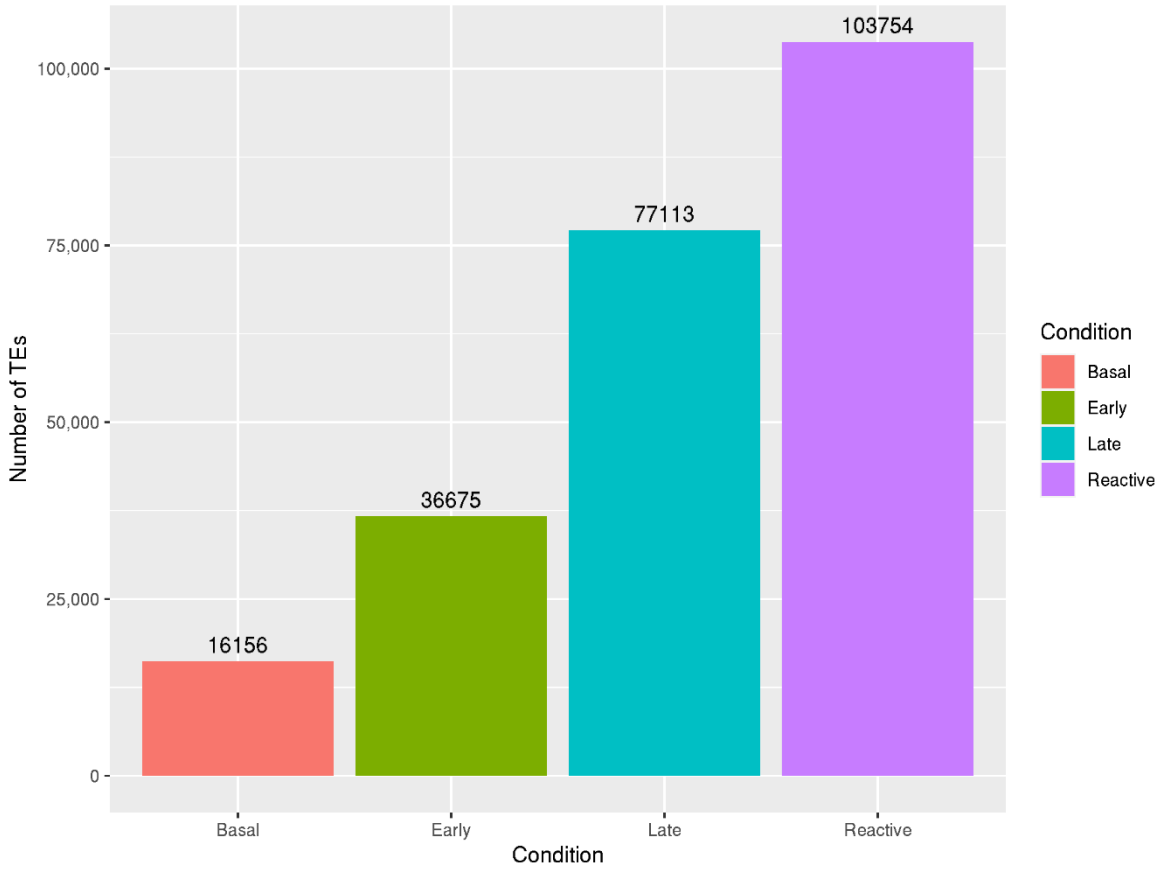

Figura 18: Grafica de barras con la cantidad total de TEs predichos por SQuIRE que se encuentran en zonas enriquecidas de ATAC-Seq estimada por los software MACS3 y HMMRATAC para la cada condición. La gráfica se divide para la condición Basal (Rojo), Early (Verde), Late (Celeste) y Reactivated (Purpura).

Luego, se exponen los TEs predichos por el software SQuIRE, que se encuentran en las regiones de eucromatina (MACS3 y HMMRATAC) divididos por el tipo de TE para cada una de las condiciones. Esto se realizó con los software BEDtools y R.

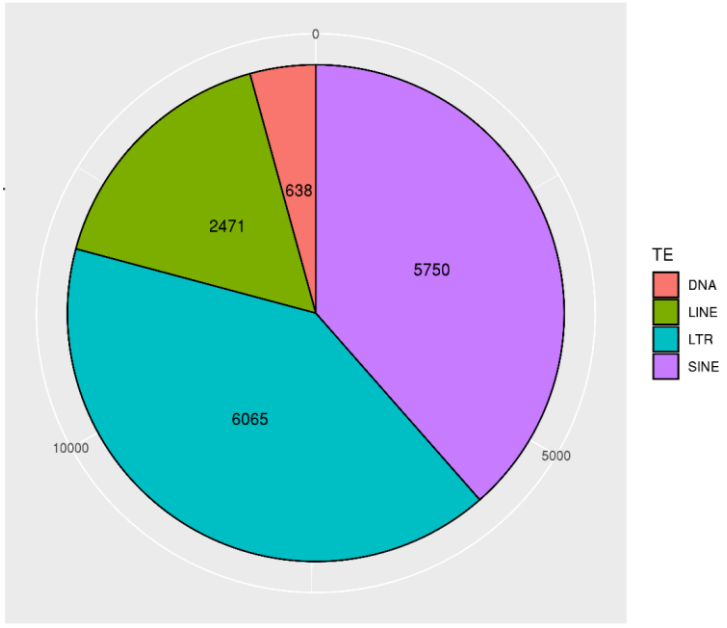

Figura 19: Grafica circular con la cantidad total de TEs predichos por SQuIRE que se encuentran en zonas enriquecidas de ATAC-Seq estimada por los software MACS3 y HMMRATAC para la condición Basal. La gráfica se divide por los tipos de TEs, los cuales son de DNA (rojo), LINE (verde), LTR (celeste) y SINE (purpura) para la condición Basal.

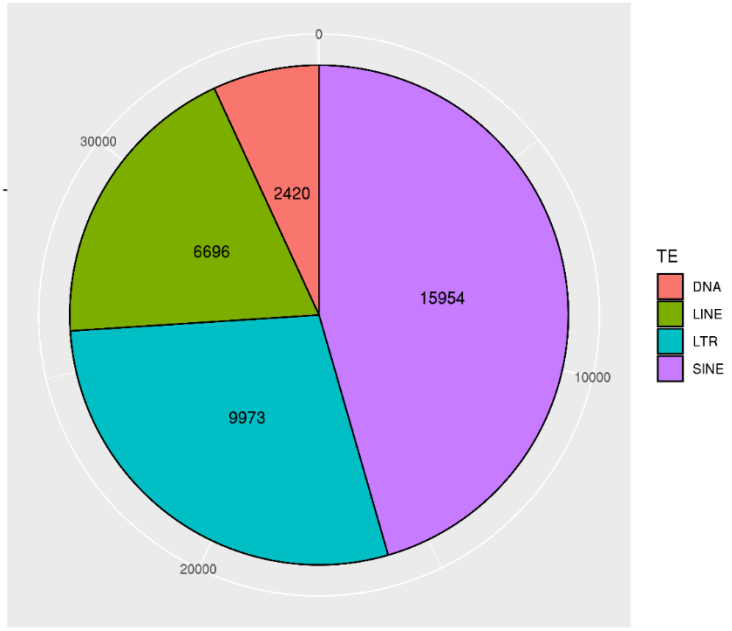

Figura 20: Grafica circular con la cantidad total de TEs predichos por SQuIRE que se encuentran en zonas enriquecidas de ATAC-Seq estimada por los software MACS3 y HMMRATAC para la condición Early. La gráfica se divide por los tipos de TEs, los cuales son de DNA (rojo), LINE (verde), LTR (celeste) y SINE (purpura) para la condición Early.

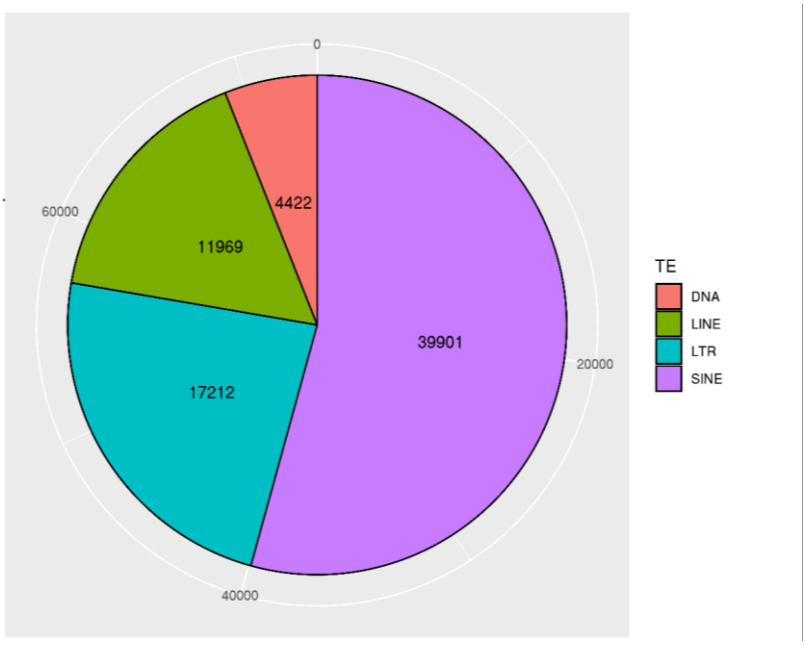

Figura 21: Grafica circular con la cantidad total de TEs predichos por SQuIRE que se encuentran en zonas enriquecidas de ATAC-Seq estimada por los software MACS3 y HMMRATAC para la condición Late. La gráfica se divide por los tipos de TEs, los cuales son de DNA (rojo), LINE (verde), LTR (celeste) y SINE (purpura) para la condición Late.

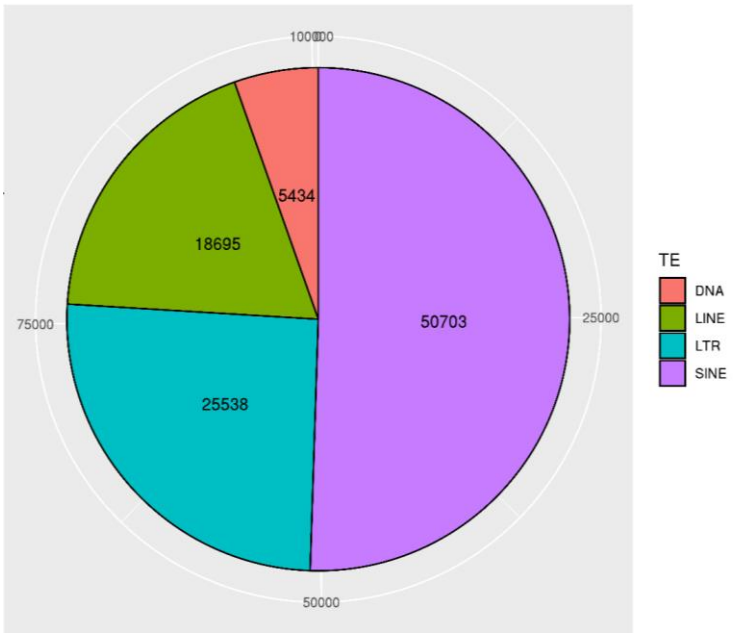

Figura 22: Grafica circular con la cantidad total de TEs predichos por SQuIRE que se encuentran en zonas enriquecidas de ATAC-Seq estimada por los software MACS3 y HMMRATAC para la condición Reactivated. La gráfica se divide por los tipos de TEs, los cuales son de DNA (rojo), LINE (verde), LTR (celeste) y SINE (purpura) para la condición Reactive.

Posteriormente, se realizó la intersección de los TEs predichos por TEcandidates con los resultados mostrados anteriormente. Para ello se utilizaron los software BEDtools y R.

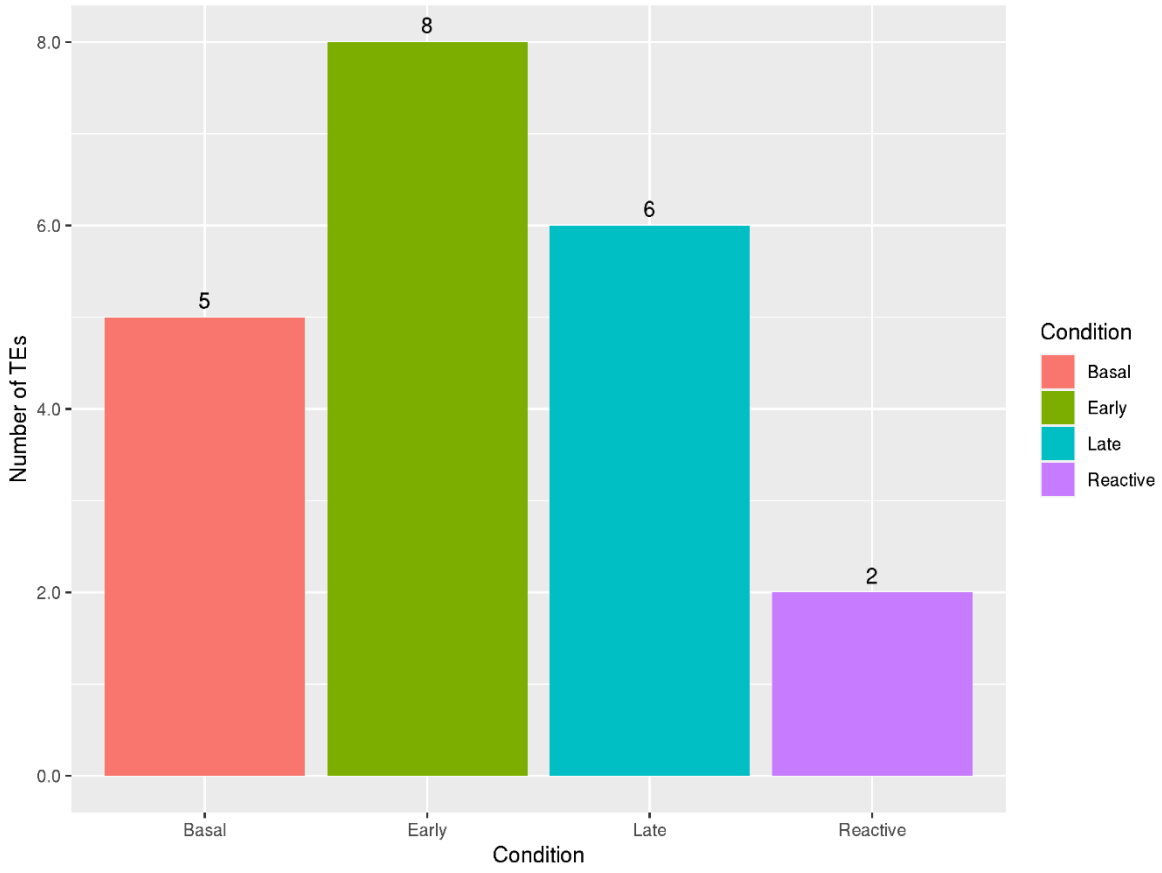

Figura 23: Grafica de barras con la cantidad total de TEs presentes en SQuIRE y TEcandidates que se encuentran en zonas enriquecidas de ATAC-Seq estimada por los software MACS3 y HMMRATAC para la cada condición. La gráfica se divide para la condición Basal (Rojo), Early (Verde), Late (Celeste) y Reactivated (Purpura).

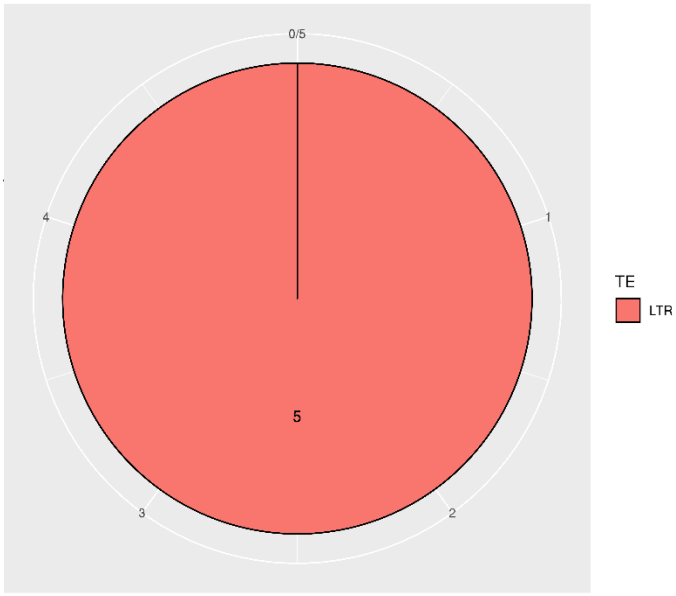

Figura 24: Grafica circular con la cantidad total de TEs presentes en SQuIRE y TEcandidates que se encuentran en zonas enriquecidas de ATAC-Seq estimada por los software MACS3 y HMMRATAC para la condición Basal. La gráfica muestra TEs del tipo LTR (rojo) para la condición Basal.

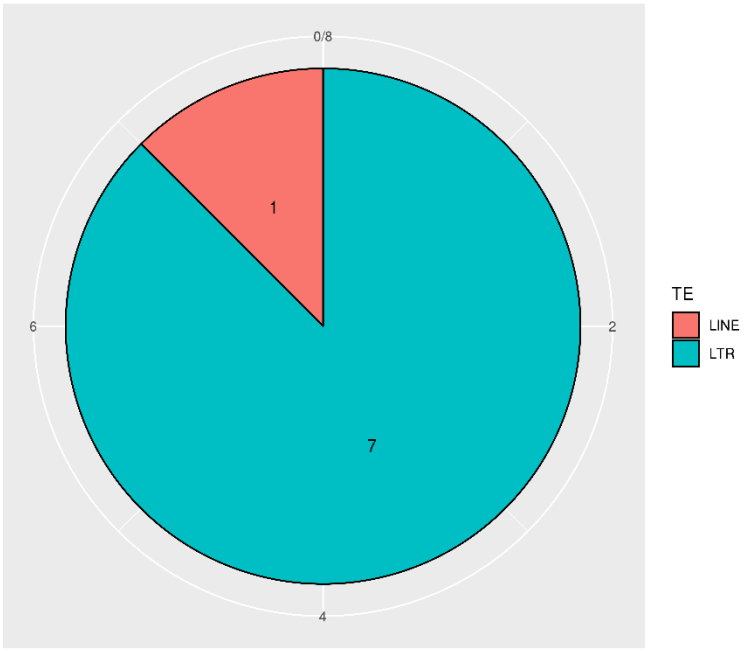

Figura 25: Grafica circular con la cantidad total de TEs presentes en SQuIRE y TEcandidates que se encuentran en zonas enriquecidas de ATAC-Seq estimada por los software MACS3 y HMMRATAC para la condición Early. La gráfica muestra TEs del tipo LTR (rojo) para la condición Early.

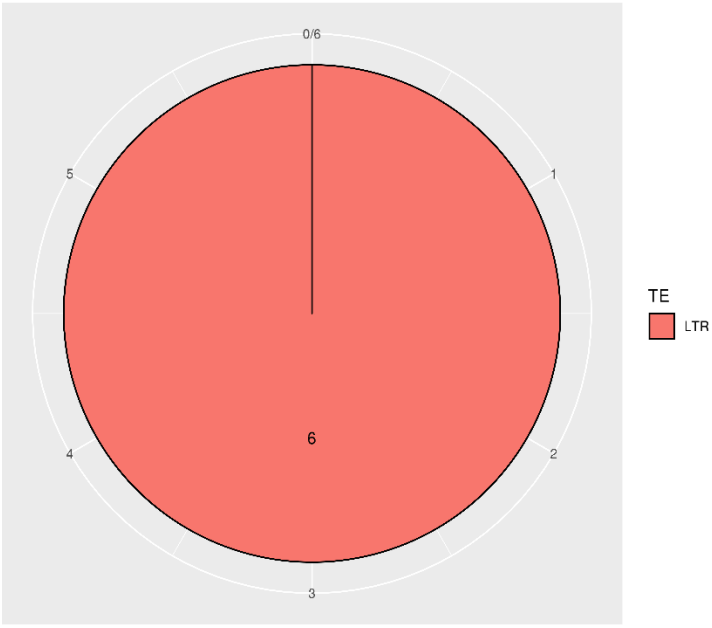

Figura 26: Grafica circular con la cantidad total de TEs presentes en SQuIRE y TEcandidates que se encuentran en zonas enriquecidas de ATAC-Seq estimada por los software MACS3 y HMMRATAC para la condición Late. La gráfica muestra TEs del tipo LTR (rojo) para la condición .Late.

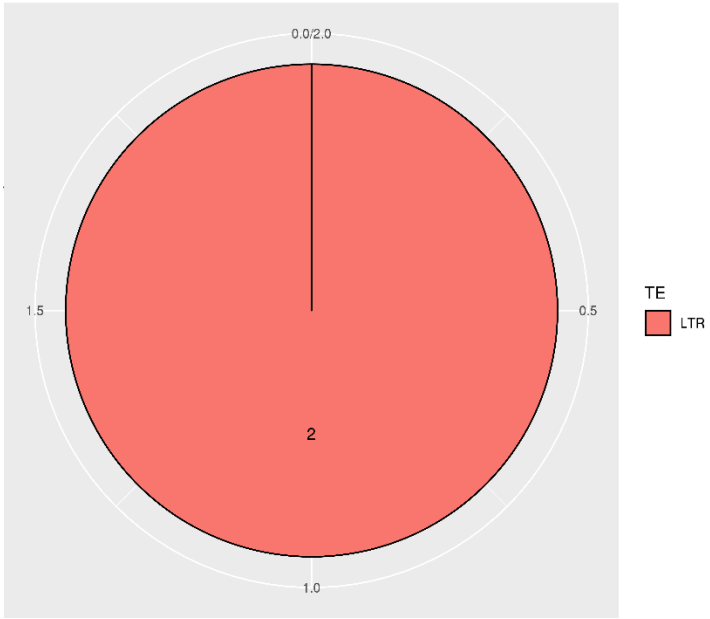

Figura 27: Grafica circular con la cantidad total de TEs presentes en SQuIRE y TEcandidates que se encuentran en zonas enriquecidas de ATAC-Seq estimada por los software MACS3 y HMMRATAC para la condición Reactivated. La gráfica se divide por los tipos de TEs, los cuales son de LINE (rojo) y LTR (celeste) para la condición Reactive.

Finalmente se obtiene el set de TEs que están en zonas de eucromatina y que fueron predichos SQuIRE y TEcandidates, con ello se asegura que los TEs predicho se encuentran en zonas de cromatina abierta.

4. Análisis de la expresión de TEs a lo largo del proceso de formación de memoria

A continuación, se presentan los volcano plots generados con los datos del inciso anterior (SQuIRE + TEcandidates + filtro eucromatina) comparando las condiciones Basal vs Early, Early vs Late y Late vs Reactive. Para ello se utilizaron los software BEDtools y R.

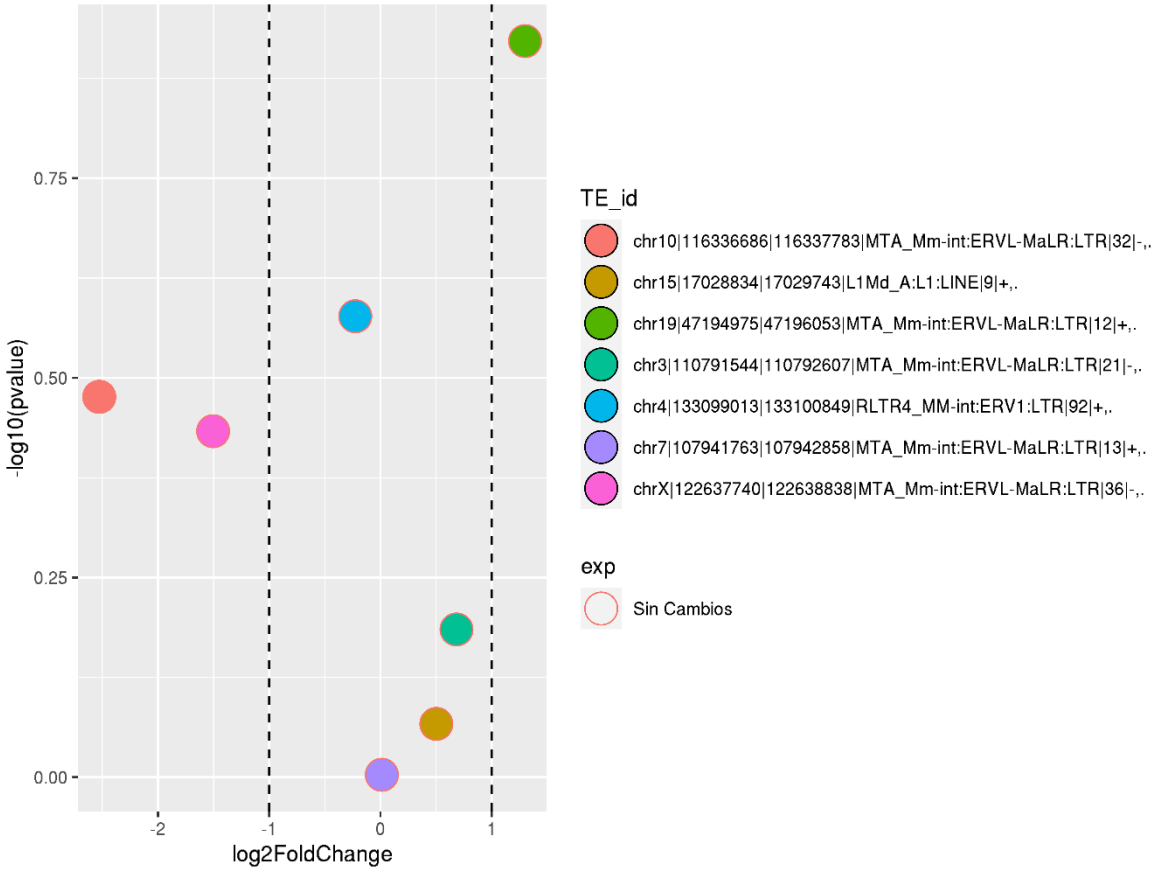

Figura 26: Volcano plot con los TEs (SQuIRE y TEcandidates) que se encuentran en zonas enriquecidas de ATAC-Seq (software MACS3 y HMMRATAC) para la condición Basal vs Early. En el eje Y esta el log10(pvalue), mientras que en el eje X está el log2foldchange, de borde rojo están los TEs que no presentan cambios.

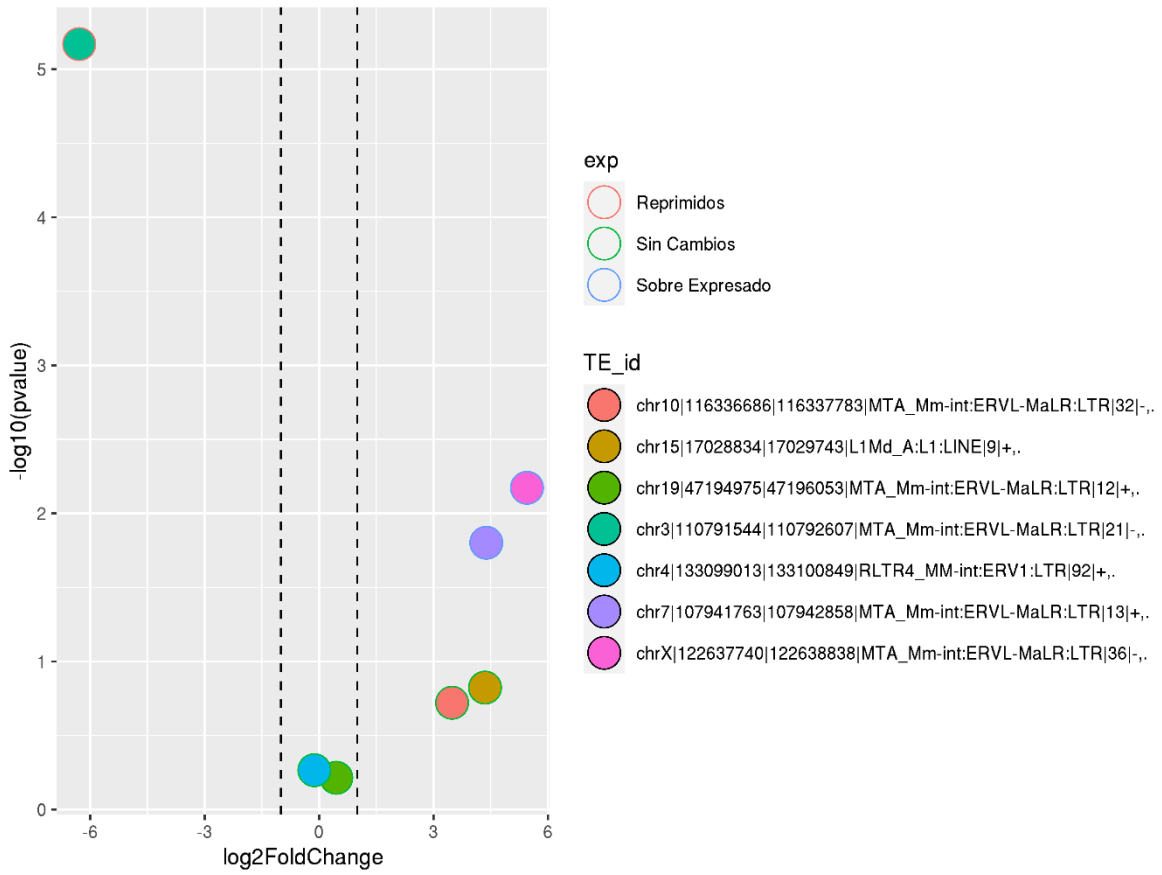

Figura 27: Volcano plot con los TEs (SQuIRE y TEcandidates) que se encuentran en zonas enriquecidas de ATAC-Seq (software MACS3 y HMMRATAC) para la condición Early vs Late. En el eje Y esta el log10(pvalue), mientras que en el eje X está el log2foldchange, de borde rojo estas los TEs sub expresados o reprimidos, de borde azul los sobre expresados y de borde verde los que no presentan cambios.

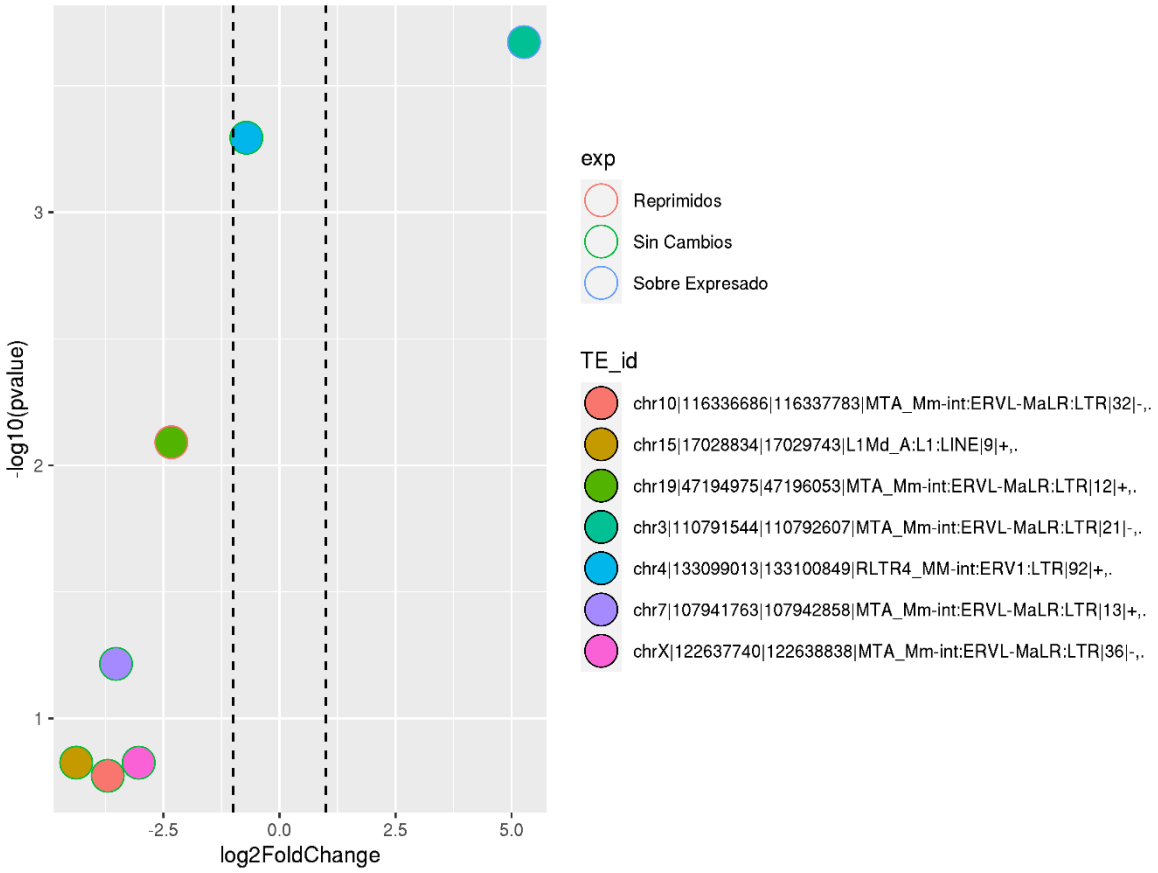

Figura 28: Volcano plot con los TEs (SQuIRE y TEcandidates) que se encuentran en zonas enriquecidas de ATAC-Seq (software MACS3 y HMMRATAC) para la condición Late vs Reactive. En el eje Y esta el log10(pvalue), mientras que en el eje X está el log2foldchange, de borde rojo estas los TEs sub expresados o reprimidos, de borde azul los sobre expresados y de borde verde los que no presentan cambios.

Luego, se realizó la relación de la expresión de genes con la expresión de TEs comparando las condiciones Basal vs Early, Early vs Late y Late vs Reactive. Para ello se utilizaron los software BEDtools y R.

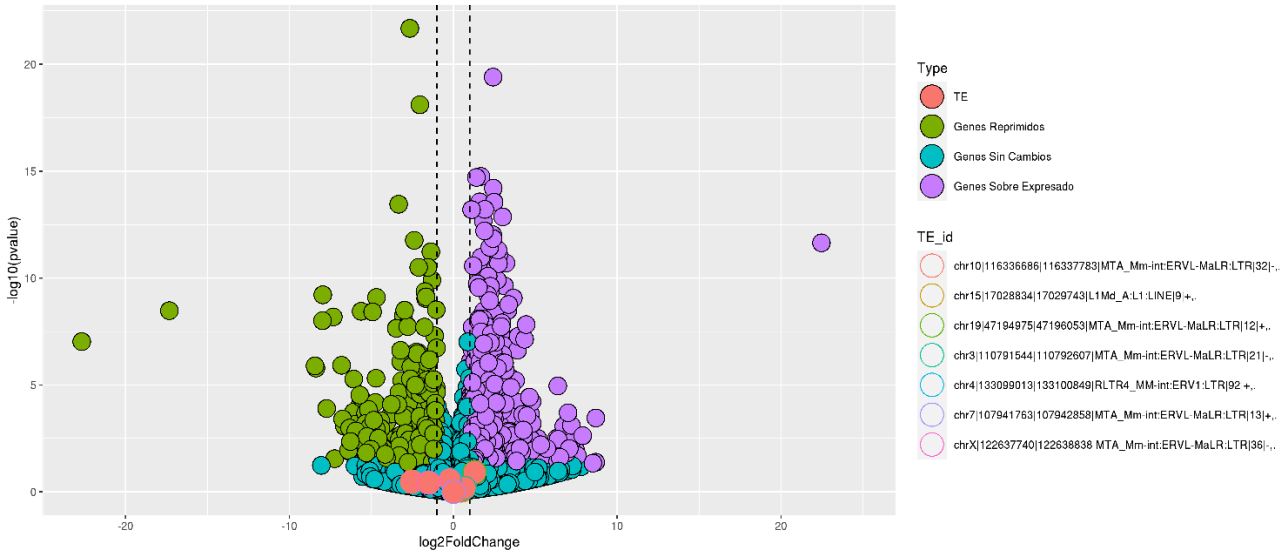

Figura 29: Volcano plot de la expresión de TEs junto a la expresión de genes para la condición Basal vs Early. En el eje Y esta el log10(pvalue), mientras que en el eje X está el log2foldchange, de color rojo estas los TEs, de color verde los genes reprimidos, de color azul los genes que no presentan cambios y de color purpura los genes que se están sobre expresando.

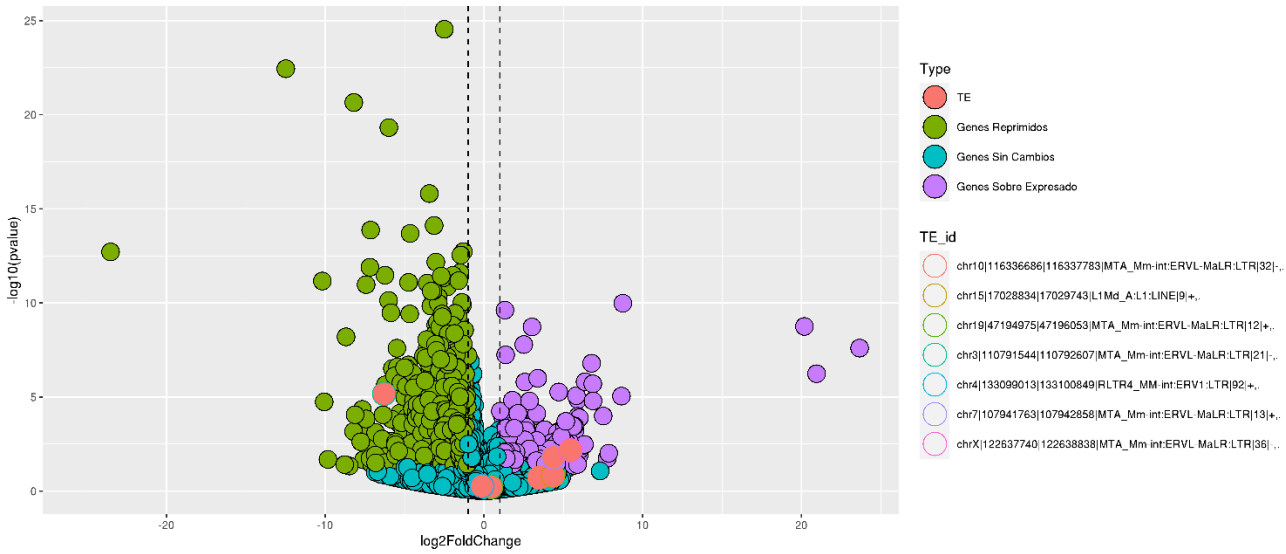

Figura 30: Volcano plot de la expresión de TEs junto a la expresión de genes para la condición Early vs Late. En el eje Y esta el log10(pvalue), mientras que en el eje X está el log2foldchange, de color rojo estas los TEs, de color verde los genes reprimidos, de color azul los genes que no presentan cambios y de color purpura los genes que se están sobre expresando.

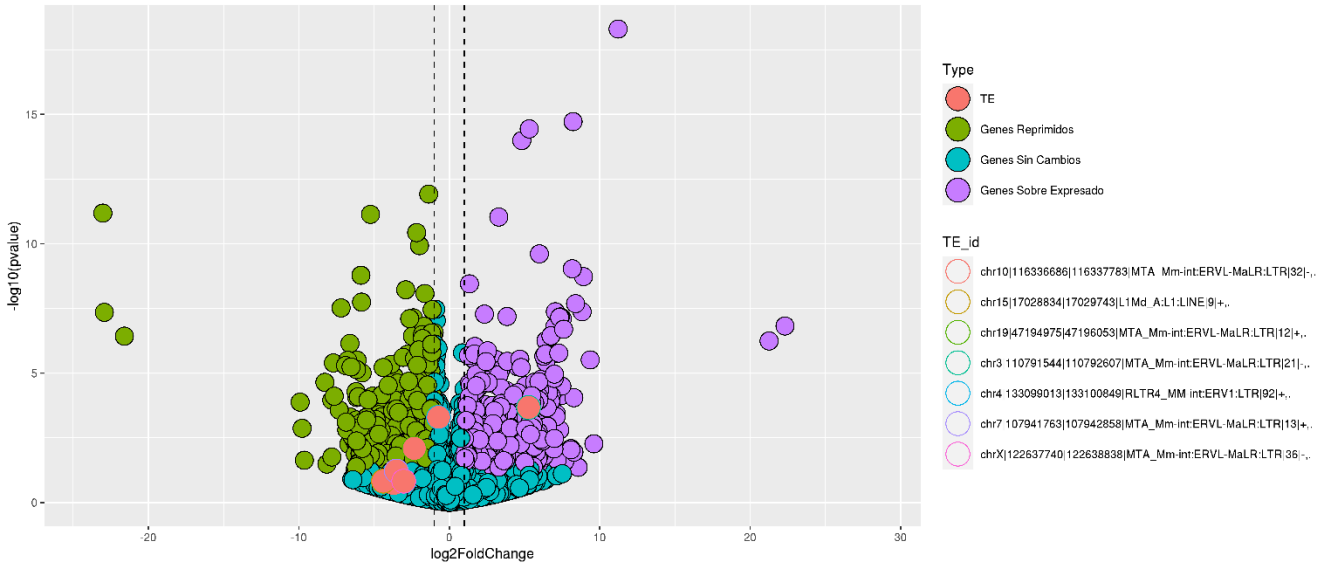

Figura 31: Volcano plot de la expresión de TEs junto a la expresión de genes para la condición Late vs Reactive. En el eje Y esta el log10(pvalue), mientras que en el eje X está el log2foldchange, de color rojo estas los TEs, de color verde los genes reprimidos, de color azul los genes que no presentan cambios y de color purpura los genes que se están sobre expresando.

Finalmente, se realizó el filtro de los genes para la relación con la expresión de los TEs filtrados. Para ello se utilizó Bedtools para encontrar los genes que estuvieran cercanos a los TEs predichos en zonas de eucromatina. Luego se filtró por los que tuvieran expresión (reprimidos/ sobre expresados) para la comparativa de condición. Con esto se obtuvo el siguiente listado:

Gen	**Expresión**
Ntng2-	Reprimido/ Reprimido
Vav3+	Reprimido
Fam102b-	Reprimido
Col11a1+	Reprimido
Tub+	Reprimido
Nap1l3-	Reprimido
Neurl1b+	Reprimido
Pcgf6-	Reprimido / Sobre Expresado
Pdcd11+	Reprimido / Sobre Expresado

Tabla 3. Se exponen los 9 genes relacionados con el set de TEs predichos.

Discusión

En este estudio se evidencia la abismal diferencia que poseen los software encargados en la predicción TEs, como se observa en el diagrama de Venn con los resultados de los TEs predicho por la herramienta SQuIRE (Figura 8) del orden de los millones de TEs predicho por condición. Una explicación para estos números obtenidos es que los TE que disponen de escasas lecturas son reportados de todas formas por esta herramienta. Por el contrario, para el software TEcandidates, en el diagrama de Venn (Figura 7) se vislumbra la baja cantidad de TEs predichos, incluso en algunas uniones de las condiciones se obtienen 0 TEs predichos por la herramienta. Además, la cantidad de los TEs predichos es del orden de las decenas. Optar por una intersección (Figuras 9 a 13) de los TEs predichos por ambas herramientas parece otorgar una mayor certeza en la predicción de los TEs.

Se realizó el proceso de peak call para la identificación de zonas enriquecidas con los datos de ATAC-Seq para cada condición que se evaluó en este estudio. La comparación de los resultados obtenidos por los software MACS3 y HMMRATAC permitió determinar las diferencias que estos poseían para cada condición. Tanto para la condición Basal (Figura 14) como la condición Early (Figura 15) demuestran una mayor región de cobertura para cada cromosoma por parte de MACS3, salvo por el cromosoma 9 el cual tiene una mayor cobertura por la herramienta HMMRATAC. Esta tendencia descrita cambia para la condiciones Late (Figura 16) y Reactivate (Figura 17) en donde se evidencia una mayor cobertura en el genoma de *Mus musculus* con la herramienta HMMRATAC. En base a esto se tomó la decisión de realizar la intersección de los resultados obtenidos por ambos software para mantener una mayor seguridad de las zonas enriquecidas, cabe señalar que al ser datos de ATAC-Seq y utilizar el mapeo de reads paired-end con tamaños de fragmento largo, es poco probable que regiones identificadas como enriquecidas estén incorrectas.

Al integrar los datos de las zonas enriquecidas proporcionados por los software MACS3 y HMMRATAC, con los TEs predichos por SQuIRE se alcanzó una gran variedad de tipos de TEs por cada condición. En base a ello, de los millones de TEs

predichos por SQuIRE, se lograron reducir aproximadamente a los 16.000 para la condición Basal, compuestos mayoritariamente por LTR y SINE. Mientras que para la condición Early los TEs predichos fueron de aproximadamente 36.000, de los cuales en su mayoría se componían de LTR y SINE. Para el caso de las condiciones Late y Reactive, estas obtuvieron aproximadamente 77.000 y 103.000 TEs predichos respectivamente (Figura 18), y estos se componían en su mayoría de LTR y SINE al igual que en los casos anteriores (Figuras 19 a 22). Luego se realizó la unión de los datos anteriormente mencionados con los TEs predicho por la herramienta TEcandidates. Esto produjo para las condiciones basal (Figura 24), Late (Figura 26) y Reactivated (Figura 27) un numero de 5, 6 y 2 TEs predichos respectivamente (Figura 23), los cuales se componen del tipo LTR. Para la condición Early (Figura 25) se tuvo 8 TEs predichos, los cuales eran 7 del tipo LTR y 1 del tipo LINE. Con estos resultados se obtuvo el set final de los TEs que se utilizaron para el posterior análisis de expresión diferencial. Cabe señalar que se confirmó lo expuesto al inicio de este estudio sobre la predicción de TEs sobre SQuIRE, los cuales generaron una cantidad colosal de predicciones por cada condición, en los cuales se argumenta que serían en su mayoría falsos positivos. Gracias a la intersección de los datos arrojados por TEcandidates, se pudo reducir este número considerablemente para mejorar el grado de certeza de los resultados.

Finalmente, para el análisis de expresión diferencial se trabajo con los datos del set final de los TEs (estos datos se obtuvieron a partir de unión de los datos de RNA-Seq y ATAC-Seq), Con estos se procedió a generar los volcano plots comparando la condición Basal vs Early (Figura 27) en la cual los TEs no presentaron cambios de expresión. Mientras que la comparación de la condición Early vs Late (Figura 28) se evidenció 2 TEs sobre expresados (chrX y chr7) y 1 TE reprimido (chr3). Por su parte la comparación de la condición Late vs Reactive (Figura 29) arrojó 1 TE sobre expresado (chr3) y 1 TE reprimido (chr19). Este análisis se relacionó con la expresión de genes (Figura 30-32) con la misma comparación de condiciones. Por último, se usó el set final TEs para la asociación con genes: para el TE ubicado en el cromosoma 7 se detectó al gen Tub+ el cual mostro actividad reprimida para la comparación de Early vs Late, mientras que el gen Nap1l3- localizado en las

proximidades de el TE en el cromosoma X demostró actividad reprimida. En cuanto al TE situado en el cromosoma 3 tenía asociados a los genes Ntng2, Vav3, Fam102b y Col11a1 con actividad reprimida para todos. Finalmente, para el TE radicado en el cromosoma 19 estaba relacionado con los genes Neurl1b (reprimida), Pcgf6 (reprimida en una etapa y sobre expresado en otra) y Pdcd11 (reprimida en una etapa y sobre expresado en otra) en las comparaciones de Early vs Late y Late vs Reactivated.

Conclusiones

Mediante la utilización de enfoques multiomicos para evaluar la expresión específica de locus de TE en este estudio, se encontraron resultados prometedores en 3 TEs específicos. Estos tenían una relación que variaba según las fases que se estaban comparando. Para el TE localizado en el cromosoma 3 tipo LTR, se encontró una relación con el gen Ntng2 el cual se encontraba reprimido para Early vs Late y Late vs Reactivated. Mientras que para el TE LTR localizado en el cromosoma 19, se encontró conexiones con los genes Pcgf6 y Pdcd11 los cuales se encontraba reprimido para Early vs Late y sobre expresado para Late vs Reactivated Las mutaciones del gen Pcgf6 que afectan la región codificante de este gen o el empalme de la transcripción se han asociado con el síndrome de Börjeson-Forssman-Lehmann (BFLS), un trastorno caracterizado por retraso mental, epilepsia, etc. Por otro lado, el gen Ntng2 está involucrado en el control de la formación de patrones y circuitos neuronales a nivel laminar, celular, subcelular y sináptico. Promueve el crecimiento de neuritas tanto de axones como de dendritas. Por último, el gen Pdcd11 es una pieza esencial para la generación de ARNr 18S maduro.

Con estos resultados prometedores, se sugiere integrar a la metodología de este técnicas como pc-HiC, para comprender de forma más completa el papel específico del locus de los TE durante la formación de memoria y recuerdo en el ratón. Con ello podría proponerse un modelo de regulación por TEs más claramente en trabajos futuros. Por el momento, el trabajo realizado destaca la importancia de este tipo de

análisis para ampliar nuestro conocimiento sobre cómo las TE pueden estar involucrados en la regulación de la formación de la memoria y el recuerdo.

Referencias:

Britten, R. J. (1996). Cases of Ancient Mobile Element DNA Insertions That Now Affect Gene Regulation. *Molecular Phylogenetics and Evolution*, *5*(1), 13–17. https://doi.org/10.1006/mpev.1996.0003

Carmona, A. (2013). LOS ELEMENTOS GENÉTICOS MOVILES EN LA CÉLULA TUMORAL: EL DESPERTAR DE UN GIGANTE. *EUBACTERIA*, *32*.

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*, *29*(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635

Elbarbary, R. A., Lucas, B. A., & Maquat, L. E. (2016). Retrotransposons as regulators of gene expression. *Science*, *351*(6274), aac7247–aac7247. https://doi.org/10.1126/science.aac7247

Estécio, M. R. H., Gallegos, J., Dekmezian, M., Lu, Y., Liang, S., & Issa, J.-P. J. (2012). SINE Retrotransposons Cause Epigenetic Reprogramming of Adjacent Gene Promoters. *Molecular Cancer Research*, *10*(10), 1332–1342. https://doi.org/10.1158/1541-7786.MCR-12-0351

Evan D. Tarbell and Tao Liu, HMMRATAC: a Hidden Markov ModeleR for ATAC-seq, Nucleic Acids Res. 2019 Jun 14. doi: 10.1093/nar/gkz533

Ferrari, R., Grandi, N., Tramontano, E., & Dieci, G. (2021). Retrotransposons as Drivers of Mammalian Brain Evolution. Life (Basel, Switzerland), 11(5), 376. https://doi.org/10.3390/life11050376

Feschotte, C. (2008). Transposable elements and the evolution of regulatory networks. *Nature Reviews Genetics*, *9*(5), 397–405. https://doi.org/10.1038/nrg2337

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B. W., Nusbaum, C., Lindblad-Toh, K., … Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, *29*(7), 644–652. https://doi.org/10.1038/nbt.1883

Huang, C. R. L., Burns, K. H., & Boeke, J. D. (2012). Active Transposition in Genomes. *Annual Review of Genetics*, *46*(1), 651–675. https://doi.org/10.1146/annurev-genet-110711-155616

Josselyn, S. A., Köhler, S., & Frankland, P. W. (2015). Finding the engram. *Nature*

Reviews Neuroscience, *16*(9), 521–534. https://doi.org/10.1038/nrn4000

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

Lev-Maor, G. (2003). The Birth of an Alternatively Spliced Exon: 3' Splice-Site Selection in Alu Exons. *Science*, *300*(5623), 1288–1291. https://doi.org/10.1126/science.1082588

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

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

Marco, A., Meharena, H. S., Dileep, V., Raju, R. M., Davila-Velderrain, J., Zhang, A. L., Adaikkan, C., Young, J. Z., Gao, F., Kellis, M., & Tsai, L.-H. (2020). Mapping the epigenomic and transcriptomic interplay during memory formation and recall in the hippocampal engram ensemble. *Nature Neuroscience*, *23*(12), 1606–1617. https://doi.org/10.1038/s41593-020-00717-0

Mariño-Ramírez, L., Lewis, K. C., Landsman, D., & Jordan, I. K. (2005). Transposable elements donate lineage-specific regulatory sequences to host genomes. *Cytogenetic and Genome Research*, *110*(1–4), 333–341. https://doi.org/10.1159/000084965

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

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

R Core Team. (2021). *No Title* (3.6). https://www.r-project.org/

Seberg, O., & Petersen, G. (2009). A unified classification system for eukaryotic transposable elements should reflect their phylogeny. *Nature Reviews Genetics*, *10*(4), 276–276. https://doi.org/10.1038/nrg2165-c3

Stajich, J. E. (2002). The Bioperl Toolkit: Perl Modules for the Life Sciences. *Genome Research*, *12*(10), 1611–1618. https://doi.org/10.1101/gr.361602

Sun, Y., Miao, N., & Sun, T. (2019). Detect accessible chromatin using ATAC-sequencing, from principle to applications. *Hereditas*, *156*(1), 29. https://doi.org/10.1186/s41065-019-0105-9

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

Valdebenito-Maturana B, Arancibia E, Riadi G, Tapia JC, Carrasco M. Locus-specific analysis of Transposable Elements during the progression of ALS in the SOD1G93A mouse model. PLoS One. 2021 Oct 6;16(10):e0258291. doi: 10.1371/journal.pone.0258291. PMID: 34614020; PMCID: PMC8494334.

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

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