Facultad de Ingeniería

Ingeniería Civil en Bioinformática

Relación entre marcas epigenéticas y actividad transcripcional en elementos transponibles.

Alumno: Matías Ezekiel Véjar Reyes

Prof. Tutor: Gonzalo Riadi Mahias

Prof. informante: Braulio Valdebenito Maturana

Talca - Chile

Junio, 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

ÍNDICE

Índice de figuras y tablas

Abstract

Transposable elements (TEs) are DNA segments that can move from one region to another in a genome. They transpose, having unknown effects on gene expression. The ENCODE project has shown that about 70% of the genome can be transcribed. It is also known that about half of the human genome corresponds to TEs, so it is deduced that about 35% of the genome corresponds only to TEs that are being transcribed.

The cell, trying to defend itself against the spread of TEs, sets in motion regulatory mechanisms at the epigenetic level. Epigenetic regulations are a set of chemical modifications of nitrogenous bases or histones, which do not change the DNA sequence, and which allow the expression of genes, including TEs, to be regulated. Thus, one might expect to find, for certain epigenetic marks, a proportional relationship between the number of marks in a predicted active TE and its expression.

Although the relationship between the transcriptional activity of TEs with epigenetic regulation is unknown, some marks have been associated with DNA activity. In this work, we propose to search for this relationship between differential expression of active TEs and the three epigenetic marks.

As most TEs are inactive fossils and fragments, it is difficult to identify the locus of the element from which it was transcribed because there are multiple paralogs in the genome; meanwhile, there are 2 computational tools, TEcandidates and SQUiRE, that allow predicting the expression of transcriptionally active TEs in a genome without losing their location. We will use these tools to select a set of active TEs from different superfamilies.

After predicting transcriptionally active TEs, it was observed that as expected, most of the TEs that were predicted to be active did not exhibit epigenetic marks. However, none of the overexpressed TEs showed epigenetic marks either. The only one of the 3 marks to show binding to TEs was H3K27ac. However, all TEs with the H3K27ac mark were not active, suggesting that this epigenetic mark might have repressive effects on TEs. In addition, this mark appeared to prefer LTR and non-LTR subclasses in retrotransposons.

Considering the results obtained, it can be established that there are signs of a relationship between H3K27ac marks and retrotransposons predicted to be active in human prostate cancer. In addition, none of the other marks showed any dependence on TEs.

It is suggested for future studies, in order to have a better understanding of the relationship between TEs and epigenetic marks, the use of a larger repertoire of marks. In addition, additional RNA-Seq dataset on different conditions and organisms would be desirable.

Resumen

Los elementos transponibles (TE) son segmentos de ADN que tienen la capacidad de moverse de una región a otra, en un genoma. Éstos se transponen, teniendo efectos desconocidos en la expresión de los genes. El proyecto ENCODE ha mostrado que sobre el 70% del genoma puede transcribirse. También se sabe que cerca de la mitad del genoma humano corresponde a TEs, por lo que se deduce que cerca de un 35% del genoma corresponde únicamente a TEs que están siendo transcritos.

La célula, intentando defenderse de la propagación de los TEs, pone en marcha mecanismos de regulación a nivel epigenético que actúan a nivel transcripcional. Las regulaciones epigenéticas son modificaciones químicas de las bases nitrogenadas o de las histonas, que no cambian la secuencia del ADN, y que permiten regular la expresión de genes y TEs. Se esperaría, para ciertas marcas epigenéticas, una relación proporcional entre el número de marcas en un TE predicho como activo y su expresión.

Aunque se desconoce la relación que existe entre la actividad transcripcional de los TEs con la regulación epigenética, se han asociado algunas marcas con la actividad del ADN. En este trabajo, se propone buscar esta correlación entre la expresión diferencial de TEs activos y las mencionadas tres marcas epigenéticas.

Como la mayoría de los TEs son fósiles inactivos y fragmentos, es difícil identificar el locus del elemento del cual fue transcrito debido a que existen múltiples parálogos en el genoma; Entretanto, hay 2 herramientas computacionales, TEcandidates y SQUiRE, que permiten predecir la expresión de TEs transcripcionalmente activos en un genoma sin perder su ubicación. Se usarán esas herramientas para seleccionar un conjunto de TEs activos de diferentes superfamilias.

Tras predecir los TEs activos se observó como esperado que gran parte de los TEs predichos como activos no presentan marcas epigenéticas. Entretanto, ningún TE sobre-expresado presentó marcas epigenéticas tampoco. La única de las 3 marcas en presentar unión con TEs fue H3K27ac. Sin embargo, todos los TEs con la marca H3K27ac, no fueron activos, lo que sugiere que esta marca podría tener efectos represivos en TEs. Además, esta marca pareciera tener una preferencia por las subclases LTR y no-LTR, en retrotransposones.

Considerando los resultados, se puede establecer que hay indicios de una relación entre la marca H3K27ac y retrotransposones predichos como activos en cáncer de próstata en humano. Además, ninguna de las otras marcas presentó ninguna dependencia con TEs.

Se sugiere para futuros estudio, para tener un mejor entendimiento de la relación entre TEs y marcas epigenéticas, la utilización de un mayor repertorio de marcas. Además, sería deseable sets de datos de RNA-Seq en estudios sobre diferentes condiciones y especies.

Introducción

Entender los misterios que encierra la información genética permitiría descubrir muchos de los fenómenos que ocurren en la vida y los elementos que están involucrados en cada una de estas condiciones. Cada vez se hacen más esfuerzos para entender qué función cumplen cada uno de los elementos que componen un organismo. Se sabe que dentro del genoma se almacena la información de genes codificantes, ADN no codificante, elementos transponibles (TE) y otros elementos, los que podrían ser transcritos en determinadas circunstancias (Dunham et al., 2012).

Los TEs en particular, son segmentos de ADN que codifican enzimas u otras proteínas involucradas en el movimiento de su ADN dentro de los genomas o entre células bacterianas (Frost et al., 2005). Existen dos clases de elementos móviles: transposones de ADN y retrotransposones, como se aprecia en la figura 1. La mayoría de los transposones de ADN se mueve como segmentos de ADN, que se cortan (dejando su sitio de origen) y pegan de forma autónoma en una nueva localización genómica. Por otro lado, los retrotransposones emplean un mecanismo de replicación, en donde se copian mediante una molécula de RNA para lograr su diseminación. Así, el retrotransposón se mantiene en el mismo sitio de donde es transcrito y una secuencia de RNA se transcribe de forma reversa respecto a la secuencia del transposón original; el cual será insertado de otra región del genoma.

Los retrotransposones se subdividen en 2 tipos: Los que poseen repetición terminal larga (LTR) y los que no (non-LTR), que comprenden tanto elemento nuclear intercalado largo (*Long Interspersed Nuclear Element,* abreviado como LINE) y elemento nuclear intercalado corto (*Short Interspersed Nuclear Element,* abreviado como SINE)) y superfamilias (por ejemplo, retrovirus endógenos (ERV), Copia y LINE-1), siguiendo el sistema de clasificación propuesto por (Wicker et al., 2007). Por ejemplo, los LTR utilizan una integrasa para insertar el ADN complementario del TE en el genoma, mientras que los elementos non-LTR utilizan una endonucleasa para realizar la transcripción inversa dirigida al locus de la inserción Por otro lado, el mecanismo de transposición de TIR (un tipo de transposón de ADN) comienza con la unión de 2 monómeros de transposasas para formar el complejo de final simple, es decir con un monómero de transposasa en cada extremo. Luego, estas transposasas se juntan formando el complejo de final apareado, donde el transposón es cortado. Finalmente este dímero de transposasas reconoce un dinucleótido TA en donde se inserta (Munoz-Lopez & Garcia-Perez, 2010).

Los retrotransposones non-LTR comprenden sobre un 33.7% del genoma humano, mientras que los LTR alrededor del 8.3% y los transposones de ADN cerca de un 2.8% (Cordaux and Batzer, 2009). Estos porcentajes dan a entender que los TE son elementos abundantes del genoma humano. Al comienzo no existía mucha información sobre estos, se pensaban que hay "basura" dentro del ADN. Con el tiempo se fue aumentando el conocimiento que se tenía sobre ellos, a tal punto que hoy en día son asociados a la regulación de genes.

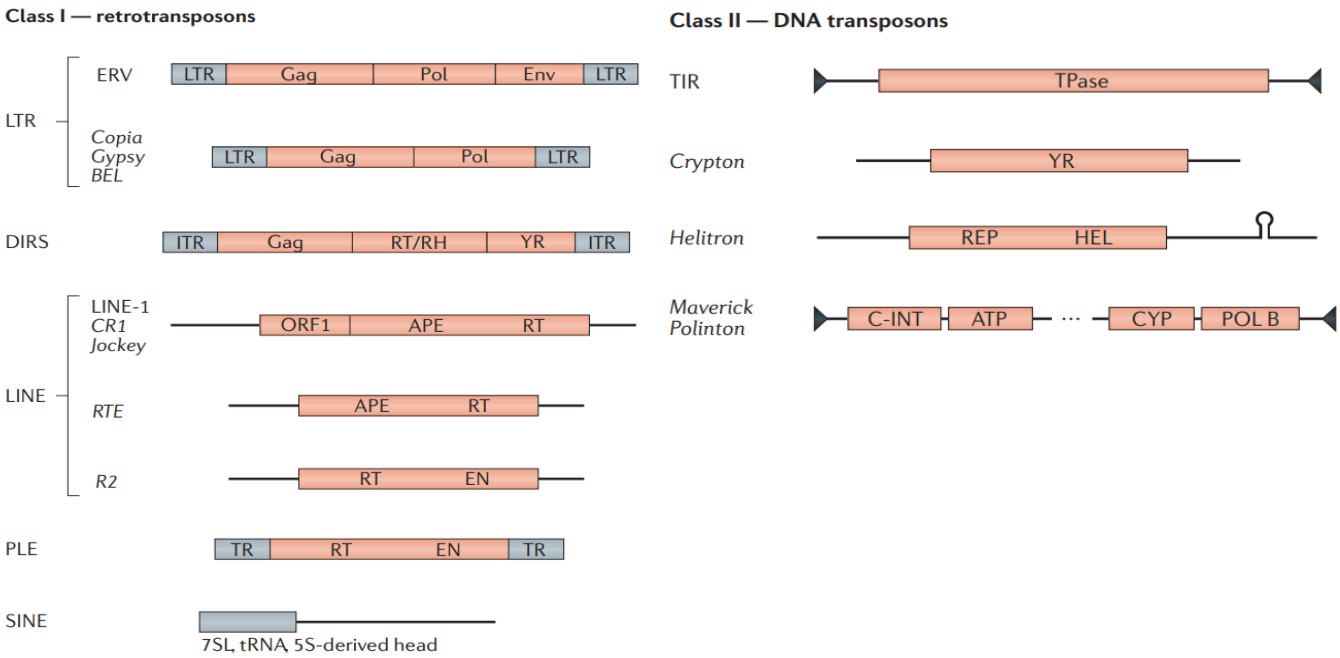

Figura 1 - Clasificación de elementos transponibles. Los TE se clasifican en retrotransposones (Clase 1), los que se identifican por la transcripción reversa de un intermediario de ARN, y transposones de ADN (Clase 2). APE, Endonucleasa apurinica; ATP, ATPasa; C-INT, c-integrasa; CYP, Cisteína proteasa; DIRS, Secuencia Repetida intermediaria de Dictyostelium ; EN, endonucleasa; HEL, helicasa; ITR, Repetición terminal repetida; ORF, Marco de lectura abierto; POL B, ADN polimerasa B; REP, proteína iniciadora de replicación; RH, RNasa H; RT, transcriptasa reversa; RTE, elemento retrotransponible; TPase, transposasa; TR, Repetición terminal (estructura variable); YR, tirosina recombinasa. Adaptado de (Deniz et al., 2019).

Según el proyecto Encode, en humanos, sobre el 85% del genoma se transcribe (Agostini et al, 2021). De esa porción, tan solo el 5% corresponde a genes que codifican a proteínas. Por tanto, se puede especular que del 65% restante que se transcribe, los TEs pueden representar al menos el 42%. Similar a lo que ocurre en el genoma humano, en otros organismos eucariontes se posee del orden de millones de copias de TEs. Es más, alrededor de mitad de las secuencias de los genomas más estudiados en mamíferos tienden a anotarse como TE. Por otro lado, entre el 5%-10% de las secuencias en genomas de mamíferos y vertebrados comprenden genes y elementos funcionales conocidos; que corresponde

principalmente a mRNAs y RNAs estructurales y regulatorios. (de Koning et al., 2011). Si bien corresponden a una importante porción de los genomas, su investigación ha sido relativamente reciente.

Los TE fueron por mucho tiempo denominados como "ADN chatarra", y no fue hasta que entre la década del 40 y 50 que Barbara McClintock descubrió los "genes saltarines" que hoy se conocen como TEs.

En la actualidad, este tipo de componentes genómicos suelen ser descartados, ya que es difícil determinar el elemento del cual fue expresado dada su naturaleza repetitiva. Sin embargo, la evidencia actual sugiere que los TE están asociados al movimiento de fragmentos de ADN en el genoma. Esto produce eventos de inserciones, deleciones, Inversiones, duplicaciones, etc. de ADN. De la misma manera, al cortar o insertar fracciones de ADN, se pueden producir mutaciones que afecten a el funcionamiento de genes u otros elementos del genoma.

Los efectos que puede tener el movimiento de TEs pueden ir desde cambios en la expresión de un gen codificante producto de una mutación al ser insertado, hasta activación o represión de genes. Las transposiciones de TE en células somáticas han recibido más atención, ya que cada vez se recopila más información de efectos perjudiciales que tiene la transposición de TE. Un ejemplo de esto se presenta en TEs de *Drosophila Melanogaster* (mosca de la fruta), que puede transponerse durante su ciclo de vida y afectar negativamente su conducta y capacidad reproductiva. La transposición de TE en células somáticas se ha relacionado con algunas enfermedades como cáncer de ovarios, cáncer colorrectal y anemia Fanconi, la cual provoca a una insuficiencia de la médula ósea para producir suficientes células sanguíneas nuevas. En adición a las enfermedades anteriores, también se han relacionado los TE con diferentes enfermedades mentales como esquizofrenia y autismo (Misiak, B. *et al.*, 2019 ; Nie, Y. *et al.*, 2020).

En pacientes esquizofrénicos, la secuenciación de sus genomas completos ha revelado inserciones específicas de TEs del tipo LINE-1 mediante retrotransposición en neuronas, los que probablemente son desencadenados por el ambiente y/o factores de riesgo genético, y localizados preferentemente en genes relacionados con la sinapsis y esquizofrenia (Bundo et al., 2014;Lozano et al., 2015). Así, se podría asociar qué TEs estén contribuyendo a la susceptibilidad o desarrollo de la enfermedad. Recientemente, el movimiento de TEs ha sido asociado a enfermedades neurodegenerativas como el Alzheimer (Guo et al., 2018). Dado que la expresión de los TE provoca múltiples efectos perjudícales, los genomas han diseñado ciertos mecanismos para evitar el daño. Es decir, para que estos elementos no puedan insertarse en zonas importantes como podrían ser genes que codifican a proteínas esenciales para el organismo.

Existen componentes en el genoma que están involucrados en la expresión o movimiento de TEs, como son la secuencia, la cromatina y los contextos nucleares, que explica la distribución de estos elementos en el genoma. Para que, de esta forma, sean insertados mayoritariamente en zonas donde no produzcan daño.

Regulación de TE

La expresión de TEs en un organismo puede producir variados efectos adversos al no ser regulada. Por esto, la célula ha incorporado mecanismo de regulación para estos elementos. La principal forma de controlarlos es por medio de la compactación y apertura de la cromatina. Los remodeladores de la cromatina que dependen de ATP son actores importantes en la regulación TEs tanto en mamíferos como en plantas. Estos dan acceso a las metiltransferasas para añadir modificaciones represivas en el ADN y cromatina. Las metilaciones del ADN, específicamente la 5-metilcitosina (5mC), es posiblemente la estrategia más utilizada en organismos eucariontes para mantener los TEs en un estado reprimido. Se ha hipotetizado que la necesidad de mantener el silenciamiento de los TEs ha impulsado la evolución de la metilación del ADN como mecanismo represivo, que más tarde se adaptó para actuar en otros contextos, como el genomic imprinting (Bajrami and Spiroski, 2016). Esta última puede definirse como la expresión selectiva de un gen según el origen parental del alelo (paterno o materno).

Aunque la 5mC ha sido la modificación de ADN más estudiada hasta la fecha, existen otras modificaciones que también se han relacionado con TE en diferentes especies. Además, de 5mC, los tipos más comunes de modificación de ADN incluyen N4-metilcitosina (4mC) y N6-metiladenina (6mA), que están muy extendidas en todas las bacterias. En particular, 6mA también se encuentra en cantidades variables en eucariontes (Deniz et al., 2019). Pero en algunos casos estas modificaciones sobre ADN no se perpetúan a lo largo de las generaciones, incluso hay circunstancias en las que es necesario que sean quitadas.

Durante el desarrollo, los mamíferos sufren dos grandes periodos de reprogramación epigenética en la que el genoma se desmetila: primero para formar las células pluripotentes del cigoto durante el desarrollo preimplantacional y segundo para producir gametos. En ratones, la hipometilación del ADN coincide con la regulación transitoria de varios TEs durante ambos períodos de reprogramación (Molaro et al., 2014). En estas etapas críticas del desarrollo, particularmente en la línea germinal, el riesgo y las consecuencias de las inserciones nuevas de TEs son elevadas. Para minimizar los riesgos asociados, el organismo utiliza numerosas estrategias complementarias para restringir la movilidad de TE. Por ejemplo, en la línea germinal masculina, la expresión de TEs que se gatilla al eliminar las 5mC conduce rápidamente a la activación de la vía piRNA. Esto a su vez produce que estos elementos sean capturados, y posteriormente degradados (Aravin et al., 2008)

Tanto en plantas como en mamíferos, los residuos de citosina pueden metilarse en un contexto simétrico (CpG), y esta metilación se copia en la nueva cadena de ADN tras la replicación del ADN, proporcionando un mecanismo para la herencia del silenciamiento de TE. En el ratón, la metiltransferasa de ADN (DNMT1) es un responsable de este mantenimiento de la metilación del ADN. Se ha reportado que niveles elevados de TEs en embriones con deficiencia de DNMT1, que mueren después de 9-10 días. Sin embargo, esta metilación parece estar dirigida a los TEs *de novo* con cada generación. Los que corresponden a TEs generados producto de la replicación del ADN. En el ratón, las metiltransferasas llevan a cabo la metilación del ADN nuevo, es decir las metilaciones que ocurren por primera vez en un miembro de una familia (principalmente CpG); y son necesarias para la metilación de TEs y el silenciamiento en células germinales. Estos hallazgos sugieren que la metilación del ADN podría ser necesaria para el silenciamiento epigenético de TE en etapas específicas del desarrollo de mamíferos (Slotkin and Martienssen, 2007).

Las metilaciones de ADN no es la única forma de regulación epigenética de TEs. Existen varios grupos de modificaciones de cromatina que reprimen la transcripción de TE. Las colas amino-terminales de las histonas sobresalen de los nucleosomas y están sujetas a una variedad de modificaciones postraduccionales (por ejemplo, acetilaciones, metilaciones y fosforilaciones, entre otras), combinaciones de las cuales se ha predicho que participan en la regulación de la transcripción. Las modificaciones de los aminoácidos N terminal de las histonas alteran la unión de factores proteicos y transmiten información a factores de transcripción. En los nucleosomas (subunidad de la cromatina compuesta de aproximadamente 146 pares de bases de ADN, que se enrollar en un octámero de proteínas histonas) asociados con los TE, se enriquecen estos factores para la metilación de la histona H3 en la lisina 9 (H3K9), que es una señal de represión transcripcional lo que inactiva la cromatina. Las mutaciones en genes necesarios para la represión de las modificaciones de las colas de histonas conducen a la reactivación de TEs; por ejemplo, las mutaciones en el gen de la metiltransferasa Suv39 (que tiene la función de metilar la lisina 9 de la histona H3) que actúa en la histona H3K9 (asociado a la inhibición de genes), dan como resultado una disminución en el silenciamiento de la transcripción de TE en la célula embrionarias de ratón (Peters et al., 2001)

Por ejemplo, la metilación de Lisina 9 de histona H3 (H3K9me1) está involucrada en la formación de heterocromatina represiva estable, mientras que la metilación de H3K4, H3K36 y H3K79 están asociadas con la actividad de transcripción. En *S. cerevisiae*, Set1p (H3K4 metiltransferasa) y Set2p (H3K36 metiltransferasa) se asocian con ARN polimerasa II (ARN Pol II), implicando en la iniciación transcripcional o elongación. Sin embargo, actualmente se sabe poco sobre la importancia de estas modificaciones en organismos

eucariontes más complejos. Los residuos de Lisina in vivo pueden ser mono- di-, o tri-metilados, haciendo su función en la transcripción más compleja. H3K4me3 es una marca asociada con genes activos en levaduras, pero la función de H3K4me2 es menos clara. Se ha sugerido que el H3K4me2 se correlaciona con un estado "permisivo" de cromatina, en el que los genes pueden ser transcritos. Para identificar las diferencias entre lisina di- y trimetil, se ha utilizado la técnica de inmunoprecipitación de cromatina seguida de por secuenciación (ChIP-seq) para mapear por separado estas dos formas metiladas de H3K4 en las regiones promotoras y transcritas de genes de pollo que están regulados por el desarrollo, constitutivamente activos e inactivos (Schneider et al., 2004).

ChIP-seq es una técnica de secuenciación que se usa principalmente para identificación de los sitios de proteínas al ADN, y es uno de los métodos más utilizados para la detección de modificaciones en colas de histonas. ChIP-seq para modificaciones de histonas utiliza la digestión de nucleasa micrococica (MNase) o sonicación para cortar el ADN. Luego, se puede ejecutar en muestras de baja cantidad cuando se combina con la amplificación posterior a ChIP adicional, como se puede ver en la Figura 2. Cabe mencionar que el éxito de la secuenciación depende fundamentalmente del desarrollo y validación de la unión entre un anticuerpo diseñado para unirse de forma específica a la proteína de interés y la modificación de la histona. De otra forma, no se podrá detectar de forma correcta la modificación (Furey, 2012).

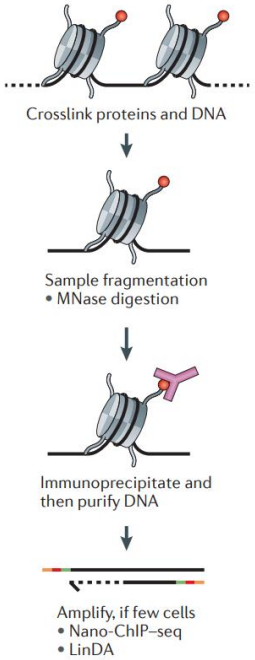

Figura 2 - Protocolo experimental CHIP-seq para modificaciones de histonas. Nano-ChIP-seq, Secuenciación de Inmunoprecipitación de Cromatina para un número limitado de células ; MNase, Nucleasa micrococica; LinDA, amplificación de ADN lineal de tubo único. Adaptado de (Furey, 2012).

Si bien existe información acerca de la clasificación de TEs, posibles efectos y formas de regulación, aun no hay una relación clara entre estos elementos móviles y las marcas epigenéticas que los estarían regulando. Pero, aun si se pudieran definir de forma clara las modificaciones que presentan las histonas de un genoma, no es simple identificar el origen de TE que estén siendo expresado. De hecho, las continuas y sucesivas mutaciones a lo largo del tiempo provocan que, las copias de los TEs se vayan fragmentando y siendo cada vez más diferentes; esto hace más difícil su identificación mediante similitud de secuencias. En ocasiones, los TEs se insertan en copias de otros TEs, por lo que se genera aún más divergencia de la secuencia original (Kaminker et al., 2002).

Por esto, identificar elementos móviles no es una tarea trivial. De hecho, suele ser un problema recurrente en algoritmos computacionales y en la detección automatizada de tales elementos. Antiguamente, se utilizaban estrategias que comparaban secuencias de interés contra TEs de bases de datos, de este modo se determinaba si existía una similitud suficiente para ser considerado TE. Sin embargo, hacer relaciones biológicas que involucren TE sólo se pueden establecer si existe un alto grado de confianza en la anotación de estos.

De esta manera, se busca que los métodos de identificación de TE sean cada vez más robustos y con una menor tasa de falsos positivos y falsos negativos. Sin embargo, muchos protocolos se basan en programas como Repeatmasker. Este fue indicado en su tiempo como "ni el enfoque más eficiente ni el más sensible" (Juretic et al., 2004). Al contrario, se fueron haciendo avances que indicaban que los modelos genéticos robustos eran aquellos que combinaban múltiples fuentes de información.

Con el desarrollo de varios métodos para la detección de TEs y elementos repetidos, es posible aplicar un enfoque similar a lo que podría llamarse "evidencia combinada" para incrementar la calidad de la anotación de los TE (Quesneville et al., 2005)

En la actualidad, cuando se quiere hacer el ensamble de un genoma, se espera encontrar estos elementos a fin de enmascarar sus secuencias. El ocultamiento de estas repeticiones se hace para que no se vean afectados los análisis realizados con estos datos, como la expresión diferencial de genes. (Biscotti et al., 2015). El estudio de estos elementos puede ser relevante para entender fenómenos biológicos que actualmente no tienen explicación. Por un lado, se tienen efectos dañinos que afectan a genes que producen enfermedades, pero por otro se añaden mutaciones que van promoviendo la evolución del genoma. Por esto, se ve como un arma de doble filo, que involucra tanto fenómenos beneficiosos como perjudiciales (Saze, 2018). Por ello, que al identificar los TE expresados en una cierta condición se debe tener una alta confianza en la predicción, ya que al no ser así podría provocar que se lleguen a conclusiones erróneas.

En la actualidad existen dos herramientas que buscan predecir TE expresados a partir de datos de RNA-seq, sin perder su ubicación en el genoma: SQUiRE (Yang et al., 2019) y TEcandidates (Valdebenito-Maturana & Riadi, 2018).

Por un lado, TEcandidates alinea lecturas de RNA-Seq en el genoma de referencia y luego contra la anotación de TEs. Luego realiza un ensamble de novo con Trinity. Los contigs generados son ensamblados al genoma de referencia y anotados con Bedtools. TEcandidates usa el largo y porcentaje de identidad de las secuencias resultantes como criterio de elección del candidato a ser transcripcionalmente activo. Por otro lado, se tiene SQUiRE el cual realiza un análisis de la expresión diferencial del conteo de genes y TEs que fueron encontrados. Para, de esta forma, poder reportar TEs sobre-expresados y reprimidos.

Si bien, estos programas realizan una predicción que va más allá de alinear secuencias con una base de datos de TE; el desempeño que se tiene no es óptimo, ya que siempre cabe la posibilidad de que existan falsos positivos, que la anotación de TEs tuviese algún problema, que el alineamiento no sea correcto, etc. Por lo que se plantea la posibilidad de añadir información de marcas epigenéticas de modo de agregan una validación extra a las predicciones en ambas herramientas.

Problemática

Por esta razón, se plantea la idea de añadir más información para identificar correctamente los TE que están siendo expresados. Como se planteó anteriormente, los TE tienen una marcada regulación por medio de mecanismos epigenéticos. Incluso, se menciona que hay factores, como las modificaciones de las colas de histonas, que intervienen directamente en la expresión de genes. De esta forma, que se busca establecer la relación entre marcas epigenéticas y la expresión de TEs al incorporar datos de marcas de histonas como H3K4me3 y H3K27me3. Si bien, se sabe que hay presencia de TEs transcripcionalmente activos asociados a ciertas enfermedades, aún se desconoce qué relación tienen estos elementos y regulación ejercida por ciertas marcas epigenéticas.

Hipótesis

La expresión diferencial de marcas epigenéticas se correlaciona directamente con la expresión diferencial de elementos transponibles.

Objetivos

Objetivo general

Correlacionar la expresión diferencial de marcas epigenéticas con la expresión diferencial de elementos transponibles.

Objetivos específicos

1. Predecir TEs activos a partir datos de estudios de SRA que incluyan experimentos de RNA-seq información ChIP-seq, en condiciones similares.
2. Identificar regiones enriquecidas en marcas epigenéticas con los respectivos sets de ChIP-seq.
3. Evaluar, para las distintas familias de TEs, la correlación de la expresión de marcas epigenéticas con la expresión diferencial de TEs.

Materiales

Para la realización de este trabajo se utilizará un clúster que tiene las siguientes características:

- 2 CPUs Intel Xeon E7-8867 v3 @ 2.50GHz.
- 64 cores (28 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)

Junto a los requerimientos en hardware, se tendrá que hacer el procesamiento de un conjunto de muestras de RNA-seq y ChIP-seq, las que serán obtenidas de la base de datos de SRA (Sequence Read Archive). Estas muestras corresponden a los BioProjects PRJNA494345 y PRJNA494344, respectivamente (Leinonen et al., 2011). De estos se seleccionarán 3 subconjuntos de muestras los que serán compuestos de un RNA-seq, ChIP-Seq de modificadores de colas de histonas y sus respectivos inputs o control de ChIP-seq. Las muestras usadas pertenecen a 25 pacientes, los que se clasifican en 17 sanos y 8 pacientes con cáncer de próstata. Por cada paciente hubo una muestra de RNA-seq y 3 datos de ChIP-seq (uno de cada una de las muestras analizadas). En el apéndice 1 se detallan las muestras que fueron utilizadas del conjunto total de datos mencionados.

Metodología

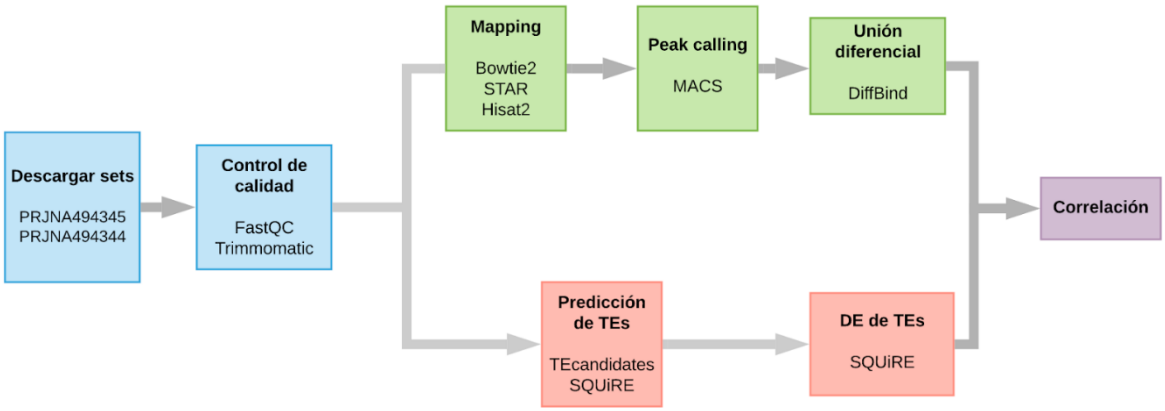

Figura 3 - Flujo de trabajo. Este gráfico cubre las 4 etapas de la metodología, identificadas por el color que rellena cada cuadro. Donde se tiene la descarga de muestras y archivos necesarios de las bases de datos. Luego se hace un control de calidad de los datos descargados. Después, hay 2 áreas enmarcadas, las que representan los procesos que corresponden al análisis de marcas epigenéticas (verde) y el otro a TEs (rosado). Finalmente, se realiza la correlación de los datos.

En la figura 3 se observa un diagrama con los pasos que fueron necesarios para la realización de este proyecto. Primeramente, se buscó sets de datos que se encuentren en la base de datos de SRA de NCBI. Los objetos de información que se buscaron fueron de RNA-seq y ChIP-seq, las que se encontraron en condiciones similares. En los datos de RNA-seq (Ozsolak and Milos, 2011) se obtuvo lecturas de la información de los transcritos, en ChIP-seq (Furey, 2012) se tuvieron lecturas de la posición en que se encontrarán modificaciones especificas en histonas. Posterior a eso, se escogió las condiciones, marcas epigenéticas y organismos que se utilizaron para este trabajo en base a la disponibilidad que existía. De este grupo de datos, se usó un set resulte relevante para este trabajo. Una vez definidos los sets de datos que se usaron, se procedió a descargarlos con la herramienta SRA Explorer (Ewels, n.d). Esta permitió, utilizando los *accession numbers*, descargar secuencias de diferentes experimentos formando una colección de archivos. Estos fueron descargados de forma consecutiva con un script que la misma herramienta entregará.

Una vez obtenidos los archivos, se procedió a hacer un control de calidad de las lecturas descargadas usando la herramienta FastQC (Andrews, 2010). Esta herramienta permite ver la calidad de las lecturas, mediante el puntaje de calidad Phred por base, que está asociado a la probabilidad de error de que la base sea denominada correctamente. También, es posible ver la presencia de adaptadores, los que de estar presentes, se tienen que eliminar. Si hay lecturas de baja calidad o adaptadores se utilizará Trimmomatic (Bolger et al., 2014) para filtrar las lecturas que no tengan buena calidad o adaptadores. Para ello se establecerá una mínima calidad de 20.

Una vez verificada la calidad del set de datos, se procedió a predecir los TEs transcripcionalmente activos, mediante el uso de TEcandidates. Este es un script que se basa en la realización de un ensamble *de novo* del transcriptoma para encontrar los elementos transponibles (TE) que están siendo expresados y su localización. TEcandidates recibe como entrada los datos de RNA-seq, la secuencia del genoma, y un archivo de anotación de TE. Con esto, el programa entrega una lista de posiciones de cada uno de los TEs que están siendo expresados. Por otro lado se usó SQUiRE para predecir los TEs transcripcionalmente activos, de modo de añadir y/o corroborar la información entregada por TEcandidates (Yang et al., 2019). Además, se ocupó para realizar la expresión diferencial de los TEs con comando squire call.

Por otro lado, se identificaron las zonas enriquecidas con marcas epigenéticas. En primer lugar se hizo el alineamiento de los archivos de ChIP-Seq en el genoma de *Homo sapiens* correspondiente a la versión 38 (Accession de NCBI: 5800238). Para realizar el alineamiento se usó Bowtie2. Para los datos correspondientes a ChIP-seq, luego se hizo un *peak calling*. Para esto se usó el programa MACS (Zhang et al., 2008) en su versión 2.2.7.1. Este programa modela el tamaño del desplazamiento de las de los reads de un ChIP-Seq, y lo usa para

mejorar la resolución de los sitios de unión predichos. También, usa una distribución de Poisson, donde λ será el número de lecturas esperadas en una cierta ventana de un tamaño dado. De este modo poder capturar de manera efectiva los sesgos locales en el genoma, lo que permite una mejor predicción. Esto quiere decir que se seleccionaran los peaks que se expresen más con respecto al resto del fragmento analizado. Este programa se ejecutó con el parámetro -t como el archivo de ChIP-Seq, -c para el archivo input correspondiente, el tamaño de genoma como hs (correspondiente a homo sapiens) y --extsize con un valor siendo obtenido de forma predeterminada por medio del comando predictd de MACS, y tamaño de shift . Para seleccionar los peaks de las muestras se utilizó un valor q menor o igual a 0.05. Una vez obtenida la posición de los peaks seleccionados en las muestras, se procedió a analizar la unión diferencial de las mismas en diferentes condiciones.

Esto se realizó con la herramienta DiffBind (Stark and Brown ,2011). DiffBind es un paquete de R, el cual recibe como entrada un conjunto de peaks en formato csv, que son conjuntos de coordenadas en el genoma que representan un sitio de unión a una proteína predicho por MACS. Cada intervalo consiste en un cromosoma, un comienzo y un final. Cada peak, esté asociado a un tipo de puntaje que indica confianza en ese peak. Cuando ya se tienen los peaks de cada condición se procede a contarlos y ver las diferencias que existen entre ellos, lo que ve reflejado en la unión diferencial de la marca. Esto posteriormente se puede ver reflejado en un gráfico como podría ser un volcano plot. El cual sirve para visualizar cambios en el logaritmo del fold change y asociarlo al logaritmo negativo en base 10 del valor p (Li, 2012).

Finalmente se buscó establecer relaciones entre los datos obtenidos por lo que se hicieron intersecciones de los TEs encontrados de TEcandidates y las marcas epigenéticas diferencialmente encontradas en las condiciones de cada conjunto de datos. Para esto se usó el programa BEDTools (Quinlan and Hall, 2010). Este programa tiene un conjunto de herramientas para el procesamiento de archivos en formato BED, BAM, GFF/GTF, VCF. Entre esas utilidades está el hacer intersección, sustracción, unión, etc. Con las posiciones de elementos anotados en un genoma. De esta forma se pudieron saber los TEs que se predijeron como activos y saber si contienen una marca epigenética asociada a esa zona del genoma.

Resultados preliminares

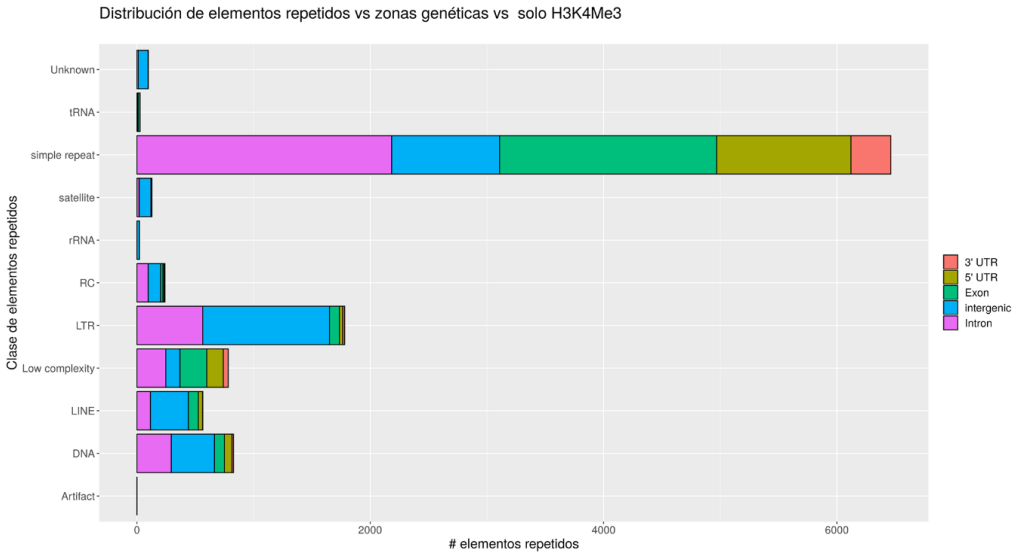

Figura 4 – Gráfico de barras de la intersección entre elementos repetidos en *Drosophila Melanogaster*, zonas genéticas y la marca epigenética H3K4me3.

La Figura 4 forma parte de un estudio previo donde se analizó la intersección de marcas epigenéticas (H3K27me3 y H3K4me3) con elementos repetidos anotados en RepeatMasker y elementos repetidos en *Drosophila Melanogaster*. En este caso solo se muestra con intersección con la marca H3K4me3. A pesar de que el genoma humano no es igual que el de la mosca de la fruta, se suele usar como modelo de estudio para la investigación de variadas enfermedades. Por lo tanto, podría servir como un precedente. Si bien, aún no se comprueba que pueda existir una relación entre TEs y marcas epigenéticas, se espera que pueda haber una intersección entre estos elementos. Esto se basa unos resultados previos realizados en *Drosophila Melanogaster* (figura 4), en donde se logra apreciar una cantidad considerable de TEs del tipo LTR que intersectan con la marca H3K4me3. Un resultado similar se observó al ser intersectado con la marca H3K27me3. Por lo que se hubiera esperado encontrar elementos repetidos del tipo Simple Repeats, seguido de LTRs, que estén relacionados con marcas epigenéticas.

Resultados

Predicción de TEs activos

En primer lugar se procedió a analizar los TEs que fueron predichos con actividad transcripcional. De este modo conocer la cantidad de TEs que fueron reportados por cada programa usado, y los que pertenecían a cada condición; que, para este trabajo, fueron muestras de pacientes con cáncer de próstata y pacientes sanos.

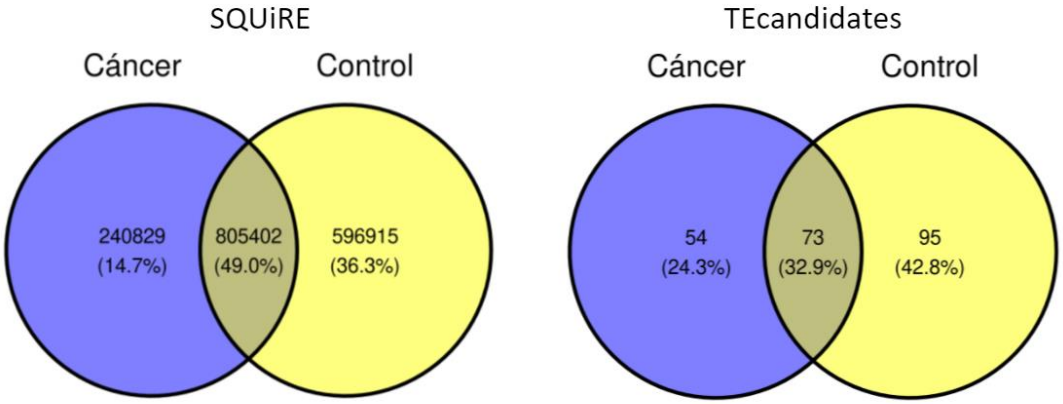

Figura 5 – Diagrama de Venn con el número de TEs predichos de SQUiRE y TEcandidates en las condiciones analizadas y los elementos en común en cada intersección.

Luego de la predicción de TEs activos, resultante por cada programa, se puede observar que ambos reportar proporciones similares (teniendo una mayor proporción de TEs para la condición de control); aunque las escalas de estos varían considerablemente.

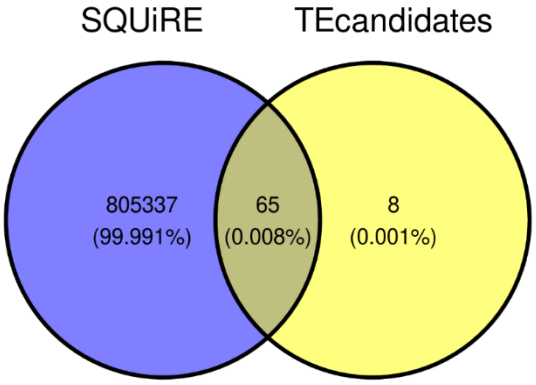

Figura 6 – Diagrama de Venn con el número de TEs predichos como activos por cada programa y la cantidad de TEs en común entre cáncer y control de cada programa.

A pesar de la diferencia en el número de TEs predichos, entregados por cada programa, se logra observar que la mayoría de las predicciones de TEcandidates están contenidas por las reportadas en SQUiRE (Figura 6).

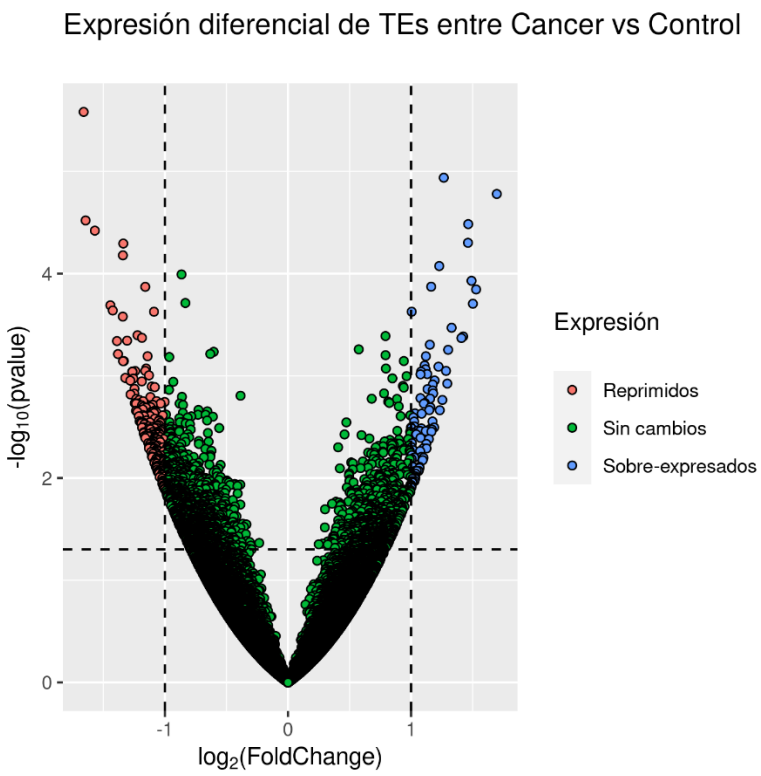

Figura 7 – Volcano plot de la expresión diferencial vs. valor p de TEs predichos por SQUiRE entre muestras de pacientes con cáncer de próstata y pacientes sanos. La línea punteada vertical corresponde a un fold change de -1 y 1, respectivamente. La línea punteada horizontal corresponde a un valor P de 0.05.

En la Figura 7 se aprecia que el cambio en la expresión de los TEs, en su mayor parte, no presentan cambios significativos. La mayoría de los TEs no sobrepasa un umbral mínimo de un fold change de 1; y los que si lo hacen, son están bastante cercanos al umbral.

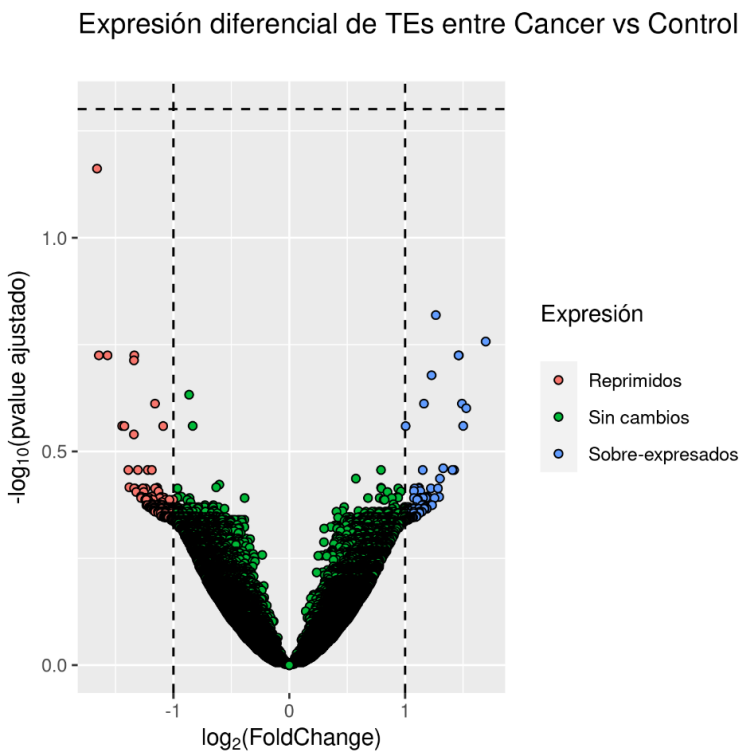

Figura 8 – Volcano plot de la expresión diferencial vs. valor p ajustado de TEs predichos por SQUiRE entre muestras de pacientes con cáncer de próstata y pacientes sanos. La línea punteada vertical corresponde a un fold change de -1 y 1 respectivamente. La línea punteada horizontal corresponde a un valor p ajustado de 0.05 en escala logarítmica.

También, se analizó el fold change de los TEs encontrados contra el valor P ajustado, donde se observaron valores por debajo un umbral mínimo de significancia. Esto es debido a la considerable cantidad de TEs que reportaron valores P bajos.

Después se procedió a realizar este mismo análisis de expresión diferencial con los TEs que ambos programas predijeron como transcripcionalmente activos. Si bien, ambos programas realizan la predicción de TEs activos de formas diferentes, se desearía que los TEs fuese entregados ambos softwares sean más confiables.

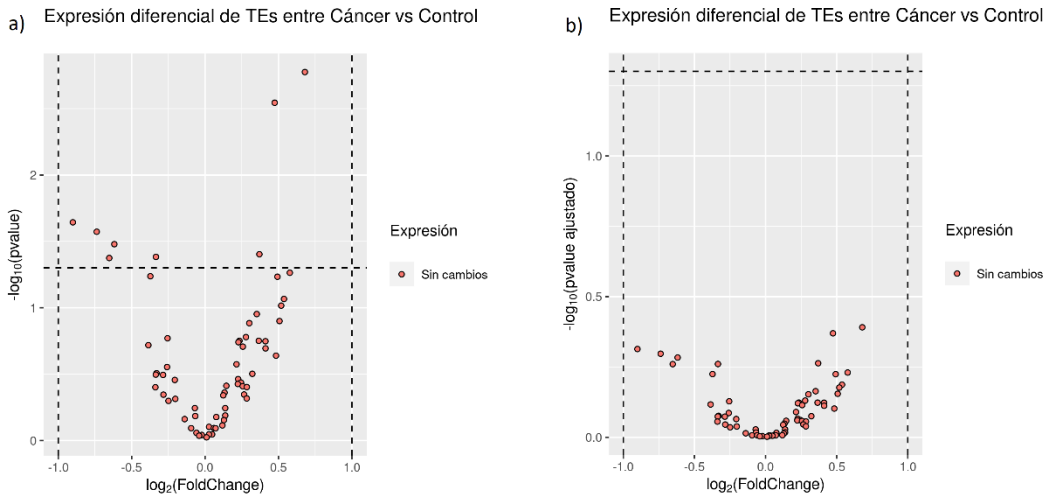

Figura 9 – Volcano plot de la expresión diferencial vs valor p (a) y vs. FDR (b) de TEs predichos como activos en común por TEcandidates y SQUiRE. La línea punteada vertical corresponde a un fold change de -1 y 1 respectivamente. Las líneas punteadas horizontales corresponden a un valor p y valor p ajustado de 0.05, respectivamente en cada gráfico.

Al igual que con los resultados de SQUiRE, los TEs que fueron reportados por ambos programas una baja significancia estadística en algunos casos.

Como parte de este trabajo, se analizaron las marcas de modificaciones de colas de histonas (H3K4me3, H3K27me3 y H3K27ac), presentes en pacientes con cáncer de próstata y sanos, donde las muestras pertenecían a los mismos pacientes que fueron usados para el análisis de TEs.

Luego de analizar los resultados entregados por DiffBind que el programa reporta el número de marcas epigenéticas consenso entre las muestras entregadas. Por lo que, a un mayor número de muestras de pacientes entregadas, menor será la cantidad de peaks encontrados.

Una vez obtenidos, tanto los TEs y peaks de marcas epigenéticas, se pasó a analizar si había una relación entre estos elementos.

Identificación marcas epigenéticas.

		LTR		LINE		SINE		DNA	
		Activo		Activo		Activo		Activo	
		Si	No	Si	No	Si	No	Si	No
H3K4me3	Si	0	0	0	0	0	0	0	0
	No	24	5087	49	9571	12	11823	5	3050
H3K27me3	Si	0	0	0	0	0	0	0	0
	No	24	5087	49	9571	12	11823	5	3050
H3K27ac	Si	0	18	0	10	0	10	0	0
	No	24	5069	49	9561	12	11813	5	3050

Tabla 1 – Esquema de la intersección de TEs, activos o no, con la presencia o ausencia de marcas epigenéticas. La primera columna corresponde a cada una de las marcas estudiadas. La primera fila define las familias de TEs analizados, los que se dividen dependiendo la expresión diferencial obtenida para cada uno, siendo activo si tuvo un fold change mayor a 1, o no activo si este valor estaba entre -1 y 1. La escala de colores representa la cantidad de TEs, que va desde rojo a verde, donde rojo presenta la menor cantidad y verde la mayor.

A raíz de la información presentaba en la Tabla 1, se puede observar que la mayor parte de los TEs que fueron predichos como activos no presentan marcas epigenéticas. Es más, ningún TE clasificado como diferencialmente sobre-expresados presentó marcas epigenéticas. Y todos los que presentan marcas, solamente H3K27ac, son no activos, sugiriendo que esta marca epigenética podría ser represiva. Además, esta marca pareciera ser específica de las subclases LTR y no-LTR, de los retrotransposones.

Debido al alto número de TEs que no tienen marcas y no son activos, un test exacto de Fisher no presenta significancia estadística. Este se utiliza para examinar la significación de la asociación (de contingencia) entre los dos tipos de clasificación

De las 3 marcas estudiadas, únicamente H3K27ac presenta interacción con TEs de la clase retrotransposon, los que no están siendo sobre-expresados. Además, se aprecia que ninguna de las otras marcas presenta una relación.

Debido a que la marca H3K27ac fue la única en la que se observó interacción con TEs, y lo logro apreciar una proporcionalidad entre TEs activos sin marca y TE no activos con la marca, se decidió seleccionar estos TEs para los análisis posteriores.

Evaluar correlación entre TEs y Marcas epigenéticas

Clase/Subclase	Cantidad	Proporción
SINE	14934	46%
LINE	9677	29%
LTR	5148	16%
DNA	3079	9%

Tabla 2 – Cantidad de TEs analizados clasificados por familia y la proporción de cada uno.

Se aprecia que a mayor parte de los TEs predichos como activos corresponde a la familia SINE.

LTR Activos Sin Marca		LTR Inactivos Con Marca	
Superfamilia	Número de TEs	Tipo	Número de TEs
ERV1	6	ERV1	8
ERVK	1	ERVK	5
ERVL	4	ERVL	2
ERVL-MaLR	12	ERVL-MaLR	3
Gypsy	1		

SINE Activos Sin Marca		SINE Inactivos Con Marca	
Superfamilia	Número de TEs	Tipo	Número de TEs
Alu	6	Alu	9
MIR	6	MIR	1

LINE Activos Sin Marca		LINE Inactivos Con Marca	
Superfamilia	Número de TEs	Superfamilia	Número de TEs
CR1	1	L1	10
L2	7		

DNA Activos Sin Marca	
Superfamilia	Número de TEs
TcMar-Tigger	2
hAT-Charlie	3

Tabla 3 – Número de TEs por Superfamilia de clase (DNA) y subclase (LINE, SINE) seleccionadas de la tabla 1. En ese caso fueron TEs inactivos con marca H3K27ac y TEs activos sin marca H3K27ac.

Se logró determinar que las superfamilias de los TEs encontradas para la subclase SINE se concentraban en Alu y MIR. De la misma forma, para la subclase LTR, se observa una predominancia de la superfamilia de ERVs (Tabla 3).

Con los TEs detallados, se analizó si estos presentaban interacción con genes. Además, se estableció cuantos de esos estaban siendo diferencialmente expresados.

	Número de genes	sobre-expresados*	Reprimidos*
LTR Activos Sin Marca	9	1	0
LINE Activos Sin Marca	28	3	0
SINE Activos Sin Marca	9	0	0
DNA Activos Sin Marca	3	0	0
LTR Inactivos Con Marca	11	0	1
LINE Inactivos Con Marca	5	0	0
SINE Inactivos Con Marca	6	0	1

Tabla 4 – Número de genes que intersectan con los TEs de cada subclase de la tabla 1. (*) Cantidad de genes, de los recién mencionados, con fold change mayor a 1 y menor a -1, respectivamente; independiente de su significancia estadística.

Se puede observar, a partir de la tabla 4, que hay una gran proporción, cercano a la mitad, de TEs de la tabla 3 que intersectan con genes.

Genes sobre-expresados de LTR activos-Sin Marca			
Nombre	Fold change	Pvalue	padj
WFDC12	1.0906	0.0068	0.4500

Genes reprimidos de LTR Inactivos-Con Marca			
Nombre	Fold change	Pvalue	padj
HSD17B6	-1.1800	0.0031	0.4306

Genes sobre-expresados de LINE activos-Sin Marca			
Nombre	Fold change	Pvalue	padj
BTNL9	1.0980	0.0005	0.3494
ZNF208	1.3045	0.0008	0.3859
KCNJ3	1.2643	0.0010	0.3923

Genes reprimidos de SINE Inactivos-Con Marca			
Nombre	Fold change	Pvalue	padj
HSD17B6	-1.1800	0.0031	0.4306

Tabla 5 – Expresión diferencial de genes sobre-expresados y reprimidos que intersectan con TEs por Superfamilia de clase (DNA) y subclase (LINE, SINE) seleccionadas de la tabla 1.

Finalmente, se realizó un análisis gene Ontology en pantherdb de los genes de la tabla 5 de forma de establecer los procesos biológicos en los que están involucrados (Thomas, P. D. *et al.*, 2003).

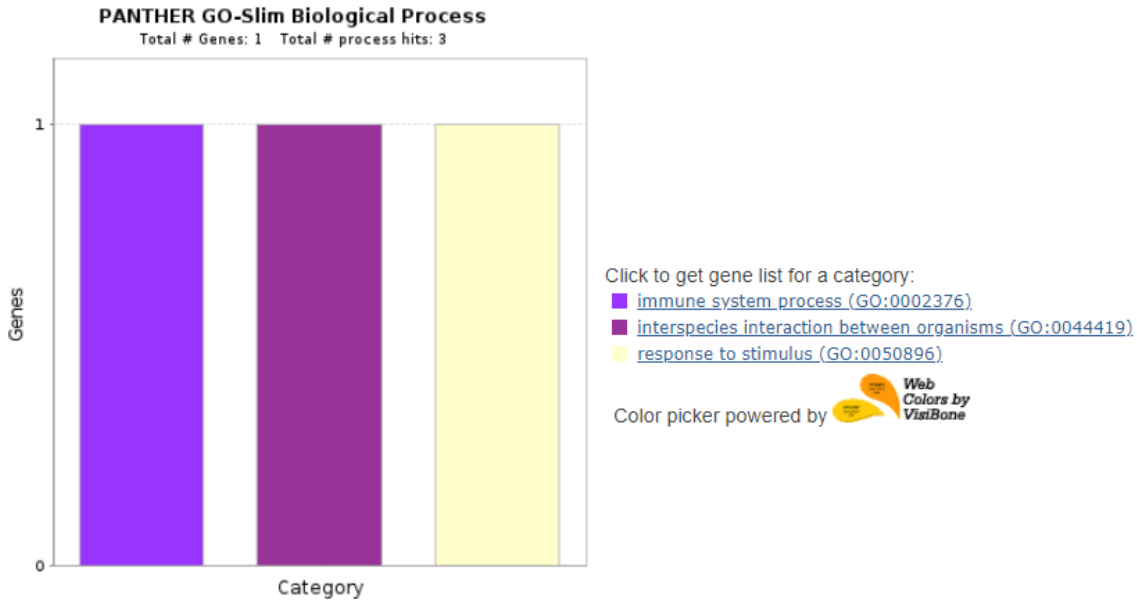

Figura 10 - Análisis genes sobre-expresados de LTR activos sin marca H3K27ac.

Como se puede comprobar en la Figura 10, el gen WFD12 está involucrado en procesos del sistema inmune, respuesta ante estímulos e interacción entre organismos.

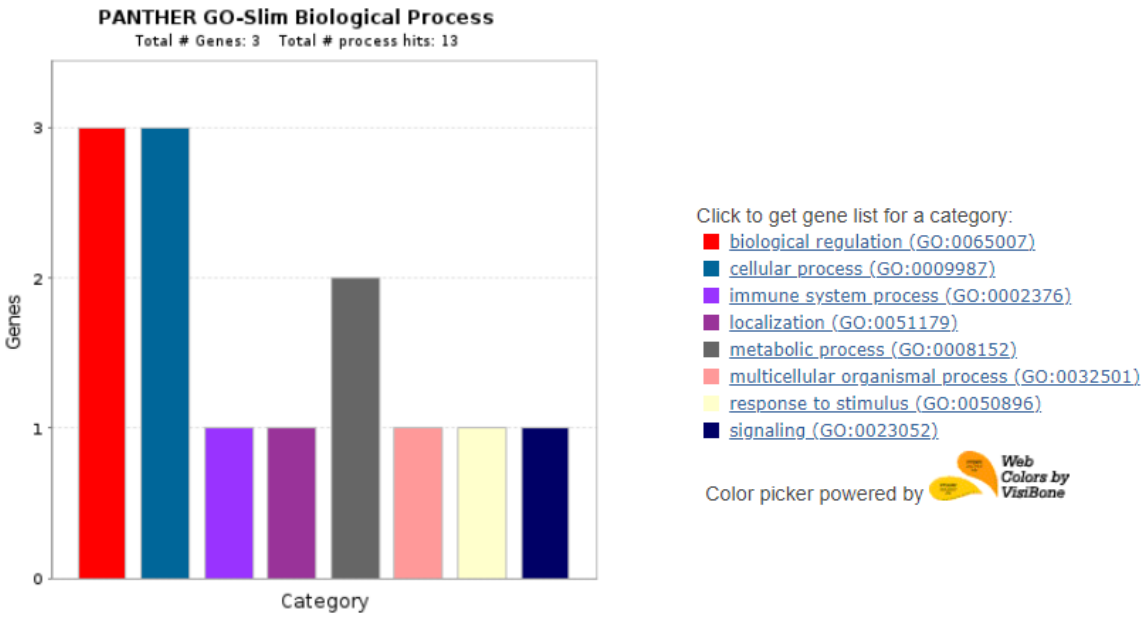

Figura 11 – Análisis Gene Ontology (GO) de genes sobre-expresados de LINE activos sin marca H3K27ac.

En la Figura 11 se observa que los genes sobre-expresados que intersectan con TEs de la familia LINE están involucrados con la regulación biológica.

En el caso del gen HSD17B6, este posee una función de deshidrogenasa. Se asocia a la biosíntesis de andrógeno (testosterona).

Discusión y conclusiones

Para este trabajo se decidió usar un subconjunto de los datos, excluyendo las muestras de ChIP-Seq correspondientes a receptor de andrógeno; dado que estaba fuera de lo que se quería analizar en este estudio.

Para esta investigación, se optó en primera instancia por analizar un fenómeno caótico en términos regulatorios a nivel celular, como puede ser la aparición de cáncer. En esa condición hay evidencia de la presencia de TEs activos.

La mayor parte de los TEs predichos no tiene una expresión diferencial significativa según el umbral normalmente establecido para genes. Además, la mayoría de los TEs no presenta marcas epigenéticas.

La marca H3K27ac es la única que presenta intersección con TEs predichos, específicamente con los pertenecientes a la clase retrotransposones. Las otras marcas analizas no presentaron relación con TEs. Esto podría sugerir que las marcas epigenéticas actúan de forma distinta dependiendo del organismo.

Se esperaba que el número de TEs con marca por familia fuese proporcional a la cantidad de TEs en el genoma. Pero, los TEs de la familia LTR presentan un mayor número de marcas de H3K27ac. Lo que podría dar indicios de una prefería de esta marca por unirse a retrotransposones LTR. Además, esto podría ser una comprobación de los resultados preliminares, en donde los TEs del tipo LTR predominan.

Se observa que el número de TEs activos encontrados sin marca H3K27ac y los TEs no activos con marca H3K27ac presentan cantidades similares. Se podría esperar que exista una relación entre estos elementos, o bien, que pertenezcan a al mismo tipo de TE. Esto se puede comprobar en LTRs y SINEs, donde hubo familias particulares de TEs de las superfamilias ERVs y SINE.

Los resultados de la tabla 3 sugieren que la marca H3K27ac podría estar relacionadas a las Superfamilias Alu y MIR, de la subclase SINE. Además, se indica que la marca H3K27ac puede estar relacionada con ERVs, en particular con ERV1, ERVK y ERVL.

De estos genes encontrados, hubo 4 genes sobre-expresados que intersectaron con TEs activos que no tenían la marca. Los genes sobre-expresados fueron asociado a regulación biología y respuesta inmune que podría deberse a la condición analizada. También, hubo 1 gene reprimidos que fue intersectado por 2 TEs no activos que tenían la marca. De hecho, los Si bien, los genes encontrados fueron localizados en la zona promotora o bien en el último exón de los genes, la cantidad de ejemplos es muy poca para realizar una inferencia al respecto.

Los TEs que se encontraron reprimidos por la marca H3K27me3, pertenecen a zonas intergénicas. Esto podría sugerir que esta marca actúa en los TEs de forma inversa a como se reporta para los genes.

Cabe mencionar, que en este trabajo se seleccionaron 3 marcas particulares, existiendo una gran variedad (sobre 30) de tipos de modificaciones postraduccionales en las colas de histonas. Junto a esto, también hay otras marcas epigenéticas como metilaciones, fosforilaciones, acetilaciones, etc. Es por esto, que no se puede descartar una relación directa de otra marca con una familia de TEs en específico. De hecho, se han reportaron TEs de la familia LTR, del tipo ERV que son regulados por la marca H3K9me3 en células madre embrionarias de ratón (Bulut-Karslioglu, A. et al., 2014). Por lo que se podría necesitar una mayor variedad de marcas al momento de establecer una relación entre TEs activos y marcas epigenéticas.

Considerando, que todo lo planteado en este trabajo sea correcto, se puede establecer que hay indicios de una relación entre las marcas H3K27ac y retrotransposones predichos como activos en cáncer de próstata, en humano.

Referencias

1. Agostini, et al. (2021). Intergenic RNA mainly derives from nascent transcripts of known genes. *Genome Biol.* 22, 1–19
2. Andrews, S. (2010). FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
3. Aravin, A. A. *et al.* (2008). A piRNA Pathway Primed by Individual Transposons Is Linked to De Novo DNA Methylation in Mice. *Mol. Cell* 31, 785–799
4. Bajrami, E., Spiroski, M. (2016). Genomic imprinting. *Open Access Macedonian Journal of Medical Sciences* 4, 181–184
5. Biscotti, M. A., Olmo, E., Heslop-Harrison, J. S. (Pat. (2015). Repetitive DNA in eukaryotic genomes. *Chromosom. Res.* 23, 415–420
6. Bolger, A. M., Lohse, M., Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120
7. Bulut-Karslioglu, A. et al. (2014). Suv39h-Dependent H3K9me3 Marks Intact Retrotransposons and Silences LINE Elements in Mouse Embryonic Stem Cells. *Mol. Cell* 55, 277–290
8. Bundo, M. *et al.* (2014). Increased L1 retrotransposition in the neuronal genome in schizophrenia. *Neuron* 81, 306–313
9. Chang, T. C. *et al.* (2019). Investigation of somatic single nucleotide variations in human endogenous retrovirus elements and their potential association with cancer. *PLoS One* 14,
10. Cordaux, R., Batzer, M. A. (2009). The impact of retrotransposons on human genome evolution. *Nature Reviews Genetics* 10, 691–703
11. Deniz, Ö., Frost, J. M., Branco, M. R. (2019). Regulation of transposable elements by DNA modifications. *Nature Reviews Genetics* 20, 417–431
12. Ewels, P. SRA Explorer. https://sra-explorer.info/
13. Frost, L. S., Leplae, R., Summers, A. O., Toussaint, A. (2005). Mobile genetic elements: The agents of open source evolution. *Nature Reviews Microbiology* 3, 722–732
14. Furey, T. S. (2012). ChIP-seq and beyond: New and improved methodologies to detect and characterize protein-DNA interactions. *Nature Reviews Genetics* 13, 840–852
15. Furey, T. S. (2012). ChIP-seq and beyond: New and improved methodologies to detect and characterize protein-DNA interactions. *Nature Reviews Genetics* 13, 840–852
16. Guo, C. *et al.* (2018). Tau Activates Transposable Elements in Alzheimer's Disease.

Cell Rep. 23, 2874–2880
17. Juretic, N., Bureau, T. E., Bruskiewich, R. M. (2004). Transposable element annotation of the rice genome. *Bioinformatics* 20, 155–160
18. Kaminker, J. S. *et al.* (2002). The transposable elements of the Drosophila melanogaster euchromatin: a genomics perspective. *Genome Biol.* 3, RESEARCH0084
19. de Koning, A. P. J., Gu, W., Castoe, T. A., Batzer, M. A., Pollock, D. D. (2011). Repetitive elements may comprise over Two-Thirds of the human genome. *PLoS Genet.* 7, e1002384
20. Lander, S. *et al.* (2001). Initial sequencing and analysis of the human genome. *Nature* 409, 860–921
21. Leinonen, R., Sugawara, H., Shumway, M. (2011). The sequence read archive. *Nucleic Acids Res.* 39,
22. Li, W. (2012). Volcano plots in analyzing differential expressions with mRNA microarrays. *J. Bioinform. Comput. Biol.* 10, 1231003
23. Lozano, R., Marin, R., Santacruz, M.-J., Pascual, A. (2015). Long interspersed retrotransposable elements and susceptibility to schizophrenia. *Acta Neuropsychiatr.* 27, 195–6
24. Luning Prak, E. T., Kazazian, H. H. (2000). Mobile elements and the human genome. *Nature Reviews Genetics* 1, 134–144
25. Molaro, A. *et al.* (2014). Two waves of de novo methylation during mouse germ cell development. *Genes Dev.* 28, 1544–1549
26. Misiak, B., Ricceri, L., Sasiadek, M. M. (2019). Transposable elements and their epigenetic regulation in mental disorders: Current evidence in the field. *Front. Genet.* 10.
27. Munoz-Lopez, M., Garcia-Perez, J. (2010). DNA Transposons: Nature and Applications in Genomics. *Curr. Genomics* 11, 115–128
28. Nie, Y. *et al.* (2020). FANCD2 is required for the repression of germline transposable elements. *Reproduction* 159, 659–668.
29. Oshita, H. et al. (2013). RASEF is a novel diagnostic biomarker and a therapeutic target for lung cancer. *Mol. Cancer Res*. 11, 937–951
30. Ozsolak, F., Milos, P. M. (2011). RNA sequencing: Advances, challenges and opportunities. *Nature Reviews Genetics* 12, 87–98
31. Peters, A. H. F. M. *et al.* (2001). Loss of the Suv39h histone methyltransferases impairs mammalian heterochromatin and genome stability. *Cell* 107, 323–337
32. Quesneville, H. *et al.* (2005). Combined evidence annotation of transposable elements in genome sequences. *PLoS Comput. Biol.* 1, 0166–0175
33. Quinlan, A. R., Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing

genomic features. *Bioinformatics* 26, 841–842
34. Saze, H. (2018). Epigenetic regulation of intragenic transposable elements: a two-edged sword. *J. Biochem.* 164, 323–328
35. Schneider, R. *et al.* (2004). Histone H3 lysine 4 methylation patterns in higher eukaryotic genes. *Nat. Cell Biol.* 6, 73–77
36. Slotkin, R. K., Martienssen, R. (2007). Transposable elements and the epigenetic regulation of the genome. *Nature Reviews Genetics* 8, 272–285
37. Thomas, P. D. et al. (2003). PANTHER: A library of protein families and subfamilies indexed by function. *Genome Res*. 13, 2129–2141
38. Valdebenito-Maturana, B., Riadi, G. (2018). TEcandidates: prediction of genomic origin of expressed transposable elements using RNA-seq data. *Bioinformatics* 34, 3915–3916
39. Wicker, T. *et al.* (2007). A unified classification system for eukaryotic transposable elements. *Nature Reviews Genetics* 8, 973–982
40. 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 Res.* 47, e27
41. Zhang, Y. *et al.* (2008). Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* 9, R137

Apéndice 1

Muestras de ChIP-Seq usadas de la marca H3K27ac.

SRR ID	Patient ID	Epigenetic mark	Sequencing Technique
SRR7949163	P223T	H3K27ac	ChIP-Seq
SRR7949171	P229T	H3K27ac	ChIP-Seq
SRR7949252	P230T	H3K27ac	ChIP-Seq
SRR7949195	P261T	H3K27ac	ChIP-Seq
SRR7949234	P265T	H3K27ac	ChIP-Seq
SRR7949216	P268T	H3K27ac	ChIP-Seq
SRR7949251	P274T	H3K27ac	ChIP-Seq
SRR7949173	P276T	H3K27ac	ChIP-Seq
SRR7949235	P294T	H3K27ac	ChIP-Seq
SRR7949203	P301T	H3K27ac	ChIP-Seq
SRR7949215	P309T	H3K27ac	ChIP-Seq
SRR7949214	P310T	H3K27ac	ChIP-Seq
SRR7949236	P355T	H3K27ac	ChIP-Seq
SRR7949198	P431T	H3K27ac	ChIP-Seq
SRR7949191	P449T	H3K27ac	ChIP-Seq
SRR7949193	P500T	H3K27ac	ChIP-Seq
SRR7949213	P506T	H3K27ac	ChIP-Seq
SRR7949184	P527T	H3K27ac	ChIP-Seq
SRR7949217	P536T	H3K27ac	ChIP-Seq
SRR7949218	P643T	H3K27ac	ChIP-Seq
SRR7949186	P710T	H3K27ac	ChIP-Seq
SRR7949219	P717T	H3K27ac	ChIP-Seq
SRR7949187	P730T	H3K27ac	ChIP-Seq
SRR7949220	P737T	H3K27ac	ChIP-Seq
SRR7949207	P823T	H3K27ac	ChIP-Seq

Muestras de ChIP-Seq usadas de la marca H3K4me3.

SRR ID	Patient ID	Epigenetic mark	Sequencing Technique
SRR7949254	P223T	H3K4me3	ChIP-Seq
SRR7949256	P229T	H3K4me3	ChIP-Seq
SRR7949257	P230T	H3K4me3	ChIP-Seq
SRR7949260	P261T	H3K4me3	ChIP-Seq
SRR7949262	P265T	H3K4me3	ChIP-Seq
SRR7949263	P268T	H3K4me3	ChIP-Seq
SRR7949265	P274T	H3K4me3	ChIP-Seq
SRR7949266	P276T	H3K4me3	ChIP-Seq
SRR7949270	P294T	H3K4me3	ChIP-Seq
SRR7949274	P301T	H3K4me3	ChIP-Seq
SRR7949275	P309T	H3K4me3	ChIP-Seq
SRR7949276	P310T	H3K4me3	ChIP-Seq
SRR7949279	P355T	H3K4me3	ChIP-Seq
SRR7949280	P431T	H3K4me3	ChIP-Seq
SRR7949282	P449T	H3K4me3	ChIP-Seq
SRR7949285	P500T	H3K4me3	ChIP-Seq
SRR7949286	P506T	H3K4me3	ChIP-Seq
SRR7949290	P527T	H3K4me3	ChIP-Seq
SRR7949293	P536T	H3K4me3	ChIP-Seq
SRR7949298	P643T	H3K4me3	ChIP-Seq
SRR7949302	P710T	H3K4me3	ChIP-Seq
SRR7949303	P717T	H3K4me3	ChIP-Seq
SRR7949304	P730T	H3K4me3	ChIP-Seq
SRR7949305	P737T	H3K4me3	ChIP-Seq
SRR7949308	P823T	H3K4me3	ChIP-Seq

Muestras de ChIP-Seq usadas de la marca H3K27me3.

SRR ID	Patient ID	Epigenetic mark	Sequencing Technique
SRR7949311	P223T	H3K27me3	ChIP-Seq
SRR7949314	P229T	H3K27me3	ChIP-Seq
SRR7949375	P230T	H3K27me3	ChIP-Seq
SRR7949335	P261T	H3K27me3	ChIP-Seq
SRR7949369	P265T	H3K27me3	ChIP-Seq
SRR7949352	P268T	H3K27me3	ChIP-Seq
SRR7949380	P274T	H3K27me3	ChIP-Seq
SRR7949315	P276T	H3K27me3	ChIP-Seq
SRR7949370	P294T	H3K27me3	ChIP-Seq
SRR7949340	P301T	H3K27me3	ChIP-Seq
SRR7949351	P309T	H3K27me3	ChIP-Seq

SRR7949350	P310T	H3K27me3	ChIP-Seq
SRR7949371	P355T	H3K27me3	ChIP-Seq
SRR7949336	P431T	H3K27me3	ChIP-Seq
SRR7949332	P449T	H3K27me3	ChIP-Seq
SRR7949333	P500T	H3K27me3	ChIP-Seq
SRR7949349	P506T	H3K27me3	ChIP-Seq
SRR7949326	P527T	H3K27me3	ChIP-Seq
SRR7949353	P536T	H3K27me3	ChIP-Seq
SRR7949354	P643T	H3K27me3	ChIP-Seq
SRR7949328	P710T	H3K27me3	ChIP-Seq
SRR7949355	P717T	H3K27me3	ChIP-Seq
SRR7949329	P730T	H3K27me3	ChIP-Seq
SRR7949356	P737T	H3K27me3	ChIP-Seq
SRR7949345	P823T	H3K27me3	ChIP-Seq

Muestras de RNA-Seq usadas.

SRR ID	Patient ID	Sequencing Technique
SRR7949386	P223T	RNA-Seq
SRR7949393	P229T	RNA-Seq
SRR7949468	P230T	RNA-Seq
SRR7949445	P261T	RNA-Seq
SRR7949387	P265T	RNA-Seq
SRR7949396	P268T	RNA-Seq
SRR7949473	P274T	RNA-Seq
SRR7949399	P276T	RNA-Seq
SRR7949400	P294T	RNA-Seq
SRR7949446	P301T	RNA-Seq
SRR7949409	P309T	RNA-Seq
SRR7949402	P310T	RNA-Seq
SRR7949459	P355T	RNA-Seq
SRR7949435	P431T	RNA-Seq
SRR7949423	P449T	RNA-Seq
SRR7949428	P500T	RNA-Seq
SRR7949429	P506T	RNA-Seq
SRR7949477	P527T	RNA-Seq
SRR7949441	P536T	RNA-Seq
SRR7949455	P643T	RNA-Seq
SRR7949408	P710T	RNA-Seq
SRR7949442	P717T	RNA-Seq
SRR7949443	P730T	RNA-Seq
SRR7949414	P737T	RNA-Seq
SRR7949418	P823T	RNA-Seq