UNIVERSIDAD DE COSTA RICA

SISTEMA DE ESTUDIOS DE POSGRADO

COMPARACIÓN EVOLUTIVA DE RETROVIRUS ENDÓGENOS (ERV) TRANSCRIPCIONALMENTE ACTIVOS EN TEJIDO TESTICULAR DE TRES ESPECIES DE PRIMATES

Tesis sometida a la consideración de la Comisión del Programa de Estudios de Posgrado en Ciencias Biomédicas para optar al grado y título de Maestría Académica en Bioinformática y Biología de Sistemas

IZAYANA SANDOVAL CARVAJAL

Ciudad Universitaria Rodrigo Facio, Costa Rica

2020

A mi madre

Agradecimientos

Agradezco en primer lugar a Dios por permitirme llegar hasta aquí y poder concluir esta etapa, así como por darme las fuerzas para superar cada nuevo reto que se me ha presentado a lo largo de mi formación académica.

Les agradezco a mis papás que siempre me han brindado todo su apoyo y que con mucho esfuerzo y sacrificio me han dado todo para que yo pudiera dedicarme a mis estudios, todo se lo debo a ellos.

A los miembros de mi comité de tesis. A mi tutora Rebeca, gracias por permitirme ser parte de este proyecto y por toda su ayuda, paciencia y compromiso durante este tiempo. Siempre que la necesité estuvo ahí para ayudarme y aclarar mis dudas. Gracias a mis lectores por todas sus valiosas sugerencias y ayuda. Agradezco también a Edgardo Camacho por ayudarme con el análisis estadístico de los datos.

Gracias a mis amigos que siempre han estado para motivarme y darme su apoyo en todo momento. A mis compañeros de la maestría, que fueron un gran apoyo durante el periodo de clases y que sin duda alguna hicieron que esos años de estudio fueran más fáciles y amenos. A mis compañeros del CIBCM que siempre estuvieron pendientes de mi trabajo, me apoyaron, me aconsejaron y me motivaron durante este tiempo.

A mi querida UCR, que a lo largo de 13 años me ha dado tantas cosas, desde académicas hasta lecciones de vida, nuevos amigos, experiencias, muchísimas cosas que llevaré conmigo por siempre. A todos los profesores que a lo largo de mis años de estudio me han inspirado y motivado con sus enseñanzas. Gracias a Elvira y a Denisse del posgrado de ciencias biomédicas por su disposición de ayudarme en todo lo que necesité durante estos años que fui parte del posgrado.

Al CNCA del CENAT por facilitarme el uso del cluster, por toda su colaboración y paciencia siempre que necesité de la ayuda de ellos. Gracias a la disponibilidad de uso de la infraestructura computacional del CNCA pude realizar todos mis análisis bioinformáticos. Es un recurso muy valioso al que por suerte tenemos acceso libremente. Gracias al CICIMA también por facilitarme el acceso al cluster de este centro.

En general gracias a todas las personas que de una u otra forma me ayudaron a lo largo de este proceso.

¡Muchas gracias a todos!

"Esta tesis fue aceptada por la Comisión del Programa de Estudios de Posgrado en Ciencias Biomédicas de la Universidad de Costa Rica, como requisito parcial para optar al grado y título de Maestría Académica en Bioinformática y Biología de Sistemas"

CECILIA DIAZ OREIRO (FIRMA)
Firmado digitalmente por CECILIA DIAZ OREIRO (FIRMA)
Fecha: 2020.06.26 15:27:06 -06'00'

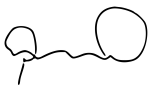
Dra. Cecilia Díaz Oreiro
Representante del Decano
Sistema de Estudios de Posgrado

REBECA CAMPOS SANCHEZ (FIRMA)
Digitally signed by REBECA CAMPOS SANCHEZ (FIRMA)
Date: 2020.06.25 16:51:43 -06'00'

Dra. Rebeca Campos Sánchez
Directora de tesis

JOSE ARTURO MOLINA MORA (FIRMA)
Firmado digitalmente por JOSE ARTURO MOLINA MORA (FIRMA)
Fecha: 2020.06.26 16:48:46 -06'00'

M.Sc. José Arturo Molina Mora
Lector


Firmado digitalmente por César Rodríguez
Fecha: 2020.06.29 09:18:45 -06'00'

Dr. César Rodríguez Sánchez
Lector

MAHMOUD SASA MARIN (FIRMA)
Firmado digitalmente por MAHMOUD SASA MARIN (FIRMA)
Fecha: 2020.07.06 20:16:28 -06'00'

Dr. Mahmood Sasa Marín
Representante de la directora del Posgrado en Ciencias Biomédicas

IZAYANA SANDOVAL CARVAJAL (FIRMA)
Firmado digitalmente por IZAYANA SANDOVAL CARVAJAL (FIRMA)
Fecha: 2020.07.06 20:39:09 -06'00'

Izayana Sandoval Carvajal
Sustentante

ÍNDICE

RESUMEN

Los elementos transponibles (ET) son secuencias que tienen la capacidad de cambiar su posición en el genoma y regular la expresión de genes cercanos. Los retrovirus endógenos (ERVs) son un tipo de elemento transponible que se cree se originaron a partir de infecciones ancestrales de retrovirus exógenos que lograron fijarse en el genoma de primates hace millones de años. Estos han originado ciertos genes mediante el proceso de exaptación y han contribuido en gran medida a la variabilidad genética y diferenciación de los humanos de las otras especies de primates. Sin embargo, aún se desconocen los detalles precisos sobre cómo ocurrió esta separación. Dado a que los procesos evolutivos que son heredados a los descendientes ocurren en la línea germinal, el estudio de la expresión de ERVs en tejido testicular puede generar información valiosa sobre como ocurrió esta separación en primates. En este contexto, la expresión de ERVs en tejido germinal y su potencial integración en nuevos sitios genómicos podrían ser la clave para elucidar aspectos evolutivos entre especies relacionadas de primates.

Este estudio tuvo como objetivos: estandarizar un protocolo bioinformático para identificación y estudio de la expresión de ERVs en tejido testicular, determinar la influencia del contexto genómico en la expresión de estos elementos y realizar una comparación evolutiva de los ERVs identificados en librerías de RNA-Seq de tejido testicular humano (n=10), gorila (n=1) y orangután (n=1). Para el ensamblaje del transcriptoma se probaron diferentes parámetros de ensamblaje (diferentes tamaños de Kmers y contigs) y dos estrategias: de *novo* y guiado por genoma de referencia. Solamente transcritos con tamaño >3 kb y cobertura > 5x se consideraron ERVs expresados (eERVs). Estos eERVs fueron anotados con la base de datos RepBase y su localización en el genoma de referencia hg38 fue determinada con la herramienta BLAT. Posterior a su anotación se determinaron características genómicas de los flancos de los eERVs en humano (como proporción de islas CpG, genes, lncRNAs, regiones conservadas, entre otras) y se compararon con las de regiones control (flancos de LTRs no expresados) para determinar la influencia del contexto genómico en la expresión de ERVs. Además se identificaron regiones sinténicas en las tres especies analizadas.

Se encontraron diferencias entre resultados obtenidos con diferentes parámetros y enfoques de ensamblaje. En total se encontraron 19 tipos de ERVs en humanos, de los cuales los más comunes fueron HERV17, HERVK22I y HERVS71, que se identificaron en cinco de las diez librerías de origen humano. En el caso de orangután solamente se identificaron los tipos HARLEQUIN, HERV1_I, HERVE, HERVK y HERVK22I, mientras que en gorila no se logró identificar ningún eERV de tamaño superior a 3kb.

La mayor cantidad de eERVs humanos se identificaron en el cromosoma 7 y la mayor densidad de eERVs en el cromosoma 19, mientras que, por el contrario, en los cromosomas 3, 6, 8,9,15, 20, 21 y Y, no se logró identificar ningún ERV > 3 kb. Cerca del 58% de los ERVs presentó traslape con genes y aproximadamente un 32% presentó traslape con Long noncoding RNAs (lncRNAs). Se encontró que ocurre una replicación más temprana en los flancos de los ERVs expresados en comparación con los flancos de LTRs no expresados y además poseen una mayor proporción de SINEs corriente abajo. Se encontraron ocho copias de ERVs homólogos (misma copia de ERV en mismo locus) entre las tres especies de primates, pero solamente se encontró potencial codificante en dos de

éstas, correspondientes al elemento HERVK en el cromosoma 1 y 11 humano. A partir del análisis de diversidad utilizando este elemento se encontró una mayor divergencia genética entre humano y orangután, entre los cuales se encontró que ocurrió selección purificadora, al igual que entre gorila y orangután. Se encontró además que entre humano y gorila existe neutralidad, es decir que la cantidad de mutaciones sinónimas es igual a la cantidad de mutaciones no sinónimas en la región analizada (dN = dS).
En conclusión, se encontró evidencia de que el contexto genómico influye en la expresión de ERVs, y se demostró homología entre algunos ERVs de humanos y algunos ERVs en gorila y orangután, lo cual demuestra su importancia evolutiva en estas especies.

Palabras clave: elementos transponibles, retrovirus endógenos, HARLEQUIN, HERV17, HERVK22I, HERVS71, HERV1_I, HERVE, HERVK, humano, gorila, orangután lncRNAs, RNA-Seq, sintenia, transcriptoma

LISTA DE CUADROS

LISTA DE FIGURAS

LISTA DE ANEXOS

LISTA DE ABREVIATURAS

ADN	Ácido desoxirribonucleico
ARN	Ácido ribonucleico
BAM	Por sus siglas en inglés, “Binary Alignment Map”
BED	Por sus siglas en inglés, “Browser Extensible Data”
BLAST	Por sus siglas en inglés, “Basic Local Alignment Search Tool”
BLAT	Por sus siglas en inglés, “BLAST-like alignment tool”
Chr	Cromosoma
CENAT	Centro Nacional de Alta Tecnología
CNCA	Colaboratorio Nacional de Computación Avanzada
eERV	ERV expresado
ERV	Por sus siglas en inglés, “Endogenous retrovirus”
ET	Elemento transponible
FDA	Por sus siglas en inglés, “Functional data análisis”
HERV	Por sus siglas en inglés, “Human Endogenous retrovirus”
HTDV	Por sus siglas en inglés, “human teratocarcinoma derived virus”
IGV	Por sus siglas en inglés, “Integrative Genomics Viewer”
lncRNA	Por sus siglas en inglés, “Long noncoding RNAs”
LINE	Por sus siglas en inglés, “Long interspersed nuclear element”
LTR	Por sus siglas en inglés, “Long terminal repeat”
ORF	Por sus siglas en inglés, “Open Reading frame”
RIN	Por sus siglas en inglés, “RNA integrity number”
SAM	Por sus siglas en inglés, “Sequence Alignment Map”
SBS	Por sus siglas en inglés, “Sequencing By Synthesis”
SINE	Por sus siglas en inglés, “Short interspersed nuclear element”
TPM	Por sus siglas en inglés, “Transcripts per million”
UCSC	University of California Santa Cruz

UNIVERSIDAD DE COSTA RICA

SEP Sistema de Estudios de Posgrado

Autorización para digitalización y comunicación pública de Trabajos Finales de Graduación del Sistema de Estudios de Posgrado en el Repositorio Institucional de la Universidad de Costa Rica.

Yo, Izayana Sandoval Carvajal, con cédula de identidad 1-1382 0518, en mi condición de autor del TFG titulado COMPARACIÓN EVOLUTIVA DE RETROVIRUS ENDÓGENOS (ERV) TRANSCRIPCIONALMENTE ACTIVOS EN TEJIDO TESTICULAR DE TRES ESPECIES DE PRIMATES

Autorizo a la Universidad de Costa Rica para digitalizar y hacer divulgación pública de forma gratuita de dicho TFG a través del Repositorio Institucional u otro medio electrónico, para ser puesto a disposición del público según lo que establezca el Sistema de Estudios de Posgrado. **SI** [x] **NO** *[ ]

***En caso de la negativa favor indicar el tiempo de restricción:** _____ **año (s).**

Este Trabajo Final de Graduación será publicado en formato PDF, o en el formato que en el momento se establezca, de tal forma que el acceso al mismo sea libre, con el fin de permitir la consulta e impresión, pero no su modificación.

Manifiesto que mi Trabajo Final de Graduación fue debidamente subido al sistema digital Kerwá y su contenido corresponde al documento original que sirvió para la obtención de mi título, y que su información no infringe ni violenta ningún derecho a terceros. El TFG además cuenta con el visto bueno de mi Director (a) de Tesis o Tutor (a) y cumplió con lo establecido en la revisión del Formato por parte del Sistema de Estudios de Posgrado.

INFORMACIÓN DEL ESTUDIANTE:

Nombre Completo: Izayana Sandoval Carvajal

Número de Carné: A76075 Número de cédula: 1-13820518

Correo Electrónico: izayana.sandoval@ucr.ac.cr

Fecha: 22/06/2020 Número de teléfono: 8533-6660

Nombre del Director (a) de Tesis o Tutor (a): Rebeca Campos Sánchez

FIRMA ESTUDIANTE

Nota: El presente documento constituye una declaración jurada, cuyos alcances aseguran a la Universidad, que su contenido sea tomado como cierto. Su importancia radica en que permite abreviar procedimientos administrativos, y al mismo tiempo genera una responsabilidad legal para que quien declare contrario a la verdad de lo que manifiesta, puede como consecuencia, enfrentar un proceso penal por delito de perjurio, tipificado en el artículo 318 de nuestro Código Penal. Lo anterior implica que el estudiante se vea forzado a realizar su mayor esfuerzo para que no sólo incluya información veraz en la Licencia de Publicación, sino que también realice diligentemente la gestión de subir el documento correcto en la plataforma digital Kerwá.

1. INTRODUCCION

Los elementos transponibles son secuencias de ADN que tienen la capacidad de cambiar su posición en el genoma (Rebollo et al., 2012). Estos pueden influenciar la expresión génica del hospedero, afectándola de manera neutral, positiva o negativamente (Rebollo et al., 2012). Los elementos transponibles pueden sufrir un proceso de "domesticación" también conocido como exaptación, en el que un elemento móvil se integra al material genético del hospedero y desarrolla una nueva función. Algunos genes de importancia biológica son producto de este proceso. Un ejemplo de ello es la telomerasa humana, que se originó como producto de la domesticación de la transcriptasa inversa (Tomás-Loba et al., 2008). En procariotas, gracias a la domesticación de elementos transponibles, se han originado algunos mecanismos de resistencia a los antibióticos y otras innovaciones evolutivas (Lynch & Conery, 2003).

Los retrovirus son un ejemplo de elementos transponibles. Estos pertenecen a la familia *Retroviridae,* la cual es una familia de virus muy diversa (Bannert & Kurth, 2006; Mager & Medstrand, 2005). Los retrovirus normalmente infectan células somáticas, pero en algunos casos pueden infectar células de la línea germinal e insertarse en su genoma, lo que posibilita su herencia vertical. A estos retrovirus se les conoce como retrovirus endógenos o ERVs (Bannert & Kurth, 2006; Mager & Medstrand, 2005).

Diferentes técnicas permiten estudiar la diversidad de transcritos en una muestra e identificar ERVs. Una de ellas es la secuenciación de ARNs o RNA-Seq, la cual, a diferencia de las técnicas basadas en hibridación o secuenciación de Sanger, permite cuantificar su nivel de expresión (Griffith et al., 2010). En este proyecto se analizaron datos obtenidos con la técnica RNA-Seq, en donde inicialmente se realizó una comparación de ensamblajes con y sin genoma de referencia. Posteriormente, se identificaron los transcritos derivados de ERVs según la base de datos RepBase. Se determinó la posición en el genoma de cada uno de los transcritos ensamblados e identificados como ERVs y se evaluó su contexto genómico para determinar su influencia en la expresión de estos elementos.

Esta investigación es relevante porque aporta información que permite comprender mejor los procesos evolutivos influenciados por ERVs en la línea germinal de primates. Particularmente, debido a promotores desmetilados del tejido testicular, los ERVs tienen la flexibilidad de expresarse en este tejido mientras que en el resto del organismo permanecen inactivos en su mayoría, lo que les permitiría modificar el genoma y que estas modificaciones se hereden a la progenie. Comprender estos mecanismos evolutivos que ocurren en la línea germinal puede contribuir también a comprender el papel en ciertas patologías influenciadas por ERVs como el cáncer, esclerosis múltiple y otras en las que se ha visto expresión de ERVs.

2. ANTECEDENTES

2.1 Elementos transponibles

Los transposones o elementos transponibles (ET) fueron descubiertos en la década de 1940 por la genetista Barbara McClintock en el genoma del maíz (Rebollo et al., 2012). Actualmente se sabe que la proporción de elementos transponibles en el genoma de diferentes especies varía desde un 85% en maíz, 44% en humanos, 10% en *Arabidopsis thaliana,* o 0% en otras especies, como *Plasmodium falciparum* (Lander et al., 2001; Rebollo et al., 2012). En algunas especies están involucrados en funciones biológicas muy importantes, como por ejemplo en el mantenimiento de la integridad de centrómeros y telómeros (Rebollo et al., 2012).

La complejidad de los genomas eucariotas se debe, al menos en parte, a la acción de elementos transponibles (Lynch & Conery, 2003). Los elementos móviles tienen una gran capacidad de mutar, la cual es una fuente de innovación evolutiva. Sin embargo, también pueden causar efectos deletéreos en el genoma de sus hospederos, como por ejemplo cuando se insertan en regiones codificantes y afectan la expresión de ese gen (Lynch & Conery, 2003). Debido a esta gran variabilidad de los elementos transponibles es difícil encontrar un criterio de clasificación apropiado y actualmente no existe un comité taxonómico que se encargue de la clasificación de los ET.

La clasificación de los ET ha estado en constante cambio, dado que con los avances de secuenciación y el aumento en la cantidad de información en los últimos años se han descrito nuevas familias (Capy, 2005). Los ET fueron clasificados por primera vez por David Finnegan en el año 1989, quien los separó en dos clases según su mecanismo de transposición y características estructurales (Capy, 2005). La clase I (retrotransposones) agrupa a todos aquellos ET que se transponen mediante transcripción reversa de un intermediario de ARN; mientras que la clase II (transposones ADN) agrupa a todos aquellos elementos que no requieren retrotranscripción y se pueden transponer directamente de ADN a ADN. Más adelante, esta clasificación fue modificada por Capy et al. 1997, en donde se dividió la clase I en dos subclases con base en la presencia/ausencia de LTRs (Long Terminal Repeats): Subclase I o LTR retrotransposones y la Subclase II o no-LTR retrotransposones. A su vez la subclase I se divide en 3 superfamilias y la subclase II se divide en dos superfamilias (Figura 1).

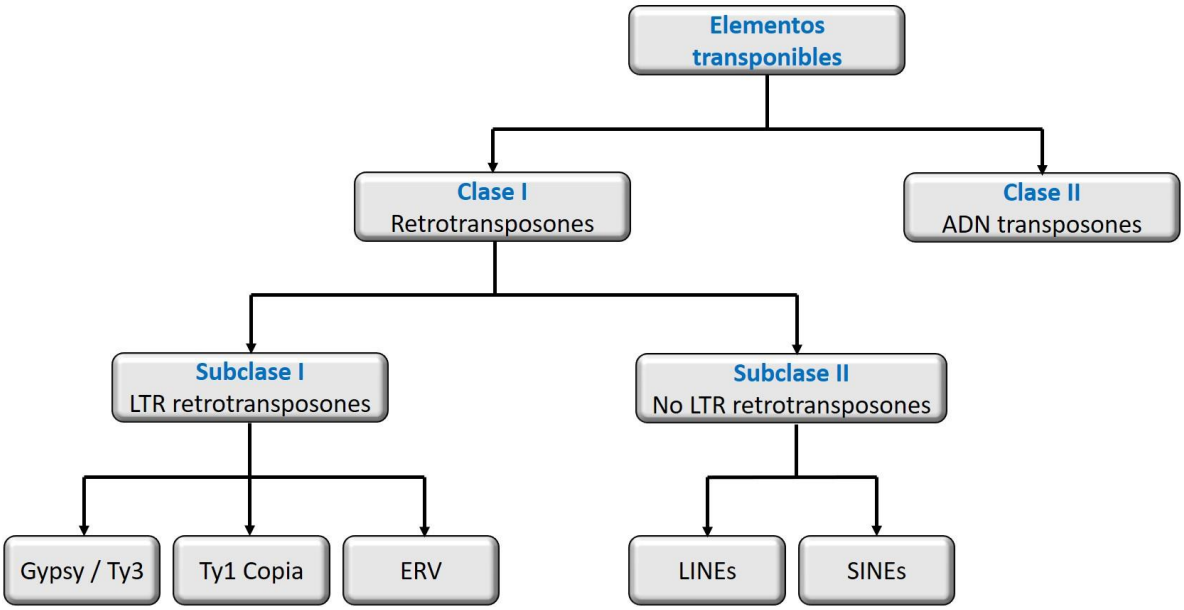


Figura 1. Representación gráfica de la clasificación de los elementos transponibles

2.2 Retrovirus

Los genomas de los virus clasificados en la familia *Retroviridae* poseen cuatro genes principales en el orden 5´-*gag-pro-pol-env*-3´. El gen *gag* codifica para las proteínas de la cápside del virus, *pro* codifica para una proteasa, *pol* para una transcriptasa reversa y *env* codifica para proteínas superficiales (Bannert & Kurth, 2006; Mager & Medstrand, 2005). Estos genes se encuentran delimitados/flanqueados por Repeticiones Terminales Largas (LTR, por sus siglas en inglés) (Cowley & Oakey, 2013). Tras la infección, los retrovirus revierten el flujo usual de la información; ya que convierten el ARN en ADN doble banda mediante la actividad de una transcriptasa reversa y el ADN resultante seguidamente se inserta en el genoma del huésped con la ayuda de una integrasa codificada por el mismo retrovirus (Bannert & Kurth, 2006).

Los retrovirus en algunos casos pueden infectar células de la línea germinal en tejidos como la placenta o testículos y se alojan en el ADN de estas células (Bannert & Kurth, 2006). Esto ocasiona que los individuos que se originen a partir de estas células sexuales sean portadores del provirus y que éste sea heredado verticalmente a los descendientes siguiendo las leyes de herencia mendeliana (Bannert & Kurth, 2006; Mager & Medstrand, 2005). Los retrovirus que llegan a ser parte del genoma del hospedero de esta manera, se les conoce como retrovirus endógenos (ERV por sus siglas en inglés) (Bannert & Kurth, 2006).

2.3 Retrovirus endógenos

Los ERV fijados en el genoma son restos de antiguas infecciones de retrovirus exógenos que se insertaron y se acumularon en el genoma de un hospedero. La mayoría de retrovirus endógenos que permanecen insertos en el genoma no son codificantes, debido a que a través del tiempo han sufrido una serie de mutaciones que han alterado el marco abierto de lectura (ORF - open reading frame) de sus genes. No obstante, aún se conservan ERVs intactos y activos en ciertos mamíferos como koalas (Tarlinton et al., 2006), ratones (Maksakova et al., 2006), gatos (Roca

et al., 2004), ovejas (Chessa et al., 2009) y en algunos primates, aunque en estos últimos los ejemplos son escasos.

Las repeticiones retrovirales pueden interferir en la expresión de otros genes (Mager & Medstrand, 2005) debido a que los ERV pueden inactivar o desregular genes esenciales e incluso pueden ocasionar varios tipos de desórdenes en el organismo (Bannert & Kurth, 2006). Asimismo, los ERV pueden incrementar la plasticidad del genoma y la tasa evolutiva de las poblaciones de su hospedero (Bannert & Kurth, 2006).

Se ha visto que para que los ERVs puedan desempeñar un papel en la regulación génica de su hospedero, requieren sufrir ciertas sustituciones tras su inserción (Emera & Wagner, 2012). Una vez insertos en el genoma del hospedero, los ERVs pueden aportar secuencias genómicas que en un futuro pueden convertirse en nuevos genes, o pueden también regular la expresión de los genes del hospedero aportando secuencias promotoras de genes o mediante elementos reguladores. A su vez los ERVs necesitan de elementos que regulan su transcripción y transposición, los cuales son codificados por el propio retrovirus (Emera & Wagner, 2012).

Los ERVs humanos se clasifican en tres familias o grupos con base en la similitud de sus secuencias y la de ciertos tipos de retrovirus infecciosos conocidos (Mager & Medstrand, 2005). Las tres clases son: a) Clase I o Gamma retrovirus-like los cuales tienen similitud en la secuencia con los retrovirus de mamíferos del tipo C; b) Clase II, que tiene similitud con los del tipo B o Betaretrovirus; y c) Clase III que son similares a los Espumaretrovirus (Mager & Medstrand, 2005). La variabilidad de los ERVs hace también que los criterios de clasificación sean complicados, variables y heterogéneos. En la actualidad no hay un sistema de clasificación definido y hay evidencia de que probablemente los grupos de ERVs de Gamma retrovirus-like y Espumaretrovirus-like no son grupos monofiléticos (Escalera-Zamudio & Greenwood, 2016). Por otra parte, hay evidencia que otros como los Betaretrovirus-like sí lo son (Escalera-Zamudio & Greenwood, 2016).

2.4 Retrovirus endógenos en primates

Los detalles precisos sobre cómo ocurrió la separación del linaje de los humanos de otros grupos de primates todavía se desconocen. Sin embargo, los sitios de inserción de retrovirus endógenos podrían ser útiles para elucidar relaciones filogenéticas entre especies relacionadas (Barbulescu et al., 2001).

Algunos autores sugieren que los elementos transponibles son los principales responsables de la diferenciación de los humanos de las otras especies de primates, y que de todos los elementos transponibles, los retrovirus endógenos son los que más han contribuido a esta diferenciación (Khodosevich et al., 2002). Esto debido a que a nivel genético los primates son similares, sin embargo, los retrovirus endógenos en primates han aportado LTRs que regulan la expresión génica tras su integración en regiones codificantes del genoma y además son capaces de inducir rearreglos genómicos y/o pueden alterar la función de los genes (Khodosevich et al., 2002).

Los retrovirus endógenos han sido encontrados en todas las especies de vertebrados. Por lo tanto, humanos y otros mamíferos poseen ERVs como parte de su genoma. Estos ERVs se han visto involucrados en la evolución del genoma, alterando su estructura y función (Harris, 1998), ya que pueden actuar como promotores y enhancers de genes o sitios de splicing alternativo en humanos (Akopov et al., 1998). Esto hace que los retrotransposones hayan contribuido considerablemente a la diversidad genética de humanos (Stewart et al., 2011). Se cree que se integraron por medio de infecciones ancestrales de retrovirus exógenos en la línea germinal, los cuales con el tiempo se llegaron a fijar en las especies. Alternativamente, existe la posibilidad de que algunos ERVs se originaran a partir de retrotransposones (Bannert & Kurth, 2006; Mager & Medstrand, 2005).

Muchos ERVs presentes en humanos actualmente también han sido encontrados en otros primates, lo cual prueba que estos retrovirus estuvieron presentes en los ancestros de los humanos desde hace millones de años, muy temprano en la evolución (Mager & Medstrand, 2005; Venter et al., 2001). Estos retrovirus que se integraron antes de la separación de linajes están presentes en posiciones ortólogas en los genomas de diferentes primates (Barbulescu et al., 1999).

No obstante, hay otros ERVs que se integraron posteriormente a la separación del linaje de humanos con respecto al linaje de los chimpancés y gorilas (Barbulescu et al., 1999; Medstrand & Mager, 1998).

En la Figura 2 se muestra una representación de la evolución de los ERVs humanos (HERVs) en el linaje de los primates. El grupo de los HERV-S/L es considerado el más ancestral, y es uno de los que invadió el genoma de vertebrados antes del origen de los primates y ha sido detectado en reptiles y peces (Escalera-Zamudio & Greenwood, 2016). A través de la historia han habido diferentes inserciones de HERVs, justo antes de la divergencia de los homínidos y los primates del viejo mundo ocurrió un pico de inserción de retrovirus en el genoma entre aproximadamente 45 y 30 millones de años atrás (Escalera-Zamudio & Greenwood, 2016). Existen algunas especies de ERVs exclusivos de humanos, por ejemplo HERVK (HML-2), y otros como PtERV1 que están presentes solo en el genoma de chimpancé y gorila (Escalera-Zamudio & Greenwood, 2016).

Actualmente, un 8% del genoma humano está compuesto de retrovirus endógenos. Se han encontrado alrededor de 100 000 loci de ERVs distribuidos en aproximadamente 50 familias (Belshaw et al., 2004; Mayer et al., 2011). Cerca del 90% de estos ERVs han perdido sus secuencias de genes por mecanismos de recombinación y permanecen como solo-LTR. No obstante, en ciertas especies, se han encontrado ERVs capaces de replicarse y la recombinación entre ERVs defectuosos podría llevar a la producción de virus infecciosos (Holloway et al., 2019). No obstante, el único ERV que actualmente se encuentra activo y conserva la capacidad de transponerse en humanos es el Retrovirus Endógeno Humano de tipo K (HERV-K o HML2), también conocido como HTDV ("human teratocarcinoma derived virus") (Bieda et al., 2001). Se estima que el HTDV invadió la línea germinal hace aproximadamente 100 000 años (Bieda et al., 2001; Moyes et al., 2007). Su expresión es inducida por el Virus de la Inmunodeficiencia Humana (VIH) y este ERV regula la expresión de genes cercanos, implicados en el desarrollo de diversas enfermedades neurológicas, autoinmunes y tumorales (Laderoute et al., 2007; Voisset et al., 2008). Otros ERV en humanos han sido implicados en enfermedades

como la esquizofrenia y la diabetes (Crowell & Kiessling, 2007), y más frecuentemente se han asociado a diferentes tipos de cáncer (Laderoute et al., 2007).

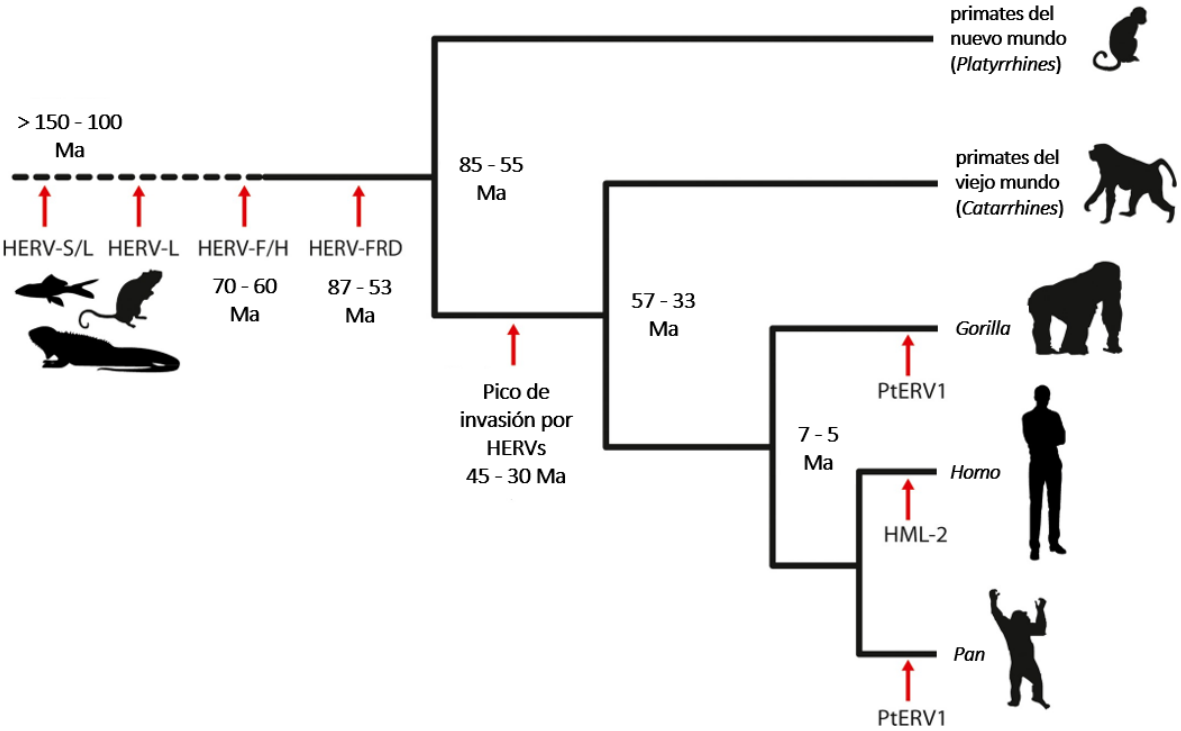


Figura 2. Representación de la evolución de los HERVs dentro del linaje de los primates. Tomado de Escalera-Zamudio & Greenwood, 2016.

2.5 Expresión de ERV en tejidos testiculares de primates

Los tejidos testiculares presentan niveles de transcripción mucho más altos en comparación con otros tejidos. Esto se debe a que los promotores presentan un alto grado de desmetilación, permitiendo la expresión de genes (Schmidt, 1996). Este alto nivel de transcripción que ocurre en tejido testicular hace que se expresen también secuencias repetitivas como transposones que normalmente no se expresan en otros tejidos (Huang et al., 2009; Ko et al., 2000; Ma, 2001). Debido a esta alta e inusual expresión genética que ocurre en tejidos testiculares, se ha propuesto que la evolución genómica se lleva a cabo durante la espermatogénesis y la embriogénesis.

Estudios previos han reportado la expresión de diferentes familias de ERVs en la línea germinal y tejido testicular de humano. Flockerzi et al., 2008, encontraron 23

HERVK transcripcionalmente activos, algunos de los cuales se había reportado previamente su potencial para formar variantes infecciosas y que se encuentran activos en células de tejido germinal. Crowell y Kiessling (2007) detectaron la expresión del gen *pol* de HERVK, HERV9 y HERVE a través de RT-PCR en muestras de testículo humano. Otros estudios han encontrado también la expresión en testículo de 16 genes de la cubierta viral provenientes de nueve familias de HERV utilizando PCR en tiempo real (de Parseval et al., 2003). Se ha visto que estos genes de cubierta de origen viral insertos en el genoma humano pueden tener diferentes funciones como: formación de la placenta mediante adhesión celular que origina el sincitiotrofoblasto (Mi et al., 2000), ejercer efectos inmunosupresores para proteger al feto del sistema inmune de la madre (Mangeney & Heidmann, 1998) y proteger contra infecciones de retrovirus exógenos a través de la interferencia con el receptor (Best et al., 1997).

Se ha visto una gran variabilidad en la expresión de ERVs incluso entre tejidos reproductivos. Estudios previos han visto diferencias entre los ERVs expresados en testículo y en epidídimo, ambos tejidos reproductivos y que comparten su dependencia a estímulos de la hormona testosterona (Crowell & Kiessling, 2007). En este estudio encontraron 6 diferentes familias de ERVs en epidídimo, mientras que solamente 3 en testículo (Crowell & Kiessling, 2007). Estas diferencias en HERVs encontrados en ambos tipos de tejidos reproductivos se cree que pueden deberse al estado de madurez de las células, ya que en epidídimo las células están completamente diferenciadas mientras que en testículo la población de células germinales son inmaduras (Crowell & Kiessling, 2007).

2.6 Estudio de la expresión génica

El estudio de todos los ácidos ribonucleicos (transcriptoma) de una muestra proveniente de una célula o tejido de un organismo se conoce como transcriptómica (Elliott & Ladomery, 2011). El estudio de transcritos específicos se ha realizado con qRT-PCR (Freeman et al., 1999), como aplicación específica posterior a la invención de la PCR. A nivel de alto rendimiento, el transcriptoma se ha estudiado a través de

aplicaciones de secuenciación de Sanger (como SAGE y MPSS) o hibridación (microarreglos), sin embargo, con el surgimiento de la secuenciación masiva actualmente se puede obtener aún más información del transcriptoma.

La técnica más utilizada en la actualidad para el estudio del transcriptoma completo es la técnica de RNA-seq, la cual se basa en el uso de secuenciación masiva paralela (Hrdlickova et al., 2017). Esta es una técnica que permite estudiar la abundancia y diversidad de transcriptos en una célula o tejido en un momento dado (Griffith et al., 2010; Marioni et al., 2008; Wang et al., 2009). Tiene la ventaja de que proporciona medidas de los niveles expresión de una manera precisa (Griffith et al., 2010; Marioni et al., 2008; Wang et al., 2009).

La técnica consiste en extraer ARNm o total a partir de una muestra biológica de interés, realizar una retrotranscripción para obtener ADN copia (ADNc), fragmentarlo, combinar los fragmentos resultantes con adaptadores terminales y secuenciarlos (Zhe Wang et al., 2009). Esta técnica se puede llevar a cabo utilizando diferentes tecnologías de secuenciación y dependiendo de la tecnología empleada se obtienen lecturas de diferentes tamaños. Por ejemplo, con la tecnología de secuenciación HiSeq (Illumina), se pueden obtener secuencias de una longitud de 30-150pb para luego ser reconstruidos por algoritmos bioinformáticos (Zhe Wang et al., 2009), o con la tecnología PacBio se pueden secuenciar ARNm completos sin necesidad de fragmentar.

En el caso de la plataforma de secuenciación Illumina, se basa en el principio de amplificación en puente y marcaje de nucleótidos utilizando fluorescencia llamado Secuenciación por síntesis (SBS) (Metzker, 2010). Los millones de secuencias generadas son analizadas por métodos bioinformáticos. Para ensamblar el transcriptoma se puede realizar un ensamblaje guiado realizando un alineamiento contra un transcriptoma o genoma de referencia, o se puede realizar un ensamblaje de *novo*. Al final se llevan a cabo análisis estadísticos para determinar la abundancia de los transcritos como ADNc y de esta forma inferir la cantidad original de ARN presente en la muestra (Griffith et al., 2010).

El RNA-Seq tiene diversas aplicaciones en el estudio de transcriptomas, entre ellas: descubrimiento de transcriptos que no habían sido descritos anteriormente, estudio de los mecanismos de regulación génica, análisis de expresión diferencial de genes, análisis de splicing alternativo, análisis de expresión alelo-específica, detección de editaje del ARN, detección viral, entre otras aplicaciones (Griffith et al., 2010; Trapnell et al., 2010). Esta técnica también cuenta con múltiples ventajas con respecto a las técnicas utilizadas tradicionalmente para el estudio de la expresión génica. Actualmente es la técnica más precisa para medir los niveles de expresión de un organismo, tejido o célula (Wang et al., 2009). Cuenta con una resolución tan alta que permite determinar variaciones de hasta un nucleótido en la secuencia de un transcrito (Trapnell et al., 2010; Wang et al., 2009). A diferencia de las técnicas tradicionales, requiere una cantidad muy baja de muestra, tiene un amplio rango dinámico y no requiere conocimiento previo del genoma o transcriptoma por analizar (Wang et al., 2009). Además, permite distinguir isoformas que son indistinguibles mediante otros métodos, a excepción de métodos de secuenciación de tercera generación como PacBio (Trapnell et al., 2010; Wang et al., 2009). No tiene el riesgo de hibridación cruzada, como ocurre con los microarreglos. También posee un alto nivel de reproducibilidad, lo que hace que sea el primer método basado en secuenciación que permite estudiar el transcriptoma entero de manera cuantitativa y muy eficiente, lo cual no era posible hasta hace algunos años con el uso de los métodos basados en secuenciación de Sanger (Wang et al., 2009)

En este proyecto se utilizarán datos de RNA-Seq para detectar expresión de ERVs en tejido testicular de tres especies de primates y correlacionar las características genómicas de los sitios cercanos a los ERVs con la expresión de estos.

3. JUSTIFICACIÓN

Esta investigación es de importancia biológica ya que aporta información acerca de la genética evolutiva y función de los ERVs en tres especies de primates, utilizando bases de datos públicas y herramientas bioinformáticas de libre acceso.

Actualmente se desconocen los detalles precisos sobre cómo ocurrió la separación de las especies de primates. No obstante, el estudio de la expresión de ERVs en tejido testicular podría ser la clave para elucidar aspectos evolutivos y comprender como ocurrió esta separación entre especies de primates relacionadas. Esto debido a que los procesos evolutivos que son heredados a los descendientes ocurren en la línea germinal, por lo tanto la expresión de ERVs en este tejido es una fuente potencial de innovación evolutiva debido a su capacidad de integración en nuevos sitios genómicos e influir en la expresión génica.

Esta investigación también genera información importante acerca de cómo influye el contexto genómico en la expresión de ERVs e información acerca de la expresión de ERVs que previamente han sido relacionados con ciertas patologías humanas. Aunque este no es el fin de esta investigación, genera información valiosa para estudios posteriores que busquen comprender la expresión de ERVs asociados a patologías humanas. Lo cual es valioso ya que hasta la fecha se conoce que la expresión de ERVs está asociada con ciertas patologías humanas, pero esta relación aún no ha sido bien caracterizada (Gröger & Cynis, 2018).

A nivel bioinformático esta investigación permitió comparar dos protocolos de ensamblaje de transcriptomas, mediante lo cual se pudo determinar el protocolo de ensamblaje más adecuado para este estudio. Realizar este tipo de comparaciones en análisis bioinformáticos es de gran importancia, ya que en este campo los resultados obtenidos de un análisis suelen depender en gran medida del protocolo utilizado y se ha visto que protocolos diseñados para desarrollar un mismo análisis pueden generar resultados muy variables.

4. HIPÓTESIS

El contexto genómico influye en la expresión de ERVs a nivel de tejido testicular en humanos, y la presencia de ERVs homólogos en las tres especies de primates refleja su papel evolutivo.

5. OBJETIVOS

5.1 Objetivo general

Determinar la influencia del contexto genómico en la expresión de ERVs a nivel de tejido testicular humano, y encontrar ERVs homólogos en tres especies de primates lo que indicaría su papel evolutivo.

5.2 Objetivos específicos

5.2.1 Identificar y clasificar los transcritos ensamblados derivados de ERVs en tres especies de primates para evaluar su papel evolutivo en la línea germinal.

5.2.2 Determinar los *loci* de los cuales se expresan los ERV y obtener las características genómicas cercanas a los sitios de integración para analizar su asociación con la expresión de ERVs.

5.2.3 Identificar y comparar las secuencias de los eERVs homólogos entre las especies estudiadas para dilucidar su papel en la evolución de los primates.

6. METODOLOGÍA

6.1 Origen de las librerías

Las librerías utilizadas corresponden a tres especies de primates: humano (*Homo sapiens*), orangután (*Pongo pygmaeus*) y gorila (*Gorilla gorilla*) (Cuadro 1). En total son 12 librerías de RNA-seq que originalmente fueron secuenciadas en ambas direcciones mediante "next generation sequencing" (paired-end sequencing) con el objetivo de determinar la especificidad de tejido de todos los genes que codifican proteínas (Cuadro 1). Las librerías de humano y gorila fueron descargadas de bases de datos públicas y la librería de orangután son datos que no han sido publicados aún y fueron proporcionados por la Dra. Kateryna Makova de la Universidad de PennState. Las librerías de humano seleccionadas eran de las pocas que se encontraban accesibles en el momento en el que dio inicio esta investigación, no

obstante, actualmente hay muchas más librerías de RNA-Seq de tejido testicular humano disponibles en bases de datos públicas.

Los datos crudos de RNA-seq de tejido testicular de orangután (sin publicar) y gorila (Vegesna et al. accepted) fueron obtenidos de animales en cautiverio del zoológico de San Diego (Estados Unidos) que fueron sacrificados por razones de salud. El ARN de testículo de estas muestras fue extraído utilizando el RNeasy Mini kit (Qiagen) y las librerías de secuenciación fueron generadas con el TruSeq RNA sample prep kit (Illumina). La secuenciación se realizó en las plataformas HiSeq y MiSeq (Illumina). Los datos de humano de Ruiz-Orera et al., 2015 provienen de tejido testicular histológicamente sano proveniente de pacientes quirúrgicos de 25 y 71 semanas de edad que fueron secuenciados como paired-end en el Illumina HiSeq 2000. Mientras que los datos de Fagerberg et al., 2014, provienen de tejido histológico testicular normal de pacientes quirúrgicos con edades de 26, 34, 37, 50, 56 y 62 años. Los ARNs fueron extraídos utilizando el RNeasy Mini kit (Qiagen) y la secuenciación se realizó utilizando los equipos HiSeq2000 y 2500 (Illumina) y el protocolo estándar de Illumina para RNA-Seq.

Todas las librerías fueron subidas al clúster del Colaboratorio Nacional de Computación Avanzada (CNCA) del Centro Nacional de Alta Tecnología (CENAT) en el cual se llevaron a cabo la mayor parte de los análisis haciendo uso de las herramientas bioinformáticas instaladas en dicha infraestructura computacional.

Cuadro 1. Información de librerías RNA-seq que se utilizaron en este proyecto

Especie	Nombre de librería	Origen	Longitud de reads	Phred score promedio
Gorila	SRR3053573	Vegesna et al. accepted	150	36
Orangután	3405	sin publicar	150	37
Humano	SRR2040581	Ruiz-Orera et al., 2015	100	38
Humano	SRR2040582	Ruiz-Orera et al., 2015	100	38
Humano	ERR315350	Fagerberg et al., 2014	100	38
Humano	ERR315351	Fagerberg et al., 2014	100	38
Humano	ERR315352	Fagerberg et al., 2014	100	38
Humano	ERR315391	Fagerberg et al., 2014	100	37
Humano	ERR315415	Fagerberg et al., 2014	100	37
Humano	ERR315446	Fagerberg et al., 2014	100	37
Humano	ERR315456	Fagerberg et al., 2014	100	37
Humano	ERR315492	Fagerberg et al., 2014	100	37

En la Figura 3 se muestra el protocolo general de análisis de datos implementado con cada una de las librerías analizadas en este estudio con el propósito de identificar y anotar transcritos originados a partir de eERVs en tejido testicular humano.

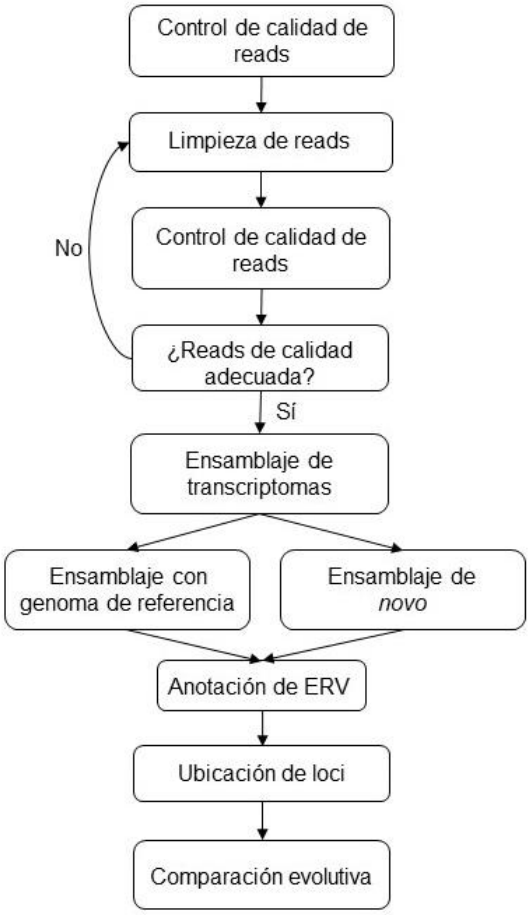


Figura 3. Protocolo de análisis de datos de RNA-Seq que se implementó para identificar eERVs y realizar una comparación evolutiva de estos elementos en las tres especies de primates.

6.2 Control de calidad de los reads

Se realizó una evaluación de la calidad de las lecturas (reads en inglés) de cada librería utilizando la herramienta FastQC (Andrews, 2010). Con este análisis se evaluó: la presencia de adaptadores, secuencias sobrerrepresentadas, contenido de GC, frecuencias de K-meros, distribución por nucleótido, nivel de duplicación, entre

otros factores. Esta evaluación de calidad se realizó en las librerías crudas y en las librerías después del preprocesamiento.

6.3 Limpieza de los reads

Como parte del preprocesamiento de los datos, se eliminaron reads con una longitud menor a 50 pb, se eliminaron adaptadores, contaminación y reads de baja calidad en todas aquellas librerías que lo requirieron. Únicamente se conservaron aquellos reads que presentaron un nivel de calidad por encima de 30. Para esto se utilizó la herramienta Trimmomatic 0.36 (Bolger et al., 2014) con los parámetros SLIDINGWINDOW:4:20, MAXINFO:50:0.8 y MINLEN:50. Al finalizar la limpieza de los reads, se realizó nuevamente un análisis con la herramienta FastQC para corroborar que la calidad fuera la adecuada (nivel de calidad > 30 y el contenido de Ns, secuencias sobrerrepresentadas y de adaptadores = 0).

6.4 Estandarización de protocolos bioinformáticos para ensamblaje de transcriptomas

Como parte del proceso de estandarización del protocolo para ensamblar los transcriptomas se probaron diferentes parámetros de ensamblaje utilizando la herramienta Trinity 2.6.2 (Grabherr et al., 2011), con el objetivo de determinar cuál generaba el mejor ensamblaje, es decir, con el que se obtuviera la mayor cantidad de transcritos y de mayor longitud.

Cada uno de los transcriptomas ensamblados en estas pruebas fue evaluado utilizando la herramienta TransRate (Smith-Unna et al., 2016) para determinar el total de contigs ensamblados, su longitud y N50, con el objetivo de determinar cuáles fueron los parámetros que generaron el mejor ensamblaje y emplear dichos parámetros en los ensamblajes de *novo* y guiados por genoma de cada una de las librerías.

6.5 Ensamblaje de transcriptomas

6.5.1 Ensamblaje guiado por genoma

Se descargaron los genomas de referencia de humano, gorila y orangután, incluyendo tanto cromosomas como scaffolds, esto debido a que en los scaffolds se pueden encontrar muchas secuencias repetitivas que aún no han podido ser asignadas a un determinado cromosoma. Estos genomas de referencia son de libre acceso en el UCSC Genome Browser (humano-hg38, gorila-gorGor3 y orangután-ponAbe2) (Cuadro 2). Cada uno de estos genomas fue respectivamente indexado utilizando la herramienta HISAT2 (D. Kim et al., 2015). Se utilizó esta herramienta para realizar un mapeo de los reads de cada una de las librerías (previamente preprocesadas) contra su respectivo genoma de referencia (humano, gorila u orangután) utilizando los parámetros default de la herramienta.

El archivo en formato SAM obtenido se convirtió al formato BAM utilizando la herramienta Samtools (Li et al., 2009). Estos archivos extensión BAM fueron utilizados como input para realizar el ensamblaje guiado por genoma utilizando la herramienta Trinity.

Finalmente, utilizando la herramienta TransRate, se evaluó la calidad de cada transcriptoma obtenido con el objetivo de comparar estas estadísticas con las de los ensamblajes de *novo* del apartado 6.5.2.

Cuadro 2. Información de los genomas de referencia de gorila (gorGor3), humano (hg38) y orangután (ponAbe2) utilizados.

	Gorila (gorGor3)	**Humano (hg38)**	**Orangután (ponAbe2)**
Versión	05/10/2011	17/12/2013	13/11/2008
Número de accesión GenBank	GCA_000151905.1	GCA_000001405.15	GCA_000001545.3
Longitud total	3,029,537,234	3,099,734,149	3,437,863,358
Nivel de ensamblaje	Cromosoma	Cromosoma	Cromosoma
N50	913,958	67,794,873	747,367

6.5.2 Ensamblaje de *novo*

Se realizó un ensamblaje de *novo* de los transcriptomas de humano, gorila y orangután a partir de las lecturas pareadas tanto en dirección "forward" como en "reverse". Se utilizó la herramienta Trinity con los parámetros --KMER_SIZE 25 y --min_contig_length 200 paralelamente en 64 cores, dado a que con esos parámetros fue que se obtuvieron los mejores resultados. Seguidamente se realizó un análisis de la calidad de cada uno de los transcriptomas ensamblados, utilizando la herramienta TransRate. Se realizó una prueba U de Mann Whitney para comparar los resultados de los ensamblajes realizados de *novo* y guiados por genoma (obtenidos en el apartado 6.5.1), con el objetivo de determinar si existían diferencias estadísticas entre los resultados obtenidos con ambos métodos de ensamblaje.

6.6 Anotación de ERVs

Se descargó la base de datos RepBase (v. RepBase24.12) en formato fasta desde la página del Genetic Information Research Institute, California, USA (https://www.girinst.org/server/RepBase/index.php, descargada el 22/2/2018). Esta es una base de datos que está compuesta por secuencias de ADN correspondientes a elementos repetitivos del genoma de diferentes especies de eucariontes. Aunque existen otras bases de datos, esta es una de las más utilizadas, ya que es muy completa y constantemente está siendo actualizada (Bao et al., 2015; Jurka et al., 2005). Debido a esto, esta base de datos es las más adecuada para la anotación de secuencias de ADN provenientes de elementos repetitivos. Esta base de datos está compuesta por prototipos de secuencias consenso de familias de elementos repetitivos de diferentes especies de eucariotas.

Debido a la naturaleza de los objetivos de esta investigación, se creó una base de datos personalizada a partir de RepBase, en la cual únicamente se incluyeron los elementos transponibles de humanos y de primates. Esto con el fin de tener en la base de datos únicamente secuencias de ADN de los elementos repetitivos de interés y eliminar el posible ruido en el análisis que pudiera generar el emplear la base de datos completa. La nueva base de datos customizada contenía un total de 782

elementos (longitudes entre 199 y 9084 pb) categorizados como: retrovirus endógenos (47), ERV1 (424), ERV2 (113) y ERV3 (198). Cada uno de los transcriptomas ensamblados a través del abordaje de *novo* y guiado por genoma, fueron alineados contra esta base de datos utilizando la herramienta BLASTn 2.7.1 (Altschul et al., 1990) con el objetivo de identificar los transcritos derivados de ERVs.

Para realizar el alineamiento con BLASTn se consideraron transcritos derivados de ERVs (eERV) solamente a aquellos que presentaron un porcentaje de identidad superior al 70% (-perc_identity 70) con alguno de los elementos de la base de datos y una longitud mayor a 3000 pb, esto debido a que se buscaba analizar únicamente los eERVs completos o casi completos. Se eligió este porcentaje de identidad considerando que los elementos transponibles tienen porcentajes de divergencia genética variables, que dependen de la longitud y edad del elemento (Giordano et al., 2007), además de que sufren mutaciones y evolucionan rápidamente por lo cual van acumulando diferencias con respecto a los elementos consenso de la base de datos Repbase. De los resultados obtenidos de este alineamiento se eligió el hit que contaba con el mejor score y el menor E-value.

Se evaluó la capacidad codificante de cada uno de los transcritos ensamblados mediante un alineamiento con BLASTx contra la base de datos nr (non-redundant protein sequences). Se eligió esta estrategia debido a que pueden existir ERVs que han perdido algunos o la mayoría de genes internos por lo cual otras estrategias como la búsqueda de codones de iniciación no serían funcionales en casos de este tipo. Otras estrategias que se basan en la búsqueda de promotores en la región 3' UTR, codones de iniciación y terminación, no pueden ser empleadas en este caso en donde únicamente se cuenta con transcritos y no con el genoma completo. Se determinó que los eERVs con capacidad codificante eran aquellos que presentaron un porcentaje de similitud superior al 75% con alguna proteína de la base de datos y que esta similitud fue en un fragmento superior a 1500 pb. Con los resultados obtenidos de este alineamiento también se eligió el hit que contaba con el score más alto y el menor E-value.

A pesar de que existen diferentes herramientas para realizar alineamientos de secuencias, se utilizó la herramienta BLAST dado a que ésta es una de las más utilizadas debido a sus múltiples ventajas (McGinnis & Madden, 2004). Esta herramienta creada hace 30 años y que ha sido mejorada a través de los años, es muy versátil, ya que entre sus múltiples aplicaciones permite realizar alineamientos de nucleótidos y proteínas, además es mucho más rápida que otras herramientas, permite realizar alineamientos locales con bastante exactitud y tiene la ventaja de que es de libre acceso (Camacho et al., 2009; McGinnis & Madden, 2004).

Finalmente, se utilizó el paquete ggplot (0.4.2) de R para realizar un gráfico de burbujas con la representación gráfica de la cantidad de eERVs de cada tipo (Wickham, 2016).

6.7 Determinación del nivel de soporte de cada transcrito ensamblado

Para determinar el nivel de soporte o cobertura de cada uno de los transcritos ensamblados, se extrajeron las secuencias en formato fasta de cada uno. Estos transcritos fueron indexados utilizando la herramienta HISAT2 y posteriormente se alinearon los reads forward y reverse (los obtenidos después del trimming) contra estos transcritos ensamblados. Para determinar el nivel de cobertura en cada una de las bases del transcrito se utilizó la herramienta Samtools depth y tras obtener este dato se estimó el promedio de cobertura en todo el transcrito. Se consideraron con suficiente soporte aquellos transcritos con una cobertura mínima de 5X y los que no alcanzaron este nivel de soporte umbral fueron descartados.

Se utilizó la herramienta Kallisto para calcular el número de TPM ("Transcripts per million"). Para esto se realizó un index de todos los transcritos ensamblados y este se utilizó para estimar los TMP utilizando las librerías preprocesadas (obtenidas después del trimming). La estimación de este parámetro permite normalizar los resultados de cobertura dado que las librerías utilizadas fueron secuenciadas a diferente profundidad. El valor de TPM indica cuántas moléculas corresponden a ERVs por cada millón de moléculas de RNA en los datos de RNA-seq de la muestra. Para esto se divide el número total de reads que mapearon con cada librería

ensamblada entre la longitud del transcrito para obtener un nivel de expresión normalizado de cada transcrito reconstruido, luego la suma de los valores de expresión normalizados es dividida entre 1 000 000 para obtener un factor de escala. Finalmente se divide el nivel de expresión normalizado de cada transcrito entre el factor de escala calculado, lo cual da como resultado el valor de TPM. Existen otras métricas para realizar esta normalización de los transcritos de RNA-Seq de diferentes muestras; sin embargo se eligió estimar TPMs ya que este parámetro es comúnmente utilizado porque tiene la ventaja de que normaliza tomando en cuenta las diferencias en composición de transcritos debido a la profundidad de secuenciación, en lugar de solo dividir entre el número de reads de la librería, esto hace que este parámetro sea más comparable entre muestras de diferentes orígenes y composición (Conesa et al., 2016).

6.8 Ubicación en el genoma y características genómicas de los flancos de eERVs humanos

6.8.1 Ubicación de la posición de cada transcrito

Los eERV con suficiente soporte se alinearon contra los genomas de referencia respectivos para identificar los *loci* en los que se encuentran insertos. Para esto se utilizó la herramienta BLAT, en donde se seleccionó una identidad mínima del 80% (-minIdentity=80) y se verificó que el alineamiento incluyera al menos 80% de la longitud del transcrito (>2400bp). Se eligieron estos parámetros tomando en cuenta que los ERVs pueden sufrir mutaciones y splicing alternativo que pueden hacer que existan diferencias entre los eERVs ensamblados a partir de las librerías utilizadas y las secuencias homólogas presentes en el genoma de referencia hg38.

Con esto se obtuvieron las coordenadas genómicas de los eERV, las cuales fueron almacenadas en formato BED. En los casos en que un mismo transcrito mapeó en diferentes partes del genoma, se asumió como correcta la posición en la cual el alineamiento fue de mayor longitud y con mayor porcentaje de similitud. Se eliminaron isoformas (transcritos que varían en longitud o en pocos nucleótidos pero

que corresponden a un mismo elemento dado a que mapearon en una misma región del genoma). Después de obtener el *locus* de cada eERVs en el genoma, se estimó la densidad de eERVs encontrada en cada cromosoma, y se comparó con la densidad esperada según lo reportado en la literatura. Se realizó una prueba t de student de dos colas para determinar si existían diferencias entre los valores de densidad de ERvs observados y esperados.

Se utilizó el paquete ggplot (0.4.2) de R para realizar un gráfico de burbujas con la representación gráfica de la cantidad de eERVs en cada cromosoma (Wickham, 2016).

6.8.2 Identificación de transcritos anidados

Con el fin de determinar si los transcritos identificados contenían otros elementos transponibles anidados, se determinó el intercepto de las coordenadas de los ERVs con las coordenadas de todos los elementos transponibles de la base de datos RepBase. Para esto se utilizó la herramienta "Intersect the intervals of two datasets" de la plataforma web de Galaxy (Version 1.0.0) utilizando el servidor público en usegalaxy.org para analizar los datos (Afgan et al., 2016). Finalmente se visualizaron los interceptos con la herramienta Integrative Genomics Viewer (IGV).

6.8.3 Determinación de traslape de ERVs expresados con genes, lncRNAs y LTRs

Se determinó el intercepto de las coordenadas de los eERVs de cada librería con genes, lncRNAs (reportados en tejido testicular de humano) y LTRs presentes en el genoma de referencia obtenidos de bases de datos públicas (Cuadro 2). Esto se realizó con el objetivo de determinar la proporción de traslape entre los eERV y elementos de cada uno de los sets de datos mencionados. Esto se realizó utilizando la herramienta "Intersect the intervals of two datasets" de la plataforma web Galaxy (Version 1.0.0) y se visualizaron los interceptos con la herramienta Integrative Genomics Viewer (IGV).

Se utilizó la herramienta "*Join the intervals of two datasets side-by-side*" de la plataforma web de Galaxy (Version 1.0.0), para unir el archivo que contenía las

coordenadas de los interceptos de los elementos genómicos con el archivo que contenía las ubicaciones genómicas de los eERV. Finalmente se realizó un conteo del total de interceptos de los ERVs con LTRs, genes y lncRNAs.

6.8.4 Determinación de las características genómicas de los flancos de eERV en el genoma humano

Los flancos de cada eERV se definieron como 100 kb corriente arriba y 100 kb corriente abajo a partir del inicio y final de la coordenada genómica de cada eERV, respectivamente.

Cada flanco de 100 kb fue dividido en ventanas no traslapantes de 1 kb en las que se cuantificó la proporción de: genes, islas CpG, otros ET (como SINEs, LINEs, LTRs y ADN transposones), lncRNAs, regiones con cromatina abierta (sensibles a ADNasa I), regiones conservadas, orígenes de replicación y tasa de recombinación, los cuales podrían estar modulando la expresión de los ERVs en testículo. Para cuantificar la proporción de traslape de cada característica genómica con cada una de las ventanas se utilizó la herramienta de la plataforma web de Galaxy (Version 1.0.0) llamada *Feature coverage*. Datos de estas características genómicas se descargaron de diversas bases de datos descritas en el Cuadro 3.

Cuadro 3. Descripción de las bases de datos utilizadas para cuantificar características genómicas en los flancos de eERVs

Base de datos	Cuantificación	Fuente	Referencia
SINE	Contenido	RepBase	Jurka et al., 2005
LINE	Contenido	RepBase	Jurka et al., 2005
HS de recombinación	Contenido	UCSC Genome Browser	Myers et al., 2005
Tasa de recombinación	Promedio	UCSC Genome Browser	Myers et al., 2005
Tiempo de replicación	Promedio	DECODEdcc	Ryba et al., 2011
Orígenes de replicación	Contenido	UCSC Genome Browser	Besnard et al., 2012
Sitios sensibles a DNasa I	Contenido	UCSC Genome Browser	Sabo et al., 2006
Islas CpG	Contenido	UCSC Genome Browser	Gardiner-Garden & Frommer, 1987
Regiones conservadas	Contenido	UCSC Genome Browser	-
lncRNAs	Contenido	Human lncRNA catalog	Cabili et al., 2011
ADN transposones	Contenido	RepBase	Jurka et al., 2005
Genes	Contenido	UCSC RefSeq (gene)	Pruitt et al., 2014

Para obtener el set de datos control que incluía a todos los LTRs que no se expresan, se filtró la base de datos RepBase y se extrajeron las coordenadas genómicas de todos los LTRs mayores a 3000pb, que correspondieron a un total de 3060 elementos. Seguidamente se eliminaron todas aquellas coordenadas genómicas que traslaparon con las coordenadas de los ERVs expresados. Además se eliminaron las coordenadas genómicas correspondientes a centrómeros (Miga et al., 2014) y télomeros utilizando la base de datos de gaps de hg38 que se obtuvo descargando el track "Gap" disponible en UCSC Genome Browser (https://genome.ucsc.edu/cgi-bin/hgTables, 28 Marzo 2020 03:55:52 (UTC)). Para eliminar estas coordenadas genómicas se utilizó la herramienta "*bedtools Intersect intervals*" de la plataforma web de Galaxy (Version 1.0.0) y finalmente se obtuvo un total de 3000 posiciones genómicas, correspondientes al set de datos control. Para cada una de las 3000 coordenadas genómicas se estimaron los flancos de la misma manera que se estimaron para los ERVs expresados y se evaluaron las características genómicas como se detalla arriba.

Con el objetivo de determinar si existían diferencias entre las características genómicas de las regiones control y de los flancos de los ERVs detectados, se realizó un análisis funcional de los datos o FDA (por sus siglas en inglés de "Functional data analysis") utilizando el paquete IWTomics (V.1.12.0) en R (V.3.6.0). Esta herramienta permite analizar y comparar datos ómicos medidos con alta resolución en regiones genómicas definidas. Este paquete permite probar la hipótesis nula de que las distribuciones de dos curvas son iguales a través de una prueba no paramétrica utilizando permutaciones, siendo ideal para utilizarse en este tipo de datos genómicos que generalmente violan el supuesto de normalidad (Cremona et al., 2018).

Esta herramienta genera una curva con los valores de *p* ajustados en cada ventana. A diferencia de otras herramientas que realizan análisis similares, ésta realiza una corrección de los valores de *p* contemplando todos los intervalos que contienen esa ventana (Cremona et al., 2018).

6.9 Comparación evolutiva de eERVs entre especies de primates

Para ubicar las regiones homólogas de los ERVs en humano con respecto a los genomas de gorila y orangután se extrajeron las secuencias de ADN (en formato fasta) del genoma humano correspondientes a cada tipo de ERV identificado. Para esto se utilizaron las coordenadas consenso (obtenidas a partir de la posición de inicio más baja y la posición de final más alta, esto para los casos en los que un mismo ERV fue ensamblado a partir de diferentes librerías y que existieron diferencias en los tamaños de estos) de cada eERV (41 en total) y se utilizó la función "GetFastaBed" de la herramienta bedtools para extraer las secuencias de ADN en formato fasta. Seguidamente se utilizó la herramienta BLAT para alinear cada una de estas secuencias en el genoma de gorila y de orangután considerando solo los alineamientos con porcentaje de similitud superior al 80% (-minIdentity=80). Con cada una de las posiciones genómicas obtenidas del alineamiento con BLAT, se extrajeron las secuencias nucleotídicas de los genomas de gorila (GorGor3) y orangután (PonAbe2) utilizando la función "GetFastaBed" de la herramienta bedtools.

Cada una de las secuencias de ADN extraídas de los genomas de gorila y orangután fue alineada mediante un megablast contra la base de datos RepBase (customizada previamente) utilizando la herramienta BLASTn 2.7.1 con un porcentaje de identidad superior al 70% (-perc_identity 70) con el objetivo de identificar los transcritos derivados de ERVs. Este porcentaje de identidad fue seleccionado tomando en cuenta los aspectos mencionados en el apartado 6.6. Para la identificación de cada copia de ERV se consideró únicamente el hit del BLAST con score más alto y el menor E-value.

Se realizó un recuento de los ERVs compartidos entre las tres especies de primates y se seleccionaron aquellos ERVs identificados en las tres especies y que además se había determinado que poseían capacidad codificante en humano. Utilizando estos ERVs se estimó la Ka (tasa relativa de mutaciones no sinónimas) y la Ks (tasa relativa de mutaciones sinónimas). Se realizó una prueba de selección purificadora entre las secuencias de las tres especies de primates utilizando el método de Nei-Gojobori en el software MEGA 7 (Kumar et al., 2016).

7. RESULTADOS

7.1. Control de calidad de las librerías y preprocesamiento de datos

Los datos crudos de todas las librerías mostraron valores de Phred score superiores a 36. Sin embargo, fue necesario realizar una limpieza y preprocesamiento de los datos debido a la presencia de restos de adaptadores y secuencias sobrerrepresentadas en algunos casos. Tras realizar el trimming de las librerías se logró eliminar reads de baja calidad, secuencias sobrerrepresentadas y adaptadores. Con esto se logró obtener librerías con reads de una calidad óptima, aunque se disminuyó ligeramente la cantidad de reads tras realizar el preprocesamiento (Cuadro 4).

Cuadro 4. Resultados análisis de calidad de las librerías crudas antes y después del trimming

Librería	Organismo	Antes del trimming			Después del trimming	
		No. reads	Promedio de calidad	Otros*	No. Remanente de reads (%)	Promedio de calidad
ERR315350_1.fastq	Humano	8881861	38	A	7351435 (82,8%)	38
ERR315350_2.fastq	Humano	8881861	37	A	7351435 (82,8%)	38
ERR315351_1.fastq	Humano	8808454	38	A	7395681 (83,9%)	38
ERR315351_2.fastq	Humano	8808454	38	A	7395681 (83,9%)	38
ERR315352_1.fastq	Humano	35915919	37	A	31656635 (88,1%)	38
ERR315352_2.fastq	Humano	35915919	37	A	31656635 (88,1%)	38
ERR315391_1.fastq	Humano	39007238	37	A	34081813 (87,4%)	37
ERR315391_2.fastq	Humano	39007238	37	A	34081813 (87,4%)	37
ERR315415_1.fastq	Humano	44085084	37	A	38696902 (87,8%)	37
ERR315415_2.fastq	Humano	44085084	37	A	38696902 (87,8%)	37
ERR315446_1.fastq	Humano	33605790	37	A	29395904 (87,5%)	37
ERR315446_2.fastq	Humano	33605790	37	A	29395904 (87,5%)	37
ERR315456_1.fastq	Humano	32054218	37	A	28115822 (87,7%)	37
ERR315456_2.fastq	Humano	32054218	37	A	28115822 (87,7%)	37
ERR315492_1.fastq	Humano	44436072	37	A, D, O	38546727 (86,7%)	37
ERR315492_2.fastq	Humano	44436072	37	A, D, O	38546727 (86,7%)	37
SRR2040581_1.fastq	Humano	14782097	38		14768715 (99,9%)	38
SRR2040581_2.fastq	Humano	14782097	38		14768715 (99,9%)	38
SRR2040582_1.fastq	Humano	85271817	38		85137618 (99,8%)	38
SRR2040582_2.fastq	Humano	85271817	38		85137618 (99,8%)	38
SRR3053573_1.fastq	Gorila	12193614	36	A	10984239 (90,1%)	36
SRR3053573_2.fastq	Gorila	12193614	37	A	10984239 (90,1%)	37
3405F.fastq	Orangután	99090483	37	A, D, O	89271670 (90.1%)	37
3405R.fastq	Orangután	99090483	37	A, D, O	89271670 (90.1%)	37

*A: Presencia de Adaptadores, D: Secuencias duplicadas, O: secuencias sobrerrepresentadas.

7.2 Estandarización de protocolos bioinformáticos para ensamblaje de transcriptomas

7.2.1 Comparación de ensamblaje utilizando diferentes parámetros con la herramienta Trinity

Tras realizar tres pruebas de ensamblajes de *novo* de la librería ERR315391 (humano) utilizando diferentes parámetros con el ensamblador Trinity, se encontró que el ensamblaje que generó los mejores resultados fue el que se realizó utilizando --min_contig_length de 200 y --KMER_SIZE 25 (Cuadro 5). Por lo tanto, se decidió utilizar este parámetro para realizar el ensamblaje de todas las librerías, con el objetivo de que los parámetros utilizados fueran consistentes entre todas las librerías y que las diferencias que se pudieran observar entre estas no fueran producto del empleo de diferentes parámetros de ensamblaje.

Cuadro 5. Comparación de calidad de ensamblajes de *novo* de la librería ERR315391 utilizando diferentes parámetros

Parámetros	N secuencias	Sec más larga	N bases	Longitud promedio	N50
--KMER_SIZE 20 --min_contig_length 150	331721	10755	196663862	548	1447
--KMER_SIZE 50 --min_contig_length 80	0	0	0	0	0
--KMER_SIZE 25 --min_contig_length 200	409681	20264	433390140	1057	2197

7.2.2 Comparación entre ensamblajes guiados por genoma y de *novo*

Al realizar mapeo de reads de cada una de las librerías contra el respectivo genoma de referencia utilizando la herramienta HISAT2, se comprobó que en todas las librerías, el 100% de las lecturas fueron pareadas y cerca del 70% de las lecturas de la mayoría de las librerías de humano mapearon al menos una vez contra el genoma de referencia. En el caso de orangután y gorila aproximadamente un 50% de las secuencias mapearon al menos una vez. En general las tasas de alineamiento fueron bastante altas para las librerías de humano (cerca del 90%), pero para las librerías de orangután y gorila fueron más bajas (cercanas al 75%) (Cuadro 6). El

archivo en formato SAM obtenido mediante este alineamiento fue utilizado como input para realizar el ensamblaje guiado por genoma (Cuadro 6).

Cuadro 6. Resultados del alineamiento de las lecturas de cada librería contra el respectivo genoma de referencia (hg38, GorGor3 o PonAbe2) utilizando la herramienta HISAT2

Librería	Especie	No. lecturas	Alinearon una vez (%)	Alinearon más de una vez (%)	No alinearon (%)	Alineamiento (%)
ERR315350	Humano	7351435	5157748 (70,2%)	716014 (9,7%)	1477673 (20,1%)	86,0%
ERR315351	Humano	7395681	5116953 (69,2%)	773708 (10,5%)	1505020 (20,3%)	85,8%
ERR315352	Humano	31656635	24752322 (78.2%)	4311633 (13.6%)	2592678 (8.2%)	95.6%
ERR315391	Humano	34081813	23148367 (67.9%)	3783081 (11.1%)	7150364 (21.0%)	84.3%
ERR315415	Humano	38696902	29194427 (75.4%)	4745725 (12.3%)	4756750 (12.3%)	91.8%
ERR315446	Humano	29395904	22587656 (76.8%)	3597535 (12.3%)	3210713 (10.9%)	93.0%
ERR315456	Humano	28115822	20982450 (74.6%)	3485616 (12.4%)	3647756 (13.0%)	91.3%
ERR315492	Humano	38546727	28864063 (74.9%)	5093324 (13.2%)	4589340 (11.9%)	92.1%
SRR2040581	Humano	14768715	12299596 (83.3%)	1945124 (13.2%)	523995 (3.5%)	97.1%
SRR2040582	Humano	85137618	69979278 (82.2%)	11063609 (13.0%)	4094731 (4.8%)	98.2%
SRR3053573	Gorila	10984239	6206481 (56.5%)	1328567 (12.1%)	3449191 (31.4%)	76.9%
3405	Orangután	89271670	48929894 (54.8%)	9403411 (10.5%)	30938365 (34.7%)	72.1%

Posteriormente, al comparar los resultados obtenidos de ensamblajes realizados con un enfoque guiado por genoma y ensamblaje de *novo*, se observó que se logró ensamblar una mayor cantidad de secuencias con el ensamblaje de *novo* (U = 37; p = 0.045), pero no se encontraron diferencias en términos de la longitud máxima de las secuencias ensambladas (U = 65.5; p = 0.713), longitud promedio (U = 52; p = 0.266) ni en el N50 (U = 76; p = 0.843) (Figura 4, Anexo 1).

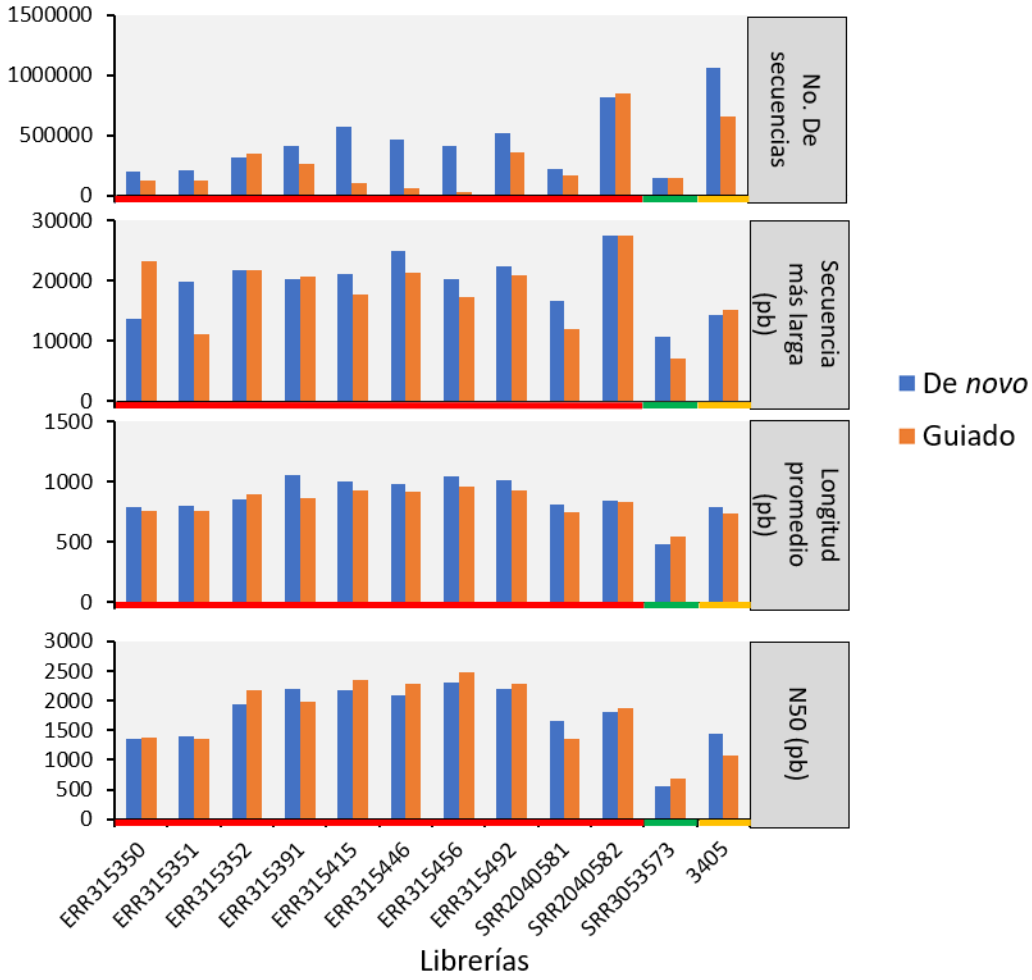


Figura 4. Comparación de diferentes parámetros de calidad para cada ensamblaje *de novo* (azul) y guiado por genoma (naranja) de cada una de las librerías de humano (rojo), gorila (verde) y orangután (amarillo).

Entre todas las librerías de humano y ambas estrategias de ensamblaje, se logró ensamblar un total de 247 transcritos de longitud superior a 3000 pb. Sin embargo, gran parte de estos corresponden a isoformas, lo cual fue corroborado en pasos posteriores del análisis. En el caso de la librería de orangután se obtuvo solo 17 transcritos mayores a 3000 pb (Cuadro 7). Mientras que en gorila los transcritos más largos fueron de alrededor de 2000 pb.

Pese a que se logró ensamblar una mayor cantidad de secuencias con el ensamblaje de *novo*, fue con el ensamblaje guiado por genoma con el cual se logró ensamblar un mayor número de transcritos de longitud superior a 3000 pb. Con este

tipo de ensamblaje se logró ensamblar un total de 194 transcritos, mientras que con el ensamblaje de *novo* se obtuvo solamente 70 transcritos de más de 3000 pb. En las librerías ERR315350 y ERR315351 de humano y en la GT911 de gorila, no se logró obtener transcritos de más de 3000 pb con ninguno de los dos tipos de ensamblaje (Cuadro 7).

Entre las diferentes librerías de humano, se encontró gran variabilidad en cuanto a la cantidad de transcritos de más de 3000 pb ensamblados guiado por genoma y de *novo*. Un ejemplo de esto es la librería SRR2040582, en la cual se obtuvieron 91 transcritos con el ensamblaje guiado por genoma y sólo nueve transcritos con el ensamblaje de *novo*, por lo tanto algunos de los transcritos solamente fueron reconstruidos por uno de los dos enfoques de ensamblaje (Cuadro 7).

Cuadro 7. Total de transcritos de humano, gorila y orangután (>3000 pb) ensamblados mediante ensamblaje guiado por genoma y de *novo*

Librería	**Organismo**	**Tipo de ensamblaje**	
		Guiado por genoma	De *novo*
ERR315350	Humano	0	0
ERR315351	Humano	0	0
ERR315352	Humano	41	5
ERR315391	Humano	4	0
ERR315415	Humano	6	0
ERR315446	Humano	12	14
ERR315456	Humano	0	2
ERR315492	Humano	27	34
SRR2040581	Humano	0	2
SRR2040582	Humano	91	9
SRR3053573	Gorila	0	0
3405	Orangután	13	4
Total		194	70

7.3 Determinación del nivel de soporte de cada transcrito ensamblado

De todos los 247 transcritos reconstruidos (incluyendo las isoformas) a partir de las librerías de RNA-seq de humano, el 22,45% (55 transcritos) presentaron cobertura inferior a 5x y el 77,55% restantes presentaron coberturas entre 5x y 169x. Para los análisis posteriores solamente se conservaron los 192 transcritos que presentaron un nivel de cobertura superior a 5x.

Al comparar la cobertura de los transcritos de cada una de las librerías utilizadas, se encontró que los transcritos de las librerías ERR315352 y SRR2040582 presentaron una gran variabilidad en los valores de cobertura. En las librerías restantes se observó que la cobertura fue muy similar tanto entre transcritos de la misma librería, así como entre diferentes librerías (Figura 5).

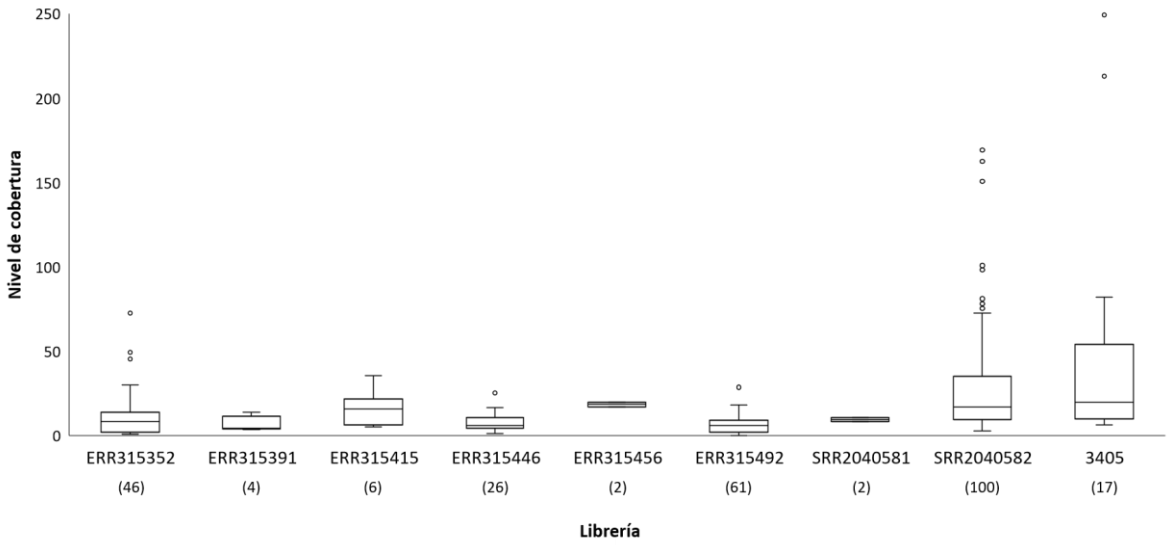


Figura 5. Nivel de cobertura estimado para cada uno de los transcritos de las nueve librerías con ERV >3000 pb. Entre paréntesis se muestra la cantidad de transcritos reconstruidos en cada librería.

Dado que las librerías fueron secuenciadas a diferente profundidad, se estimó el valor de TPM (Transcripts per million), con el propósito de disminuir el efecto de la profundidad de secuenciación en la cobertura de las librerías analizadas. Se observó

que los transcritos de las librerías ERR315415, ERR315456 y SRR2040581 poseen promedios de TPM mayores con respecto al resto de las librerías (Figura 6).

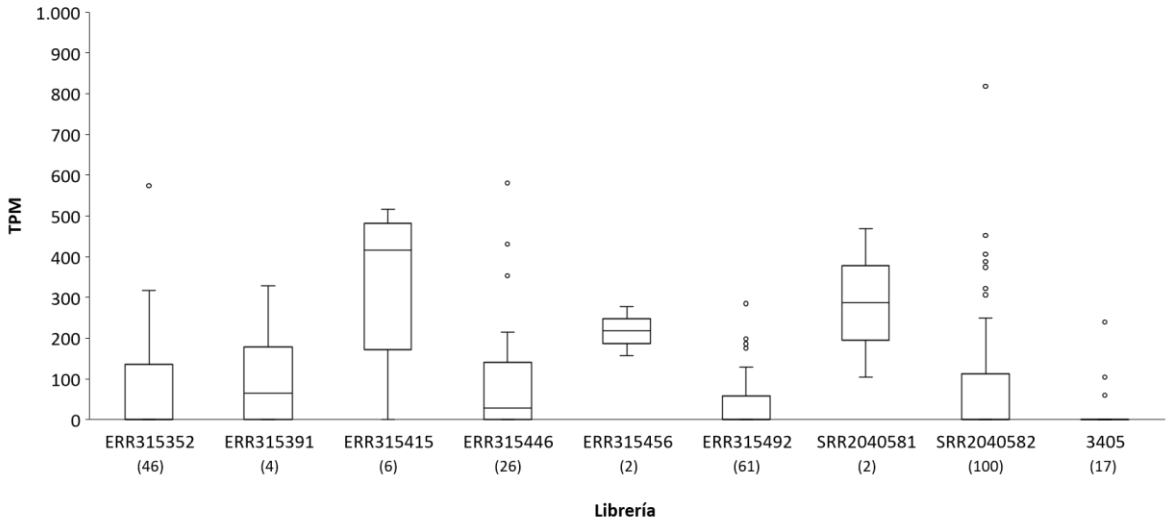


Figura 6. Valores de transcritos por millón (TPM) estimados para cada uno de los transcritos de las diferentes librerías usando Kallisto. Entre paréntesis se muestra la cantidad de transcritos (> 3000 pb) reconstruidos en cada librería.

7.4 Anotación de transcritos ensamblados

Al realizar la anotación de cada uno de los transcritos de tamaño superior a 3000 pb identificados en las especies de primates analizadas, se encontró que la mayoría de los transcritos ensamblados correspondían a isoformas (transcritos pertenecientes a un mismo ERV pero con ligeras variaciones de tamaño o en pocos nucleótidos). Al eliminar todas las isoformas, se pasó de tener 192 transcritos a 77.

Se encontraron diferencias entre los transcritos reconstruidos por ambos métodos de ensamblaje. En múltiples ocasiones ciertos ERVs fueron reconstruidos solamente con uno de los dos métodos de ensamblaje utilizados. Este fue el caso de la librería SRR2040582, la cual fue la librería de humano en la cual se detectaron la mayor cantidad transcritos (n=40), pero 35 de 40 transcritos fueron ensamblados solamente con el enfoque guiado por genoma. Por esta razón a partir de este punto se decidió combinar los resultados obtenidos con ambos métodos de ensamblaje.

Se encontró que los eERVs más frecuentes en las librerías de humano fueron HERV17, HERVK221 y HERVS71, encontrados en cinco de las diez librerías (Figura 7). Mientras que los tipos de ERVs de los cuales se ensambló una mayor cantidad de transcritos fueron HERVK, HERVS71, HERVK31 y HERV17 (Figura 7).

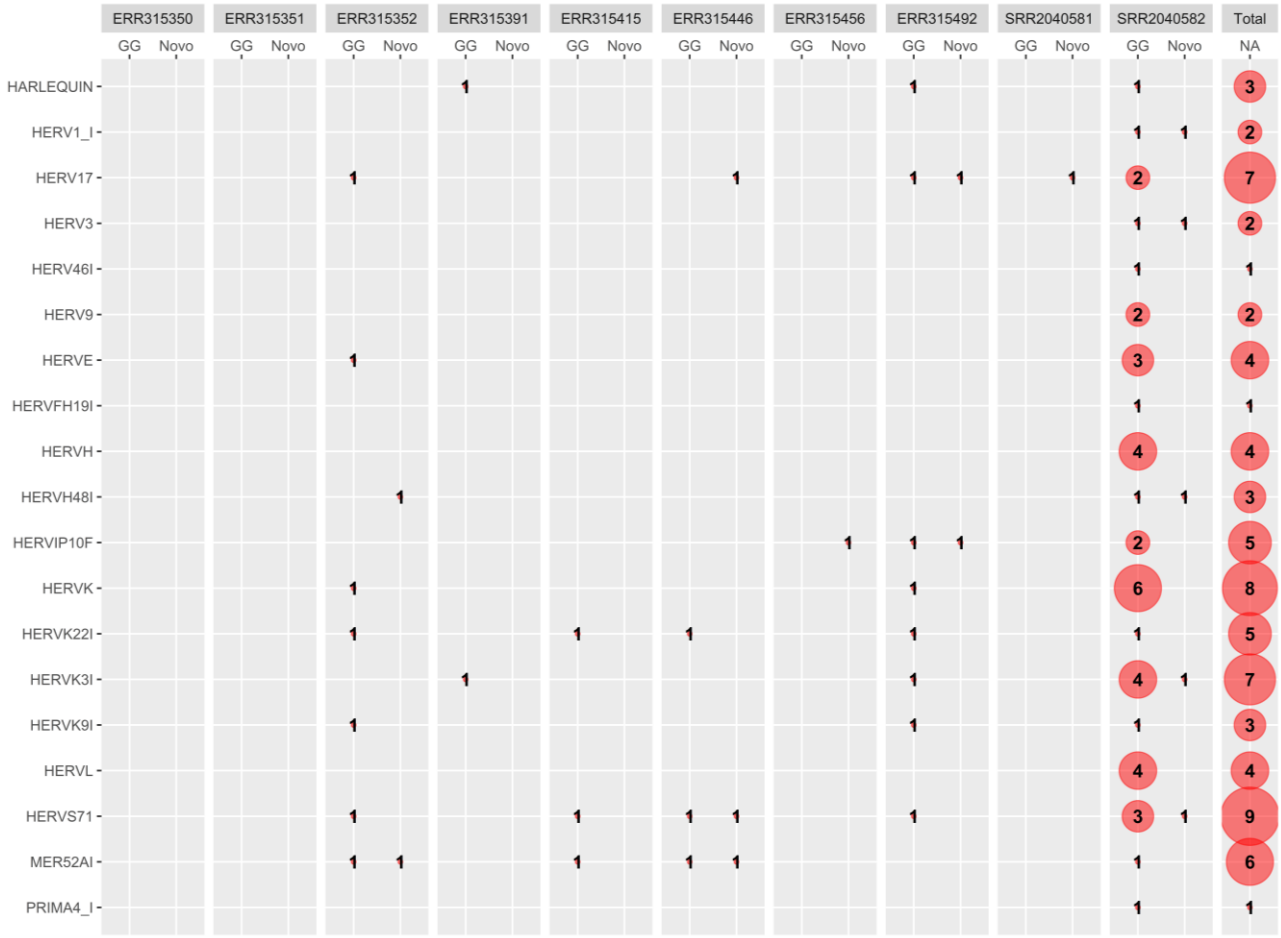


Figura 7. Total de eERVs (> 3000pb) anotados por familia detectados en cada librería de humano analizada.

En el caso de la librería de Orangután también se encontraron diferencias entre los tipos y cantidad de ERVs obtenidos con el ensamblaje de *novo* y con el ensamblaje guiado por genoma. Con el ensamblaje guiado por genoma se detectó la expresión de cinco tipos de ERVs (HARLEQUIN, HERV1_I, HERVE, HERVK y HERVK22I); mientras que con el ensamblaje de *novo* se encontraron solamente dos tipos de ERVs (HERV46I y HERVS71), pero ninguno fue identificado con ambos

enfoques de ensamblaje (Figura 8). Se encontró que HERVE fue el retrovirus que presentó más transcritos en orangután.

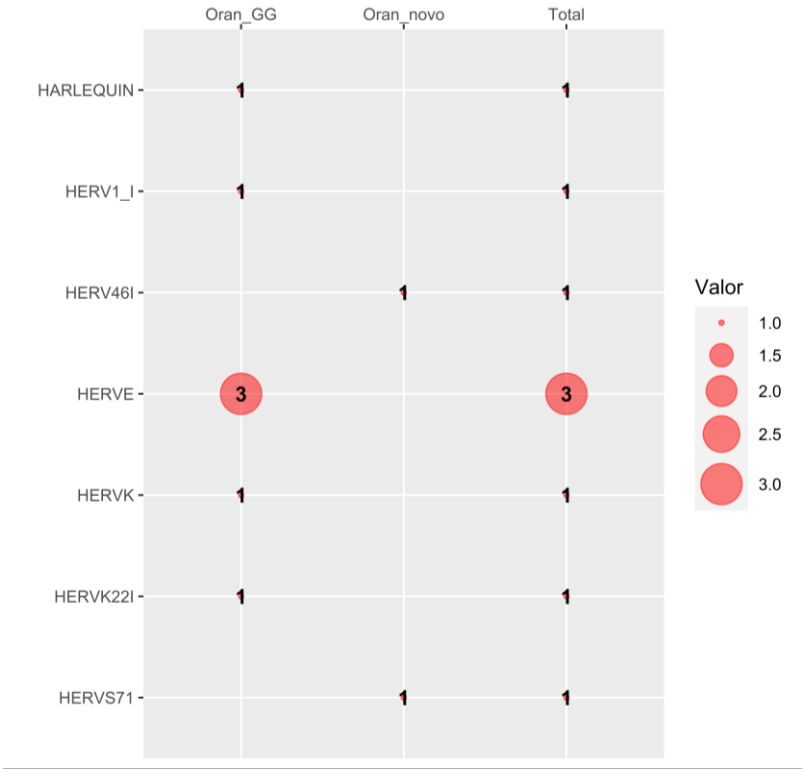


Figura 8. Total de eERVs (> 3000pb) anotados por familia detectados en la librería de orangután (3405).

Se encontró que de los eERVs ensamblados en humano, un 64,9% pertenecían a la familia ERV1, un 29,9% a la familia ERVK y solamente un 5,2% pertenecían a la familia ERVL (Cuadro 8). En el caso de la librería de orangután, solamente se detectaron ERVs de las familias ERV1 y de la familia ERVK, en un 77,8% y un 22,2% respectivamente. Tanto en humano como en orangután la familia de ERVs predominante fue ERV1.

Cuadro 8. Total de transcritos (> 3 kb) correspondientes a cada una de las familias de ERV identificadas en humano, orangután y gorila

Familia de ERV	Humano (%)	Orangután	Gorila
ERV1	50 (64,9 %)	7 (77,8%)	0 (0%)
ERVK	23 (29,9 %)	2 (22,2 %)	0 (0%)
ERVL	4 (5,2 %)	0 (0%)	0 (0%)
Total	77 (100%)	10 (100%)	0 (0%)

7.5 Determinación de los loci de los eERV

Al realizar el alineamiento contra el genoma de referencia utilizando la herramienta BLAT, se determinó la posición en el genoma de cada uno de los eERVs a partir de cada una de las librerías. Se logró determinar la posición de todos los transcritos ensamblados en humano. La Figura 9 muestra la posición de cada uno de los 68 eERVs en cada cromosoma del genoma de referencia hg38.

Figura 9. Posición de cada uno de los 68 eERVs en cada cromosoma del genoma humano. Entre paréntesis se muestra la cantidad de librerías en las que se encontró cada uno de estos elementos.

Se observó que en algunos casos, un mismo tipo de ERV se encontró en diferentes librerías. Esto se observó en ERVs de los cromosomas 2, 5, 7, 10, 11, 12,

14, 17, 18, 19, 22 y X. Se encontraron dos copias expresadas del retrovirus HERVIP10F, una se ubica en el brazo largo del cromosoma 2 y la otra se ubica en el brazo corto del cromosoma 18, en este último se ubica muy cerca del centrómero, esto fue encontrado en dos (ERR315492 y SRR2040582) y tres (ERR315456, ERR315492 y SRR2040582) librerías, respectivamente (Figura 10).

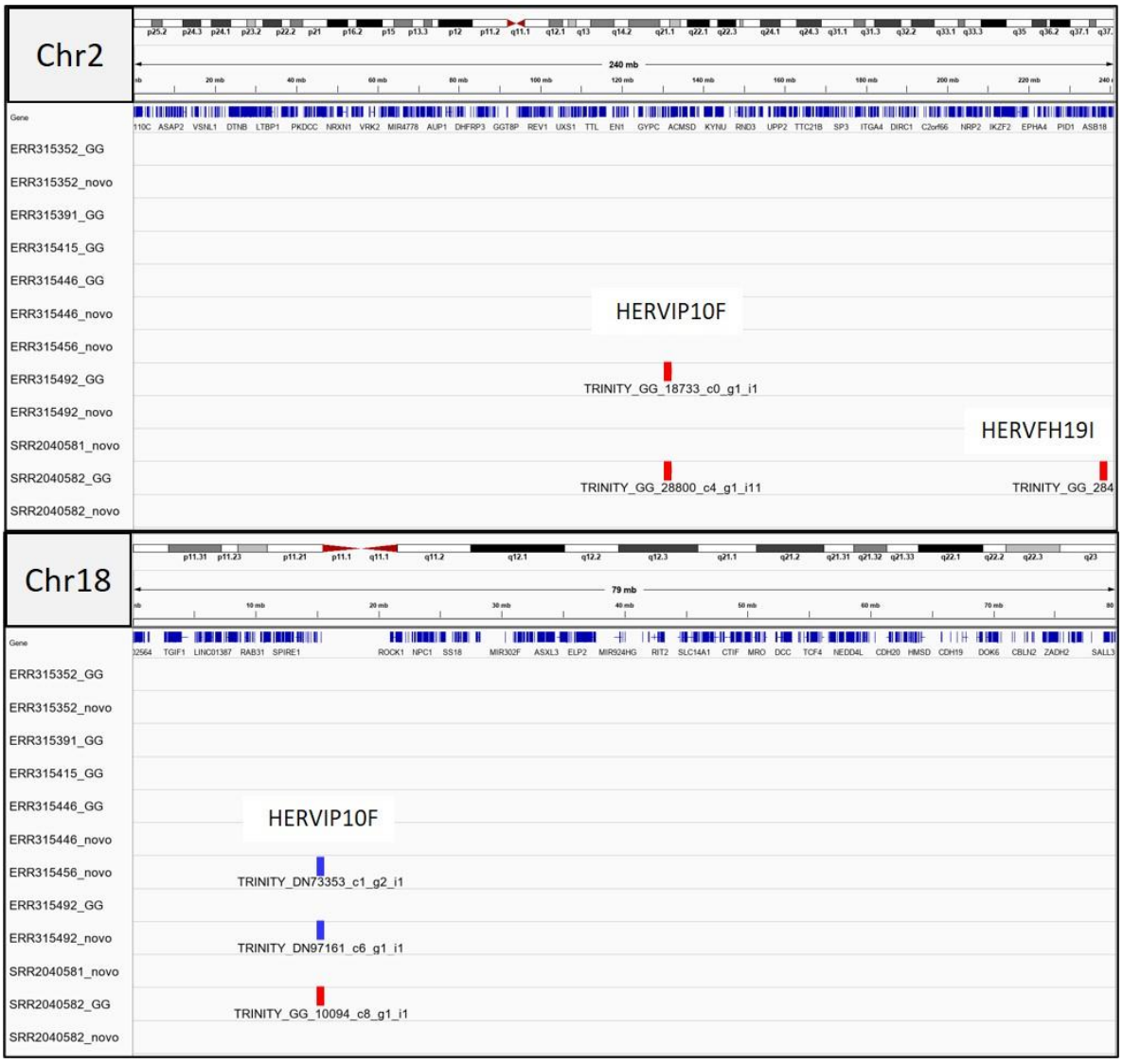


Figura 10. Ubicación en el genoma de copias no idénticas de HERVIP10F reconstruidas a partir de diferentes librerías de humano mediante ensamblaje de *novo* (azul) y guiado por genoma (rojo).

Lo mismo se encontró en el caso de HERVH48I, en el cual también se observó que existen dos copias expresándose de este elemento en diferentes cromosomas. Una copia se encuentra en el cromosoma 14 y otra copia en el cromosoma X, y esto se observó en dos librerías y en una librería, respectivamente (Figura 11).

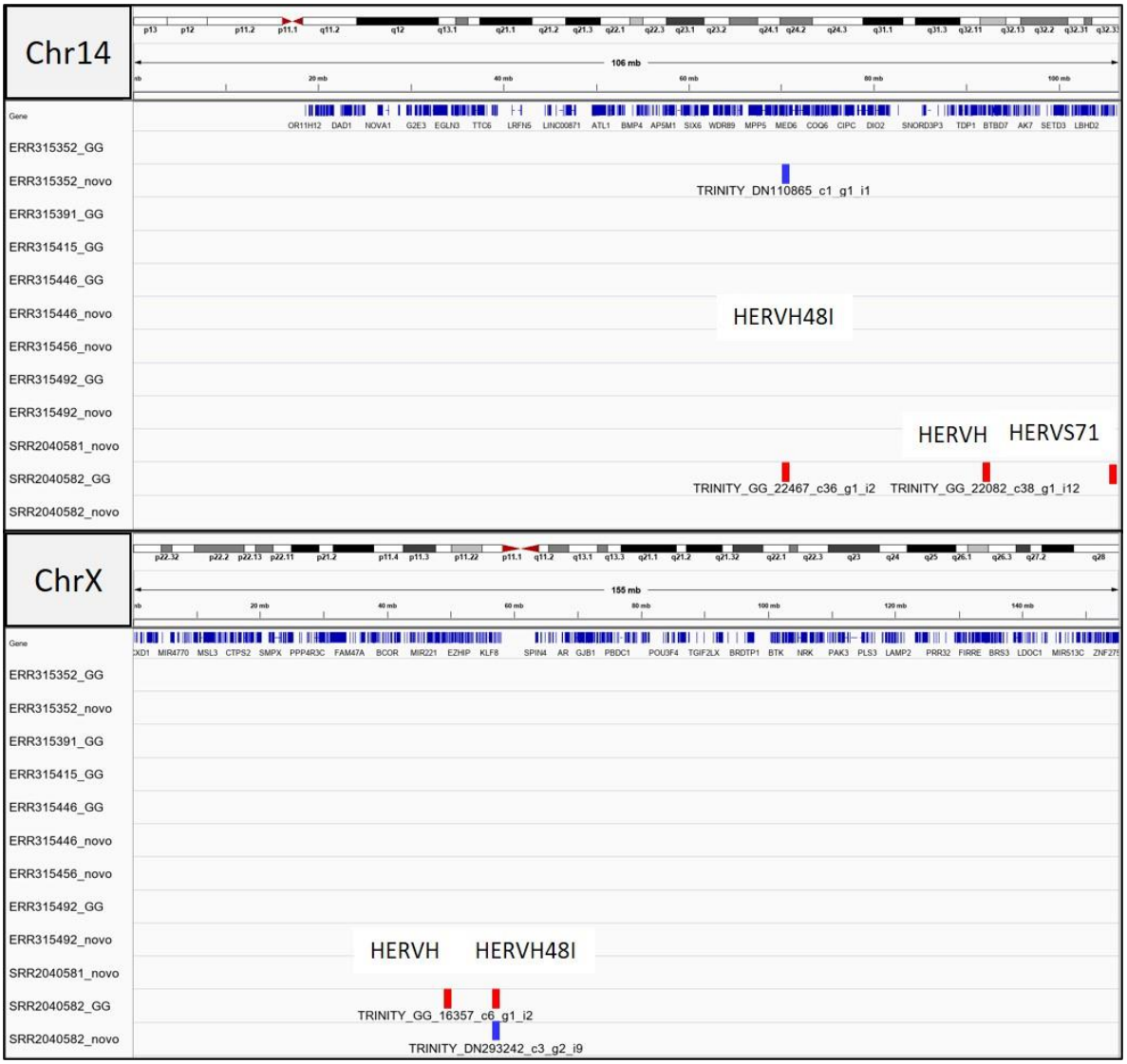


Figura 11. Ubicación en el genoma de copias no idénticas de HERVH48I reconstruidas a partir de librerías de humano mediante ensamblaje de *novo* (azul) y guiado por genoma (rojo).

Se encontraron dos copias del elemento HERVS71, una en el cromosoma 7 y la otra en el cromosoma 10. Esto se encontró en tres (ERR315352, ERR315492 y SRR2040582) y en cuatro (ERR315352, ERR315391, ERR315446 y SRR2040582) librerías, respectivamente. Las librerías ERR315352 y SRR2040582 presentaron ambas copias de HERVS71 en los cromosomas 7 y 10 (Figura 12). En el brazo largo del cromosoma 7 se encontró también una copia del elemento HERV17, el cual fue encontrado en cinco de las 10 librerías de humano analizadas (Figura 12). De estas cinco librerías en las que se encontró este elemento, en las librerías ERR315352, ERR315446 y SRR2040581 solo se logró ensamblarlo con uno de los dos tipos de ensamblaje, y solo en las librerías ERR315492 y SRR2040582 se logró ensamblar dicho elemento tanto por ensamblaje de *novo* así como con el ensamblaje guiado por genoma (Figura 12).

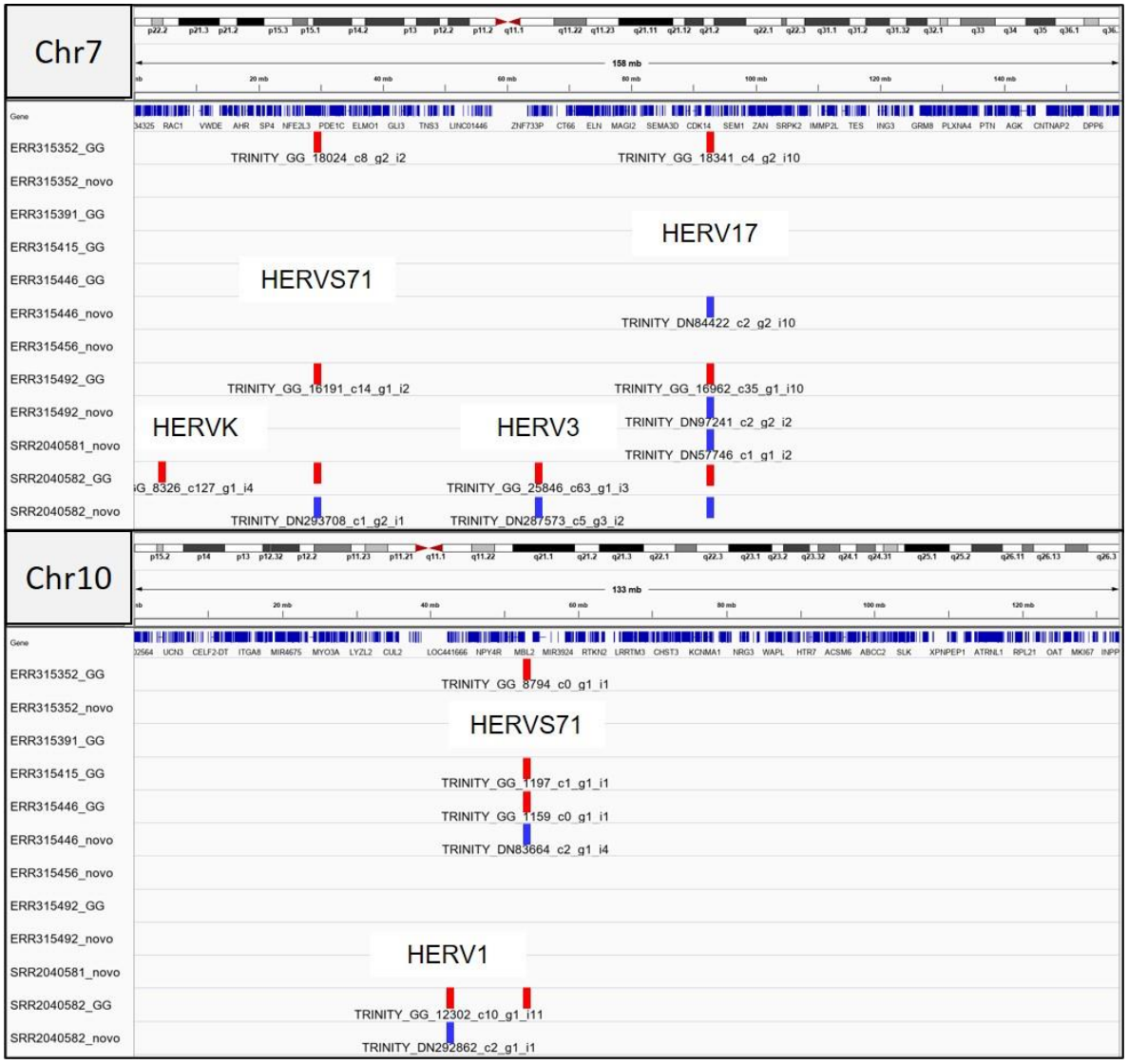


Figura 12. Ubicación en el genoma de copias no idénticas de HERVS71, HERV17, HERV3 y HERV1 reconstruidas a partir de librerías de humano mediante ensamblaje de *novo* (azul) y guiado por genoma (rojo).

Otro elemento que se logró ensamblar a partir de cinco de las librerías de humano fue el elemento HERVK22I ubicado en el cromosoma 12. Este elemento fue encontrado en las librerías ERR315352, ERR315391, ERR315446, ERR315492 y SRR2040582. Y en todas estas librerías solamente se logró reconstruir por uno de los dos métodos de ensamblaje (Anexo 3).

El elemento MER52A se encontró en el brazo corto del cromosoma 11 en cuatro de las librerías de humano. Este elemento fue ensamblado por ambos métodos de ensamblaje a partir de las librerías ERR315352 y ERR315446, y solamente por uno de los métodos de ensamblaje a partir de las librerías ERR315391 y SRR2040582 (Anexo 4).

El elemento HERVK9I, fue ensamblado en tres de las librerías de humano (ERR315352, ERR315492 y SRR2040582) y únicamente por uno de los métodos de ensamblaje. Al localizarlo en el genoma se encontró que se ubica al inicio del brazo corto del cromosoma 5 (Anexo 5).

Se encontró a los elementos Harlequín y HERVK3I en el cromosoma 17 (Anexo 6) y 19 (Anexo 7), respectivamente. Estos elementos fueron ensamblados a partir de las librerías ERR315391, ERR315492 y SRR2040582 de humano.

En el caso del elemento HERVK este fue reconstruido a partir de las mismas librerías de humano ERR315352, ERR315492 y SRR2040582. Al determinar la posición del elemento HERVK en el genoma, se ubicó en el brazo largo del cromosoma 22 (Anexo 8).

Aunque se logró ensamblar transcritos correspondientes a un mismo ERV a partir de diferentes librerías, se observaron diferencias en el tamaño de estos según la librería o el tipo de ensamblaje. Un ejemplo de esto se muestra en la Figura 13, en donde se puede observar el alineamiento de transcritos correspondientes al elemento HERV17 (del cromosoma 7) reconstruidos a partir de diferentes librerías y que presentan grandes variaciones de longitud.

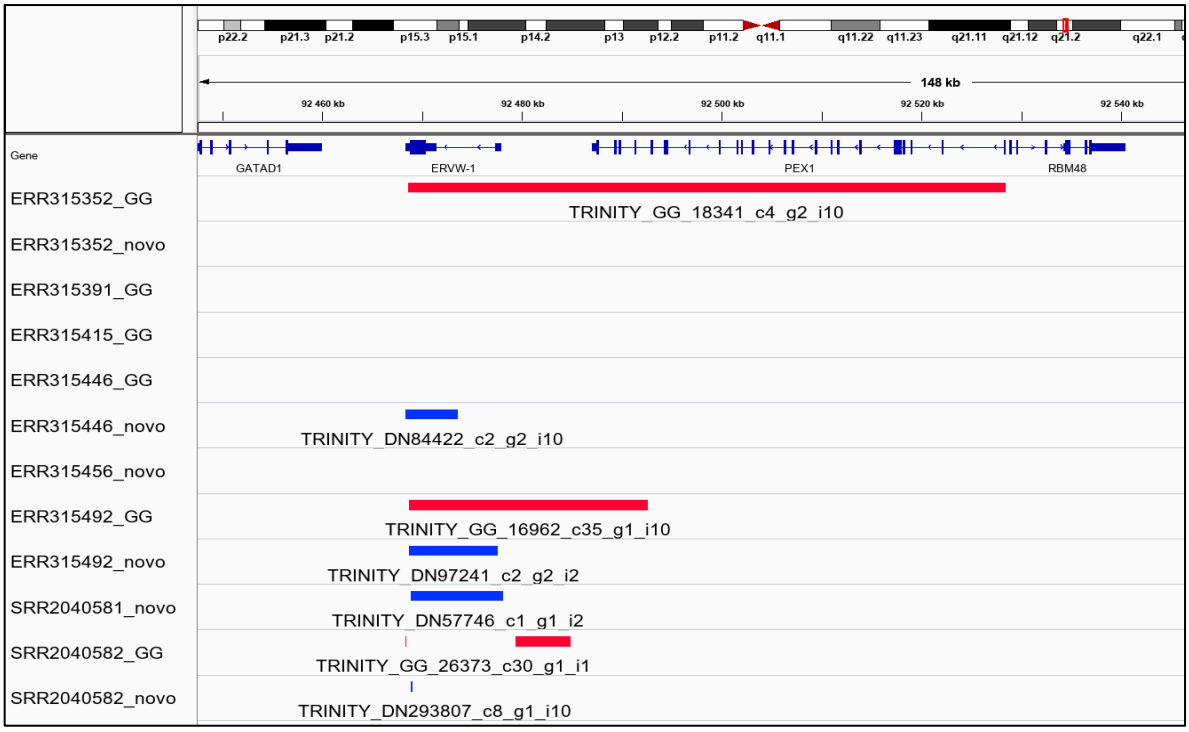


Figura 13. Alineamiento múltiple de las copias de HERV17 del cromosoma 7 reconstruidas a partir de diferentes librerías y a partir de ensamblaje de *novo* (azul) y guiado por genoma (rojo) contra el genoma humano de referencia hg38.

Al realizar un recuento de la cantidad y tipos de ERVs encontrados en cada cromosoma humano entre todas las librerías analizadas, se identificó un total de diez eERVs correspondientes a cinco tipos de ERVs en el cromosoma 7. Los cromosomas 1, 11, 14, 17 y 19 presentaron tres tipos de ERVs expresados. Por el contrario, en los cromosomas 3, 6, 8, 9,15, 20, 21 y Y, no se identificó ningún ERV (> 3 kb) en ninguna de las librerías de humano analizadas (Figura 14).

Se encontró que la mayor densidad (calculada al dividir el total de eERVs entre el tamaño del cromosoma) de ERVs (>3000 pb) expresados se observó en los cromosomas 19 y 22 (cuadro 9). Esto se realizó con el fin de comprobar si el número de ERVs identificados en cada cromosoma es proporcional a las cantidades de ERVs reportadas en UCSC Genome Browser en cada uno de los cromosomas de hg38. En términos generales, se encontró diferencia entre la densidad de eERVs esperada con respecto a la densidad reportada (t=2.77; gl=23; p=0.0109).

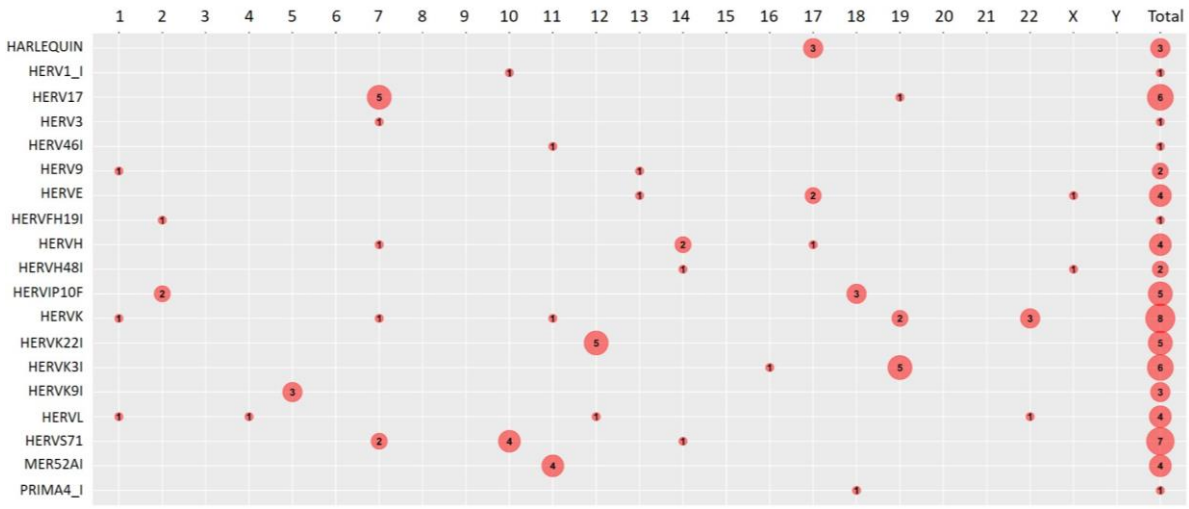


Figura 14. Total de transcritos correspondientes a retrovirus endógenos (>3000pb) en cada cromosoma humano

Cuadro 9. Densidad de ERVs >3kb observada vs esperada en cada cromosoma (Chr) humano

Chr	Tamaño del cromosoma (Mpb)	Total de eERVs	LTRs reportados en hg38	Densidad de eERVs observada	Densidad de LTRs reportada en hg38
1	257	3	228	0.012	0.887
2	242	3	176	0.012	0.723
3	205	0	233	0	1.137
4	192	1	256	0.005	1.333
5	186	3	168	0.016	0.903
6	179	0	214	0	1.196
7	163	10	141	0.061	0.865
8	146	0	144	0	0.979
9	132	0	74	0	0.561
10	142	5	108	0.035	0.761
11	142	6	150	0.042	1.056
12	141	6	138	0.043	0.979
13	116	2	86	0.017	0.741
14	106	4	90	0.038	0.849
15	100	0	42	0	0.420
16	93	1	35	0.011	0.376
17	84	3	32	0.036	0.381
18	82	3	68	0.036	0.829
19	77	7	79	0.091	1.026
20	63	0	22	0	0.349
21	45	0	35	0	0.778
22	48	4	12	0.083	0.250
X	152	2	258	0.013	1.691
Y	59	0	169	0	2.864

*Densidad de ERVs, Total de ERVs / Longitud del cromosoma (Mbp)

Al determinar la capacidad codificante de los 19 tipos de eERVs por medio de un alineamiento con blastx contra la base de datos nr (non-redundant protein sequences), se consideraron como codificantes a aquellos eERVs que tuvieran similitud >75% con alguna proteína depositada en la base de datos nr, que no tuvieran codones de terminación en medio de su secuencia, e-value igual a cero y que tuvieran potencial codificante en fragmentos superiores a 1500 pb. Se encontró que siete tipos de ERV tienen potencial codificante, cuatro de ellos (HERV3, HERVE, HERVH y HERVL) codifican para proteínas de origen retroviral. En el caso de elementos de los cuales se encontraron varias copias en diferentes cromosomas, por ejemplo en HERVK, algunas de estas copias fueron codificantes pero otras no (cuadro 10).

De estos siete transcritos con capacidad codificante solamente los elementos HERV17 (chr7), HERVK (chr22), HERVL (chr12) y el MER52AI (chr11) corresponden a transcritos del gen completo (cuadro 10).

Cuadro 10. Análisis de BLASTx realizado para determinar capacidad codificante de transcritos correspondientes a cada uno de los tipos de ERVs identificados.

ERV	No.[1]	Chr	Primer hit de Blastx	Blastx nr	Cobertura (%) [2]	Identidad (%)	e-value	Longitud eERV	Longitud codificante
HARLEQUIN	3	17	ERV group 3 member Env polyprotein	XP_011741830.1	27	73.6	0	7225	1951
HERV1_I	1	10	uncharacterized protein LOC114603757	XP_028598772.1	64	52.45	0	7149	4575
HERV17	**4**	**7**	**peroxisome biogenesis factor 1 isoform 1**	**NP_000457.1**	**32**	**99.91**	**0**	**10744**	**3438**
HERV17	1	19	ERV polyprotein [Multiple sclerosis associated]	AAB66528.1	47	67	0	4715	2216
HERV3	1	7	endogenous retroviral sequence 3	BAJ21154.1	20	99.8	0	9099	1820
HERV46I	1	11	hypothetical protein DUI87_01908	RMC21052.1	60	26.86	1E-94	4807	2884
HERV9	1	1	hCG2041486	EAW82343.1	37	72	0	4685	1733
HERV9	1	13	hCG2040415	EAW68108.1	30	77.25	2E-141	3740	1122
HERVE	1	13	retroviral gag protein [HERV HCML-ARV]	AAP06676.1	22	68	0	6889	1515
HERVE	2	17	pol protein [HERV HCML-ARV]	AAP06677.1	38	91.8	0	9156	3479
HERVE	1	X	ERV group 3 member 1 Env polyprotein	XP_011741830.1	58	79	0	3509	2035
HERVFH19I	1	2	uncharacterized protein LOC115307850	XP_029819589.1	63	34.4	5E-89	3966	2499
HERVH	1	7	gag-pol precursor polyprotein	YP_009109689.1	57	36	2E-166	5619	3202
HERVH	1	14	Env protein HERV-H	Q9N2K0.1	24	91	0	6989	1677
HERVH	1	17	Env protein [Homo sapiens]	AAD34324.1	31	85	0	5148	1595
HERVH48I	1	14	uncharacterized protein LOC109118148	XP_019063396.1	79	35.6	0	3255	2571
HERVH48I	1	X	uncharacterized protein LOC114684040	XP_028714305.1	55	41.85	1E-102	5180	2849
HERVIP10F	2	2	Select seq gb|RMB88968.1|	RMB88968.1	60	50.8	0	4119	2471
HERVIP10F	3	18	hypothetical protein DUI87_18850	RMC04603.1	70	51.5	0	3854	2698
HERVK	1	1	ERV group K member 7 Pol protein	P63135.1	50	91	0	8526	4263
HERVK	1	7	ERV group K member 7 Pol protein	P63135.1	53	93.7	0	8389	4446
HERVK	1	11	ERV group K member 7 Gag polyprotein	P63130.2	28	75	0	7051	1974
HERVK	1	19	ERV group K member 8 Pol protein	P63133.1	32	54	0	8299	2655
HERVK	**3**	**22**	**Gag-Pro-Pol protein**	**AAD51797.1**	**44**	**84**	**0**	**5924**	**2606**
HERVK22I	5	12	unnamed protein product	BAD18490.1	10	100	1E-97	4377	438
HERVK3I	3	16	ERV group K member 6 Env polyprotein	XP_023059357.1	31	69.8	0	4274	1324
HERVK3I	1	19	Gag-Pro-Pol protein	AAD51797.1	54	44	0	7704	4160
HERVK9I	3	5	Gag-Pro-Pol protein	AAD51796.1	61	65	0	4576	2791
HERVL	1	1	Deoxyuridine triphosphatase	XP_028690305.1	56	54	0	5206	2915
HERVL	1	4	pro-pol-dUTPase polyprotein-murine ERV-L	T29097	65	58.4	0	4757	3092
HERVL	**1**	**12**	**activating transcription factor 7 protein**	**EAW96319.1**	**20**	**99.9**	**0**	**17404**	**3481**
HERVL	1	22	ribonuclease H	SFW04545.1	82	75	0	3503	2872
HERVS71	2	7	uncharacterized protein LOC106699118	XP_023619823.1	67	56	0	7295	4888
HERVS71	4	10	uncharacterized protein LOC102257441	XP_014403533.1	62	54	0	5529	3427
HERVS71	1	14	uncharacterized protein LOC106699118	XP_023619823.1	61	60	0	6198	3780
MER52AI	**4**	**11**	**natural cytotoxicity triggering precursor**	**NP_001189368.1**	**21**	**99.77**	**0**	**6118**	**1285**
PRIMA4_I	1	18	Retrovirus-related Pol polyprotein	TKS65232.1	50	35.9	2E-99	5279	2640

* En gris se muestran todos aquellos ERVs considerados con capacidad codificante y resaltados en negrita los transcritos que corresponden a un gen completo

[1] Número de librerías en las que se encontró ese transcrito

[2] Porcentaje del eERV que alineó con una proteína depositada en la base de datos nr

En orangután se encontraron eERVs en siete cromosomas (6, 8, 10, 11, 13, 17 y 19) como se muestra en el Cuadro 11. Al ubicar la posición de cada uno de los transcritos ensamblados, se encontraron algunas isoformas que fueron eliminadas, de manera tal que al final quedaron únicamente siete tipos de transcritos. En el caso de HERVE, se encontraron dos copias de este elemento en el cromosoma 8 y una copia en el cromosoma 17.

Cuadro 11. Ubicación de eERVs (> 3kb) en el genoma de referencia de orangután PonAbe2 mediante alineamiento con la herramienta BLAT

Retrovirus	Posición genómica		
	Chr	Inicio	final
HERVK22I	6	57686525	57690791
HERVE	8	29665562	29670393
HERVE	8	92617709	92623523
HERV1_I	8	37206503	37210164
HERVS71	10	82838970	82842870
HERV46I	11	89090833	89095876
HARLEQUIN	13	4352239	4372788
HERVE	17	20439059	20443048
HERVK	19	33244565	33247790

7.6 Traslape de ERVs expresados con otros elementos genómicos en humano

Debido a que ciertos ERVs y porciones de estos han sido exaptados a través de la historia evolutiva para desarrollar nuevas funciones en genomas de primates, se investigó cuáles de los eERVs traslapaban con secuencias de genes, lncRNAs y otros LTRs.

Se encontró que un 58,4% de los eERVs traslapó con genes. Algunos elementos como HERV17, HERVH48I, HERVK9I, HERVK3I y HERV3 presentaron traslape con dos genes. Por ejemplo la copia de HERV17 expresada en el cromosoma 7 traslapa con los genes ERVW-1 y PEX1, sin embargo, la copia de este mismo elemento ubicada en el cromosoma 19 no traslapa con ningún gen, lncRNA ni con algún otro LTR. De los diez eERVs que previamente se encontró que poseen capacidad codificante: cuatro de estos (HERVL (Chr12), HERV17 (Chr7), HERV3 (Chr7) y MER52AI (Chr11)) presentaron concordancia con los resultados del blastx y

los genes con los que traslaparon; cuatro de estos (HERVE (Chr17), HERVK (Chr7 y 11) y HERVH (Chr17)) no presentaron traslape con ningún gen; mientras que los dos elementos restantes (HERVK en Chr1 y 22) traslaparon con el gen CD48 que codifica para proteínas receptoras de inmunoglobulinas mientras que con el blastx los transcritos presentaban similitud con proteínas de origen retroviral (Cuadro 12, Anexo 9).

Al evaluar el traslape con lncRNAs se observó que el 32,5% de los ERVs presentaron traslape con estos. Los elementos HERVK, HERVIP10F, HERVS71, HERVK9I, HERVH48I y HERVH, presentaron traslape con más de un lncRNA. En este caso también se observaron diferencias en el traslape según la librería a partir de la cual fueron reconstruidos los eERVs (Cuadro 12, Anexo 9). El elemento HERVK9I fue el que presentó el traslape con la mayor cantidad de lncRNAs, el cual traslapó con un total de 16 lncRNAs.

Al evaluar el traslape entre cada uno de los ERVs identificados con otros tipos de LTRs, se encontró que tres de los ERVs expresados traslaparon con otros LTRs, lo cual sugiere que estos elementos se encuentran anidados. Esto se observó solamente en los elementos HERV17 (Chr 7), HERVK (Chr 19), HERV46I (Chr 11) y en MER52AI (Chr 11). En este caso también se observó que el traslape con LTRs varía entre elementos reconstruidos a partir de diferentes librerías (Cuadro 12, Anexo 9).

En el cuadro 13 se muestra la información de cada uno de los genes que traslaparon con ERVs expresados en humano y que se encontró que poseen capacidad codificante.

Cuadro 12. Traslape de la posición de los 41 eERVs con genes, lncRNAs y LTRs reportados en hg38

ERV	Chr	Tamaño de *locus* (pb)	No. librerías	Genes	lncRNAs	LTRs
HARLEQUIN	chr17	8127	3	NR_110868	0	0
HERV1_I	chr10	8743	1		0	0
HERV17	chr7	60126	5	ERVW-1 /PEX1	0	HERVH
HERV17	chr19	4715	1		0	0
HERV3	chr7	19280	1	ZNF117 / ERV3-1	0	0
HERV46I	chr11	4807	1	JRKL-AS1	1	LTR46
HERV9	chr1	4684	1	WARS2-AS1	0	0
HERV9	chr13	3744	1		0	0
HERVE	chr13	6889	1		1	0
HERVE	chr17	9413	2		1	0
HERVE	chrX	3510	1		1	0
HERVFH19I	chr2	3971	1		0	0
HERVH	chr7	5618	1		0	0
HERVH	chr14	7259	1		0	0
HERVH	chr14	16607	1	ATXN3	2	0
HERVH	chr17	5153	1	RNF213-AS1	0	0
HERVH48I	chr14	3249	1	SYNJ2BP/COX16	0	0
HERVH48I	chrX	5271	1		2	0
HERVIP10F	chr2	22243	2	ANKRD30BL	4	0
HERVIP10F	chr18	21707	3	LOC644669	5	0
HERVK	chr1	8969	1	CD48	0	0
HERVK	chr7	8681	1		0	0
HERVK	chr11	8670	1		0	0
HERVK	chr19	26878	1		0	HERVK9
HERVK	chr19	6408	1	ZNF420	0	0
HERVK	chr22	10939	3	PCAT14	2	0
HERVK22I	chr12	6948	5	GOLGA2P5	0	0
HERVK3I	chr16	4273	1		1	0
HERVK3I	chr19	3515	1	ZNF439	0	0
HERVK3I	chr19	611139	3	ZNF528 / FPR3 / ZNF577	0	0
HERVK3I	chr19	10631	1	NR_144447 / ERVK3-1	0	0
HERVK9I	chr5	60423	3	LOC728613 / SDHAP3	16	0
HERVL	chr1	155109	1		0	0
HERVL	chr4	30588	1		0	0
HERVL	chr12	137303	1	cD48	0	0
HERVL	chr22	3502	1	FBLN1	0	0
HERVS71	chr7	9001	2		3	0
HERVS71	chr10	9128	4		1	0
HERVS71	chr14	6061	1		0	0
MER52AI	chr11	40416	4	NCR3LG1	0	HUERS-P3
PRIMA4_I	chr18	8553	1	TXNL4A	0	0

*Resaltados en gris se muestran los eERVs que poseen capacidad codificante

Cuadro 13. Descripción de las funciones y otras características de cada gen que traslapó con un eERVs con capacidad codificante

eERV[1]	Chr	Gen	Descripción	Función	Expresión en testículo[2]	Cadena	Tamaño nt/aa
HERV17	7	ERVW-1	Codifica sincitina, proteína originada de un gen de cubierta de ERV-W. Este gen es parte de un provirus al que por mutación se le inactivaron los genes gag y pol	Formación del sincitiotrofoblasto y fusión de gametos. La expresión de este gen se ha asociado a esclerosis múltiple	ARNm	-	9607pb /538aa
HERV17	7	PEX1	Codifica un tipo de ATPasa	Tiene diversas funciones, por ejemplo importar proteínas a los peroxisomas y biogénesis de peroxisomas	ARNm y proteínas	-	41512pb / 1283aa
HERV3	7	ZNF117	Codifica para una proteína con dedos de zinc.	Actúa como factor de transcripción de unión a ADN	ARNm	-	34913pb /483aa
HERV3	7	ERV3-1	Codifica para proteína *env* del ERV del grupo 3. La proteína codificada por este gen está sobre expresada en cáncer colorectal y otros tipos de cáncer	Esta proteína ha perdido sus capacidades fusogénicas y ahora inhibe el crecimiento celular a través de disminución de la expresión de ciclina B	ARNm	-	16392pb /604aa
HERVH	17	RNF213-AS1	Codifica un lncRNA	Desconocida	NI	-	79840pb
HERVK	1	CD48	Codifica para un receptor de inmunoglobulinas miembro de la familia CD2 que se encuentra en la superficie de linfocitos, células dendríticas y células endoteliales	Participan en vía de activación y diferenciación celular	ARNm	-	33106pb /243aa
HERVK	22	PCAT14	lncRNA asociado a cáncer de próstata	Desconocida	NI	+	10917pb
HERVL	12	CD48	Codifica para un receptor de inmunoglobulinas miembro de la familia CD2 que se encuentra en la superficie de linfocitos, células dendríticas y células endoteliales	Participan en vía de activación y diferenciación celular	ARNm	-	33106pb /243aa
MER52AI	11	NCR3LG1	Codifica un activador de las células natural killer. Se expresa selectivamente en tumores	Procesamiento y presentación del antígeno, involucrado en el sistema de inmunidad innata	ARNm	+	29930pb /454aa

*Información tomada de GenCards V4.14 (https://www.genecards.org/),
[1] eERVs que corresponden a un gen completo
[2] NI= No existe información

Se visualizó en IGV el traslape entre los eERVs con los genes, lncRNAs y LTRs de RepBase. Se encontraron diferentes escenarios, en algunos casos se encontró que un eERV coincidió con el ERV reportado en el genoma de referencia hg38, pero

no se observó traslape ni con genes ni con lncRNAs. Por ejemplo, en la Figura 15, se muestra al elemento HERVFH19 ubicado en el cromosoma 2, que no presenta traslape ni con genes ni con lncRNAs y se encuentra en una región intergénica.

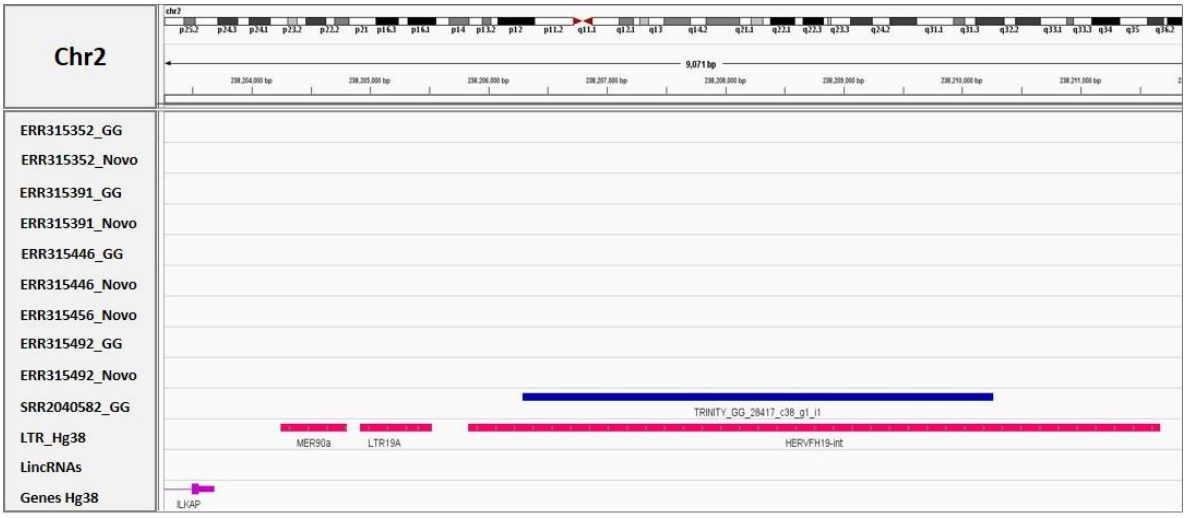


Figura 15. Visualización del traslape del eERV anotado como HERVFH19 (azul) con el ERV reportado en RepBase (rosado).

En otros casos, se observó un traslape entre el eERV, un lncRNA y un intrón. Por ejemplo, en la Figura 16 se muestra al elemento HERVE, ubicado en el cromosoma 13, que traslapó con la región no codificante del gen TPTE2P5 y el lncRNA TCONS_00022264.

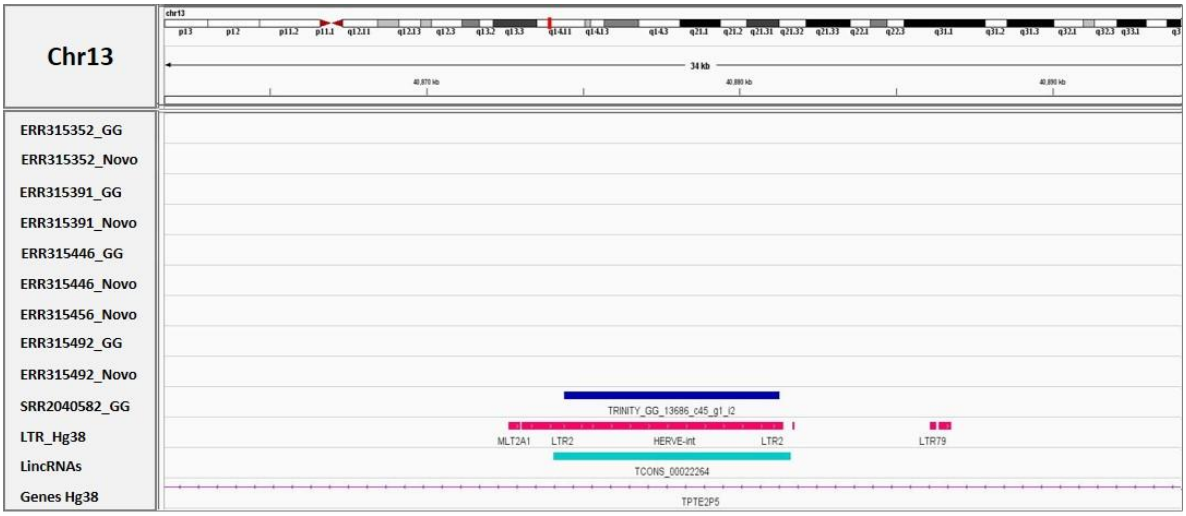


Figura 16. Visualización del traslape del eERV anotado como HERVE (azul) con una región no codificante del gen TPTE2P5 (morado) y el lncRNA (verde) TCONS_00022264 en el cromosoma 13.

Mientras que en otros casos se observó traslape del eERV con un gen y con un lncRNA. Por ejemplo en la figura 17 en donde se muestra al elemento HERVK en el cromosoma 22 que presentó traslape con el gen PCAT14 y con los lncRNAs TCONS_l2_00017644 y TCONS_l2_00017645.

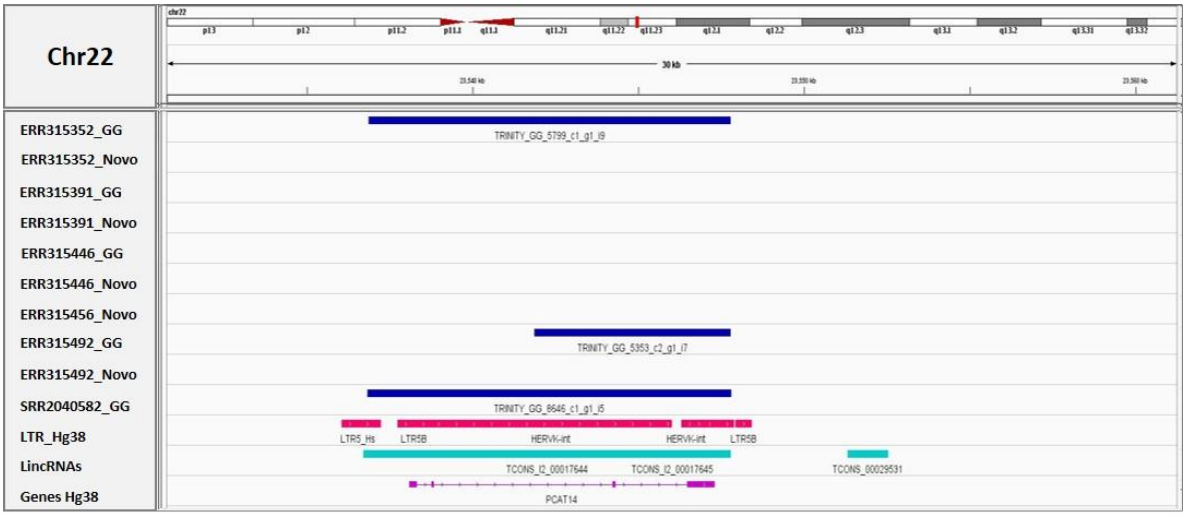


Figura 17. Visualización del intercepto de HERVK-int con el gen PCAT14 (morado) y con los lncRNAs (verde) TCONS_l2_00017644 y TCONS_l2_00017645 en el cromosoma 22.

7.7 Características genómicas de los loci cercanos a los sitios de integración (flancos) de los eERVs

Dado a que un mismo eERV fue detectado en diferentes librerías y la reconstrucción no es completa, se obtuvieron variaciones en los tamaños de estos y por ende en las coordenadas genómicas. Para eliminar la redundancia de tener diferentes coordenadas genómicas para el mismo elemento proveniente de diferentes librerías, se determinaron las coordenadas genómicas consenso correspondientes a la mayor longitud, eligiendo la posición de inicio menor y la posición final más distante. Al realizar esta simplificación de los datos se obtuvieron 41 posiciones de ERVs. En el cuadro 14 se muestra un ejemplo de la forma en la que se seleccionaron las coordenadas representativas de cada eERV. En este caso el elemento MER52AI que fue ensamblado a partir de cuatro diferentes librerías y presentó variaciones en el tamaño, por lo cual las coordenadas consenso se obtuvieron tomando la posición de inicio más baja y la posición final más distante (sombreado en color gris).

Cuadro 14. Ejemplo de obtención de coordenadas genómicas consenso para eERVs de diferentes librerías que presentaron variaciones en coordenadas

Chr	inicio	final	transcrito	ERV	Librería	Longitud
chr11	17336959	17377375	TRINITY_GG_36351_c62_g1_i2	MER52AI	582_GG	40416
chr11	17351965	17377248	TRINITY_GG_7893_c2_g1_i1	MER52AI	415_GG	25283
chr11	17352507	17377290	TRINITY_DN108979_c3_g2_i1	MER52AI	352_novo	24783
chr11	17352511	17377341	TRINITY_GG_26568_c5_g1_i4	MER52AI	352_GG	24830
chr11	17356648	17377242	TRINITY_GG_4089_c7_g1_i2	MER52AI	446_GG	20594
chr11	17356648	17376558	TRINITY_DN80847_c2_g5_i2	MER52AI	446_Novo	19910

En el primer panel de cada recuadro de la Figura 18 se muestra el heatmap que representa la curva de valores de p ajustados, en el segundo panel se muestra la curva de valores de p ajustados con los valores significativos ($p < 0.05$) resaltados en gris, y en el tercer panel se muestran las curvas promedio (señales promedio en cada ventana de 1 kb) para los dos tipos de regiones (flancos vs control).

Se observó que corriente abajo (extremo 3') de los ERVs expresados tiende a haber una mayor proporción de orígenes de replicación (recuadro A, Figura18) y una

mayor proporción de SINEs (recuadro B, Figura 18) con respecto a los flancos de LTRs no expresados. En el caso de los datos de orígenes de replicación, solamente fue posible ajustar unos pocos valores de *p*, sin embargo todos los valores de *p* que se lograron ajustar fueron significativos. En el caso de los SINEs, se lograron ajustar todos los valores de *p* de cada ventana, sin embargo solamente una minoría fueron significativos.

Se encontró que en los flancos de los eERVs existe una mayor proporción de lncRNAs con respecto a los controles (recuadro C, Figura 18). En este caso en particular se encontró una mayor proporción de lncRNAs justo en las regiones donde están los eERVs, lo cual coincide con lo que se observó en la sección 7.6 de traslape con características genómicas.

También se encontró que los flancos de eERV en tejido testicular presentan valores de tiempo de replicación mayores con respecto a los controles ($p < 0.05$) (recuadro D, Figura 18). Estos resultados fueron consistentes en casi todas las 200 ventanas adyacentes a los elementos, tal y como se muestra en el segundo panel del recuadro C de la figura 18.

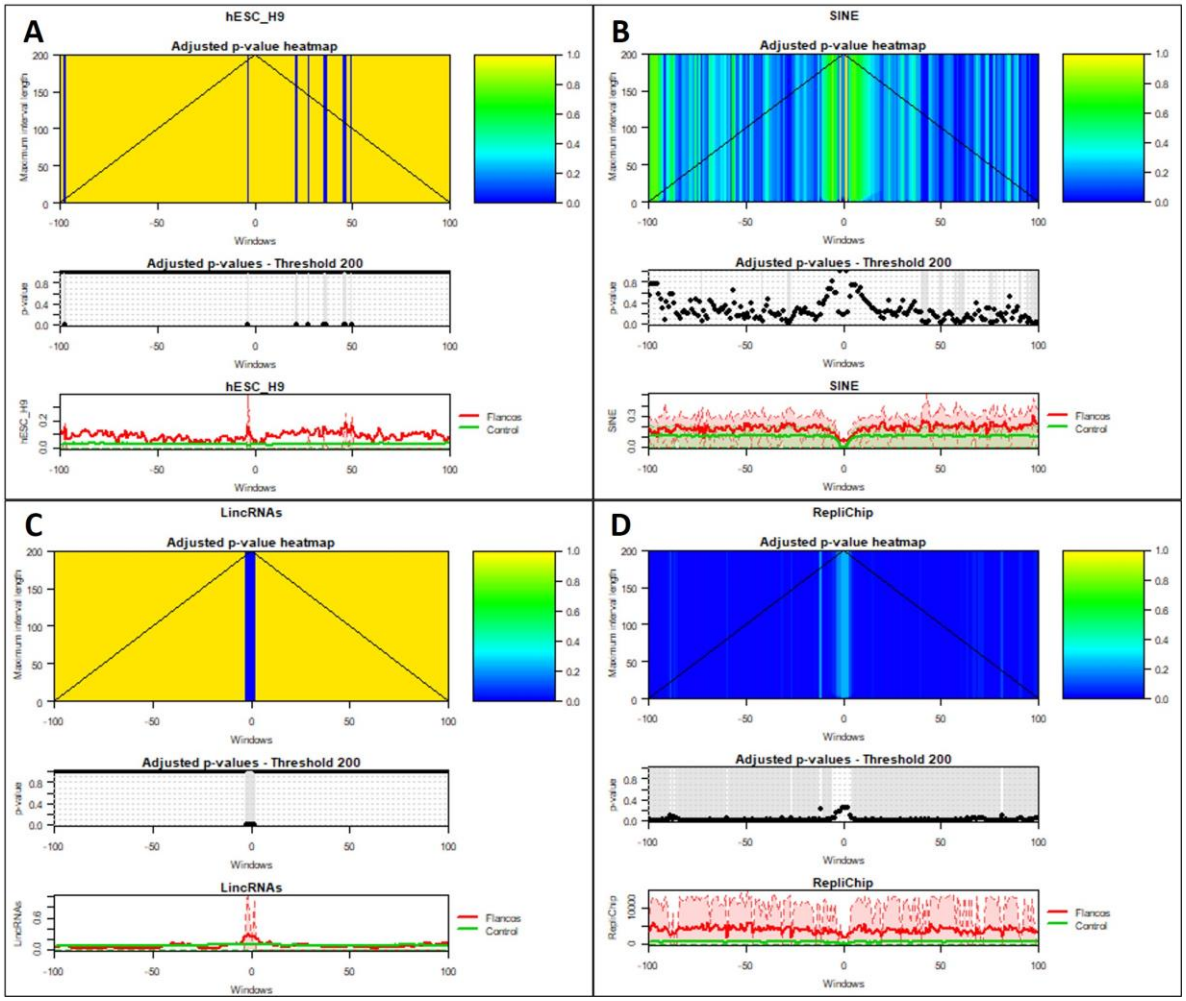


Figura 18. Resultados del análisis funcional de la proporción de orígenes de replicación (A), SINEs (B), lncRNAs (C), y tiempo de replicación (D) estimados en flancos de ERVs expresados y en regiones control.

En la figura 19 se muestra la representación gráfica de los valores de *p* que se lograron ajustar para cada una de las características genómicas pero que no resultaron significativos al realizar la comparación entre flancos de ERVs expresados y flancos de LTRs no expresados (control). Se puede observar como en la proporción de islas CpG y en hot spots de recombinación no fue posible ajustar el valor de *p* en ninguna de las 200 ventanas. Por el contrario, en otras características como por ejemplo la proporción de genes, sí fue posible ajustar el valor de *p* de todas las ventanas, pero estos valores de *p* ajustados no fueron significativos.

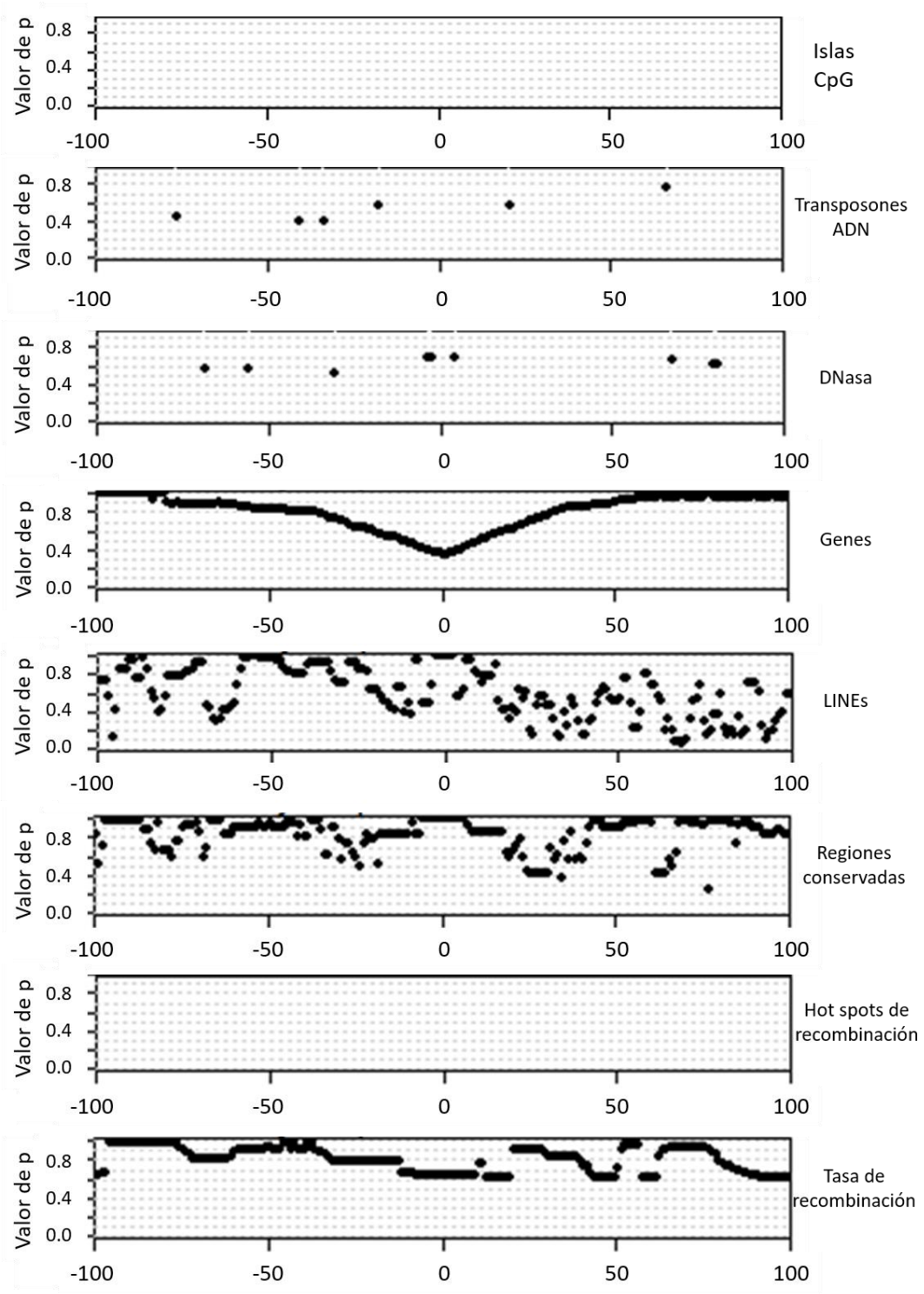


Figura 19. Representación gráfica de los valores de *p* ajustados para cada característica genómica estimada en 200 ventanas y que no mostraron diferencias significativas al hacer la comparación entre flancos de ERVs expresados y flancos de LTRs no expresados.

7.8 Comparación evolutiva de los ERVs presentes en primates

Las 41 secuencias fasta extraídas a partir del genoma humano hg38 utilizando las coordenadas genómicas correspondientes a las diferentes copias de eERVs, se alinearon contra el genoma de gorila y orangután utilizando la herramienta BLAT con el propósito de localizarlas en dichos genomas. Se encontraron 24 de estas secuencias en el genoma de gorila y 15 fueron ubicadas en el genoma de orangután. En total se encontraron diez copias de ERVs compartidas entre las tres especies de primates, pero solamente se encontró potencial codificante en dos de estas correspondientes al elemento HERVK (cuadro 15).

Cuadro 15. Identificación de eERVs de humano en el genoma de gorila (GorGor3) y orangután (PonAbe2)

Chr	Inicio	Final	Humano	Gorila	Orangutan
chr1	55580995	55736104	HERVL		
chr1	119174301	119178985	HERV9	HERV9	HERV9
chr1	**160690937**	**160699906**	**HERVK**	**HERVK**	**HERVK**
chr2	132139713	132161956	HERVIP10F		HERVIP10F
chr2	238206291	238210262	HERVFH19I	HERVFH19I	HERVFH19I
chr4	139417093	139447681	HERVL	HERVL	
chr5	1571908	1632331	HERVK9I		
chr7	4587241	4595922	HERVK		
chr7	29636370	29645371	HERVS71	HERVS71	
chr7	64978994	64998274	HERV3		HERV3
chr7	92468388	92528514	HERV17		
chr7	92479345	92484963	HERVH		
chr10	42643923	42652666	HERV1_I	HERV1_I	HERV1_I
chr10	52946439	52955567	HERVS71	HERVS71	
chr11	17336959	17377375	MER52AI	PTERV1c	
chr11	96501944	96506751	HERV46I	HERV46I	
chr11	**118720831**	**118729501**	**HERVK**	**HERVK**	**HERVK**
chr12	14365654	14502957	HERVL		
chr12	100156286	100163234	HERVK22I	HERVK22I	
chr13	27212761	27216505	HERV9	HERV9	HERV9
chr13	40874388	40881277	HERVE	HERVE	HERVE
chr14	70293163	70300422	HERVH		
chr14	70346788	70350037	HERVH48I	HERVH48I	
chr14	92042979	92059586	HERVH		
chr14	106196990	106203051	HERVS71		HERVS71
chr16	35520414	35524687	HERVK3I	HERVK3I	
chr17	28230587	28240000	HERVE	HERVE	
chr17	43158981	43167108	HARLEQUIN	HARLEQUIN	
chr17	80403343	80408496	HERVH		
chr18	15308637	15330344	HERVIP10F		HERVIP10F
chr18	80025407	80033960	PRIMA4_I	PRIMA4_I	PRIMA4_I
chr19	9435604	9462482	HERVK		
chr19	11854696	11858211	HERVK3I		HERVK3I
chr19	22747259	22751974	HERV17	HERV17	
chr19	37106649	37113057	HERVK	HERVK	HERVK
chr19	51804439	52415578	HERVK3I		
chr19	58305340	58315971	HERVK3I	HERVK3I	
chr22	23536842	23547781	HERVK	HERVK	
chr22	45556467	45559969	HERVL	HERVL	
chrX	49139620	49143130	HERVE	HERVE	HERVE
chrX	57219432	57224703	HERVH48I		

En gris se resaltan los ERVs compartidos entre las tres especies de primates, y en negrita los ERVs compartidos entre las tres especies y que además poseen capacidad codificante en humano

Utilizando las secuencias de las copias de HERVK de los cromosomas 1 y 11 de las tres especies de primates, se estimó la divergencia evolutiva entre las tres especies. Se encontró que en estas dos regiones la divergencia evolutiva entre humano y gorila fue de 0.09, la divergencia entre humano y orangután fue de 0.10 y la divergencia entre gorila y orangután fue de 0.05 sustituciones nucleotídicas promedio por sitio.

Se estimó también la tasa de mutaciones sinónimas (Ks) y la tasa de mutaciones no sinónimas (Ka), a partir de las secuencias de un mismo ERV compartido entre las tres especies. Con estos valores estimados se realizó una prueba basada en los codones mediante el método de Nei-Gojobori con el objetivo de determinar si ocurrió selección purificadora en dicho elemento. Con esto se encontró que ocurrió selección purificadora entre HERVK de humano y orangután (p=0.03) y entre gorila y orangután (p=0.02) (Cuadro16 y 17). Sin embargo, entre humano y gorila se encontró que existe neutralidad entre la tasa de mutaciones sinónimas y no sinónimas (dN = dS) (p>0.05), ver Cuadro 16 y 17.

Cuadro 16. Prueba de selección purificadora basada en codones entre secuencias de copias de HERVK en el cromosoma 1 en humano con respecto a sus homólogos en gorila y orangután.

	Orangután	Humano	Gorila
Orangután		1.84	2.09
Humano	**0.03**		1.42
Gorila	**0.02**	0.08	

En negrita: valores de probabilidad significativos; encima de la diagonal: valores de Ks-Ka obtenidos en la prueba de Nei-Gojobori; debajo de la diagonal: valores de probabilidad obtenidos en cada caso.

Cuadro 17. Prueba de selección purificadora basada en codones entre secuencias de copias de HERVK en el cromosoma 11 en humano con respecto a sus homólogos en gorila y orangután.

	Orangután	Humano	Gorila
Orangután		1.95	2.09
Humano	**0.03**		1.01
Gorila	**0.02**	0.16	

En negrita: valores de probabilidad significativos; encima de la diagonal: valores de Ks-Ka obtenidos en la prueba de Nei-Gojobori; debajo de la diagonal: valores de probabilidad obtenidos en cada caso.

8. DISCUSIÓN

Actualmente se sabe que los ERVs y ETs en general han contribuido en gran medida a la variabilidad genética y diferenciación de los humanos de las otras especies de primates, pero aún se desconocen los detalles precisos sobre como ocurrió esta separación. Debido a la expresión inusual de ERVs que ocurre en tejido testicular, donde tienen el potencial de modificar el genoma y que estas modificaciones sean heredadas a los descendientes, los ERVs pueden ser la clave para elucidar aspectos evolutivos entre especies relacionadas de primates. En este proyecto se buscó determinar la influencia del contexto genómico en la expresión de ERVs y realizar una comparación evolutiva de los ERVs identificados en muestras de tejido testicular de humano, gorila y orangután. Para este fin se utilizaron librerías de RNA-Seq que fueron ensambladas mediante dos estrategias (de *novo* y guiado por genoma), se identificaron los ERVs expresados en cada especie y mediante un FDA utilizando características genómicas de flancos de ERVs expresados y LTRs no expresados en humanos, se determinó la influencia del contexto genómico en la expresión. Además, utilizando secuencias de ERVs expresados en humano se realizó una búsqueda de secuencias homólogas en genomas de gorila y orangután y se analizó el tipo de selección ocurrido en dichos *loci*.

Se encontró que el contexto genómico influye en la expresión de ERVs, algunas características genómicas como proporción de SINEs, orígenes de replicación, tiempo de replicación y lncRNAs se encontraron capaces de influir en la expresión. Se encontró homología entre copias de HERVK de cromosomas 1 y 11 en las tres especies de primates. Con base en la prueba de Nei-Gojobori se determinó que ocurrió selección purificadora en esa región de humano y orangután, especies entre las cuales también se confirmó que existe la mayor divergencia genética.

8.1 Control de calidad de las secuencias y preprocesamiento de los datos

Con el trimming de las librerías fue posible eliminar restos de adaptadores, secuencias sobrerrepresentadas y reads de baja calidad. Con esto se logró obtener librerías con reads de calidad >38 y longitud mayor a 50 pb. En análisis de datos

obtenidos mediante *next generation sequencing*, el preprocesamiento de los datos es un paso clave que determinará la calidad y confiabilidad de los resultados que se obtengan en posteriores análisis. Presencia de secuencias de baja calidad pueden ocasionar resultados subóptimos en análisis posteriores (Bolger et al., 2014), como por ejemplo obtención de transcritos quiméricos.

Para el preprocesamiento de los datos utilizados en esta investigación se decidió utilizar la herramienta Trimmomatic. Esta herramienta ha mostrado resultados similares y en algunos casos superiores a los obtenidos con otras herramientas empleadas en el procesamiento de los reads. Además tiene la ventaja de que es más flexible y personalizable que muchas de las herramientas comúnmente utilizadas para este fin (Bolger et al., 2014).

8.2 Ensamblaje (Guiado por genoma y de *Novo*)

Al realizar el mapeo de las lecturas de cada librería contra el genoma de referencia, se observó que en el caso de las librerías de humano la tasa de alineamiento fue mayor que la tasa de alineamiento observada en orangután y gorila. Esto puede ser debido a que el genoma humano es uno de los genomas más estudiados y por lo tanto más completos y mejor anotados que existen actualmente, permitiendo un porcentaje de alineamiento superior (>90% en la mayoría de los casos). Ejemplo de esto es el hecho de que el cromosoma Y de humano está secuenciado pero el de orangután no y el de gorila aún está en múltiples contigs (Tomaszkiewicz et al., 2016) y no está en la referencia gorGor3 que es una hembra, esto puede ser una de las razones por las cuales se logró mapear una menor cantidad de las lecturas en estas dos especies. Otra razón puede ser la presencia de contaminación de otras especies en las librerías, lo cual no se verificó en este estudio.

Al comparar los resultados de los ensamblajes de *novo* y guiados por genoma, se determinó que se logró ensamblar una mayor cantidad de transcritos en la mayoría de casos con el ensamblaje *de novo*. Esto puede deberse a que con este enfoque de ensamblaje es posible detectar transcritos nuevos, a diferencia del ensamblaje guiado por genoma. Sin embargo fue con el enfoque guiado por genoma de referencia con el

que se obtuvo la mayor cantidad de transcritos de tamaño superior a 3000 pb. Esto se puede deber a que el genoma de referencia permite al ensamblador extender los transcritos lo más que se pueda sin probabilidad de errores, cosa que no es posible si no se cuenta con una guía como en el caso del ensamblaje de *novo*. Otro factor que pudo influir en este resultado es la presencia de otro tipo de transcritos (contaminación), los cuales se ensamblaron con el enfoque de *novo* pero no con el guiado por genoma.

Se observó una gran variabilidad en la cantidad de transcritos de más de 3000 pb entre librerías de humano. En ocho librerías se detectaron en gran cantidad mientras que en dos no se detectaron del todo. Se observó que existe una correlación muy alta entre la cantidad total de transcritos ensamblados y el tamaño de la librería (cantidad de lecturas), sin embargo, no se observó que el tamaño de la librería tuviera correlación con la cantidad de transcritos >3kb. Esto sugiere que otros factores diferentes están determinando la cantidad de transcritos superiores a 3kb, uno de estos factores puede ser el nivel de expresión génica. Dado a que se ha visto que entre los retos de la reconstrucción de transcriptomas está, la gran variabilidad en los niveles de expresión génica que existe en los tejidos. Esto origina una cobertura no uniforme de todos los transcritos (Lu et al., 2013). En donde aquellos genes que tienen un nivel de expresión muy bajo pueden estar cubiertos por una proporción muy baja de reads, haciendo que sea difícil obtener la secuencia completa de toda la longitud del gen (Lu et al., 2013). Puede que este sea el caso de las librerías en las que no se logró reconstruir ningún transcrito de tamaño superior a 3000pb con ninguno de los dos enfoques de ensamblaje. También la calidad del ARN total medida como el RIN influye en el tamaño de los transcritos ensamblados. Sin embargo, en los datos analizados aquí no parece ser determinante pues el ARN de gorila presentó un 8.2 y el de orangután un 5.2 (datos personales Dra. Campos), los ARN de humano presentaron un RIN>7.5 (Fagerberg et al., 2014; Ruiz-Orera et al., 2015).

Dentro de una misma librería también se observó variabilidad entre la cantidad de transcritos reconstruidos por ambos enfoques de ensamblaje (guiado por genoma

y de *novo*). Tanto el ensamblaje guiado por genoma así como el ensamblaje de *novo* tienen ventajas y desventajas. Por esta razón, ciertos autores recomiendan que para obtener una reconstrucción más completa de un transcriptoma se deben mezclar ensamblajes realizados con genoma de referencia y ensamblajes *de novo*, con el fin de obtener mejores resultados (Jain et al., 2013; Lu et al., 2013). La reconstrucción del transcriptoma es un paso fundamental para poder realizar posteriormente otros análisis con los transcritos ensamblados (Lu et al., 2013).

Con el objetivo de lograr ensamblar eERV de la mayor longitud posible, en estudios posteriores similares a este es recomendable realizar un paso de "scaffolding" posterior al ensamblaje utilizando herramientas para este fin como por ejemplo SSPACE (Boetzer et al., 2011). Es recomendable también realizar ensamblaje utilizando otros ensambladores, dado que ninguno de los ensambladores que se emplean en estos estudios cumple por completo las condiciones de sensibilidad, especificidad y recuperación necesarias para poder ensamblar por completo un transcriptoma (Jain et al., 2013).

8.3 Determinación del nivel de soporte de los transcritos ensamblados

El hecho de que alrededor del 22% de los transcritos ensamblados no presentaron suficiente soporte se debe al alto número de isoformas encontradas en los datos. Se observó que había presencia de transcritos idénticos en su secuencia nucleotídica pero con variaciones de tamaño. Por este motivo en este paso del protocolo, se eliminaron todas aquellas isoformas que presentaron valores de cobertura menores a 5x y se eligieron las isoformas más largas de cada uno de los transcritos.

La gran variabilidad en los valores de cobertura observados en las librerías ERR315352 y SRR2040582, probablemente se asocien a que estas dos librerías fueron en las que se logró ensamblar una mayor cantidad de transcritos y por ende es esperado que presenten mayor ámbito de cobertura en sus transcritos. Sin embargo una vez que se estimaron los valores de TMP, al realizar esta misma comparación entre librerías no se observaron diferencias. Esto debido a que con la estimación del

TMP se eliminan las diferencias en los valores de cobertura que podrían deberse a diferencias en la profundidad a la que fueron secuenciadas las librerías.

8.4 Anotación

Muchos estudios han sugerido que la actividad de los retroelementos está suprimida en el genoma con el propósito de restringir los efectos potenciales de mutaciones en el genoma y asegurar el mantenimiento de la estabilidad genómica (Goff, 2004), no obstante, estudios recientes han demostrado que la actividad de algunos ERVs está incrementada en tejido embrionario (Glinsky, 2015). Esto concuerda con lo observado en este estudio, en el cual se encontró expresión de 19 tipos diferentes de ERVs en tejido testicular de humano. En humanos se han reportado cerca de 30 grupos de retrovirus endógenos humanos (HERV) (Soygur & Sati, 2016). En esta investigación se observó mucha variabilidad en cuanto a la cantidad y tipos de transcritos expresados en cada librería. Los ERVs más frecuentes entre las librerías de humano analizadas fueron HERV17, HERVK22I y HERVS71, los cuales estaban presentes en cinco de las diez librerías analizadas.

En el caso de orangután, pese a que se lograron reconstruir una gran cantidad de transcritos sólo una proporción muy baja de estos correspondieron a eERVs. La tasa de alineamiento hace sospechar que pudo haber contaminación de la muestra, por lo que se recomienda para estudios posteriores realizar un análisis para determinar presencia de contaminación con otros organismos, así como analizar más individuos de la misma especie para poder comparar los resultados. Ya que en este caso en el que solamente se analizó la librería de un único individuo no es posible determinar si la baja cantidad de eERVs se debe a una condición particular de la librería utilizada o si por el contrario es un resultado común en esta especie.

Se ha visto que los retrovirus endógenos se encuentran silenciados por metilación en la mayor parte del genoma, sin embargo esta metilación no es homogénea a lo largo del genoma ni entre células de diferentes tejidos. En general se ha visto que en tejido testicular se presenta una baja metilación y esto favorece la expresión de ERVs (Soygur & Sati, 2016). Se ha encontrado que los retrovirus de la

familia HERV pueden actuar como promotores y que en placenta, dado a que en este tejido hay una hipometilación global, se favorece su expresión (Reiss et al., 2007).

Se ha comprobado que elementos repetitivos originados a partir de ERVs se transcriben sistemáticamente en etapas específicas de la embriogénesis humana. Además se ha visto que muchos elementos LTR7-HERVH producen lncRNAs que son requeridos para el mantenimiento del estado pluripotente, sugiriendo que los ERVs pueden tener diversas funciones aparte de su propia retrotransposición (Göke et al., 2015).

Al comparar los valores de TPM de cada uno de los ERVs ensamblados, en general se observó mucha variabilidad en los valores de copias de un mismo transcrito ensamblado a partir de diferentes librerías. Estudios previos han demostrado que el nivel de expresión de un ERV puede variar por diferentes razones, entre ellas la etapa en el desarrollo en la que se encuentre el organismo (Okahara et al., 2004). Lo cual puede estar influyendo en este estudio dado que las librerías fueron obtenidas a partir muestras de individuos con edades muy diferentes, un factor que se podría analizar más en detalle a futuro.

Cuando se anotaron los transcritos reconstruidos por alineamiento con la base de datos Repbase, se encontró que algunos de los transcritos fueron reconstruidos solamente por uno de los dos métodos de ensamblaje. Esto reafirma la necesidad de complementar este tipo de análisis con los resultados de ambos enfoques de ensamblaje, guiado por genoma y de *novo*.

Finalmente, se determinó la capacidad codificante de cada uno de los 19 tipos de ERVs encontrados. Con esto se encontraron diez elementos con capacidad codificante. Se encontró que los ERVs HERV3, HERVE, HERVH y HERVL codifican para proteínas retrovirales, según el blastx realizado. Lo cual sugiere que dichos elementos podrían estar cumpliendo funciones biológicas muy importantes dado que conservan su capacidad codificante. Sin embargo los que no tienen capacidad codificante también pueden ser importantes como reguladores de la transcripción, entre ellos los lncRNAs.

En el caso de elementos en los cuales se encontraron varias copias en diferentes cromosomas y que se observó capacidad codificante en algunas de estas copias pero en otras no, esto se debe a que copias de un mismo ERV se insertaron en diferentes momentos de la historia evolutiva de los primates y esto ha ocasionado que dichas copias hayan acumulado diferentes mutaciones, lo cual explica la divergencia genética observada entre copias de un mismo elemento.

8.5 Determinación de la posición de cada eERV en el genoma

La detección del eERV HERV17 en el chr7 en cinco de las librerías de humano, sugiere que este elemento cumple funciones biológicas muy importantes en tejido testicular. Al determinar la posición exacta de este elemento, se encontró que se ubica en una región del cromosoma que corresponde a los genes ERW-1 y PEX-1 y que una proporción del transcrito de alrededor de 3400 pb tiene potencial codificante, lo que se discute más adelante. Se encontró variabilidad de tamaños en los transcritos correspondientes a este elemento. Esto también se ha observado en otros estudios, lo cual se debe a que ocurre splicing alternativo que produce transcritos de este gen de diferente tamaño (RefSeq, Mar 2010).

Los retrovirus HERVK22I, HERVS71, MER52AI, HERVK9I, HARLEQUIN, HERVIP10F, HERVK3I y HERVK fueron encontrados en varias de las muestras analizadas. Esto sugiere que la expresión de estos en tejido testicular es algo común. Estudios previos han reportado altos niveles transcripcionales de HERVK en testículo humano (Kim et al., 2004), debido a lo cual es esperable haberlos encontrado expresándose en tejido testicular y que estos resultados fueran consistentes en varias de las librerías. En el caso de HERVS71, MER52AI, HARLEQUIN y HERVIP10F no se encontró reportes previos de expresión en tejido testicular.

También se detectó expresión de diferentes copias de un mismo tipo de ERV, que se encuentran localizadas en diferentes cromosomas. Esto se explica debido a que los procesos de inserción y fijación en el genoma han ocurrido en múltiples ocasiones a través de la evolución, en donde algunos ERVs se han insertado y han

logrado fijarse en el genoma incluso en varias ocasiones, mientras que otros no han logrado fijarse o han sufrido mutaciones que truncan su capacidad de expresión.

En este estudio se encontró que el cromosoma 7 fue el que presentó la mayor cantidad de eERVs con cinco tipos diferentes, seguido de los cromosomas 1, 11, 14, 17 y 19 con tres tipos de ERVs cada uno. Este último cromosoma fue el que presentó una mayor densidad de ERVs expresándose. Esto concuerda con lo observado en estudios previos, que también han encontrado que el cromosoma 19 posee una mayor cantidad de elementos en comparación con otras regiones del genoma (Katzourakis et al., 2007). Katzourakis et al., 2007 sugieren que esto puede deberse a la baja densidad de genes y elevadas tasas de integración de ERVs en este cromosoma.

Sin embargo, otros estudios que han evaluado la cantidad de ERVs en todo el genoma humano, han reportado que en el cromosoma 4 y Y es donde se encuentra la mayor cantidad (Kim et al., 2004). Debido a esto se esperaría que estos cromosomas presentaran también una alta expresión de ERVs, pero esto no fue lo que se encontró en este estudio, probablemente debido a que en este caso solo se consideraron elementos de tamaño superior a >3kb. Esto indica que no precisamente en las regiones del genoma en donde exista mayor abundancia de ERVs van a ser las regiones en donde más se expresen, esto comprueba que existen otros factores que influyen en la expresión de ERVs. Por ejemplo, la importancia de los genes presentes en un cromosoma influye en el nivel de regulación en este cromosoma (Kim et al., 2004).

Por el contrario, en los cromosomas 3, 6, 8, 9, 15, 20, 21 y Y no fue posible identificar ningún ERV expresándose. Contrario a lo esperado en el caso de los cromosomas 3, 6, 8 y 9 en los cuales se conoce que poseen gran cantidad (más de 75) de LTRs mayores a 3000pb. Mientras que en caso contrario los cromosomas 15, 20 y 21 son de los que poseen menor cantidad de LTRs mayores a 3000pb, por lo cuál es no es raro no haber podido encontrar eERVs en estos cromosomas.

Los elementos reconstruidos en las librerías analizadas pertenecían principalmente a la familia ERV1 (clase I) y a la familia ERVK (clase II), esto se observó

tanto en las librerías de humano como en la de orangután. Solamente se encontró la expresión de un elemento perteneciente a la familia ERVL en humano (clase III). Esto concuerda con lo esperado, dado que estudios previos han reportado que ERVs del grupo HERVK son los más activos transcripcionalmente debido a que fueron los últimos en fijarse en el genoma (Garcia-Montojo et al., 2018). Los elementos de las clases I y II se originaron hace aproximadamente 50 millones de años, y los elementos pertenecientes a estas clases tienden a ser transcripcionalmente más activos que los pertenecientes a la clase III dado que han acumulado menos mutaciones debido a su reciente inserción en el genoma (Barbulescu et al., 1999, 2001). Sin embargo, el único elemento de la clase III encontrado (HERVL) en este estudio posee una secuencia de cerca de 3400 pb con capacidad codificante para una proteína que interactúa con el factor activador de la transcripción 7 (ATF7IP), la cual es una proteína nuclear que se asocia con la heterocromatina. Esta proteína puede actuar como coactivador o correpresor de la transcripción (Liu et al., 2009). Esta importante función biológica adquirida por el HERVL explica por qué pese a que es un elemento antiguo aún en la actualidad posee capacidad codificante.

8.6 Traslape de ERVs con elementos genómicos

Se encontraron todas las alternativas de traslape posibles de elementos genómicos con eERVs. Se encontraron ERVs sin traslape con ninguna característica como genes o lncRNAs, traslape solo con genes, traslape solo con lncRNAs, traslape simultáneo con genes y con lncRNAs.

En este estudio se encontró que más de la mitad de los retrovirus endógenos identificados (45/77) en testículo, coinciden con regiones genómicas correspondientes a 29 genes diferentes. La alta proporción de traslape entre genes y eERVs es esperable, dado a que según estudios previos, pueden surgir nuevos genes a partir de retrotransposición de un ARNm que se retrotranscribe a ADN y se inserta en el genoma (Carelli et al., 2016). Se ha visto que en los testículos es donde evolucionan la mayoría de retrogenes, lo cual es facilitado por el carácter permisivo de la cromatina que favorece la transcripción en este tejido. Carelli y colaboradores

(2016) observaron que los genes jóvenes se expresan específicamente en testículo y sugieren que la expresión inicial de un retrogen es facilitada por promotores insertos en el genoma o bien por el surgimiento de nuevos promotores que le permiten a la nueva retrocopia expresarse. Se ha visto que en humanos islas CpG, contribuyen a la expresión de retrocopias, dado a que juegan un papel como activadoras de promotores (Carelli et al., 2016). Debido a esto, la expresión de eERVs en tejido testicular pudo haber sido una fuente de nuevos genes a través de la historia evolutiva de las especies de primates y eucariotas en general.

Estudios previos en los que se ha realizado una búsqueda para detectar genes humanos de origen retroviral, han encontrado un total de 16 genes de cubierta con capacidad codificante expresándose en diferentes tejidos sanos (de Parseval & Heidmann, 2005). Muchos HERVK aún conservan genes con capacidad codificante dado que han sido exaptados para beneficio de su hospedero (de Parseval & Heidmann, 2005). En este estudio se encontró una copia del elemento HERVK en el cromosoma 22 con capacidad codificante correspondiente a la longitud total del gen PCAT14 (según información disponible en GeneCards). Este es un gen que previamente no ha sido reportado expresándose en testículo, pero se sabe que codifica para un lncRNA de función desconocida pero que ha sido asociado a cáncer de próstata.

Nekrutenko & Li analizaron un total de 13799 genes humanos y demostraron que alrededor de 550 genes contenían elementos transponibles en sus regiones codificantes. También se han encontrado genes en regiones no codificantes, sin embargo, estos elementos son capaces de promover variación regulatoria y diversificación de genes a través de la donación de señales reguladoras transcripcionales (van de Lagemaat et al., 2003). Algunos elementos como HERV17, HERVH48I, HERVK9I, HERVK3I y HERV3 presentaron traslape incluso con dos genes diferentes.

Entre los diferentes genes encontrados traslapando con ERVs, se detectaron algunos que codifican para proteínas con dedos de zinc como por ejemplo los genes ZNF528, ZNF420, ZNF439, ZNF577, ZNF117 y RNF213. Algunos autores han

encontrado relación entre los ERVs y genes codificantes para proteínas con dedos de zinc (Gemmell et al., 2019). Se ha encontrado que la presencia de retroelementos LTR puede promover la duplicación y divergencia de los genes en tandem que codifican para las proteínas de dedos de zinc del hospedero y que el número de ambos está correlacionado (Thomas & Schneider, 2011).

Thomas & Schneider, 2011 observaron que la integración al genoma de nuevas familias de retrovirus endógenos está correlacionada con la aparición de nuevas repeticiones de genes ZF (zinc fingers) (Thomas & Schneider, 2011). Debido a esto existe la hipótesis de que los genes en tándem con dedos de zinc evolucionaron en represores transcripcionales para reprimir la actividad retroviral (Thomas & Schneider, 2011). Sin embargo, al evaluar la capacidad codificante de estos eERV que a nivel de secuencia nucleotídica tienen similitud con proteínas con dedos de zinc, a nivel de proteína tienen más similitud con proteínas de origen retroviral, dos de tres con capacidad codificante. Lo cual sugiere que estos genes fueron exaptados a partir de retrovirus para cumplir nuevas funciones en humanos. Solo la copia del HERV3 del cromosoma 7 fue similar con una de estas proteínas con dedos de zinc según los resultados del Blastx, dicha proteína ha sido previamente reportada en testículo en donde actúa como un factor de transcripción.

Se observó traslape entre el HERV17 y el gen ERW-1. Este gen codifica para la sincitina 1, este gen fue exaptado de un retrovirus endógeno hace 25 millones de años y actualmente cumple funciones biológicas muy importantes (Voisset et al., 1999). Este gen es parte de un provirus que sufrió inactivación mediante mutaciones en los genes *gag* y *pol*, pero el gen ERW-1 se mantiene transcripcionalmente activo y codifica para una glicoproteína de la envoltura viral. La proteína codificada por este gen es expresada en el sincitiotrofoblasto placental y está involucrada en la fusión de células del citotrofoblasto para formar la capa sincitial de la placenta (Frendo et al., 2003). Se cree que la sincitina 1 está involucrada en la fertilización y que probablemente contribuya a la fusión de gametos, dado que se ha encontrado la expresión de esta proteína en la superficie celular de esperma mientras que en los oocitos no, en los oocitos por el contrario se expresa el receptor de sincitina 1

(SLC1A5) (Soygur & Sati, 2016). En GeneCards (https://www.genecards.org/) hay reportes de presencia de ARNm de este gen en testículo, pero no se han reportado proteínas en este tejido.

En la región del cromosoma 7 donde mapeó el transcrito identificado como HERV17, se determinó que esta región también corresponde al gen PEX1. Este gen codifica para una proteína llamada factor de biogénesis peroxisomal 1 (https://ghr.nlm.nih.gov/gene/PEX1). Esta proteína cumple funciones biológicas muy importantes dado a que es indispensable para la formación y funcionamiento normal de los peroxisomas, los cuales actúan degradando ácidos grasos y ciertas sustancias tóxicas para el organismo y se ha encontrado esta proteína en testículo. Dadas las funciones biológicas tan importantes que cumplen los genes ERW-1 y PEX-1 es esperado que su expresión sea común y se haya encontrado en cinco de las librerías de humano analizadas. Hubiera sido valioso determinar si estas cinco librerías en las que se encontró expresión del elemento tienen en común alguna característica en particular como la edad del individuo. Sin embargo, no se cuenta con información detallada sobre la edad de los individuos de cada una de las librerías.

En este estudio se observó que un 32,5% de los ERVs identificados presentó traslape con lncRNAs. Lo cual es esperado dado a que se ha visto que los ERVs pueden formar parte de ARNs regulatorios como los lncRNAs (Young et al., 2013). Existen ejemplos de lncRNAs derivados de retrovirus endógenos, como el caso de *human pluripotency-associated transcripts* (HPAT), que son una familia de lncRNAs derivados de ERVs de primates (Glinsky et al., 2018). Sin embargo en este estudio no se encontró expresión de este lncRNA. Estudios previos han encontrado que cerca del 80% de los lncRNA contienen retroelementos, de los cuales la mayoría son HERVs insertados en el sitio de inicio de la transcripción de lncRNAs, esto sugiere que los HERVs son responsables de la regulación transcripcional de los lncRNAs (Young et al., 2013). En este estudio se encontró que los elementos HERVK, HERVIP10F, HERVS71, HERVK9I, HERVH48I y HERVH presentaron traslape con más de un lncRNA.

Se encontró traslape de HERVK con el lncRNA TCONS_l2_00017644. Esto concuerda con estudios previos que han reportado una alta expresión de este lncRNA en próstata, testículos y ovarios y que contiene el ORF *gag* de HERVK, sin embargo su función es desconocida (Bhardwaj et al., 2015; Cabili et al., 2011b). No obstante, en este estudio se encontró que el elemento HERVK posee capacidad codificante para el ORF *pol* en un fragmento de cerca de 4400 pb con un 93.7% de similitud. De los otros lncRNAs encontrados en este estudio no se encontró información acerca de su función o relación con ERVs.

Se encontró que las copias de los HERK expresadas en los cromosomas 7 y 11 presentaron capacidad codificante para la proteína *pol* y *gag*, respectivamente. No obstante, ninguno de estos dos eERVs presentó traslape con genes, lncRNAs ni otros LTRs. Esto genera la incertidumbre sobre la función que están cumpliendo estos elementos. Existe la posibilidad que se estén expresando sólo por la baja metilación de los promotores que existe a nivel testicular o que estén cumpliendo una función biológica importante. En el caso de la copia del cromosoma 11 probablemente desempeñe función biológica importante, ya que como se discutirá más adelante se encontró en una región sinténica con gorila y orangután. No obstante, estudios sugieren que todos los elementos de HERVK se relacionan con diversas patologías como el cáncer, enfermedades autoinmunes y neurodegenerativas. Esto debido a que los HERVKs son los únicos ERVs que poseen inserción polimórfica, es decir, que no todos los individuos poseen a este elemento en el mismo locus. Se han encontrado diferencias en locus y en número de copias de este elemento entre individuos de la misma especie (Li et al., 2019).

8.7 Características genómicas de loci cercanos a ERVs expresados

La regulación de la transcripción de retrovirus endógenos es algo que aún no se comprende bien. Por este motivo, se decidió incluir en este estudio diferentes características genómicas que han sido reportadas en estudios previos como factores que pueden influir en la inserción/fijación de retrovirus endógenos (Campos-Sánchez et al., 2016b).

Al evaluar la proporción de las diferentes características genómicas en los flancos de ERVs, se encontró una alta proporción de lncRNAs cerca de los ERVs expresados, esto se explica por la estrecha relación que existe entre ERVs y lncRNAs que se discutió anteriormente. De hecho se observa que la alta proporción de lncRNAs ocurre únicamente cerca del sitio en donde se encuentran los ERVs y no a lo largo de todo el flanco. Esto apoya la premisa de que algunos ERVs forman parte de los lncRNAs. Necsulea y colaboradores (2014) encontraron que los lncRNAs tienen bajos niveles de transcripción, pero particularmente en testículo humano se encuentran altos niveles de expresión tanto de lncRNAs jóvenes como de otros más ancestrales. Lo cual sugieren que apoya la hipótesis de que la cromatina permisiva de los testículos favorece el origen de nuevos genes, dado a que se ha visto selección purificante en estos elementos que ha hecho que una proporción de lncRNAs hayan adquirido funciones novedosas (Necsulea et al., 2014). Aunque se desconoce la función de la mayor parte de lncRNAs, se sabe que algunos están involucrados en procesos fundamentales como espermatogénesis, desarrollo de placenta, diferenciación celular, entre otros (Necsulea et al., 2014).

Otra de las características genómicas que se evaluó fue el tiempo de replicación. Estudios previos habían determinado que el tiempo de replicación está relacionado con la preferencia de integración y fijación de ETs en el genoma (Campos-Sánchez et al., 2016a). Sin embargo, según los resultados obtenidos en este estudio parece que también se relacionan con la expresión de ERVs. Dado a que en este caso se encontró que los flancos de ERVs expresados mostraron promedios de tiempo de replicación más elevados que los observados en los flancos de LTRs no expresados. El tiempo de replicación es medido como el log2 de poblaciones de células en fase S temprana/tardía en cultivo, por lo tanto valores altos de tiempo de replicación indican que estas regiones se replican más temprano (Ryba et al., 2010). Según los resultados que se obtuvieron, los flancos de ERVs expresados se replican más temprano en comparación con flancos de LTRs no expresados. En general, la replicación temprana es característica de regiones ricas en GC y regiones con mayor densidad de genes expresados y que por el contrario replicación tardía está asociada

con genes silenciados, no obstante existen algunas excepciones (Chakalova et al., 2005).

Se encontró que en los flancos ubicados corriente abajo del eERV (en el extremo 3') existe una mayor proporción de orígenes de replicación y de SINEs. Estudios previos han encontrado que elementos presentes en las regiones flanqueantes ubicadas corriente abajo de los sitios de inserción de ERVs, tienen la capacidad de modular la expresión de estos (Baust et al., 2001). Esto apoya los resultados obtenidos en este estudio en donde se observó que los orígenes de replicación y los SINES ubicados corriente debajo de los ERVs pueden estar influyendo positivamente en la expresión de estos. No obstante estudios previos han reportado que los orígenes de replicación se encuentran subrepresentados en los flancos de HERVKs fijados en el genoma (Campos-Sánchez et al., 2016). Los orígenes de replicación se han visto asociados a regiones del genoma en donde hay abundancia de regiones ricas en GC y a sitios en donde hay elementos que regulan la expresión génica (Cadoret et al., 2008). Lo cual puede ser una razón por la que se encontraron sobrexpresados en flancos de eERVs, dado a que los LTRs de estos podrían estar modulando la expresión de genes cercanos.

En este estudio no se observaron diferencias entre la proporción de genes en los flancos de ERVs expresados y en los flancos de LTRs no expresados. Sin embargo estudios previos han encontrado que pese a que los ERVs se encuentran distribuidos a través de todo el genoma, tienden a ser más abundantes en regiones con pocos genes, de manera tal que no interfieran con la transcripción (Brady et al., 2009). Por otra parte, se ha visto que en integraciones de *novo* sucede lo contrario, los ERVs tienden a insertarse preferiblemente en regiones ricas en genes, pero estos sufren un proceso de purificación y se conservan solo los que no están dentro de unidades de transcripción, esto explica la tendencia en el sitio de inserción observada en los ERVs ya fijados en el genoma (Brady et al., 2009). Una minoría de ERVs se han observado dentro de regiones ricas en genes, sin embargo, estos elementos usualmente se encuentran orientados en dirección opuesta a los genes del hospedero (Medstrand et al., 2002; Smit, 1999; van de Lagemaat et al., 2006).

Dentro de este estudio se incluyó la sensibilidad a la ADNasa I como una de las características genómicas evaluadas. Dado a que las regiones con cromatina abierta son sensibles a ADNasa I, hacer esta evaluación permitiría por ende determinar la proporción de regiones en los flancos que presentan cromatina abierta. Estudios previos han sugerido que el estado de la cromatina contribuye a la regulación de la expresión de ERVs (Lavie et al., 2005). No obstante en este caso, no se observó diferencia en la proporción de regiones sensibles a ADNasa I de los flancos de los ERVs expresados y de los flancos de LTRs no expresados, por lo cual en este caso parece no estar jugando un papel clave en la regulación de la expresión de los ERVs.

En el caso de las islas CpG, no se logró ajustar el valor de p por lo cual no fue posible comparar estadísticamente ambas curvas (la curva de los flancos de ERVs expresados y la curva de los flancos de los LTRs no expresados). Sin embargo estudios previos han reportado que las islas CpG son más abundantes en regiones ricas en genes (Kazanets et al., 2016). Dado que este caso no se encontraron diferencias entre las proporciones de genes presentes en flancos de ERVs y flancos de LTRs no expresados, podría pensarse que es probable que tampoco existan diferencias en la proporción de islas CpG. Sin embargo, se ha visto que regulaciones epigenéticas en islas CpG juegan un papel fundamental en la regulación de la expresión de retrovirus endógenos (Lavie et al., 2005). Por lo cual en futuros estudios es recomendable analizar bases de datos de metilación en islas CpG.

De forma similar, mediante el análisis de FDA tampoco fue posible ajustar los valores de *p* para el predictor *Hot Spots* de recombinación, por los tanto no fue posible determinar si existían diferencias entre las curvas de los flancos de ERVs expresados y flancos de LTRs no expresados. La imposibilidad de ajustar los valores de *p* se debe a la gran cantidad de ceros en ambos sets de datos. No obstante, estudios previos han determinado que en *Hot Spots* de recombinación hay una alta cantidad de ERVs humanos (Kent et al., 2017). Esto debido a que los ERVs pueden experimentar recombinación homóloga entre LTRs flanqueantes en 5' y 3' (Hughes & Coffin, 2004), y pueden también sufrir recombinación ectópica entre copias de un mismo ERVs insertadas en cromosomas no homólogos. Debido a esta capacidad de

recombinación de los elementos transponibles en general, es que se piensa que han contribuido en gran medida a la diversidad genética, ya que tienen el potencial de cambiar la estructura y función del genoma (Lower et al., 1996).

Con respecto a las demás características genómicas evaluadas, no se encontraron diferencias entre la proporción de regiones conservadas y la tasa de recombinación de flancos de ERVs expresados y flancos de LTRs no expresados. Esto puede deberse a que en general se ha visto que los ETs se han fijado en el genoma en regiones poco conservadas y con tasas de recombinación bajas (Campos-Sánchez et al., 2016; Kent et al., 2017). Por lo que se esperaría que las curvas de flancos de eERVs y controles tengan valores muy similares de estas características genómicas. Tampoco se encontraron diferencias entre la proporción de LINEs y ADN transposones de flancos de ERVs expresados y de LTRs no expresados.

La expresión de los ERVs es un mecanismo complejo que puede ser influenciado por diferentes características genómicas simultáneamente. Se ha visto que la expresión de ERVs es susceptible a estímulos provenientes de la célula o tejido o factores externos como por ejemplo citoquinas, esteroides y ciertos químicos (Taruscio & Mantovani, 2004). Además, pueden existir muchos otros factores que estén influyendo en la expresión de los retrovirus endógenos en tejido testicular. Por ejemplo se ha visto que la posición de un retrovirus endógeno dentro del genoma es un factor determinante en la expresión génica, debido a que dependiendo del locus en el que se encuentre va a ser tratado de manera diferente por la maquinaria de metilación de ADN de la célula (Reiss et al., 2007).

También pueden afectar factores metodológicos del FDA. Por ejemplo, para poder observar diferencias entre las regiones comparadas es importante considerar el tamaño de la ventana seleccionada para realizar las mediciones de las características genómicas. Si se elige un tamaño de ventana muy grande, se estaría viendo el efecto de regiones aledañas al elemento, aunque esto podría impedir la detección de diferencias localizadas ya que los valores de las características genómicas evaluadas se acumulan y se promedian a través de la ventana (Cremona

et al., 2018). Ventanas más pequeñas permiten determinar de manera más detallada la señal, pero no contempla el efecto en regiones cercanas (Cremona et al., 2018). Por este motivo se debe de ser muy cuidadoso al momento de seleccionar la ventana y esta selección debe ir acorde a la evaluación que se quiere realizar.

En general se ha encontrado que los retrovirus endógenos en humanos se expresan de manera específica al tejido en el cual se encuentran. Estudios en los que se ha investigado la expresión de ERVs en humanos, han encontrado que la mayor expresión de estos ocurre en tejidos de órganos reproductivos y placenta (Seifarth et al., 2005). Tal como en este caso en donde se lograron determinar un total de 41 tipos de ERVs en tejido testicular de humano, mucho más de los reportados en otro tipo de tejidos. Aunque en otros tejidos la expresión es más baja, no se han encontrado tejidos sin expresión de ERVs (Seifarth et al., 2005).

8.8 Comparación evolutiva de familias de retrovirus en especies de primates

En este estudio se buscó determinar si existía homología en humano, gorila y orangután en las regiones genómicas en donde se encontraron eERVs, esto como un medio para generar información que elucide aspectos relevantes de la historia evolutiva de los primates. Se encontró que de los eERVs en humano se encontraron 58.5% regiones homólogas en gorila y 36.5% en orangután. Estos resultados concuerdan con lo esperado dado a que genéticamente el humano es más similar al gorila, que con el genoma del orangután (Perelman et al., 2011).

Se encontró que solamente diez (24%) de los eERVs identificados en humano eran compartidos entre las tres especies de primates. Aunque, de estos elementos compartidos entre las tres especies de primates, solamente dos copias del elemento HERVK, una en el cromosoma 1 y otra en el 11 tienen capacidad codificante.

La homología del elemento HERVK observada en las tres especies, coincide con lo esperado ya que la familia HERVK fue la que se insertó más recientemente en el genoma de los primates del viejo mundo hace aproximadamente 32-44 millones de años, después de la separación de los primates del nuevo mundo. Es la única familia que en humanos se continuó replicando hasta hace aproximadamente 100 000 años

(Escalera-Zamudio & Greenwood, 2016; Marchi et al., 2013). En humanos se encuentran funcionalmente inactivos, pero los gorilas aún poseen virus con capacidad infectiva, ya que poseen secuencias completas de todos los genes, esto debido a que la infección en gorilas fue más reciente que la infección en humanos (Holloway et al., 2019). Esta es una familia muy estudiada dado a que se ha visto que se encuentra sobreexpresada en ciertas patologías y se estudia la posibilidad de utilizarla como un posible blanco terapéutico en ciertos tipos de cáncer (Magiorkinis et al., 2015).

Se utilizaron las secuencias de los HERVK codificantes para estimar la divergencia evolutiva de las tres especies y se determinó que la mayor divergencia evolutiva se encontró entre humano y orangután (0.10). Lo cual significa que entre estas dos especies ocurrió un mayor número de sustituciones nucleotídicas en esta región del genoma. Lo cual concuerda con las diferencias genéticas que presentan actualmente los genomas de ambas especies y que se conoce que orangután y humano no son genéticamente tan similares entre sí como humano y gorila (Perelman et al., 2011). No obstante, estudios han determinado que humano y orangután comparten la particularidad de que no poseen nuevas familias de ERVs en su genoma, a diferencia de otras especies de primates como chimpancé y gorila (Magiorkinis et al., 2015).

Se determinó también que esta región homóloga en donde se ubica el elemento HERVK, experimentó una selección purificadora, lo cual se determinó al comparar las secuencias de humano y orangután. Esto indica que hay una mayor cantidad de mutaciones sinónimas que de mutaciones no sinónimas entre ambas especies. En el caso de humano y gorila no se encontraron diferencias entre la cantidad de mutaciones sinónimas y no sinónimas. Estudios previos sugieren que la selección purificadora tiene como fin reducir las inserciones de TEs del genoma (Hollister & Gaut, 2009).

También se encontró homología en la posición en la que se ubican los elementos HERV9 (Chr1 y 13), HERVFH19I (Chr2), HERV1 (Chr10), HERVE (Chr13 y X) y PRIMA4 (Chr18), pero ninguno de estos presentó capacidad codificante en

humano. En el caso del elemento HERVE, se había determinado anteriormente su traslape con un lncRNA, y dado a que se ha visto que cerca de un 80% de los lncRNAs contienen elementos transponibles que conforman alrededor del 40 % de su secuencia, sugiere que los ERVs pueden haber sido exaptados y estar cumpliendo un rol en la regulación de la transcripción de estos (Kelley & Rinn, 2012). Esto explica la homología observada en las tres especies de primates en la posición en la que se ubica este elemento. Los elementos HERV9 (Chr1) y PRIMA4 previamente se había encontrado su traslape con los genes WARS2-AS2 y TXNL4A, respectivamente. Lo cual explica que se haya observado homología en las posiciones de estos elementos. Sin embargo, otros como HERVE (Chr 13) y HERV1 no se encontró que tuvieran traslape con algún gen o lncRNA que pudiera explicar la homología en esta región.

De acuerdo con los resultados obtenidos parece que el humano y el gorila tienen una mayor similitud en cuanto a los ERVs de tamaños superiores a 3 kb, ya que cerca del 58.5% de los ERVs identificados en humano son compartidos con gorila, lo cual sugiere que estos elementos o se insertaron recientemente en el genoma que aún conservan grandes fragmentos con capacidad codificante o bien que sus genes fueron exaptados y adquirieron nuevas funciones en estas especies.

Existen muchos factores que hacen complejo elucidar la evolución de los ERVs, por ejemplo, el que pudieron originarse por duplicación a partir de un locus o por inserciones en diferentes momentos de la historia, también la posibilidad de que pudieron experimentar mutaciones o recombinaciones, los cuales representan factores de confusión en el estudio de su evolución (Belshaw et al., 2005; Henzy et al., 2014). Por lo cual es recomendable en estudios posteriores incluir otras especies de primates para poder tener un panorama más completo de cómo ocurrió la evolución de los ERVs en estas especies. El hecho de considerar únicamente ERVs de más de 3000 pb genera un sesgo porque puede haber muchos otros ERVs insertos parcialmente en el genoma. Pero al elegir ERVs de longitud mayor a 3 kb se estaría beneficiando a aquellos elementos que se insertaron recientemente en el genoma y que aún no han sufrido tantas mutaciones, recombinación y pérdida de sus genes. Esta puede ser la causa por la que casi no se observan eERVs en gorila.

Adicionalmente, existen una serie de factores metodológicos que pueden afectar los resultados de un estudio de este tipo. Entre estos se puede mencionar el tejido del cual se tomó la muestra, la edad o estado de desarrollo del organismo, estado de salud, manipulación y preservación de la muestra, integridad del ARN (RIN), preparación de librerías y secuenciación. Además de las diferentes etapas del análisis bioinformático a través de las cuales los parámetros elegidos en cada una determinarán los resultados que se obtengan al final del estudio.

Para este tipo de análisis bioinformático, es muy importante también contar con la metadata asociada a cada una de las muestras que se están utilizando. Dado que no contar con la metadata completa limita las conclusiones del análisis. Por ejemplo, en este estudio no se lograron asociar ciertas diferencias de los eERVs a características de las librerías como edad del individuo, porque dicha información no está disponible en la base de datos ni en las publicaciones de las cuales se obtuvieron las librerías. Esto limita las conclusiones que se puedan generar.

Otro factor importante para considerar es que algunas características genómicas funcionales son determinadas por los tipos de células. En este estudio se utilizaron algunas bases de datos de características genómicas que fueron obtenidas a partir de líneas celulares embrionarias dado a que estas eran lo más similar a células reproductivas que se encontraban disponibles en bases de datos públicas. Por lo cual existe la posibilidad de que no sean representativas del contexto genómico de células reproductivas.

9. CONCLUSIONES

Al comparar los resultados obtenidos con el ensamblaje *novo* y guiado por genoma, se encontraron diferencias considerables en la cantidad y tipos de transcritos ensamblados para una misma librería. Se observó también gran variabilidad en la cantidad y tipos de ERVs identificados entre diferentes librerías humanas. En estas librerías se encontró expresión de un total de 19 tipos de ERVs distintos, pero no se encontró ninguno que fuera codificante en su totalidad. En orangután solo se encontraron siete tipos de eERVs, mientras que en gorila no se

encontraron elementos de más de 3000pb. Lo cual probablemente se deba a la baja profundidad a la que se secuenció esta librería.

Se encontró que más de la mitad (58.4%) de los ERVs expresados presentó traslape con genes y una tercera parte con lncRNAs. Además, se observó que los orígenes de replicación, SINEs y el promedio de replicación de los flancos influyen en la expresión de los ERVs.

Un 58,5% y un 36,5% de los eERVs encontrados en humanos presentaron homología con ERVs de gorila y orangután, respectivamente. De los cuales un 24% eran homólogos en las tres especies, pero solo dos copias del elemento HERVK presentaron capacidad codificante. Con el análisis de este elemento codificante (HERVK) homólogo en las tres especies, se determinó que en esta región ocurrió mayor divergencia entre humano y orangután debido a selección purificadora en este locus.

Dado que las especies de primates utilizadas en este estudio son consideradas primates del viejo mundo, los resultados de esta investigación no pueden ser generalizados a las especies del nuevo mundo. Sin embargo, la metodología de análisis bioinformático estandarizada en este trabajo puede ser utilizada para analizar cualquier especie de primate e incluso otras especies.

Puede haber muchos factores como presencia de contaminación, integridad del ARN, diferencia en la cobertura de las librerías, edad de los individuos, entre otros que influyan en los resultados que se obtienen en un estudio de este tipo. Todas las etapas son cruciales, desde la selección de la muestra (tipo de tejido, forma de colecta y preservación del tejido, momento del desarrollo, entre otros), preparación de las librerías, preprocesamiento de los datos, ensamblaje, herramientas elegidas para análisis bioinformático, tipo de ensamblaje, cobertura, nivel de expresión de cada gen, análisis realizados, entre muchos otros factores que determinan los resultados que se obtengan. Por esto es recomendable tener varias réplicas de las muestras empleadas en este tipo de estudios.

Los resultados obtenidos en esta investigación aportan información valiosa acerca de factores genéticos que influyen en la expresión de ERVs en humanos,

genética evolutiva y función de los ERVs en tres especies de primates. Dicha información podría ser utilizada en estudios posteriores que tengan como fin estudiar cómo ocurrió la separación de los primates y estudios relacionados con la expresión de ERVs.

10. RECOMENDACIONES

- Antes de iniciar el análisis utilizando librerías disponibles en bases de datos públicas, asegurarse de que estás cuenten con la metadata respectiva. De manera tal que una vez finalizado el análisis, sea posible correlacionar esta metadata con los resultados obtenidos.
- Se recomienda también realizar un análisis de la contaminación que pudiera estar presente en las librerías que se van a utilizar en determinado estudio.
- Dado que se observaron diferencias entre los transcritos ensamblados en librerías ensambladas de *novo* y las ensambladas utilizando un genoma de referencia, se recomienda para futuros estudios de este tipo utilizar siempre ambos métodos de ensamblaje. Debido a que cada método tiene sus propias ventajas y desventajas, por lo tanto complementar el análisis con ambos enfoques permitirá obtener resultados más confiables.
- Para estudios posteriores se recomienda realizar un paso de "scaffolding" posterior al ensamblaje, con el objetivo de mejorar el ensamblaje de los transcritos y lograr obtener eERV de la mayor longitud posible.
- Extender este tipo de análisis a otras especies de primates, incluidas especies de primates del nuevo mundo. Y tener varias réplicas por especie cuando sea posible.
- Siempre realizar una estandarización del script de análisis.
- En futuros estudios se podrían incorporar otras características genómicas adicionales a las que utilizamos en este caso como por ejemplo: porcentaje de GC, nivel de metilación, localización en el cromosoma, composición nucleotídica, etc.

- Se podría hacer un análisis de expresión génica en tejido testicular para determinar el nivel de expresión de los ERVs ensamblados. Lo cual podría compararse con otros estudios de expresión diferencial de ERVs en diferentes tejidos y en diferentes tipos de células. Esto podría ser de gran utilidad para entender cambios en patrones de expresión asociados a diferentes patologías humanas.
- Para estudios futuros se recomienda utilizar las versiones más recientes de los genomas de referencia, dado a que estás se espera que sean las más completas y permitan realizar una mejor anotación y mapeo de los transcritos ensamblados.

11. LITERATURA CITADA

Afgan, E., Baker, D., van den Beek, M., Blankenberg, D., Bouvier, D., Čech, M., Chilton, J., Clements, D., Coraor, N., Eberhard, C., Grüning, B., Guerler, A., Hillman-Jackson, J., Von Kuster, G., Rasche, E., Soranzo, N., Turaga, N., Taylor, J., Nekrutenko, A., & Goecks, J. (2016). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Research*, *44*(1), 3-10.

Akopov, S. B., Nikolaev, L. G., Khil, P. P., Lebedev, Y. B., & Sverdlov, E. D. (1998). Long terminal repeats of human endogenous retrovirus K family (HERV-K) specifically bind host cell nuclear proteins. *FEBS Letters*, *421*(3), 229-233.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology*, *215*(3), 403-410.

Andrews, S. (2010). *FastQC: a quality control tool for high throughput sequence data*.

Bannert, N., & Kurth, R. (2006). The evolutionary dynamics of human endogenous retroviral families. *Annual Review Genomics and Human Genetics*, *7*, 149-173.

Bao, W., Kojima, K. K., & Kohany, O. (2015). Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA*, *6*(1), 11.

Barbulescu, M., Turner, G., Seaman, M. I., Deinard, A. S., Kidd, K. K., & Lenz, J. (1999). Many human endogenous retrovirus K (HERV-K) proviruses are unique to humans. *Current Biology*, *9*(16), 861-S1.

Barbulescu, M., Turner, G., Su, M., Kim, R., Jensen-Seaman, M. I., Deinard, A. S., Kidd, K. K., & Lenz, J. (2001). A HERV-K provirus in chimpanzees, bonobos and gorillas, but not humans. *Current Biology*, *11*(10), 779-783.

Baust, C., Seifarth, W., Schön, U., Hehlmann, R., & Leib-Mösch, C. (2001). Functional Activity of HERV-K-T47D-Related Long Terminal Repeats. *Virology*, *283*(2), 262-272.

Belshaw, R., Dawson, A. L. A., Woolven-Allen, J., Redding, J., Burt, A., & Tristem, M. (2005). Genomewide screening reveals high levels of insertional polymorphism in the human endogenous retrovirus family HERV-K(HML2): Implications for present-day activity. *Journal of Virology*, *79*(19), 12507-12514.

Belshaw, R., Pereira, V., Katzourakis, A., Talbot, G., Paces, J., Burt, A., & Tristem, M. (2004). Long-term reinfection of the human genome by endogenous retroviruses. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(14), 4894-4899.

Best, S., Le Tissier, P. R., & Stoye, J. P. (1997). Endogenous retroviruses and the evolution of resistance to retroviral infection. *Trends in Microbiology*, *5*(8), 313-318.

Bhardwaj, N., Montesion, M., Roy, F., & Coffin, J. (2015). Differential Expression of HERV-K (HML-2) Proviruses in Cells and Virions of the Teratocarcinoma Cell Line Tera-1. *Viruses*, *7*(3), 939-968.

Bieda, K, Hoffmann, A., & Boller, K. (2001). Phenotypic heterogeneity of human endogenous retrovirus particles produced by teratocarcinoma cell lines. *The Journal of general virology*, *82*(3), 591-596.

Bieda, Katrin, Hoffmann, A., & Boller, K. (2001). Phenotypic heterogeneity of human endogenous retrovirus particles produced by teratocarcinoma cell lines. *Journal of General Virology*, *82*(3), 591-596.

Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D., & Pirovano, W. (2011). Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics*, *27*(4), 578-579.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114-2120.

Brady, T., Lee, Y. N., Ronen, K., Malani, N., Berry, C. C., Bieniasz, P. D., & Bushman, F. D. (2009). Integration target site selection by a resurrected human endogenous retrovirus. *Genes & Development*, *23*(5), 633-642.

Cabili, M. N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., & Rinn, J. L. (2011). Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes & Development*, *25*(18), 1915-1927.

Cadoret, J. C., Meisch, F., Hassan-Zadeh, V., Luyten, I., Guillet, C., Duret, L., ... & Prioleau, M. N. (2008). Genome-wide studies highlight indirect links between human replication origins and gene regulation. *Proceedings of the National Academy of Sciences*, *105*(41), 15837-15842.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: Architecture and applications. *BMC Bioinformatics*, *10*(1), 421.

Campos-Sánchez, R., Cremona, M. A., Pini, A., Chiaromonte, F., & Makova, K. D. (2016). Integration and Fixation Preferences of Human and Mouse Endogenous Retroviruses Uncovered with Functional Data Analysis. *Plos Computational Biology*, *12*(6), e1004956.

Capy, P. (2005). Classification and nomenclature of retrotransposable elements. *Cytogenetic and Genome Research*, *110*(1-4), 457-461.

Carelli, F. N., Hayakawa, T., Go, Y., Imai, H., Warnefors, M., & Kaessmann, H. (2016). The life history of retrocopies illuminates the evolution of new mammalian genes. Genome Research, 26(3), 301-314.

Chakalova, L., Debrand, E., Mitchell, J. A., Osborne, C. S., & Fraser, P. (2005). Replication and transcription: shaping the landscape of the genome. *Nature Reviews Genetics*, 6(9), 669-677.

Chessa, B., Pereira, F., Arnaud, F., Amorim, A., Goyache, F., Mainland, I., Kao, R. R., Pemberton, J. M., Beraldi, D., Stear, M. J., Alberti, A., Pittau, M., Iannuzzi, L., Banabazi, M. H., Kazwala, R. R., Zhang, Y. P., Arranz, J. J., Ali, B. A., Wang, Z., … Palmarini, M. (2009). Revealing the history of sheep domestication using retrovirus integrations. *Science*, *324*(5926), 532-536.

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M. W., Gaffney, D. J., Elo, L. L., Zhang, X., & Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biology*, *17*(1), 13.

Cowley, M., & Oakey, R. J. (2013). Transposable elements re-wire and fine-tune the transcriptome. *PLoS Genet*, *9*(1), e1003234.

Cremona, M. A., Pini, A., Cumbo, F., Makova, K. D., Chiaromonte, F., & Vantini, S. (2018). IWTomics: Testing high-resolution sequence-based 'Omics' data at multiple locations and scales. *Bioinformatics*, *34*(13), 2289-2291.

Crowell, R. C., & Kiessling, A. A. (2007). Endogenous retrovirus expression in testis and epididymis. *Biochemical Society transactions*, *35*(3), 629-633.

de Parseval, N., & Heidmann, T. (2005). Human endogenous retroviruses: From infectious elements to human genes. *Cytogenetic and Genome Research*, *110*(1-4), 318-332.

de Parseval, N., Lazar, V., Casella, J. F., Benit, L., & Heidmann, T. (2003). Survey of Human Genes of Retroviral Origin: Identification and Transcriptome of the Genes with Coding Capacity for Complete Envelope Proteins. *Journal of Virology*, *77*(19), 10414-10422.

Elliott, D., & Ladomery, M. (2011). *Molecular Biology of RNA*. OUP Oxford. https://books.google.co.cr/books?id=iUicAQAAQBAJ

Emera, D., & Wagner, G. P. (2012). Transposable element recruitments in the mammalian placenta: Impacts and mechanisms. *Briefings in functional genomics*, *11*(4), 267-276.

Escalera-Zamudio, M., & Greenwood, A. D. (2016). On the classification and evolution of endogenous retrovirus: Human endogenous retroviruses may not be 'human' after all. *APMIS*, *124*(1-2), 44-51.

Fagerberg, L, Hallstrom, B. M., Oksvold, P., Kampf, C., Djureinovic, D., Odeberg, J., Habuka, M., Tahmasebpoor, S., Danielsson, A., Edlund, K., Asplund, A., Sjostedt, E., Lundberg, E., Szigyarto, C. A. K., Skogs, M., Takanen, J. O., Berling, H., Tegel, H., Mulder, J., … Uhlen, M. (2014). Analysis of the Human Tissue-specific Expression by Genome-wide Integration of Transcriptomics and Antibody-based Proteomics. *Molecular & Cellular Proteomics*, *13*(2), 397-406.

Flockerzi, A., Ruggieri, A., Frank, O., Sauter, M., Maldener, E., Kopper, B., Wullich, B., Seifarth, W., Müller-Lantzsch, N., Leib-Mösch, C., Meese, E., & Mayer, J. (2008). Expression patterns of transcribed human endogenous retrovirus HERV-K(HML-2) loci in human tissues and the need for a HERV Transcriptome Project. *BMC Genomics*, *9*(1), 354.

Freeman, W. M., Walker, S. J., & Vrana, K. E. (1999). Quantitative RT-PCR: Pitfalls and Potential. *BioTechniques*, *26*(1), 112-125.

Frendo, J.-L., Olivier, D., Cheynet, V., Blond, J.-L., Bouton, O., Vidaud, M., Rabreau, M., Evain-Brion, D., & Mallet, F. (2003). Direct Involvement of HERV-W Env Glycoprotein in Human Trophoblast Cell Fusion and Differentiation. *Molecular and Cellular Biology*, *23*(10), 3566-3574.

Garcia-Montojo, M., Doucet-O'Hare, T., Henderson, L., & Nath, A. (2018). Human endogenous retrovirus-K (HML-2): A comprehensive review. *Critical Reviews in Microbiology*, *44*(6), 715-738.

Gardiner-Garden, M., & Frommer, M. (1987). CpG Islands in vertebrate genomes. *Journal of Molecular Biology*, *196*(2), 261-282.

Gemmell, P., Hein, J., & Katzourakis, A. (2019). The Exaptation of HERV-H: Evolutionary Analyses Reveal the Genomic Features of Highly Transcribed Elements. *Frontiers in Immunology*, *10*, 1339.

Giordano, J., Ge, Y., Gelfand, Y., Abrusán, G., Benson, G., & Warburton, P. E. (2007). Evolutionary History of Mammalian Transposons Determined by Genome-Wide Defragmentation. *PLoS Computational Biology*, *3*(7), e137.

Glinsky, G., Durruthy-Durruthy, J., Wossidlo, M., Grow, E. J., Weirather, J. L., Au, K. F., Wysocka, J., & Sebastiano, V. (2018). Single cell expression analysis of primate-specific retroviruses-derived HPAT lincRNAs in viable human blastocysts identifies embryonic cells co-expressing genetic markers of multiple lineages. *Heliyon*, *4*(6), e00667.

Glinsky, G. V. (2015). Viruses, stemness, embryogenesis, and cancer: A miracle leap toward molecular definition of novel oncotargets for therapy-resistant malignant tumors? *Oncoscience*, *2*(9), 751.

Goff, S. P. (2004). Retrovirus Restriction Factors. *Molecular Cell*, *16*(6), 849-859.

Göke, J., Lu, X., Chan, Y.-S., Ng, H.-H., Ly, L.-H., Sachs, F., & Szczerbinska, I. (2015). Dynamic Transcription of Distinct Classes of Endogenous Retroviral Elements Marks Specific Populations of Early Human Embryonic Cells. *Cell Stem Cell*, *16*(2), 135-141.

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., & Zeng, Q. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature biotechnology*, *29*(7), 644.

Griffith, M., Griffith, O. L., Mwenifumbo, J., Goya, R., Morrissy, A. S., Morin, R. D., Corbett, R., Tang, M. J., Hou, Y.-C., & Pugh, T. J. (2010). Alternative expression analysis by RNA sequencing. *Nature methods*, *7*(10), 843-847.

Gröger, V., & Cynis, H. (2018). Human Endogenous Retroviruses and Their Putative Role in the Development of Autoimmune Disorders Such as Multiple Sclerosis. *Frontiers in Microbiology*, *9*, 265.

Harris, J. R. (1998). Placental endogenous retrovirus (ERV): Structural, functional, and evolutionary significance. *Bioessays*, *20*(4), 307-316.

Henzy, J. E., Gifford, R. J., Johnson, W. E., & Coffin, J. M. (2014). A Novel Recombinant Retrovirus in the Genomes of Modern Birds Combines Features of Avian and Mammalian Retroviruses. *Journal of Virology*, *88*(5), 2398-2405.

Hollister, J. D., & Gaut, B. S. (2009). Epigenetic silencing of transposable elements: A trade-off between reduced transposition and deleterious effects on neighboring gene expression. *Genome Research*, *19*(8), 1419-1428.

Holloway, J. R., Williams, Z. H., Freeman, M. M., Bulow, U., & Coffin, J. M. (2019). Gorillas have been infected with the HERV-K (HML-2) endogenous retrovirus much more recently than humans and chimpanzees. *Proceedings of the National Academy of Sciences*, *116*(4), 1337-1346.

Hrdlickova, R., Toloue, M., & Tian, B. (2017). RNA-Seq methods for transcriptome analysis: RNA-Seq. *Wiley Interdisciplinary Reviews: RNA*, *8*(1), e1364.

Huang, C.-J., Lin, W.-Y., Chang, C.-M., & Choo, K.-B. (2009). Transcription of the rat testis-specific Rtdpoz-T1 and -T2 retrogenes during embryo development: Co-transcription and frequent exonisation of transposable element sequences. *BMC Molecular Biology*, *10*(1), 74.

Hughes, J. F., & Coffin, J. M. (2004). Human endogenous retrovirus K solo-LTR formation and insertional polymorphisms: implications for human and viral evolution. *Proceedings of the National Academy of Sciences*, 101(6), 1668-1672.

Jain, P., Krishnan, N. M., & Panda, B. (2013). Augmenting transcriptome assembly by combining *de novo* and genome-guided tools. *PeerJ*, *1*, e133.

Jurka, J., Kapitonov, V. V., Pavlicek, A., Klonowski, P., Kohany, O., & Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic and Genome Research*, *110*(1-4), 462-467.

Katzourakis, A., Pereira, V., & Tristem, M. (2007). Effects of Recombination Rate on Human Endogenous Retrovirus Fixation and Persistence. *Journal of Virology*, *81*(19), 10712-10717.

Kazanets, A., Shorstova, T., Hilmi, K., Marques, M., & Witcher, M. (2016). Epigenetic silencing of tumor suppressor genes: Paradigms, puzzles, and potential. *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer*, *1865*(2), 275-288.

Kelley, D., & Rinn, J. (2012). Transposable elements reveal a stem cell-specific class of long noncoding RNAs. *Genome Biology*, *13*(11), R107.

Kent, T. V., Uzunović, J., & Wright, S. I. (2017). Coevolution between transposable elements and recombination. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1736), 20160458.

Khodosevich, K., Lebedev, Y., & Sverdlov, E. (2002). Endogenous Retroviruses and Human Evolution. *Comparative and Functional Genomics*, *3*(6), 494-498.

Kim, D., Langmead, B., & Salzberg, S. L. (2015). HISAT: A fast spliced aligner with low memory requirements. *Nature Methods*, *12*(4), 357-360.

Kim, T. H., Jeon, Y. J., Yi, J. M., Kim, D. S., Huh, J. W., Hur, C. G., & Kim, H. S. (2004). The distribution and expression of HERV families in the human genome. *Molecules and Cells*, *18*(1), 87-93.

Ko, M. S., Kitchen, J. R., Wang, X., Threat, T. A., Wang, X., Hasegawa, A., Sun, T., Grahovac, M. J., Kargul, G. J., Lim, M. K., Cui, Y., Sano, Y., Tanaka, T., Liang, Y., Mason, S., Paonessa, P. D., Sauls, A. D., DePalma, G. E., Sharara, R., … Doi, H. (2000). Large-scale cDNA analysis reveals phased gene expression patterns during preimplantation mouse development. *Development (Cambridge, England)*, *127*(8), 1737-1749.

Kumar, S., Stecher, G., & Tamura, K. (2016). MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Molecular Biology and Evolution*, *33*(7), 1870-1874.

Laderoute, M. P., Giulivi, A., Larocque, L., Bellfoy, D., Hou, Y., Wu, H.-X., Fowke, K., Wu, J., & Diaz-Mitoma, F. (2007). The replicative activity of human endogenous retrovirus K102 (HERV-K102) with HIV viremia. *Aids*, *21*(18), 2417-2424.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., & FitzHugh, W. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*(6822), 860-921.

Lavie, L., Kitova, M., Maldener, E., Meese, E., & Mayer, J. (2005). CpG Methylation Directly Regulates Transcriptional Activity of the Human Endogenous Retrovirus Family HERV-K(HML-2). *Journal of Virology*, *79*(2), 876-883.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078-2079.

Li, W., Lin, L., Malhotra, R., Yang, L., Acharya, R., & Poss, M. (2019). A computational framework to assess genome-wide distribution of polymorphic human endogenous retrovirus-K In human populations. *PLoS Computational Biology*, *15*(3), e1006564.

Liu, L., Ishihara, K., Ichimura, T., Fujita, N., Hino, S., Tomita, S., Watanabe, S., Saitoh, N., Ito, T., & Nakao, M. (2009). MCAF1/AM Is Involved in Sp1-mediated Maintenance of Cancer-associated Telomerase Activity. *Journal of Biological Chemistry*, *284*(8), 5165-5174.

Lower, R., J. Lower & R. Kurth, (1996). The viruses in all of us: characteristics and biological significance of human endogenous retrovirus sequences. *Proceedings of the National Academy of Sciences*. *93*, 5177–5184.

Lu, B., Zeng, Z., & Shi, T. (2013). Comparative study of de novo assembly and genome-guided assembly strategies for transcriptome reconstruction based on RNA-Seq. *Science China Life Sciences*, *56*(2), 143-155.

Lynch, M., & Conery, J. S. (2003). The origins of genome complexity. *Science*, *302*(5649), 1401-1404.

Ma, J. (2001). Regulation of Zygotic Gene Activation in the Preimplantation Mouse Embryo: Global Activation and Repression of Gene Expression. *Biology of Reproduction*, *64*(6), 1713-1721.

Mager, D. L., & Medstrand, P. (2005). Retroviral repeat sequences. *Encyclopedia of Life Sciences*, *1*, 7.

Magiorkinis, G., Blanco-Melo, D., & Belshaw, R. (2015). The decline of human endogenous retroviruses: Extinction and survival. *Retrovirology*, *12*(1), 8.

Maksakova, I. A., Romanish, M. T., Gagnier, L., Dunn, C. A., de Lagemaat, L. N. V., & Mager, D. L. (2006). Retroviral elements and their hosts: Insertional mutagenesis in the mouse germ line. *PLoS Genetics*, *2*(1), e2.

Mangeney, M., & Heidmann, T. (1998). Tumor cells expressing a retroviral envelope escape immune rejection *in vivo*. *Proceedings of the National Academy of Sciences*, *95*(25), 14920.

Marchi, E., Kanapin, A., Byott, M., Magiorkinis, G., & Belshaw, R. (2013). Neanderthal and Denisovan retroviruses in modern humans. *Current Biology*, *23*(22), R994-5.

Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M., & Gilad, Y. (2008). RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome research*, *18*(9), 1509-1517.

Mayer, J., Blomberg, J., & Seal, R. L. (2011). A revised nomenclature for transcribed human endogenous retroviral loci. *Mobile DNA*, *2*(1), 7.

McGinnis, S., & Madden, T. L. (2004). BLAST: At the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Research*, *32*, 20-25.

Medstrand, P., & Mager, D. L. (1998). Human-specific integrations of the HERV-K endogenous retrovirus family. *Journal of Virology*, *72*(12), 9782-9787.

Medstrand, P., van de Lagemaat, L. N., & Mager, D. L. (2002). Retroelement distributions in the human genome: Variations associated with age and proximity to genes. *Genome Research*, *12*(10), 1483-1495.

Metzker, M. L. (2010). Sequencing technologies—The next generation. *Nature reviews genetics*, *11*(1), 31.

Mi, S., Lee, X., Li, X., Veldman, G. M., Finnerty, H., Racie, L., LaVallie, E., Tang, X.-Y., Edouard, P., Howes, S., Keith, J. C., & McCoy, J. M. (2000). Syncytin is a

captive retroviral envelope protein involved in human placental morphogenesis. *Nature*, *403*(6771), 785-789.

Miga, K. H., Newton, Y., Jain, M., Altemose, N., Willard, H. F., & Kent, W. J. (2014). Centromere reference models for human chromosomes X and Y satellite arrays. *Genome Research*, *24*(4), 697-707.

Moyes, D., Griffiths, D. J., & Venables, P. J. (2007). Insertional polymorphisms: A new lease of life for endogenous retroviruses in human disease. *Trends in Genetics*, *23*(7), 326-333.

Necsulea, A., Soumillon, M., Warnefors, M., Liechti, A., Daish, T., Zeller, U., Baker, J. C., Grützner, F., & Kaessmann, H. (2014). The evolution of lncRNA repertoires and expression patterns in tetrapods. Nature, 505(7485), 635-640.

Okahara, G., Matsubara, S., Oda, T., Sugimoto, J., Jinno, Y., & Kanaya, F. (2004). Expression analyses of human endogenous retroviruses (HERVs): Tissue-specific and developmental stage-dependent expression of HERVs. *Genomics*, *84*(6), 982-990.

Perelman, P., Johnson, W. E., Roos, C., Seuánez, H. N., Horvath, J. E., Moreira, M. A. M., Kessing, B., Pontius, J., Roelke, M., Rumpler, Y., Schneider, M. P. C., Silva, A., O'Brien, S. J., & Pecon-Slattery, J. (2011). A Molecular Phylogeny of Living Primates. *PLoS Genetics*, *7*(3), e1001342.

Pruitt, K. D., Brown, G. R., Hiatt, S. M., Thibaud-Nissen, F., Astashyn, A., Ermolaeva, O., Farrell, C. M., Hart, J., Landrum, M. J., McGarvey, K. M., Murphy, M. R., O'Leary, N. A., Pujar, S., Rajput, B., Rangwala, S. H., Riddick, L. D., Shkeda, A., Sun, H., Tamez, P., … Ostell, J. M. (2014). RefSeq: An update on mammalian reference sequences. *Nucleic Acids Research*, *42*(1), 756-763.

Rebollo, R., Romanish, M. T., & Mager, D. L. (2012). Transposable Elements: An Abundant and Natural Source of Regulatory Sequences for Host Genes. *Annual Review of Genetics*, *46*(1), 21-42.

Reiss, D., Zhang, Y., & Mager, D. L. (2007). Widely variable endogenous retroviral methylation levels in human placenta. *Nucleic Acids Research*, *35*(14), 4743-4754.

Roca, A. L., Pecon-Slattery, J., & O'brien, S. J. (2004). Genomically intact endogenous feline leukemia viruses of recent origin. *Journal of virology*, *78*(8), 4370-4375.

Ruiz-Orera, J., Hernandez-Rodriguez, J., Chiva, C., Sabidó, E., Kondova, I., Bontrop, R., Marques-Bonet, T., & Albà, M. M. (2015). Origins of De Novo Genes in Human and Chimpanzee. *PLoS Genetics*, *11*(12), e1005721.

Ryba, T., Hiratani, I., Lu, J., Itoh, M., Kulik, M., Zhang, J., Schulz, T. C., Robins, A. J., Dalton, S., & Gilbert, D. M. (2010). Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types. *Genome Research*, *20*(6), 761-770.

Ryba, Tyrone, Battaglia, D., Pope, B. D., Hiratani, I., & Gilbert, D. M. (2011). Genome-scale analysis of replication timing: From bench to bioinformatics. *Nature Protocols*, *6*(6), 870-895.

Sabo, P. J., Kuehn, M. S., Thurman, R., Johnson, B. E., Johnson, E. M., Cao, H., Yu, M., Rosenzweig, E., Goldy, J., Haydock, A., Weaver, M., Shafer, A., Lee, K., Neri, F., Humbert, R., Singer, M. A., Richmond, T. A., Dorschner, M. O., McArthur, M., … Stamatoyannopoulos, J. A. (2006). Genome-scale mapping of DNase I sensitivity in vivo using tiling DNA microarrays. *Nature Methods*, *3*(7), 511-518.

Schmidt, E. E. (1996). Transcriptional promiscuity in testes. *Current Biology*, *6*(7), 768-769.

Seifarth, W., Frank, O., Zeilfelder, U., Spiess, B., Greenwood, A. D., Hehlmann, R., & Leib-Mosch, C. (2005). Comprehensive Analysis of Human Endogenous Retrovirus Transcriptional Activity in Human Tissues with a Retrovirus-Specific Microarray. *Journal of Virology*, *79*(1), 341-352.

Smit, A. F. (1999). Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Current Opinion in Genetics & Development*, *9*(6), 657-663.

Smith-Unna, R., Boursnell, C., Patro, R., Hibberd, J. M., & Kelly, S. (2016). TransRate: Reference-free quality assessment of de novo transcriptome assemblies. *Genome Research*, *26*(8), 1134-1144.

Soygur, B., & Sati, L. (2016). The role of syncytins in human reproduction and reproductive organ cancers. *Reproduction*, *152*(5), 167-178.

Stewart, C., Kural, D., Strömberg, M. P., Walker, J. A., Konkel, M. K., Stütz, A. M., Urban, A. E., Grubert, F., Lam, H. Y. K., Lee, W.-P., Busby, M., Indap, A. R., Garrison, E., Huff, C., Xing, J., Snyder, M. P., Jorde, L. B., Batzer, M. A., Korbel, J. O., … 1000 Genomes Project. (2011). A Comprehensive Map of Mobile Element Insertion Polymorphisms in Humans. *PLoS Genetics*, *7*(8), e1002236.

Tarlinton, R. E., Meers, J., & Young, P. R. (2006). Retroviral invasion of the koala genome. *Nature*, *442*(7098), 79-81.

Taruscio, D., & Mantovani, A. (2004). Factors regulating endogenous retroviral sequences in human and mouse. *Cytogenetic and Genome Research*, *105*(2-4), 351-362.

Thomas, J. H., & Schneider, S. (2011). Coevolution of retroelements and tandem zinc finger genes. *Genome Research*, *21*(11), 1800-1812.

Tomás-Loba, A., Flores, I., Fernández-Marcos, P. J., Cayuela, M. L., Maraver, A., Tejera, A., Borrás, C., Matheu, A., Klatt, P., & Flores, J. M. (2008). Telomerase reverse transcriptase delays aging in cancer-resistant mice. *Cell*, *135*(4), 609-622.

Tomaszkiewicz, M., Rangavittal, S., Cechova, M., Sanchez, R. C., Fescemyer, H. W., Harris, R., Ye, D., O'Brien, P. C. M., Chikhi, R., Ryder, O. A., Ferguson-Smith, M. A., Medvedev, P., & Makova, K. D. (2016). A time- and cost-effective strategy to sequence mammalian Y Chromosomes: An application to the de novo assembly of gorilla Y. *Genome Research*, *26*(4), 530-540.

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., Van Baren, M. J., Salzberg, S. L., Wold, B. J., & Pachter, L. (2010). Transcript assembly and

quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology*, *28*(5), 511-515.

van de Lagemaat, L. N., Landry, J.-R., Mager, D. L., & Medstrand, P. (2003). Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions. *Trends in Genetics*, *19*(10), 530-536.

van de Lagemaat, L. N., Medstrand, P., & Mager, D. L. (2006). Multiple effects govern endogenous retrovirus survival patterns in human gene introns. *Genome Biol*, *7*(9), R86.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., & Holt, R. A. (2001). The sequence of the human genome. *science*, *291*(5507), 1304-1351.

Voisset, Cecile, Blancher, A., Perron, H., Mandrand, B., Mallet, F., & Paranhos-Baccala, G. (1999). Phylogeny of a Novel Family of Human Endogenous Retrovirus Sequences, HERV-W, in Humans and Other Primates. *AIDS Research and Human Retroviruses*, *15*(17), 1529-1533.

Voisset, Cécile, Weiss, R. A., & Griffiths, D. J. (2008). Human RNA "rumor" viruses: The search for novel human retroviruses in chronic disease. *Microbiology and Molecular Biology Reviews*, *72*(1), 157-196.

Wang, Zhong, Gerstein, M., & Snyder, M. (2009). RNA-Seq: A revolutionary tool for transcriptomics. *Nature reviews genetics*, *10*(1), 57-63.

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis* (Second edition). Springer.

Young, J. M., Whiddon, J. L., Yao, Z., Kasinathan, B., Snider, L., Geng, L. N., Balog, J., Tawil, R., van der Maarel, S. M., & Tapscott, S. J. (2013). DUX4 Binding to Retroelements Creates Promoters That Are Active in FSHD Muscle and Testis. *PLoS Genetics*, *9*(11), e1003947.

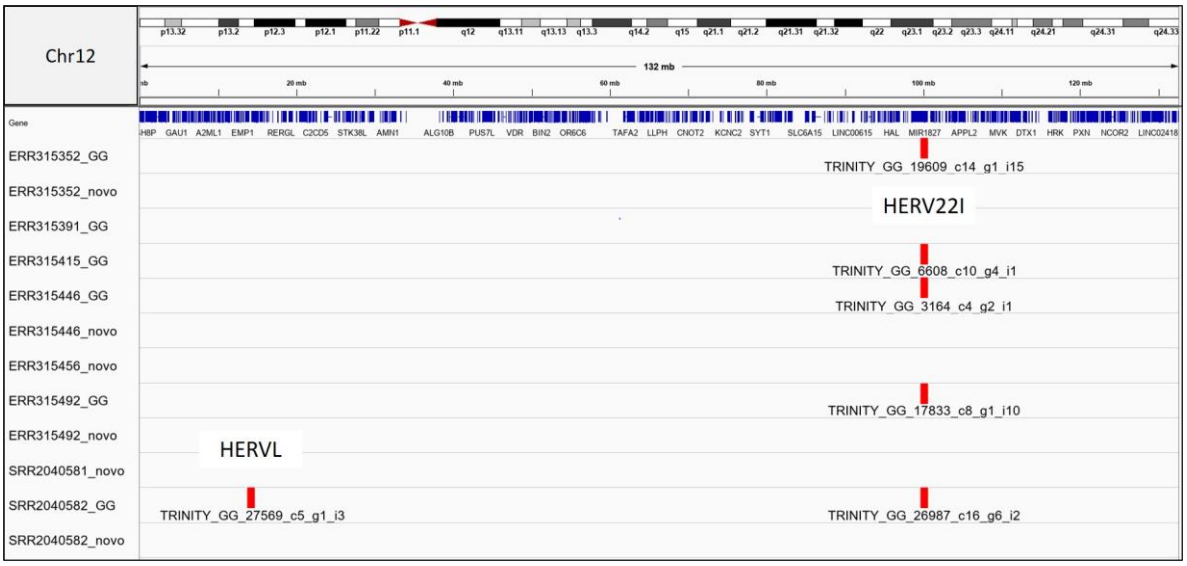
12. ANEXOS

Anexo 1. Comparación de calidad de ensamblajes de *novo* y guiado por genoma realizados para cada una de las librerías

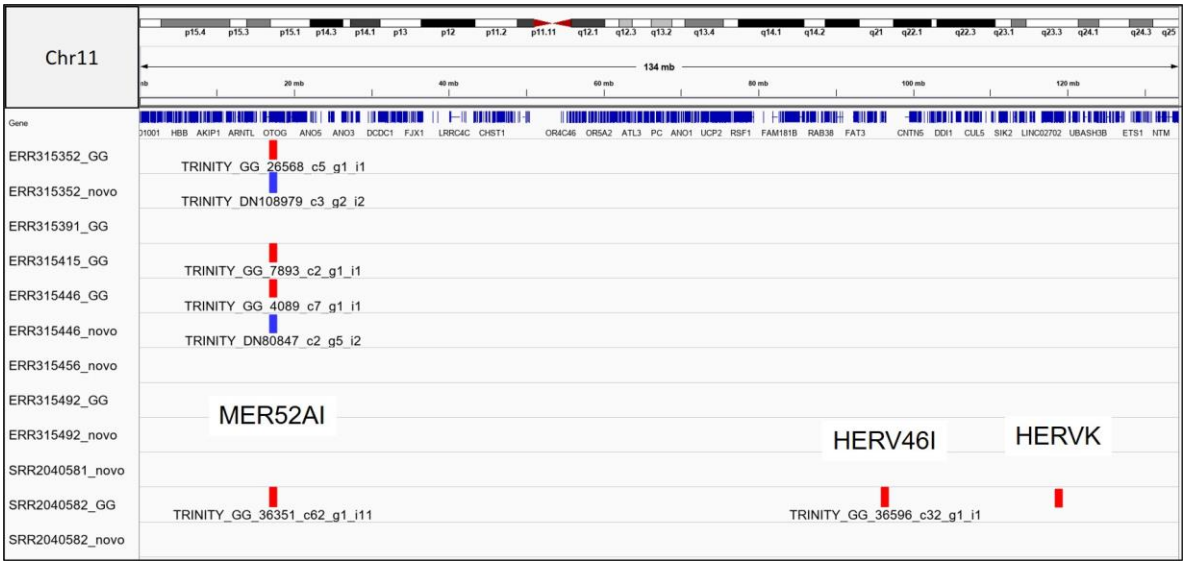
Librería	Especie	Ensamblaje	N secuencias	Sec más larga	Longitud promedio	N50
SRR2040581	Humano	Guiado	166423	12072	750	1363
		De *Novo*	219597	16576	814	1667
SRR2040582	Humano	Guiado	852718	27571	831	1875
		De *Novo*	811894	27566	845	1815
ERR315350	Humano	Guiado	123188	23357	754	1372
		De *Novo*	203478	13706	788	1362
ERR315351	Humano	Guiado	124060	11221	752	1365
		De *Novo*	205370	19806	799	1400
ERR315352	Humano	Guiado	349184	21693	896	2169
		De *Novo*	313933	21693	850	1930
ERR315391	Humano	Guiado	267525	20660	865	1978
		De *Novo*	409681	20264	1057	2197
ERR315415	Humano	Guiado	107186	17727	930	2357
		De *Novo*	571589	21070	1002	2168
ERR315446	Humano	Guiado	63527	21450	917	2278
		De *Novo*	462938	25015	984	2101
ERR315456	Humano	Guiado	24810	17383	959	2486
		De *Novo*	409061	20195	1046	2301
ERR315492	Humano	Guiado	360748	20869	924	2289
		De *Novo*	518593	22319	1009	2207
SRR3053573	Gorila	Guiado	142105	7155	546	687
		De *Novo*	141724	10795	478	563
3405	orangután	Guiado	658742	15281	733	1070
		De *Novo*	1056084	14263	788	1447

Anexo 2. Cantidad y tipos de transcritos reconstruidos a partir de las librerías de humano y orangután

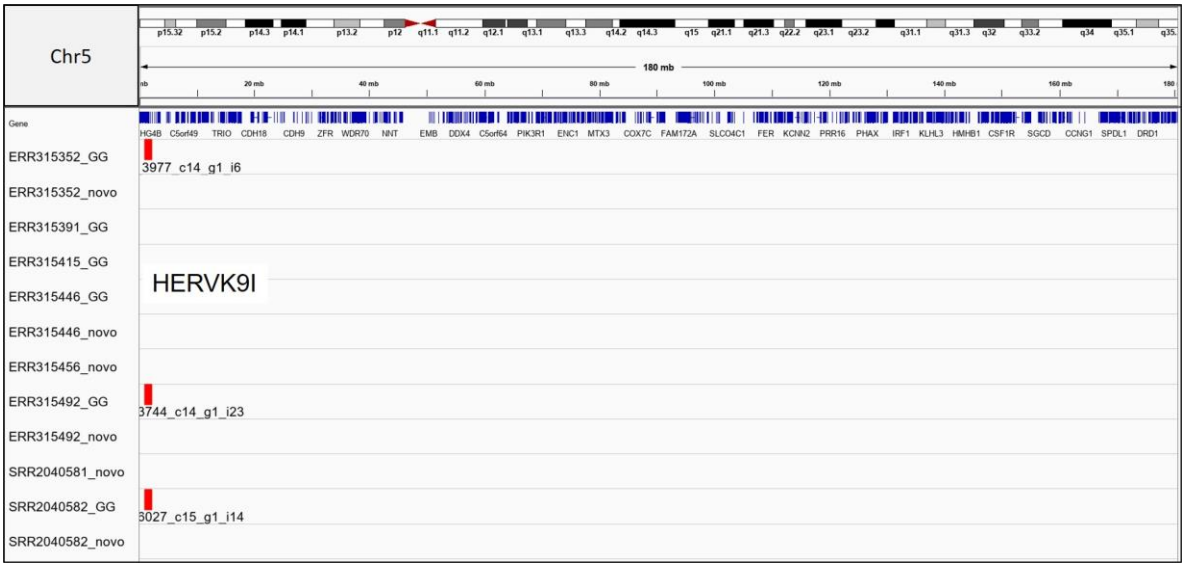
ERV	Humano	orangután	Clase	Familia
HARLEQUIN	3	1	I	ERV1
HERV1_I	2	1	I	ERV1
HERV17	7		I	ERV1
HERV3	2		I	ERV1
HERV46I	1	1	I	ERV1
HERV9	2		I	ERV1
HERVE	4	3	I	ERV1
HERVFH19I	1		I	ERV1
HERVH	4		I	ERV1
HERVH48I	3		I	ERV1
HERVIP10F	5		I	ERV1
HERVK	8	1	II	ERVK
HERVK22I	5	1	II	ERVK
HERVK3I	7		II	ERVK
HERVK9I	3		II	ERVK
HERVL	4		III	ERVL
HERVS71	9	1	I	ERV1
MER52AI	6		I	ERV1
PRIMA4_I	1		I	ERV1



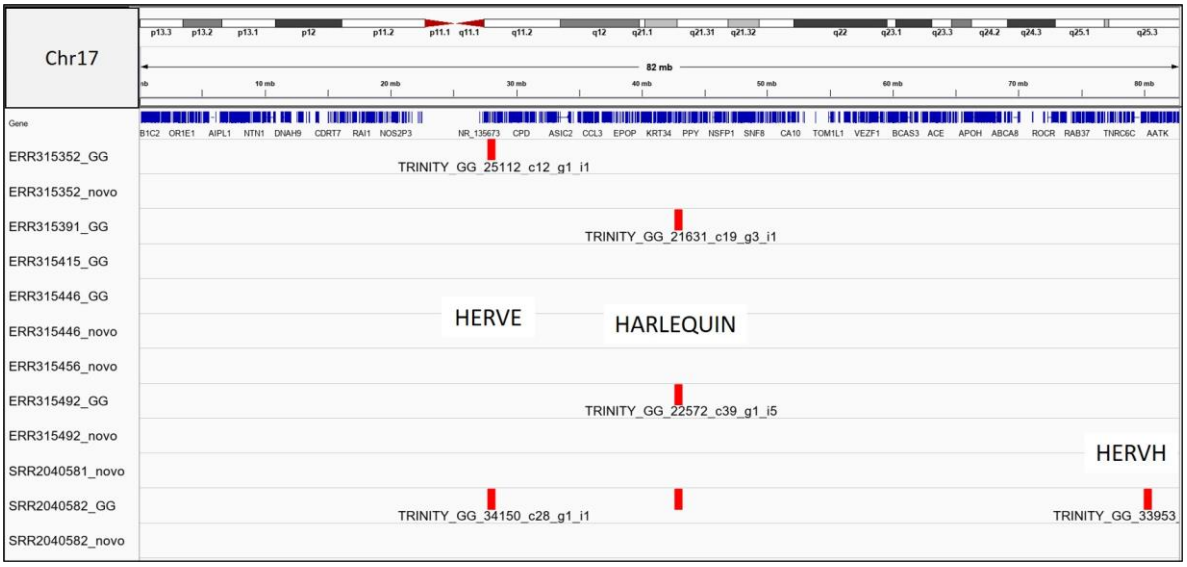
Anexo 3. Copias no idénticas de HERV22I ubicadas en el cromosoma 11 que fueron reconstruidas a partir de librerías de humano mediante ensamblaje de *novo* (azul) y guiado por genoma (rojo).



Anexo 4. Copias no idénticas de MER52AI ubicadas en el cromosoma 11 que fueron reconstruidas a partir de librerías de humano mediante ensamblaje de *novo* (azul) y guiado por genoma (rojo).

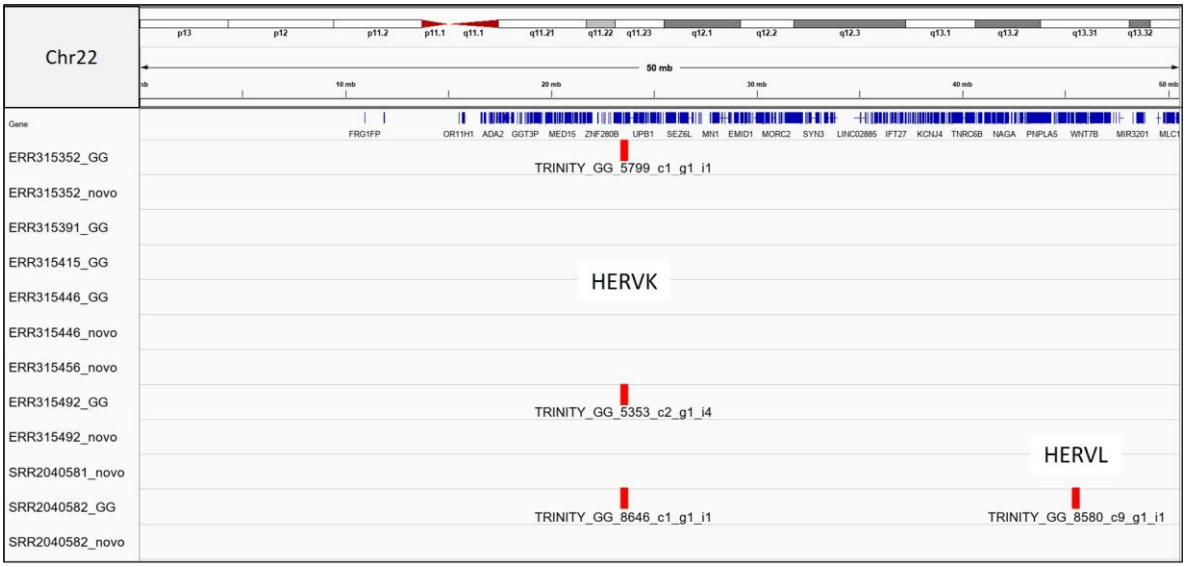


Anexo 5. Ubicación en el genoma de copias no idénticas del retrovirus endógeno HERVK9 reconstruidas a partir de librerías de humano mediante ensamblaje de *novo* (azul) y guiado por genoma (rojo).



Anexo 6. Copias no idénticas de retrovirus endógenos ubicadas en el cromosoma 17 que fueron reconstruidas a partir de librerías de humano mediante ensamblaje de *novo* (azul) y guiado por genoma (rojo).

Anexo 7. Copias no idénticas de retrovirus endógenos ubicadas en el cromosoma 19 que fueron reconstruidas a partir de librerías de humano mediante ensamblaje de *novo* (azul) y guiado por genoma (rojo).



Anexo 8. Copias no idénticas de HERVKs ubicadas en el cromosoma 22 que fueron reconstruidas a partir de librerías de humano mediante ensamblaje de *novo* (azul) y guiado por genoma (rojo).

Anexo 9. Descripción de intercepto de ERVs identificados con otros elementos del genoma.

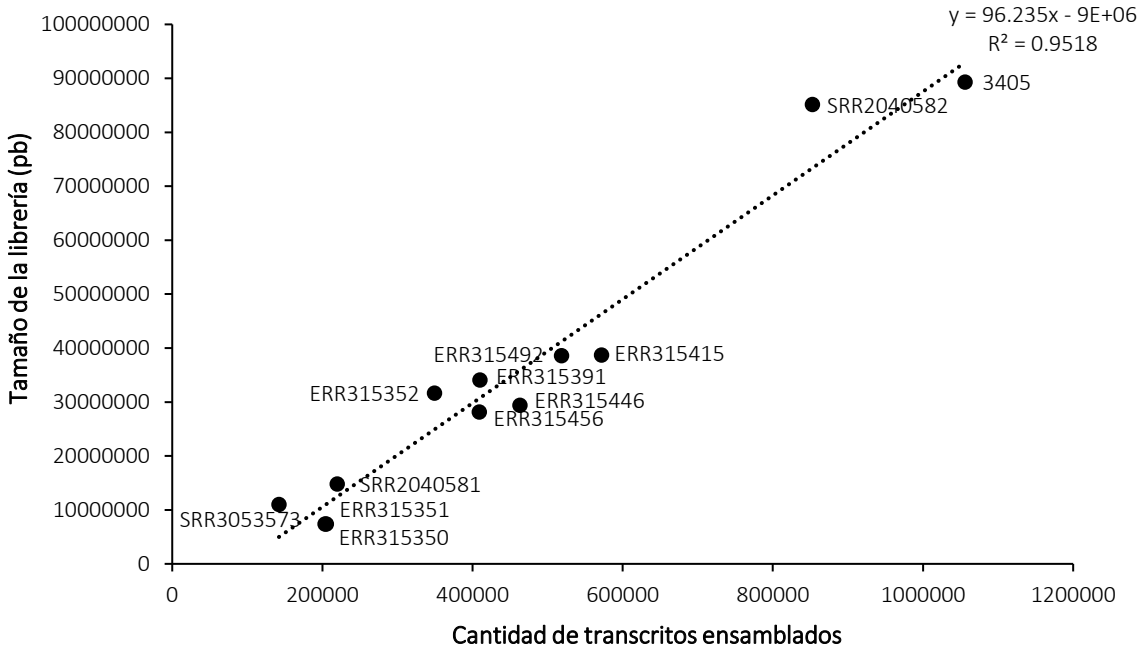
Transcrito	Chr	orientación	TPM	ERV	Genes	lncRNAs	LTRs
TRINITY_GG_18341_c4_g2_i8	7	-	286	HERV17	ERVW-1 PEX1		HERV17-int y HERVH-int
TRINITY_DN84422_c2_g2_i10	7	-	11	HERV17	ERVW-1		HERV17-int
TRINITY_GG_16962_c35_g1_i14	7	-	0.02	HERV17	ERVW-1 PEX1		HERV17-int
TRINITY_DN97241_c2_g2_i2	7	+	0	HERV17	ERVW-1		HERV17-int
TRINITY_DN57746_c1_g1_i2	7	-	469	HERV17			HERV17-int
TRINITY_GG_18529_c34_g1_i1	19	+	0.632	HERV17			HERV17-int
TRINITY_GG_26373_c30_g1_i1	7	-	115	HERV17	ERVW-1		HERV17-int
TRINITY_GG_19609_c14_g1_i8	12	-	248	HERVK22I	GOLGA2P5		HERVK22-int
TRINITY_GG_6608_c10_g4_i1	12	+	516	HERVK22I	GOLGA2P5		HERVK22-int
TRINITY_GG_3164_c4_g2_i3	12	+	30	HERVK22I	GOLGA2P5		HERVK22-int
TRINITY_GG_17833_c8_g1_i10	12	-	286	HERVK22I	GOLGA2P5		HERVK22-int
TRINITY_GG_26987_c16_g6_i3	12	-	0.122	HERVK22I	GOLGA2P5		HERVK22-int
TRINITY_GG_5799_c1_g1_i4	22	-	0	HERVK	PCAT14	TCONS_l2_00017644 TCONS_l2_00017645	HERVK-int
TRINITY_GG_5353_c2_g1_i4	22	-	104	HERVK	PCAT14	TCONS_l2_00017644 TCONS_l2_00017645	HERVK-int
TRINITY_GG_8646_c1_g1_i1	22	+	452	HERVK	PCAT14	TCONS_l2_00017644 TCONS_l2_00017645	HERVK-int
TRINITY_GG_18027_c57_g1_i2	19	+	103	HERVK			HERVK9-int
TRINITY_GG_18121_c244_g2_i1	19	+	0	HERVK	ZNF420		HERVK-int
TRINITY_GG_20185_c2_g1_i3	1	+	136	HERVK	CD48		HERVK-int
TRINITY_GG_36369_c10_g1_i4	11	+	0.06	HERVK			HERVK-int
TRINITY_GG_8326_c127_g1_i4	7	+	0.455	HERVK			HERVK-int
TRINITY_GG_26568_c5_g1_i1	11	+	177	MER52AI	NCR3LG1		HUERS-P3-int
TRINITY_GG_7893_c2_g1_i1	11	+	466	MER52AI	NCR3LG1		HUERS-P3-int
TRINITY_GG_4089_c7_g1_i2	11	+	431	MER52AI	NCR3LG1		HUERS-P3-int

TRINITY_DN80847_c2_g5_i2	11	-	33	MER52AI	NCR3LG1		HUERS-P3-int
TRINITY_GG_36351_c62_g1_i2	11	-	0.0001	MER52AI	NCR3LG1		HUERS-P3-int
TRINITY_GG_25112_c12_g1_i1	17	+	138	HERVE		TCONS_00025599	HERVE-int
TRINITY_GG_13686_c45_g1_i2	13	+	230	HERVE		TCONS_00022264	HERVE-int
TRINITY_GG_16357_c6_g1_i2	X	+	0.893	HERVE		TCONS_l2_00030488	HERVE-int
TRINITY_GG_34150_c28_g1_i1	17	-	110	HERVE		TCONS_00025599	HERVE-int
TRINITY_GG_8794_c0_g1_i1	10	-	21	HERVS71			HERVS71-int
TRINITY_GG_1197_c1_g1_i1	10	+	106	HERVS71		TCONS_l2_00003584	HERVS71-int
TRINITY_GG_1159_c0_g1_i1	10	+	128	HERVS71		TCONS_l2_00003584	HERVS71-int
TRINITY_DN83664_c2_g1_i4	10	-	0.052	HERVS71		TCONS_l2_00003584	HERVS71-int
TRINITY_GG_11191_c0_g1_i1	10	+	250	HERVS71		TCONS_l2_00003584	HERVS71-int
TRINITY_GG_22501_c30_g2_i3	14	+	0.416	HERVS71			HERVS71-int
TRINITY_GG_16191_c14_g1_i2	7	-	289	HERVS71		TCONS_l2_00026455 TCONS_l2_00027302 TCONS_l2_00026456	HERVS71-int
TRINITY_GG_26157_c44_g2_i1	7	-	111	HERVS71		TCONS_l2_00026455 TCONS_l2_00027302 TCONS_l2_00026456	HERVS71-int
TRINITY_DN293708_c1_g2_i1	7	-	121	HERVS71		TCONS_l2_00026455 TCONS_l2_00027302 TCONS_l2_00026456	HERVS71-int
TRINITY_GG_21631_c19_g3_i1	17	-	329	HARLEQUIN	NR_110868		Harlequin-int
TRINITY_GG_22572_c39_g1_i5	17	-	285	HARLEQUIN	NR_110868		Harlequin-int
TRINITY_GG_34805_c64_g3_i9	17	-	388	HARLEQUIN	NR_110868		Harlequin-int
TRINITY_GG_3977_c14_g1_i6	5	+	317	HERVK9I			HERVK9-int
TRINITY_GG_6027_c15_g1_i2	5	-	0.048	HERVK9I	SDHAP3 LOC728613	TCONS_l2_00022747 TCONS_l2_00023634 TCONS_l2_00023635 TCONS_l2_00023637 TCONS_l2_00022749 TCONS_l2_00022748 TCONS_l2_00023638 TCONS_l2_00022750 TCONS_l2_00022139 TCONS_l2_00022752 TCONS_l2_00022753 TCONS_l2_00023639 TCONS_l2_00023640 TCONS_l2_00023642 TCONS_l2_00022754 TCONS_l2_00022755	HERVK9-int
TRINITY_GG_3744_c14_g1_i23	5	+	0.066	HERVK9I	SDHAP3 LOC728613	TCONS_l2_00022747 TCONS_l2_00023634 TCONS_l2_00023635 TCONS_l2_00023637 TCONS_l2_00022749 TCONS_l2_00022748 TCONS_l2_00023638 TCONS_l2_00022750 TCONS_l2_00022139 TCONS_l2_00022752 TCONS_l2_00022753 TCONS_l2_00023639 TCONS_l2_00023640 TCONS_l2_00023642 TCONS_l2_00022754 TCONS_l2_00022755	HERVK9-int
TRINITY_DN110865_c1_g1_i1	14	-	231	HERVH48I	SYNJ2BP COX16		HERVH48-int
TRINITY_GG_16183_c6_g2_i1	X	-	135	HERVH48I		TCONS_l2_00030703 TCONS_l2_00030228	HERVH48-int
TRINITY_DN293242_c3_g2_i9	X	+	0.376	HERVH48I		TCONS_l2_00030703 TCONS_l2_00030228	HERVH48-int
TRINITY_GG_11273_c97_g2_i1	19	+	129	HERVK3I	ZNF528		HERVK3-int
TRINITY_GG_11961_c26_g1_i1	19	+	199	HERVK3I	ZNF528		HERVK3-int
TRINITY_GG_18180_c153_g1_i1	19	-	0.99	HERVK3I	ZNF439		HERVK3-int

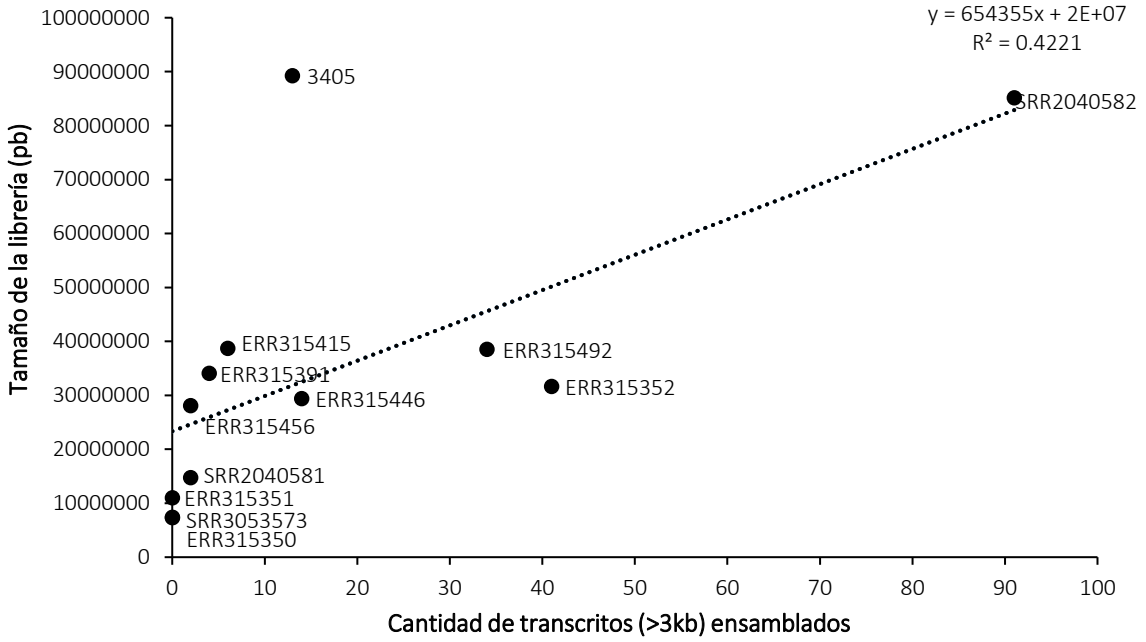
TRINITY_GG_18291_c173_g1_i6	19	-	194	HERVK3I	NR_144447 ERVK3-1		
TRINITY_GG_18519_c16_g1_i3	19	+	0.45	HERVK3I	FPR3 ZNF577		HERVK3-int
TRINITY_DN292066_c7_g1_i3	19	+	374	HERVK3I	ZNF528		HERVK3-int
TRINITY_GG_6867_c9_g1_i1	16	+	167	HERVK3I		TCONS_l2_00009710	HERVK3-int
TRINITY_GG_18733_c0_g1_i1	2	-	177	HERVIP10F			HERVIP10F-int
TRINITY_DN73353_c1_g2_i2	18	+	278	HERVIP10F	LOC644669	TCONS_l2_00012203 TCONS_l2_00012204 TCONS_l2_00012205	HERVIP10F-int
TRINITY_DN97161_c6_g1_i1	18	-	175	HERVIP10F	LOC644669	TCONS_l2_00012203 TCONS_l2_00012204 TCONS_l2_00012205	HERVIP10F-int
TRINITY_GG_10094_c8_g1_i1	18	-	174	HERVIP10F	LOC644669	TCONS_l2_00012203 TCONS_l2_00012204 TCONS_l2_00012205 TCONS_l2_00011997 TCONS_l2_00011998	HERVIP10F-int
TRINITY_GG_28800_c4_g1_i15	2	+	143	HERVIP10F	ANKRD30BL	TCONS_l2_00014914 TCONS_l2_00014916 TCONS_l2_00015978 TCONS_l2_00014917	HERVIP10F-int
TRINITY_GG_13764_c36_g1_i1	13	-	0.828	HERV9			HERV9NC-int
TRINITY_GG_19787_c5_g1_i1	1	+	0.956	HERV9	WARS2-AS1		HERV9NC-int
TRINITY_GG_22467_c36_g1_i2	14	-	0.814	HERVH			HERVH-int
TRINITY_GG_33953_c15_g1_i1	17	-	0.415	HERVH	RNF213-AS1		HERVH-int
TRINITY_GG_22082_c38_g1_i9	14	+	0.77	HERVH	ATXN3	TCONS_00022381 TCONS_00022382	HERVH-int
TRINITY_GG_26373_c50_g7_i1	7	+	0.712	HERVH			HERVH-int
TRINITY_GG_25846_c63_g1_i3	7	+	0.954	HERV3	ZNF117 ERV3-1		HERV3-int
TRINITY_DN287573_c5_g3_i2	7	-	156	HERV3	ZNF117 ERV3-1		HERV3-int
TRINITY_GG_20435_c7_g1_i8	1	+	0.857	HERVL			HERVL-int
TRINITY_GG_32241_c15_g1_i1	4	-	0.85	HERVL			HERVL-int
TRINITY_GG_8580_c9_g1_i1	22	-	0.70	HERVL	FBLN1		HERVL-int
TRINITY_GG_27569_c5_g1_i3	12	+	213	HERVL	cD48		HERVL-int
TRINITY_GG_12302_c10_g1_i7	10	-	0.727	HERV1_I			HERV1_I-int
TRINITY_DN292862_c2_g1_i1	10	-	0.957	HERV1_I			HERV1_I-int
TRINITY_GG_28417_c38_g1_i1	2	-	0.528	HERVFH19I			HERVFH19-int
TRINITY_GG_36596_c32_g1_i1	11	-	0.853	HERV46I	JRKL-AS1	TCONS_00019737	LTR46-int
TRINITY_GG_9698_c54_g4_i3	18	+	0.755	PRIMA4_I	TXNL4A		PRIMA4-int

Anexo 10. Traslape de la posición de las 41 copias de eERV con genes, lncRNAs y LTRs reportados en hg38.

ERV	Chr	Inicio	Final	No. librerías	Genes	lncRNAs	LTRs
HERVL	chr1	55580995	55736104	1		0	0
HERV9	chr1	119174301	119178985	1	WARS2-AS1	0	0
HERVK	chr1	160690937	160699906	1	CD48	0	0
HERVIP10F	chr2	132139713	132161956	2	ANKRD30BL	4	0
HERVFH19I	chr2	238206291	238210262	1		0	0
HERVL	chr4	139417093	139447681	1		0	0
HERVK9I	chr5	1571908	1632331	3	LOC728613 / SDHAP3	16	0
HERVK	chr7	4587241	4595922	1		0	0
HERVS71	chr7	29636370	29645371	2		3	0
HERV3	chr7	64978994	64998274	1	ZNF117 / ERV3-1	0	0
HERV17	chr7	92468388	92528514	5	ERVW-1 /PEX1	0	HERVH
HERVH	chr7	92479345	92484963	1		0	0
HERV1_I	chr10	42643923	42652666	1		0	0
HERVS71	chr10	52946439	52955567	4		1	0
MER52AI	chr11	17336959	17377375	4	NCR3LG1	0	HUERS-P3
HERV46I	chr11	96501944	96506751	1	JRKL-AS1	1	LTR46
HERVK	chr11	118720831	118729501	1		0	0
HERVL	chr12	14365654	14502957	1	cD48	0	0
HERVK22I	chr12	100156286	100163234	5	GOLGA2P5	0	0
HERV9	chr13	27212761	27216505	1		0	0
HERVE	chr13	40874388	40881277	1		1	0
HERVH	chr14	70293163	70300422	1		0	0
HERVH48I	chr14	70346788	70350037	1	SYNJ2BP/COX16	0	0
HERVH	chr14	92042979	92059586	1	ATXN3	2	0
HERVS71	chr14	106196990	106203051	1		0	0
HERVK3I	chr16	35520414	35524687	1		1	0
HERVE	chr17	28230587	28240000	2		1	0
HARLEQUIN	chr17	43158981	43167108	3	NR_110868	0	0
HERVH	chr17	80403343	80408496	1	RNF213-AS1	0	0
HERVIP10F	chr18	15308637	15330344	3	LOC644669	5	0
PRIMA4_I	chr18	80025407	80033960	1	TXNL4A	0	0
HERVK	chr19	9435604	9462482	1		0	HERVK9
HERVK3I	chr19	11854696	11858211	1	ZNF439	0	0
HERV17	chr19	22747259	22751974	1		0	0
HERVK	chr19	37106649	37113057	1	ZNF420	0	0
HERVK3I	chr19	51804439	52415578	3	ZNF528 / FPR3 / ZNF577	0	0
HERVK3I	chr19	58305340	58315971	1	NR_144447 / ERVK3-1	0	0
HERVK	chr22	23536842	23547781	3	PCAT14	2	0
HERVL	chr22	45556467	45559969	1	FBLN1	0	0
HERVE	chrX	49139620	49143130	1		1	0
HERVH48I	chrX	57219432	57224703	1		2	0
		Total		68	22	78	4



Anexo 11. Análisis de correlación entre el tamaño de la librería después del trimming y la cantidad total de transcritos ensamblados.



Anexo 12. Análisis de correlación entre el tamaño de la librería después del trimming y la cantidad total de transcritos ensamblados de tamaño superior a 3kb.