Caracterización *in silico* de elementos transponibles en el genoma de *Trypanosoma cruzi*.

Gaston Rijo

Tutora: Dra. Luisa Berná

Co-tutor: Dr. Sebastián Pita

Unidad de Bioinformatica, Institut Pasteur de Montevideo

11 de Mayo de 2018

A mis viejos

Tabla de contenidos

Resumen

En el presente trabajo se realiza una caracterización bioinformática de los elementos transponibles CZAR, L1Tc, NARTc, VIPER y SIRE de *Trypanosoma cruzi.* En un primer acercamiento, se utilizaron los genomas de las cepas TCC y Dm28c secuenciadas con tecnología SMRT de PacBio, a los cuales se les realizó una búsqueda por similaridad (BLAST) utilizando las secuencias de retroelementos de referencia depositadas en NCBI. Se determinaron estadísticas generales de las distintas copias de los retroelementos (*i.e.* número de copias, contenido GC), los elementos potencialmente activos de los retrotransposones autónomos, y se construyeron filogenias con los elementos de ambas cepas. Se determinaron las anotaciones próximas a cada elemento con el objetivo de elucidar contextos genómicos preferenciales para la inserción. Para obtener un panorama más amplio sobre la historia evolutiva de los retroelementos, se aplicó la búsqueda a genomas de otras cepas disponibles en distintas bases de datos genómicas y se construyeron filogenias con las secuencias obtenidas.

Los resultados obtenidos confirman resultados previos sobre algunos de los retrotransposones. Adicionalmente, las filogenias construidas sugieren que hubieron amplificaciones de CZAR, L1Tc y VIPER posteriores a la divergencia entre las distintas DTUs, mientras que en el caso de NARTc y SIRE, las filogenias sugieren que la amplificación y posterior inactivación fue previa a la divergencia.

Introducción

Enfermedad de Chagas

La enfermedad de Chagas en una enfermedad neotropical endémica transmitida principalmente a través de insectos vector de la subfamilia Triatominae (vinchucas) siendo su agente etiológico el protozoario *Trypanosoma cruzi*. El número de personas infectadas por Chagas es alrededor de los 8 millones, y se estima que produce alrededor de diez mil muertes por año(1). En 2013, se estimó que la enfermedad de Chagas produce un costo financiero de 627 millones de dólares anuales(2). La enfermedad de Chagas no tiene cura. Las principales estrategias de control diseminación de la enfermedad son a través de la erradicación del vector en el ambiente doméstico y prevención de transmisión vertical y por transfusión de sangre. Actualmente existen dos medicamentos para el tratamiento de casos agudos (Nifurtimox y Benznidazol), que debido a sus efectos secundarios no son efectivos para pacientes crónicos. Considerando estos factores, es relevante estudiar la biología de *T. cruzi* para poder desarrollar nuevas tecnologías de diagnóstico, prevención y cura(1,3).

Biología de *Trypanosoma cruzi*

Trypanosoma cruzi es un protozoario del orden Trypanosomatidae, clase Kinetoplastida, phylum Euglenozoa, dominio Eukarya. El orden Trypanosomatidae contiene a varios parásitos de insectos, de los cuales algunos utilizan a los humanos como hospederos secundarios; *Leishmania spp., Trypanosoma brucei* y *Trypanosoma cruzi*, responsables de la Leishmaniasis, enfermedad del sueño y de Chagas, respectivamente.

El parásito posee un ciclo de vida que se alterna entre el insecto vector y el mamífero hospedero. A lo largo del ciclo, el parásito pasa por diferentes etapas morfológicas adaptadas a las diferentes condiciones. El ciclo comienza cuando el insecto se alimenta de la sangre del individuo a infectar, dejando una herida en la piel. De forma concomitante con la succión de la sangre, el insecto vector defeca heces en las cuales se encuentra el parásito en su forma denominada tripomastigota metacíclico. Los tripomastigotas metacíclicos ingresan en el torrente sanguíneo, lo cual les permite invadir células, infectando diferentes tejidos. Una vez ocurrida la invasión celular, los tripomastigotas metacíclicos se transforman en amastigotas, y comienzan una fase de multiplicación por fisión binaria. Los amastigotas se transforman en tripomastigotas, para luego abandonar la célula mediante lisis. Estos tripomastigotas pueden infectar

más células, reanudando el ciclo replicativo, o pueden ser ingeridos por un insecto vector al alimentarse de la sangre. Los tripomastigotas se adhieren al intestino medio del insecto, se transforman en epimastigotas y entran en una fase replicativa aumentando su número por fisión binaria. Parte de estos epimastigotas migran hacia el intestino posterior transformándose en tripomastigotas metacíclicos, y si el insecto vector defeca sobre la herida que causo en un mamífero, se reanuda el ciclo de vida del parásito(4).

Trypanosoma cruzi es un organismo predominantemente diploide(5), y se reproduce de forma asexuada por fisión binaria(6). Ya que la cromatina no se condensa a lo largo de su ciclo celular, es necesario realizar cariotipos moleculares usando electroforesis de campo pulsado. Posee una alta variabilidad en cuanto al número de cromosomas(7) y se han reportado casos de aneuploidías(8). A pesar de la naturaleza clonal de su mecanismo reproductivo, *T. cruzi* posee los genes necesarios para llevar a cabo la reproducción sexuada(5). Sumado a esto, varios de los linajes existentes del parásito poseen genomas híbridos(9), y se han obtenido individuos recombinantes en experimentos in vitro(10). Estos hechos sugieren que *T. cruzi* retiene una capacidad potencial para la reproducción sexual.

Actualmente se sostiene que existen seis linajes principales, o Discrete Typing Units (DTUs) en el sentido más estricto: TcI, TcII, TcIII, TcIV, TcV, TcVI, con algunas diferencias en sus nichos, distribución geográfica, hospederos, vectores e implicaciones clínicas(11). Debido a la naturaleza asexuada de la reproducción, y a eventos de recombinación meiótica en el pasado, las DTUs se pueden clasificar en homocigotas: TcI, TcII, TcIII, TcIV, y heterocigotas: TcV y TVI. Actualmente, se sostienen dos modelos de evolución de los linajes en los cuales tanto TcI como TcII provienen de un linaje ancestral, mientras que TcIII, TcV y TcVI provienen de eventos de hibridación entre TcI y TcII (Fig 1)(12).

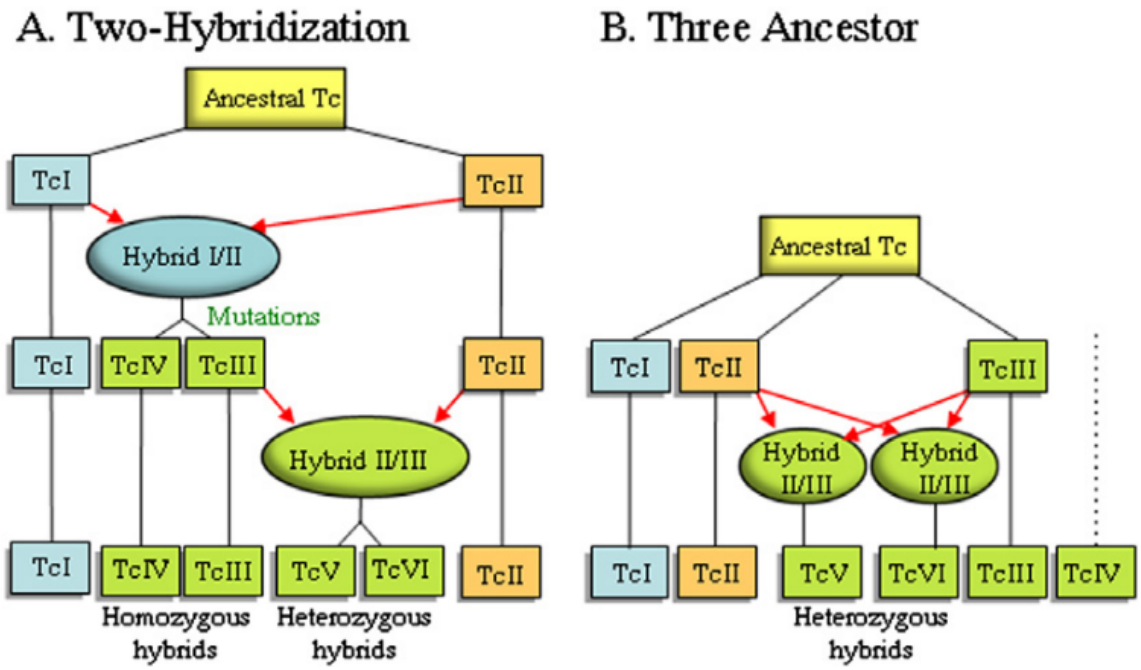


Figura 1. Modelos de intercambio genético entre linajes de Trypanosoma cruzi. Adaptado de Zingales et al 2012.

T. cruzi, así como los kinetoplástidos en general, tiene la particularidad de no regular de forma apreciable la tasa de iniciación de la transcripción, regulando la expresión de genes codificantes a nivel post transcripcional. Otra particularidad es que transcriben sus genes codificantes de forma policistrónica, sin mostrar sesgos por la función del gen, salvo algunas excepciones(13). Los policistrones están organizados en *Directional Gene Clusters* (DGCs), conjunto de genes ubicados en la misma hebra, y que se transcriben en un mismo RNA precursor. Estos DGCs están delimitados por regiones en la cual cambia la hebra transcripta, llamadas *Strand Switch Regions* (SSRs). Como consecuencia las SSR pueden ser cabeza-cabeza (ambos DGCs adyacentes terminan la transcripción), o cola-cola (ambos DGCs adyacentes inician la transcripción)(14). Estas regiones están enriquecidas en secuencias derivadas de retrotransposones, sugiriendo que estos podrían cumplir un rol importante en la organización genómica de los tripanosomátidos(15).

La maduración hacia ARNm a partir de un ARN policistrónico involucra un proceso de *trans*-splicing en el cual se agrega un ARN conteniendo un cap 5' del exón del gen *Spliced Leader*(SL), con la concomitante poliadenilación del extremo 3'(14).

Una excepción a la transcripción policistrónica son los genes *Spliced Leader*, que poseen su promotor transcripcional corriente arriba del sitio de inicio(16). Contienen un

exón de 35 pb, el cual es ligado a los mRNA en el proceso de *trans*-splicing(17). Se encuentran ordenados en tándem, separados por secuencias intergénicas de ~400 pb, las cuales han sido utilizadas como marcador para estudios filogenéticos(18).

El genoma de *T. cruzi* se secuenció por primera vez en 2005 por El-Sayed y colaboradores. Este trabajo fue de suma importancia, permitiendo revelar aspectos clave de la biología del parásito. Se logró estimar el número de genes presentes, y muchas de sus funciones; la mayoría de estos pertenecen a familias multigénicas y ARN ribosomal. Fueron determinadas las secuencias de genes involucrados en procesos de reparación de ADN, vías de señalización y generación de respuesta inmune. Subsecuentes análisis genómicos permitieron ampliar el conocimiento sobre las capacidades metabólicas y patológicas del organismo. Sumado a esto, fue caracterizado un gran número de secuencias repetitivas, estimando que constituyen el 50% del genoma. Estas secuencias repetitivas incluyen las familias multigénicas, repetidos en tándem, y retrotransposones(5).

Clasificación de elementos transponibles

Las secuencias repetidas fueron consideradas por mucho tiempo “ADN basura”. Sin embargo, hoy se sabe que participan en la evolución de los genomas, contribuyendo a su plasticidad, moldeando y dando luz a nuevas funciones e innovaciones biológicas. Por su naturaleza repetitiva, estas secuencias tienden a sufrir recombinaciones homólogas y ectópicas. En particular, los elementos transponibles, tienen la capacidad de insertarse en ORFs y secuencias regulatorias, alterando los procesos celulares establecidos y dotando de materia prima a los procesos evolutivos para la generación de complejidad biológica. Cabe destacar que numerosas enfermedades son causadas por inserciones de elementos móviles o por los rearreglos cromosómicos que estos pueden promover (19).

En 2007, Wicker y colaboradores (20) propusieron un sistema de clasificación jerárquico para los elementos transponibles eucarióticos, con el fin de proveer un consenso para facilitar el trabajo de la anotación y el análisis evolutivo de estos elementos. Este trabajo hace uso de esa clasificación.

Los elementos transponibles (ET) se dividen en dos clases: los retrotransposones (ET de clase I), y los transposones de ADN (ET de clase II). Los transposones de DNA, se escinden del genoma de forma temporal, para luego insertarse nuevamente. Esta clase de elementos transponibles no se encuentra en el genoma de *T. cruzi*. Los retrotransposones se replican usando un intermediario de ARN (transcripto de su secuencia), para luego insertarse en otro punto del genoma.

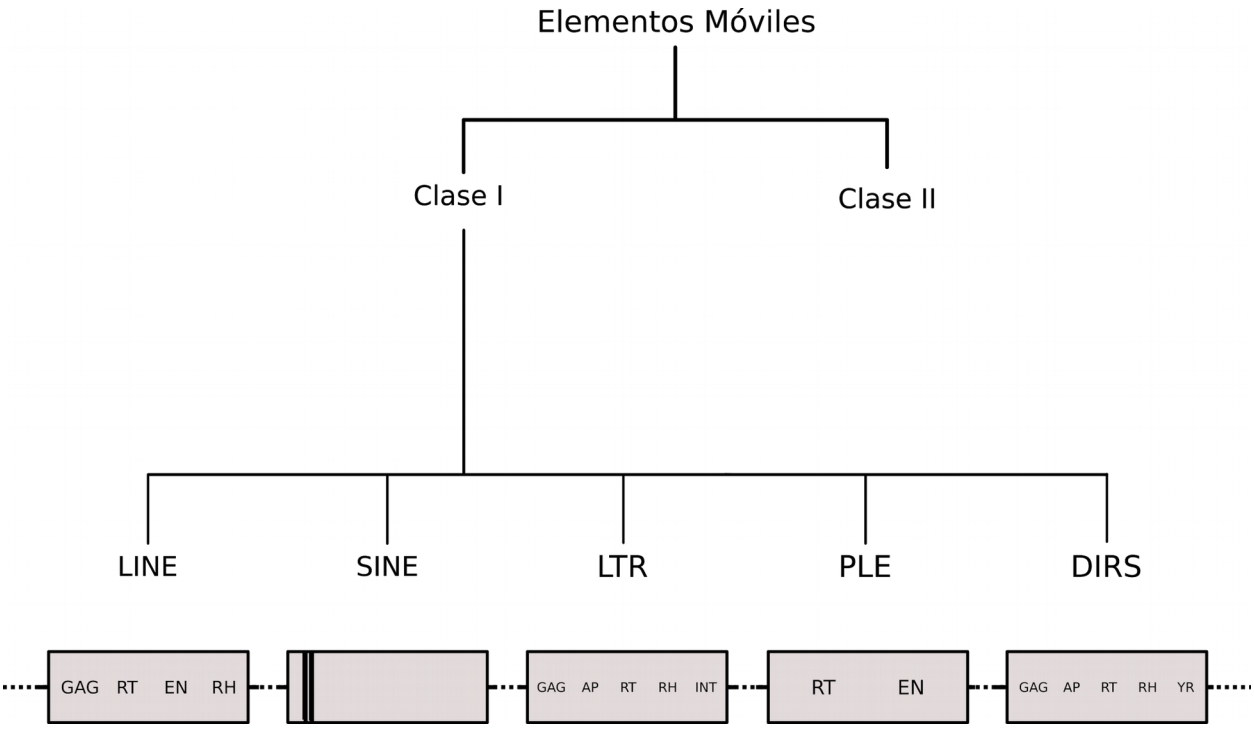


Fig 2 Clasificación de elementos transponibles de Clase II, junto con los dominios proteicos más comunes en sus ORFs. Los retroelementos SINE no tienen ORFs, suelen tener en cambio secuencias de promotores en sus extremos 3'.

Cada clase se divide en subtaxones del nivel orden, y el criterio de clasificación a este nivel depende del mecanismo específico de retrotransposición (desde el punto de vista mecanístico y enzimológico), de su organización (qué marcos abiertos de lectura poseen, como están organizados), y la relación filogenética de sus dominios transcriptasa reversa.

Los retrotransposones son subdivididos en cinco órdenes; Long Terminal Repeat (LTR), Penelope Like Element (PLE), Long Interspersed Nuclear Element (LINE), Short Interspersed Nuclear Element (SINE) y Dictyostelium Intermediate Repeat Sequence (DIRS).

Los retrotranposones del orden LTR son predominantes en los genomas de plantas, y presentes en algunos genomas de animales. Su longitud se encuentra en el intervalo de los cientos de pares de bases hasta 25 kb. Contienen marcos abiertos de lectura (ORFs) que codifican para una proteína GAG, necesaria para el ensamblaje de *virus-like particles* (necesarias para el proceso de retrotransposición) y una proteína POL que

posee dominios proteasa aspártica (AP) cuya función es clivar la proteína POL; transcriptasa reversa(RT), necesaria para transcribir el mRNA del TE; RNasa-H (RH), cuya función es degradar el ARN en la hebra ARN-ADN formada en la retrotranscripción; y DDE integrasa(INT), necesaria para realizar la recombinación entre el ADN intermediario y el ADN genómico. La característica principal de los LTR, a la cual debe su nombre, es la presencia de repetidos largos de 100-700 pb a cada extremo del elemento. Esta particularidad es esencial para un mecanismo de retrotransposición exitoso, y además es clave para comprender el patrón de degradación de este orden.

Descritos por primera vez en *Drosophila virilis*, los elementos móviles del orden PLE suelen presentar dos ORFs, uno con un dominio RT, y otro con un dominio endonucleasa EN. Suelen presentar secuencias flanqueantes similares a los LTR en sentido directo o inverso. Poseen una distribución extensa, al estar presentes en genomas de animales multi y unicelulares, plantas y hongos (20).

Los retroelementos del orden LINE se encuentran en la mayoría de organismos eucariotas. Sus ORFs codifican al menos un dominio RT y un dominio nucleasa (endonucleasa o endonucleasa apurínica/apirimidinica), necesarios para la transposición. Suelen poseer ORFs similares a la proteína GAG, y los LINEs de la superfamilia I suelen poseer un ORF codificante para una RNAsa H. Se ha reportado que los ARN intermediarios tienen afinidad por las proteínas que codifican; una vez que los ORFs son traducidos en el citoplasma, las proteínas se asocian con el ARN intermediario y este complejo ARN-proteína es importado hacia el núcleo, lo que contribuye a la inserción copia-específica. Suelen presentar duplicaciones en sus extremos que se originan a partir del mecanismo de inserción (duplicaciones del sitio de inserción, DSI).

Los elementos del orden SINE son elementos transponibles no autónomos, dependen de la maquinaria de retrotransposición de otros elementos móviles para su replicación. Se originan a partir de la fusión de un promotor de pol III localizado en el extremo 5’ del SINE y un segmento hacia el extremo 3’ que suelen poseer homología con ETs del orden LINE. Presentan duplicaciones en sus extremos, ya que su mecanismo de replicación es necesariamente igual a los LINEs. Son más pequeños que los elementos autónomos, cubriendo un rango de 80 a 500 pb. Uno de los elementos SINE más estudiados es *Alu*, que se encuentra exclusivamente en el genoma humano con aproximadamente un millón de copias (21), y hay abundante evidencia de su rol en el origen de enfermedades hereditarias(22).

Los retroelementos del orden DIRS, de manera similar a los retroelementos LTR, poseen ORFs que codifican para proteínas GAG, y POL con dominios AP, RT y RH. Tienen en cambio la particularidad de poseer un dominio tirosina recombinasa (YR), a diferencia de los LTR que poseen dominio integrasa, que cumple la misma función. Otra particularidad es que sus extremos presentan repetidos inversos o *split direct repeats*. Estas diferencias implican que los retroelementos del orden DIRS poseen un

mecanismo de retrotransposición distinto al de otros órdenes de retrotransposones(20).

Bajo la taxonomía propuesta por Wicker et al, las superfamilias dentro de un orden comparten estrategias de replicación, y su clasificación depende de características estructurales a gran escala, como puede ser la estructura de las proteínas que codifican sus ORFs, estructura de secuencias no codificantes, y tamaño de DSI. Generalmente no hay conservación de secuencia a este nivel. El nivel familia está determinado por conservación en secuencia de ADN, en general restringida a segmentos conservados en regiones codificantes. Las subfamilias están definidas por datos filogenéticos y son útiles para diferenciar entre elementos autónomos y no autónomos. El último nivel, inserción, está definido como un evento de transposición particular en el genoma, con coordenadas genómicas particulares. El término inserción se usa de forma intercambiable con el término copia en este trabajo.

En el genoma de *T. cruzi* se han encontrado únicamente retrotransposones de los órdenes LINE, SINE y DIRS. Consecuentemente, este trabajo está enfocado en la biología de este subconjunto de elementos transponibles.

Biología de los elementos transponibles

Los TEs poseen la capacidad de autorreplicarse dentro del genoma de un individuo, y evolucionar al enfrentarse a presiones selectivas ejercidas por el hospedero. Como consecuencia de estas presiones selectivas, muchas familias de TEs se encuentran altamente conservadas a lo largo de distintos clados y existe una gran variedad de mecanismos de replicación, sitios de inserción y patrones de degradación (23).

Un modelo de mecanismo de replicación ampliamente aceptado de los órdenes LINE y SINE depende del uso del ARN del retrotransposon como molde para su retrotranscripción directa en el genoma del hospedero. Este mecanismo (Figura 3) depende de la presencia en su extremo 3’ de una cola poly(A), un repetido en tándem, o una región rica en adenina. La endonucleasa cliva una de las hebras del ADN genómico exponiendo una secuencia complementaria al extremo 3’ del ARN intermediario, para luego formarse un complejo ARN-ADN (ii). Una vez establecido el apareamiento, comienza la retrotranscripción por parte de la transcriptasa reversa del retrotransposon (iii). A continuación, el ADN genómico es clivado por la endonucleasa en la hebra de ADN genómico complementaria corriente arriba del complejo ARN-ADN, exponiendo un extremo 3’-OH en el ADN genómico. Este extremo 3’-OH es susceptible a una ligación que involucre a la secuencia retrotranscripta (v), lo cual tendría como consecuencia la síntesis dependiente de ADN de la secuencia complementaria a la hebra retrotranscripta y luego su ligación (vi). Este tipo de integración produce DSI en los extremos flanqueantes del retrotransposon (vii)(24).

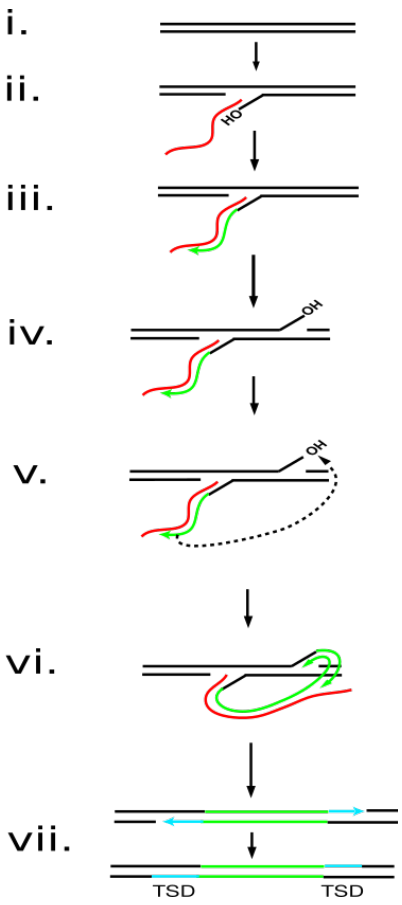


Figura 3. Modelo de retrotransposición de LINEs. Adaptado de Han et al 2010.

Las particularidades de este mecanismo de retrotranscripción tiene algunas implicancias en cuanto a los patrones que se observan a nivel de conservación de secuencia, y al momento de inferir sobre la autonomía de las copias. En el proceso de retrotranscripción, la retrotranscriptasa puede disociarse sin haber llegado hacia el final del ARN intermediario, lo cual tendría como resultado la integración de un elemento truncado en su extremo 5', y posiblemente no autónomo debido a la fragmentación del primer ORF. A este tipo de integración de elementos inactivos se le suele llamar "*dead on arrival*" y es muy común encontrar en los genomas elementos truncados por su extremo 5', aunque podrían seguir activos al utilizar la maquinaria de retrotransposición de sus parientes activos(20).

Los LINEs y SINEs poseen especificidad de inserción variada; los elementos R1 y R2 de insectos se insertan dentro del locus 28S rDNA mediante la acción de endonucleasas (AP endonucleasa en el caso de R1)(25,26). Otros ejemplos son los elementos DRE y Tdd3 de *Dictyostelium discoideum*, que se insertan corriente arriba y

abajo de genes de ARNt, respectivamente. Estudios demuestran que el retroelemento L1 con una endonucleasa inactiva puede integrarse en sitios con lesiones en el ADN y en telómeros disfuncionales en células CHO deficientes de p53(27,28).

Se han propuesto dos mecanismos de replicación distintos para los retrotransposones del orden DIRS, cuya diferencia radica en el tipo de regiones repetitivas que poseen en sus secuencias. Los retrotransposones denominados DIRS-*like* poseen repetidos terminales largos invertidos (*inverted terminal repeats*, ITRs) parcialmente complementarios entre sí en sus extremos 5' y 3', y repetidos complementarios a los ITRs en una región interna no codificante denominados *internal complementary regions*. Otros retrotransposones del orden DIRS, denominados PAT-*like*, poseen cuatro regiones de secuencias repetidas; A1, se encuentra en el extremo 5', B1 y A2 se encuentran en orden en una región interna no codificante y B2 se encuentra en el extremo 3', en donde las secuencias A1 y A2 son complementarias entre sí, al igual que B1 y B2(29–31).

Según el modelo de replicación de los elementos DIRS-like (Fig 4A), el proceso comienza por la transcripción del retrotransposon (a), este ARN es luego retrotranscripto en un ADN copia (ADNc) por la transcriptasa reversa del elemento(b). Los ITRs del ADNc se aparean parcialmente en donde uno de los ITRs sobresale con respecto al otro, acto siguiente, una ADN polimerasa utiliza uno de los ITRs como molde para completar el más corto (c,d). A continuación, ambos ITRs se aparean con sus respectivos ICRs (e), esta estructura es luego convertida en un ADN circular de simple cadena mediante el uso de ADN polimerasas y ligasas (f)(30).

El modelo propuesto para la replicación de los elementos PAT-*like* (Fig 4B) comienza con la transcripción del retroelemento en ARN(a), y su subsecuente retrotranscripción en un ADNc (b). El proceso de circularización de ADNc depende de las secuencias repetidas A y B; las secuencias A1 y B2 del ADNc se aparean con las secuencias A2 y B1 del ARN del retrotransposon(c), respectivamente. En este complejo híbrido, el ARN es utilizado como molde para la circularización del ADNc (d)(31).

Los ADN circulares en ambos modelos son luego sujetos a una síntesis de doble hebra. El ADN de doble hebra circular resultante, sufre una recombinación catalizada por la tirosina recombinasa codificada por uno de los ORFs del retroelemento para integrarse en el genoma(29). Se ha reportado que los retrotransposones del orden DIRS poseen especificidad de inserción por secuencias GTT y TAA(29), además de una preferencia de inserción por secuencias DIRS pre-existentes(30).

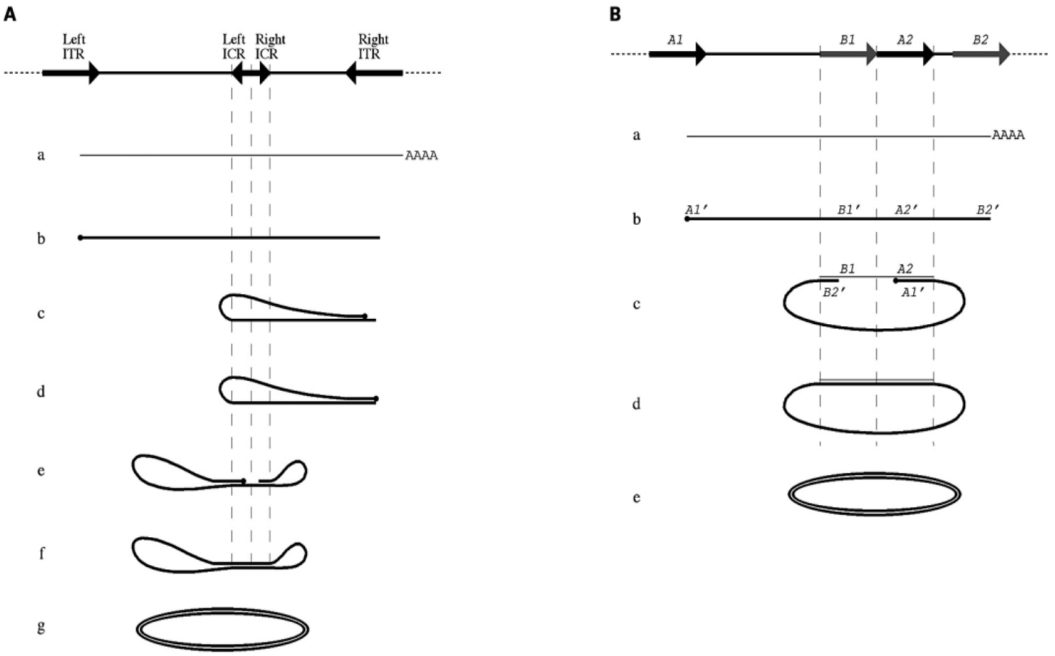


Figura 4. Modelo de retrotransposición de DIRS. Adaptado de Poulter et al 2005.

Otras características son compartidas por diversos órdenes de ETs. La mayoría de los ETs poseen promotores para Pol II o Pol III, lo que garantiza que el inicio de su transcripción sea independiente del hospedero. En eucariotas multicelulares, es necesario que los ETs sufran períodos de expansión en la línea germinal para poder transmitirse de manera vertical, y de hecho, se ha encontrado evidencia en plantas y animales de un aumento en la actividad de transposición en células de la línea germinal(32).

Existen modelos que proponen como los ETs podrían modelar la arquitectura genómica y la regulación génica. Los ETs poseen secuencias de unión a proteínas de unión a ADN (DBP) que podrían interactuar con genes en *cis*, modificando su transcripción. Si en un evento de expansión, las copias del ET se insertan cerca de un set de genes particular, o dentro de sus UTRs, este set de genes podría comenzar a ser corregulado por las DBP que se unen a las secuencias del ET, estableciendo una red regulatoria. Un caso similar podría suceder si el mRNA del ET poseyera secuencias de unión a proteínas de unión a ARN (RBP); los genes cercanos a los ETs podrían ser transcriptos junto con un segmento del ET que posea el sitio de unión a RBP, estableciendo así una red regulatoria post-transcripcional. Otra posibilidad es la interacción de los ETs con el silenciamiento génico mediado por microARNs; de forma similar al caso anterior, si los genes son transcriptos con un segmento de ARN perteneciente al ET, y a su vez las secuencias de ETs expresadas son procesadas por la maquinaria de silenciamiento para su uso como ARN guía, el set de genes puede ser elegido como blanco para su degradación, controlando sus niveles de expresión(33).

Algunas de las funciones y mecanismos propuestos para los TE se han confirmado con el desarrollo de la biología molecular clásica junto con las nuevas tecnologías de secuenciación y herramientas bioinformáticas. Un ejemplo clásico de exaptación de ETs para cumplir un rol funcional lo encontramos en los genomas de *Drosophila spp.*, en donde el retrotransposon HeT-A cumple un rol similar a la telomerasa en el mantenimiento de los telómeros(34). Hay evidencia diversa que refuerza la idea de que los ETs pueden proveer secuencias reguladoras en *cis*(35); estudios de ChiP-Seq demuestran que los ETs contribuyen a un porcentaje considerable de sitios de unión a factores de transcripción(36,37), además hay familias de ETs que están enriquecidas en eventos de unión a factores de transcripción(32,37–39). Algunos SINEs contienen sitios de unión que actúan como aisladores, alterando los patrones de formación de heterocromatina(32), mientras que otros (L1) forman ensamblados eucromatínicos(35).

A partir de su transcripción, los ETs también pueden actuar como reguladores en *trans*, produciendo miRNAs, lncRNAs y sufriendo trans-splicing; en *L. major*, el ET no autónomo SIRE participa en la regulación post-transcripcional de los ARNm, en *H. sapiens* el SINE Alu/B1 que codifica para un lncRNA que regula los niveles transcripcionales mediante un mecanismo de *RNA decoy*(35), y el uso de secuencias derivadas de ETs de ADN para la modulación de vida media de ARNm mediada por miRNAs en algunas plantas(40).

En mamíferos, los ETs tienden a agruparse cerca de genes involucrados en el desarrollo y la regulación de la transcripción (33). En un ensayo de ChIP-Seq comparativo con 26 factores de transcripción ortólogos entre dos líneas celulares comparables de ratón y humano, más del 98% de los picos derivados de ETs fueron especie específico(37), indicando un rol en el establecimiento de redes regulatorias linaje específicas.

Otro descubrimiento interesante es que algunas características asociadas a la domesticación de plantas y animales son producto de selección artificial que favorecen la presencia de secuencias reguladoras en *cis* provenientes de ETs(41–44).

Vale la pena aclarar el hecho de que la conservación filogenética y a nivel de secuencia de elementos regulatorios originadas de ETs no son suficientes para determinar su importancia funcional. En general es necesario hacer estudios de *knock-out* para determinar si verdaderamente la ausencia de estos elementos tienen un efecto fenotípico apreciable(32).

Es de suma importancia estudiar los elementos transponibles, ya que son un factor esencial a tomar en cuenta en el estudio de la evolución de los genomas, alterando la regulación transcripcional, post-transcripcional, la formación de heterocromatina, y la organización genómica, además de ser modeladores de redes regulatorias linaje y especie específicas.

Elementos transponibles en *Trypanosoma cruzi*

Hasta el momento se han reportado cinco retrotransposones en el genoma de *Trypanosoma cruzi*, dos del orden LINE; L1Tc y CZAR, dos del orden SINE; NARTc y SIRE, y uno del orden DIRS; VIPER (45–47).

L1Tc es el retroelemento más estudiado de *Trypanosoma cruzi*, de ~4.9 kb y con especificidad de inserción(48). Con 320 copias estimadas(5), se encuentra ampliamente distribuido en el genoma. Se encuentra flanqueado por duplicaciones de sitio de inserción de aproximadamente 12 pb. Posee un promotor dependiente de pol II hacia su extremo 5' (49) y un tramo poly-A hacia su extremo 3'.

Su único ORF de 4722 nt codifica para una proteína con dominios AP endonucleasa, transcriptasa reversa, RNAsa H, y un dominio de unión a ADN hacia el extremo C-terminal.

Hay evidencia abundante de que L1Tc está activo; se ha evidenciado la presencia de transcriptos en ensayos *in vitro*(49) y de actividades endonucleasa AP, chaperona de ácidos nucleicos y transcriptasa reversa en la proteína que codifica su ORF(50–52)

L1Tc pertenece a la familia de retrotransposones *ingi*. Esta familia abarca retrotransposones presentes en los genomas de Tripanosomátidos. A partir de un análisis filogenético con el dominio RT, Bringaud y colaboradores definen seis subclados; en el subclado más basal, *ingi1*, L1Tc agrupa junto con L1Tco (*T. congolens*e), y retrotransposones inactivos de *L. brasiliensis* (LbDIRE), los otros cinco subclados (*ingi2-6*) contienen los retrotransposones *ingi* y DIRE ("degenerate *ingi*/L1Tc-related elements") de *T. brucei*, *T. congolense*, *T. vivax*, *L. major*, y *L. brasiliensis* (15,48,50,53,54).

Como ejemplo de "parasitismo anidado", tenemos a NARTc ("non-autonomous retrotransposon in *T. cruzi*"), un retroelemento no autónomo, que depende de la maquinaria de transposición de L1Tc para su replicación. De un largo de ~0.26 kb y con 133 copias estimadas(5), NARTc posee cierto grado de homología con L1Tc y ambos retroelementos poseen las mismas duplicaciones de sitio de inserción, sugiriendo el mismo mecanismo de transposición.

L1Tc y NARTc poseen una homología del 100% en las primeras 77 bases. Desde el nucleótido 78 al 252, NARTc posee una secuencia que ha divergido bastante de L1Tc, con un 52% de homología. Hacia el extremo 3' posee una secuencia poly-A al igual que L1Tc.

De forma similar a L1Tc, el elemento *ingi* de *T. brucei* posee un compañero no autónomo, RIME. Considerando la similaridad entre *ingi* y L1Tc, y las relaciones *ingi*/RIME y L1Tc/NARTc, se propusieron dos escenarios evolutivos posibles; el primero es un escenario en el cual los elementos no autónomos se formaron a partir de

deleciones de sus compañeros autónomos, el segundo escenario posible es que tanto RIME como NARTc se hayan originado a partir del ancestro común de los elementos *ingi* y L1Tc (54).

El retrotransposon CZAR, reportado en 1991 por Villanueva y colaboradores, es un elemento móvil de ~7.9 kb y entre 30 y 40 copias estimadas. Posee dos ORFs; el primero, de 1158 nt codifica para una proteína con dominio de unión a ADN, el segundo ORF, de 3951 nt codifica para una proteína con un dominio transcriptasa reversa. Este elemento posee la particularidad de tener especificidad de inserción por gen de Spliced Leader RNA (SL), insertándose en la posición 11 del exón, generando DSI de 22 nucleótidos tras su inserción. Estas DSI fueron reportadas como muy conservadas en las copias caracterizadas, sugiriendo una actividad de retrotransposición reciente. En su extremo 5', corriente arriba del primer ORF, posee 2.5 copias de un repetido en tándem (TR) de 185 pb. Hacia el extremo 3', posee un tramo polyA, de aproximadamente 42 nucleótidos(55).

SLACS y CRE1, retrotransposones de *Trypanosoma brucei* y *Crithidia fasciculata*, respectivamente, comparten la particular especificidad de inserción. Estos elementos móviles poseen DSI de entre 22 y 49 nucleótidos, particularmente largas considerando que la mayoría de los retrotransposones no-LTR producen DSI de entre 4 y 14 nucleótidos. Sus ORFs muestran un porcentaje de similaridad considerable. Además, están presentes en bajo número de copias en sus genomas hospederos, y están todos asociados con los tándem de SL. Estas características compartidas sugieren fuertemente que son elementos cercanamente emparentados (56,57).

Trypanosoma cruzi posee un único representante del orden DIRS; VIPER ("vestigial interposed retroelement"). VIPER tiene un largo de 4480 pb y 275 copias estimadas(5). A partir de las secuencias de varias copias de VIPER, Lorenzi y colaboradores reconstruyeron tres ORFs. Este procedimiento fue necesario para determinar el orden y naturaleza de los ORFs ya que todos presentaban múltiples codones STOP. El primer ORF posee un dominio gag, el segundo, de 842 nt, posee un dominio YR, y el tercero, de 1470 nt, posee dominios transcriptasa reversa y RNasa H. Estudios de su distribución en el cromosoma tres de *T. cruzi* (TcCh3) indican que el elemento se localiza frecuentemente en regiones que contienen otros retroelementos como L1Tc y DIRE (57). Secuencias homólogas a VIPER han sido encontradas en *Trypanosoma brucei* y *Trypanosoma vivax* (58).

VIPER posee un compañero no autónomo denominado SIRE ("short interspersed repetitive element"), de ~0.42 kb y 480 copias estimadas. Presentan una alta homología en sus extremos 5' y 3', el extremo 3' de SIRE posee parte del dominio RNasa H. SIRE posee tres regiones altamente conservadas; dos hacia sus extremos 5' y 3', y una entre los nucleótidos 190 y 260, con alto contenido GC. Hacia su extremo 3' posee un sitio AG aceptor de splicing. SIRE ha sido encontrado en los 3' UTR de histona H2A, 2-hidroxiácido deshidrogenasa, oligopeptidasa, alfa-manosidasa

lisosomal, aldehído deshidrogenasa y en otros ORF no identificados. Además su secuencia codifica para el extremo C-terminal de una RNAsa H y un ORF de función desconocida(59). Esta evidencia, sumada al hecho de que SIRE presenta una alta conservación de secuencia a lo largo de las distintas DTUs indican fuertemente un rol funcional en la biología del parásito(60).

Secuenciación y ensamblado de datos NGS

La secuenciación y ensamblaje del genoma humano en 2001 dio paso a una nueva era en los estudios genómicos (61). Distintas organizaciones comenzaron a desarrollar tecnologías de secuenciación y software de ensamblado eficientes para poder dar sentido a la gran cantidad de información que se encuentra en los genomas de los seres vivos.

La mayoría de los organismos eucariotas poseen secuencias repetitivas en su genoma en distintas cantidades. Estas secuencias repetitivas suelen ser repetidos en tándem, familias multigénicas, duplicaciones segmentales, y elementos transponibles. Tecnologías de Next Generation Sequencing de reads cortos proveen reads de 50 a 900 pb (62). Ya que las secuencias repetitivas suelen tener tamaños mucho más grandes, los reads cortos no son capaces de resolver por completo estas secuencias. Los algoritmos diseñados para el ensamblaje de estos reads cortos(basados en grafos de de Bruijn, u *Overlapping Consensus Motifs*) suelen colapsar todos los reads provenientes de secuencias en unas pocas copias debido a la alta identidad de los reads, subestimando el número y variabilidad de las secuencias repetitivas. Otros problemas que surgen son la fragmentación y el mal ensamblado de contigs, que tienen como consecuencia bajos valores de N50 y errores en el orden de genes y sintenia; estos problemas afectan la calidad de los genomas y los análisis subsecuentes. Estos obstáculos que presentan las secuencias repetitivas han podido ser parcialmente eludidos gracias a protocolos de secuenciación que permiten estimar las distancias y el ligado entre contigs (mate pair sequencing), y que hacen posible la preparación de reads largos sintéticos, como la preparación de cromosomas bacterianos artificiales (BACs) en conjunto con DNA barcoding (63).

El advenimiento de tecnologías de secuenciación de reads largos logró saltar en mayor medida el problema de las secuencias repetitivas; con PacBio SMRT sequencing es posible obtener una longitud máxima de reads de 60 kb y un promedio de 10 kb (64), y con Oxford Nanopore una longitud máxima de 1 Mb (65). Con estas tecnologías, muchas secuencias repetidas pueden ser determinadas en su totalidad con baja ambigüedad, además de obtener sus secuencias flanqueantes, que podrían brindar información relevante para comprender la importancia de la organización genómica.

Usando un genoma secuenciado con esta tecnología, es posible caracterizar copias

específicas de elementos transponibles de interés. Se podría en teoría trazar su patrón de amplificación y distribución genómica, determinar especificidad de inserción, patrones de degradación, y determinar la localización genómica específica de elementos transponibles potencialmente activos.

Materiales y métodos

Datos

Los genomas y archivos de anotación de Trypanosoma cruzi TCC y Dm28c ensamblados a partir de reads PacBio e Illumina fueron proveídos por Berná y colaboradores. Los genomas de las cepas Esmeraldo, Sylvio y Bug2148 fueron obtenidos de NCBI, assembly ids: GCA_000327425.1, GCA_000188675.2, GCA_002749415, respectivamente. El genoma de la cepa CLBrener separado por haplotipos fue obtenido de TriTryp (release 34); bajo los identificadores TcruziCLBrener, TcruziCLBrenerEsmeraldo-like y TcruziCLBrenerNon-Esmeraldo-like.

Las secuencias de referencia de los elementos transponibles fueron descargadas de la base de datos de NCBI, números de acceso: CZAR; M62862.1, L1Tc; AF208537.2, NARTc; AF215898.1, VIPER; AY747608.1, SIRE; ver Anexo.

Software

Búsqueda por similaridad

La búsqueda de ET se realizó con el software BLAST+ (ver. 2.6.0)(66). Se construyeron bases de datos de nucleótidos con el comando:

```
makeblastdb -input_type fasta -dbtype nucl
```

Este comando construye una base de datos de nucleótidos a partir de un archivo fasta. Dicha base de datos es utilizada por el comando blastn, lo cual permite un acceso rápido y eficiente al gran número de secuencias.

El comando utilizado para la búsqueda fue:

```
blastn          -db $database \
                   -query $reference \
                   -out blastnout/${strain}_${element}_blastn.out \
                   -evalue 1e-5 \
                   -outfmt "6 qseqid sseqid qstart qend sstart send  length pident qcovhsp sstrand evalue"\
                   -num_threads 5
```

El cual toma un archivo fasta y la base de datos de nucleótidos como input. El archivo fasta contiene la secuencia de referencia (query) a buscar en la base de datos (subject). La opción `-outfmt` determina que el output del programa sea una tabla personalizada que contenga información sobre los *High-scoring Segment Pairs* (HSPs) encontrados en la base de datos. Estos HSPs son subsecuentemente filtrados con el script `blastout_to_gff.py` segun su largo y porcentaje de identidad del *query*; en este caso se utilizó un criterio de 70% de ambos para obtener los ETs menos degradados y de esta manera trabajar con secuencias más informativas.

El hecho de que esta tabla personalizada posea información sobre la hebra en la cual se encuentra cada ET (parámetro `sstrand`), permite obtener las secuencias de los ET de la hebra sentido, necesaria para los subsecuentes análisis.

Búsqueda de dominios conservados en ORFs

A partir de los archivos fasta obtenidos en la búsqueda de ET, se buscaron ORFs con el software getorf de EMBOSS suite (ver. 6.6.0.0)(67). Este programa toma como *input* una secuencia en formato fasta, y otros parámetros que determinan el criterio que define un ORF. Los parámetros utilizados fueron:

```
getorf -find 1 -reverse 0, -minsize 600 ,-maxsize 6000
```

los cuales determinan que sólo sean devueltos los ORFs que comienzan con un codón START (AUG) y terminan en un codón STOP (UAA, UAG o UGA), en la hebra sentido, con un número de nucleótidos mínimo de 600 nt y máximo de 6000 nt. La salida del programa es un archivo fasta con las secuencias aminoacídicas de los ORFs encontrados.

Para la búsqueda de dominios conservados, se utilizó la plataforma Web Batch Conserved Domain Search de NCBI (CD-Search)(68), que predice la presencia de dominios conservados en un conjunto de secuencias aminoacídicas. La predicción es realizada comparando la secuencia problema con bases de datos curadas que contienen secuencias consenso para los distintos dominios conservados a lo largo del árbol evolutivo. En esta búsqueda se utilizó la base de datos Conserved Domain Database (69).

Alineamiento de secuencias

El alineamiento de secuencias se realizó con el software MAFFT (ver. 7.222)(70), utilizando la opción:

```
mafft --auto
```

que indica al programa que elija un algoritmo de alineamiento óptimo dado el tamaño de los datos y la complejidad de las secuencias. MAFFT toma como input un archivo multifasta, y devuelve un archivo multifasta con las secuencias alineadas.

Debido al gran número de secuencias involucradas, se utilizó TrimAl (ver. 1.4.rev22) (71), un programa que automatiza la curación de alineamientos múltiples. Se utilizaron las opciones:

```
trimal -gt 0.9 -cons 60
```

donde -gt 0.9 especifica que se conserven solo las columnas con gaps en menos del 10% de las posiciones. En caso de que este procesamiento devuelva un alineamiento reducido a menos del 60% de sus columnas, la opción -cons 60 especifica que se conserven las mejores columnas para llegar al 60% de las columnas del alineamiento original.

La visualización de alineamientos múltiples se realizó con Jalview (ver. 2.10.1)(72)

Construcción de filogenias

La construcción de árboles filogenéticos realizó con el software MUSCLE (ver. 3.8.31) (73) implementando el algoritmo Neighbor Joining (74) con las opciones:

```
muscle -maketree -cluster neighborjoining
```

MUSCLE toma como input un archivo multifasta alineado, y devuelve un archivo de árbol filogenético en formato Newick.

Scripts

Para un uso automatizado y organizado de los datos, se escribieron scripts en BASH. Además, se escribieron scripts en Python (ver. 2.7.12) que permitieron un uso personalizado de las secuencias, datos de anotación y archivos intermedios. Los gráficos finales y reportes intermedios se realizaron con R (ver. 3.4.2), utilizando los paquetes ape (ver. 5.0), ggtree (ver. 1.10.4), tidyverse (ver. 1.2.1), y knitr (ver. 1.19) (75–79).

Debajo se encuentra una lista de los scripts creados con un breve resumen de su funcion, su código se encuentra en el repositorio de github https://github.com/gaxyz/scripts-tesina.

Python

blastout_to_gff.py: Convierte un output de blast tabular en un archivo de anotación gff.

fasta_extract_from_gff.py: Extrae secuencias correspondientes a un archivo de anotación de un archivo multifasta.

retroelements_stats.py: Calcula caracteristicas generales de las secuencias brindadas como input (número de secuencias, contenido gc promedio, largo promedio, identidad promedio).

run_getorf.py: Wrapper de getorf. Usa getorf en un archivo multifasta y asigna identificadores comprensibles.

cdsearch_parse.py: Procesa un archivo de output de Batch CD Search y genera una tabla que especifica si un elemento dado es potencialmente activo.

get_up_and_downstream_annotations.py: Determina anotaciones corriente arriba y corriente abajo de un conjunto de features en un archivo gff. Es necesario proveer el archivo gff correspondiente al genoma con información redundante.

R

build_two_strain_trees.R: Construcción de árboles de dos cepas a partir de un archivo formato Newick y una tabla con observaciones para cada copia.

cis_annotations.R: Filtra tablas de anotaciones en cis a partir de un umbral de proximidad.

full_table_construction.R: Construye tablas de observaciones de cada copia a partir de tablas de observaciones individuales.

all_trees.R: Construcción de árboles a partir de un archivo formato Newick y una tabla con observaciones para cada copia.

Resultados y discusión

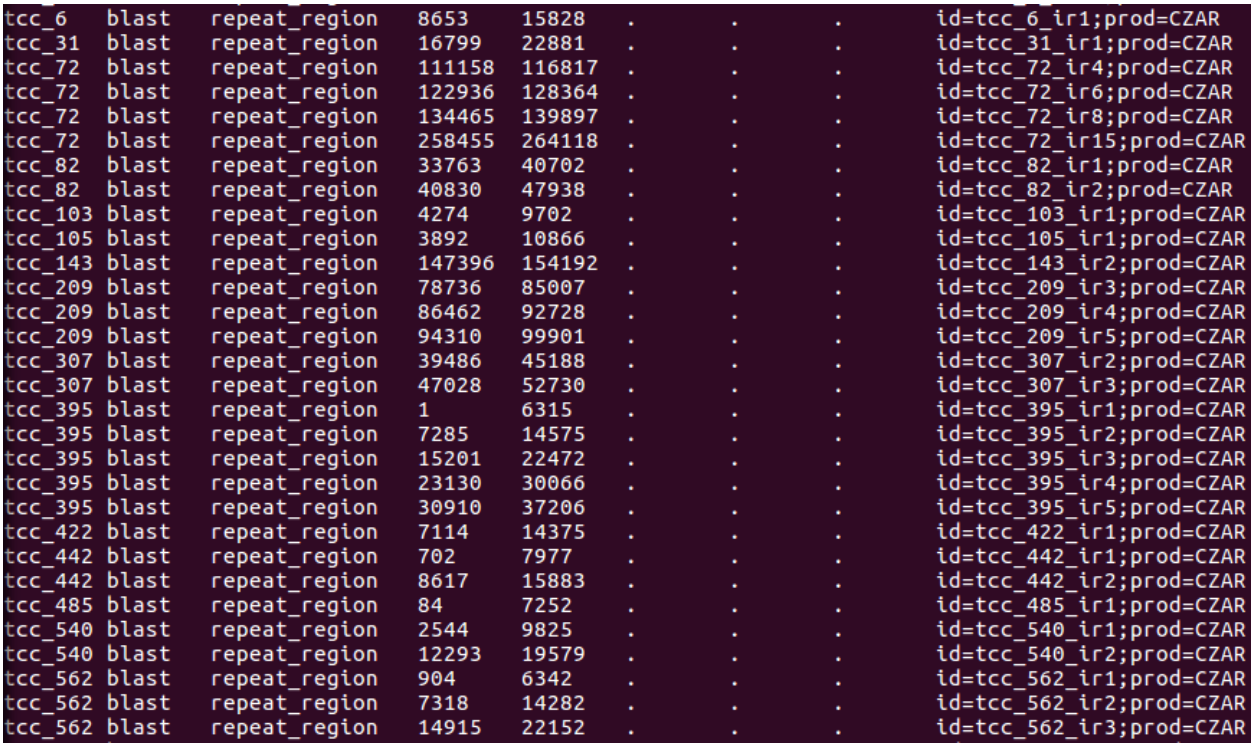
Búsqueda y extracción de secuencias

La búsqueda de secuencias de retroelementos se realizó mediante búsqueda por similaridad (BLAST). El resultado de la búsqueda es una tabla que contiene información sobre cada HSP; el contig en el cual se encuentra y sus coordenadas, la región de la referencia con la cual posee similaridad, la longitud del HSP, el porcentaje de identidad, porcentaje de cobertura y e-value (Figura 5). Estos datos fueron utilizados para generar un archivo de anotación de formato General Feature Format (.gff) con el script `blastout_to_gff.py`, el cual filtra la tabla resultante de la búsqueda de blast para HSPs con porcentaje de identidad y cobertura mayor al 70% con respecto a la referencia y reorganiza la información para adecuarla al formato de anotación (Figura 6). A partir del archivo de anotación generado, se extrajeron las secuencias nucleotídicas correspondientes del genoma ensamblado con el script `fasta_extract_from_gff.py` y se escribieron en un archivo fasta para su uso en subsecuentes análisis.

```
M62862.1        tcc_442 1018    1131    9425    9538    114     92.105  2       plus    1.26e-37
M62862.1        tcc_442 1018    1131    1499    1613    115     91.304  2       plus    5.86e-36
M62862.1        tcc_442 1       43      32      74      43      93.023  1       plus    3.66e-08
M62862.1        tcc_442 1       43      16501   16543   43      93.023  1       plus    3.66e-08
M62862.1        tcc_442 1       43      17116   17158   43      93.023  1       plus    3.66e-08
M62862.1        tcc_442 1       43      17731   17773   43      93.023  1       plus    3.66e-08
M62862.1        tcc_442 1       43      18345   18387   43      93.023  1       plus    3.66e-08
M62862.1        tcc_442 1       43      19508   19550   43      93.023  1       plus    3.66e-08
M62862.1        tcc_442 1       43      20110   20152   43      93.023  1       plus    3.66e-08
M62862.1        tcc_442 1       43      20722   20764   43      93.023  1       plus    3.66e-08
M62862.1        tcc_442 1       43      21334   21376   43      93.023  1       plus    3.66e-08
M62862.1        tcc_442 1       43      21896   21938   43      93.023  1       plus    3.66e-08
M62862.1        tcc_442 1       43      22506   22548   43      93.023  1       plus    3.66e-08
M62862.1        tcc_442 1       43      23119   23161   43      93.023  1       plus    3.66e-08
M62862.1        tcc_442 1       43      23685   23727   43      93.023  1       plus    3.66e-08
M62862.1        tcc_442 1       43      24301   24343   43      93.023  1       plus    3.66e-08
M62862.1        tcc_442 1       43      24918   24960   43      93.023  1       plus    3.66e-08
M62862.1        tcc_442 1       43      25535   25577   43      93.023  1       plus    3.66e-08
M62862.1        tcc_442 1       43      26150   26192   43      93.023  1       plus    3.66e-08
M62862.1        tcc_442 1       43      26765   26807   43      93.023  1       plus    3.66e-08
M62862.1        tcc_442 1       43      27380   27422   43      93.023  1       plus    3.66e-08
M62862.1        tcc_442 1       43      27996   28038   43      93.023  1       plus    3.66e-08
M62862.1        tcc_442 1       43      28610   28652   43      93.023  1       plus    3.66e-08
M62862.1        tcc_442 1       43      29225   29267   43      93.023  1       plus    3.66e-08
M62862.1        tcc_442 1       43      29839   29881   43      93.023  1       plus    3.66e-08
M62862.1        tcc_442 1       43      30450   30492   43      93.023  1       plus    3.66e-08
M62862.1        tcc_442 1       32      18950   18981   32      100.000 0       plus    4.73e-07
M62862.1        tcc_682 42      7270    16471   9216    7318    90.763  99      minus   0.0
M62862.1        tcc_682 1030    1167    15659   15522   138     86.957  2       minus   5.86e-36
M62862.1        tcc_682 1       43      640     598     43      93.023  1       minus   3.66e-08
```

Figura 5. Ejemplo de archivo de output de BLAST tabular, visualizado en terminal de

Linux. De izquierda a derecha: identificador de query, identificador de subject, comienzo de HSP en query, final de HSP en query, comienzo de HSP en subject, final de HSP en subject, longitud de HSSP, porcentaje de identidad, cobertura del query, sentido de la hebra, e-value.



tcc_6	blast	repeat_region	8653	15828	.	.	.	id=tcc_6_ir1;prod=CZAR
tcc_31	blast	repeat_region	16799	22881	.	.	.	id=tcc_31_ir1;prod=CZAR
tcc_72	blast	repeat_region	111158	116817	.	.	.	id=tcc_72_ir4;prod=CZAR
tcc_72	blast	repeat_region	122936	128364	.	.	.	id=tcc_72_ir6;prod=CZAR
tcc_72	blast	repeat_region	134465	139897	.	.	.	id=tcc_72_ir8;prod=CZAR
tcc_72	blast	repeat_region	258455	264118	.	.	.	id=tcc_72_ir15;prod=CZAR
tcc_82	blast	repeat_region	33763	40702	.	.	.	id=tcc_82_ir1;prod=CZAR
tcc_82	blast	repeat_region	40830	47938	.	.	.	id=tcc_82_ir2;prod=CZAR
tcc_103	blast	repeat_region	4274	9702	.	.	.	id=tcc_103_ir1;prod=CZAR
tcc_105	blast	repeat_region	3892	10866	.	.	.	id=tcc_105_ir1;prod=CZAR
tcc_143	blast	repeat_region	147396	154192	.	.	.	id=tcc_143_ir2;prod=CZAR
tcc_209	blast	repeat_region	78736	85007	.	.	.	id=tcc_209_ir3;prod=CZAR
tcc_209	blast	repeat_region	86462	92728	.	.	.	id=tcc_209_ir4;prod=CZAR
tcc_209	blast	repeat_region	94310	99901	.	.	.	id=tcc_209_ir5;prod=CZAR
tcc_307	blast	repeat_region	39486	45188	.	.	.	id=tcc_307_ir2;prod=CZAR
tcc_307	blast	repeat_region	47028	52730	.	.	.	id=tcc_307_ir3;prod=CZAR
tcc_395	blast	repeat_region	1	6315	.	.	.	id=tcc_395_ir1;prod=CZAR
tcc_395	blast	repeat_region	7285	14575	.	.	.	id=tcc_395_ir2;prod=CZAR
tcc_395	blast	repeat_region	15201	22472	.	.	.	id=tcc_395_ir3;prod=CZAR
tcc_395	blast	repeat_region	23130	30066	.	.	.	id=tcc_395_ir4;prod=CZAR
tcc_395	blast	repeat_region	30910	37206	.	.	.	id=tcc_395_ir5;prod=CZAR
tcc_422	blast	repeat_region	7114	14375	.	.	.	id=tcc_422_ir1;prod=CZAR
tcc_442	blast	repeat_region	702	7977	.	.	.	id=tcc_442_ir1;prod=CZAR
tcc_442	blast	repeat_region	8617	15883	.	.	.	id=tcc_442_ir2;prod=CZAR
tcc_485	blast	repeat_region	84	7252	.	.	.	id=tcc_485_ir1;prod=CZAR
tcc_540	blast	repeat_region	2544	9825	.	.	.	id=tcc_540_ir1;prod=CZAR
tcc_540	blast	repeat_region	12293	19579	.	.	.	id=tcc_540_ir2;prod=CZAR
tcc_562	blast	repeat_region	904	6342	.	.	.	id=tcc_562_ir1;prod=CZAR
tcc_562	blast	repeat_region	7318	14282	.	.	.	id=tcc_562_ir2;prod=CZAR
tcc_562	blast	repeat_region	14915	22152	.	.	.	id=tcc_562_ir3;prod=CZAR

Figura 6. Ejemplo de archivo de anotación gff, visualizado en terminal de Linux. De izquierda a derecha: contig, programa utilizado para anotación, tipo de feature, coordenada de inicio de feature, coordenada de fin de feature, puntaje, sentido de la hebra, marco, atributos de anotación.

Con el objetivo de obtener características generales de las secuencias extraídas, se escribió un script (retroelements_stats.py) que calcula el número de retroelementos encontrados, longitud, contenido GC, y porcentaje de identidad promedios. El cálculo de porcentaje de identidad promedio se realiza mediante blast “todos contra todos”. Los resultados obtenidos se encuentran resumidos en la Tabla 1.

El número de secuencias encontradas para elementos de familias autónomas (CZAR, L1Tc, VIPER) es considerablemente más bajo que el de familias no autónomas como SIRE y NARTc(Tabla 1). CZAR posee 43 copias en el genoma de TCC, y 57 copias en el genoma de Dm28c, lo cual concuerda con los primeros estudios de caracterización molecular de CZAR, donde se estimó que su número de copias se encontraba cerca de 40. En cambio, para L1Tc (43 en TCC, 54 en Dm28c), NARTc (110, 55) y VIPER (244, 194) se encontró que su número de copias es menor a estimativos *in silico* previos (5). Atribuimos esta discrepancia a los umbrales conservativos de la búsqueda. El número

de copias encontradas de SIRE fue de 851 en TCC y 669 en Dm28c, un número considerablemente mayor a estimativos previos.

Dado que TCC es una cepa heterocigota, es esperable que se encuentre un mayor número de copias al considerar que es posible diferenciar haplotipos en algunas regiones genómicas una vez realizado el ensamblado. Esta predicción se cumple para NARTc, VIPER y SIRE, en cambio, hay un mayor número de copias de CZAR y L1Tc en el genoma de Dm28c. Cabe la posibilidad de que estos últimos retrotransposones se encuentren en regiones del genoma de TCC que no pudieron separarse en sus respectivos haplotipos, o, de manera no excluyente, que el software de ensamblado haya colapsado más copias en el genoma de TCC.

Las longitudes medias encontradas son similares a las reportadas en la bibliografía, aunque en general de menor tamaño, lo cual se debe a que muchas de las copias encontradas son retrotransposones degradados. Las identidades medias son altas como es de esperar al usar un umbral de identidad alto, todos los valores se encuentran cerca del 90%. El contenido GC de las secuencias extraídas se encuentra cerca contenido GC genómico promedio (55%)(5), con la excepción de las secuencias de SIRE, que poseen un contenido GC más bajo, alrededor de 44%. Se ha sugerido que SIRE podría cumplir un rol funcional en el genoma de *Trypanosoma cruzi* (59). La diferencia marcada en el contenido GC se podría deber a una colonización reciente; el tiempo transcurrido no fue suficiente para volver al valor promedio del genoma. Una explicación alternativa sería la presencia de presiones selectivas que restringen la regresión al valor promedio.

Tabla 1. Características generales del resultado de la búsqueda de ETs en los genomas de TCC y Dm28c.

ET	Cepa	N° copias	Largo medio (nt)	Identidad media (%)	%GC media
CZAR	TCC	43	6496	93	55
	Dm28c	57	6441	91	56
L1Tc	TCC	43	4873	90	53
	Dm28c	54	4748	96	53
NARTc	TCC	110	256	92	51
	Dm28c	55	258	90	51
VIPER	TCC	244	3454	84	55
	Dm28c	194	3422	87	54
SIRE	TCC	851	440	87	44
	Dm28c	669	441	88	44

Búsqueda de retroelementos potencialmente activos

Siguiendo la definición de copia autónoma de Wicker y colaboradores, consideramos que una copia de un elemento es autónoma si sus ORFs codifican todos los dominios necesarios para su transposición. Este criterio nos permite reducir el número de candidatos autónomos mediante métodos bioinformaticos, mientras que para confirmar su actividad de retrotransposición es necesario recurrir a ensayos experimentales.

Con el objetivo de encontrar copias autónomas en nuestro set de datos, se realizó una predicción *ab initio* de ORFs en las secuencias de CZAR, L1Tc y VIPER, y una subsecuente búsqueda de dominios conservados en las secuencias proteicas de estos ORFs. La búsqueda de ORFs en las secuencias de los retroelementos se realizó con el programa getorf de EMBOSS. La salida del programa es un archivo fasta con las secuencias aminoacídicas de los ORFs encontrados.

Se escribió un *wrapper* en python (`run_getorf.py`) que hizo posible editar los *headers* para mantener el identificador de cada secuencia y el número de ORFs encontrados, y

automatizar la búsqueda de ORFs para todas las secuencias.

Una vez obtenidas las secuencias aminoacídicas, se utilizó la plataforma Web Batch Conserved Domain Search de NCBI (CD-Search) para predecir la presencia de dominios conservados.

Se descargaron las tablas resultantes de CD-Search (Figura 7), y se escribió un script para parsear los resultados y determinar copias potencialmente autónomas (cdsearch_parse.py). Los hits con dominios parciales fueron descartados. Los criterios utilizados para definir un ET como potencialmente autónomo fueron formulados en base a la literatura disponible de cada uno de ellos (53,55,58,59),

Se consideró que una copia de CZAR es autónoma si posee por lo menos un ORF que contenga un dominio transcriptasa reversa. En el caso de L1Tc, se considera autónoma si posee al menos un dominio transcriptasa reversa, RNasa H y endonucleasa entre todos sus ORFs. Por último, para VIPER se consideró que una copia es potencialmente autónoma si posee al menos un dominio transcriptasa reversa, RNAsa H y tirosina recombinasa entre todos sus ORFs.

Query	Hit type	PSSM-ID	From	To	E-Value	Bitscore	Accession	Short name	Incomplete	Superfamily
Q#1 - >tcc_266_ir9_ORF1	specific	238827	636	898	4.79008e-56	192891	cd01650	RT_nLTR_like	-	cl02808
Q#1 - >tcc_266_ir9_ORF1	superfamily	295487	636	898	4.79008e-56	192891	cl02808	RT_like superfamily	-	-
Q#1 - >tcc_266_ir9_ORF1	non-specific	306564	644	883	2.18221e-30	118558	pfam00078	RVT_1	-	cl26764
Q#1 - >tcc_266_ir9_ORF1	superfamily	331585	644	883	2.18221e-30	118558	cl26764	RVT_1 superfamily	-	-
Q#1 - >tcc_266_ir9_ORF1	non-specific	197311	54	271	9.28578e-23	96.9772	cd09077	R1-I-EN	-	cl00490
Q#1 - >tcc_266_ir9_ORF1	superfamily	321002	54	271	9.28578e-23	96.9772	cl00490	EEP superfamily	-	-
Q#1 - >tcc_266_ir9_ORF1	specific	308788	69	264	3.3865e-16	78.8676	pfam03372	Exo_endo_phos	-	cl00490
Q#1 - >tcc_266_ir9_ORF1	non-specific	316995	153	264	6.63751e-15	71633	pfam14529	Exo_endo_phos_2	-	cl00490
Q#1 - >tcc_266_ir9_ORF1	non-specific	238828	656	887	1.10151e-14	74.1596	cd01651	RT_G2_intron	-	cl02808
Q#1 - >tcc_266_ir9_ORF1	non-specific	197310	60	191	8.01607e-12	65.8357	cd09076	L1-EN	C	cl00490
Q#1 - >tcc_266_ir9_ORF1	non-specific	275209	740	872	1.81622e-08	57.0824	TIGR04416	group_II_RT_mat	NC	cl26764
Q#1 - >tcc_266_ir9_ORF1	non-specific	197306	56	271	8.38732e-07	50.9429	cd08372	EEP	-	cl00490
Q#1 - >tcc_266_ir9_ORF1	non-specific	238185	797	873	2.77445e-05	43.4936	cd00304	RT_like	-	cl02808
Q#1 - >tcc_266_ir9_ORF1	non-specific	197314	30	194	4.27746e-05	45.7975	cd09080	TDP2	C	cl00490
Q#1 - >tcc_266_ir9_ORF1	non-specific	238824	741	874	4.96208e-05	44.2423	cd01646	RT_Bac_retron_I	C	cl02808
Q#1 - >tcc_266_ir9_ORF1	non-specific	225565	52	272	0.000586504	42.8136	COG3021	YafD	N	cl00490
Q#1 - >tcc_266_ir9_ORF1	non-specific	197336	72	193	0.000650194	42.2143	cd10281	Nape_like_AP-endo	C	cl00490
Q#1 - >tcc_266_ir9_ORF1	non-specific	197307	62	193	0.00125185	41.5045	cd09073	ExoIII_AP-endo	C	cl00490
Q#1 - >tcc_266_ir9_ORF1	non-specific	238825	830	874	0.0073623	38.3453	cd01647	RT_LTR	N	cl02808
Q#2 - >tcc_266_ir9_ORF2	non-specific	260008	160	288	1.88105e-14	69554	cd09276	Rnase_HI_RT_non_LTR	-	cl14782
Q#2 - >tcc_266_ir9_ORF2	superfamily	326352	160	288	1.88105e-14	69554	cl14782	RNase_H_like superfamily	-	-
Q#2 - >tcc_266_ir9_ORF2	non-specific	306562	157	290	1.95523e-08	52.7433	pfam00075	RNase_H	-	cl14782
Q#2 - >tcc_266_ir9_ORF2	non-specific	260012	170	290	9.3726e-08	50.6421	cd09280	RNase_HI_eukaryote_like	-	cl14782
Q#2 - >tcc_266_ir9_ORF2	non-specific	259998	160	288	1.39027e-06	46.9236	cd06222	RNase_H_like	-	cl14782
Q#2 - >tcc_266_ir9_ORF2	non-specific	260010	200	293	1.09245e-05	44.7811	cd09278	RNase_HI_prokaryote_like	N	cl14782
Q#2 - >tcc_266_ir9_ORF2	non-specific	223405	200	294	0.000586473	40.0324	COG0328	RnhA	N	cl14782
Q#3 - >tcc_199_ir2_ORF1	specific	238827	636	898	1.48776e-55	192506	cd01650	RT_nLTR_like	-	cl02808
Q#3 - >tcc_199_ir2_ORF1	superfamily	295487	636	898	1.48776e-55	192506	cl02808	RT_like superfamily	-	-
Q#3 - >tcc_199_ir2_ORF1	non-specific	306564	644	883	2.78859e-30	118943	pfam00078	RVT_1	-	cl26764
Q#3 - >tcc_199_ir2_ORF1	superfamily	331585	644	883	2.78859e-30	118943	cl26764	RVT_1 superfamily	-	-
Q#3 - >tcc_199_ir2_ORF1	non-specific	197311	54	271	3.82312e-22	95.8216	cd09077	R1-I-EN	-	cl00490

Figura 7. Ejemplo de archivo de output de CD Search, visualizado en libreoffice.

Los resultados se encuentran resumidos en la Tabla 2. El número de CZAR potencialmente autónomos encontrados fue de 25 y 34 en TCC y Dm28c, respectivamente. Estos números son un límite superior del número de CZAR activos, ya que el criterio de búsqueda requiere de la presencia de al menos un dominio RT entre sus ORFs . Se utilizó este criterio debido a que ninguno de los ORFs mostraron homología con dominios que no fueran RT, lo cual se contradice con evidencia de que posee un dominio de unión a ADN en su primer ORF(55).

En la búsqueda de L1Tc, se encontraron 26 y 30 elementos potencialmente autónomos en TCC y Dm28c, respectivamente.

No se encontraron elementos VIPER potencialmente autónomos, tal como había sido descrito en análisis previos(60).

Tabla 2. ETs potencialmente autonomos.

	CZAR		L1Tc	
	Totales	Activos	Totales	Activos
TCC	43	25	110	26
Dm28c	57	34	55	30

Análisis filogenético

Se realizaron alineamientos múltiples con las secuencias de los retrotransposones para ambas cepas en conjunto. Se utilizó el software de alineamiento MAFFT con la opción --auto. Los alineamientos resultantes poseían diversos gaps, y muchas de las secuencias estaban fragmentadas hacia sus extremos, por lo que se procedió a curarlos con el software TrimAl (Figura 8).

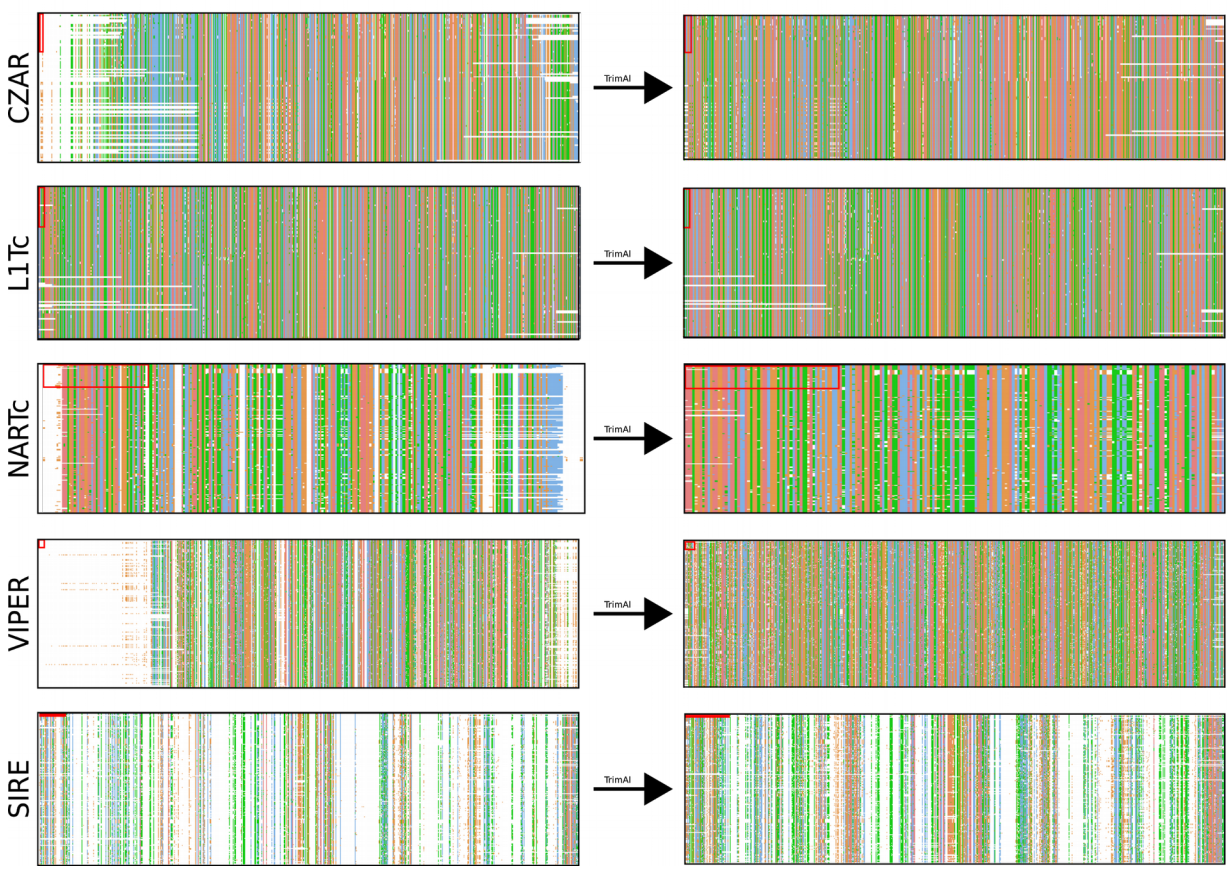


Figura 8. Alineamientos de retroelementos de TCC y Dm28c. Previos al curado (columna izquierda), y posteriores al curado con TrimAl (columna derecha).

A partir de los alineamientos múltiples curados de cada ET, para ambas cepas en conjunto, se construyeron árboles filogenéticos con el método Neighbor Joining implementado en el software MUSCLE.

CZAR

En el caso de CZAR (Figura 9), las secuencias se agrupan claramente por cepa, con la excepción de un clado en el cual se encuentran secuencias de ambas cepas. La estructura del árbol resultante podría ser explicada por eventos de amplificación de CZAR luego de la divergencia entre ambas cepas; el clado mixto representaría a las copias “ancestrales” (previas a la divergencia), y los clados cepa-específicos a las copias originadas en los eventos de amplificación. La determinación de CZAR potencialmente activos concuerda con este modelo de amplificación; aquellos sin dominios RT completos en sus ORFs se agrupan en el clado mixto; al ser más ancestrales, la degeneración de sus secuencias es mayor.

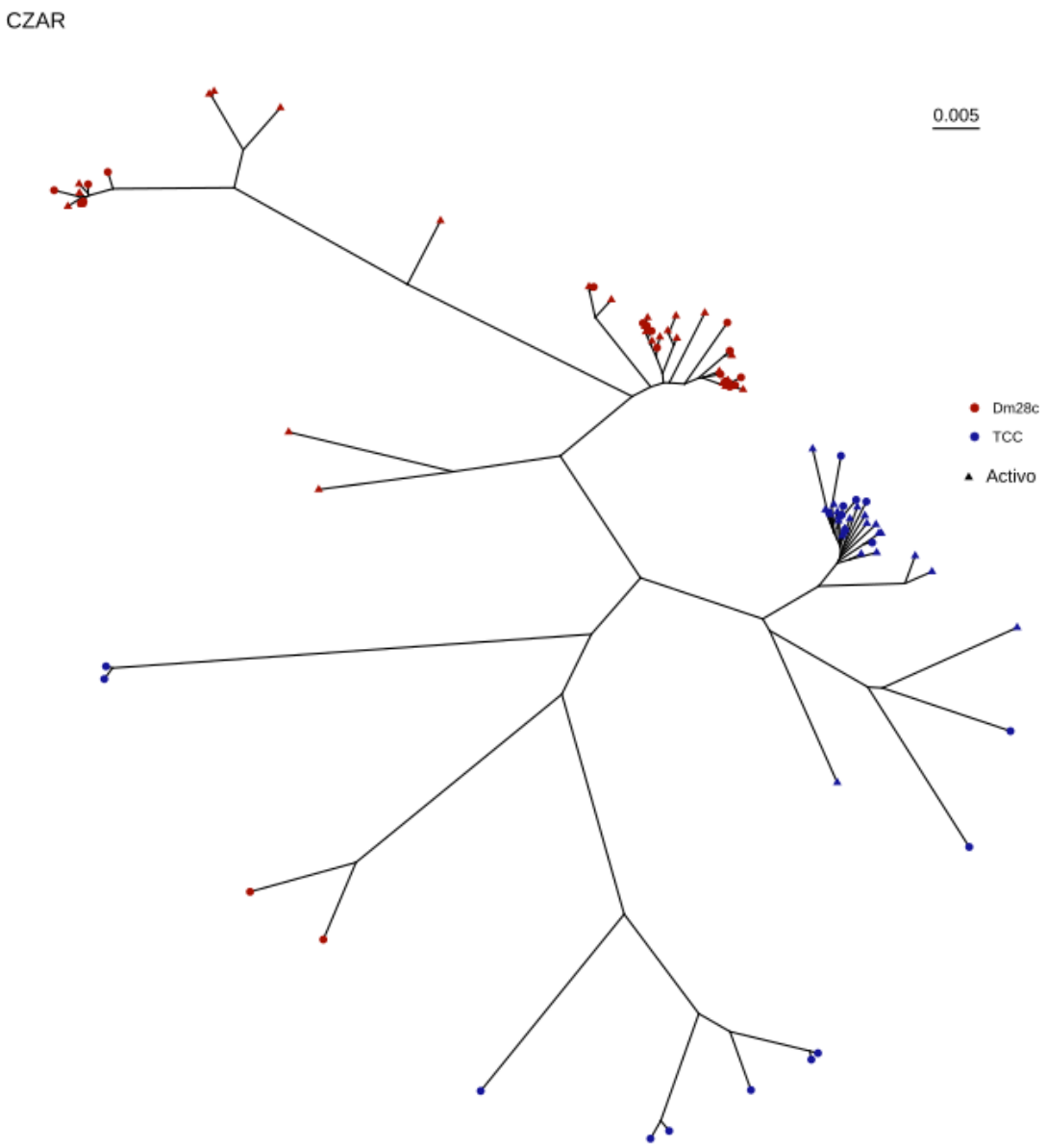


Figura 9. Filogenia generada por NJ de copias CZAR en TCC (azul) y Dm28c (rojo). Las copias potencialmente activas están representadas como triángulos.

L1Tc

La filogenia de L1Tc (Figura 10) presenta una estructura similar a CZAR en cuanto a su agrupación por cepa. La distribución de copias potencialmente activas no sigue ningún patrón particular, estando presentes en todos los clados. Las copias en el clado mixto podrían ser las copias previas a la divergencia entre las cepas, mientras que los clados cepa especifico corresponderían a las amplificaciones post-divergencia al igual que CZAR.

Teniendo en cuenta que de las cuatro copias de Dm28c presentes en el clado mixto, dos son activas y que además, todas presentan un porcentaje de identidad alto con respecto a TCC (98%), una explicación alternativa a esta topología podría ser un evento de transferencia horizontal entre las cepas.

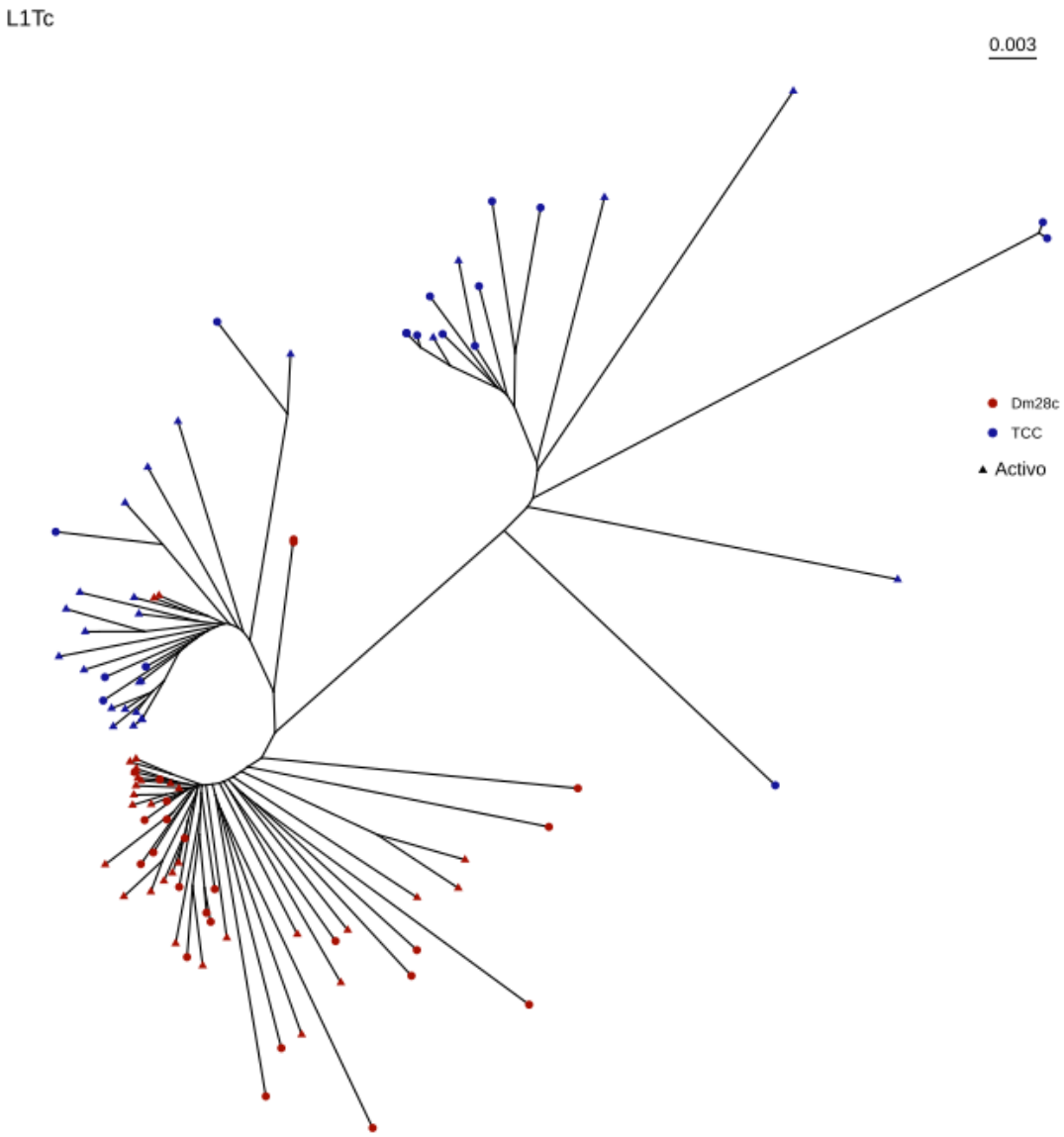


Figura 10. Filogenia de copias L1Tc en TCC (azul) y Dm28c (rojo). Las copias potencialmente activas están representadas como triángulos.

NARTc

Se podría esperar que NARTc siguiera un patrón similar a L1Tc; al ser el “compañero” no autónomo de L1Tc podría tomar ventaja de los eventos de amplificación. Los resultados (Figura 11) indican que esto no ha sucedido, ya que la distribución de cepas es relativamente homogénea a lo largo de los clados.

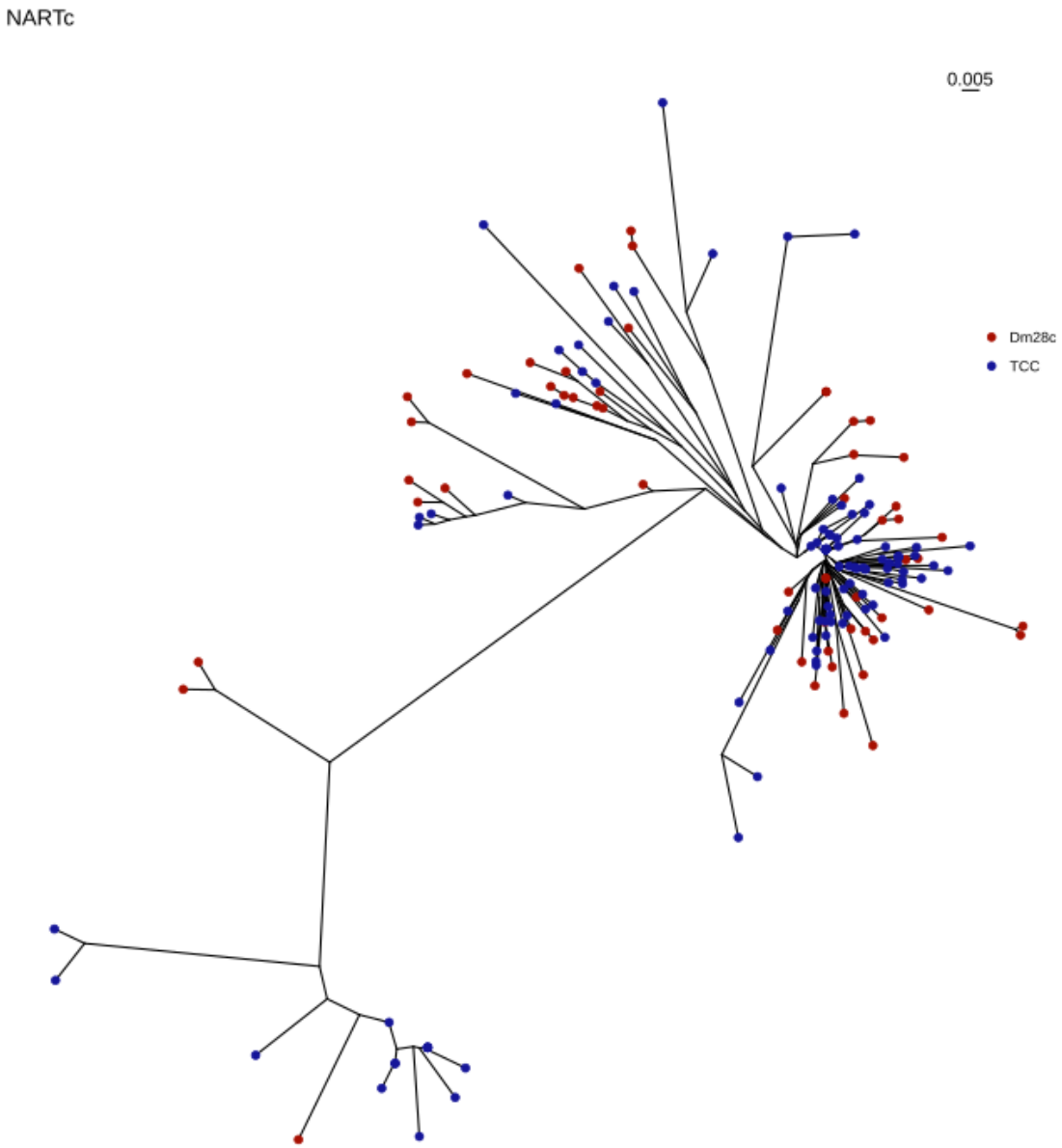


Figura 11. Figura 9. Filogenia generada por NJ de copias NARTc en TCC (azul) y Dm28c (rojo)

VIPER

En la filogenia de VIPER (Figura 12) la mayoría de los clados están compuestos por copias de ambas cepas, con la excepción de algunos clados exclusivos de TCC y Dm28c. La mayoría de los clados mixtos, corresponden a un patrón en el cual las copias de TCC y Dm28c serían ortólogas, es decir, heredadas del ancestro previo a la divergencia. Algunas de las ramas de estas copias ortólogas contienen clados con copias parálogas entre si, los cuales corresponderian a eventos de amplificación. Teniendo en cuenta que estas copias parálogas tienen muy alto porcentaje de identidad entre sí, estos eventos de amplificación podrían ser explicados, además de por eventos de retrotransposición, por duplicaciones segmentales o conversión génica.

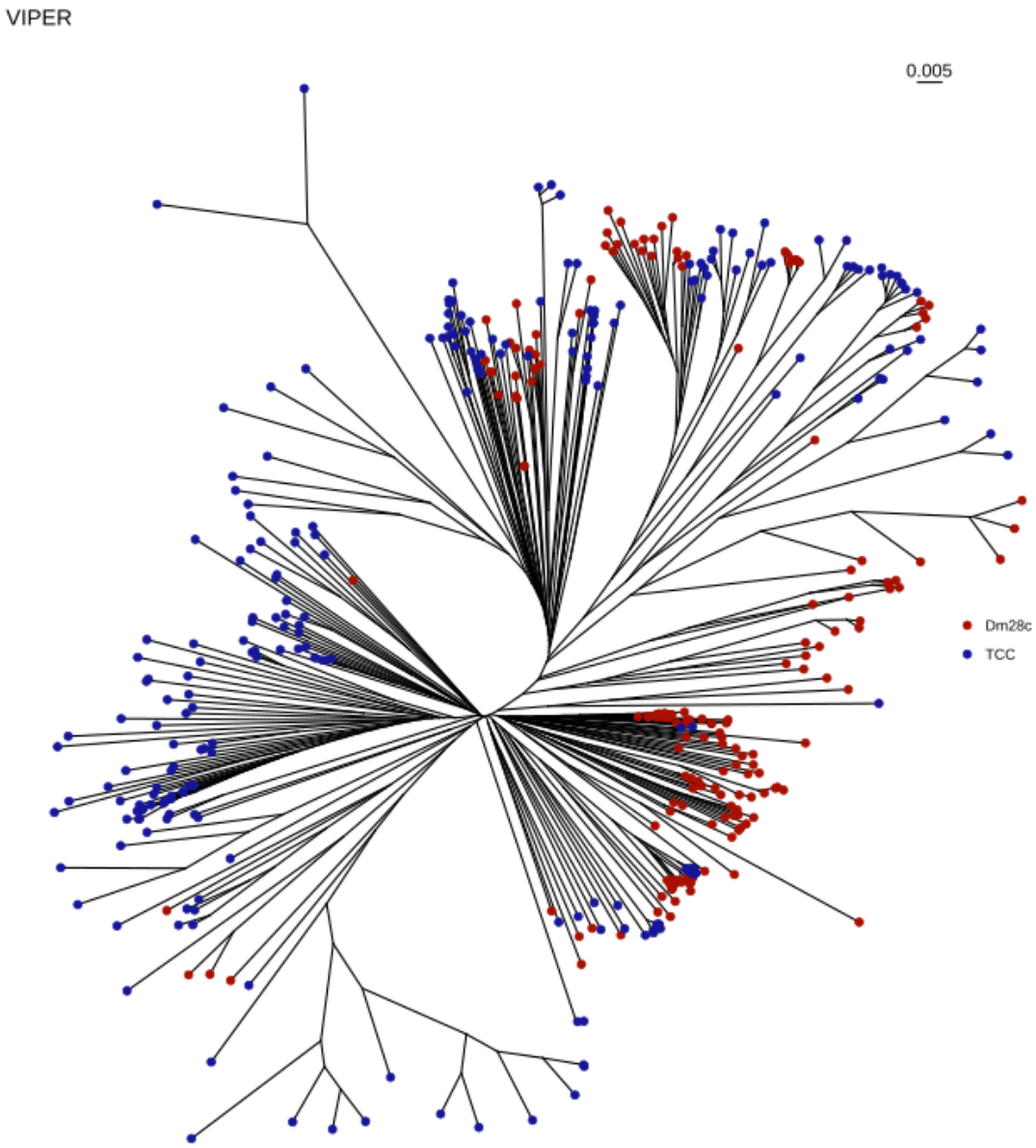


Figura 12. Figura 9. Filogenia generada por NJ de copias VIPER en TCC (azul) y Dm28c (rojo)

SIRE

El retrotransposon SIRE muestra una distribución de cepas homogénea a lo largo de los clados (Figura 13), tal como se observa para los elementos NARTc. De forma análoga al par L1Tc-NARTc, VIPER y SIRE no muestran topologías similares. La ubicuidad de ambas cepas en los clados indica que no hubieron eventos de amplificación posteriores a la divergencia entre las distintas cepas.

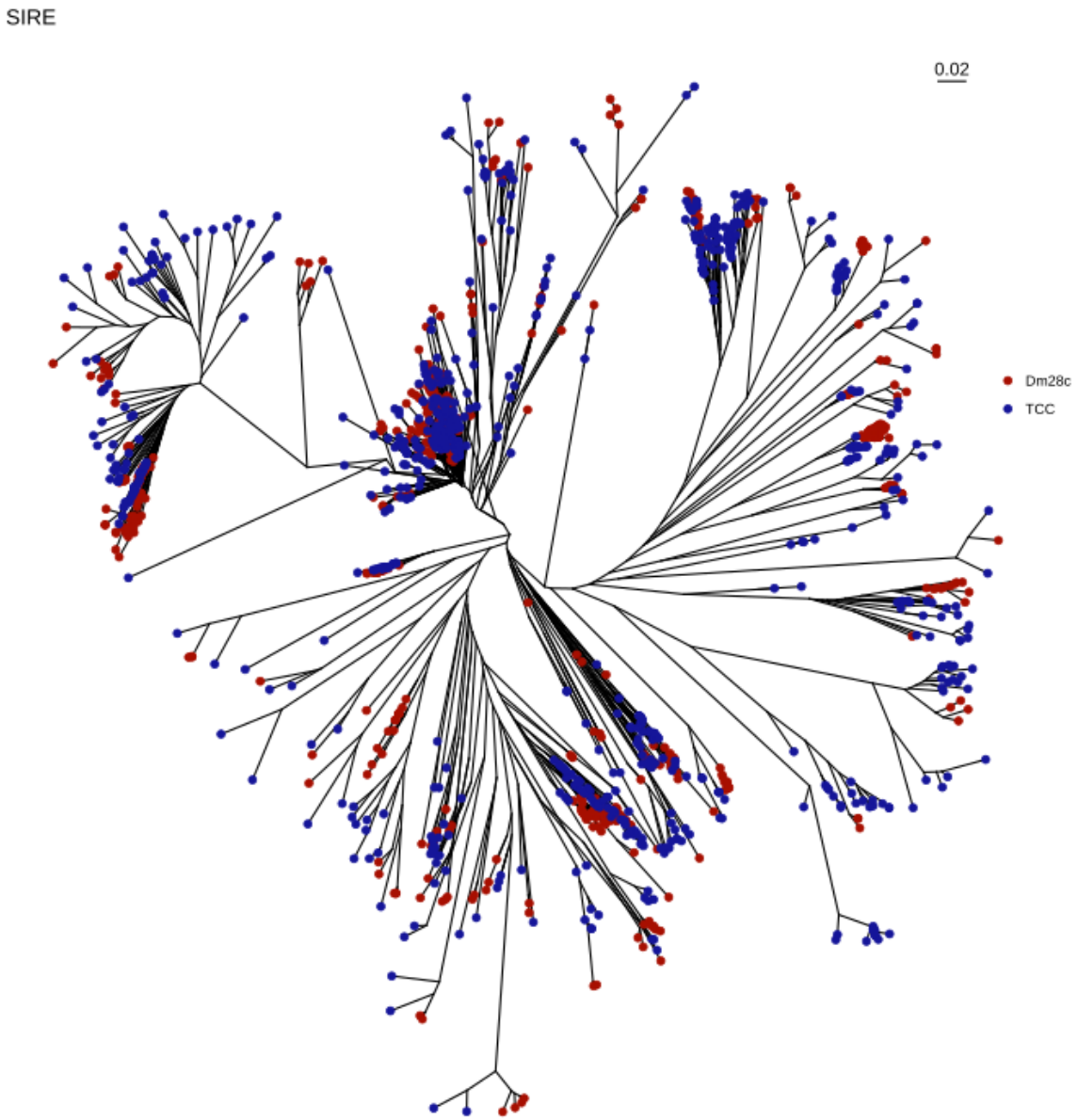


Figura 13. Figura 9. Filogenia generada por NJ de copias SIRE en TCC (azul) y Dm28c (rojo)

Anotaciones en cis

Con el objetivo de determinar posibles especificidades de inserción por regiones genómicas particulares, se analizaron las anotaciones adyacentes a cada una de las copias de los retroelementos estudiados.

Se escribió un script, `get_up_and_downstream_annotations.py`, que toma como *input* dos archivos de anotación en formato .gff; el archivo generado a partir de la búsqueda mediante BLAST, y un archivo de anotación del genoma, en donde también se encuentran las anotaciones del primer archivo. El *output* es una tabla que especifica las anotaciones corriente arriba y corriente abajo de cada copia, y su distancia física en nucleótidos. En el caso de que no haya una anotación adyacente, como puede ser en el caso de una copia en el borde de un contig, se asigna un “NA” a la entrada correspondiente en la tabla. El script descarta las anotaciones que se encuentran solapadas con cada elemento evaluado. La tabla generada se procesó con el script `cis_annotations.R`, con un umbral designado tal que solo sean consideradas las anotaciones que se encuentren a menos de 1000 nucleótidos de distancia.

Los resultados relevantes se encuentran resumidos en la Tabla 3. Muchas de las anotaciones en *cis* de gran número coinciden con genes altamente representados en el genoma, como mucinas, retrotransposon hotspot proteins (RHS), DGF-1, GP63 y trans-sialidasas, o con genes hipotéticos conservados.

Tabla 3. Anotaciones en cis relevantes.

	CZAR			
	Upstream		Downstream	
	Anotacion	N°	Anotacion	N°
TCC	SL	21	SL	18
	SIRE	3	VIPER	3
	VIPER	1		
DM28c	SL	28	SL	43
	SIRE	4	VIPER	1
	L1Tc			
	Upstream		Downstream	
	Anotacion	N°	Anotacion	N°
TCC	Beta tubulina	4	Beta tubulina	7
	Hipotetica	6	Hipotetica	5
	Otras	22	Otras	14
	NARTc			
	Upstream		Downstream	
	Anotacion	N°	Anotacion	N°
TCC	Rnasa H	4	TR	6
	Hipotetica	11	Hipotetica	11
	Otras	21	Otras	32

El estudio de anotaciones en *cis* de CZAR indica que las copias de CZAR se encuentran mayoritariamente adyacentes a otras copias de CZAR y a genes *Spliced Leader*. Esto es de esperar, ya que los genes de *Spliced Leader* se encuentran ordenados en tándem, y CZAR tiene especificidad de inserción por el exón.

En el genoma de TCC se encontraron numerosas anotaciones en *cis* de L1Tc que corresponden a genes putativos de beta-tubulina. Este hecho fue verificado en el explorador web del genoma construido por Berná y colaboradores (80); en todos los arreglos en tándem de genes de beta tubulina se puede encontrar por lo menos una copia de L1Tc (Figura 14). El hecho de que estas regiones no son sintéticas permite descartar la posibilidad de que una de las regiones fue originada a partir de la otra mediante duplicación.

En el caso de NARTc en el genoma de TCC se encontraron anotaciones que corresponden a endonucleasas y transcriptasas reversas; esto puede deberse a un problema de anotación en el cual una copia de L1Tc fue anotada como NARTc, debido a los parámetros de similaridad utilizados.

No se encontraron anotaciones en *cis* relevantes en el caso de SIRE y VIPER.

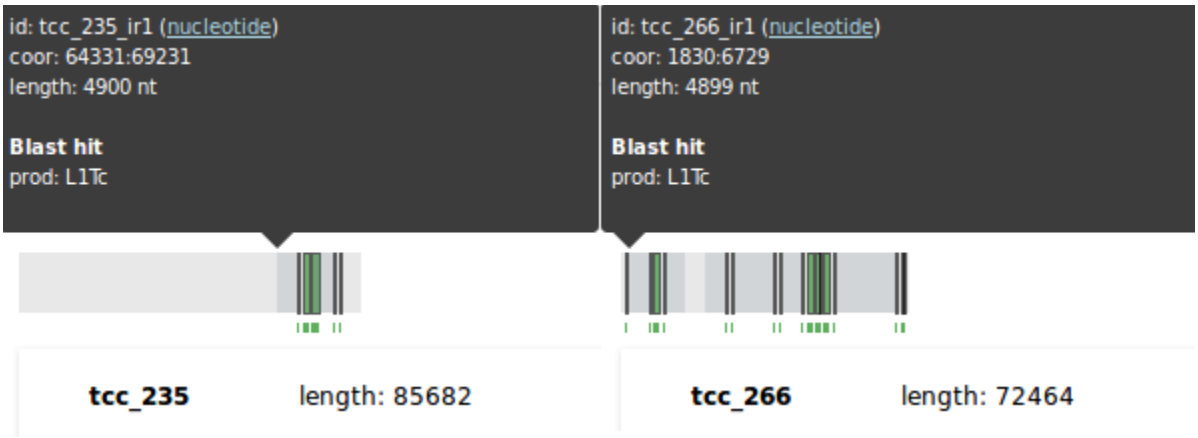


Figura 14. Proximidad de copias L1Tc con genes en tándem de beta tubulina en el genoma de TCC. Los segmentos en verde corresponden a genes de beta tubulina, los segmentos en gris oscuro corresponden a copias de L1Tc. Visualizado en explorador de genoma de *Trypanosoma cruzi* construido por Berná y colaboradores(80).

Análisis filogenético con otras cepas

Con el objetivo de obtener una visión del panorama evolutivo de los retrotransposones de interés en *Trypanosoma cruzi*, se procedió a implementar el proceso de búsqueda y anotación de CZAR, L1Tc, NARTc, VIPER y SIRE en genomas disponibles en la base de datos de NCBI. Utilizando los mismos criterios descritos anteriormente, se construyeron bases de datos con los genomas de las cepas Sylvio, Bug2148, Esmeraldo y CLBrener (divididos en sus haplotipos derivados de Esmeraldo, No-Esmeraldo y sin asignar) para luego realizar búsqueda por BLAST.

El número de retroelementos encontrados depende, además de los parámetros de búsqueda, de la tecnología con la cual fueron secuenciadas los genomas y la homocigosis de las cepas. Los genomas secuenciados con tecnologías de segunda generación (CLBrener, Esmeraldo) van a tender a generar ensamblados con elementos repetidos colapsados (y como consecuencia, menor cantidad), mientras que los secuenciados con tecnologías de tercera generación (Sylvio, Bug2148) permitirían encontrar una mayor cantidad de retroelementos. Este fenómeno se cumple para CZAR y L1Tc, en donde se encontró mayor cantidad de retroelementos en Sylvio y Bug2148 que en CLBrener y Esmeraldo (Tabla 4).

Tabla 4. Número de copias de ETs encontrados, agrupados por orden, cepa y DTU.

DTU	Cepa	CZAR	L1Tc	NARTc	VIPER	SIRE
TcI	Dm28c	57	54	55	669	194
	Sylvio	20	104	60	82	485
TcII	Esmeraldo	2	0	30	45	246
TcV	Bug2148	24	91	62	159	716
TcVI	CLBrener	12	28	296	275	852
	TCC	43	43	110	851	244

CZAR

La filogenia de CZAR (Figura 15) muestra un clado (i) mayoritariamente compuesto por copias de cepas pertenecientes a la DTU TcVI (CL Brener y TCC), con la excepción de dos copias de Esmeraldo, perteneciente a DTU TcII. Los modelos actuales de intercambio genético de DTUs sostienen que TcVI se originó a partir un evento de hibridación entre TcII y TcIII, lo cual explica la presencia de copias de una cepa TcII en un clado casi exclusivo para cepas de TcVI. Se observa una separación clara por haplotipos; dentro del clado i se encuentra un subclado (ii) que corresponde al haplotipo “Non-Esmeraldo like” de TCC y CLBrener.

En el clado iii, las copias de Bug2148, perteneciente a DTU TcV agrupan junto con copias de Dm28c y Sylvio, pertenecientes a DTU TcI. Esto es inesperado, ya que los modelos actuales sostienen que TcV se originó a partir de una hibridación entre TcII y TcIII. Se esperaría, en cambio, que algunas copias de Bug2148 agruparan junto con el clado correspondiente al haplotipo Esmeraldo-*like*. Es posible que la cepa secuenciada bajo el nombre Bug2148 sea en realidad una cepa perteneciente a la DTU TcI. Experimentos y análisis subsecuentes son necesarios para confirmar o descartar esta hipótesis.

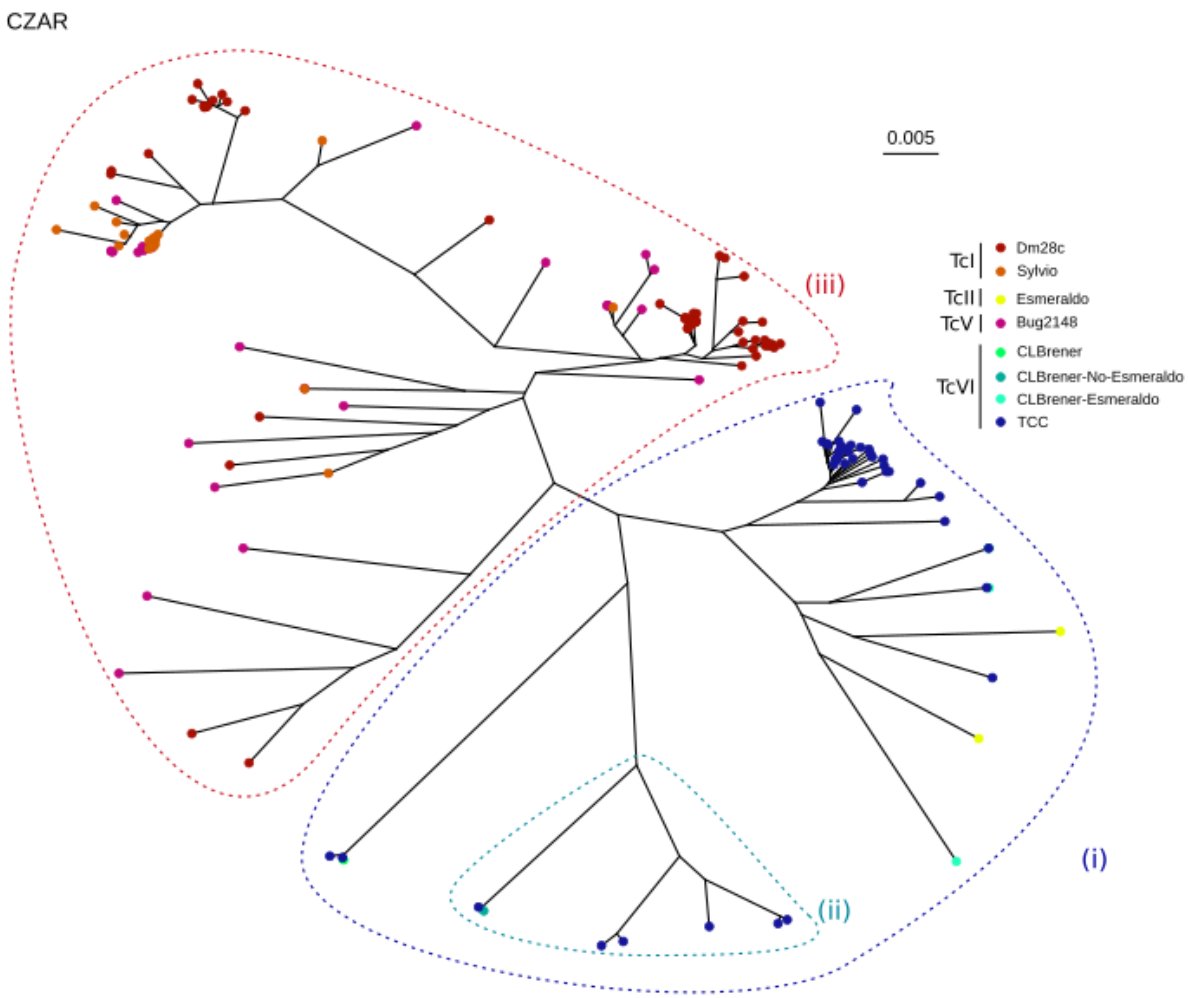


Figura 15. Filogenia de copias de CZAR de las cepas Dm28c, Sylvio, Esmeraldo, Bug2148, CLBrener (separado por haplotipos) y TCC.

L1Tc

En la filogenia de L1Tc (Figura 16) se pueden reconocer tres clados principales; un clado compuesto exclusivamente por copias de las cepas pertenecientes a TcVI (i), uno compuesto en su totalidad por copias de Dm28c, Sylvio y Bug2148 (ii), y un clado mixto, compuesto por copias de TcI, TcV y TcVI (ii). Este clado mixto puede ser explicado por la agrupación de secuencias de L1Tc ancestrales, previas a la divergencia entre las distintas DTUs, mientras que los clados DTU-específicos contienen secuencias que fueron sujetas a una expansión posterior a la divergencia.

Nuevamente, se observa que las secuencias pertenecientes a Bug2148 agrupan en su mayoría con secuencias pertenecientes a DTU TcI.

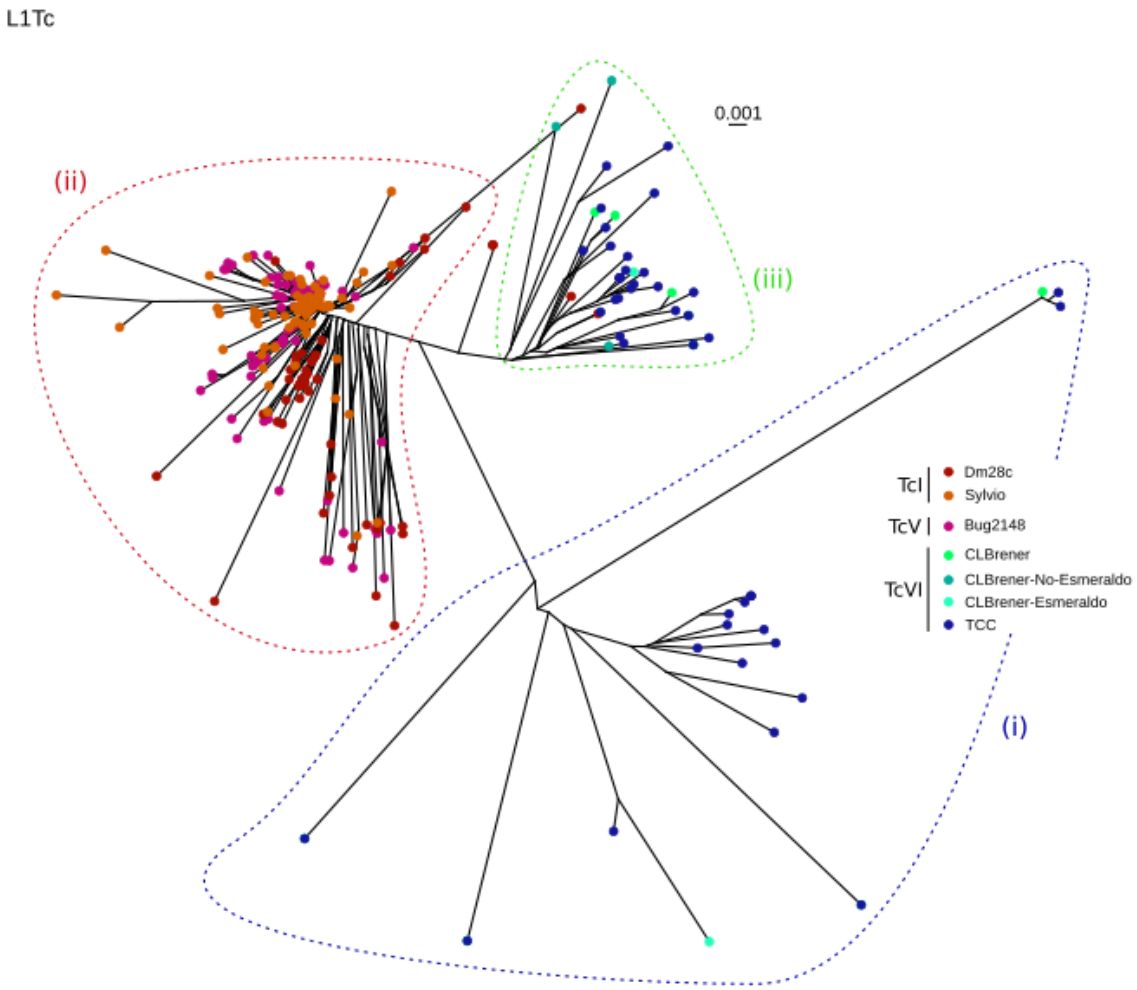


Figura 16. Filogenia de copias de L1Tc de las cepas Dm28c, Sylvio, Esmeraldo, Bug2148, CLBrener (separado por haplotipos) y TCC.

NARTc

Las cepas muestran una distribución homogénea a lo largo de los clados en la filogenia de NARTc (Figura 17). Esta distribución sugiere que los eventos de amplificación de NARTc se dieron previos a la divergencia entre las cepas. De acuerdo con las filogenias de TCC y Dm28c, NARTc no sigue el mismo patrón de agrupación que L1Tc.

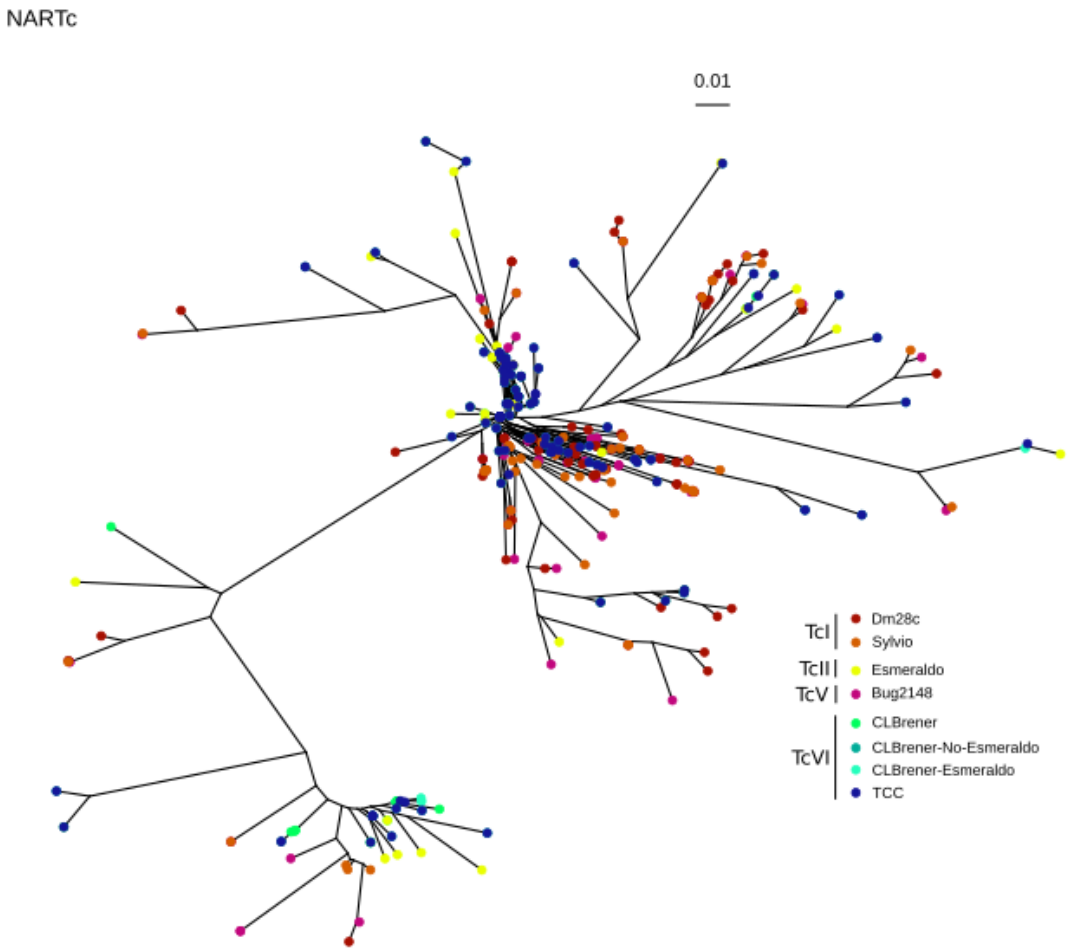


Figura 17. Filogenia de copias de NARTc de las cepas Dm28c, Sylvio, Esmeraldo, Bug2148, CLBrener (separado por haplotipos) y TCC.

VIPER

La filogenia de VIPER no muestra un patrón claro con respecto a la distribución de las cepas en los clados. Esta topología se debe a que las copias de VIPER se encuentran posiblemente inactivas, y como consecuencia han sufrido una mayor divergencia en sus secuencias al compararla con los otros elementos potencialmente activos.

De todas maneras, es posible observar algunos clados especificos para las DTUs TcI (Dm28c y Sylvio) y TcII (Esmeraldo, y los haplotipos Esmeraldo-like de TCC y CL Brener). La presencia de estos clados podría ser explicada por eventos de amplificación específicos para los ancestros comunes de cada DTU.

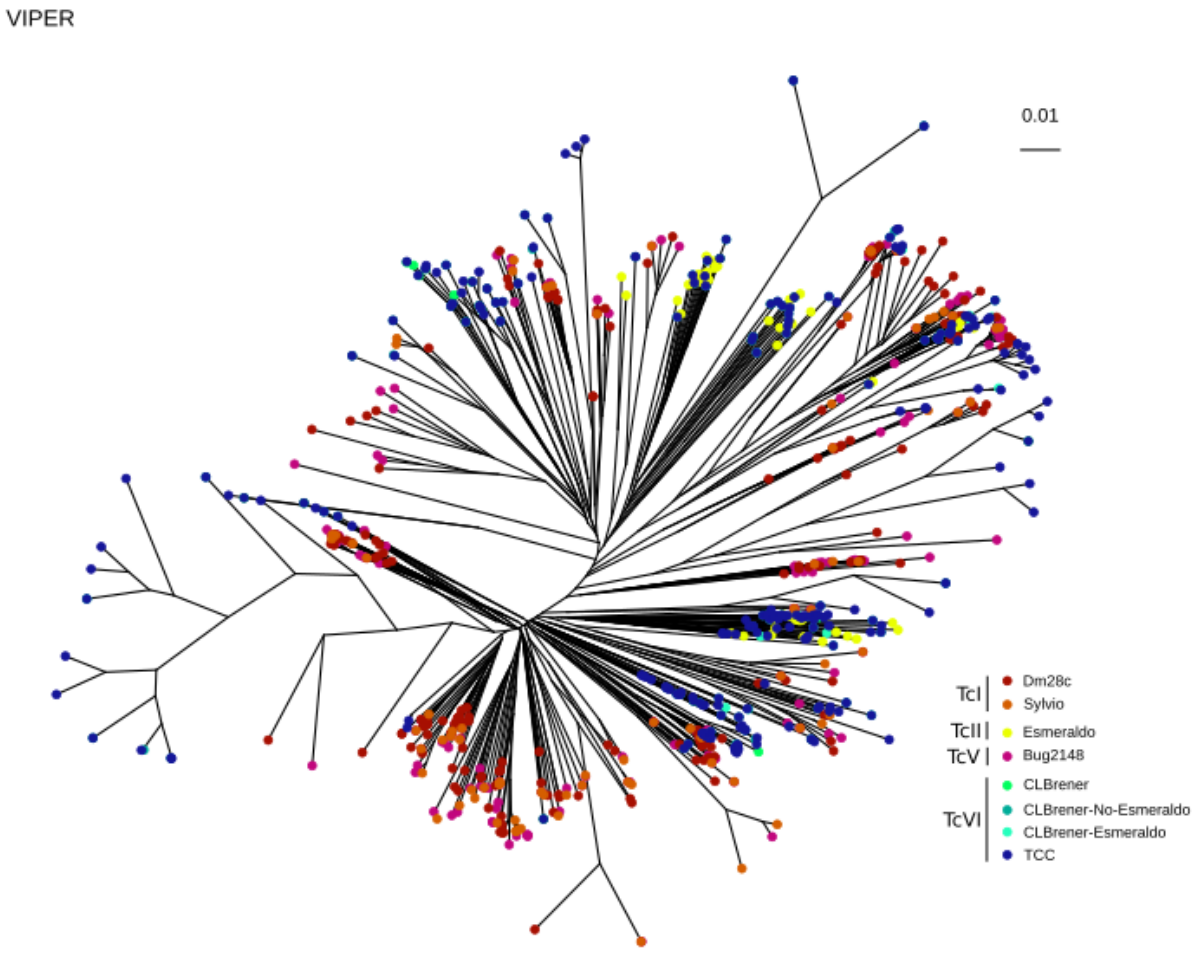


Figura 18. Filogenia de copias de VIPER de las cepas Dm28c, Sylvio, Esmeraldo, Bug2148, CLBrener (separado por haplotipos) y TCC.

SIRE

Las cepas muestran una distribución homogénea a lo largo de los clados en la filogenia de SIRE (Figura 19). Esta topología concuerda con los resultados obtenidos con las secuencias de TCC y Dm28c. En conjunto, las filogenias de SIRE indican que los eventos de amplificación se dieron previos a la divergencia entre todas las cepas.

Se confirma el hecho de que los ETs no autónomos (NARTc y SIRE), no siguieron el mismo patrón de amplificación que sus correspondientes ETs autónomos (L1Tc y VIPER). Esto sugiere que estos ETs perdieron la capacidad de utilizar la maquinaria de retrotransposición de sus “compañeros” autónomos previamente a la divergencia entre las cepas, o que alternativamente nunca la poseyeron.

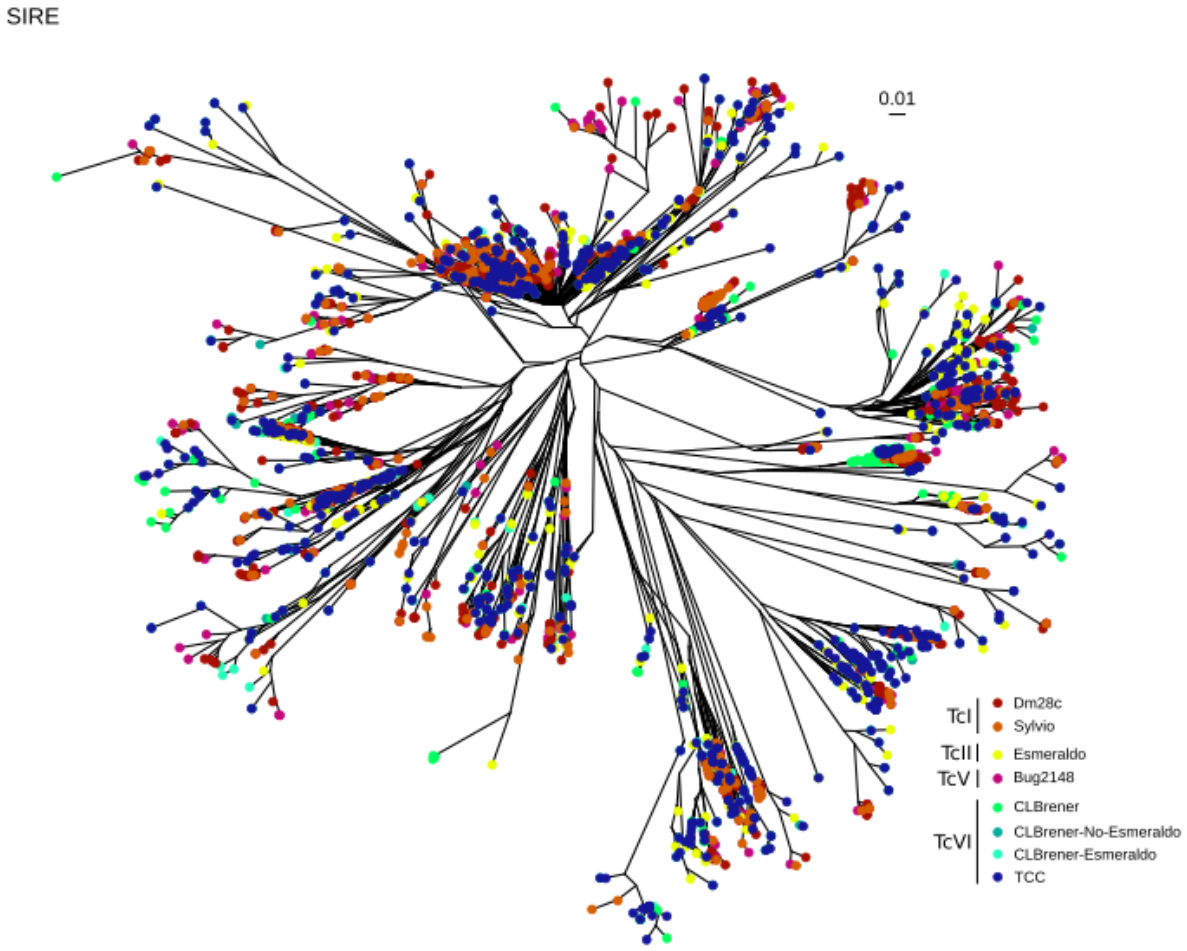


Figura 19. Filogenia de copias de SIRE de las cepas Dm28c, Sylvio, Esmeraldo, Bug2148, CLBrener (separado por haplotipos) y TCC.

Conclusiones

Fue posible encontrar una gran cantidad de copias con alto porcentaje de identidad de los distintos retroelementos en los genomas de TCC y Dm28c. Se encontró un mayor número de retrotransposones largos (CZAR y L1Tc) en los genomas secuenciados con tecnologías de tercera generación (TCC, Dm28c, Sylvio).

Se determinó el número de CZAR y L1Tc potencialmente autónomos, siendo una fracción considerable de los elementos totales, mientras que no se encontraron copias VIPER potencialmente autónomos. Estos resultados deben ser verificados en ensayos de biología molecular, en donde se confirme la actividad de las proteínas que codifican sus ORFs, y su capacidad para insertarse en una hebra de ADN.

El estudio de contexto genómico de las copias confirma la especificidad de CZAR por los tandem de *Spliced Leader*. Además, se encontró que hay copias de L1Tc en todos los tándem de genes de beta tubulina.

Las filogenias construidas indican que CZAR y L1Tc sufrieron amplificaciones posteriores a la divergencia entre las distintas DTUs. Particularmente en CZAR, es posible observar dos clados distintos que corresponden a los dos haplotipos de la DTU TcVI. En el caso de L1Tc se puede observar un clado que corresponde a las copias previas a la divergencia, mientras que las copias de CZAR se agrupan en su totalidad por DTU.

Las filogenias de VIPER, indican que este elemento se encuentra inactivo, ya que hay una gran presencia de clados mixtos para las DTUs, y que la divergencia entre sus copias es mayor que la de los elementos potencialmente activos. De todas maneras es posible observar clados que corresponden a eventos de amplificación que tuvieron lugar en los ancestros comunes de algunas DTUs, lo cual sugiere que VIPER todavía se encontraba activo al momento de la divergencia de todas las DTUs.

En cuanto a NARTc y SIRE, no se observan clados específicos para DTUs, lo cual sugiere que los eventos de amplificación y “muerte” de estos retrotransposones fueron previos a la divergencia. El hecho de que NARTc y SIRE no presenten un patrones de agrupación similar a L1Tc y VIPER, respectivamente, sugiere que estos ET no autónomos ya no poseen la capacidad de utilizar la maquinaria de retrotransposición de L1Tc, o nunca la tuvieron.

Las secuencias provenientes de la cepa Bug2148, en las filogenias que se observan clados DTU específico, agrupan con secuencias correspondientes a DTU TcI, algo totalmente inesperado ya que Bug2148 pertenece a la DTU híbrida TcV. Es posible que esto se deba a un error de manipulación de las cepas al momento de la construcción de librerías.

Los resultados de este trabajo reflejan la utilidad de los datos de secuenciación de tercera generación para determinar la variabilidad intragenómica de elementos transponibles.

Los umbrales de búsqueda utilizados para la extracción de secuencias de las distintas copias introducen un sesgo al momento de la construcción de filogenias, y como consecuencia, en las conclusiones obtenidas a partir de estas. Utilizar umbrales de búsqueda más laxos, aplicar buenos criterios de clasificación, e incluir secuencias de numerosas DTUs podría permitir confirmar o descartar los escenarios propuestos en este trabajo, además de esclarecer los mecanismos que dictaron la historia evolutiva de los ETs y los eventos de amplificación a los cuales estuvieron sujetos los genomas de *Trypanosoma cruzi*.

Bibliografía

1. WHO | Chagas disease (American trypanosomiasis) [Internet]. WHO. [cited 2018 May 10]. Available from: http://www.who.int/chagas/en/

2. Lee BY, Bacon KM, Bottazzi ME, Hotez PJ. Global economic burden of Chagas disease: a computational simulation model. The Lancet infectious diseases. 2013;13(4):342–348.

3. Jannin J, Villa L. An overview of Chagas disease treatment. Memórias do Instituto Oswaldo Cruz. 2007;102:95–98.

4. Prevention C-C for DC and. CDC - Chagas Disease [Internet]. 2017 [cited 2018 May 10]. Available from: https://www.cdc.gov/parasites/chagas/index.html

5. El-Sayed NM, Myler PJ, Bartholomeu DC, Nilsson D, Aggarwal G, Tran A-N, et al. The Genome Sequence of Trypanosoma cruzi, Etiologic Agent of Chagas Disease. Science. 2005 Jul 15;309(5733):409–15.

6. Tibayrenc M, Ward P, Moya A, Ayala FJ. Natural populations of Trypanosoma cruzi, the agent of Chagas disease, have a complex multiclonal structure. PNAS. 1986 Jan 1;83(1):115–9.

7. Henriksson J, \AAslund L, Pettersson U. Karyotype variability in Trypanosoma cruzi. Parasitology today. 1996;12(3):108–114.

8. Minning TA, Weatherly DB, Flibotte S, Tarleton RL. Widespread, focal copy number variations (CNV) and whole chromosome aneuploidies in Trypanosoma cruzi strains revealed by array comparative genomic hybridization. BMC Genomics. 2011 Mar 7;12:139.

9. Machado CA, Ayala FJ. Nucleotide sequences provide evidence of genetic exchange among distantly related lineages of Trypanosoma cruzi. PNAS. 2001 Jun 19;98(13):7396–401.

10. Gaunt MW, Yeo M, Frame IA, Stothard JR, Carrasco HJ, Taylor MC, et al. Mechanism of genetic exchange in American trypanosomes. Nature. 2003 Feb 27;421:936.

11. Tibayrenc M. Genetic epidemiology of parasitic protozoa and other infectious agents: the need for an integrated approach. International journal for parasitology. 1998;28(1):85–104.

12. Zingales B, Miles MA, Campbell DA, Tibayrenc M, Macedo AM, Teixeira MMG, et al. The revised Trypanosoma cruzi subspecific nomenclature: Rationale,

epidemiological relevance and research applications. Infection, Genetics and Evolution. 2012 Mar;12(2):240–53.

13. Nara T, Gao G, Nakajima-Shimada J, Aoki T. Localization of Carbamoyl-Phosphate Synthetase II (Cps Ii) and Aspartate Carbamoyltransferase (Act) Genes in <Emphasis Type="Italic">trypanosoma Cruzi</Emphasis> Chromosomal DNA. In: Purine and Pyrimidine Metabolism in Man IX [Internet]. Springer, Boston, MA; 1998 [cited 2018 Mar 29]. p. 227–30. (Advances in Experimental Medicine and Biology). Available from: https://link.springer.com/chapter/10.1007/978-1-4615-5381-6_44

14. Campbell DA, Thomas S, Sturm NR. Transcription in kinetoplastid protozoa: why be normal? Microbes and Infection. 2003 Nov;5(13):1231–40.

15. Ghedin E, Bringaud F, Peterson J, Myler P, Berriman M, Ivens A, et al. Gene synteny and evolution of genome architecture in trypanosomatids. Molecular and Biochemical Parasitology. 2004 Apr;134(2):183–91.

16. Campbell DA, Sturm NR, Yu MC. Transcription of the Kinetoplastid Spliced Leader RNA Gene. Parasitology Today. 2000 Feb 1;16(2):78–82.

17. McCarthy-Burke C, Taylor ZA, Buck GA. Characterization of the spliced leader genes and transcripts in Trypanosoma cruzi*. In: RNA: Catalysis, Splicing, Evolution [Internet]. Amsterdam: Elsevier; 1989 [cited 2018 May 10]. p. 177–89. Available from: https://www.sciencedirect.com/science/article/pii/B9780444812100500221

18. Tomasini N, Lauthier JJ, Rumi MMM, Ragone PG, D'Amato AAA, Brandan CP, et al. Interest and limitations of Spliced Leader Intergenic Region sequences for analyzing Trypanosoma cruzi I phylogenetic diversity in the Argentinean Chaco. Infection, Genetics and Evolution. 2011 Mar;11(2):300–7.

19. Levin HL, Moran JV. Dynamic interactions between transposable elements and their hosts. Nature Reviews Genetics. 2011 Aug 18;12(9):615–27.

20. Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, et al. A unified classification system for eukaryotic transposable elements. Nature Reviews Genetics. 2007;8(12):973.

21. Bennett EA, Keller H, Mills RE, Schmidt S, Moran JV, Weichenrieder O, et al. Active Alu retrotransposons in the human genome. Genome Res. 2008 Dec 1;18(12):1875–83.

22. Deininger PL, Batzer MA. Alu Repeats and Human Disease. Molecular Genetics and Metabolism. 1999 Jul 1;67(3):183–93.

23. Lynch M, Walsh B. The origins of genome architecture. Vol. 98. Sinauer Associates Sunderland (MA); 2007.

24. Han JS. Non-long terminal repeat (non-LTR) retrotransposons: mechanisms,

recent developments, and unanswered questions. Mobile Dna. 2010;1(1):15.

25. Eickbush TH, Malik HS. Origins and Evolution of Retrotransposons. Mobile DNA II. 2002 Jan 1;1111–44.

26. Luan DD, Korman MH, Jakubczak JL, Eickbush TH. Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: A mechanism for non-LTR retrotransposition. Cell. 1993 Feb 26;72(4):595–605.

27. Winckler T, Dingermann T, Glöckner G. Dictyostelium mobile elements: strategies to amplify in a compact genome. Cellular and Molecular Life Sciences CMLS. 2002;59(12):2097–2111.

28. Hofmann J, Schumann G, Borschet G, Gösseringer R, Bach M, Bertling WM, et al. Transfer RNA genes from Dictyostelium discoideum are frequently associated with repetitive elements and contain consensus boxes in their 5′ and 3′-flanking regions. Journal of Molecular Biology. 1991;222(3):537–52.

29. Poulter RTM, Goodwin TJD. DIRS-1 and the other tyrosine recombinase retrotransposons. Cytogenetic and Genome Research. 2005;110(1–4):575–88.

30. Cappello J, Handelsman K, Lodish HF. Sequence of Dictyostelium DIRS-1: an apparent retrotransposon with inverted terminal repeats and an internal circle junction sequence. Cell. 1985;43(1):105–115.

31. Goodwin TJD, Poulter RTM. A New Group of Tyrosine Recombinase-Encoding Retrotransposons. Molecular Biology and Evolution. 2004 Apr;21(4):746–59.

32. Chuong EB, Elde NC, Feschotte C. Regulatory activities of transposable elements: from conflicts to benefits. Nature Reviews Genetics. 2017 Feb;18(2):71–86.

33. Feschotte C. Transposable elements and the evolution of regulatory networks. Nature Reviews Genetics. 2008;9(5):397.

34. Mason JM, Biessmann H. The unusual telomeres of Drosophila. Trends in genetics. 1995;11(2):58–62.

35. Bodega B, Orlando V. Repetitive elements dynamics in cell identity programming, maintenance and disease. Current Opinion in Cell Biology. 2014 Dec;31:67–73.

36. Bourque G, Leong B, Vega VB, Chen X, Lee YL, Srinivasan KG, et al. Evolution of the mammalian transcription factor binding repertoire via transposable elements. Genome Res. 2008 Nov 1;18(11):1752–62.

37. Sundaram V, Cheng Y, Ma Z, Li D, Xing X, Edge P, et al. Widespread contribution of transposable elements to the innovation of gene regulatory networks. Genome Res. 2014 Dec 1;24(12):1963–76.

38. Kunarso G, Chia N-Y, Jeyakani J, Hwang C, Lu X, Chan Y-S, et al. Transposable elements have rewired the core regulatory network of human embryonic stem cells. Nature Genetics. 2010 Jul;42(7):631–4.

39. Wang T, Zeng J, Lowe CB, Sellers RG, Salama SR, Yang M, et al. Species-specific endogenous retroviruses shape the transcriptional network of the human tumor suppressor protein p53. PNAS. 2007 Nov 20;104(47):18613–8.

40. Piriyapongsa J, Jordan IK. Dual coding of siRNAs and miRNAs by plant transposable elements. RNA. 2008 Mar 27;14(5):814–21.

41. Lisch D. How important are transposons for plant evolution? Nature Reviews Genetics. 2012 Dec 18;14:49.

42. Hirsch CD, Springer NM. Transposable element influences on gene expression in plants. Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms. 2017;1860(1):157–65.

43. Studer A, Zhao Q, Ross-Ibarra J, Doebley J. Identification of a functional transposon insertion in the maize domestication gene tb1. Nature Genetics. 2011 Sep 25;43:1160.

44. Butelli E, Licciardello C, Zhang Y, Liu J, Mackay S, Bailey P, et al. Retrotransposons Control Fruit-Specific, Cold-Dependent Accumulation of Anthocyanins in Blood Oranges. Plant Cell. 2012 Mar 1;24(3):1242.

45. Wickstead B, Ersfeld K, Gull K. Repetitive elements in genomes of parasitic protozoa. Microbiology and Molecular Biology Reviews. 2003;67(3):360–375.

46. Bhattacharya S, Bakre A, Bhattacharya A. Mobile genetic elements in protozoan parasites. Journal of genetics. 2002;81(2):73–86.

47. Thomas MC, Macias F, Alonso C, López MC. The biology and evolution of transposable elements in parasites. Trends in Parasitology. 2010;26(7):350–62.

48. Bringaud F, Bartholomeu DC, Blandin G, Delcher A, Baltz T, El-Sayed NMA, et al. The Trypanosoma cruzi L1Tc and NARTc Non-LTR Retrotransposons Show Relative Site Specificity for Insertion. Molecular Biology and Evolution. 2006 Feb 1;23(2):411–20.

49. Heras SR, Lopez MC, Olivares M, Thomas MC. The L1Tc non-LTR retrotransposon of Trypanosoma cruzi contains an internal RNA-pol II-dependent promoter that strongly activates gene transcription and generates unspliced transcripts. Nucleic Acids Research. 2007 Mar 11;35(7):2199–214.

50. Olivares M, Alonso C, López MC. The open reading frame 1 of the L1Tc retrotransposon ofTrypanosoma cruzi codes for a protein with apurinic-apyrimidinic nuclease activity. Journal of Biological Chemistry. 1997;272(40):25224–25228.

51. Heras SR, Thomas MC, Macias F, Patarroyo ME, Alonso C, López MC. Nucleic-acid-binding properties of the C2-L1Tc nucleic acid chaperone encoded by L1Tc retrotransposon. Biochemical Journal. 2009 Dec 15;424(3):479–90.

52. Gonz�lez CI, Thomas MC, Olivares M, L�pez MC, Garc�a-P�rez JL. Characterization of reverse transcriptase activity of the L1Tc retroelement from Trypanosoma cruzi. Cellular and Molecular Life Sciences (CMLS). 2003 Dec 1;60(12):2692–701.

53. Martín F, Marañón C, Olivares M, Alonso C, López MC. Characterization of a non-long terminal repeat retrotransposon cDNA (L1Tc) fromTrypanosoma cruzi: Homology of the first ORF with the APE family of DNA repair enzymes. Journal of molecular biology. 1995;247(1):49–59.

54. Bringaud F, García-Pérez JL, Heras SR, Ghedin E, El-Sayed NM, Andersson B, et al. Identification of non-autonomous non-LTR retrotransposons in the genome of Trypanosoma cruzi. Molecular and Biochemical Parasitology. 2002 Sep 10;124(1):73–8.

55. Villanueva MS, Williams SP, Beard CB, Richards FF, Aksoy S. A new member of a family of site-specific retrotransposons is present in the spliced leader RNA genes of Trypanosoma cruzi. Molecular and cellular biology. 1991;11(12):6139–6148.

56. Aksoy S, Williams S, Chang S, Richards FF. SLACS retrotransposon from Trypanosoma brucei gambiense is similar to mammalian LINEs. Nucleic acids research. 1990;18(4):785–792.

57. Gabriel A, Yen TJ, Schwartz DC, Smith CL, Boeke JD, Sollner-Webb B, et al. A rapidly rearranging retrotransposon within the miniexon gene locus of Crithidia fasciculata. Mol Cell Biol. 1990 Feb 1;10(2):615–24.

58. Lorenzi HA, Robledo G, Levin MJ. The VIPER elements of trypanosomes constitute a novel group of tyrosine recombinase-enconding retrotransposons. Molecular and Biochemical Parasitology. 2006 Feb;145(2):184–94.

59. Vázquez M, Ben-Dov C, Lorenzi H, Moore T, Schijman A, Levin MJ. The short interspersed repetitive element of Trypanosoma cruzi, SIRE, is part of VIPER, an unusual retroelement related to long terminal repeat retrotransposons. Proceedings of the National Academy of Sciences. 2000;97(5):2128–2133.

60. Pavia PX, Thomas MC, López MC, Puerta CJ. Molecular characterization of the short interspersed repetitive element SIRE in the six discrete typing units (DTUs) of Trypanosoma cruzi. Experimental Parasitology. 2012 Oct;132(2):144–50.

61. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The Sequence of the Human Genome. Science. 2001 Feb 16;291(5507):1304–51.

62. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-

generation sequencing technologies. Nature Reviews Genetics. 2016 Jun;17(6):333–51.

63. Treangen TJ, Salzberg SL. Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nature Reviews Genetics. 2012 Jan;13(1):36–46.

64. SMRT Sequencing - PacBio [Internet]. [cited 2018 May 10]. Available from: https://www.pacb.com/smrt-science/smrt-sequencing/

65. Oxford Nanopore Technologies [Internet]. [cited 2018 May 10]. Available from: https://nanoporetech.com/

66. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009 Dec 15;10:421.

67. Rice P, Longden I, Bleasby A. EMBOSS: the European molecular biology open software suite. Elsevier Current Trends; 2000.

68. Marchler-Bauer A, Bryant SH. CD-Search: protein domain annotations on the fly. Nucleic Acids Res. 2004 Jul 1;32(suppl_2):W327–31.

69. Marchler-Bauer A, Bo Y, Han L, He J, Lanczycki CJ, Lu S, et al. CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. Nucleic Acids Research. 2017 Jan 4;45(D1):D200–3.

70. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability | Molecular Biology and Evolution | Oxford Academic [Internet]. [cited 2018 May 11]. Available from: https://academic.oup.com/mbe/article/30/4/772/1073398

71. Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009 Aug 1;25(15):1972–3.

72. Jalview Version 2—a multiple sequence alignment editor and analysis workbench | Bioinformatics | Oxford Academic [Internet]. [cited 2018 May 11]. Available from: https://academic.oup.com/bioinformatics/article/25/9/1189/203460

73. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004 Mar 1;32(5):1792–7.

74. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol. 1987 Jul 1;4(4):406–25.

75. R: The R Project for Statistical Computing [Internet]. [cited 2018 May 11]. Available from: https://www.r-project.org/

76. Paradis E, Claude J, Strimmer K. APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics. 2004 Jan 22;20(2):289–90.

77. Yu G, Smith DK, Zhu H, Guan Y, Lam TT-Y. ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods in Ecology and Evolution. 2017 Jan 1;8(1):28–36.

78. Wickham H, RStudio. tidyverse: Easily Install and Load the “Tidyverse” [Internet]. 2017 [cited 2018 May 11]. Available from: https://CRAN.R-project.org/package=tidyverse

79. Xie Y. knitr: A general-purpose tool for dynamic report generation in R. R package version. 2013;1(1).

80. Berná L, Rodriguez M, Chiribao ML, Parodi-Talice A, Pita S, Rijo G, et al. Expanding an expanded genome: long-read sequencing of Trypanosoma cruzi. Microbial Genomics [Internet]. 2018 Apr 30 [cited 2018 May 10]; Available from: http://www.microbiologyresearch.org/content/journal/mgen/10.1099/mgen.0.000177.v1

Anexo

ID secuencias de referencia SIRE.

NW_001849025
NW_001849027
NW_001849037
NW_001849040
NW_001849043
NW_001849044
NW_001849046
NW_001849047
NW_001849049
NW_001849051
NW_001849054
NW_001849058
NW_001849065
NW_001849081
NW_001849097
NW_001849103
NW_001849118
NW_001849121
NW_001849122
NW_001849130
NW_001849132
NW_001849157
NW_001849168
NW_001849170
NW_001849180
NW_001849185
NW_001849187
NW_001849191
NW_001849194
NW_001849197
NW_001849205
NW_001849208
NW_001849212
NW_001849216
NW_001849217
NW_001849222
NW_001849223

NW_001849228
NW_001849235
NW_001849240
NW_001849245
NW_001849251
NW_001849252
NW_001849257
NW_001849260
NW_001849263
NW_001849265
NW_001849271
NW_001849272
NW_001849273
NW_001849274
NW_001849283
NW_001849284
NW_001849292
NW_001849293
NW_001849297
NW_001849298
NW_001849301
NW_001849303
NW_001849304
NW_001849309
NW_001849310
NW_001849315
NW_001849317
NW_001849321
NW_001849322
NW_001849326
NW_001849327
NW_001849328
NW_001849330
NW_001849334
NW_001849335
NW_001849336
NW_001849337
NW_001849338
NW_001849340
NW_001849343
NW_001849348
NW_001849357
NW_001849358
NW_001849361

NW_001849365
NW_001849366
NW_001849369
NW_001849370
NW_001849375
NW_001849379
NW_001849380
NW_001849382
NW_001849383
NW_001849389
NW_001849391
NW_001849392
NW_001849395
NW_001849396
NW_001849397
NW_001849398
NW_001849401
NW_001849402
NW_001849404
NW_001849405
NW_001849406
NW_001849407
NW_001849409
NW_001849411
NW_001849412
NW_001849413
NW_001849414
NW_001849416
NW_001849418
NW_001849420
NW_001849421
NW_001849422
NW_001849423
NW_001849426
NW_001849427
NW_001849428
NW_001849430
NW_001849431
NW_001849433
NW_001849439
NW_001849440
NW_001849441
NW_001849447
NW_001849450

NW_001849453
NW_001849459
NW_001849461
NW_001849463
NW_001849465
NW_001849467
NW_001849468
NW_001849471
NW_001849472
NW_001849473
NW_001849474
NW_001849475
NW_001849476
NW_001849477
NW_001849480
NW_001849481
NW_001849482
NW_001849484
NW_001849486
NW_001849487
NW_001849488
NW_001849489
NW_001849492
NW_001849493
NW_001849494
NW_001849496
NW_001849498
NW_001849499
NW_001849500
NW_001849501
NW_001849504
NW_001849505
NW_001849506
NW_001849507
NW_001849509
NW_001849510
NW_001849511
NW_001849512
NW_001849514
NW_001849516
NW_001849518
NW_001849520
NW_001849521
NW_001849522

NW_001849523
NW_001849527
NW_001849528
NW_001849530
NW_001849532
NW_001849534
NW_001849536
NW_001849538
NW_001849539
NW_001849540
NW_001849542
NW_001849543
NW_001849544
NW_001849545
NW_001849546
NW_001849547
NW_001849548
NW_001849549
NW_001849550
NW_001849552
NW_001849553
NW_001849554
NW_001849555
NW_001849556
NW_001849557
NW_001849558
NW_001849559
NW_001849560
NW_001849561
NW_001849562
NW_001849563
NW_001849564
NW_001849565
NW_001849566
NW_001849567
NW_001849568
NW_001849569
NW_001849570
NW_001849571
NW_001849572
NW_001849573
NW_001849574
NW_001849575
NW_001849576

NW_001849577