Mecanismos regulatorios de familias moduladas a nivel traduccional en *Trypanosoma cruzi*

Tesina de Maestría del Lic. Santiago Radío Lalanne

Orientador: Dr. Pablo Smircich

Co-orientador: Dr. Gustavo Guerberoff

PEDECIBA – Maestría en Bioinformática

Departamento de Genómica - Instituto de Investigaciones Clemente Estable – Ministerio de Educación y Cultura

Laboratorio de Interacciones Moleculares - Facultad de Ciencias - Universidad de la República

Mayo de 2019

Índice

Resumen

Trypanosoma cruzi es el agente etiológico de la enfermedad de Chagas, patología de alta prevalencia en América Latina. El parásito divergió tempranamente dentro de los eucariotas, exhibiendo procesos moleculares distintivos. En particular, los mecanismos de expresión génica son excepcionales, habiendo un número importante de interrogantes con respecto a su funcionamiento. Teniendo en cuenta que la regulación post-transcripcional es el principal nivel de control en tripanosomátidos, la determinación de los ARNm activamente traducidos es especialmente adecuada para analizar los perfiles de expresión génica en estos organismos. Particularmente, dos familias están fuertemente controladas a través de la modulación de su eficiencia traduccional: los genes que codifican las proteínas ribosomales y los genes pertenecientes a la superfamilia de las trans-sialidasas.

En el presente trabajo, diseñamos una metodología capaz de determinar con precisión los genes con eficiencia traduccional diferencial durante la metaciclogénesis de *T. cruzi*. La mejora es producto de una mayor precisión en el alineamiento de lecturas cortas en genomas con un gran porcentaje de secuencias repetidas, y de la aplicación de nuevos métodos de determinación de eficiencia traduccional diferencial.

A su vez, desarrollamos herramientas que nos permitieron profundizar en la compresión de listas de genes producto de análisis de expresión diferencial, DARK e IdMiner. DARK es una herramienta de interfaz gráfica en la cual se puede visualizar e interrogar las anotaciones de proteínas producidas por estrategias de comparación HMM-HMM. Estas comparaciones son muy sensibles y como resultado se obtuvo anotación para más de 2500 proteínas con función desconocida en tripanosomátidos. Por otra parte, IdMiner es un programa de *text-mining* que también cuenta con interfaz gráfica y permite encontrar términos sobrerrepresentados en la literatura asociados a proteínas de interés, así como relaciones entre los términos y entre las proteínas.

Para profundizar en los mecanismos regulatorios de estas proteínas, obtuvimos regiones UTR de *T. cruzi* a través del desarrollo de UTRme, el cual utiliza datos de RNA-Seq para identificar sitios de procesamiento. Cada sitio de procesamiento identificado cuenta con

un puntaje asociado el cual tiene en cuenta características genómicas de los tripanosomátidos.

Posteriormente y a partir de las regiones 5' UTR identificadas se determinaron la presencia de uORFs y se observó cómo son capaces de modular la eficiencia traduccional de los genes asociados. La eficacia para regular la eficiencia traduccional se asoció a características intrínsecas al uORF como su posición en el 5' UTR, tamaño y codón iniciador.

Finalmente, se identificaron motivos regulatorios, tanto a nivel de secuencia primaria como estructura secundaria, en las regiones no traducidas de las proteínas ribosomales y de la superfamilia trans-sialidasas, los cuales pueden jugar un importante papel en la regulación de estas familias.

Por lo tanto, los resultados obtenidos aportan conocimiento a aspectos de la regulación post-transcripcional en *T. cruzi* tanto a nivel general como específico para familias reguladas traduccionalmente. Por otra parte, se generaron varias herramientas bioinformáticas que pueden ser usadas por investigadores del área para el estudio de la biología del parásito a nivel genómico.

1. Introducción

1.1. Los kinetoplástidos

Los organismos de la familia de kinetoplástidos son protistas que pertenecen al filo Euglenozoa. Están categorizados por la presencia de un organelo denominado kinetoplasto, del cual toman su nombre y que representa la apomorfía del grupo. Este organelo es fácilmente identificable como una gran masa de ADN mitocondrial. Los estilos de vida que presentan estos organismos pueden ser muy diversos, por ejemplo, pueden tener un estilo parasítico o de vida libre, pueden ser monoexénicos (un solo hospedero) o diexénicos (dos hospederos) y pueden ser intra o extracelulares. La diferencia entre estos, conjuntamente con las manifestaciones de la enfermedad que algunos producen, rasgos morfológicos e históricos han sido determinantes para la clasificación taxonómica de estos organismos (Lukes et al., 2014). Recientemente, mediante estudios de ARN ribosomal (ARNr) 18S la clase Kinetoplastea ha sido dividida en dos subclases: Porkinetoplastina y Metakinetoplastina. Este último presenta 4 órdenes siendo el más estudiado el orden de los Trypanosomatida, Figura 1.1 (Moreira et al., 2004; d'Avila-Levy et al., 2015).

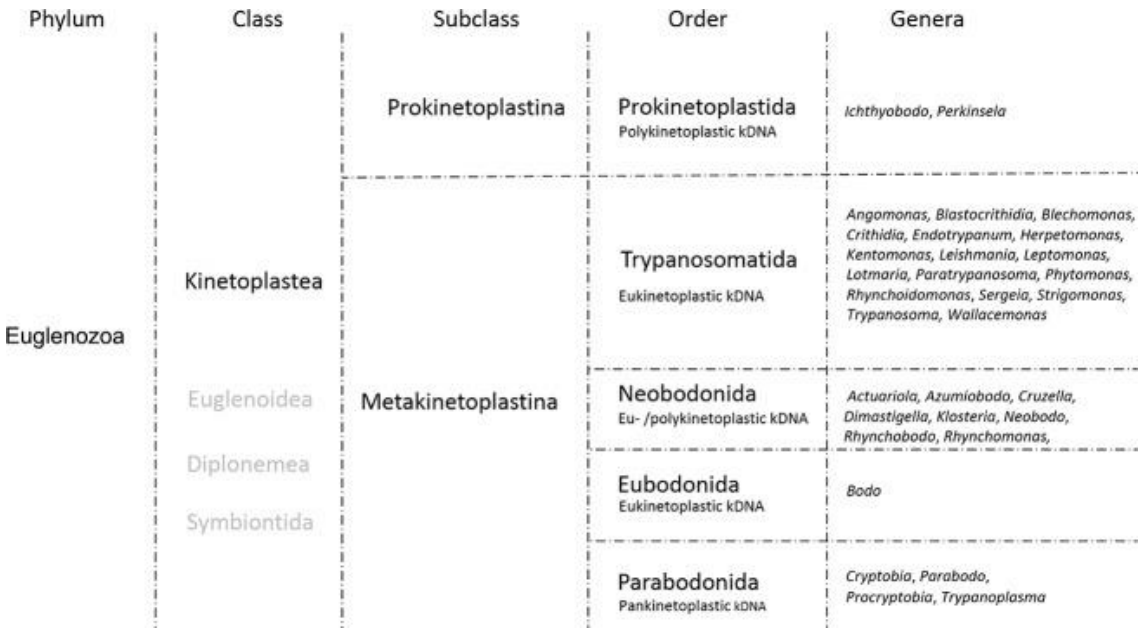


Figura 1.1. Taxonomía de la clase Kinetoplastea. Tomado de (d'Avila-Levy et al., 2015).

Los miembros del orden Trypanosomatidia son todos pertenecientes a la familia Trypanosomatidae. Los miembros diexénicos de esta familia son parásitos obligatorios e incluyen los causantes de la enfermedad del sueño africana (*Trypanosoma brucei*), la enfermedad de Chagas (*Trypanosoma cruzi*) y varias formas de leishmaniasis (*Leishmania*). Presentan ciclos de vida complejos involucrando el pasaje a un hospedero vertebrado desde un vector (principalmente insecto), presentando dramáticas diferencias en cuanto al ciclo de vida y estrategias de supervivencia (Lukes et al., 2018). Además de humanos, existe una amplia gama de organismos que pueden ser infectados por miembros de esta familia, entre los que encontramos animales domésticos y salvajes, y también plantas, como por ejemplo varias especies de *Pythomonas* (Jaskowska et al., 2015).

1.2. Enfermedad de Chagas

La enfermedad de Chagas, también conocida como tripanosomiasis americana, es una zoonosis que afecta a las poblaciones rurales pobres de Latinoamérica. Su agente etiológico es el parásito protozoario *Trypanosoma cruzi* (*T. cruzi*) el cual se transmite al hospedero mamífero, en el que se desarrolla la patología, a través de insectos triatominos hematófagos que funcionan como vectores. Esta enfermedad fue descrita por primera vez por el investigador Carlos Chagas en el año 1909, quien fue el pionero tanto en la identificación como la caracterización de este parásito (Chagas, 1909). Constituye un problema para la población americana en donde se estima que hay entre 6 y 7 millones de personas afectadas (World Health Statistics (WHO), https://www.who.int/news-room/fact-sheets/detail/chagas-disease-(american-trypanosomiasis)). Por esta razón la Organización Mundial de la Salud, la declaró como una de las 17 enfermedades tropicales desatendidas.

No existen actualmente tratamientos eficaces contra esta enfermedad, utilizando principalmente las drogas Benznidazol y Nurtifimox. Estas quimioterapias son poco eficientes, y altamente dañinas para el paciente, por lo que la búsqueda de blancos moleculares drogables sigue siendo un tema de principal importancia (Molyneux, 2014). La patología suele presentar dos fases claramente distinguibles, una fase aguda que

tiende a tener una duración de entre 40 y 60 días, y una fase crónica que puede persistir durante toda la vida del individuo. La fase de infección aguda se caracteriza por la presencia de síntomas de intensidad variable que pude incluir, disnea, fatiga, fiebre, vómitos, diarrea, edemas en las extremidades inferiores o en la cara, dolores abdominales y/o torácicos entre otros. La muerte en la fase aguda raramente se produce en adultos (en general es asintomática), pero en niños puede llevar a severas complicaciones neurológicas (WHO). Las manifestaciones de la enfermedad aguda se resuelven espontáneamente en alrededor del 90% de los individuos infectados, incluso si la infección no se trata con fármacos tripanocidas (Rassi et al., 2010). Durante el período de infección aguda todos los tipos de células nucleadas pueden ser blancos donde se aloje el parásito. Esta fase presenta alta parasitemia, pero con el desarrollo de la respuesta inmune se logra disminuirla, aunque no acabar con ella, marcando el final de la fase. Debido a que la eliminación no es completa, la enfermedad puede avanzar hacia la fase crónica, donde en un 20-40% de los casos se observan trastornos cardíacos y alteraciones digestivas (Coura and Borges-Pereira, 2010).

1.3. Generalidades de *T. cruzi* y su ciclo de vida

T. cruzi es un organismo que pertenece al género Trypanosoma y constituye una población heterogénea, donde las diferentes cepas usualmente usadas a nivel experimental presentan perfiles genéticos y proteicos polimórficos. Si bien, a lo largo del tiempo, las distintas cepas de *T. cruzi* han sido catalogadas en distintas agrupaciones, actualmente, se reconoce la existencia de seis linajes principales. Estos linajes están clasificados en unidades de tipificación discretas (DTUs), y se nombran del I al VI (Zingales et al., 2009). Recientemente se ha agregado un séptimo grupo muestreado en murciélagos y llamado TcBat (Cosentino and Aguero, 2012). Debido a la evolución predominantemente clonal del parásito, estas DTUs son bastante estables, constituyendo un marco útil para el análisis epidemiológico y evolutivo (Burgos et al., 2013).

T. cruzi presenta un ciclo de vida muy complejo, que involucra un hospedero mamífero y un insecto hematófago (géneros Triatoma, Rhodinius o Panstrongylus), que actúa como

vector. Durante su ciclo de vida presenta formas infectivas y no infectivas, como así también, formas replicativas y no replicativas. En el hospedero mamífero podemos encontrar las formas amastigotas y tripomastigotas sanguíneos, mientras que en el vector se encuentran los estadios epimastigotas y tripomastigotas metacíclicos. Estos 4 estadios, presentan diferencias notorias tanto a nivel metabólico como morfológico (tamaño y forma celular, posición del núcleo y kinetoplasto, largo y forma del flagelo, etc.).

En el momento que el insecto se alimenta de la sangre de un mamífero infectado incorpora mayormente tripomastigotas sanguíneos, los cuales se localizan en el tracto digestivo medio del vector diferenciándose en la forma epimastigota. Luego, los parásitos avanzan a través del tracto digestivo hasta ubicarse en la ampolla rectal del vector. Estos se encuentran ahora en su estadio tripomastigota metacíclicos, forma infectiva y quiescente, responsable de la transmisión de la enfermedad. Si el insecto se alimenta de un hospedero mamífero, los tripomastigotas metacíclicos presentes en las heces pueden penetran al torrente sanguíneo a través de mucosas o de la piel dañada. Una vez dentro del hospedero son internalizados por las células cercanas al sitio de entrada a través de vacuolas endocíticas (parasitóforas). Mediante la acidificación y ruptura de estas vacuolas los parásitos serán liberados al citoplasma celular (la ruptura se da por la acción de la proteína lítica TcTox en conjunto con la actividad trans-sialidasa (TS)) (Tyler and Engman, 2001). Una vez libres, los parásitos terminan su transformación a amastigotas (replicativas, intracelulares). Luego de varias replicaciones intracelulares, ocurre la lisis celular y liberación de los parásitos los cuales son nuevamente tripomastigotas sanguíneos, que tienen la posibilidad de infectar nuevas células (Tyler and Engman, 2001) (Figura 1.2). La posibilidad de infectar células que presenta el estadio amastigota ha sido revisada y discutida (Mortara et al., 2005).

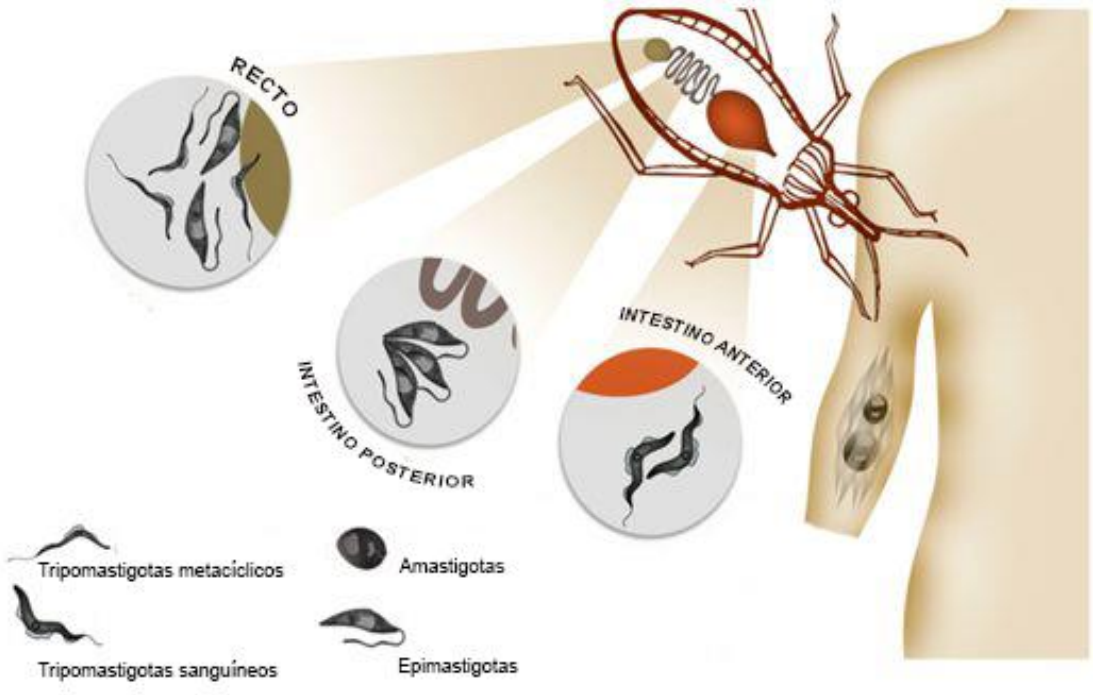


Figura 1.2. Ciclo de vida de *T. cruzi*. En el hospedero vertebrado, el tripomastigota metacíclico infectivo entra en contacto con los mamíferos a través de heridas o exposición a la mucosa y, una vez internalizado, se diferencia a su forma replicativa amastigota. Después de una multiplicación intensa, rompe las células del hospedero vertebrado y la forma tripomastigota sanguínea se expone al torrente sanguíneo. Posteriormente, un triatomino se alimenta del hospedero infectado, y la forma de parásito tripomastigota sanguíneo se diferencia en la forma epimastigota en su intestino posterior. En la zona rectal del triatomino, la forma epimastigota se adhiere y luego se diferencia en la forma de tripomastigota metacíclico cerrando el ciclo. Adaptado de (Amorim et al., 2017).

1.4. Características estructurales

Debido a la distancia filogenética respecto a otros eucariotas, los tripanosomátidos presentan características biológicas excepcionales, siendo uno de los géneros más ancestrales que se han estudiado (Smith and Parsons, 1996). Estas características hacen que se sitúe a estos organismos en el límite genético entre los organismos eucariotas y procariotas.

Entre las características más sobresalientes de estos parásitos se encuentra la presencia de una única pero muy desarrollada mitocondria, que abraca gran parte del cuerpo celular. La mitocondria presenta los compartimentos clásicos (membrana (externa e interna), espacio intermembrana, matriz mitocondrial, etc.) y puede ocupar un volumen variable celular, de acuerdo a la condición en que se encuentre la célula (de Souza et al.,

2009). De particular importancia, son las variaciones morfológicas y composicionales a la que está sometida la mitocondria durante el ciclo de vida. En los estadios amastigota y epimastigota la mitocondria adopta una estructura tipo pesuña, en tripomastigotas sanguíneos presentan cómo una doble estructura, mientras que en los metacíclicos adquiere una forma tubular opuesta a la membrana ondulante (Maria et al., 1972; Newberry and Paulin, 1989). El ADN mitocondrial representa aproximadamente un 25% del ADN celular. Este ADN extra nuclear se localiza en una estructura distintiva denominada kinetoplasto. Contiene dos tipos de ADN circular, los maxi y mini círculos, que se encuentran conectados entre sí y físicamente ligados al cuerpo basal, el cual se encuentra en la base del flagelo y se localiza de forma perpendicular a su eje (Souto-Padron et al., 1984). Coocurren miles de mini círculos y unas pocas decenas de maxi círculos, que varían en el primer caso de un largo aproximado de 1kb hasta en el caso de los segundos unos 25kb (Hoffmann et al., 2016). Los mini círculos contienen las secuencias que codifican para los ARN guías, los cuales son necesarios para la edición (*editing*) de transcriptos mitocondriales, fenómeno descrito por primera vez en tripanosomátidos (Benne et al., 1986; Aphasizhev et al., 2003; Read et al., 2016). Mientras que los maxi círculos codifican los ARNr y proteínas mitocondriales, al igual que en eucariotas superiores (Hoffmann et al., 2016).

El núcleo, a diferencia de la mitocondria, presenta una organización estructural semejante al de las células eucariotas típicas, midiendo cerca de 2.5 µm de diámetro. En los estadios replicativos, *T. cruzi* se reproduce mediante fisión binaria, presentando un núcleo esférico y un nucléolo central. Si bien la forma es similar en estos estadios, presentan gran diferencia en tamaño. Aún no se conoce con exactitud a que se debe esta diferencia, pero se sugiere que es por un cambio de ploidía. Mientras que en tripomastigotas, donde existe una clara disminución de la actividad transcripcional y son no replicativos, ocurre una dramática reducción del tamaño celular. El núcleo presenta una forma alargada con alto contenido de heterocromatina y carente de nucléolo (Elias et al., 2001; Schenkman et al., 2011). La envoltura nuclear es conservada durante la división celular, determinando que la segregación cromosómica ocurra dentro del núcleo no observándose condensación del ADN nuclear a nivel de cromosoma metafásicos (Solari, 1995).

La superficie celular de *T. cruzi* está conformada por la membrana celular, con su bicapa lipídica, y un conjunto de compuestos azucarados que conforman el glicocálix del parásito. La membrana no es homogénea, existiendo al menos tres macrodominios (el cuerpo celular, flagelo, y la bolsa paraflagelar).

Existen muchas otras particularidades de estos organismos que no serán profundizadas aquí, pero simplemente y a modo de ejemplo se mencionan dos de ellas: el acidocalsisoma y glicosoma. Los acidocalsisomas son organelos que tienen la capacidad de transportar protones y calcio y han sido identificados en todos los miembros de la familia Trypansomatidae (Docampo et al., 2005; Docampo and Moreno, 2011). Participan en diversas funciones como el almacenamiento de calcio, magnesio, sodio, potasio, en la homeostasis del pH y en la osmorregulación junto a la vacuola contráctil. Por otra parte, los glicosomas son organelos que marcan una particularidad metabólica de estos organismos jugando un importante rol en la adaptación metabólica del parásito a los diferentes entornos a los que se expone durante su ciclo de vida. Varias etapas de la glucolisis ocurren aquí adentro (Cazzulo, 1994; Hannaert and Michels, 1994; Hannaert et al., 2003).

1.5. Organización genómica

Muchas de estas particularidades han podido ser caracterizadas más a fondo dada la disponibilidad de los genomas de tripanosomátidos modelos (*T. brucei*, *T. cruzi* y *L. major*) publicados en el año 2005 por el grupo de El-Sayed (El-Sayed et al., 2005b). Al conjunto de estos tres organismos se lo conoce popularmente como TriTryps, y a partir de ellos se originó la base de datos más ampliamente utilizada en estos organismos, TriTrypDB (Aslett et al., 2010).

La cepa elegida de *T. cruzi* para ser secuenciada fue la cepa CL Brener (TcVI), que presentó la particularidad de ser una cepa híbrida entre la cepa Esmeraldo (TcII) y No-Esmeraldo (TcIII) (El-Sayed et al., 2005b). El ensamblado del genoma (~55 Megabases (Mb) para el genoma haploide) presentó dificultades extras a la previamente mencionada, debido a la gran cantidad de secuencias repetidas. El genoma haploide contiene aproximadamente

12000 genes que codifican para proteínas (1994 genes para ARN y 3590 pseudogenes) y en base a homología con genes correspondientes a otras proteínas previamente caracterizadas o por dominios funcionales conocidos, se pudo asignar una función probable al 50% de los genes que codificaban para estas proteínas (Choi and El-Sayed, 2012). Esta situación es similar al resto de los tripanosomátidos lo que implica que la mitad de las proteínas de estos organismos no tienen función asignada. También vale la pena resaltar aquí que al menos el 50% del genoma consiste en regiones de secuencias repetidas, que consisten en retrotransposones, repetidos en tándem, subteloméricos y familias multigénicas. Las familias multigénicas más expandidas, son proteínas tipo TS, mucinas, metaloproteasas, DGF-1, proteínas RHS y las proteínas de superficie asociadas a mucinas (MASP). Cada una de estas familias incluye varios cientos de genes los cuales pueden ser expresados simultáneamente. Las familias multigénicas que codifican para antígenos de superficie son parte de una estrategia clave de evasión al sistema inmune y otros procesos relacionados con la infección.

Si bien existe una gran distancia filogenética entre los TriTryps, se determinaron mediante búsqueda de ortología aproximadamente 6200 genes comunes (Figura 1.3). Se observaron una identidad a nivel de aminoácidos del 57% entre *T. cruzi* y *T. brucei* mientras que las proteínas de *L. major* tienen una identidad del 44% con los otros dos tripanosomátidos (El-Sayed et al., 2005b).

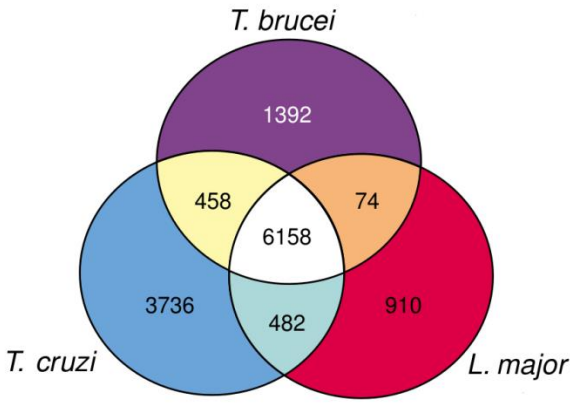


Figura 1.3. Diagrama de Venn mostrando el número de genes compartidos entre las diferentes especies secuenciadas. Extraído de (El-Sayed et al., 2005b).

La secuenciación de los genomas ha podido confirmar la organización de sus genes en agrupamientos largos y direccionados (DGCs, *direccional gene clusters*) codificados sobre la misma hebra, los cuales a pesar de que su estructura recuerda a los operones procariotas, no presentan agrupación funcional (Palenchar et al., 2006). Esta observación fue inicialmente comprobada en el cromosoma 1 de *L. major*, que fue el primer cromosoma secuenciado de tripanosomátidos (Myler et al., 1999). Interesantemente, existe una alta conservación en la sintenia de los DGCs integrados por genes comunes entre los tres organismos (94%) (El-Sayed et al., 2005b).

Dada esta organización particular, la expresión de los genes nucleares que codifican proteínas ocurre por transcripción policistrónica dando lugar a transcritos primarios que incluyen varios genes en una misma molécula precursora (PTU, *policistronic transcription units*). Estos ARN primarios son procesados por un mecanismo intermolecular llamado trans-empalme que fue descrito por primera vez en los tripanosomátidos e involucra la adición de un miniexón (SL, *spliced leader*) de 39 pares de bases (pb) con estructura de CAP (caperuza) a las regiones 5‘ de los diferentes ARNm (Laird, 1989). La estructura CAP presenta un mayor número de modificaciones que en los eucariotas superiores, es denominada CAP-4 y consiste en una 7-metilguanosina además de grupos 2’ O-metilo en los cuatro primeros nucleótidos (Bangs et al., 1992; Tschudi and Ullut, 2002). La secuencia del miniexón proviene del extremo 5’ de un ARN nuclear pequeño (snRNA), el ARN SL, que está compuesto por 120 nucleótidos y no está poliadenilado (Agabian, 1990). La adición de la secuencia SL se produce en un sitio consenso constituido por un dinucleótido AG localizado corriente arriba del codón de iniciación a distancias variables (Agabian, 1990). Este proceso está acoplado a la poliadenilación en donde, a diferencia con los eucariotas superiores, en tripanosomátidos no se ha podido describir una secuencia consenso que actúe como señal. El procesamiento de ARN es co-transcripcional y se rige por un tracto de polipirimidina, ubicado entre dos marcos de lectura abiertos vecinos, que es la secuencia señal reconocida por los mecanismos de trans-empalme y poliadenilación (LeBowitz et al., 1993; Matthews et al., 1994). Este proceso co-transcripcional convierte los ARN policistrónicos en ARNm maduros monocistrónicos traducibles. Por lo anterior, el mecanismo de trans-empalme resulta

indispensable para que se generen los ARNm maduros traducibles (Blumenthal et al., 2002; Chen et al., 2013).

Una implicancia de la organización en DGCs es la existencia de sitios en el genoma donde se invierte el sentido transcripcional. Estas regiones que se encuentran entre dos DGCs consecutivos, se denominan regiones de cambio de hebra (SSR, *strand switch regions*) y juegan roles en el inicio de la transcripción (Smircich et al., 2017).

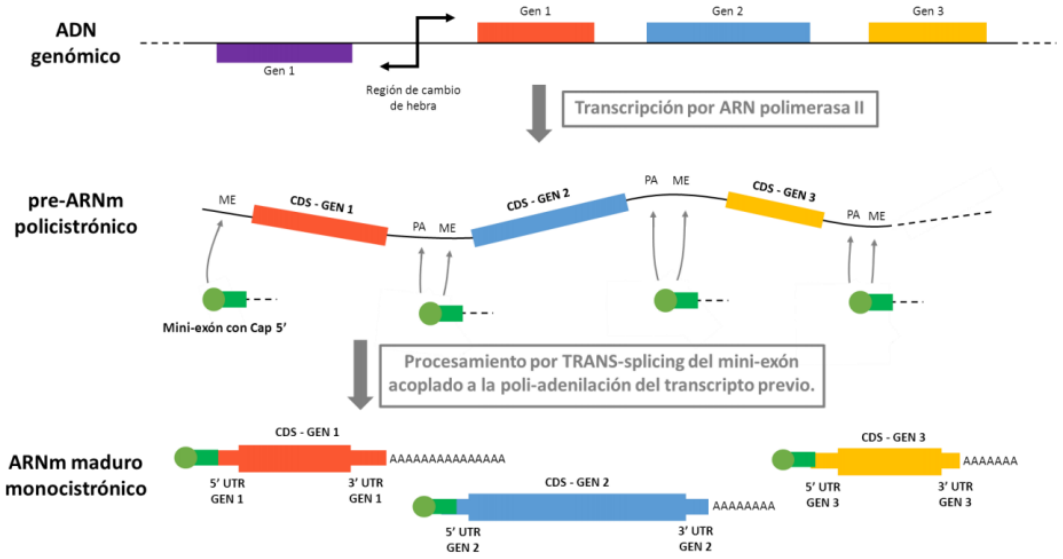


Figura 1.4. Expresión génica en tripanosomátidos. Se esquematizan las etapas de transcripción y procesamiento de los transcriptos primarios. El ARN del mini-exón también es transcripto por la ARN polimerasa II a partir de otra región genómica. Tomado de la tesina de maestría de Santiago Chavez, 2016.

1.6. Regulación de la expresión génica

La regulación de la expresión génica es necesaria para equilibrar la síntesis de los componentes proteicos necesarios para cumplir cualquier función celular. En eucariotas, la primera capa de regulación está a nivel de transcripción génica. Más allá de esta capa, existen varios mecanismos que regulan la expresión a nivel postranscripcional, siendo este, el principal nivel de control en tripanosomátidos. Consecuentemente con esta afirmación, se presume que la mayoría de los genes son transcritos constitutivamente observándose muy pocos posibles factores transcripcionales reguladores (Palenchar and Bellofatto, 2006). Sin embargo, y a pesar de la transcripción primaria común, los genes

individuales que pertenecen a la misma unidad policistrónica a menudo muestran diferentes niveles de transcripto en estado estacionario, evidenciando el hecho de que la regulación de la expresión génica opera principalmente a nivel post-transcripcional. En efecto, estudios de transcriptoma de *T. cruzi* a lo largo del ciclo de vida, mediante hibridación en microarreglos (Minning et al., 2009), han encontrado diferencias relativas de transcripto en estado estacionario.

Particularmente, las regiones no traducidas (UTRs) son esenciales para la estabilización o degradación del ARNm, modulando su vida media en diferentes estadios y/o condiciones (Vanhamme and Pays, 1995; Furger et al., 1997). Si bien ambas regiones son susceptibles a regulación, las regiones 3' UTRs no presentan el mismo nivel de presión selectiva que las 5' UTRs ya que estás últimas requieren acomodar la maquinara traduccional, lo que permite cierto grado de libertad para acomodar nuevas secuencias reguladoras (De Gaudenzi et al., 2003). En este sentido son fundamentales los motivos de secuencias y/o estructuras presentes en los mismos, permitiendo que proteínas de unión a ARN (RBPs, *RNA Binding Proteins*) regulen los niveles de estado estacionario de ARNm.

Hay muchos dominios de unión a ARN identificables, de los cuales el motivo de reconocimiento de ARN (RRM, *RNA Recognition Motif*) es el mejor caracterizado (Romaniuk et al., 2018). El genoma de *T. cruzi* codifica una larga variedad de RBPs que tienen un rol importante en la regulación post-transcripcional de los niveles de ARNm, así como en otros aspectos del metabolismo del ARN (El-Sayed et al., 2005b). Las proteínas RBPs pueden establecer interacciones con grupos de ARNs que comparten elementos en *cis* entre sí (ya se secuencia primaria o estructuras secundarias) que permiten la co-regulación de un grupo de ARNm que pueden cumplir funciones similares (Keene, 2007). Los dominios RRM pueden unirse a elementos en *cis* que varían entre 2 a 8 nucleótidos y pueden presentarse en proteínas involucradas en diversos eventos postranscripcionales, como el procesamiento, transporte, traducción, degradación y estabilidad del ARN (Maris et al., 2005). Varias líneas de evidencia apuntan hacia el papel de las RBPs en la regulación del inicio de los programas que actúan en los tripanosomas al comienzo de la diferenciación (Hendriks et al., 2001; Kolev et al., 2012; Wurst et al., 2012; Mugo and Clayton, 2017). Es así, que en tripanosomátidos se propone la figura de operones post-transcripcionales co-regulados por grupos de RBPs específicas (De

Gaudenzi et al., 2003; Ouellette and Papadopoulou, 2009; Li et al., 2012; De Gaudenzi et al., 2013).

Teniendo en cuenta los genomas de los TriTryp, se ha podido identificar un total de 139 proteínas de tipo RRM en *T. cruzi*, 75 en *T. brucei* (de las cuales únicamente a 18 pudieron asignarle función) y 80 en *L. major*, existiendo 77 proteínas comunes a los tres (De Gaudenzi et al., 2005). Actualmente, algunas de las proteínas con dominio RRM han sido caracterizadas (Manger and Boothroyd, 1998; Clayton, 2002; D'Orso and Frasch, 2002; Perez-Diaz et al., 2007; Guerra-Slompo et al., 2012; Perez-Diaz et al., 2013). Uno de los mecanismos de regulación más caracterizado son los mediados por motivos UREs (*Uridine Rich Elements*), los cuales han sido encontrados en varios ARN afectando su estabilidad (Haile et al., 2007). Los UREs parecen tener estructura y función similares a las secuencias AREs (*AU rich elements*) presentes en los mensajeros de otros organismos (von Roretz et al., 2011). Las proteínas implicadas en la regulación de la estabilidad de los ARN que presentan AREs se dividen en varias clases, según los dominios de unión a RNA: incluyen dominios KH, dedos de zinc y RRMs. Los ARE similares están presentes en varios ARNm regulados en los kinetoplástidos (De Gaudenzi et al., 2005), como los codificantes para las proteínas de superficie mucinas (D'Orso and Frasch, 2001) y las proteínas TcUBPs (D'Orso and Frasch, 2002). También se han observado otros motivos ricos en U, por ejemplo, la presencia de un elemento rico en U de 43 nucleótidos en un gran número de ARNm que controla su abundancia en el estadio amastigota del parásito (Li et al., 2012); los genes de la fosfoglicerato quinasa son regulados por un elemento rico en AU presente en el 3'UTR que desestabiliza el ARNm en la forma sanguínea (Quijada et al., 2002). Uno de los ejemplos más recientes, es el de la RBP10 de *T. brucei*, descrito por el grupo de Clayton. RBP10 se une a los ARNm específicos de procíclicos que contienen un motivo UAUUUUUU, dirigiéndolos a la represión traduccional y a su destrucción, promoviendo a su vez la diferenciación de la forma procíclica a la forma sanguínea (Mugo and Clayton, 2017). Los cambios de estadio disparados por la acción de RBPs ya había sido descrito previamente en *T. brucei* (Kolev et al., 2012). En este sentido, se ha evidenciado recientemente, que la proteína TcUBP1 promueve el inicio del proceso de diferenciación de epimastigotas a tripomastigotas metacíclicos en *T. cruzi* (Romaniuk et al., 2018).

En otros casos, la identidad de la señal regulatoria es menos clara, postulándose que es la combinación de motivos presentes en el UTR la responsable de la regulación estadio específica (Hotz et al., 1997; Coughlin et al., 2000; Pastro et al., 2013; Chavez et al., 2017).

Aunque, como hemos mencionado, normalmente se asume que los motivos regulatorios se concentran en las regiones 3'UTR, las regiones no traducidas en el 5' son también importantes en los procesos de control traduccional (Clayton, 2002). De hecho, se ha descrito que los genes de expresión constitutiva como los que codifican las proteínas ribosomales poseen regiones 5'UTR muy cortas, estando el codón de inicio exactamente después del final de la secuencia del miniexón (Jensen et al., 2014). Se postula que, de esta forma, se evitaría la presencia de reguladores negativos en esta región (Greif et al., 2013).

1.7. Regulación mediada por uORF

La mayor parte del conocimiento que tenemos acerca de la regulación traduccional dependiente de secuencia surge de experimentos que involucran mutagénesis sitio-dirigida de genes particulares, implicando que la información obtenida es contexto específica y difícil de generalizar. Sin embargo, recientemente y determinado por la aparición de los trabajos de *Ribosome Profiling* (o estudio de perfiles de huellas ribosomales) es posible estudiar todos los ARNm al mismo tiempo (Ingolia et al., 2011; Fritsch et al., 2012; Lee et al., 2012; Smircich et al., 2015). La regulación de este proceso estaría dada a nivel de la formación del complejo de iniciación de la traducción y del paso subsiguiente de elongación (McCarthy, 1998). Aunque normalmente se asume que los motivos regulatorios se concentran en las regiones 3'UTR, las regiones no traducidas en el 5' son también importantes en los procesos de control traduccional, existiendo mecanismos específicos como la presencia de marcos de lectura 5' río arriba del CDS principal (*uORF: upstream ORF*).

En particular, los uORF se definen como marcos de lectura abiertos, generalmente pequeños, que se inician dentro de la región no traducida 5 ' de un ARNm maduro. El inicio traduccional dentro del 5' UTR del ARNm da lugar a un marco abierto de lectura

que solapa, o bien, termina antes del codón de inicio del CDS. Si bien la eficiencia traduccional disminuye en la mayoría de los genes que contienen uORF, existen evidencias de algunos que aumentan la eficiencia de los genes asociados, aunque son casos más bien excepciones (Griffin et al., 2001; Vattem and Wek, 2004; Chen et al., 2010).

Los uORF pueden servir como elementos de respuesta rápida, permitiendo que las células adopten inmediatamente una producción de proteínas acorde a condiciones ambientales alteradas (Calvo et al., 2009; Lawless et al., 2009). En esta línea, se ha visto mediante técnicas de *Ribosome Profilling* que el 50% de los transcriptos humanos poseen inicios traduccionales río arriba del CDS principal lo que sugiere un fuerte rol regulatorio mediado por uORFs. En los estudios pioneros se observó que la mayoría (más de un 70%) de los inicios traduccionales estaban mediados por inicios no canónicos (no AUG), principalmente por codones que difieren en una base con este (Ingolia et al., 2011; Fritsch et al., 2012; Lee et al., 2012). Esto implicó una gran sorpresa debido a que se consideraba que la maquinaria de inicio de traducción favorecía fuertemente el inicio en AUG y que los inicios no-AUG eran raros y gen específicos. Incluso se había determinado que la eficiencia de iniciación no AUG, en un contexto de nucleótidos óptimo, es al menos 20 veces menor que la de AUG (Clements et al., 1988). Recientemente, a través del metaanálisis de datos de *Ribosome Profiling* se sugirió que la alta proporción de codones no-AUG que iniciaban la traducción eran artefactuales, y que la eficiencia de traducción de los codones no-AUG era muy débil. Esto implica que el potencial regulatorio de los uORF depende en gran manera de la presencia de codones AUG 5’ río arriba del AUG del CDS principal (Michel et al., 2014).

Existen muchas otras propiedades estructurales y funcionales que determinan el impacto regulatorio de un uORF. Su éxito se basa en el modo que afectan al mecanismo de escaneo. El rol predominante del escaneo en el inicio de la traducción fue establecido primariamente por Kozak, que determinó a través de varios estudios que la mayor parte de los ARNm eucarióticos son monocistrónicos, no poseen AUG en el extremo 5’ del sitio iniciador, y qué el extremo CAP-7-metil-guanosil (m7G-CAP) estimula la traducción. Estas observaciones fueron claves para la formulación de la hipótesis del escaneo (Kozak, 1978). Este mecanismo comienza con la disociación del complejo 80S ribosomal en las

subunidades libres 40S y 60S. La subunidad menor se une al complejo de pre-iniciación 43S (PIC), que contiene el ARNt iniciador en un complejo ternario (TC), con el factor de inicio de la traducción eucariótico 2 (eIF2) en su forma de unión a GTP (eIF2-GTP-Met-ARNti). Esta formación está estimulada por otros factores de inicio de la traducción, entre los que se encuentran, eIFs 1, 1A, 5 y el complejo eIF3. El complejo 43S PIC se une a la región 5' del ARNm a través de la estructura m7G-CAP, estimulado por el complejo eIF4F, el cual está compuesto por eIF4E (proteína de unión a caperuza), eIF4G, la helicasa de ARN eIF4A y por la proteína de unión a poli-A (PABP) (Hinnebusch, 2011). Una vez ensamblado, el complejo de pre-iniciación escanea el ARNm en dirección 3' hasta que el anticodón de metionina encuentra un inicio de codón funcional en un contexto de secuencia favorable. El reconocimiento del codón AUG determina el cese del escaneo mediante la hidrólisis irreversible del GTP unido a eIF2 en el TC, mediado por la proteína de activación de GTPasa (eIF5-GAP) produciendo un complejo 48S PIC estable. Posteriormente, se une la subunidad 60S ribosomal favorecido por la liberación del complejo eIF5-GDP y otros factores de inicio de la traducción, formando el complejo de iniciación (IC) 80S el cual contiene el ARNt iniciador pareado al codón AUG en el sitio P ribosomal y pronto para iniciar la fase de elongación en la síntesis proteica (Wethmar, 2014).

En la base de datos uORFdb (Wethmar et al., 2014), la cual categoriza todas las publicaciones relacionadas a uORF, se mencionan numerosos mecanismos mediante los cuales pueden afectar la eficiencia traduccional (TE, definida como el número de huellas ribosomales normalizadas por molécula de mensajero) de los genes asociados. Por ejemplo, Kozak propuso que la TE estaba fuertemente determinada por el contexto del AUG iniciador. En particular describió la presencia de una secuencia nucleotídica denominada secuencia consenso Kozak (GCCGCC(A/G)CCAUGG (A/G)) en los inicios eficientes. Sin embargo, esta secuencia en realidad es poco conservada en eucariotas, conteniendo únicamente un 0.2% de los vertebrados la misma secuencia y la divergencia es mayor a medida que nos alejamos de este grupo (Kozak, 1978; 2002; Nakagawa et al., 2008). En particular en kinetoplástidos no se observa la secuencia Kozak conservada, pero si está reportado un enriquecimiento en adeninas en el extremo 5' al AUG (Nakagawa et al., 2008). También se ha reportado que el contexto del AUG iniciador a nivel de

estructura secundaria influye en la eficiencia traduccional del uORF y del CDS principal, sin embargo, este efecto parece ser determinante en organismos multicelulares complejos (i.e. humanos) y no así en levadura (Vilela and McCarthy, 2003; Chew et al., 2016).

Se han reportado numerosas observaciones de que el potencial regulatorio de un uORF es dependiente de la posición dentro de la región 5' UTR del ARNm en cuestión (Kozak, 1987; Wethmar, 2014; Chew et al., 2016; Fervers et al., 2018). Por ejemplo, la eficiencia de iniciación del codón AUG se deteriora progresivamente si la longitud al extremo 5' UTR se reduce por debajo de 15-20 nucleótidos. Además, la distancia del codón de parada del uORF al codón AUG del CDS principal tiene un gran impacto en el potencial regulatorio, debido a que una vez que se traduce un uORF la subunidad ribosomal menor puede seguir unida al ARNm y continuar escaneado hasta reiniciar la traducción en el codón iniciador principal (Kozak, 2002; Chew et al., 2016). También, se ha observado que la TE decrece a medida que la distancia entre el uORF y el CDS decrece (Kozak, 1987; Chew et al., 2016). Cuanto menor es la distancia menor va a ser el tiempo para readquirir el ARNt iniciador, por ende, una mayor distancia entre el codón de parada y el AUG del CDS principal facilita el reinicio al brindar ese tiempo extra necesario para la captación del ARNt iniciador (Fervers et al., 2018). Finalmente, la disminución de la TE parece ser mayor si el uORF solapa al CDS principal. Si bien, transformaciones de uORF no solapantes en solapantes lleva frecuentemente a un aumento en la represión de la TE del CDS, aún no se ha visto una correlación positiva entre solapamiento y reducción de la TE en uORF que ocurran naturalmente (Wethmar, 2014). El largo del uORF es otro de los factores que parece influir en el potencial represivo. Si un ribosoma inició la traducción en un uORF, la capacidad de reiniciar rio abajo está fuertemente influenciada por el número de codones que debe atravesar antes de llegar al codón de parada. En *S. cerevisiae* se ha observado la capacidad de reiniciar de los ribosomas cae a cero para uORF con más de 35 codones de largo (Kozak, 2001; Rajkowitsch et al., 2004).

En kinetoplástidos, los uORF quedan determinados entre el sitio de trans-empalme y el siguiente codón de parada en marco, que puede ser solapante o no solapante al CDS principal. El posible rol regulatorio que tienen los uORF en kinetoplástidos ha sido estudiado a través de la utilización de genes reporteros y mediante el análisis de datos

de *Ribosome Profiling* y proteómica (Siegel et al., 2005; Vasquez et al., 2014; Fervers et al., 2018). La determinación de uORF en tripanosomátidos se ha basado en la presencia de un uAUG y el siguiente codón parada, sin restricciones de otro tipo (Jensen et al., 2014; Fervers et al., 2018). Jensen et al., determinó que el 11% de los genes analizados (7331) de *T. brucei* presentan uORFs comparado al 22% (4909 genes analizados) reportado por Siegel et al., mientras que Fervers et al., encontró que el 29% de los genes los presentan en *T. congolense*. Estas variaciones pueden deberse a diferencias en la forma de determinar los 5' UTRs y definir los uORFs reguladores.

1.8. Familias proteicas diferencialmente traducidas entre los estadios epimastigota y tripomastigota metacíclico.

Los análisis de los traductomas, realizados por el grupo, han detectado diferencias significativas en los niveles de expresión y eficiencias traduccionales de varios genes, que permiten la identificación de grupos regulados específicamente a este nivel. Particularmente, dos grupos resultaron fuertemente controlados a través de la modulación de su eficiencia traduccional: los genes que codifican las proteínas ribosomales (PR) y los genes pertenecientes a la superfamilia de las trans-sialidasas (Smircich et al., 2015).

1.8.1. Proteínas trans-sialidasas

La habilidad de *T. cruzi* para sobrevivir en el hospedero mamífero, se debe en parte a la presencia de una diversa membrana de superficie. En gran medida, la diversidad está dada por la expansión masiva de genes que codifican para proteínas polimórficas de superficie, como las trans-sialidasas, proteínas asociadas a mucinas (MASP), mucinas (MUC), entre otras. Las TS son una de las familias que presentan mayor expansión en *T. cruzi* contándose más de 1400 genes (se ha visto que su número varía mucho de acuerdo a la cepa analizada) (Freitas et al., 2011).

Las TS son proteínas unidas a motivos GPI presentes en la superficie de *T. cruzi*, aunque también aparecen en la superficie de *T. brucei*. Estás proteínas cumplen una función clave dada la incapacidad del parásito de sintetizar ácido siálico (AS) *de novo*. Su rol consiste en transferir residuos de AS (proceso de sialilación) gracias a la actividad de una sialidasa modificada que, en lugar de hidrolizar el ácido siálico, transfiere residuos de sialilo unidos a alfa (2-3) de sialoglicoconjugados y proteínas del hospedero, a las proteínas MUC (Schenkman et al., 1991; Parodi et al., 1992; Frasch, 2000; Vercelli et al., 2005; Buscaglia et al., 2006; Mucci et al., 2006). Constituyendo uno de los mecanismos más elegantes de sialilación de la superficie celular en la naturaleza. Sin embargo, solo un subgrupo pequeño de esta familia conserva actividad enzimática.

Los glicoconjugados sialilados decoran la superficie de todas las células de mamíferos. Debido a su presencia y abundancia ubicuas, los AS tienen muchos efectos biofísicos importantes. AS es un grupo de monosacáridos de 9 carbonos estructuralmente diversos con estructuras anulares heterocíclicas. Tiene una carga negativa a través de un grupo ácido carboxílico unido al anillo, así como otros grupos químicos, incluidos los grupos N-acetilo y N-glicolil, que confieren diversas actividades biológicas a las glicoproteínas y glicolípidos, como la promoción de interacciones entre células o el enmascaramiento de los sitios de reconocimiento debido a su carga negativa.

Aunque los AS son los principales azúcares de la superficie celular en el linaje de los deuterostomados, también se encuentran en otras ramas de la vida (Cohen and Varki, 2010), incluidos microorganismos como virus, bacterias y protozoos (Els et al., 1989; Matrosovich et al., 2015).

Las proteínas de la familia de las TS han sido consideradas como un importante factor de virulencia, ya sea por su capacidad para amortiguar la inmunidad como para mediar la interacción entre el parásito y las células hospederas (Buschiazzo et al., 2012; Mendonca-Previato et al., 2013). Varios estudios han demostrado que la forma tripomastigota de *T. cruzi* presenta proteínas que pueden modular la respuesta inmunitaria del hospedero durante las primeras etapas de la infección (DosReis, 2011). Es un hecho bien establecido que en las formas tripomastigotas la TS se libera dinámicamente al medio extracelular. Esto conduce a una distribución sistémica de la enzima a través del torrente sanguíneo. La vida media de esta proteína liberada es bastante extensa debido a la presencia de un

dominio C-terminal denominado SAPA (*Shed Acute Phase Antigen*) (Alvarez et al., 2004). Interesantemente, Ribeirao et al. propuso que la presencia de este dominio C-terminal es una estrategia evolutiva adoptada por el parásito para prevenir la producción temprana de anticuerpos contra el dominio N-terminal, responsable de la actividad catalítica (Ribeirao et al., 1997).

Si bien, la actividad enzimática de las proteínas TS se propuso hace más de 30 años como un factor clave para la patogénesis de *T. cruzi*, ahora se sabe que las TS enzimáticamente inactivas (iTS) pueden jugar un papel importante en la biología del parásito. Como se ha mencionado previamente, existe una mayor cantidad de iTS que de TS acitvas (aTS). La diferencia entre estas dos clases de TS radica únicamente en el cambio en un único aminoácido, como demostró el grupo de Cremona (Cremona et al., 1995). La comparación entre las secuencias de aminoácidos de aTS e iTS muestra variaciones en 20 residuos, aunque la inactivación se debe totalmente a la única sustitución crucial (T/C); mientras que aTS tiene un residuo de tirosina en la posición 342 (Tyr342), iTS muestra un residuo de histidina (His342) en la misma posición (Cremona et al., 1995). Se ha demostrado, que si bien inactiva, las iTS tienen propiedades de adhesina, jugando un rol importante en la interacción hospedero-patógeno (Freire-de-Lima et al., 2015).

Trabajos iniciales para determinar la variabilidad polimórfica de este grupo han descrito cuatro tipos de TS según la similitud de secuencia y propiedades funcionales. El grupo I contiene trans-sialidasas activas, denominadas TCNA y SAPA (antígeno de fase aguda eliminada), y proteínas TS expresadas en epimastigotas. Los miembros del grupo II, no tienen actividad TS pero son capaces de unirse a β-galactosa, laminina, fibronectina, colágeno, y su función está vinculada a la unión e invasión de la célula hospedera. El tercer grupo abarca proteínas involucradas en la regulación del sistema del complemento. Finalmente, el grupo IV, no tiene una función descripta y está incluido en la superfamilia TS porque contiene el motivo VTVxNVxLYNR conservado (Freitas et al., 2011). Más recientemente, el grupo de Bartholomeu estudió la diversidad de este grupo, mediante el análisis de agrupamientos considerando la secuencia codificante de 505 miembros (y 300 pares de base río abajo del codón de parada) (

Figura ***1.5***). Como resultado determinaron la existencia de 8 grupos de proteínas TS. Los grupos I-IV coincidieron con los previamente identificados, definiéndose además 4

nuevos grupos (V-VIII). De los 505 miembros analizados se observó que en un 96% contenían el motivo VTVxNVxLYNR (o una versión degenerada de este; siendo x cualquier aminoácido). Por lo tanto, confirmaron que este motivo es una firma de la familia TS que se encuentra en todos sus miembros. Otros motivos que usaron para caracterizar los distintos miembros de esta familia fueron el motivo ASP-Box, común en virus y bacterias, y el motivo FIRP (xRxP) involucrado en la unión del grupo carboxilato del ácido siálico y secuencias repetidas en la región C-terminal (i.e., SAPA). Además, observaron que los niveles de expresión de las proteínas TS no son homogéneos entre y dentro de los distintos grupos. Pero si pudieron determinar la existencia, para unos pocos genes, de una relación entre el perfil de expresión y la similitud de sus regiones 3' UTRs (Freitas et al., 2011).

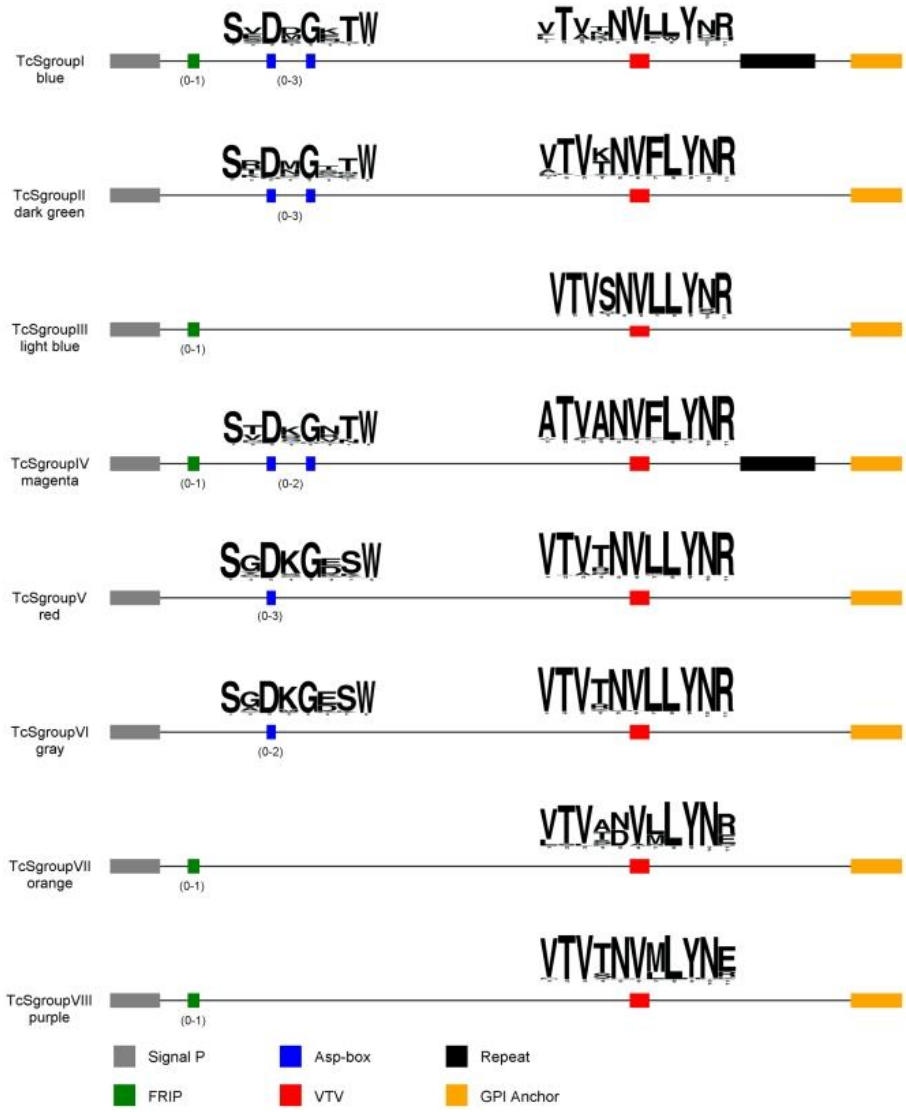


Figura 1.5. Estructura funcional de los agrupamientos de proteínas pertenecientes a la superfamilia de trans-sialidasas. Tomada de (Freitas et al., 2011).

Sin embargo, más recientemente, mediante el estudio de genomas ensamblados con tecnologías de secuenciación de lecturas largas, se puso en duda la clasificación previamente mencionada, indicando que esta no es lo suficientemente robusta o que podría ser específica de la especie (Berna et al., 2018).

1.8.2. Proteínas ribosomales

El ribosoma es una gran partícula ribonucleoproteica que sintetiza proteínas en toda la célula, utilizando a los mensajeros de ARN como molde y a los ARN de transferencia conjugados a aminoacil como substratos. El ribosoma eucariota se encuentra compuesto por dos subunidades. La subunidad mayor (LSU) 60S, típicamente, contiene tres tipos de ARN ribosomales (ARNr) (5S, 25S (levadura; 28S en humanos) y 5.8S) y más de 40 (46 levadura; 47 en humanos) proteínas, mientras que la subunidad menor (SSU) 40S tiene una sola molécula de ARNr (18S) y más de 30 proteínas diferentes, que conjuntamente forma el ribosoma 80S.

La biogénesis del ribosoma está muy conservada a través del árbol filogenético eucariótico, sin embargo existen características únicas que han sido descritas en kinetoplástidos (Hashem et al., 2013; Liu et al., 2016; Shalev-Benami et al., 2016). El ARNr equivalente al 25S (en levaduras) es procesado en seis fragmentos conocidos como LSU-α, LSU-β, y ARNsr (pequeños ARNr) 1-4 (Hashem et al., 2013; Liu et al., 2016; Shalev-Benami et al., 2016). Las interacciones entre proteínas ribosomales y ARNr tienen características únicas favorecidas por segmentos expandidos (ES) en ambas moléculas (Liu et al., 2016) determinando que los ribosomas de tripanosomátidos sean particularmente grandes (1.3 veces más grande que el promedio) (Ayub et al., 2009). Este resultado fue comprobado en (Ayub et al., 2009), observando que las PR de la subunidad mayor de *T. cruzi* son, en promedio, 20 aminoácidos más largas que las proteínas correspondientes de *S. cerevisiae*. Para la subunidad menor, las proteínas de *T. cruzi* son en promedio 10 aminoácidos más largas. Las regiones adicionales están generalmente en

los extremos N o C-terminales (Ayub et al., 2009). Un caso notable de extensión son las proteínas que forman parte del canal de salida ribosomal, en particular L19 la cual presenta una extensión C-terminal de 168 aminoácidos. Se sugiere que las extensiones son en respuesta a la falta de una región plana (clásica de eucariotas) en esta región (Ayub et al., 2009). Otras proteínas que presentan diferencias interesantes son las proteínas P ribosomales. Estas presentan aspectos distintivos en la región C-terminal con respecto a las proteínas P de otros eucariontes. Pl y P2 se desvían del consenso C-terminal de eucariontes superiores ya que la típica serina es reemplazada por un ácido glutámico, mientras que la región C-terminal de PO difiere completamente de este consenso y se parece a la proteína ribosomal de arquibacterias (Levin et al., 1993).

En *T. cruzi*, la mayor parte de las PR tienen su contraparte en levadura (excepto eL41 y eS31). Interesantemente, la proteína eL28 está presente en tripanosomátidos pero no así en levadura, lo que sugiere que el genoma eucariótica ancestral si la contenía. La identidad promedio de aminoácidos entre las PR de *S. cerevisiae* y *T. cruzi* es notablemente baja (~ 50%), teniendo en cuenta el alto grado de conservación del ribosoma 80S a través de la evolución (Ayub et al., 2009). Sorprendentemente, RACK1 (recientemente identificada como PR) no ha sido visualizado como parte de la estructura ribosomal como si pasa en el resto de los organismos eucarióticos estudiados, en donde se ha identificado su rol como regulador directo del inicio traduccional a través del reclutamiento de la proteína quinasa c (Sengupta et al., 2004). En tripanosomátidos se ha identificado un homólogo funcional de RACK1, lo que sugiere un aspecto más ancestral del funcionamiento ribosomal.

2. Objetivos

2.1. Objetivo general

Profundizar en la comprensión de la expresión génica en *T.cruzi*, caracterizando aspectos de la modulación traduccional en familias de expresión génica diferencial estadio específica.

2.2. Objetivos específicos

1. Optimización del análisis de datos de *Ribosome Profiling* de los estadios epimastigotas y tripomastigotas metacíclicos.

2. Desarrollo de herramientas bioinformáticas que permitan mejorar el análisis de genes diferencialmente expresados.

3. Generación de un *software* enfocado en kinetoplástidos que permita definir regiones UTRs de los ARNm.

4. Determinación de la regulación mediada por la presencia de uORFs en las regiones 5' UTRs de los mensajeros de *T. cruzi*.

5. Búsqueda de motivos de secuencia primaria y secundaria en regiones UTR en genes co-modulados de la familia de trans-sialidasas y de proteínas ribosomales.

3. Resultados y discusión

3.1. Optimización del análisis de datos de *Ribosome Profiling*

Utilizando la metodología de *Ribosome Profilling* nuestro grupo ha detectado diferencias significativas en los niveles de expresión y eficiencias traduccionales entre los estadios epimastigota y tripomastigota metacíclico de *T. cruzi* (Smircich et al., 2015). Brevemente, la mencionada metodología consiste en la purificación de la fracción polisomal seguida por un ensayo de protección a nucleasas. La digestión controlada genera pequeños fragmentos de ARNm, denominados huellas ribosomales (RFPs) de unos aproximadamente 30 nucleótidos, los cuales son posteriormente secuenciados y mapeados a los transcritos, obteniéndose un perfil de huellas ribosomales sobre cada mensajero (Ingolia et al., 2009; Eastman et al., 2018).

Dada las dificultades que presenta el alineamiento de huellas en genomas repetitivos como el de *T. cruzi*, en esta sección nos enfocamos en la puesta a punto de condiciones óptimas para esta tarea (objetivo específico 1). Esto es particularmente importante ya que los resultados obtenidos anteriormente por el grupo resaltan el rol de proteínas de familias multigénicas, las cuales son particularmente sensibles a errores de mapeo.

3.1.1. Tratamiento inicial de los datos

Una vez descargados los datos del SRA, se utilizó el programa cutadapt (Martin, 2011) para pre-procesarlos (ver sección 3.1.5 Estrategia). Estos pasos fueron realizados de forma equivalente a (Smircich et al., 2015). Seguidamente, se procedió a eliminar los fragmentos de ARN ribosomal que se purifican conjuntamente durante el aislamiento de huellas protegidas con ribosomas, lo que disminuye el rendimiento de lecturas de secuenciación útiles que se pueden obtener en los experimentos de Ribo-Seq. En total se eliminaron alrededor del 40% en concordancia con lo reportado en (Beaupere et al., 2017).

3.1.2. Selección de alineadores: Bowtie o ShortStack?

En los últimos años ha habido un incremento en la disponibilidad de técnicas que permiten secuenciar pequeños ARNs (sRNA-Seq, *small RNA Sequencing*), por ejemplo, microARNs, pequeños ARNs de interferencia, ARNs asociados a Piwi, etc. Sin embargo, el alineamiento de pequeños fragmentos de ARNm a un genoma de referencia sigue siendo un desafío (Johnson et al., 2016). La principal problemática radica en la prevalencia de lecturas de mapeo múltiple (MMAP). Este fenómeno sucede cuando para una misma lectura pueden existir varias regiones de alineación con la mejor puntuación en el genoma de referencia.

En *T. cruzi*, la naturaleza repetitiva del genoma (El-Sayed et al., 2005b; Berna et al., 2018) hace que la determinación de niveles individuales, en especial al de la superfamilia de TS, no sea una tarea trivial. Los MMAP a menudo se tratan de una manera simplista, ya sea ignorando la lectura, seleccionando al azar una posición o seleccionando y cuantificando todos los sitios. En nuestro caso ignorarla no es una opción ya que nuestro objetivo es poder mejor la determinación de conteos en las regiones repetidas, seleccionar al azar solamente incrementaría la tasa de error y contándola tantas veces como aparece conduciría a una representación excesiva de la abundancia de *loci*, y en realidad está en contra de la asunción de todos los paquetes que realizan expresiones diferenciales con los datos de conteo (Pantano et al., 2011).Por otra parte, los datos de secuenciación masiva producidos en (Smircich et al., 2015), fueron obtenidos mediante la tecnología SOLiD, la cual produce un formato de lecturas basados en espacio de color (*colorspace*) que virtualmente ha dejado de utilizarse. Esto implica que la mayoría de los programas de reciente desarrollo no presentan métodos para el alineamiento de lecturas de este tipo.

En el presente trabajo, hemos decidido analizar dos alineadores distintos, Bowtie (Langmead et al., 2009) y ShortStack (Johnson et al., 2016), para optimizar la estrategia de mapeo. Bowtie es uno de los alineadores más popularmente usados y que aún tiene soporte de lecturas *colorspace*. Cuenta con la ventaja de haber sido desarrollado para lecturas cortas (50 pb), lo cual es conveniente para nuestros datos de RNA-Seq (50 pb) y especialmente para las huellas ribosomales (~30 pb). Además, ya ha sido sugerido como un buen alineador para el mapeo de huellas ribosomales (Chung et al., 2015). Dado lo

anteriormente dicho, es que consideramos que el algoritmo de alineamiento utilizado por Bowtie se adapta a nuestras necesidades, con la salvedad del tratamiento de los MMAP. Bowtie selecciona por defecto una posición al azar para estas lecturas, pudiendo configurarse para ignorarlas. Por otro lado, ShortStack utiliza Bowtie en los primeros pasos de mapeo al identificar las mejores regiones de alineamiento para cada lectura. Este primer paso implica que, si la lectura no es MMAP, la determinación de Bowtie y ShortStack será la misma. En caso de que la lectura sea MMAP, ShortStack tiene diversas formas de definir el origen de cada lectura. Nosotros, y por sugerencia de Michael J. Axtell (desarrollador de ShortStack) decidimos que el modo de ponderación única era el más adecuado para nuestros datos (ver sección 3.1.5 Estrategia).

Con el fin de comprobar si la estrategia empleada por ShortStack implica una mejora en la precisión de mapeo con respecto a Bowtie, se elaboró un programa (material suplementario) capaz de simular lecturas, con las características de huellas ribosomales, a partir de todos los transcritos de *T. cruzi* CL Brener Esmerlado-Like (versión 30). Los datos simulados se generan en conjunto con una tabla de referencia de conteos que nos permite saber con exactitud de que transcrito proviene cada una de las lecturas (ver sección 3.1.5 Estrategia). El archivo de lecturas simuladas fue mapeado contra el archivo fasta de los transcritos del cual se originaron. Cada archivo de mapeo resultante fue cuantificado utilizando el programa FeatureCount de SubRead (Liao et al., 2013) (ver sección 3.1.5 Estrategia).

Los resultados observados en la Tabla 3.1 y en la Figura 3.1, permiten concluir que el algoritmo que emplea ShortStack define con mayor precisión el origen de las huellas ribosomales. La mejora en la precisión es más evidente cuando se comparan únicamente los conteos asociados a los transcritos codificantes para TS, Figura 3.1 C y D.

Tabla 3.1. Valores del test estadístico Spearman para la correlación entre la cuantificación de niveles de ARN generadas a partir del uso de un alineador contra el conteo de referencia (*gold standard*).

	Spearman – Todos	Spearman – TS
Bowtie	0.978	0.807
ShortStack	0.993	0.901

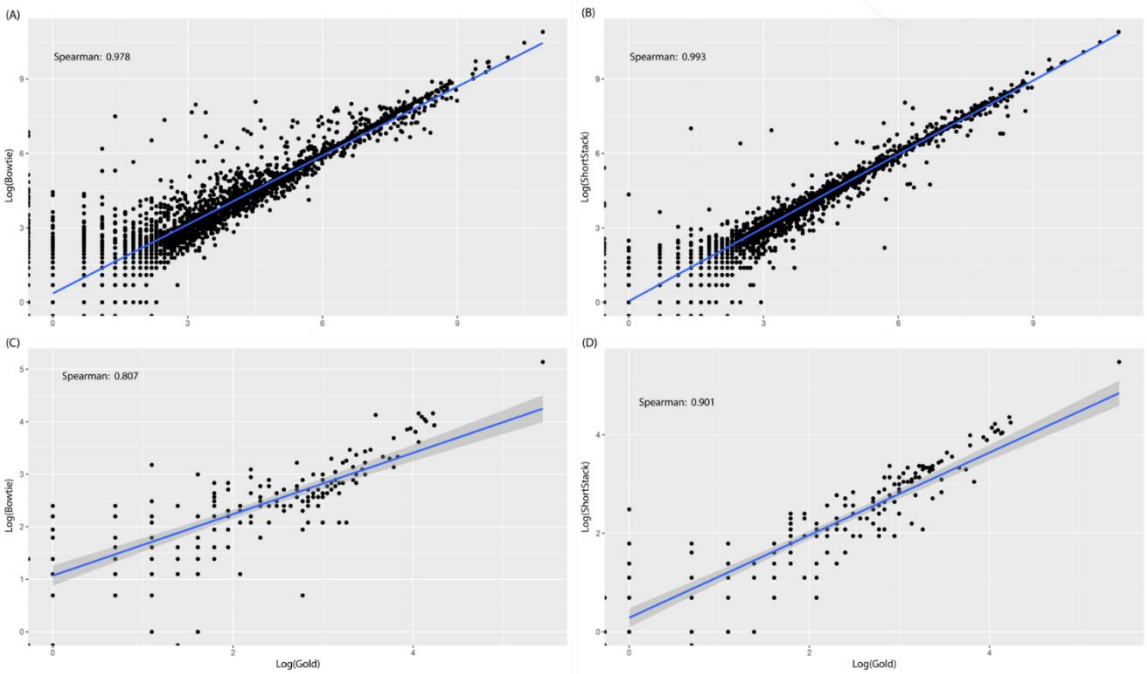


Figura 3.1 Correlación entre la cuantificación de niveles de ARN generadas a partir del uso de un alineador contra una referencia. A) Correlación entre Bowtie y la referencia utilizando todos los transcritos. B) Correlación entre ShortStack y la referencia utilizando todos los transcritos. C) Correlación entre Bowtie y la referencia utilizando solo transcritos codificantes para TS. D) Correlación entre ShortStack y la referencia utilizando solo transcritos codificantes para TS. La correlación de los datos se evaluó mediante el test estadístico de Spearman.

3.1.3. Alineamiento y cuantificación.

Una vez procesadas las lecturas se procedió a alinearlas con ShortStack sobre la referencia genómica. El primer paso para usar este alineador, es indexar el genoma de referencia a utilizar con la herramienta bowtie-build (versión 1.2.2). En este caso, se usó como referencia el haplotipo Esmerlado-Like del genoma de *T. cruzi* CL Brener (versión 30). A continuación, se utilizó el programa FeatureCount para contar cuántas lecturas mapearon sobre cada gen y construir una tabla con los identificadores de cada uno de los genes (ID) y el número de lecturas asignadas al gen. Los resultados obtenidos (Tabla 3.2) están en el orden de lo reportado por Smircich et al. (Smircich et al., 2015).

Tabla 3.2. Resultados de alineamiento y cuantificación de los datos producidos en (Smircich et al., 2015), mediante la utilización de ShortStack y FeatureCounts.

Datos	Estadio	Lecturas Totales	Lecturas mapeadas	Lecturas asignadas a CDS
Huellas	Epimastigota	590017392	31266150	1924763
Huellas	Tripomastigota metacíclico	372656719	12561684	574905
RNA-Seq	Epimastigota	179429362	161952512	11962810
RNA-Seq	Tripomastigota metacíclico	47931449	45729912	944710

3.1.4. Determinación de genes con eficiencia traduccional diferencial

El número de huellas ribosomales asociados a un gen o transcripto está influido por la actividad de traducción específica del gen, así como por la abundancia del ARNm. Sin tener en cuenta los niveles de ARNm de fondo, no se pueden distinguir las diferencias asociadas con la traducción de las que surgen a través de la regulación del estado estacionario del ARN. Se ha definido entonces a la eficiencia traduccional (TE) como la relación entre la tasa de traducción (derivada de los recuentos de huellas por ARNm) sobre el nivel de estadio estacionario del mismo (derivada de los niveles de ARNm medidos por ARN-Seq) (Ingolia et al., 2009). El problema clave del análisis de eficiencia traduccional diferencial se convierte entonces en la identificación de genes cuya diferencia en la abundancia de huellas ribosomales no puede explicarse por las diferencias en la abundancia de ARNm de fondo (Li et al., 2017). En los últimos años ha habido una explosión de algoritmos capaces de analizar experimentos de *Ribosome Profiling* (Eastman et al., 2018) que tienen en cuenta las particularidades de la técnica y en especial el cálculo de la TE diferencial. Con el fin de reducir el número de falsos positivos decidimos emplear distintas herramientas para quedarnos con los genes detectados en común. Si bien esta estrategia es conservadora, nos permite definir un claro panorama de cuáles son las proteínas y funciones con una señal más fuerte en cuanto cambio en la eficiencia traduccional. Para esto utilizamos RiboRex (Li et al., 2017) y RiboDiff (Zhong et al., 2017). RiboRex integra los conteos de los datos de RNA y Ribo-Seq en un único modelo generalizado lineal, al cual le aplica los métodos de análisis clásicos de RNA-Seq, con las herramientas edgeR (McCarthy et al., 2009), DESeq2 (Love et al., 2014) y Voom (Law et al., 2014). Por otra parte, RiboDiff también utiliza modelos

lineales generalizados para estimar la sobre-dispersión de datos de RNA-Seq y Ribo-Seq por separado y luego realiza test estadísticos para determinar la eficiencia traduccional haciendo uso de la abundancia de ARNm y la ocupancia ribosomal.

Smircich et al. ha reportado que un gran subconjunto de genes que cambian su TE en la diferenciación de epimastigota a tripomastigota metacíclico. En particular destacaron la regulación positiva de los genes pertenecientes a la superfamilia TS y la regulación negativa de las proteínas ribosomales. Para comprobar si mediante la nueva estrategia de alineamiento implementada y las nuevas herramientas de análisis de TE se siguen observando las mismas familias y para determinar si se detectan nuevos grupos de genes regulados, procedimos a realizar el análisis de eficiencia traduccional diferencial.

Los resultados observados en la Figura 3.2, muestran los genes que regulan su eficiencia traduccional durante la metaciclogénesis. Para considerar un gen como regulado establecimos un aumento (o disminución) de su valor de TE en dos veces (*Fold Change* de 2), y un valor de FDR (*False Discovery Rate*) < 0.01. Los filtros aplicados son bastante estrictos, pero nos asegura limitar aún más el número de falsos positivos.

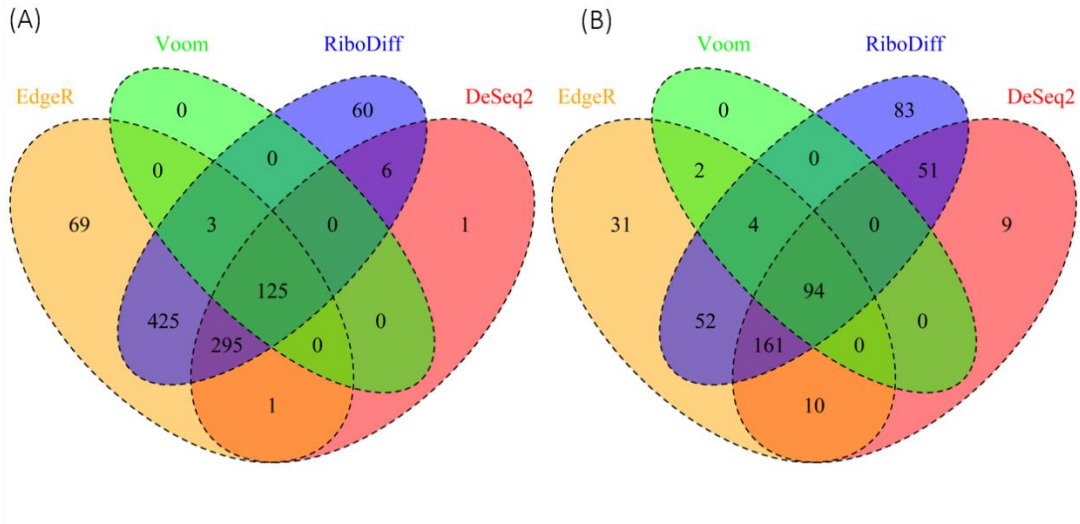


Figura 3.2. Diagrama de Venn mostrando los resultados de eficiencia traduccional diferencial obtenido utilizando las herramientas RiboRex (edgeR, DESeq2 y Voom) y Ribodiff. Las condiciones establecidas para determinar a un gen como regulado fueron FDR < 0.01 y para los sobre-regulados un *Fold Change* >= 2, mientras que para los sub-regulados un *Fold Change* <= -2. A) Diagrama de Venn mostrando los genes que bajan su eficiencia en el pasaje de epimastigota a tripomastigota metacíclico. B) Diagrama de Venn mostrando los genes que suben su eficiencia en el pasaje de epimastigota a tripomastigota metacíclico.

En total se observaron 125 genes que, en todos los análisis realizados, disminuyeron su eficiencia traduccional en el pasaje de epimastigota a tripomastigota metacíclico. En particular, 914 se detectaron en RiboDiff, 918 en edgeR, 428 en DESeq2 y 128 en Voom. Mientras que 94 genes, detectados por todas las herramientas, aumentaron su eficiencia traduccional durante la metaciclogénesis. Se observaron 445 genes utilizando RiboDiff, 354 edgeR, 325 DESeq2 y 100 Voom. La Figura 3.2 nos permite evidenciar que Voom representa el algoritmo más conservador. Los resultados obtenidos mediante los cuatro métodos mencionados se pueden encontrar en el material suplementario.

Los genes sobre-expresados (Tabla 6.1), están fuertemente asociados a procesos traduccionales, habiendo más de un 30% de proteínas ribosomales. Los resultados observados son equivalentes a los reportados en (Smircich et al., 2015). Por otro lado, los 94 genes que presentaron un aumento de eficiencia traduccional en el cambio de estadio, fueron en su gran mayoría proteínas TS (> 70%) (Tabla 6.2). Este resultado confirma también lo observado en (Smircich et al., 2015).

En resumen, los resultados obtenidos aquí nos permiten validar la estrategia de optimización diseñada y, gracias a la disponibilidad de nuevas herramientas de análisis, asignar valores estadísticos a las diferencias de eficiencia traduccional observada entre ambos estadios.

Para profundizar en el análisis de genes regulados traduccionalmente tomamos los valores obtenidos por la herramienta RiboDiff ya que además de presentar resultados de TE de cada estadio independiente, ha sido utilizado ampliamente en la literatura (Hassan et al., 2017; Kiss et al., 2017; Tuorto et al., 2018). Como hemos mencionado, obtuvimos 914 genes sobre-expresados en epimastigota y 445 sub-expresados, como se puede observar en la Figura 3.3.

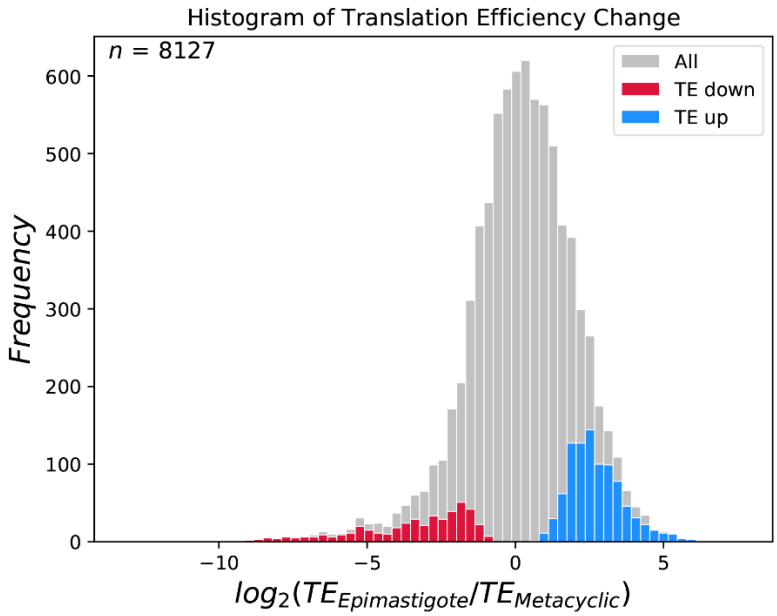


Figura 3.3. Histograma del cambio en la eficiencia traduccional entre los estadios epimastigota y tripomastigota metacíclico. En rojo se marcan los genes que presentan mayor TE en el estadio epimastigota, mientras que en azul se marcan aquellos genes que presentaron menor eficiencia en este estadio. Para considerar que un gen presenta cambios estadísticamente significativos establecimos un cambio de FC absoluto mayor a 2 asociado a un valor de FDR menor a 0.01.

Del total de genes que se encuentran sobre-expresados en el estadio epimastigota, más de un 40% de ellos no presentan anotación funcional (anotados como proteínas hipotéticas) en TriTrypDB. Mientras que de los genes sub-expresados más de un 25 % son consideradas proteínas hipotéticas. Dada esta situación, decidimos desarrollar una herramienta capaz de realizar una anotación profunda de las proteínas de los kinetoplástidos presentes en el TriTrypDB. Se incluyeron todas las proteínas salvo aquellas que estaban anotados como pseudogenes o representen fragmentos. El programa desarrollado lo llamamos DARK y será introducido en el siguiente capítulo.

Está nueva estrategia de anotación nos permitió, a su vez, profundizar en la anotación de ontología génica (GO) (ver próximo capítulo). Para evaluar el impacto de la mejora en la anotación se comparó las categorías génicas disponibles en TriTrypDB y las asignadas por DARK, mediante el programa WEGO (Ye et al., 2018). En total se compararon 7744 genes (el conjunto de estos genes se le llama *background*). Con respecto al *background*, DARK logró anotar un ~75% de los genes mientras que TriTrypDB un ~35%, lo cual representa una mejora de más del doble en el número de genes anotados. Si comparamos las

categorías de ontología génica de forma individual también encontramos notorias diferencias. En cuanto a la categoría componentes celulares (CC), TriTrypDB presenta un ~10% de los genes con esta categoría de anotación, mientras que DARK encuentra un 70%. En la categoría Función Molecular (MF) la relación fue 26-63% y en proceso biológico (BP) fue de 18-65%. Esta diferencia queda representada en la Tabla 3.3 y

Figura 3.4.

Tabla 3.3. Comparación de anotación de términos de ontología génica entre TriTrypDB y DARK. Se comparan los genes anotados totales en las categorías componentes celulares (CC), función molecular (MF) y procesos biológicos (BP).

Herramienta	Condición	# Genes	Genes anotados	CC	MF	BP
TriTrypDB	*Background*	7744	2794	736	2028	1388
Dark	*Background*	7744	5740	5459	4872	4972
TriTrypDB	*Up*	774	383	125	585	177
Dark	*Up*	774	580	538	551	537
TriTrypDB	*Down*	302	101	25	69	61
Dark	*Down*	302	244	234	221	220

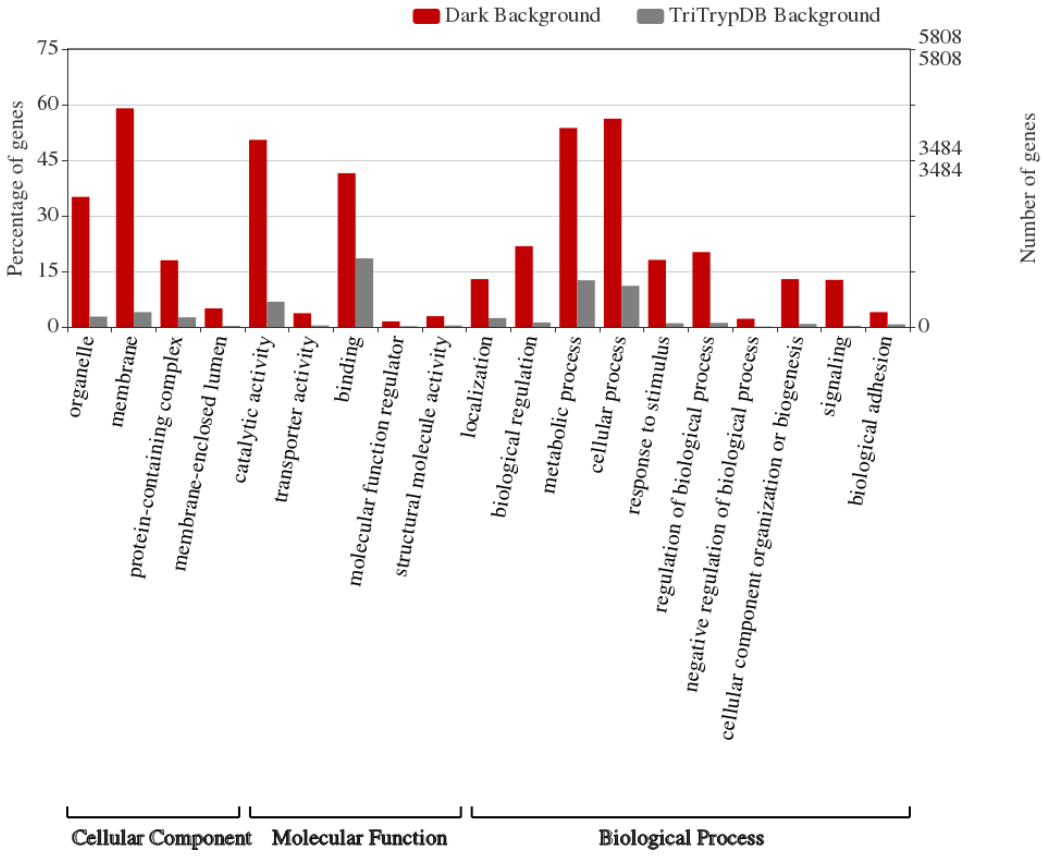


Figura 3.4. Términos de ontología génica que presentan diferencias significativas (p-valor < 0.01) entre DARK y TriTrypDB, para las categorías componentes celulares, función molecular y procesos biológicos.

Con esta nueva anotación, analizamos las categorías génicas de los genes que presentan diferencias en cuanto a la regulación de eficiencia traduccional entre los estadios analizados. De los 914 genes sobre-expresados en el estadio epimastigota, se analizaron 774 ya que el resto fue eliminado por no cumplir requerimientos de DARK. Análogamente se analizaron 302 genes sub-expresados. Al comparar los resultados obtenidos por las dos herramientas (Figura 3.5 A y B) podemos observar categorías comunes como traducción, unión a ácidos nucleicos, biosíntesis de proteínas (macromoléculas) y procesos de modificación del ARN. Sin embargo, podemos ver como DARK permite una mayor comprensión de los resultados obtenidos, contando con una mayor cantidad de términos ontológicos diferenciales (Figura 3.5 A).

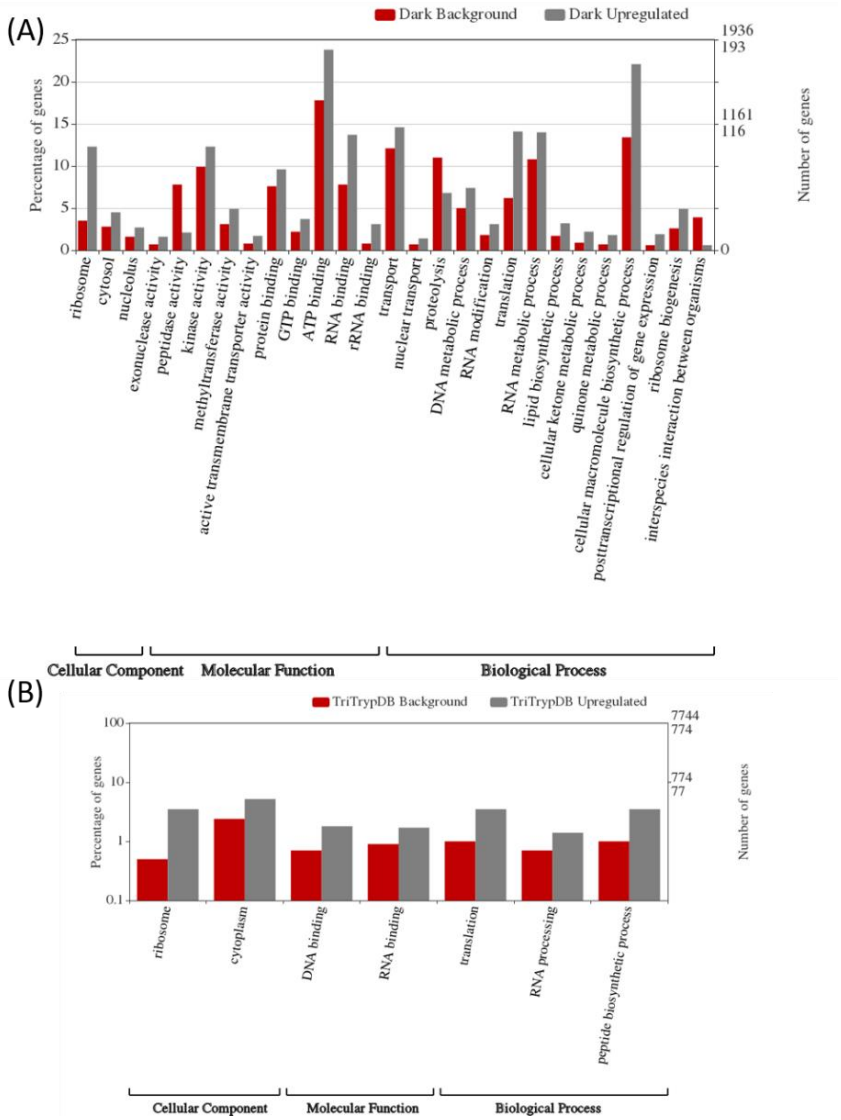


Figura 3.5. Estudio de sobrerrepresentación de categorías génicas realizadas utilizando DARK (A) y TriTrypDB (B), para los genes que disminuyeron su eficiencia traduccional en el pasaje de epimastigota a tripomastigota metacíclico.

Dentro de los nuevos términos surgidos en el análisis de DARK, en la categoría CC, podemos destacar el nucléolo, el cual ya ha sido visto como enriquecido en (Santos et al., 2018) en la fase de crecimiento exponencial de epimastigota en comparación con la fase estacionaria, donde el porcentaje de tripomastigotas metacíclicos es mayor. Además, el enriquecimiento observado en la actividad citosólica ya ha sido reportada en experimentos de proteómica (de Godoy et al., 2012). Dentro de la categoría MF podemos destacar la aparición de categorías específicas como la presencia de transportadores transmembrana cuya sobreexpresión en epimastigota ya ha sido reportada en (Li et al., 2016) y juegan un rol importante en la adquisición de nutrientes (por ejemplo, transportadores de azúcares de hexosa, nucleósidos y aminoácidos). También podemos observar un enriquecimiento en el término metilación, el cual ya se ha observado en (Santos et al., 2018). Finalmente, en cuanto a los términos enriquecidos en procesos biológicos, se encuentra la biosíntesis lipídica, la cual ha sido reportado en (Berna et al., 2017), también y en concordancia con el mismo trabajo, podemos ver que la actividad proteolítica en epimastigota está subrepresentada (Berna describe la sobrerrepresentación de este proceso en el estadio tripomastigota sanguíneo) y como era esperado, una subrepresentación de genes que participan en la interacción interespecie (proteínas TS principalmente).

DARK también realiza una búsqueda de la bibliografía asociada a los genes de la lista analizada (ver próximo capítulo) los que nos permitió expandir los resultados de ontología génica a genes individuales. A modo de ejemplo desarrollaremos algunos casos. El gen TcCLB.510601.30, que codifica para la proteína exonucleasa XRN 5'-3' ha sido descrito en *T. brucei* como clave para el procesamiento del ARNr, participando en la biogénesis ribosomal (término sobrerrepresentado en epimastigotas)(Sakyiama et al., 2013; Fiebig et al., 2015). La regulación al alza de esta proteína en el estadio epimastigota, concuerda con la abundante actividad traduccional observada. También, en el mismo sentido, se observa una sobreexpresión del gen TcCLB.506495.10, parálogo de TbMTase37 de *T. brucei*, el cual codifica una proteína que se localiza en el nucléolo y su depleción da como resultado la acumulación de partículas ribosómicas que carecen de srARN 4 y niveles reducidos de ribosomas asociados a polisomas (Fleming et al., 2016). También, se observan varios genes asociados a la vía de reparación del ADN (TcCLB.506945.80,

TcCLB.507711.320, TcCLB.506147.180, TcCLB.508837.180, TcCLB.509099.70, TcCLB.511867.110, TcCLB.503955.20, TcCLB.511803.20, TcCLB.508277.150 y TcCLB.506743.180), los cuáles han sido descritos en (Genois et al., 2014). Dado que el estadio epimastigota es replicativo y no infectivo (al contrario que el estadio tripomastigota metacíclico) este resultado era esperable. Sin embargo, el análisis permite individualizar los genes involucrados. Finalmente observamos proteínas de unión a ARN (RBPs) como son TcCLB.506773.130, TcCLB.504089.60 y TcCLB.509167.140 que codifican para Pumilio 7 (PUF7), Alba 1 y RBP42, respectivamente. Las proteínas de la familia pumilio, y en particular PUF7, son necesarias para la maduración del ARNr en el nucléolo. Por otra parte, las proteínas Alba están involucradas en los proceso de traducción (Mani et al., 2011; Clayton, 2013). Finalmente, la proteína RBP42 en *T. brucei* está parcialmente asociado con los polisomas y está involucrado con muchos sitios diana de ARNm relacionados con el metabolismo energético (Clayton, 2013).

Los términos de ontología sobrerrepresentados en los genes que están subexpresados en epimastigota se observan en la Figura 3.6. Los resultados muestran una clara sobrerrepresentación de componentes asociados a membrana celular, y una subrepresentación de componentes asociados al citoesqueleto y mitocondria. Se destaca la actividad sialidasa, asociadas a las ya discutidas proteínas TS (determinan también la sobrerrepresentación de genes involucrados en la patogénesis), y al aumento de actividad catabólica celular. Este último aspecto ya ha sido reportado en (de Godoy et al., 2012; Berna et al., 2017). El término transducción de señales intracelulares, se asocia a las actividades de invasión celular y evasión de la respuesta inmune de los tripomastigotas. Análogamente al estudio de los genes sobrerrepresentados en epimastigotas, se utilizó el DARK para encontrar referencias bibliográficas asociados a los genes sobreexpresados en los parásitos tripomastigotas metacíclicos. A modo de ejemplo, observamos el gen TcCLB.510247.20, codificante para la proteína caseína quinasa 1 (CK1), la cual se reportó como una proteínas esencial de la forma sanguínea de *T. brucei* (Urbaniak, 2009). La importancia de esta proteína también ha sido evaluada en Leishmania (Dan-Goor et al., 2013; Rachidi et al., 2014). La sobreexpresión de las proteínas metaciclinas II y III (TcCLB.506529.600 y TcCLB.509351.6) ha sido reportada en (Yamada-Ogatta et al., 2004; de Godoy et al., 2012). Los transportadores de

folato/pteridina juegan un rol clave en la diferenciación del parásito. En particular, se ha visto que estos compuestos son claves para la adhesión, proceso vital para que ocurra la metaciclogénesis (Santos et al., 2018). Por ende, la sobreexpresión de TcCLB.511575.130, acompaña estos cambios. Interesantemente, encontramos regulada la proteína hipotética TcCLB.511071.190, la cual ha sido previamente reportada como un transportador del grupo hemo. En ausencia de hemo citosólico se activa deteniendo el crecimiento celular e induciendo la diferenciación de epimastigota a tripomastigota metacíclico (da Silva Augusto et al., 2015). Finalmente, se observan proteínas de la familia TcTASV (Trypomastigote Alanine, Valine y Serine), TcCLB.511877.10 y TcCLB.509123.10. La familia TcTASV se conserva entre todos los linajes de *T. cruzi* analizados hasta el momento y no tiene ortólogos en otras especies, incluidos los tripanosomátidos estrechamente relacionados. Si bien el rol se desconoce se cree que están involucradas en el establecimiento de la infección inicial de *T. cruzi*, expresándose principalmente en los estadios tripomastigota (Bernabo et al., 2013; Caeiro et al., 2018).

Por lo tanto, las nuevas estrategias utilizadas para el análisis de los datos de Ribo-Seq de estadios epimastigotas y tripomastigotas metacíclicos, han permitido confirmar los resultados obtenidos previamente, y además mejorar ampliamente el análisis de los mismos revelando nuevas proteínas y procesos diferenciales. Esto ha sido resultado de la mejora en la sensibilidad de la determinación de los cambios en la eficiencia traduccional (mejora en el mapeo y aplicación de RiboDiff) así como en la significativa mejora en la anotación del genoma dada por la herramienta DARK.

La optimización de los mapeos y los conteos asociados serán utilizados en los capítulos 3.4 y 3.5.

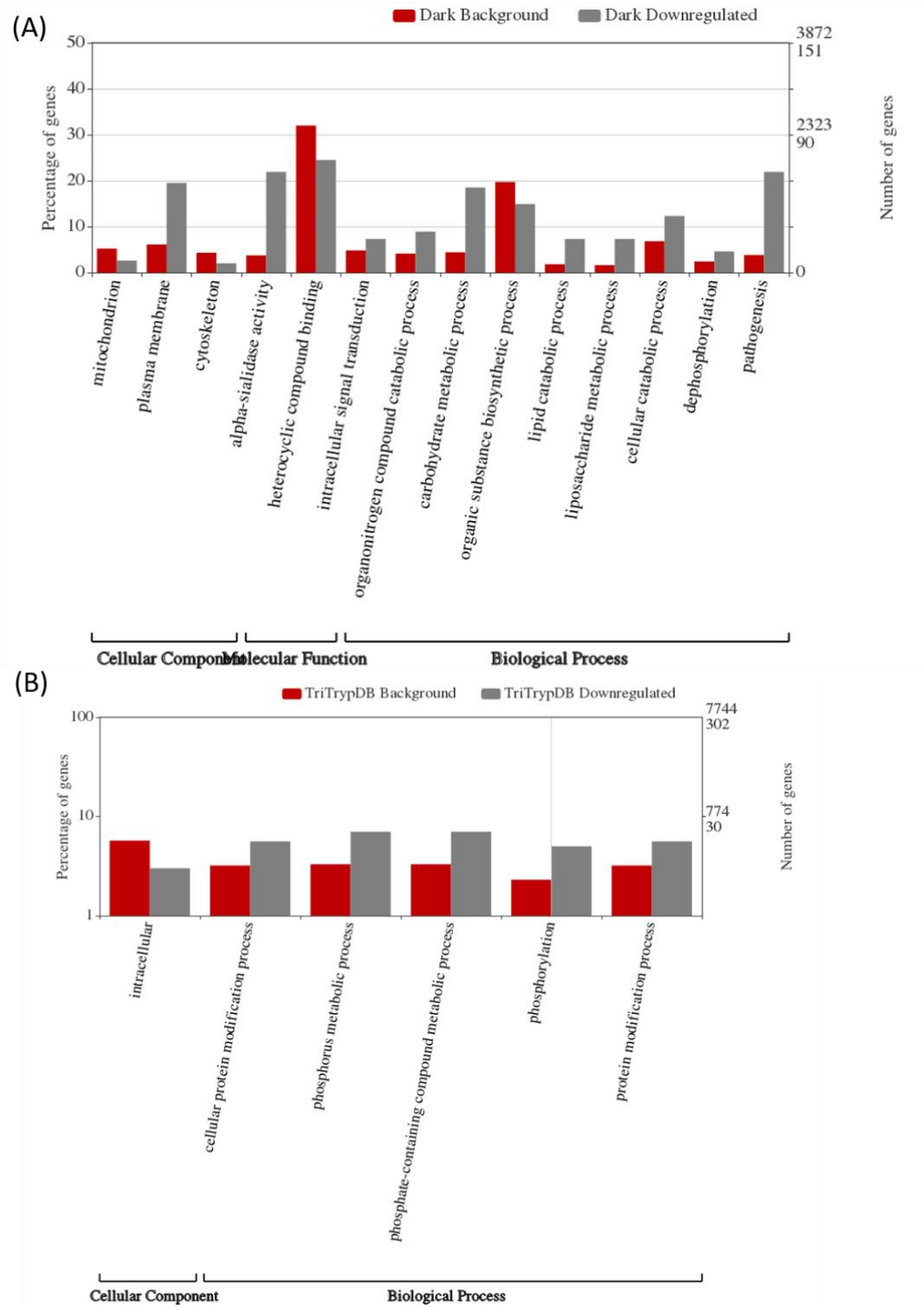


Figura 3.6. Estudio de sobrerrepresentación de categorías génicas realizadas utilizando DARK (A) y TriTrypDB (B), para los genes que aumentaron su eficiencia traduccional en el pasaje de epimastigota a tripomastigota metacíclico.

3.1.5. Estrategia

3.1.5.1. Obtención y pre-procesamiento de los datos de *Ribosome Profiling*

Los datos de *Ribosome Profiling* fueron obtenidos del SRA: PRJNA260933 (Smircich et al., 2015) y corresponden a los estadios epimastigota y tripomastigota metacíclico de *T. cruzi*. En conjunto los datos representan doce muestras. Seis corresponden a datos de ARNm completos (datos RNA-Seq) tres réplicas biológicas por estadio. Los restantes 6 representan muestras provenientes de fragmentos de ARNm presentes en los polisomas (datos Ribo-Seq), también tres réplicas biológicas por estadio.

Se utilizó el programa cutadapt para remover los adaptadores y filtrar por calidad. Se utilizaron los mismos parámetros para los datos RNA-Seq y para los datos Ribo-Seq, con la excepción de la limitación de los largos, en donde en el primero se definió un tamaño mayor a 18 pares de bases (pb), mientras que en el segundo se permitió un rango entre 25 y 40 pares de bases. El resto de los parámetros que se utilizaron consistieron en la especificación del adaptador (5' - CGCCTTGGCCGTACAGCAG - 3'), calidad mínima permitida (13 "*phred score*"), máxima tasa de error permitida (0.1), y el modo *colorspace*.

Se seleccionó el alineador Bowtie (version 1.2.2) para remover contaminación producida por lecturas de origen ARN ribosomal. Los ARNr de *T. cruzi* fueron descargados de la base de datos TriTrypDB.

3.1.5.2. Alineamiento de las lecturas

Para alinear las lecturas previamente obtenidas se utilizaron dos alineadores distintos, Bowtie y ShortStack (version 3.6). Bowtie se ejecutó por defecto, cambiando únicamente dos parámetros (solamente al alinear datos Ribo-Seq; l:20 y --best) recomendados para el mapeo de lecturas cortas (http://bowtie-bio.sourceforge.net/manual.shtml).

Por su parte, el programa ShortStack tuvo que ser adaptado para obtener un correcto funcionamiento con los datos disponibles. Se adaptó el código para que aceptara datos provenientes de la tecnología SOLiD (basados en espacio de color) (Figura 6.1). Además,

se deshabilitó la predicción de estructuras secundarias (--align_only) y de micro ARNs (--nohp). El modo de mapeo elegido fue el modo de ponderación única (U) en donde, únicamente las frecuencias de las asignaciones de lecturas alineadas de forma única en la vecindad de la alineación en cuestión se toman en cuenta en la ponderación final.

3.1.5.3. Cuantificación de transcritos

Se utilizó el módulo FeatureCounts, del paquete SubRead (v1.5.2) para cuantificar el número de lecturas originadas en cada transcrito.

3.1.5.4. Creación de simulador de mapeo

Utilizando el lenguaje de programación Python se elaboró un programa para simular lecturas provenientes de datos de Ribo-Seq y a partir de ellas generar una tabla de conteos de referencia. Brevemente, el programa utiliza un archivo en formato fasta de transcriptos, los cuales toma de referencia para generar lecturas de un largo variable. En nuestro caso el rango elegido toma en cuenta el tamaño de las huellas ribosomales por lo que tomamos largos entre 20 y 35 nucleótidos. Las regiones del transcrito seleccionadas y el número de lecturas por transcrito fueron al azar. Este último dato fue registrado para poder determinar con exactitud cuantas lecturas provienen de cada uno de los transcritos. Generando así un archivo de lecturas simuladas con su correspondiente tabla de conteos de referencia.

3.1.5.5. Comparación de alineadores

Los gráficos que comparan la precisión de mapeo de los alineadores fue realizado mediante la utilización del lenguaje estadístico R (Team, 2013), en conjunto con el IDE RStudio (Team, 2015). La correlación entre los conteos producidos por los alineadores y el conteo de referencia se evaluó mediante el test estadístico Spearman.

3.1.5.6. Análisis de sobrerrepresentación génica.

Los estudios de sobrerrepresentación de ontología génica fueron realizados utilizando la herramienta online WEGO, tomando en cuenta la anotación ontológica realizada por DARK. Para considerar una categoría como sobrerrepresentada se estableció un límite de 0.05 de p-valor.

3.2. Desarrollo de herramientas bioinformáticas que permitan mejorar el análisis de genes diferencialmente expresados

3.2.1. DARK

Actualmente, la clasificación de las proteínas se realizada en función de su estructura, funciones, propiedades fisicoquímicas y, especialmente de su contribución a las vías metabólicas. La secuenciación masiva de genomas microbianos y organismos multicelulares, han allanado el camino en la caracterización rápida de los genes y sus productos proteicos. Sin embargo, aún quedan muchos genes no caracterizados y sus productos son conocidos como proteínas hipotéticas (PH). En general, una PH se define como un marco de lectura que ha sido predicho mediante herramientas computacionales como codificante, pero no existe evidencia experimental de su expresión. Muchas proteínas se muestran como hipotéticas cuando un genoma es secuenciado, se predice el gen pero la anotación automática no le pude encontrar función. Algunas proteínas son análogas a las proteínas que tienen una función desconocida, por lo que están conservadas en varios linajes y son denominadas PH conservadas.

De acuerdo a los datos disponibles en el GeneBank, en el 2019 podemos encontrar 2898031 PH en eucariotas, 833607 en bacteria y 239273 en arqueas, dando idea de la importancia de un método que permita mejorar la anotación de los organismos en general. De las seis subfamilias de Trypanosomatidae, en la base de datos TriTrypDB se encuentran representados organismos de las subfamilias Blechomona, Paratrypanosoma, Leishmaniinae y Trypanosoma. En particular la subfamilia Leishmaniinae presenta parásitos monoexénicos de insectos (géneros Crithidia, Leptomonas) y parásitos diexénicos de insectos y vertebrados (géneros Leishmania y Endotrypanum).

La divergencia temprana de estos organismos determina que una gran cantidad de proteínas no tengan una función específica asignada por búsqueda de homología por secuencia. Por lo tanto, en este trabajo nos planteamos el desafío de obtener información a partir de homología remota. Para esto es aconsejable utilizar tanta información sobre las proteínas en cuestión como sea posible. Los perfiles de secuencia contienen para cada

columna de un alineamiento múltiple las frecuencias de los 20 aminoácidos. Por lo tanto, contienen información detallada sobre la conservación de residuos en cada posición, lo que permite inferir qué tan importante es cada posición para definir a otros miembros de la misma familia de proteínas. Esta es la razón por la que la comparación secuencia-secuencia (o secuencia-perfil) es inferior a la comparación perfil-perfil. En particular, los modelos de Markov ocultos (HMM) son similares a los perfiles de secuencia simples, pero además de las frecuencias de aminoácidos en las columnas del alineamiento múltiple, contienen información sobre la frecuencia de inserciones y eliminaciones en cada columna. Por lo tanto, el uso de perfiles HMM en lugar de perfiles de secuencia simples debería mejorar aún más la sensibilidad.

En la presente sección pretendemos hacer uso de las recientemente desarrolladas metodologías de comparación HMM-HMM para anotar los genomas de kinetoplástidos. Desarrollamos una interfaz gráfica que hemos denominado DARK (*Deep Annotation of Representative Kinetoplastids*) que permite de una forma intuitiva analizar las anotaciones generadas. DARK permite una mejora general en la anotación de las proteínas de los parásitos, incluyendo las que actualmente no cuentan con ningún tipo de anotación (PH) (objetivo específico 2).

3.2.1.1. Obtención de las proteínas presentes en TriTrypDB

El primer paso para mejorar la anotación de las proteínas presentes en la base de datos TriTrypDB, es obtener la secuencia aminoacídica de todas ellas. Dado que esta base de datos proporciona acceso programático a sus búsquedas, a través de los servicios web REST, se creó un programa capaz de descargar toda la información de las proteínas presentes en la base de datos (ver sección 3.2.1.9 Estrategia, disponible en material suplementario). Se obtuvo tanto la información como la secuencia de todas las proteínas presentes en la versión 40 de la base. En total se descargaron 380468 proteínas pertenecientes a 46 organismos (Tabla 3.4) observándose que más de la mitad de las proteínas no presentan anotación funcional.

Tabla 3.4 .Lista de organismos cuyas proteínas fueron obtenidas a través de la versión 40 de TriTrypDB.

SubFamilia	Género	Especie	Cepa
Trypanosomatinae	Trypansoma	Trypanosoma vivax	Y486
Trypanosomatinae	Trypansoma	Trypanosoma theileri	isolate Edinburgh
Trypanosomatinae	Trypansoma	Trypanosoma rangeli	SC58
Trypanosomatinae	Trypansoma	Trypanosoma grayi	ANR4
Trypanosomatinae	Trypansoma	Trypanosoma evansi	strain STIB 805
Trypanosomatinae	Trypansoma	Trypanosoma cruzi	Tula cl2
Trypanosomatinae	Trypansoma	Trypanosoma cruzi	Dm28c
Trypanosomatinae	Trypansoma	Trypanosoma cruzi	strain CL Brener
Trypanosomatinae	Trypansoma	Trypanosoma cruzi	strain Esmeraldo
Trypanosomatinae	Trypansoma	Trypanosoma cruzi	JR cl. 4
Trypanosomatinae	Trypansoma	Trypanosoma cruzi	CL Brener Esmeraldo-like
Trypanosomatinae	Trypansoma	Trypanosoma cruzi	CL Brener Non-Esmeraldo-like
Trypanosomatinae	Trypansoma	Trypanosoma cruzi	Sylvio X10/1
Trypanosomatinae	Trypansoma	Trypanosoma cruzi	Sylvio X10/1-2012
Trypanosomatinae	Trypansoma	Trypanosoma cruzi	cruzi strain Dm28c
Trypanosomatinae	Trypansoma	Trypanosoma cruzi	marinkellei strain B7
Trypanosomatinae	Trypansoma	Trypanosoma congolense	IL3000
Trypanosomatinae	Trypansoma	Trypanosoma brucei	brucei TREU927
Trypanosomatinae	Trypansoma	Trypanosoma brucei	gambiense DAL972
Trypanosomatinae	Trypansoma	Trypanosoma brucei	Lister strain 427
Paratrypanosomatinae	Paratrypanosoma	Paratrypanosoma confusum	CUL13
Leishmaniinae	Leptomonas	Leptomonas seymouri	ATCC 30220
Leishmaniinae	Leptomonas	Leptomonas pyrrhocoris	H10
Leishmaniinae	Leishmania	Leishmania turanica	strain LEM423
Leishmaniinae	Leishmania	Leishmania tropica	L590
Leishmaniinae	Leishmania	Leishmania tarentolae	Parrot-TarII
Leishmaniinae	Leishmania	Leishmania sp.	Leishmania sp. MAR LEM2494
Leishmaniinae	Leishmania	Leishmania panamensis	MHOM/COL/81/L13
Leishmaniinae	Leishmania	Leishmania panamensis	strain MHOM/PA/94/PSC-1
Leishmaniinae	Leishmania	Leishmania mexicana	MHOM/GT/2001/U1103
Leishmaniinae	Leishmania	Leishmania major	strain Friedlin
Leishmaniinae	Leishmania	Leishmania major	strain LV39c5
Leishmaniinae	Leishmania	Leishmania major	strain SD 75.1
Leishmaniinae	Leishmania	Leishmania infantum	JPCM5
Leishmaniinae	Leishmania	Leishmania gerbilli	strain LEM452
Leishmaniinae	Leishmania	Leishmania enriettii	strain LEM3045
Leishmaniinae	Leishmania	Leishmania donovani	BPK282A1
Leishmaniinae	Leishmania	Leishmania donovani	strain BHU 1220

Leishmaniinae	Leishmania	Leishmania braziliensis	MHOM/BR/75/M2903
Leishmaniinae	Leishmania	Leishmania braziliensis	MHOM/BR/75/M2904
Leishmaniinae	Leishmania	Leishmania arabica	strain LEM1108
Leishmaniinae	Leishmania	Leishmania amazonensis	MHOM/BR/71973/M2269
Leishmaniinae	Leishmania	Leishmania aethiopica	L147
Leishmaniinae	Endotrypanum	Endotrypanum monterogeii	strain LV88
Leishmaniinae	Crithidia	Crithidia fasciculata	strain Cf-Cl
Blechomonadinae	Blechomonas	Blechomonas ayalai	B08-376

3.2.1.2. Generación de agrupamientos de proteínas.

Para generar perfiles HMM de las proteínas descargadas de la base de datos es necesario generar agrupamientos (o *clusters*) de secuencias a partir de los cuales realizar alineamientos múltiples que sirvan de base para la construcción del perfil de HMM. Se decidió filtrar aquellas proteínas que estuvieran anotadas como pseudogenes o fragmentos, con el objetivo de introducir la menor cantidad de ruido posible, obteniéndose un total de 371671 proteínas, de las cuales casi la mitad (46 %) no presenta anotación funcional.

Para producir los *clusters* se utilizó el programa MMseqs2 (Steinegger and Soding, 2017). MMseq2 es un *software* que está especialmente desarrollado para realizar agrupamientos de grandes conjuntos de secuencias. La ejecución de MMseqs2 (ver sección 3.2.1.9 Estrategia) produjo 51156 *clusters*, estando un 60% integrado por una única proteína. Un 22% son agrupamientos que contienen entre 2-10 proteínas, mientras que un 17% poseen entre 11-50 y finalmente el 1% de los agrupamientos están compuestos por más de 50 proteínas (Figura 3.7).que un 17% poseen entre 11-50 y finalmente el 1% de los agrupamientos están compuestos por más de 50 proteínas (Figura 3.7).

Figura 3.7. Distribución de tamaños de agrupamientos generados por MMseqs2.

La distribución de tamaños varía de acuerdo al género y la especie analizada. Como modo de ejemplo tomaremos los *clusters* formados por los miembros del género Trypanosoma. Se establecieron 34797 agrupamientos siendo el 58% de proteínas únicas (agrupamientos únicos). El número de *clusters* formados varía de acuerdo a la especie de estudiada. Los genomas de *T. cruzi* son diversos en cuanto a su calidad lo que influye en la determinación de las proteínas reportadas. En total suman 166064 proteínas que formaron 23342 *cluster*, siendo el 49% de ellos agrupamientos únicos. En el caso de *T. brucei* donde se dispone de genomas de mejor calidad se detectaron 7720 agrupamientos, siendo un 8% únicos. Finalmente, para las especies que se cuenta con un único genoma (*T. vivax, T. congolense, T. rangeli, T. theileri, T. evansi y T. grayi*) el rango de agrupamientos únicos varía de 11-24 %. La información del género Trypansoma se presenta en la Figura 3.8.

Interesantemente, la subfamilia Paratrypanosoma, la cual cuenta con un único representante, presenta 7915 agrupamientos, con un 34 % de únicos. Es posible que el gran porcentaje de únicos presentados por estos organismos se deba a su posición distante en el árbol filogenético (Figura 6.2). Se presenta la información de todos los géneros en la Tabla 3.5.

Tabla 3.5. Resumen de los datos analizados por MMseqs2 para los organismos presentes en la base de datos TriTrypDB.

Género	Especie	# Proteínas	# Agrupamientos	% Únicos
Trypansoma	Trypansoma	166062	34797	58
Leishmaniinae	Leishmaniinae	188928	18095	38
Paratrypanosomas	Paratrypanosoma	8027	7915	34
Blechomonas	Blechomona	8653	7315	18
Trypansoma	Brucei	25631	7720	8
Trypansoma	Congolense	10065	6997	19
Trypansoma	Cruzi	83327	23342	49
Trypansoma	Evansi	8869	7508	11
Trypansoma	Grayi	11312	8695	13
Trypansoma	Rangeli	7475	6665	12
Trypansoma	Theileri	10585	9372	24
Trypansoma	Vivax	9798	6792	23
Leishmania	Aethiopica	8722	7793	4
Leishmania	Amazonensis	8133	7085	1
Leishmania	Arabica	8646	7768	3
Leishmania	Brazilensis	8824	7857	4
Leishmania	Donovani	8527	7280	1
Leishmania	Enrietti	8731	7764	9
Leishmania	Gerbilli	8599	7692	2
Leishmania	Infantum	8483	7532	6
Leishmania	Major	17315	8810	10
Leishmania	Mexicana	26095	8406	4
Leishmania	Leishmania sp.	16376	8137	5
Leishmania	Panamensis	8527	7280	1
Leishmania	Tarentolae	8452	7676	11
Leishmania	Tropica	26095	8406	4
Leishmania	Turanica	8608	7747	2
Crithidia	Fasciculata	9055	7382	3
Endotrypanum	Monterogeii	8285	7166	9
Leptomonas	Pyrrhocoris	9873	8038	6
Leptomonas	Seymouri	8485	7702	5
Blechomonas	Ayali	8027	7315	18
Paratrypanosomas	Confusum	8653	7915	34

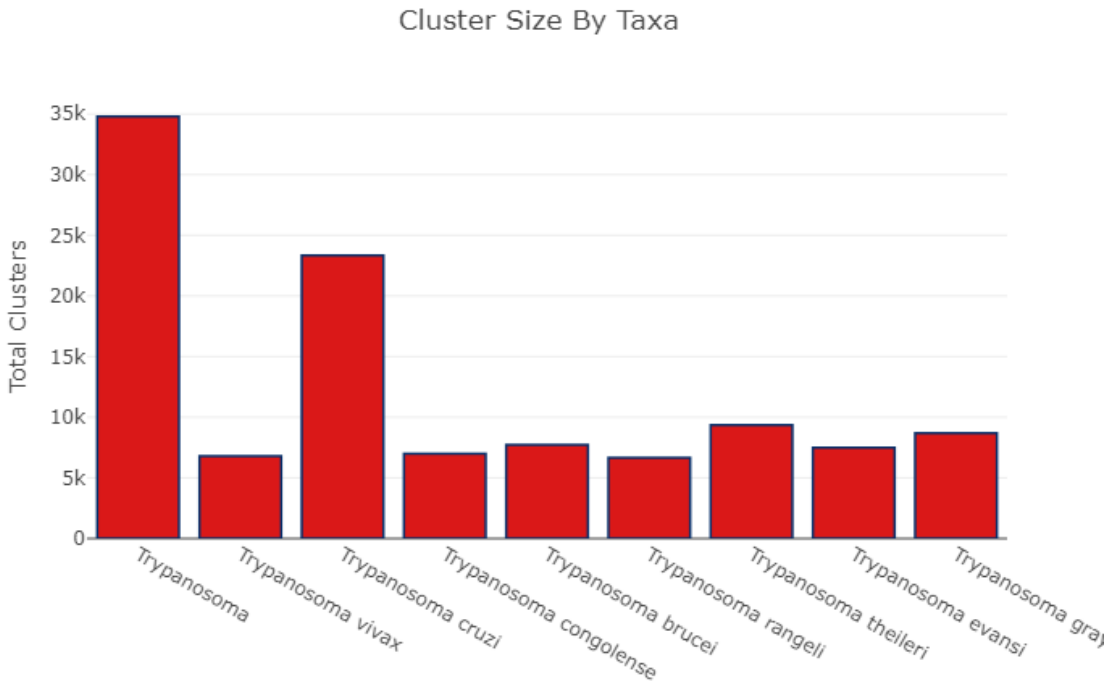


Figura 3.8. Cantidad de agrupamientos generados por los miembros del género Trypanosoma.

Para observar el relacionamiento entre las distintas subfamilias analizamos cuantos agrupamientos presentan en común. Para considerar que un agrupamiento pertenece a una subfamilia particular debe tener cinco o más proteínas en el caso de Trypansoma y de Leishmaniinae y al menos una en el caso de Blechomonas y Paratrypanosomas.

Se observaron 3785 agrupamientos compartidos por todas las subfamilias. Este número nos sirve de referencia para establecer un núcleo común de la familia Trypanosomatidae. A modo de ejemplo, si nos centramos en la subfamilia Trypanosoma, encontramos 3088 agrupamientos únicos; 4850 que comparte únicamente con Leishmaniinae; 549 que comparte con Paratrypanosoma y 437 con Blechomonas. Se presenta en la Tabla 3.6, el resumen de las relaciones observadas.

Tabla 3.6. Agrupamientos compartidos entre las subfamilias.

	Trypanosoma	Leishmaniinae	Blechomonas	Paratrypanosoma	Todos
Trypanosoma	3088	4850	437	549	3785
Leishmaniinae	4850	2078	336	5189	3785
Blechomonas	437	336	2711	1699	3785
Paratrypanosoma	549	5189	1699	1368	3785

3.2.1.3. Anotación de los agrupamientos

Cuanto mayor sea el número de miembros de un agrupamiento, más confiable será el perfil HMM producido. Por esa razón, decidimos utilizar únicamente los agrupamientos que tuvieran al menos 5 miembros; analizándose en total 12034.

Para generar un perfil de HMM es necesario contar con el alineamiento múltiple de secuencia (MSA, por su denominación en inglés) del conjunto de proteínas presentes en cada uno de los agrupamientos. Los MSA fueron producido con el programa T-Coffee (Notredame et al., 2000), los cuales fueron posteriormente transformados en perfiles de HMM, mediante el programa HHblits (Remmert et al., 2011) (ver sección 3.2.1.9 Estrategia). Posteriormente, fueron comparados contra perfiles HMM disponibles en distintas bases de datos: PDB (Berman et al., 2000), PFAM (El-Gebali et al., 2019), SCOP (Andreeva et al., 2014) y UniClust (Mirdita et al., 2017) (versión *clusterizada* de UniProtKB (UniProt, 2019)) (ver sección 3.2.1.9 Estrategia). Cada perfil de HMM contra el cual se obtuvo un *hit* en las bases de datos, se representa mediante un único miembro y tiene asociado un identificador y una función asignada por la propia base. Además de esta información, como resultado de la comparación se obtiene la probabilidad de que la asignación sea correcta, el e-valor y la región de los perfiles donde se produjo el *match*, entre otros datos.

Como máximo, cada archivo resultante, contiene 500 hits. Lo que implica que para cada agrupamiento tenemos como máximo 500 funciones posibles, por lo que es necesario tomar medidas para asignarle la función más informativa posible. En primer lugar, nos quedamos únicamente con los *hits* que posean una probabilidad mayor al 90% y un e-valor menor a 0.005, independientemente de la base de datos del cual provenga el *match*. Luego, utilizando la anotación proveniente de todos los *hits* para el *cluster* en estudio, calculamos la frecuencia de las palabras presentes en la descripción de las funciones (sentencia), ignorando palabras no informativas (ver sección 3.2.1.9 Estrategia) y determinamos cuales están sobrerrepresentadas en el conjunto. Luego se puntean las palabras de acuerdo a su frecuencia, valor que es utilizado para calcular un puntaje asociado a las sentencias. El valor asociado a cada sentencia está influido además por las veces que se repite la misma sentencia y por la probabilidad del *hit* asociado a ella. Las 3 sentencias con mayor puntaje se seleccionan y corresponderán a la anotación reportada

por DARK para el *cluster* en estudio. El algoritmo asegura que las 3 sentencias contengan diferente anotación calculando la disimilitud entre ellas. De esta manera, la segunda sentencia seleccionada será la que le sigue en puntaje a la primera si y solo si las palabras que lo componen son suficientemente distintas. Análogamente, la tercer sentencia seleccionada deberá ser suficientemente distinta a la primera y a la segunda (ver sección 3.2.1.9 Estrategia).

De esta forma se pretende otorgar la mayor información posible a cada agrupamiento. Es importante resaltar que las comparaciones de perfiles se hacen con cada base por separado, por lo que la anotación de los agrupamientos cuenta con un máximo de tres sentencias para cada base de datos (PDB, SCOP, PFAM y UniClust).

De los 12034 agrupamientos al 77 % se le asignó una función derivada de por lo menos una de las bases de datos analizadas. A 8753 agrupamientos (73 %) se le asignó función derivada de UniClust; al 51% de PDB, 54 % de SCOP y 43 % de PFAM.

Para determinar si la estrategia desarrollada es capaz de mejorar la anotación de las PH, clasificamos a los agrupamientos generados por el porcentaje de miembros cuya su función es desconocida (según TriTrypDB).

Analizamos la cantidad de agrupamientos anotados en:

1. Agrupamientos con más del 50% de PH
2. Agrupamientos con más del 80% de PH
3. Agrupamientos con el 100% de PH

Del primer grupo (7098 agrupamientos) se logró anotar el 61 %. De los 6157 agrupamientos integrados por más de un 80% de PH, se pudieron anotar un 56 %. Finalmente, de los 5098 agrupamientos donde la totalidad de las proteínas son PH, se logró anotar un 51 % (Figura 3.9). Los resultados de las anotaciones derivadas de cada una de las bases de datos para cada categoría establecida se resumen en la Tabla 3.7.

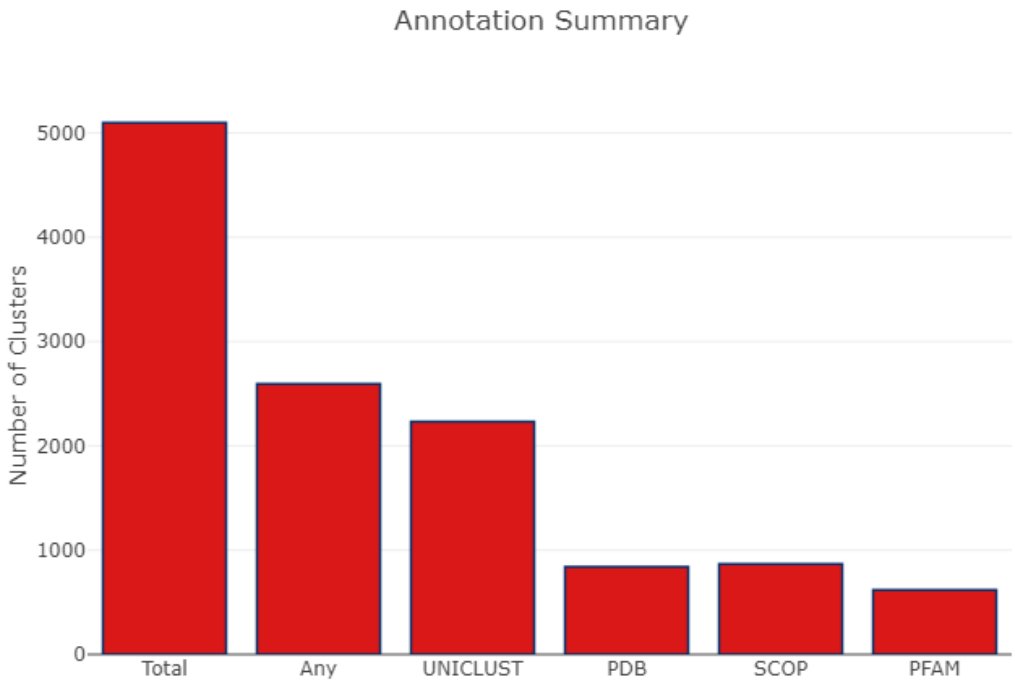


Figura 3.9 . Se representa la cantidad de agrupamientos donde la totalidad de los miembros son PH. Se puede observar la cantidad de ellos que está anotada con cada base de datos.

Tabla 3.7. Resumen de anotación de los agrupamientos producidos.

	# Agrupamientos	# Anotados	# UniClust30	# PDB	# SCOP	# PFAM
0% PH	12034	9289	8753	6157	6510	5171
50 % PH	7098	4319	3832	1968	1987	1485
80 % PH	6157	3454	3020	1342	1362	988
100 % PH	5098	2594	2233	839	867	620

Además de la información previamente mencionada, DARK genera para cada agrupamiento un reporte que incluye información sobre ontología génica, literatura científica asociada, resumen de la información derivada del TriTrypDB, etc. A continuación, pasaremos a detallar alguna de ellas.

3.2.1.4. Asignación de ontología génica

Ya hemos demostrado en el capítulo anterior, que uno de los aspectos relevantes de la presente estrategia de anotación es que nos permite obtener una mayor comprensión de los procesos biológicos que ocurren en una condición particular al dotar, con mayor profundidad, de términos GO a los genes que son regulados específicamente en esa

condición. Para poder realizar la asignación de términos de GO, utilizamos la comparación de perfiles realizada contra UniClust30. La ventaja de esta base es que está basada en UniProtKB cuyos miembros poseen anotación ontológica producida por el *Gene Ontology Project*.

A partir de los *hits* que pasaron los filtros de probabilidad y e-valor previamente mencionados, se obtuvieron los identificadores de UniProtKB de los cuales se consiguieron los términos GO para las categorías componentes celulares, funciones moleculares y procesos biológicos. Reportando, para cada categoría los términos GO más representativos (ver sección 3.2.1.9 Estrategia).La comparación de esta anotación con respecto al TriTrypDB y la aplicación práctica de estos resultados ya fue demostrada en el capítulo anterior. En la Figura 3.10, se muestra un ejemplo de asignación de ontología génica realizado por DARK.

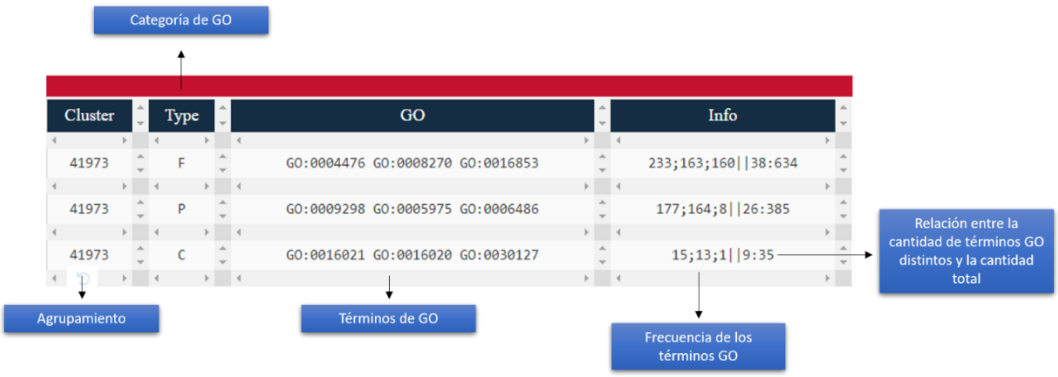


Figura 3.10. Reporte de las categorías génicas asignadas a un agrupamiento (41973) tal como se visualiza en la interfaz gráfica del DARK.

3.2.1.5. Asignación bibliográfica

Previamente utilizamos la capacidad de DARK para minar la literatura e identificar artículos asociados a genes que regulan su eficiencia traduccional durante la metaciclogénesis. No todos los artículos resultantes están relacionados directamente con los genes evaluados, sino que pueden estar vinculados a miembros de la misma agrupación. Por esta razón, la búsqueda bibliográfica realizada por DARK permite visualizar un panorama más completo del rol de un gen en la biología de los kinetoplástidos.

La minería de la literatura se realiza a través del acceso programático del European PMC (Europe, 2015), el cual permite buscar palabras claves en el texto completo (incluido pie

de figura y tablas) de artículos de libre acceso. Las palabras usadas como claves son los identificadores de las proteínas (TriTrypDB ID) miembros de cada agrupamiento. Es importante aclarar que, por problemas de accesibilidad, solo se trabaja con artículos de libre acceso y con los identificadores actuales del TriTrypDB. La actualización de la literatura científica se planea realizar de forma trimestral, siendo la última de febrero del 2019.

Como ejemplo de uso, tomamos el *cluster* 31931. Este agrupamiento está conformado por 39 proteínas (38 hipotéticas) Una de ellas (*T. brucei*) está anotada como "FGR1 oncogene partner-like protein (FOP)" (con la cual DARK coincide). DARK encontró un artículo asociado a este conjunto de proteínas donde se describe el rol de las proteínas FOP en *T. brucei* (Harmer et al., 2018), Figura 3.11.

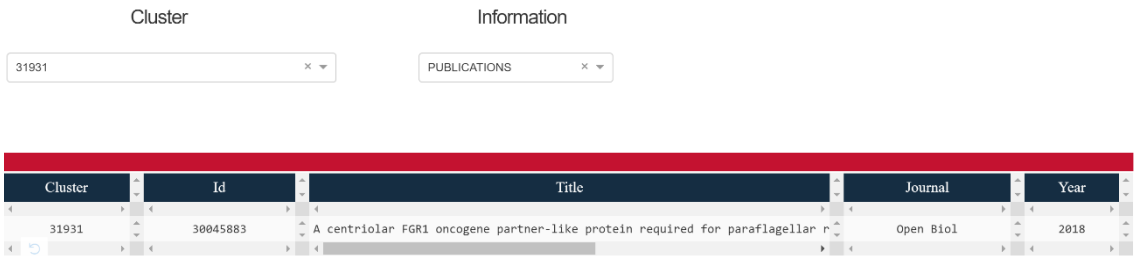
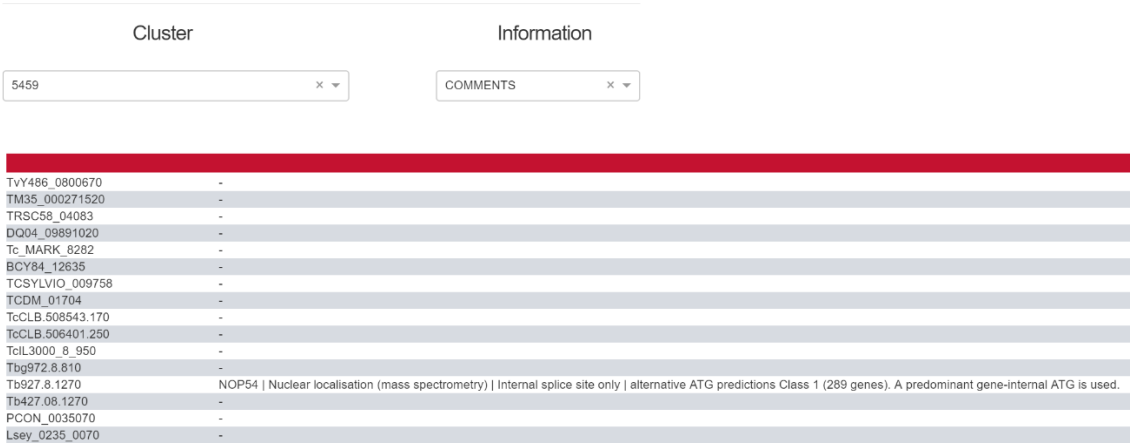


Figura 3.11. Resultado de la minería de la literatura para el agrupamiento 31931. Se presenta el nombre del agrupamiento y el identificador, título, revista y año de publicación de los artículos encontrados.

3.2.1.6. Asignación de información derivada del TriTrypDB

DARK reporta, para cada agrupamiento, información individual de las proteínas obtenida del TriTrypDB. Por ejemplo, la función de la proteína, comentarios de usuarios presentes en la base, cantidad de dominios transmembrana, la anotación funcional en las diferentes bases de datos, etc. Toda esta información centralizada a través del agrupamiento permite rápidamente tener una visión global de la información de los miembros presentes en el TriTrypDB. Como ejemplo de uso visualizamos los comentarios del *cluster* 5459 que representa una proteína conservada en la familia Trypanosomatida. El único comentario que aparece nos dice que la proteína en cuestión tiene localización nuclear, al menos en *T. brucei*.



TvY486_0800670	-
TM35_000271520	-
TRSC58_04083	-
DQ04_09891020	-
Tc_MARK_8282	-
BCY84_12635	-
TCSYLVIO_009758	-
TCDM_01704	-
TcCLB.508543.170	-
TcCLB.506401.250	-
TcIL3000_8_950	-
Tbg972.8.810	-
Tb927.8.1270	NOP54 | Nuclear localisation (mass spectrometry) | Internal splice site only | alternative ATG predictions Class 1 (289 genes). A predominant gene-internal ATG is used.
Tb427.08.1270	-
PCON_0035070	-
Lsey_0235_0070	-

Figura 3.12. Comentarios de usuarios presentes en la base de datos TriTrypDB asociados a las proteínas miembros del *cluster* 5459. El botón desplegable *cluster*, en la zona superior izquierda de la figura, permite seleccionar el agrupamiento a analizar; mientras que el botón a su derecha *information*, permite seleccionar distintos campos de los cuales se quiere obtener información. Por ejemplo, las anotaciones asignadas por TriTrypDB, a cada proteína miembro del agrupamiento analizado, utilizando diferentes bases de datos: Prosite (Sigrist et al., 2010), Interpro (Mitchell et al., 2019), Smart (Letunic and Bork, 2018), etc.

3.2.1.7. Visualización de la información producida por DARK

La página de inicio de DARK muestra información general de los *cluster*. ¿Cuántos son?, ¿Cómo es su distribución en tamaño?, ¿Cuántos fueron anotados?, ¿Qué agrupamientos tienen en común las distintas subfamilias?, etc. La información se presenta mediante gráficas interactivas y, a modo de ejemplo, en la Figura 3.13 podemos ver cómo es posible explorar los agrupamientos compartidos entre las distintas subfamilias, centrándose en cualquiera de ellas al seleccionar una subfamilia en el botón desplegable. La información de esta gráfica ya fue previamente discutida y se puede visualizar en la Tabla 3.6. El resto de las gráficas presentadas en esta sección también forman parte del reporte que presenta DARK en su página de inicio.

Figura 3.13. Gráfica interactiva presente en la página de inicio del DARK que analiza los agrupamientos compartidos entre las distintas subfamilias de tripanosomátidos.

Además, desde aquí se puede navegar a otras secciones del DARK que iremos explorando a continuación:

3.2.1.7.1. Genes

La presente sección permite visualizar la información asociada a los productos génicos presentes en TriTrypDB y conocer cuál fue el agrupamiento asignado a cada gen. La parte superior de esta sección contiene una tabla filtrable donde se reporta información básica de todas las proteínas: identificador, organismo, largo en aminoácidos, descripción y el agrupamiento al cual fue asignada. Un ejemplo de esta tabla se puede observar en la Figura 3.14. La región inferior presenta un reporte individualizado de los genes, los cuales son seleccionados a través de botones desplegables, expandiendo la información provista, como se evidencia en la Figura 3.15.

ID	ORGANISM	LENGTH	PRODUCT	CLUSTER
filter data...				
TY486_1009410	T. vivax Y486	1079	leucine-rich repeat protein (LRRP), putative	2665
TvY486_0000010	T. vivax Y486	447	hypothetical protein, conserved in T. vivax	16174
TvY486_0000030	T. vivax Y486	193	casein kinase, putative	19577
TvY486_0000050	T. vivax Y486	367	hypothetical protein, conserved in T. vivax	23041
TvY486_0000080	T. vivax Y486	338	hypothetical protein, conserved in T.vivax	45621
TvY486_0000090	T. vivax Y486	567	hypothetical protein, conserved in T. vivax	427
TvY486_0000120	T. vivax Y486	417	hypothetical protein, conserved	33163
TvY486_0000130	T. vivax Y486	217	hypothetical protein, conserved in T. vivax	30385
TvY486_0000140	T. vivax Y486	507	amino acid transporter, putative	9716
TvY486_0000150	T. vivax Y486	1230	flagellar calcium-binding-like protein, putative	16833

Figura 3.14. Tabla filtrable que permite visualizar las proteínas (ID) presentes en el TriTrypDB y determinar el *cluster* asignado a cada gen. Además, brinda información del organismo al cual pertenece cada proteína, su largo en aminoácidos y la función predicha.

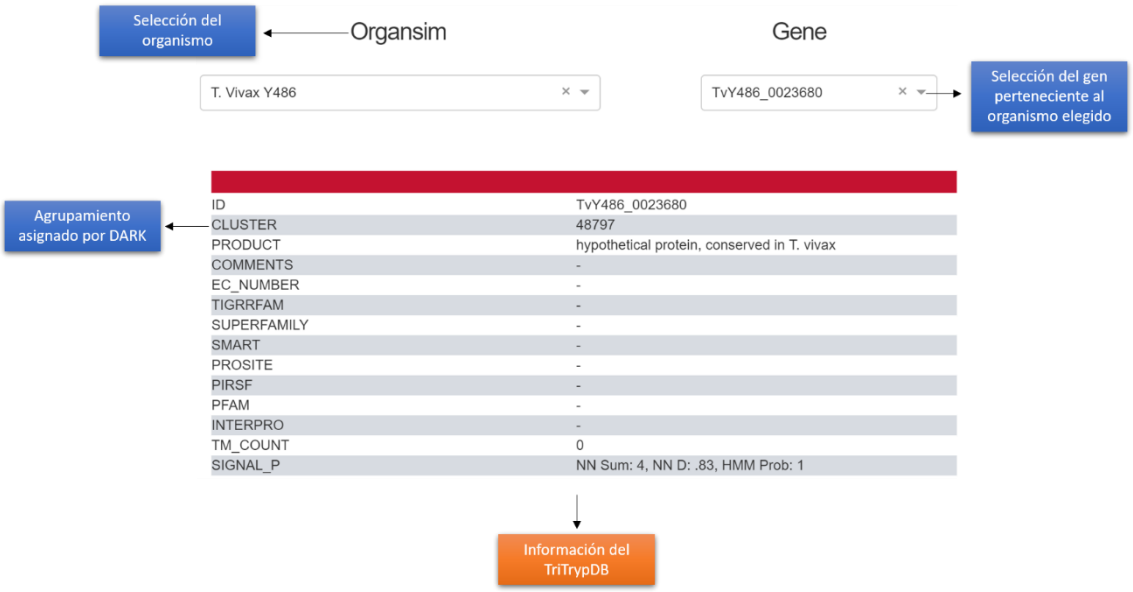


Figura 3.15. Visualización de la información del gen TvY486_0023680 de *T. vivax* Y486 producida por DARK.

3.2.1.7.2. Cluster

La sección *cluster* permite visualizar la información asociada a cada uno de los agrupamientos producidos por DARK. Análogamente a la sección anterior el informe se presenta en dos regiones. La superior es también una tabla filtrable donde se puede visualizar información general de los agrupamientos. ¿Qué tamaño tiene?, ¿Cuántos miembros de cada subfamilia lo componen? Y si fue o no anotada por DARK.

CLUSTER	SIZE	TRYPANOSOMA	LEISHMANIINAE	BLECHOMONAS	PARATRYPANOSOMA	ANNOTATED
filter data...	> 4					
50781	35	11	22	1	1	NO
9365	8	8	0	0	0	NO
43989	36	12	22	1	1	NOT COMPUTED
25996	19	1	17	0	1	YES
35783	23	0	22	1	0	YES
18913	21	0	21	0	0	NO
34329	12	12	0	0	0	YES
6799	34	11	22	1	0	YES
28311	39	15	22	1	1	YES
12317	8	8	0	0	0	YES

Figura 3.16. Información general de los *cluster* producidos por DARK. Se puede observar el número total (*size*) y por subfamilia de proteínas que integran cada *cluster* y si fue o no anotado.

La inferior, presenta información individual de los agrupamientos a través de botones desplegables que confieren la posibilidad de elegir el *cluster* y la información a desplegar asociada a este. Por ejemplo, si se selecciona ORGANISM se despliega una tabla que indica de que organismos provienen las proteínas que componen un agrupamiento. También se puede seleccionar ANNOTATION la cual pone en evidencia las tres sentencias (como máximo) seleccionadas por DARK, en cada base de datos, para asignarle una función a un *cluster*. Además, se pueden investigar las publicaciones asociadas, los términos de ontología génicos asignados, la descripción de los productos génicos y los comentarios de usuarios del TriTrypDB y más información relevante, Figura 3.17.

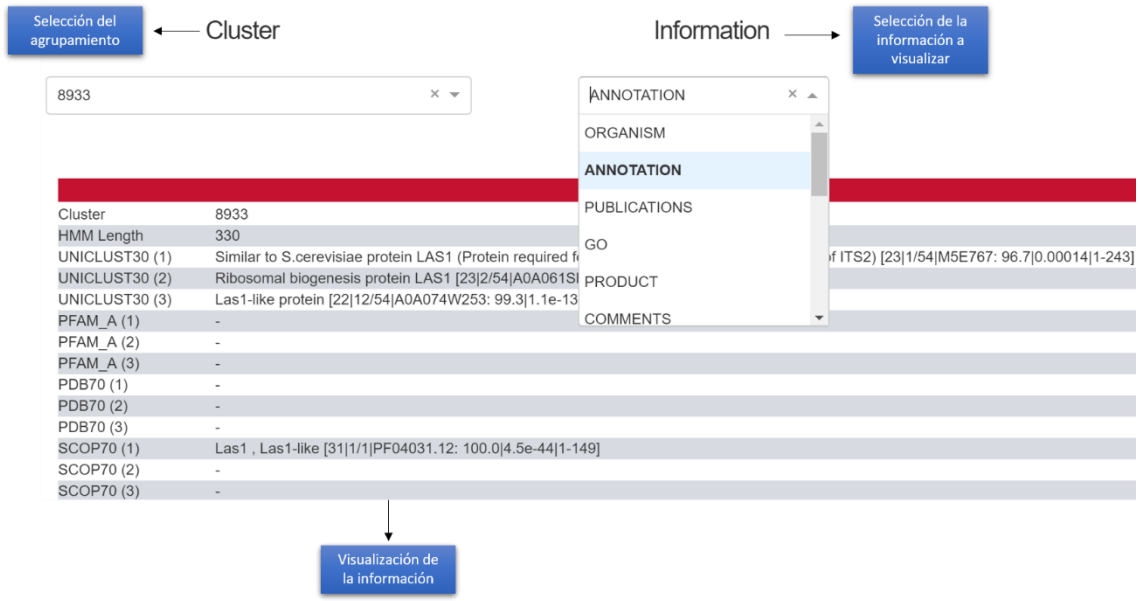


Figura 3.17. Visualización del reporte de cada agrupamiento realizado por DARK. Aquí se representa la anotación del cluster 8933. El resultado de la anotación de este agrupamiento sugiere que las proteínas que lo conforman son LAS1, el cual es un factor que participa en la biogénesis ribosomal. Apoyando está observación, la proteína Tb927.8.5820 integrante de este agrupamiento presenta localización nuclear (según comentario en el TriTrypDB).

En resumen, al tratar a cada gen como parte de un agrupamiento, se obtiene una visión más global de nuestros genes de interés, desde un punto de vista evolutivo, funcional, bibliográfico, etc.

3.2.1.7.3. Functions

La presente sección permite al usuario explorar la anotación de los *cluster* partiendo de palabras claves. La posibilidad de interrogar las funciones de los *cluster* permite explorar

de forma eficiente todos los agrupamientos que se relacionan con uno o varios términos funcionales.

Cada agrupamiento puede presentar un máximo de 12 funciones asociadas (3 por base de dato), las cuales se toman en conjunto para formar una única sentencia que será interrogada en la búsqueda de palabras claves. Hemos desarrollado dos modalidades para realizar la búsqueda: unión e intersección. La modalidad de unión simplemente es la sumatoria de los agrupamientos que poseen alguna de las palabras utilizadas en la búsqueda. Mientras que, la modalidad intersección, permite seleccionar los agrupamientos en donde en la sentencia interrogada coexistan los términos buscados. A medida que se van escribiendo los términos los resultados se van actualizando, mostrando únicamente aquellos agrupamientos que cumplan los criterios de búsqueda. Los resultados se presentan en la región inferior al recuadro de búsqueda y son análogos a los exhibidos en la sección anterior. En la Figura 3.18, se observa el resultado de la búsqueda de las palabras claves TFIIH y helicasa, para la cual se obtuvieron 8 agrupamientos cuyas funciones incluían ambas palabras. Como se puede observar, la mayoría de los agrupamientos tienen representantes en todas las subfamilias lo que indica que son proteínas muy conservadas.

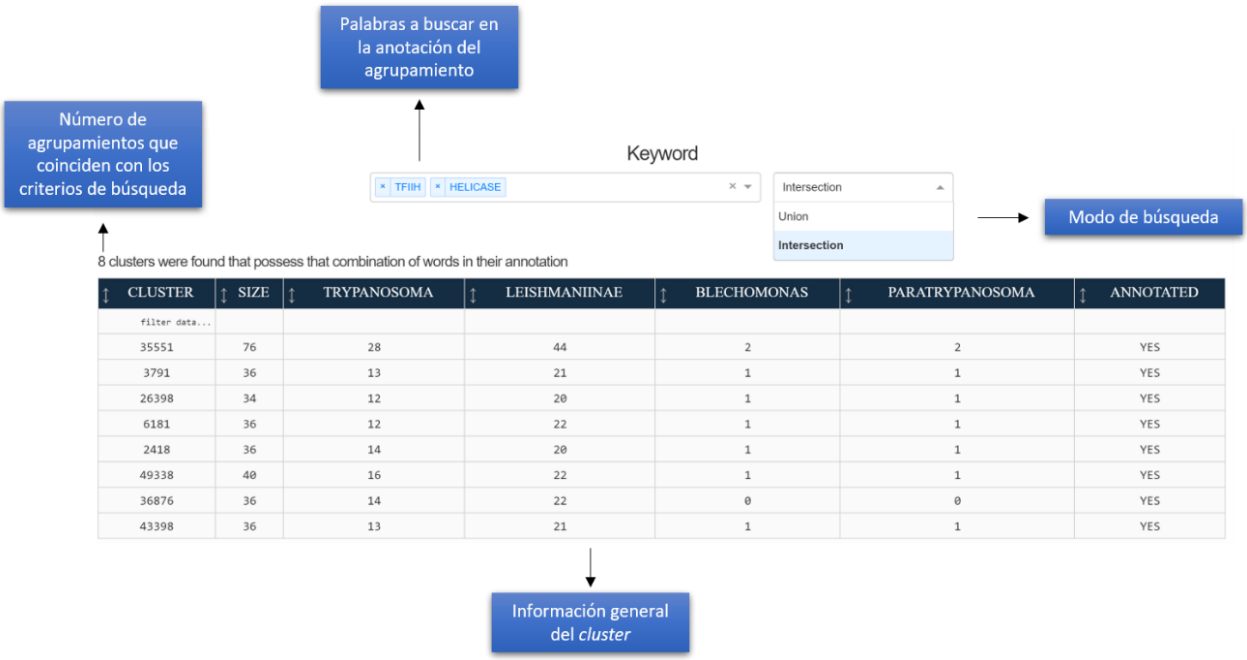


Figura 3.18. Visualización de la búsqueda de palabras claves en las funciones asignadas a los agrupamientos. En el ejemplo se exhibe la búsqueda de las palabras TFIIH y helicasa utilizando el modo intersección. Como resultado se observan 8 agrupamientos, los que pueden ser explorados en la misma forma que la Figura 3.17.

3.2.1.7.4. Gene List

Una de las funcionalidades que consideramos más importantes del DARK es la capacidad de reportar información en formato tabular partiendo de una lista de proteínas de interés. Esta funcionalidad es muy fácil de utilizar, simplemente se debe copiar una lista de identificadores de proteínas (TriTrypDB) y presionar el botón *Compute*. Como resultado se generarán, en la computadora local del usuario, distintos archivos los cuales serán brevemente introducidos, pero que reportan los resultados obtenidos para cada proteína en cuestión.

En total se generan 5 archivos, que se nombran de acuerdo a la información que aportan y la fecha en la cual se realizó el reporte. Los archivos generados son:

1. Genes no analizados. Son aquellos genes que por estar fragmentados o por ser pseudogenes no fueron analizados.
2. Agrupamientos. Brinda información básica de los agrupamientos, tamaño, distribución en subfamilias, funciones asignadas por DARK y la función del gen dada por TriTrypDB.
3. Artículos asociados. Son los artículos que se encontraron vinculados a los agrupamientos de los cuales los genes son miembro.
4. Ontología génica. Información de la ontología génica asignada a cada gen (dado el agrupamiento al cual pertenece).
5. TriTrypDB. Información del TriTrypDB asociada a los genes en cuestión.

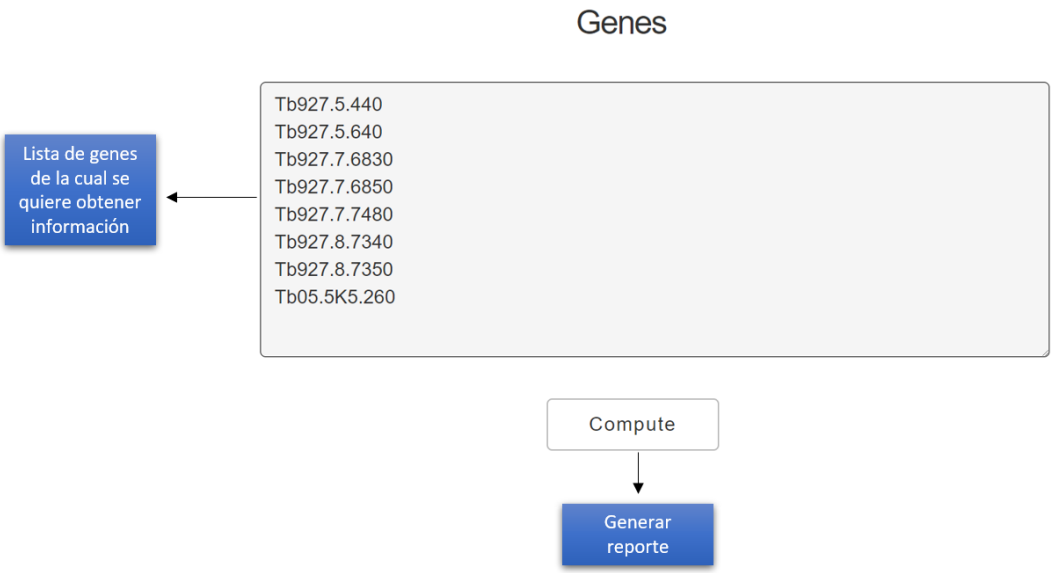


Figura 3.19. Visualización de la sección *Gene List* de DARK. Está sección permite obtener de forma sencilla un reporte de los resultados del programa para los genes de interés.

3.2.1.8. Casos de estudio

Para evaluar la sensibilidad de DARK para detectar homología remota en comparación a métodos que están basados en HMM-secuencia decidimos utilizar el programa BUSCO (Waterhouse et al., 2017). BUSCO fue desarrollado para determinar la calidad de un genoma basado en la detección de un conjunto de proteínas que deberían encontrarse en todos los organismos de un linaje determinado dada su conservación virtualmente universal. BUSCO utiliza la base de datos OrthoDB (Kriventseva et al., 2019) para seleccionar el grupo de proteínas conservadas (> 90 % de las especies analizadas) en un linaje particular, conformado grupos ortólogos de los cuales genera perfiles HMM. En particular, analizamos la presencia de 303 grupos de ortólogos, conservados en linaje de Eucariotas, en 39 genomas del TriTrypDB. Dado que la función de estos grupos de ortólogos no es dada por BUSCO, la misma la asignamos mediante la búsqueda de homología realizada con el programa HHMER (Eddy, 2011) entre los perfiles de HMM de los grupos de BUSCO y las secuencias proteicas de la base Swiss-Prot (Bairoch and Apweiler, 2000).

En total se encontraron 51 grupos ortólogos que llamativamente están ausentes en los 39 genomas estudiados y que se reportan en la Tabla 3.8. Las proteínas representadas en

esta tabla representan casos interesantes para probar la capacidad del DARK en detectar homología remota.

Tabla 3.8. Grupos de ortólogos no encontrados por el programa BUSCO en los genomas del TriTrypDB. Se muestra el identificador del grupo de ortólogo y la descripción de está basada en la búsqueda por homología al Swiss-Prot.

BUSCO ID	Descripción
EOG09370FSS	ALG6: Dolichyl pyrophosphate Man9GlcNAc2 alpha-1,3-glucosyltransferase
EOG09370K1R	ATP5F1C: ATP synthase subunit gamma, mitochondrial
EOG09370YC4	ATP5PO: ATP synthase subunit O, mitochondrial
EOG09370P7I	ATPAF2: ATP synthase mitochondrial F1 complex assembly factor 2
EOG0937129K	DAD1: Dolichyl-diphosphooligosaccharide--protein glycosyltransferase subunit
EOG093705U4	DCP2: m7GpppN-mRNA hydrolase
EOG09370DX4	DCPS: m7GpppX diphosphatase
EOG09370KGV	DCTD: Deoxycytidylate deaminase
EOG09370CTU	DDOST: Dolichyl-diphosphooligosaccharide--protein glycosyltransferase 48 kDa subunit
EOG09370DXT	DMAP1: DNA methyltransferase 1-associated protein 1
EOG09370ZI7	DNTTIP2: Deoxynucleotidyltransferase terminal-interacting protein 2
EOG0937036L	EIF3A: Eukaryotic translation initiation factor 3 subunit A
EOG09370VFL	EIF3H: Eukaryotic translation initiation factor 3 subunit H
EOG09370JW6	ELP4: Elongator complex protein 4
EOG09370CK6	GTF2E1: General transcription factor IIE subunit 1
EOG09370DS4	GTF2F2: General transcription factor IIF subunit 2
EOG09370BK5	GTF2H1: General transcription factor IIH subunit 1
EOG09370QBK	GTF2H3: General transcription factor IIH subunit 3
EOG09370E36	HAT1: Histone acetyltransferase type B catalytic subunit
EOG09370NBW	IWS1: Protein IWS1 homolog
EOG09370XFJ	KIN: DNA/RNA-binding protein KIN17
EOG093704J7	MCM5: DNA replication licensing factor MCM5
EOG093714Q2	MED31: Mediator of RNA polymerase II transcription subunit 31
EOG093710A7	MNAT1: CDK-activating kinase assembly factor MAT1
EOG093702MJ	NCBP1: Nuclear cap-binding protein subunit 1
EOG093718E9	NDUFA2: NADH dehydrogenase [ubiquinone] 1 alpha subcomplex subunit 2
EOG093712G8	NDUFA8: NADH dehydrogenase [ubiquinone] 1 alpha subcomplex subunit 8
EOG09370RRJ	NDUFS3: NADH dehydrogenase [ubiquinone] iron-sulfur protein 3, mitochondrial
EOG09370RIT	NDUFS8: NADH dehydrogenase [ubiquinone] iron-sulfur protein 8, mitochondrial
EOG09370GEO	ORC2: Origin recognition complex subunit 2
EOG09370KWF	PGAP3: Post-GPI attachment to proteins factor 3
EOG09370L77	POLA2: DNA polymerase alpha subunit B
EOG09370UV4	POLR2G: DNA-directed RNA polymerase II subunit RPB7
EOG093710JH	POLR2I: DNA-directed RNA polymerase II subunit RPB9
EOG09370QT3	POP4: Ribonuclease P protein subunit p29
EOG09370YSI	RBBP5: Retinoblastoma-binding protein 5
EOG0937106D	REXO2: Oligoribonuclease, mitochondrial
EOG0937192A	RPP30: Ribonuclease P protein subunit p30
EOG09370FKI	RTF1: RNA polymerase-associated protein RTF1 homolog

EOG09370QCX	SEC62: Translocation protein SEC62
EOG0937183G	SPCS1: Signal peptidase complex subunit 1
EOG09370QKS	SSU72: RNA polymerase II subunit A C-terminal domain phosphatase
EOG09370TTI	TAF11: Transcription initiation factor TFIID subunit 11
EOG0937122Q	TAF13: Transcription initiation factor TFIID subunit 13
EOG093703W1	TAF2: Transcription initiation factor TFIID subunit 2
EOG0937186Q	TAF9: Transcription initiation factor TFIID subunit 9
EOG09370DYK	TAMM41: Phosphatidate cytidylyltransferase, mitochondrial
EOG09370BGO	THOC1: THO complex subunit 1
EOG09370LYY	TIM44: Mitochondrial import inner membrane translocase subunit
EOG093718EG	TIMM10: Mitochondrial import inner membrane translocase subunit
EOG09370RCQ	YJU2: Splicing factor YJU2

Uno de los casos que nos llamó la atención su ausencia en tripanosomátidos, es el complejo ORC. Este es un complejo de proteínas formado por seis subunidades (1-6) que participa en el reconocimiento de los sitios de origen de replicación del ADN, a través del dominio AAA ATPasa, muy conservado en eucariotas. El proyecto de secuenciación genómica, mediante comparaciones secuencia-secuencia, logró identificar solamente una proteína de este complejo: ORC1 (El-Sayed et al., 2005a). Por lo tanto, ORC2 surge como un buen candidato para testar las capacidades de DARK.

El primer paso del análisis consistió en la búsqueda de la palabra clave ORC2, donde se obtuvieron dos posibles agrupamientos, ambos presentes en todas las subfamilias; 4036 cuenta con 38 miembros y 51088 con 36. Los resultados se exhiben en la Figura 3.20.

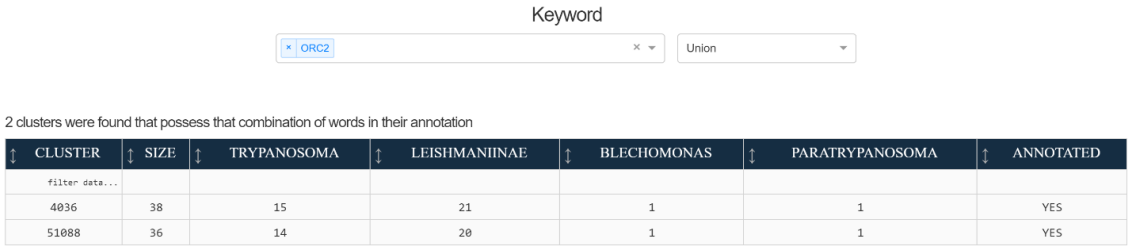


Figura 3.20. Resultados de la búsqueda de ORC2 en DARK. Se obtuvieron dos agrupamientos los cuales están conservados en todas las subfamilias estudiadas.

En el primero de ellos, DARK menciona a ORC2 como parte de una anotación derivada de PDB: 4XGC. Este identificador hace referencia a la estructura cristalográfica de las proteínas del complejo ORC de *Drosophila menalogaster*. Sin embargo, la mayor parte de las anotaciones sugieren que esta proteína es ORC1. La anotación derivada de TriTrypDB es diversa, 23 miembros son PH, 9 están anotadas como proteínas contenedoras de

dominio AAA ATPasa, mientras 5 están anotadas como ORC1. Como se ha mencionado previamente, ORC1 ya ha sido identificada y el agrupamiento analizado coincide con este resultado (El-Sayed et al., 2005a; Godoy et al., 2009).

Si analizamos el siguiente *cluster*, encontramos que el 100 % de las proteínas están anotadas como hipotéticas en TriTrypDB. DARK por otra parte sugiere que está proteína podría corresponder a ORC2, por la anotación de PDB y SCOP, tal como se observan en Figura 3.21.

Cluster	51088
HMM Length	1025
UNICLUST30 (1)	WGS project CAEQ00000000 data, annotated contig 1340 [37|1/1|F9W5L7: 100.0|2e-158|1-602]
UNICLUST30 (2)	-
UNICLUST30 (3)	-
PFAM_A (1)	-
PFAM_A (2)	-
PFAM_A (3)	-
PDB70 (1)	Orc2, Orc3, Orc5, Orc1, Orc6 [57|1/3|4XGC_B: 100.0|1.2e-43|163-353]
PDB70 (2)	DNA replication licensing factor MCM2 [54|1/3|5UDB_B: 100.0|1.1e-42|417-618]
PDB70 (3)	Origin recognition complex subunit 2 [27|1/3|5C8H_A: 99.6|3.8e-21|17-119]
SCOP70 (1)	ORC2 , Origin recognition complex subunit 2 [57|1/1|PF04084.13: 100.0|1.8e-39|134-307]
SCOP70 (2)	-
SCOP70 (3)	-

Figura 3.21. Resultados de anotación producidos por DARK para el cluster 51099. En él se sugiere que puede tratarse de la proteína ORC2.

Analizando los comentarios disponibles en TriTrypDB, encontramos que una proteína (Tb927.9.4530) de *T. brucei* presenta localización nuclear (determinado mediante espectrometría de masas). Para confirmar la localización sub-celular utilizamos el proyecto TrypTag (Dean et al., 2017) (proyecto que tiene como objetivo determinar dónde se localiza cada proteína de tripanosoma dentro de la célula). La Figura 3.22, muestra como efectivamente está proteína se encuentra mayormente localizada en el núcleo. La localización nuclear coincide con lo esperada para las proteínas del complejo ORC.

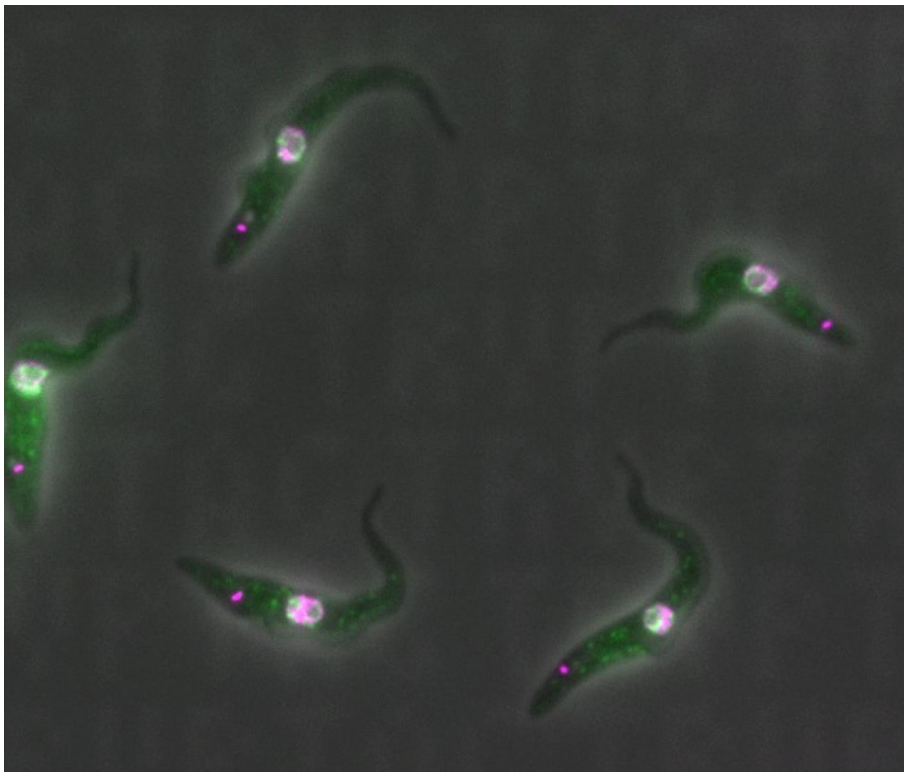


Figura 3.22. Localización principalmente nuclear de la proteína Tb927.9.4530 de *T. brucei* observada mediante el solapamiento de fluorescencia y contraste de fase, obtenida de TrypTag.

Al utilizar las capacidades de minería de texto derivadas de DARK encontramos un estudio reciente publicado donde se estudia el complejo de proteínas ORC en *T. brucei* (Marques et al., 2016). Interesantemente, en este artículo sugieren que la proteína de *T. brucei* sugerida por DARK es efectivamente ORC2 ya que co-inmunoprecipita con ORC1 (Godoy et al., 2009); contribuye a la replicación del ADN a través del mismo complejo y presenta similitudes estructurales con la estructura cristalográfica de ORC2 de *Drosophila menalogaster* (Marques et al., 2016). Interesantemente, la cadena B del 4XGC (reportada por DARK), es la misma estructura cristalográfica utilizada por Marques et al. para caracterizar ORC2

Este ejemplo demuestra como DARK es capaz de anotar correctamente una proteína cuya funcionalidad es imprescindible en eucariotas pero que las metodologías tradicionales de anotación fueron incapaces de anotarla, por la gran divergencia a nivel de secuencia primaria.

Otro ejemplo donde testeamos las capacidades de DARK se relaciona con una publicación de nuestro grupo donde se predicen proteínas mitocondriales de *T. cruzi* (Becco et al., 2019). Este trabajo identifica la proteína hipotética TcCLB.458015.4 como posible proteína mitocondrial. Esta forma parte del agrupamiento 16881 el cual está compuesto por 39 miembros. En este *cluster* están todas las subfamilias y todas las especies representadas, por lo que nos permite suponer que es una proteína ancestral de la familia Trypansomatidia. De las 39 anotaciones presentes en el TriTrypDB, 38 de ellas aparecen como proteína hipotética, mientras que Tb927.11.10780 (*T. brucei*) está anotada como canal aniónico dependiente de voltaje. DARK, lo anota como una porina mitocondrial basándose principalmente en homología encontrada a nivel de estructura terciaria (PDB y SCOP). En cuanto a la minería de la literatura este *cluster* fue vinculado con un artículo, en el cual se analiza el importoma mitocondrial de *T. brucei* (Peikert et al., 2017) en el que observan la proteína previamente mencionada. Finalmente, utilizando el proyecto TrypTag encontramos que esta proteína posee localización mitocondrial (etiquetado N-terminal) y citoplásmica (etiquetado C-terminal). La evidencia aludida sugiere que efectivamente TcCLB.458015.4 es una proteína mitocondrial.

Finalmente, decidimos estudiar la presencia de proteínas relacionados al flagelo, dado que está estructura es fácilmente identificable mediante microscopía. Utilizando nuevamente la herramienta de búsqueda de función nos llamó la atención el agrupamiento 5206, el cual se presenta como una posible proteína asociada al flagelo. Este agrupamiento formado por 40 miembros está conservado en todas las subfamilias, el 100 % está anotada como hipotética (o DUF: dominio de función desconocida) y la proteína de *T. brucei* (Tb927.7.4510) está regulada por RBP10, la cual promueve la diferenciación al estadio sanguíneo (Mugo and Clayton, 2017).

Para verificar si efectivamente los miembros del cluster son proteínas flagelares, verificamos su localización subcelular mediante TrypTag. La Figura 3.23, muestra como Tb927.7.4510 se localiza efectivamente en el flagelo. Esta proteína representa un caso muy interesante para profundizar ya que es una proteína ancestral dentro del grupo, es regulada por RBP10 y no encontramos ninguna publicación que reporte estudios sobre ella.

Figura 3.23. Localización subcelular de Tb927.7.4510 obtenida de TrypTag, mediante el solapamiento de fluorescencia y contraste de fase. Se concluye que se trata de una proteína asociada al flagelo.

Estos ejemplos demuestran la potencia de DARK para anotar proteínas con función desconocida en tripanosomátidos (se lograron anotar más de 2500 *clusters* compuestos exclusivamente por proteínas hipotéticas), así como encontrar proteínas relacionadas con funciones de interés, integrar información ya presente en el TriTrypDB y encontrar fácilmente bibliografía relacionada, entre otros usos que dependerán de las interrogantes del usuario. Cabe destacar que nos encontramos escribiendo un manuscrito para presentar esta herramienta y los resultados obtenidos durante la tesis.

3.2.1.9. Estrategia

3.2.1.9.1. Descarga de datos masiva del TriTrypDB

Para descargar toda la información presente en la base de datos se hizo uso de los documentos WADL disponibles en el sitio web de TriTrypDB. Un WADL es un documento XML que describe en detalle cómo formar una URL para realizar una búsqueda como una solicitud de servicio web. Por ende, mediante la generación de URL y haciendo uso del acceso programático se generó un *script* en Python capaz de descargar la secuencia aminoacídica e información de cada proteína. La información obtenida incluye, nombre,

organismo, largo de secuencia, producto génico, términos de ontología génica, anotaciones de distintas bases de datos, etc., de la versión 40 del TriTrypDB.

3.2.1.9.2. Generación de agrupamientos

Los agrupamientos fueron producidos con MMseqs2. Este, tiene la ventaja de estar diseñado para ser utilizado de forma paralela (múltiples núcleos) y es fácilmente escalable. MMseqs2 posee tres módulos de funcionamiento para lograr la comparación de dos conjuntos de secuencias, el módulo de pre-filtrado, el de alineamiento y el de *clusterización*. El primero, computa la similitud entre todas las secuencias de una base de datos de consulta contra las secuencias de una base de datos blanco. Esta parte utiliza una búsqueda de *k-mers* seguido por alineamientos sin *gaps*. Posteriormente todos aquellos alineamientos que pasaron cierto *cutoff* del módulo de pre-filtrado, son sujetos a alineamiento del tipo Smith-Waterman vectorizado (con *gaps*) y finalmente las proteínas son agrupadas. Todos los modos de agrupamiento transforman los resultados de alineación en un grafo no dirigido. En esta notación de grafos, los vértices representan las proteínas, las cuales están conectadas por un borde. Se introduce un borde (conexión entre nodos) entre las proteínas si los criterios de alineación son satisfechos.

El primero paso para obtener un agrupamiento es convertir un archivo fasta que contenga la información de todas las proteínas de tripanosomátidos en una base de datos de formato MMseqsDB. Una vez obtenida, se procede a realizar el agrupamiento con el comando *cluster* de MMseqs2. El comando *cluster* requiere información referente a los límites establecidos para el alineamiento y el modo de agrupación. En este caso decidimos utilizar un mínimo de 70% de cobertura, una identidad mayor al 50% y un e-valor menor a 0.001 para que sea considerado un *hit.* Además, se utilizó el modo de alineamiento número 3, el cual si bien es el más lento, es el más preciso. Por otra parte, el modo de agrupamiento elegido (modo 0), utiliza el algoritmo *Greedy Set cover* el cual presenta un tiempo de resolución polinomial. Este algoritmo, elimina el nodo con las mayores conexiones y todos los que estén conectados a él. Estos forman un agrupamiento y el procedimiento se repite hasta que todos los nodos estén presentes en algún agrupamiento (si no presenta conexiones con otro nodo se le asigna su propio

agrupamiento). *Greedy Set cover* es seguido por un paso de reasignación. Los miembros del agrupamiento se asignan a otro centroide si su puntaje de alineación fue mayor (Figura 6.3).

3.2.1.9.3. Alineamiento múltiple de los agrupamientos

Para realizar el alineamiento múltiple de los agrupamientos seleccionados se utilizó el programa T-Coffee (versión 11.00.8cbe486), utilizando parámetros por defecto, a través de una ejecución iterativa mediada por un programa escrito en el idioma Python (material suplementario).

3.2.1.9.4. Conversión a perfiles HMM

Cada alineamiento múltiple fue convertido en perfil de HMM mediante la utilización de los módulos reformat y hmmake del programa HHblits. El módulo reformat me permite cambiar el formato de los alineamientos (de fasta a a3m), mientras que mediante el módulo hmmake se generan los perfiles.

3.2.1.9.5. Descarga de perfiles HMM pertenecientes a bases de datos públicas y comparación de perfiles.

Se descargaron los perfiles de distintas bases de datos de la página http://wwwuser.gwdg.de/~compbiol/data/hhsuite/databases/hhsuite_dbs/.

En particular se utilizaron PDB70 (versión pdb70_from_mmcif_17Nov17), Scop70 (versión scop70_1.75), PFAM-A (versión 31) y UniClust30 (versión 2017_10). La base de datos PDB (Protein Data Bank) contiene información sobre estructuras de proteínas, ácidos nucleicos y ensamblajes complejos determinados experimentalmente, en particular se utilizó PDB70 (agrupamientos al 70% de identidad). La base de datos PFAM es una base de datos de dominios proteicos, de los cuales existen perfil de HMM y la base de datos SCOP, presenta proteínas que están estructuralmente caracterizadas y depositadas en el PDB, también agrupadas al 70% de identidad de secuencia. Finalmente,

se utilizó UniClust30, que corresponde a los resultados de agrupación del UniProtKB al 30% de identidad.

La comparación de perfiles HMM-HMM se realizó mediante el programa HHblits, estableciendo un máximo de 500 *hits* por comparación. A excepción de la búsqueda contra UniClust30, se utilizó el modo local de homología. Dado que el resto de las bases se basan en aspecto estructurales o en dominios, el modo local era más adecuado. Las comparaciones fueron realizadas de forma independiente para cada base de datos utilizada.

3.2.1.9.6. Anotación de agrupamientos

Sin importar la base de datos considerada, se conservaron los *hits* cuyo e-valor sea menor a 0.005 y la probabilidad asociada sea mayor a 90%. Luego, se tomó la función (sentencia) de cada uno de ellos y se individualizaron las palabras utilizando el paquete de Python NLTK (kit de herramientas de lenguaje natural) (Loper and Bird, 2002). Una vez obtenida la bolsa de las palabras se eliminaron los signos de puntuación y se determinaron las palabras no informativas (*stop words*). Las *stop words* seleccionadas son: family, protein, domain, containing, putative, like, WGS, project, data, annotated, contig, ensemble, refinement, methodology, predicted, unplaced, genomic, scaffold, whole, genome, shotgun, sequence. Estas palabras son muy frecuentes entre las sentencias y no aportan información relevante, por ende, para que no influyan en los puntajes de las sentencias, su valor asignado de frecuencia fue de 0. Una vez reducida la bolsa de palabras se calcularon las frecuencias de ellas, y se estableció un rango de frecuencias de mayor a menor seleccionando las 10 mayores. A estás se les asignó una escala de valor del 10 al 1.5, mientras que al resto de las frecuencias un valor de 0. Finalmente, a cada palabra se le estableció el puntaje asociado a su frecuencia.

La definición de puntajes de las sentencias se basó en tres aspectos:

1. Puntaje de las palabras que lo componen.
2. Número de veces que se repite la sentencia.
3. Probabilidad del *hit* asociado a la sentencia.

El primer punto corresponde a la sumatoria de los puntajes asociados a cada palabra. Cuanto mayor representación tengan las palabras que componen una sentencia, mayor puntaje tendrá. El segundo punto hace pesar cuantas veces está representada una sentencia particular en el conjunto evaluado y se traduce directamente al puntaje. Finalmente, el tercer punto, es un agregado diferencial (20 puntos) para las sentencias cuyo *hit* tengan un 100% de probabilidad. Este sistema de puntaje fue ajustado de forma manual para poder captar las funciones más representativas asociadas a un agrupamiento.

Como último paso asociado a la anotación de los agrupamientos se definió una metodología para poder reportar la mayor variabilidad posible dentro de las sentencias que resultaron más representativas. La estrategia consistió en la aplicación del algoritmo de distancia de Levenshtein mediante el paquete de Python fuzzywuzzy. Informalmente, la distancia de Levenshtein entre dos sentencias es el número mínimo de ediciones de un solo carácter (inserciones, eliminaciones o sustituciones) necesarias para cambiar una sentencia por otra. El algoritmo es una versión levemente modificada dado que individualiza las palabras, las ordena y luego las compara. Este reporta un porcentaje que indica la probabilidad de que las sentencias tengan el mismo origen, cuanto mayor sea ese porcentaje más similares serán las sentencias. En total se reportan tres sentencias, la primera es la de mayor puntaje, la segunda es la que le sigue en puntaje sí y solo sí la distancia de Levenshtein es menor al 50 % (70% en el caso de UniClust30) y finalmente, la tercer sentencia reportada será la que continué en puntaje, siempre y cuando se logre diferenciarse de la primera y segunda.

Un ejemplo del reporte presentado por DARK se presenta en la Figura 3.24.

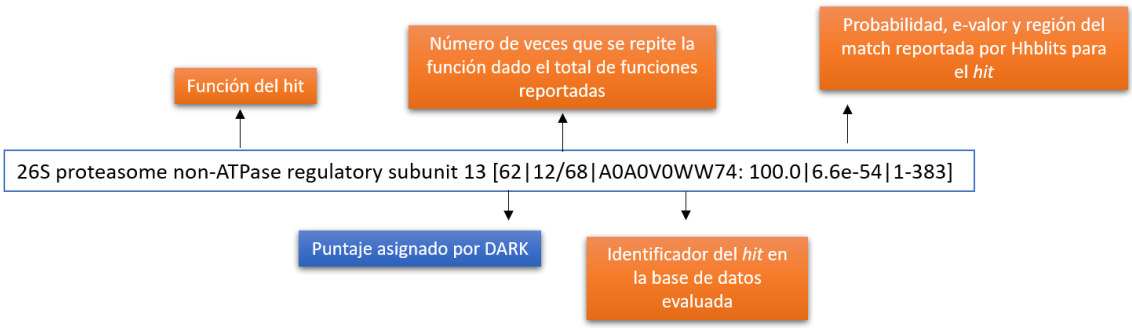


Figura 3.24. Resultado de anotación reportado por DARK del agrupamiento número 5 con la base de datos UniClust30. Los recuadras anaranjados marcan los datos extraídos directamente del resultado de HHblits, mientras que en azul se marca el puntaje asignado por DARK a ese *hit* en particular.

3.2.1.9.7. Anotación de ontología génica

Los datos de ontología génica de las proteínas de UniProtKB fueron descargados de la base de datos Gene Ontology Annotation (GOA) (Huntley et al., 2015), en su versión 14/1/19. Para asociar y reportar los términos asociados a cada agrupamiento se elaboró un programa en Python capaz de vincular los identificadores de UniProt KB del archivo GOA con los identificadores de los *hit*s obtenidos de las comparaciones con UniClust30 para cada agrupamiento. Para trabajar con tan grandes volúmenes de datos, hicimos uso del paquete Dask (Rocklin, 2015).

Para definir los términos GO asignados a cada agrupamiento se tomaron los términos más frecuentes. El valor umbral para que un término sea reportado se fijó en un 25 % del valor de frecuencia del término más abundante; estableciendo el mínimo en 2. Si con la estrategia anterior no se llega a un mínimo de 3 términos, se reportan los 3 más frecuentes.

3.2.1.9.8. Desarrollo de DARK como herramienta de interfaz gráfica.

La interfaz gráfica fue desarrollada en Python haciendo uso del paquete Dash, pudiéndose visualizar en cualquier navegador. Los gráficos interactivos que presenta la

herramienta fueron generados con Plotly. DARK ha sido testeado en Ubuntu 18 y Windows 10 y se encuentra disponible en Github (https://github.com/sradiouy/DARK).

3.2.1.9.9. Análisis de proteínas conservadas mediante Busco

La herramienta BUSCO (versión 3) fue utilizada para encontrar en los genomas de TriTrypDB proteínas que pertenecen a grupo de ortólogos conservados en el linaje eucariótico. Se utilizaron parámetros por defecto con la salvedad del establecimiento de un valor e de 0,001 cómo mínimo para establecer un *match*. Para anotar funcionalmente los grupos de ortólogos se utilizó la función hmmsearch del programa HMMER (versión 3.1) con parámetros por defecto.

3.2.2. IdMiner

En el transcurso de esta tesina nos enfrentamos en múltiples ocasiones a analizar listas de proteínas reguladas entre condiciones. El paso natural cuando se obtienen una lista de este tipo es analizar las categorías génicas sobrerrepresentadas. Este análisis nos puede dar una imagen general de la cual se pueden hilvanar hipótesis. Para testar estas hipótesis se debe conducir una búsqueda profunda de la bibliografía y muchas veces el número de proteínas presentes hace que esta tarea sea abrumadora.

En la presente sección presentaremos IdMiner (https://github.com/sradiouy/IdMiner), a través de un manuscrito en preparación (para ser presentado en formato de *short communication*) que adjuntamos a continuación. Brevemente, IdMiner es una herramienta que mediante *text-mining* permite encontrar términos sobrerrepresentados en la literatura asociados a proteínas de interés, así como relaciones entre los términos y entre las proteínas (objetivo específico 2).

IdMiner: text mining of paper abstracts to explore overrepresented terms from gene lists

Introduction

The advent of deep sequencing technologies has resulted in a rapid growth of genomic sciences. The initial bioinformatic analysis of genome wide studies frequently results in a list of a few hundred relevant gene identifiers. Extracting the biological meaning of these lists is central to understand the underlying biological process. This presents the investigator with the task of establishing functional correlations among several genes. Gene annotation enrichment analysis is a way to produce relevant relationships by classifying the genes in pre-established functional categories and finding out the ones that are present more frequently than expected by chance (Huang da et al., 2009). This method is widely used and certainly helpful, but it also presents limitations, mainly related to the quality of the annotation itself. One important drawback is that gene databases do not incorporate new annotation rapidly enough as they cannot cope with the publication rate. This means that to extract more information from gene lists the researcher must manually retrieve relevant data contained in biomedical bibliography. However, this task is complex and time-consuming as the number of pertinent published papers is large, and the trend marks that the rate of publication would be increasing exponentially. Besides, the multidisciplinary characteristics of current research projects means that important information may reside in journals not commonly related to the research under course (Khare et al., 2014). Text mining offers the possibility to ease these issues by presenting the information contained in papers in an ordered and hierarchical way.

Here we describe IdMiner a tool that aids the investigator to search and analyze relevant bibliography related to genes of interest that result from high throughput studies. The tool allows to uncover terms that are frequently mentioned in the set of papers related to gene lists and to establish relationships with between different frequent terms, the terms and

their associated genes, and between genes that are mentioned, and to easily retrieve the papers where the terms/genes (or combinations) are mentioned. As a consequence, the investigator can check previous hypothesis or generate new ones.

Results

IdMiner accepts as input a list of gene identifiers or a multifasta file containing the protein sequences that are coded by the genes of interest (Figure 1.A). Using the functionality given by the PaperBlast tool (Price and Arkin, 2017) IdMiner will obtain the PubMed ids (PMIDs) for all the papers related to each gene and its homologous (according to a sequence similarity and blast coverage cutoffs selected by the user). This PMIDs are used to retrieve the abstracts that are then analyzed by the tool using text mining approaches. By default, IdMiner will collect the 5000 most frequently used terms in the set of abstracts, excluding terms commonly present in the English language (with three cutoff levels chosen by the user in the GUI) except those that may be biologically relevant (such as "human", "cancer", etc). The user may also further refine this step by establishing specific terms to be kept or to be excluded (Figure 1.A).

The results can be explored in the GUI performing a "Term centered" or a "Gene centered" approach. When exploring the results through a Term Centered approach, IdMiner will display most frequent words with information about the number of genes that are associated to that word, its ZIPF Score (a measurement of the frequency of the word in the English language), the number of abstracts in which this word is present and its raw count. The user can then select a specific term or set of terms for further analysis. With the queried terms, IdMiner will draw a diagram showing the gene ids associated with them, colored according to the percentage of the abstracts related to each gene. In addition, this search can be either be performed in either an "intersection" mode, which will present only the articles where all the terms searched are present, or an "union" mode, which will show the articles related with any of the searched terms. Clicking on the plot's title will direct the user to a PubMed page displaying the list of papers related to the user's query. If

the user clicks on a single gene, only papers related to that gene will be displayed in PubMed (Figure 1.B). Finally, the analysis can also be performed in a "Gene Centered" way. In this case IdMiner constructs a table showing the number of articles that associate the original genes among them. Querying a specific gene, will generate a plot where all the genes sharing articles with the selected gene will be shown. In these case colors indicate the number of shared articles. Clicking on one of the plotted IDs will open the corresponding articles in PubMed (Figure 1.C).

Conclusion

Globally, IdMiner allows to rapidly portrait the biological terms most frequently found in the literature as related to the genes of interest, offering a quick and intuitive way to get to these relevant papers. Also, relationships between terms and between genes can be uncovered and explored, where again the literature linking them is readily obtained. This software provides a shortcut to a process that would be extremely arduous otherwise. IdMiner will take around 30 minutes to analyze a list of 100 genes (may vary depending on selected maximum number of terms to analyze) on a standard laptop computer (i7 intel processor with 8 GB RAM memory). IdMiner is open source, multiplatform (Linux, Windows and Mac) and user friendly (easy to install and use with a web GUI).

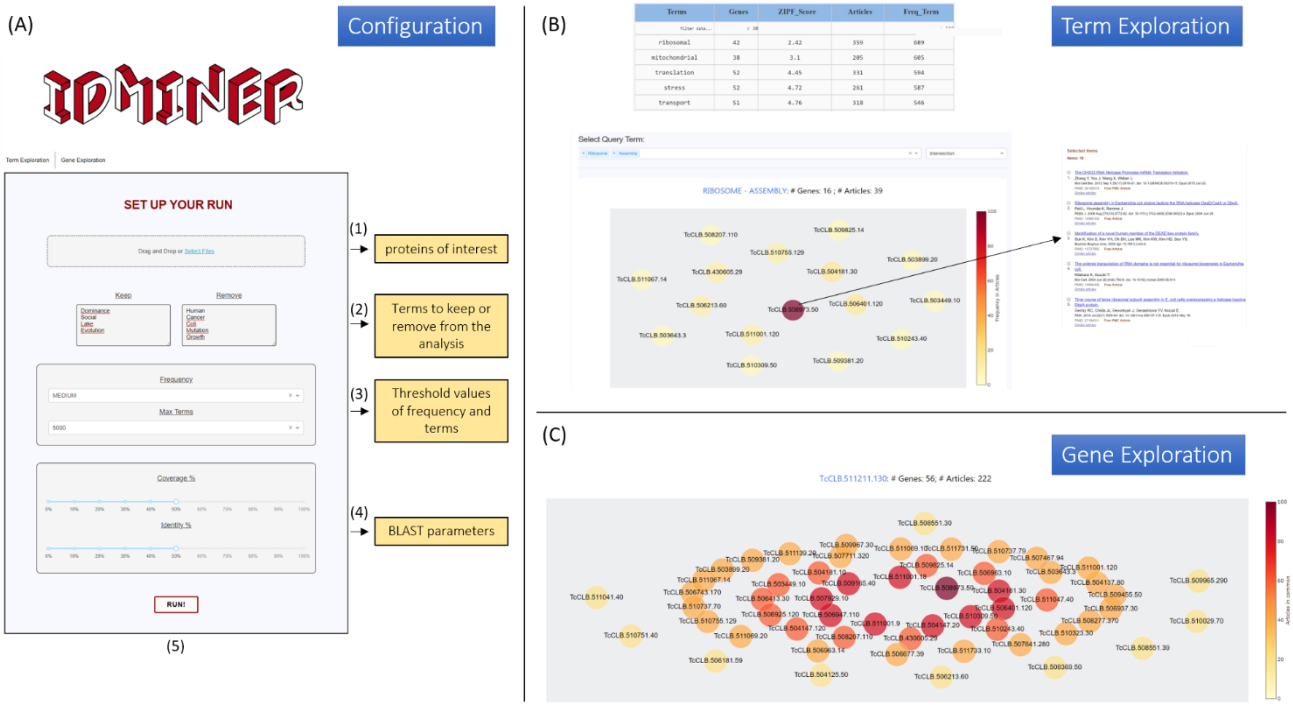


Figure 1. Case study showing the IdMiner analysis of genes that decrease their translational efficiency significantly during *T. cruzi* metacyclogenesis (data obtained from Smircich et al., 2015) A) Configuration panel as displayed in the Vivaldi web browser. IdMiner is configured in a simple way and involves five steps. The first is to upload a file that contains a list of protein identifiers (.txt extension) or a multifasta protein sequence file (.fasta extension). The second step involves selecting words to keep or remove from the analysis regardless of their ZIPF score. The third involves filtering the terms analyzed according to their ZiPF score (5-High, 3.5-Medium, 2-Low) and limiting the number of terms analyzed (by default to 5000). The fourth step is to define the coverage and identity values to filter the BLAST results obtained by PaperBlast. Finally, by pressing the "run" button IdMiner is executed with the selected parameters. B) The term exploration tab is displayed. The table above shows the 5 terms linked to the largest number of genes and below is a search that includes all the abstracts that have two of these words (ribosome and biogenesis) in them. The las panel shows that by clicking on the node (gene) with the largest number of associated abstracts, IdMiner automatically opens a PubMed tab where the abstracts associated with the gene and the terms are displayed. C) Gene exploration. The diagram shows the relationship of the gene that presents the greatest amount of abstract in common with other genes. Clicking on each of the nodes will produce a PubMed search with the articles common to both.

3.3. Generación de un *software* enfocado en kinetoplástidos que permita definir regiones UTRs de los ARNm.

Como mencionamos en la introducción, las regiones UTRs contienen los elementos fundamentales que controlan la expresión génica en tripanosomátidos. Por lo tanto, nos planteamos como objetivo desarrollar una herramienta que permita definirlas de forma precisa (objetivo específico 3). Los datos obtenidos con esta herramienta permitirán continuar con el desarrollo de la tesis mediante el análisis de regiones regulatorias a nivel general (objetivo específico 4, sección 3.4), así como de familias reguladas traduccionalmente (objetivo específico 5, sección 3.5).

Los resultados de este capítulo y la discusión de los mismos fueron publicados en el journal Frontiers in Genetics (Radio et al., 2018). El programa se encuentra disponible en GitHub (https://github.com/sradiouy/UTRme).

El material suplementario del artículo se encuentra disponible en: https://www.frontiersin.org/articles/10.3389/fgene.2018.00671/full#supplementary-material

frontiers in Genetics

METHODS
published: 18 December 2018
doi: 10.3389/fgene.2018.00671

UTRme: A Scoring-Based Tool to Annotate Untranslated Regions in Trypanosomatid Genomes

Santiago Radío[1,2], *Rafael Sebastián Fort*[1,2], *Beatriz Garat*[2], *José Sotelo-Silveira*[1,3] *and Pablo Smircich*[1,2]*

[1] Department of Genomics, Instituto de Investigaciones Biológicas Clemente Estable, MEC, Montevideo, Uruguay,
[2] Laboratory of Molecular Interactions, Facultad de Ciencias. Universidad de la República, Montevideo, Uruguay,
[3] Department of Cell and Molecular Biology, Facultad de Ciencias, Universidad de la República, Montevideo, Uruguay

Most signals involved in post-transcriptional regulatory networks are located in the untranslated regions (UTRs) of the mRNAs. Therefore, to deepen our understanding of gene expression regulation, delimitation of these regions with high accuracy is needed. The trypanosomatid lineage includes a variety of parasitic protozoans causing a significant worldwide burden on human health. Given their peculiar mechanisms of gene expression, these organisms depend on post-transcriptional regulation as the main level of gene expression control. In this context, the definition of the UTR regions becomes of key importance. We have developed UTR-mini-exon (UTRme), a graphical user interface (GUI) stand-alone application to identify and annotate 5′ and 3′ UTR regions in a highly accurate way. UTRme implements a multiple scoring system tailored to address the issue of false positive UTR assignment that frequently arise because of the characteristics of the intergenic regions. Even though it was developed for trypanosomatids, the tool can be used to predict 3′ sites in any eukaryote and 5′ UTRs in any organism where trans-splicing occurs (such as the model organism *C. elegans*). UTRme offers a way for non-bioinformaticians to precisely determine UTRs from transcriptomic data. The tool is freely available via the conda and github repositories.

Keywords: post transcriptional regulation, untranslated region, UTR prediction software, prediction score, GUI

OPEN ACCESS

Edited by:
Alfredo Pulvirenti,
Università degli Studi di Catania, Italy

Reviewed by:
Panagiotis Alexiou,
Central European Institute of Technology (CEITEC), Czechia
Xiaohui Wu,
Xiamen University, China

***Correspondence:**
Pablo Smircich
psmircich@fcien.edu.uy

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *14 September 2018*
Accepted: *04 December 2018*
Published: *18 December 2018*

Citation:
Radío S, Fort RS, Garat B, Sotelo-Silveira J and Smircich P (2018) UTRme: A Scoring-Based Tool to Annotate Untranslated Regions in Trypanosomatid Genomes. Front. Genet. 9:671. doi: 10.3389/fgene.2018.00671

INTRODUCTION

Post-transcriptional regulation is a key step to control gene expression levels in eukaryotes (Franks et al., 2017) that depends on factors recognizing signals mostly present in the UTRs of the mRNAs. These mechanisms are crucial in trypanosomatids since they lack transcription initiation control. The trypanosomatid lineage includes a variety of parasitic protozoans causing significant worldwide burden on human health (Prüss-Ustün et al., 2016). Trypanosomatids represent early divergent eukaryotes that have evolved distinctive biological features; one of the most intriguing characteristic is the apparent lack of transcription initiation control, being initiation sites characterized only by chromatin modifications and DNA structural signals (Respuela et al., 2008; Siegel et al., 2009; Thomas et al., 2009; Wright et al., 2010; Ekanayake and Sabatini, 2011; Smircich et al., 2013; Ramos et al., 2015). This implies that the gene expression patterns result mainly from post-transcriptional control. Therefore, the regulation of mRNA localization (Pastro et al., 2017), stability (Fadda et al., 2014), and translatability (Jensen et al., 2014; Vasquez et al., 2014; Smircich et al., 2015)

are key mechanisms to determine protein concentration. These processes depend on regulatory proteins which interact with RNA by recognizing either sequence or structural signals present mainly on the UTRs of the mRNAs (Clayton, 2013; De Gaudenzi et al., 2013; Pastro et al., 2013). So, to deepen our understanding of gene expression regulation and the involved signals we need to delimit these regions with high accuracy. The annotation of UTR regions has been a challenging task depending on specific experiments designed for each particular gene. However, transcriptomic approaches currently give the opportunity to annotate these sites on a global scale. Efforts have been carried out to provide tools that allow the definition of UTR boundaries in trypanosomatids (Fiebig et al., 2014; Dillon et al., 2015). Although these tools have proven useful (Dillon et al., 2015; Pastro et al., 2017), both the repetitive nature of the trypanosomatid genomes and the high abundance of poly(A) tracts present in their intergenic regions confound the algorithms. Therefore, we have developed UTRme (UTR-mini-exon), a stand-alone application to identify and annotate 5′ and 3′ UTR regions, implementing a multiple scoring system that addresses both the aforementioned and several other issues that arise during the UTR annotation process. The tool provides not only the annotation but also a score that enables to discriminate the certainty of that annotation improving the usability of the results. Additionally, UTRme offers a Graphical User Interface (GUI) which turns it user friendly to non-bioinformaticians and, as a stand-alone application, can be scaled to any project depending only on the user's hardware. UTRme reports annotation and sequence files and plots general characteristics of the resulting data (such as the distribution of UTR lengths, UTRme scores and number of processing sites per gene). The 5′ UTR prediction can be easily extended to any organism where trans-splicing occurs, like the model organism *C. elegans*, among others (Lei et al., 2016). Furthermore, UTRme can be used for 3′ UTR prediction in any eukaryote. The source code is freely available at https://github.com/sradiouy/UTRme and can be easily installed via the conda repository on a linux based systems with a single command "*conda install -c sradiouy utrme.*"

METHODS

Genome Data

Genomic and coding sequences (cds) annotation files where downloaded from TritrypDB (http://tritrypdb.org/) release 35.

Transcriptomic Data Simulation

In order to test the software accuracy, a 30x 100 bp pair-end RNA-seq run was simulated using the Piquant package (https://github.com/lweasel/piquant). This package simulates sequencing errors and platform bias. To simulate reads originating from full transcripts [including UTRs, SL, and poly(A) sequences] a random length UTR was added to each *T. cruzi* coding sequence. For 5′ UTRs a maximum length of 101 bp was allowed while for 3′ UTRS the maximum length was set to 301 pb. The SL sequence or a 35 pb poly(A) tail was added to each end accordingly.

5′ End Enriched RNA-seq Library Construction

First strand of cDNA was prepared with 3 μg of purified RNA, random hexamers and Invitrogen SuperScript® III First-Strand Synthesis System (Pub. No. MAN0001346). Second strand of cDNA was prepared using a specific SL primer (5′tacagtttctgt actatattg3′) and DNA Polymerase I Large (Klenow) Fragment (NEB M0210). Library preparation protocol included end-repair, adapters ligation, size selection (Pipping Prep SAGE System), and amplification of the library using manufacturer's recommended protocol Ion plus fragment library kit (Pub. No. MAN0009847). Qualitative and quantitative assessment of the libraries was analyzed by Agilent 2100 Bioanalyzer System, using HS DNA 1000 reagents (Agilent Technologies). Emulsion amplification of the library was performed using Ion Onetouch 2 System with the Ion PGM Template OT2 Hi-Q view 400 kit (Pub. No. MAN0014579). Ion Sphere Particles (ISPs) enrichment step was performed on the Ion OneTouch ES system (Pub. No. MAN0014579). The Ion PGM system was used for sequencing using Ion PGM Hi-Q view Sequencing Solutions and Ion 318 Chip v2, following the manufacturer's recommended protocol for 400 bp reads (Pub. No. MAN0014583). (SRA BioProject PRJNA473354).

RESULTS AND DISCUSSION

Pipeline Description

The software was written in python (version 3) and depends on cutadapt (Martin, 2011), bedtools (Quinlan and Hall, 2010), bowtie2 (Langmead and Salzberg, 2012), samtools (Li et al., 2009), and python and unix modules. All dependencies are automatically configured during installation. UTRme needs a reference genome (sequence and cds annotation) and raw reads from an RNA-seq experiment (single-end or paired-end) (**Figure 1**). These required files, and optional arguments are selected through the GUI. Documentation, including a preview of the GUI, is available at https://github.com/sradiouy/UTRme.

The pipeline starts with the removal of adapter sequences and trimming of low-quality ends from reads using cutadapt. By default, UTRme trims the Illumina TrueSeq adapter, but any sequence can be specified. Afterwards, the trimming software is also used to identify and clip the reads containing putative poly(A) tails or spliced leader (SL) sequences, allowing for mismatches. By default, an error probability 0.01 for poly(A) sequences (adjustable by the user) and one mismatch for SL sequences are defined. To correctly identify the trans-splicing sites, the organism must be specified. Currently, *Leishmania major*, *Trypanosoma brucei,* and *Trypanosoma cruzi* are available, however other species can be included by adding specific SL sequences. This trimming process allows us to define two regions on a read (**Figure 2**). The primary region is the sequence that was left after read trimming, while the secondary region is the putative poly(A) tail or SL sequence recognized by cutadapt.

The primary regions of the reads are aligned to the genome using bowtie2 applying the default very-sensitive local end-to-end alignment mode (**Figure 1**). The subset of reads aligning to

intergenic regions is selected using bedtools. The mapping of the primary regions defines the putative splice acceptor site or poly(A) addition sites. At this point, UTRme evaluates in detail each putative site to assess its reliability by reporting a score that quantitates the confidence of the UTR site definition. This metric is calculated by combining an individual score that indicates the confidence with which each read predicts a given site, and global score that considers the cumulative evidence of all the reads that support a single processing site (see **Supplementary File 1** for a detailed description of all the scores and their calculation). The individual score includes three components: the primary, secondary and accessory scores. As read mapping is not always accurate, the primary score aims to assess the likelihood that the primary region was indeed transcribed in the genomic region that it was mapped to. This is estimated based on the evaluation of their similarity using a modified version of the Damerau-Levenshtein algorithm (Levenshtein, 1966; Majorek et al., 2014) implemented in the fuzzywuzzy python library (https://github.com/seatgeek/fuzzywuzzy). This metric evaluates the minimum number of changes that are required to go from string A to string B considering mismatches and gaps. Once the primary score has been measured and the read is not discarded, the secondary score is calculated. This evaluates the difference between the secondary region [putative poly (A) tail or SL sequence] and the genomic region contiguous to the primary region [by calculating the Hamming distance; (He et al., 2004)]. A true processing event would result in a sequence that is independent of this genomic region, so the greater the difference between the secondary region and the genomic region, the higher the score. In trypansomatids, where a high number A tracts repeats are present in the intergenic regions (Duhagon et al., 2011), a poly(A) in a read could be the result of transcription and not mRNA processing. Another aspect to consider is the length of the secondary region. The longer this sequence, more likely it represents a true post transcriptional event and this is included in the score. Also, the number of adenines in the secondary genomic region is also considered; a higher proportion of As result in a smaller the score. Finally, UTRme also considers aspects that influence the reliability of the processing site determination (see **Supplementary File 1**). Most are used to fine tune the final individual score and depends on features such as the confidence that the read was not misplaced during mapping, the presence of specific splicing signals (AG acceptor and polypyrimidine tract [poly(Y)] and the existence of unannotated open reading frames (ORFs) or undetermined nucleotides (Ns) in the defined region. As an example, the presence and characteristics of a poly(Y) tracts upstream of the trans-splicing site is verified. We defined poly(Y) tracts as the longest tract of pyrimidines not interrupted by more than a single purine (Dillon et al., 2015). The presence and composition of the tract is analyzed, and scores are assigned considering their

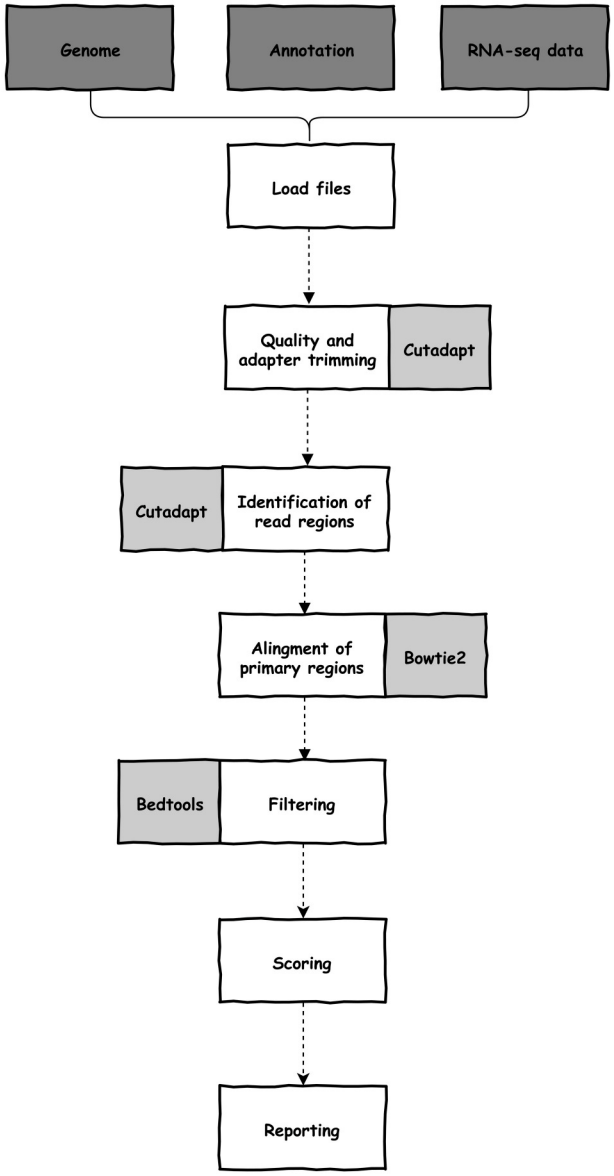


FIGURE 1 | Outline of the UTRme pipeline. Required initial files, data processing steps and software packages used during processing are depicted in dark gray, white and light gray backgrounds, respectively.

	Primary region	Secondary region
Read	AGTGTGTTGCGCTGaCGACTGACTGATCGATG	AAAAAAAAAAAAA
Genome	AGTGTGTTGCGCTGGCGACTGACTGtTCGATG	TGATGGAAAATGTT

FIGURE 2 | UTRme classification of read regions. Regions of each read and their counterparts in the genome are defined by UTRme as primary and secondary.

accordance with poly(Y) tract characteristics defined in (Siegel et al., 2005).

The global score considers the cumulative evidence of all the reads that support a single processing site giving a broader view of the accuracy of the site. For SL sites, it is proportional to the number of reads that support the site ("occurrences"). For poly(A) sites, in addition to the previous metric, the sequences of the putative poly(A) tail of all the reads that support the site is analyzed (for details see **Supplementary File 1**).

Finally, the reported score is calculated by adding the global score to the value of the third quartile of the individual's scores of all the reads that support that site. The maximum value for this score is set to 100. The higher the score the more confident is the prediction. All sites with positive scores are reported as they are supported by a reasonable amount of evidence. By default, if a site has a negative score it is not reported (this can be modified by the user).

In summary, the reported score recaps many aspects that influence the certainty that a site can be defined with the provided RNA-seq data.

Assessment of UTRme Accuracy

UTRme takes about 1 h to process 90M paired reads in a middle-sized hardware configuration (40 cores−3 Gb max. RAM footprint). The results are presented as tab—delimited text or excel files, report plots, annotation and sequence files.

Tables include a full report that details both the basic information of the site (such as associated gene, UTR length, acceptor dinucleotide for the SL, and site score) and also the different computed scores and other features of the site (information about the poly(Y) tract for the SL, maximum ORF sequence in the UTR -if its length is greater than 30 amino acids-, among others) (**Supplementary Table 1**). A summary report is also created where only basic information for the best scoring site is informed for each gene (**Table 1**).

UTRme generates both a sequence fasta file containing the sequences of the UTRs, as well as an annotation gff file that allows visualization and further analysis (**Supplementary Figure 1**).

TABLE 1 | Example of UTRme summary report output.

Gene	utr_len	acceptor	score	occurrences	# sites
TcCLB.397937.5	15	AG	89	418	4
TcCLB.398343.9	80	AG	79	2	2
TcCLB.399033.19	21	AG	90	27	4
TcCLB.400945.10	100	AG	85	39	4
TcCLB.404001.10	14	AG	95	59	3
TcCLB.404001.4	11	AG	91	75	5
TcCLB.404843.20	143	AG	92	65	2
TcCLB.405165.19	41	AG	92	54	4
TcCLB.407477.20	10	AG	91	64	2
TcCLB.407477.30	63	AG	96	51	4

Summary report of best scoring epimastigote's SL sites using epimastigote RNA-seq data from Li et al. (2016). The first 10 lines are shown.

This output is provided for all the sites and for the best scoring sites separately. Finally, the reported plots show general properties of the predicted UTRs (UTR lengths, scores, occurrences vs scores, number of sites per gene) (**Figure 3**).

To test the accuracy of the software, RNA-seq data from *T. cruzi* epimastigotes was obtained using an approach aimed to obtain a 5′ end enriched library. To improve mapping accuracy the average read size was set to 400 nt (see Methods section). 5′ processing sites were defined using UTRme and the best scoring ones where checked against previously published UTRs that were described though specific experimental approaches (**Table 2**) (Bontempi et al., 1994; Di Noia et al., 1998, 2000; Vandersall-Nairn et al., 1998; Teixeira et al., 1999; Búa et al., 2001; D'Orso and Frasch, 2001; Bartholomeu et al., 2002; Bhatia et al., 2004; Coelho et al., 2006; García et al., 2010).

Also, the availability of deep sequenced transcriptomes (Li et al., 2016) for the same *T. cruzi* stage, allowed us to check UTRme performance using reads obtained using a standard protocol RNA-seq experiment and shorter reads. As before, UTRme predictions were contrasted against the previously described UTRs. UTRme results for both approaches showed an excellent agreement with previously reported processing sites (**Table 2**). In most cases UTRme predicts the same UTR or a site that is within a few bases from the experimentally defined site, highlighting that the algorithm predicts sites with good precision. For those cases where the experimentally determined site was not identical to the best score site predicted by UTRme, the experimental site was usually present in the list of predicted sites with a lesser score. In the case of the deep sequenced transcriptome a greater number of processing sites was detected as reflected in the table.

To further validate our results in a genome wide scale, RNA-seq reads were simulated using randomly assigned UTRs. UTRme predicts 3′ UTRs for 7,116 genes, most of which (97.2%) are correctly assigned (within 5 nt distance of the real site). Considering multi mapping reads more genes are assigned a poly(A) site (7,884), but the accuracy diminishes significantly (91.4%). Taking into consideration the percentage of multi gene family members in the Tritryps genomes this is expected. This result prompted us not to consider multi-mapping reds by default. An analogous result is obtained for the miniexon addition site, assigning UTRs for 7,640 genes where 98.2% are correctly predicted, while when multi-mapping reads are considered the number of genes increases and a decrease in accuracy is observed (8,530 assigned 5′ UTRs with an accuracy of 92.5%).

It is interesting to note that when the dinucleotide of the 5′ splicing acceptor site is studied for the simulation, an overrepresentation of the AG dinucleotide is not observed. This is expected as UTRs lengths where randomly assigned. However, when this analysis is performed for real RNA-seq data, the AG dinucleotide is clearly the major acceptor site as expected (**Supplementary Figure 2**), reinforcing the accuracy of the annotations.

A key feature of UTRme is the reporting of a global score for each site. Positive scoring sites are given as they are supported by a reasonable amount of evidence. A higher

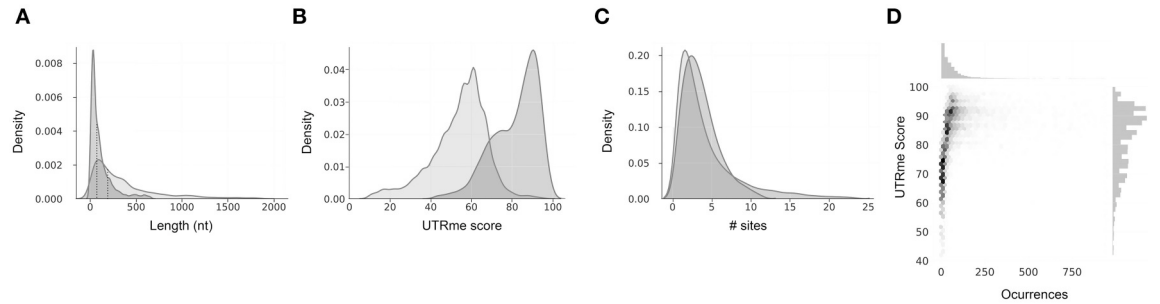


FIGURE 3 | Example of UTRme summary plots output. Reported plots for the 5′ and 3′ UTRs predicted using *T. cruzi* epimastigote Y strain RNA-seq data from Li et al. (2016). Plots for 5′ and 3′ UTRs are in dark gray and light gray, respectively. **(A)** Kernel density estimation plot of UTR lengths. **(B)** Kernel density estimation plot of both 5′ and 3′ UTR score distribution. **(C)** Kernel density estimation plot for the number of 5′ and 3′ UTR sites. In all cases the median is indicated as a dotted line. **(D)** Central panel: Scatter plot of 5′ UTR scores vs occurrences. A higher point density is indicated by a darker color for each bin. Upper panel: histogram of occurrences. Right panel: histogram of scores.

TABLE 2 | Comparison of UTRme predictions against experimentally defined processing sites.

Site	Gene	UTRme 5′ enriched	UTRme Li	UTRme Pastro	SlaP mapper pastro	Exp.	Article
5′	TcCLB.509147.50	48	51	51	54	55	Di Noia et al., 2000
5′	TcCLB.511679.10	51	51	51	54	51	Di Noia et al., 2000
3′	TcCLB.506533.142	786	786	764	–	789	Di Noia et al., 2000
3′	TcCLB.511679.10	–	375	–	–	∼353	Di Noia et al., 2000
5′	TcCLB.507485.140	–	140	137	–	137	Teixeira et al., 1999
5′	TcCLB.506407.10	93	102	101	718	103	Vandersall-Nairn et al., 1998
5′	TcCLB.509123.10	–	33	–	–	33	García et al., 2010
5′	TcCLB.505931.50	43	76	72	43	76	Bontempi et al., 1994
5′	TcCLB.507093.220	68	66	68	–	68	D'Orso and Frasch, 2001
5′	TcCLB.507639.30	42	42	42	42	42	Coelho et al., 2006
5′	TcCLB.507511.81	–	41	41	–	41	Di Noia et al., 1998
5′	TcCLB.510241.70	–	144	144	144	142	Bhatia et al., 2004
5′	TcCLB.506925.300	60	60	58	63	60	Búa et al., 2001
5′	TcCLB.506563.40	110	110	110	113	110	Bartholomeu et al., 2002

For UTRme predictions the best scoring site using T. cruzi epimastigote data is shown. UTRme 5′ enriched: UTRme predictions using In-house low pass sequencing of 5′ UTR enriched library. UTRme Li: UTRme predictions using Li et al. (2009) data. UTRme Pastro: UTRme predictions using Pastro et al. (2017) data. SLaP mapper Pastro: SLaP mapper predictions using Pastro et al. (2017) data. Exp., Experimentally defined sites. Article: Reference where the experimental prediction was described.

score indicates more evidence supporting the site. Using the simulated dataset, we explored the relationship between the UTRme score and the software performance. A plot that depicts the number of correct predictions (true positives) vs. the number of incorrect assignments (false positives) for various score cutoffs was constructed (**Figure 4A** for the 5′ UTRs results, see **Supplementary Figure 3A** for the 3′ results).

The figure clearly shows that increasing the score decreases rapidly the number of false positives. High scores (>80) show virtually no false positives; as the score decreases, both the number of both true and false positives increase, but true positives increase at a higher rate. When the score reaches a value around the average, this trend starts reverting. Even though further lowering the score accomplishes an increase in true positives, this is accompanied by an increased rate of incorrect assignments. It is important to notice that the maximum number of true positives is around 7000 sites, while the maximum number of false positives is <120, even for the lowest scores. All this indicates that, as expected, incorrect assignments tend to have lower scores. This is more clearly shown in **Figure 4B** (and **Supplementary Figure 3B** for 3′ sites) where the score and distance to the real site for incorrect assignments are plotted together with a histogram representing the score for all the sites. Most false positive annotations present low scores compared to the general distribution. All this evidence supports that UTRme is a very accurate tool and that the score reflects the reliability of the predicted sites.

To test the possibility of annotating UTRs outside trypanosomes, *Echinococcus granulosus* RNA-seq data (13 paired end data from SRA Bioproject accession PRJEB5096) was examined with UTRme. The corresponding miniexon sequence was obtained from Brehm et al. (Brehm et al., 2000).

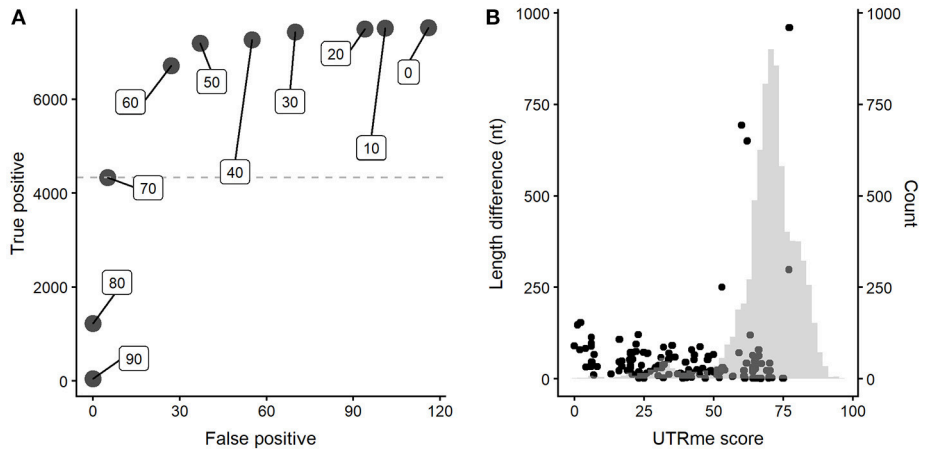


FIGURE 4 | UTRme accuracy assessment for 5′ UTRs. **(A)** Dependence of the number of true positives and false positives on the UTRme score (indicated as inserts). **(B)** False positive annotations are plotted as dots indicating their score and distance to the real processing site. The histogram shows the distribution of scores for all predicted sites.

One thousand eight hundred and ten sites in 1,369 genes were annotated with a 5′ UTR, while a polyadenylation site could be assigned for 6,841 genes presenting a total of 24,946 sites. These are expected results as SL addition is not pervasive in plathelminths as it is in trypanosomatids (Brehm et al., 2000). Analysis of the sequence of the trans splicing acceptor sites reveal a high percentage of the AG dinucleotide supporting the reliability of the annotated sites (**Supplementary Figure 4A**). A summary of UTR lengths and UTRme score distribution is shown in **Supplementary Figures 4B,C**.

Comparison With Previously Available Tools

Several groups have reported tools to identify 3′ UTRs in eukaryotes, however the algorithms consider signals not clearly present in trypanosomatids and lack the possibility of studying 5′ processing sites (Xia et al., 2014; Kim et al., 2015; Grassi et al., 2016; Ha et al., 2018). For trypanosomatids, there are reports of global identification of UTRs, but in most cases the task was performed using in-house tools (Gopal et al., 2005; Siegel et al., 2005; Kolev et al., 2010; Kelly et al., 2011; Dillon et al., 2015).

Currently, to our knowledge, the most accessible method to predict UTRs in trypanosomatid genomes is the SLaP mapper web service (Fiebig et al., 2014). To contrast UTRme results with those obtained by SLaP mapper we used 27M paired-end reads from *T. cruzi* epimastigotes (Pastro et al., 2017) (number of reads was reduced to accommodate SLaP mapper upload size limitation). In this experiment, where standard RNA-seq protocols were carried-out, UTRme was able to detect 8,448 5′ UTR regions in 5682 genes whereas SLaP mapper detected 5,343 sites in 4,061 genes. Of the genes detected by UTRme, 1878 were exclusive whereas SLaP mapper detected 257 genes exclusively. Three thousand eight hundred and four genes were detected by both software packages of which 88% had coincident predictions (**Figure 5A**). Of the 8,448 total sites identified by UTRme, 3,408 did not show matches with SLaP mapper, 71% were due to sites corresponding to genes detected exclusively by UTRme. SLaP mapper detected 536 exclusive sites, of which 56% were due to genes only detected by this software. The number of coincident sites is 5,040 for UTRme and 4805 for Slap mapper (the difference is due to the fact that a 5 pb window was implemented to define matching sites) (**Figure 5B**). The median length for the 5′ UTR regions was similar in both cases (59 and 53 bp for UTRme and SLaP mapper, respectively). While the median length for sites detected exclusively by UTRme remains around this figure (88.5), in sites detected exclusively by SLaP mapper this number increases to 786, which may be indicative of issues in these non-coincident annotations (**Supplementary Figure 5A**). For 3′ UTRs a similar situation was found (see **Supplementary Figures 5B**, **6**).

Considering the genes where both tools predicted splicing sites, a density plot shows a very good correlation (**Figures 6A,C**). Interestingly, this correlation is better for sites with high UTRme score. This is shown in **Figures 6B,D**. Here, sites were classified as coincident if their length difference was 5nt or less or non-coincident otherwise. The percentage of coincident and non-coincident sites that are above a certain score threshold is calculated and plotted. The figure shows that this percentage decreases more rapidly for non-coincident sites than for coincident sites when the UTRme score increases. This observation supports that in cases where a high score is assigned by UTRme (which suggests that the sites can be readily identified by the reads), SLaP mapper mostly reports the same site, verifying that the score is a key factor in capturing the certainty of site definition. Nonetheless, a low score in UTRme indicates that there was less evidence to support it, which in turn likely explains the decrease in correlation with SLaP mapper predictions.

We also compared the results obtained using UTRme to analyze the RNA-seq *T. brucei* data generated by Kolev, et al. in (Kolev et al., 2010) with the ones reported by the authors. These authors constructed a SL-primed library and a 3′ end-enriched library to detect 5′ and 3′ boundaries, respectively, predicting

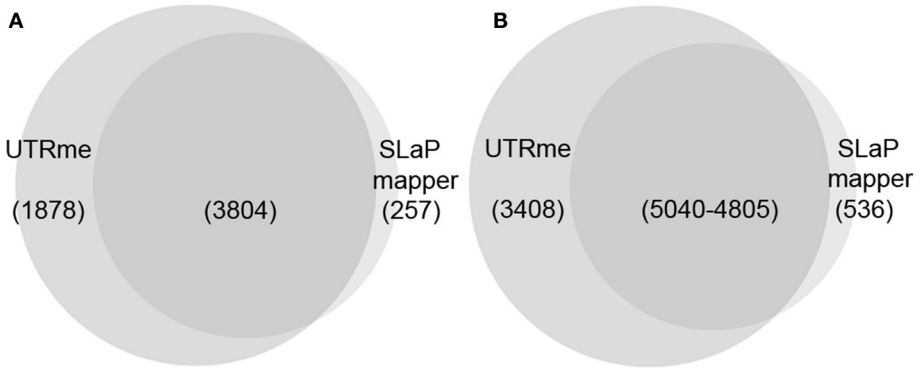


FIGURE 5 | Venn diagrams comparing the results of UTRme and SLaP mapper 5′ processing sites annotations. **(A)** The intersection of the genes predicted by each tool is shown. **(B)** For genes were annotations are available for both tools, the intersection of the sites predicted by each tool is shown.

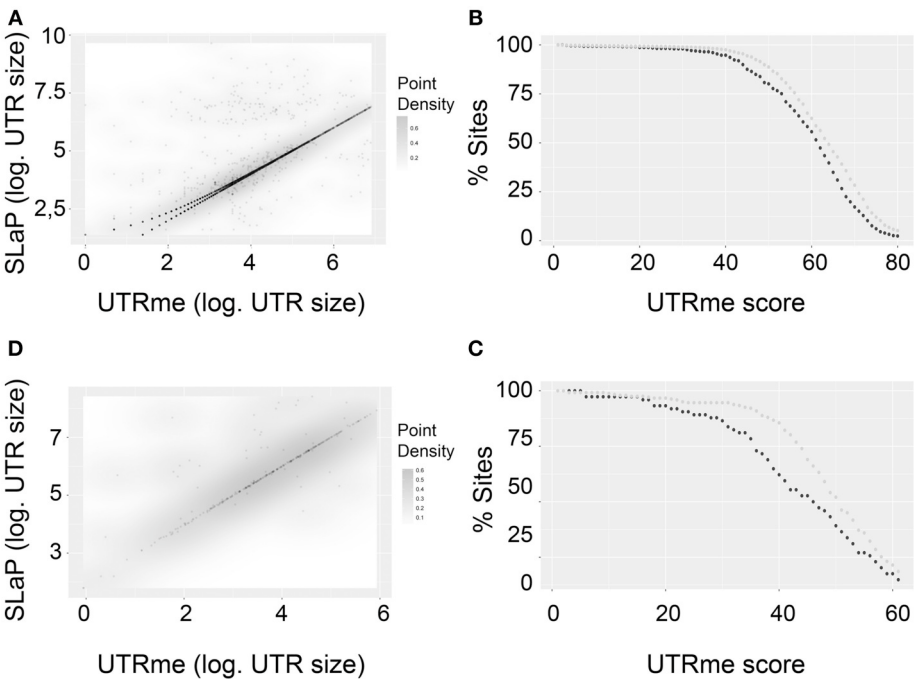


FIGURE 6 | Comparison of UTRme best scoring sites with the ones predicted by Slap mapper using Pastro et al. (2017) data. **(A)** Scatter plot of 5′ UTR lengths. Darker regions indicate higher density of points. **(B)** The percentage of points that have scores above a threshold is plotted for coincident and non-coincident sites. *Dark gray:* non-coincident sites. *Light gray:* coincident sites. The percentage was calculated until the number of sites remaining is above 10 **(C,D).** Same as **(A,B)** for 3′ UTRs.

processing sites by using an in-house pipeline. The results obtained for the comparison where similar to the ones observed for SLaP mapper (**Supplementary Figures 7**, **8**).

Interestingly, for both comparisons UTRme was able to predict a higher number of sites. This is possibly due to the inclusion by UTRme of predictions that are discarded by other tools but that UTRme does include by penalizing them with a low score. The good correlation between the results obtained through the two tools and the influence of the UTRme score on the percentage of agreement is clearly shown in both cases.

Globally, the comparison of UTRme with available data and applications supports the software accuracy and highlights the importance and usefulness of the UTRme scores.

FINAL REMARKS

Post-transcriptional mechanisms are recognized as important regulatory steps in eukaryotes. Post-transcriptional mRNA regulators most commonly bind to sequences present in UTR

regions, so their definition is critical to better understand regulatory networks. For trypansomatids, UTR delimiting algorithms are confounded by the presence of the A tracts in intergenic regions (Duhagon et al., 2011) and by the repetitive nature of the sequences that cause issues in the genomic assembly, among other reasons. This lead us to develop UTRme, a tool that allows not only the identification of processing sites from RNA-seq data but also reports their associated confidence. UTRme is easy to install in linux based systems, is provided with a GUI making it user friendly and it does not require previous expertise on RNA-seq data analysis, something we expect that will make the tool more readily available for wet lab biologists.

As shown by the excellent correlation with sites experimentally determined and considering the results obtained for the simulated RNA-seq data, we can conclude that UTRme predicts sites with excellent precision and that the scoring system is capable of reflecting the certainty of the annotations. The comparison with other tools allowed us to further support the advantage and usefulness of the UTRme scoring system which discriminates between sites that are clearly predicted, from those where evidence is less clear.

Finally, UTRme can be applied to predict 3′ processing sites not only in trypanosomatids but any eukaryotes and can be used for 5′ end determination in other organisms where trans splicing occurs.

DATA AVAILABILITY STATEMENT

The dataset generated for this study can be found in the SRA repository BioProject PRJNA473354.

AUTHOR CONTRIBUTIONS

SR and PS UTRme software development and design of the methodology. SR performed the analysis. RF and PS performed the 5′ end enriched RNAseq experiment. JS-S, BG, RF, SR, and PS wrote and reviewed the manuscript. JS-S, BG, and PS acquisition of financial support. PS coordinated the project.

FUNDING

This project was supported by ANII, FCE_3_2016_1_126317; CSIC, I+D research groups program 108725. SR and RF received scholarships from ANII. SR and PS received financial support from PEDECIBA.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2018.00671/full#supplementary-material

REFERENCES

Bartholomeu, D. C., Silva, R. A., Galvao, L. M., el-Sayed, N. M., Donelson, J. E., and Teixeira, S. M. (2002). *Trypanosoma cruzi*: RNA structure and post-transcriptional control of tubulin gene expression. *Exp. Parasitol.* 102, 123–133. doi: 10.1016/S0014-4894(03)00034-1

Bhatia, V., Sinha, M., Luxon, B., and Garg, N. (2004). Utility of the *Trypanosoma cruzi* sequence database for identification of potential vaccine candidates by in silico and in vitro screening. *Infect. Immun.* 72, 6245–6254. doi: 10.1128/IAI.72.11.6245-6254.2004

Bontempi, E. J., Porcel, B. M., Henriksson, J., Carlsson, L., Rydaker, M., Segura, E. L., et al. (1994). Genes for histone H3 in *Trypanosoma cruzi*. *Mol. Biochem. Parasitol.* 66, 147–151.

Brehm, K., Jensen, K., and Frosch, M. (2000). mRNA trans-splicing in the human parasitic cestode *Echinococcus multilocularis*. *J. Biol. Chem.* 275, 38311–38318. doi: 10.1074/jbc.M006091200

Búa, J., Aslund, L., Pereyra, N., Garcia, G. A., Bontempi, E. J., and Ruiz, A. M. (2001). Characterisation of a cyclophilin isoform in *Trypanosoma cruzi*. *FEMS Microbiol. Lett.* 200, 43–47. doi: 10.1111/j.1574-6968.2001.tb10690.x

Clayton, C. (2013). The regulation of trypanosome gene expression by RNA-binding proteins. *PLoS Pathog.* 9:e1003680. doi: 10.1371/journal.ppat.1003680

Coelho, E. R., Rodrigues Dde, C., Urmenyi, T. P., Rondinelli, E., and Silva, R. (2006). Polymorphic and differential expression of the *Trypanosoma cruzi* alleles containing universal minicircle binding protein. *Biochem. Biophys. Res. Commun.* 341, 382–390. doi: 10.1016/j.bbrc.2005.12.189

De Gaudenzi, J. G., Carmona, S. J., Aguero, F., and Frasch, A. C. (2013). Genome-wide analysis of 3'-untranslated regions supports the existence of post-transcriptional regulons controlling gene expression in trypanosomes. *PeerJ* 1:e118. doi: 10.7717/peerj.118

Di Noia, J. M., D'Orso, I., Aslund, L., Sanchez, D. O., and Frasch, A. C. (1998). The *Trypanosoma cruzi* mucin family is transcribed from hundreds of genes having hypervariable regions. *J. Biol. Chem.* 273, 10843–10850.

Di Noia, J. M., D'Orso, I., Sanchez, D. O., and Frasch, A. C. (2000). AU-rich elements in the 3'-untranslated region of a new mucin-type gene family of *Trypanosoma cruzi* confers mRNA instability and modulates translation efficiency. *J. Biol. Chem.* 275, 10218–10227. doi: 10.1074/jbc.275.14.10218

Dillon, L. A., Okrah, K., Hughitt, V. K., Suresh, R., Li, Y., Fernandes, M. C., et al. (2015). Transcriptomic profiling of gene expression and RNA processing during *Leishmania major* differentiation. *Nucleic Acids Res.* 43, 6799–6813. doi: 10.1093/nar/gkv656

D'Orso, I., and Frasch, A. C. (2001). TcUBP-1, a developmentally regulated U-rich RNA-binding protein involved in selective mRNA destabilization in trypanosomes. *J. Biol. Chem.* 276, 34801–34809. doi: 10.1074/jbc.M102120200

Duhagon, M. A., Smircich, P., Forteza, D., Naya, H., Williams, N., and Garat, B. (2011). Comparative genomic analysis of dinucleotide repeats in Tritryps. *Gene* 487, 29–37. doi: 10.1016/j.gene.2011.07.022

Ekanayake, D., and Sabatini, R. (2011). Epigenetic regulation of polymerase II transcription initiation in *Trypanosoma cruzi*: modulation of nucleosome abundance, histone modification, and polymerase occupancy by O-linked thymine DNA glucosylation. *Eukaryot. Cell* 10, 1465–1472. doi: 10.1128/EC.05185-11

Fadda, A., Ryten, M., Droll, D., Rojas, F., Farber, V., Haanstra, J. R., et al. (2014). Transcriptome-wide analysis of trypanosome mRNA decay reveals complex degradation kinetics and suggests a role for co-transcriptional degradation in determining mRNA levels. *Mol. Microbiol.* 94, 307–326. doi: 10.1111/mmi.12764

Fiebig, M., Gluenz, E., Carrington, M., and Kelly, S. (2014). SLaP mapper: A webserver for identifying and quantifying spliced-leader addition and polyadenylation site usage in kinetoplastid genomes. *Mol. Biochem. Parasitol.* 196, 71–74. doi: 10.1016/j.molbiopara.2014.07.012

Franks, A., Airoldi, E., and Slavov, N. (2017). Post-transcriptional regulation across human tissues. *PLoS Comput. Biol.* 13:e1005535. doi: 10.1371/journal.pcbi.1005535

García, E. A., Ziliani, M., Aguero, F., Bernabo, G., Sanchez, D. O., and Tekiel, V. (2010). TcTASV: a novel protein family in trypanosoma cruzi identified from a subtractive trypomastigote cDNA library. *PLoS Negl. Trop. Dis.* 4:e841. doi: 10.1371/journal.pntd.0000841

Gopal, S., Awadalla, S., Gaasterland, T., and Cross, G. A. (2005). A computational investigation of kinetoplastid trans-splicing. *Genome Biol.* 6:R95. doi: 10.1186/gb-2005-6-11-r95

Grassi, E., Mariella, E., Lembo, A., Molineris, I., and Provero, P. (2016). Roar: detecting alternative polyadenylation with standard mRNA sequencing libraries. *BMC Bioinformatics* 17:423. doi: 10.1186/s12859-016-1254-8

Ha, K. C. H., Blencowe, B. J., and Morris, Q. (2018). QAPA: a new method for the systematic analysis of alternative polyadenylation from RNA-seq data. *Genome Biol.* 19:45. doi: 10.1186/s13059-018-1414-4

He, M. X., Petoukhov, S. V., and Ricci, P. E. (2004). Genetic code, hamming distance and stochastic matrices. *Bull. Math. Biol.* 66, 1405–1421. doi: 10.1016/j.bulm.2004.01.002

Jensen, B. C., Ramasamy, G., Vasconcelos, E. J., Ingolia, N. T., Myler, P. J., and Parsons, M. (2014). Extensive stage-regulation of translation revealed by ribosome profiling of *Trypanosoma brucei*. *BMC Genomics* 15:911. doi: 10.1186/1471-2164-15-911

Kelly, S., Wickstead, B., Maini, P. K., and Gull, K. (2011). Ab initio identification of novel regulatory elements in the genome of *Trypanosoma brucei* by Bayesian inference on sequence segmentation. *PLoS ONE* 6:e25666. doi: 10.1371/journal.pone.0025666

Kim, M., You, B. H., and Nam, J. W. (2015). Global estimation of the 3' untranslated region landscape using RNA sequencing. *Methods* 83, 111–117. doi: 10.1016/j.ymeth.2015.04.011

Kolev, N. G., Franklin, J. B., Carmi, S., Shi, H., Michaeli, S., and Tschudi, C. (2010). The transcriptome of the human pathogen *Trypanosoma brucei* at single-nucleotide resolution. *PLoS Pathog.* 6:e1001090. doi: 10.1371/journal.ppat.1001090

Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9, 357–359. doi: 10.1038/nmeth.1923

Lei, Q., Li, C., Zuo, Z., Huang, C., Cheng, H., and Zhou, R. (2016). Evolutionary insights into RNA trans-splicing in vertebrates. *Genome Biol. Evol.* 8, 562–577. doi: 10.1093/gbe/evw025

Levenshtein, A. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Sov. Phys. Doklady* 10, 707–710.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352

Li, Y., Shah-Simpson, S., Okrah, K., Belew, A. T., Choi, J., Caradonna, K. L., et al. (2016). Transcriptome remodeling in *Trypanosoma cruzi* and human cells during intracellular infection. *PLoS Pathog.* 12:e1005511. doi: 10.1371/journal.ppat.1005511

Majorek, K. A., Dunin-Horkawicz, S., Steczkiewicz, K., Muszewska, A., Nowotny, M., Ginalski, K., et al. (2014). The RNase H-like superfamily: new members, comparative structural analysis and evolutionary classification. *Nucleic Acids Res.* 42, 4160–4179. doi: 10.1093/nar/gkt1414

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* 17, 10–12. doi: 10.14806/ej.17.1.200

Pastro, L., Smircich, P., Di Paolo, A., Becco, L., Duhagon, M. A., Sotelo-Silveira, J., et al. (2017). Nuclear compartmentalization contributes to stage-specific gene expression control in *Trypanosoma cruzi*. *Front. Cell Dev. Biol.* 5:8. doi: 10.3389/fcell.2017.00008

Pastro, L., Smircich, P., Perez-Diaz, L., Duhagon, M. A., and Garat, B. (2013). Implication of CA repeated tracts on post-transcriptional regulation in *Trypanosoma cruzi*. *Exp. Parasitol.* 134, 511–518. doi: 10.1016/j.exppara.2013.04.004

Prüss-Ustün, A., Wolf, J., Corvalán, C., Bos, R., and Neira, M. (2016). *Preventing Disease Through Healthy Environments: A Global Assessment of the Burden of Disease from Environmental Risks*. WHO.

Quinlan, A. R., and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841–842. doi: 10.1093/bioinformatics/btq033

Ramos, T. C., Nunes, V. S., Nardelli, S. C., dos Santos Pascoalino, B., Moretti, N. S., Rocha, A. A., et al. (2015). Expression of non-acetylatable lysines 10 and 14 of histone H4 impairs transcription and replication in *Trypanosoma cruzi*. *Mol. Biochem. Parasitol.* 204, 1–10. doi: 10.1016/j.molbiopara.2015.11.001

Respuela, P., Ferella, M., Rada-Iglesias, A., and Aslund, L. (2008). Histone acetylation and methylation at sites initiating divergent polycistronic transcription in *Trypanosoma cruzi*. *J. Biol. Chem.* 283, 15884–15892. doi: 10.1074/jbc.M802081200

Siegel, T. N., Hekstra, D. R., Kemp, L. E., Figueiredo, L. M., Lowell, J. E., Fenyo, D., et al. (2009). Four histone variants mark the boundaries of polycistronic transcription units in *Trypanosoma brucei*. *Genes Dev.* 23, 1063–1076. doi: 10.1101/gad.1790409

Siegel, T. N., Tan, K. S., and Cross, G. A. (2005). Systematic study of sequence motifs for RNA trans splicing in *Trypanosoma brucei*. *Mol. Cell Biol.* 25, 9586–9594. doi: 10.1128/MCB.25.21.9586-9594.2005

Smircich, P., Eastman, G., Bispo, S., Duhagon, M. A., Guerra-Slompo, E. P., Garat, B., et al. (2015). Ribosome profiling reveals translation control as a key mechanism generating differential gene expression in *Trypanosoma cruzi*. *BMC Genomics* 16:443. doi: 10.1186/s12864-015-1563-8

Smircich, P., Forteza, D., El-Sayed, N. M., and Garat, B. (2013). Genomic analysis of sequence-dependent DNA curvature in leishmania. *PLoS ONE* 8:e63068. doi: 10.1371/journal.pone.0063068

Teixeira, S. M., Kirchhoff, L. V., and Donelson, J. E. (1999). *Trypanosoma cruzi*: suppression of tuzin gene expression by its 5'-UTR and spliced leader addition site. *Exp. Parasitol.* 93, 143–151. doi: 10.1006/expr.1999.4446

Thomas, S., Green, A., Sturm, N. R., Campbell, D. A., and Myler, P. J. (2009). Histone acetylations mark origins of polycistronic transcription in *Leishmania major*. *BMC Genomics* 10, 152. doi: 10.1186/1471-2164-10-152

Vandersall-Nairn, A. S., Merkle, R. K., O'Brien, K., Oeltmann, T. N., and Moremen, K. W. (1998). Cloning, expression, purification, and characterization of the acid alpha-mannosidase from *Trypanosoma cruzi*. *Glycobiology* 8, 1183–1194.

Vasquez, J. J., Hon, C. C., Vanselow, J. T., Schlosser, A., and Siegel, T. N. (2014). Comparative ribosome profiling reveals extensive translational complexity in different *Trypanosoma brucei* life cycle stages. *Nucleic Acids Res.* 42, 3623–3637. doi: 10.1093/nar/gkt1386

Wright, J. R., Siegel, T. N., and Cross, G. A. (2010). Histone H3 trimethylated at lysine 4 is enriched at probable transcription start sites in *Trypanosoma brucei*. *Mol. Biochem. Parasitol.* 172, 141–144. doi: 10.1016/j.molbiopara.2010.03.013

Xia, Z., Donehower, L. A., Cooper, T. A., Neilson, J. R., Wheeler, D. A., Wagner, E. J., et al. (2014). Dynamic analyses of alternative polyadenylation from RNA-seq reveal a 3'-UTR landscape across seven tumour types. *Nat. Commun.* 5:5274. doi: 10.1038/ncomms6274

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

3.4. Determinación de la regulación mediada por la presencia de uORFs en las regiones 5’ UTRs

La eficiencia traduccional es una de las métricas que mejor determina la abundancia proteica y se regula principalmente a nivel de la formación del complejo de iniciación de la traducción y del paso subsiguiente de elongación. Uno de los mecanismos específicos capaz de regular este proceso son los marcos abiertos de lectura 5’ río arriba del CDS principal (uORF).

El reanálisis de los datos de *Ribosome Profiling* sumado al desarrollo de UTRme establece condiciones óptimas para el estudio de uORF y su rol regulatorio en *T. cruzi* (objetivo específico 4).

3.4.1. Obtención y clasificación de los marcos de lectura

Para determinar el rol regulatorio de los uORF, estos deben ser identificados y caracterizados. Dada su localización en 5’ UTR es necesario caracterizar estas regiones de forma precisa. La definición de las secuencias no traducidas fue realizada mediante el programa UTRme, desarrollado en el capítulo anterior y utilizando los datos de (Li et al., 2016) (ver sección 3.4.14 Estrategia). Luego, se utilizó el programa getORF (Rice et al., 2000), para obtener los marcos abiertos de lectura en las regiones 5’ UTR caracterizadas. Los marcos de lectura en las regiones 5’ UTR se definieron como cualquier marco mayor a tres aminoácidos que empiece en un codón de parada y termine en otro codón de parada (tomando a los inicios y finales del 5’ UTR como tales) en la misma hebra que el gen codificante. Los marcos obtenidos se clasificaron en tres niveles mediante un programa de elaboración local (ver material suplementario).

El primero depende de la presencia de codones de inicio, el segundo depende de los delimitadores del marco (inicio y fin de la región 5’UTR y codones de parada) y el tercero depende de la ubicación del marco dentro del 5’ UTR del transcripto.

3.4.2. Primer nivel de clasificación

Según la definición de los marcos de lectura expuesta anteriormente, cada uno tiene el potencial de contener múltiples codones de inicio. Por esta razón, y como ya se ha hecho con anterioridad (Chew et al., 2016)., se decidió dividir cada uno de los marcos en sub-marcos. Así cada marco de lectura predicho puede generar n sub-marcos, siendo n el número de codones de inicio presentes.

De esta manera pudimos dividir a cada marco resultante en dos categorías:

- *Full*: en donde la longitud de la secuencia del marco de lectura original se mantiene, porque no se encontraron codones de inicio dentro del marco o porque el codón de inicio es el primer triplete de este.
- *Short*: en donde la longitud de la secuencia del marco resultante es menor que la del marco original.

3.4.3. Segundo nivel de clasificación

Los marcos de lectura pueden quedar delimitados por los siguientes elementos: codón de inicio, codón de parada, inicio o fin de la región 5’ UTR. De acuerdo a la combinación de estos delimitadores los marcos se clasificaron en una de las siguientes cuatro categorías:

- *begin2end*: indica que el comienzo del marco de lectura esta dado por el comienzo de la región 5' del ARNm y que el final del marco esta dado por el comienzo de la región codificante principal.
- *begin2stop*: indica que el comienzo del marco de lectura esta dado por el comienzo de la región 5' del ARNm y que el final del marco esta dado por un codón de parada.
- *stop2stop*: indica que el comienzo del marco de lectura esta dado por un codón de parada, mientras que el final del marco esta dado por otro codón de parada distinto.

- *stop2end*: indica que el comienzo del marco de lectura esta dado por un codón de parada y que el final del marco esta dado por el comienzo de la región codificante principal.

En la Figura 3.25 podemos ver gráficamente la clasificación descrita. La región 5’ UTR representada posee tres codones de parada y cinco codones de inicio. Dada la distribución de estos elementos se establecen cuatro marcos. El primer marco (*frame-1*) queda delimitado por el inicio del 5’ UTR y el primer codón de parada, por ende, a este marco de lo puede clasificar como *full-begin2stop*. El segundo marco (*frame-2*) queda configurado por el primer y segundo codón de parada y se clasifica como *full-stop2stop*. El tercer marco también está delimitado por dos codones de parada, sin embargo, a diferencia del anterior, presenta dos codones de inicio. En estas situaciones el marco original se divide en el número de codones de inicio presentes, por lo que se generan en este caso, dos sub-marcos (*frame-3.1* y *frame-3.2*), ambos del tipo *short-stop2stop*. Finalmente, el cuatro marco (*frame-4*) queda delimitado por un codón de parada y el final de la región 5’ UTR (marcada por el codón de inicio del CDS principal), presentando dos codones de inicios en esta región. Ambos codones de inicio definen marcos (*frame-4.1* y *frame-4.2*) del tipo *short-stop2end*.

3.4.4. Tercer nivel de clasificación

El último nivel de clasificación encasilla a los marcos en tres categorías dependientes de la ubicación de los delimitadores dentro de la región 5’ UTR:

- uORF: son los marcos que están enteramente en la región 5' no traducida del ARNm. Pueden estar o no en el mismo marco que el CDS principal.
- *Overlap*: son los marcos que inician en la región 5' UTR y finalizan fuera de ella. Presentan entonces solapamiento con el inicio del CDS principal. El marco evaluado se encuentra en una fase distinta al AUG iniciador del CDS.
- *Extended*: son los marcos que inician en la región 5' UTR y finalizan en el codón de parada del CDS principal. Por ende, el codón de inicio está en fase con el AUG iniciador del CDS principal.

Utilizaremos la Figura 3.25 para ejemplificar lo anterior. En este caso el primer marco (frame-1) está enteramente definido en la región 5' UTR por lo que su clasificación es uORF. De forma similar se define a los marcos 2, 3.1 y 3.2. El marco 4 subtipo 1 (*frame-4.1*) presenta un codón de inicio en la región 5' UTR mientras que su codón de parada es el codón de parada del CDS principal, por lo tanto este marco se lo define como *Extended* del tipo *short-stop2end*. Finalmente, el marco 4 subtipo 2 (*frame-4.2*) está delimitado por un codón de inicio no en fase con el CDS principal y por un codón de parada que queda fuera de la región 5' UTR. Este marco queda definido entonces como *Overlap* del tipo *short-stop2end*. La diferentes clasificaciones asignadas a cada marco se resumen en la Tabla 3.9. Una tabla con la clasificación de los marcos de lectura (tomando como codón de inicio AUG) presentes en las regiones 5' UTR de epimastigotas y tripomastigotas se dispone en materiales suplementarios.

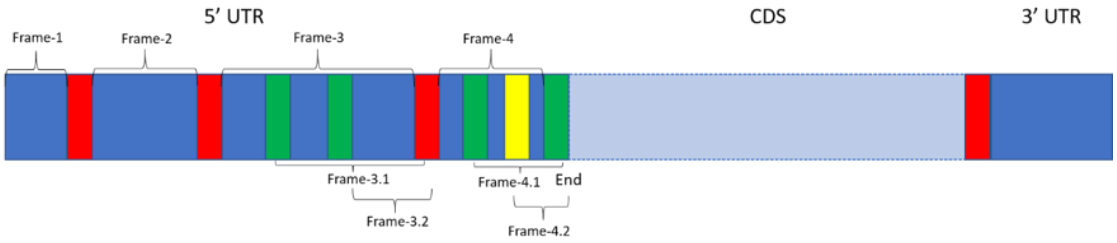


Figura 3.25. Representación de una región 5' UTR donde queda esquematizado las posibles clasificaciones que pueden adoptar los marcos de lecturas resultantes de la utilización del programa getORF. Los codones de inicio y fin quedan representados por barras verdes y rojas respectivamente. Mientras que la barra amarilla indica un codón de inicio que no está en fase con el codón de inicio del CDS principal.

Tabla 3.9. Clasificaciones asignadas a los marcos de lectura representados en Figura 3.25.

Marco	Nivel 1	Nivel 2	Nivel 3
1	*Full*	*begin2stop*	uORF
2	*Full*	*stop2stop*	uORF
3.1	*Short*	*stop2stop*	uORF
3.2	*Short*	*stop2stop*	uORF
4.1	*Short*	*stop2end*	*Overlap*
4.2	*Short*	*stop2end*	*Extended*

3.4.5. Determinación del potencial represivo de las regiones 5' UTRs

No todos los marcos de lecturas río arriba del CDS principal tienen potencial codificante, ni tampoco presentan la misma eficiencia de iniciación. Como se detalló en la

introducción (capítulo 1.7), existen una multitud de propiedades y características que en su conjunto determinan la posibilidad de que un marco dado sea represor. Si bien existen casos en donde su presencia parece influir positivamente en la TE del CDS principal, estos son minoritarios (Griffin et al., 2001; Vattem and Wek, 2004; Chen et al., 2010). Si una región 5' UTR contiene marcos con potencial represivo, nos planteamos que el gen asociado estará sometido a control traduccional mediado por uORF.

Partiendo de la clasificación descrita anteriormente, evaluaremos el potencial represor de todas las regiones 5' UTR. Las regiones que contengan marcos del tipo *Overlap* que inicien con un codón iniciador AUG serán considerados como represivas, de acuerdo a los observado en (Wethmar, 2014). Por otra parte, aquellas regiones que contengan marcos uORF podrán o no ser represoras dependiendo de las características del marco. De acuerdo a los conocimientos actuales que existen sobre uORF y su capacidad de disminuir la eficiencia traduccional establecimos criterios para definir su potencial represivo. Para considerar a un uORF como represor definimos necesaria la presencia de un codón de inicio del tipo AUG, dentro del marco definido. Esta decisión está fundamentada en los estudios recientes que demuestran que los codones iniciadores con buena eficiencia de iniciación son en una abrumadora mayoría AUGs (Clements et al., 1988). Seguidamente, establecimos límites en cuanto al posicionamiento del uORF dentro del 5' UTR. En el extremo 5' establecimos una distancia mínima de 15 nucleótidos (distancia al inicio del 5' UTR) para considerar a un uORF como represivo. Esta distancia le brinda un espacio a la maquinaria traduccional para poder acoplarse correctamente (Vilela and McCarthy, 2003). Por otra parte, limitando la distancia entre el codón de parada del uORF y el AUG del CDS principal (distancia al codón de inicio), limitamos el tiempo que tiene la maquinaria traduccional para captar un nuevo ARNt iniciador, además de otros factores de inicio de traducción (Fervers et al., 2018). Cuanto menor sea el tiempo, mayor será su potencial represor. En este caso, se estableció el límite en un máximo de 50 nucleótidos siguiendo lo hecho en (Chew et al., 2016).

Finalmente, consideramos que el largo mínimo de un uORF para tener potencial represor es de 5 aminoácidos. Previamente se ha vinculado la capacidad de reiniciar la traducción con el largo del uORF, determinando que cuanto mayor sea el largo menor será la

probabilidad de reinicio (Kozak, 2001; Rajkowitsch et al., 2004). Los valores seleccionados se resumen en la Tabla 3.10

Tabla 3.10. Condicionantes para asignar un potencial represor a un marco del tipo uORF .

	Valor
Codón Iniciador	AUG
Distancia -5'	>= 15 (nt)
Distancia - 3'	<= 50 (nt)
Largo uORF	>= 5 (aa)

Por último, los marcos de la categoría *Extended*, los cuales extienden el CDS principal en el extremo N-terminal, tiene la potencialidad de cambiar la localización subcelular de la proteína u otros efectos a nivel post traduccional, pero difícilmente afecten la eficiencia traduccional del gen asociado. Decidimos ignorar aquellas regiones UTRs que tuvieran este tipo de marco, dado que no sabemos con certeza si el codón de inicio funcional es el anotado o es el extendido.

Dado lo anterior, aquellas regiones 5' UTR que presenten marcos con potencial represivos serán clasificadas como represivas. A su vez, las 5' UTRs que no contengan marcos represivos, presenten un tamaño mayor a 50 nucleótidos y no contengan AUGs serán definidas como no represivas. Estas características nos permiten definir dos grupos bien distintos en cuanto al potencial represor. A las regiones que no integran ninguna de las dos categorías no se les asignó clasificación. Mediante la utilización de un *script* de elaboración propia se clasificaron (material suplementario), de acuerdo a lo descrito anterior, las regiones 5' UTRs de los genes expresados en los estadios epimastigota y tripomastigota.

Para el estadio epimastigota se analizaron en total las regiones 5' UTRs de 6744 ARNm. De acuerdo a los establecido anteriormente, se clasificaron 568 (8.4%) regiones 5' UTRs como represivas de las cuales, 111 fueron presentaron marcos *Overlap* y 160 *uORFs* represivos, el restante (297) presentaron marcos de ambas categorías. A su vez, 3375 (50%) regiones 5' UTR fueron clasificadas como no represivas. Finalmente, 52 5' UTRs tienen marcos de la categoría *Extended* (~1%) (Tabla 3.11). En el estadio tripomastigota, se analizaron 6750 ARNm se observándose una distribución similar. 602 regiones 5' UTR (~9%) fueron clasificadas como represivas, de las cuales 231 presentaron marcos de

lecturas de las categorías uORF represivo y *Overlap*, mientras que 204 presentaron únicamente *Overlap* y 167 únicamente uORF. 3333 (~50%) regiones 5' UTR fueron clasificadas como no represivas. Finalmente, 53 5' UTRs tienen marcos de la categoría *Extended* (~1%) (Tabla 3.11).

Tabla 3.11. Categorización de las regiones 5' UTR de los estadios de los genes expresados en los estadios epimastigota y tripomastigota.

Estadio	Represivo (uORF/Overlap)	No Represivo	No clasificado
Epimastigota	568	3375	2801
Tripomastigota	602	3333	2815

El porcentaje de genes cuya 5' UTR tiene potencial represivo es similar al detectado en (Jensen et al., 2014) (11%). Al igual que lo discutido en este artículo, creemos que la razón de la diferencia observada con respecto a otros trabajos, es debida al refinamiento en la caracterización de las regiones 5' UTR (Vasquez et al., 2014; Fervers et al., 2018; Radio et al., 2018).

3.4.6. Regulación de los niveles de eficiencia traduccional de genes mediada por marcos de lectura presentes en la región 5' UTR

Un marco de lectura con potencial represivo en la región 5' UTR de un gen debería disminuir la eficiencia traduccional del mismo. Con el fin de evaluar si efectivamente la presencia de este tipo de marcos disminuye la TE, utilizamos los datos de TE calculados en el capítulo 3.1.

Para evaluar de forma independiente el efecto represivo de los marcos de lectura, se obtuvieron los valores de TE de los genes cuyas regiones 5' UTR fueran únicamente uORF represivos (a partir de ahora represivos) u *Overlap* (definidos desde aquí como *Overlaps* represivos). A su vez, se obtuvieron los valores de TE para la categoría no represivos y para la totalidad de los genes, independientemente de la categoría (Total) (Tabla 3.12).

Tabla 3.12. Números de transcriptos analizados en cada categoría para los genes expresados en los estadios epimastigota y tripomastigota metacíclico.

Estadio	Overlap	Represivo (uORF)	No Represivo	Total
Epimastigota	111	160	3375	6744
Tripomastigota	204	167	3333	6750

En el estadio epimastigota, los resultados de la comparación de los niveles de TE de las cuatro categorías evaluadas permiten evidenciar como la TE es dependiente de la presencia de marcos de lectura con potencial represivo y de la categoría a la que estos pertenezcan (Figura 3.26). La aplicación del test no-paramétrico U de Mann-Whitney, el cual ya fue previamente aplicado para comparar diferencias de eficiencias traduccionales por Jensen et al. (Jensen et al., 2014), determinó diferencias estadísticamente significativas (<0.01) en todas las comparaciones efectuadas (Tabla 6.3).

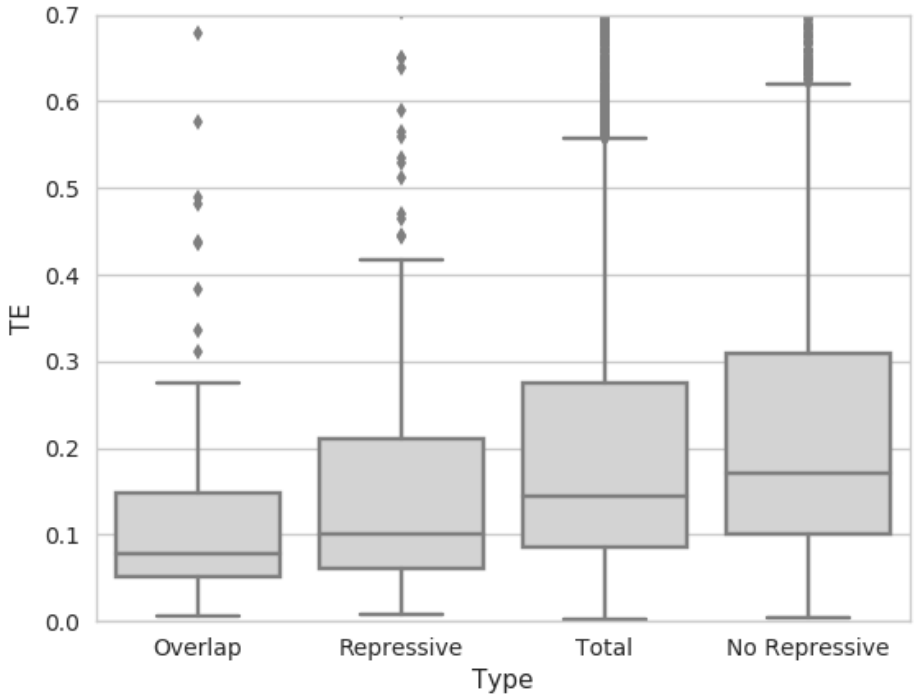


Figura 3.26. Comparación de eficiencia traduccional en el estadio epimastigota de genes con 5' UTR clasificadas como *Overlap*, represivas, no represivas y contra la totalidad de los genes independientemente de la categoría a la que pertenezcan. Los resultados obtenidos son estadísticamente significativos ($p < 0.01$) (Tabla 6.3).

Interesantemente, los genes asociados a la categoría *Overlap* presentan una menor TE que el resto. Si la maquinaria traduccional logra acoplarse correctamente al uAUG solapante, se evitará un correcto posicionamiento en el AUG iniciador del CDS, estableciendo un importante mecanismo de control traduccional. Existen muy pocos casos donde se ha visto una correlación positiva entre solapamiento y reducción de la eficiencia traduccional en uORF de forma natural (Wethmar, 2014). Sin embargo, recientemente Frevers et al. observó en *T. congolense* que a medida que la distancia

entre el uAUG y el AUG del CDS principal disminuía la eficiencia traduccional lo hacía también, llegando a la disminución máxima cuando ocurría el solapamiento (Fervers et al., 2018). La categoría que le sigue con respecto al nivel de TE (de menor a mayor) es la perteneciente a las regiones 5' UTRs contenedoras de marcos represivos del tipo uORFs. La represión de la eficiencia traduccional mediada por uORF (ubicados completamente dentro de la región 5' UTR) es de menor eficacia que la categoría anterior, sin embargo, se observa un claro efecto represivo, particularmente al comparar con los transcriptos con 5' UTRs categorizadas como no represivos. La TE de los genes pertenecientes a los integrantes del grupo no represivos fue la mayor en toda la comparación, sugiriendo que estos no presentan control traduccional mediado por uORF. Considerando todas las regiones 5' UTR se pueden observar que presentan un valor medio entre las categorías represivas y no represivas (Figura 3.26).

En el estadio tripomastigota metacíclico se evidencia una situación similar (Figura 3.27), aunque menos clara (Tabla 6.3). La disminución del efecto observado probablemente se deba a que en el estadio tripomastigota metacíclico ocurre una disminución traduccional para la mayoría de los genes, por lo que deben existir otros mecanismos que ejercen un efecto más significativo que los uORFs para regular la traducción (Smircich et al., 2015).

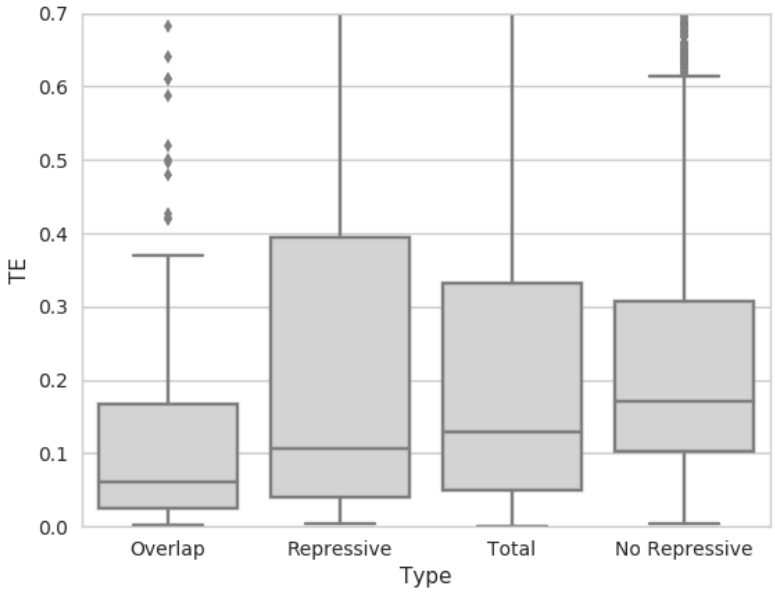


Figura 3.27. Comparación de eficiencia traduccional en el estadio tripomastigota metacíclico de genes con 5' UTR clasificadas como *Overlap*, represivas, no represivas y contra la totalidad de los genes independientemente de la categoría a la que pertenezcan. Los resultados obtenidos son estadísticamente significativos ($p < 0.01$) (Tabla 6.3).

Los resultados coinciden con los obtenidos previamente en otros tripanosomátidos, lo que sugiere que la presencia de marcos de lectura con potencial regulador en las regiones 5' UTR es un factor que contribuye al control traduccional en estos organismos (Siegel et al., 2005; Jensen et al., 2014; Fervers et al., 2018). A su vez, se observa que el efecto que ejerce el solapamiento no en fase de un marco abierto de lectura, es el que logra el mayor nivel de represión. Finalmente, podemos determinar que las características seleccionadas representan buenos indicadores para evaluar el potencial represivo de los uORFs, ya sea para las categorías represivas como para las no represivas.

3.4.7. Evaluación de la importancia del codón iniciador AUG para los marcos de lectura del tipo uORF represivos.

Debido a que la literatura con respecto al uso de codones iniciadores en uORFs es contradictoria (Clements et al., 1988; Ingolia et al., 2011; Fritsch et al., 2012; Lee et al., 2012), decidimos evaluar si la presencia de codones uAUG es determinante para el efecto represor sobre la eficiencia traduccional. Para eso, realizamos una nueva determinación de uORF represivos cambiando únicamente el codón seleccionado como iniciador. Mediante un *script* de elaboración propia (material suplementario), se determinaron los marcos de lectura uORF (represivo) para cada uno de los 61 posible codones en ambos estadios. De esta manera, pretendemos evaluar si la presencia de codones uAUG es necesaria para que un uORF sea represor, o si el resto de las propiedades establecidas ya son suficientes. A su vez, nos permite determinar si existen codones no-uAUG que generen uORFs represivos por ser buenos iniciadores de la traducción. Para esto, la eficiencia traduccional de cada grupo fue comparada contra la eficiencia traduccional de los genes con 5' UTR con uORF represivos de uAUG mediante el test no paramétrico U de Mann-Whitney.

En el estadio epimastigota, la eficiencia traduccional de los genes con 5' UTR con uORF represivos de AUG es significativamente menor con respecto al resto de los codones (Figura 3.28). El resultado es concordante con lo observado en (Clements et al., 1988) donde determina que los uORF con codones no-AUG no presentan buena eficiencia traduccional. Tampoco se evidencia un enriquecimiento particular de codones

iniciadores que varían en un base con los AUG (AGG, ACG y AAG), como previamente se ha reportado (Ingolia et al., 2011; Lee et al., 2012). Los resultados observados para el estadio tripomastigota metacíclico coinciden con lo observado en epimastigota (Figura 6.4). La comparación de la TE asociada a AUG con respecto al resto de los codones dio estadísticamente significativa en todas las comparaciones analizadas en ambos estadios (material suplementario).

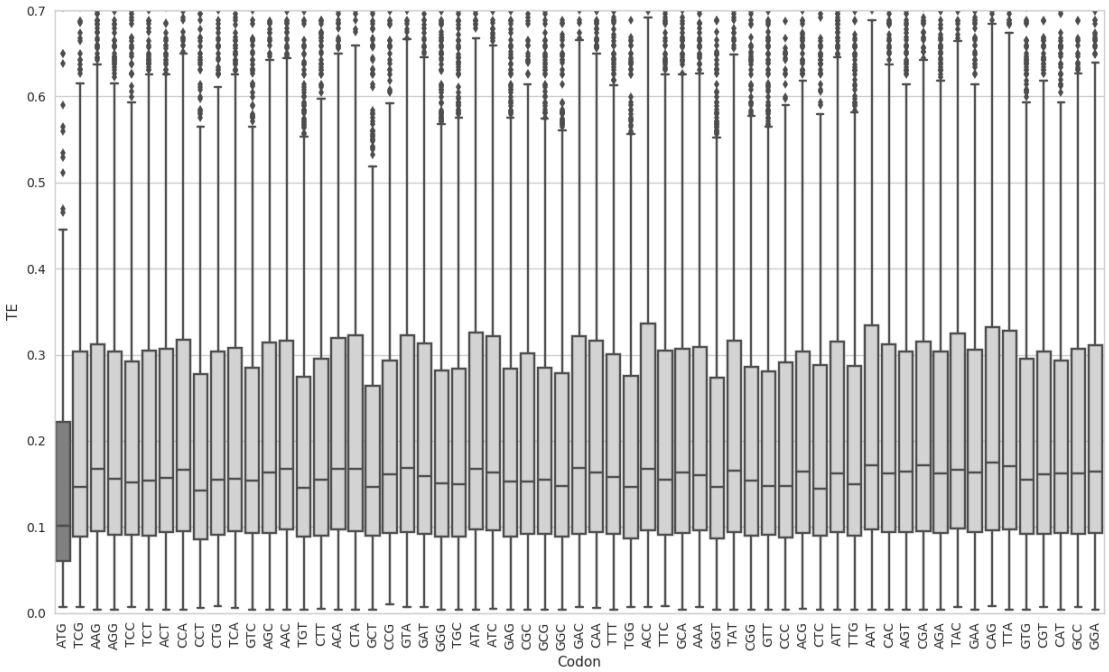


Figura 3.28. Comparación de la eficiencia traduccional de genes expresados en el estadio epimastigota, que en las regiones 5' UTR contienen uORFs con potencial represivo (sin *Overlaps*). Cada caja representa la determinación de uORFs variando únicamente el codón iniciador.

3.4.8. Tamaño de las regiones 5' UTR

Debido a los requerimientos establecidos para considerar a un uORF como represor de la TE (Tabla 3.10), las regiones 5' UTR de los genes que los poseen es más grande que la media (~80 nucleótidos). Por ende, dada esta definición, la mayor parte de los genes no tendrán regulación de este tipo. Sin embargo, podemos comparar con las regiones 5' UTR que presentan uORFs no AUG con características represivas (definidas en la sección anterior).

Como se puede observar en Figura 3.29 y Figura 6.6 el tamaño de las regiones 5' no traducidas de los genes asociados a uORF de AUG es significativamente mayor (test U de Mann-Whitney, p-valor < 0.001) que el resto de los grupos comparados, para los estadios epimastigota y tripomastigotas metacíclicos respectivamente (material suplementario). Lo que puede determinar que el largo del 5' UTR de por sí es un actor determinante en la determinación del potencial represivo de los uORFs o que el mantenimiento de uORFs largos aumente el tamaño de estas regiones.

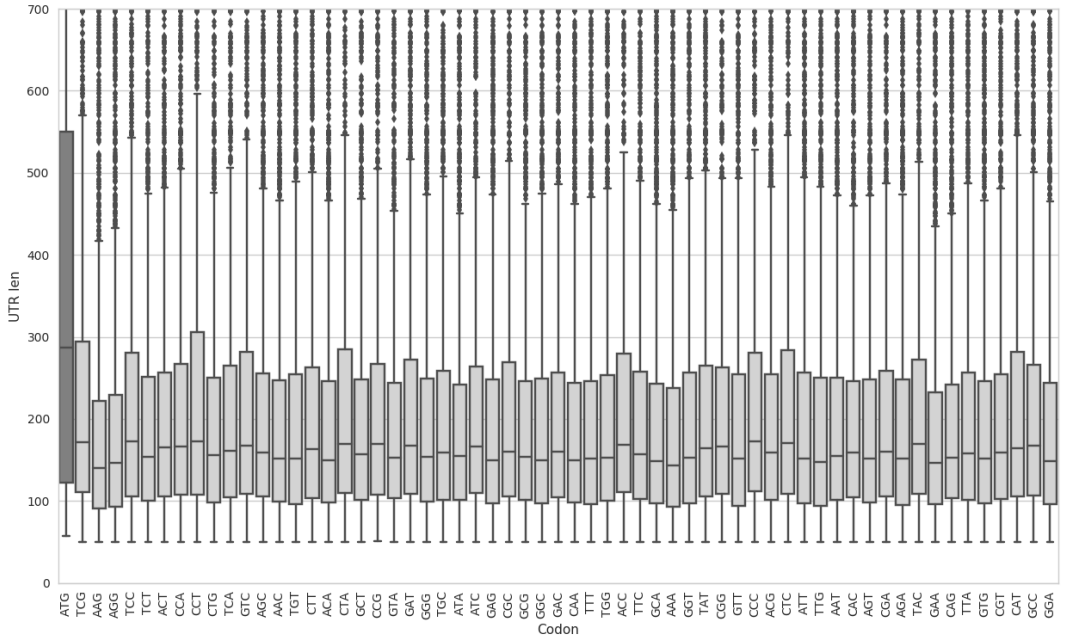


Figura 3.29. Comparación de tamaño de las regiones 5' UTR determinada en el estadio epimastigota de genes que contienen al menos un uORF con potencial represivos, cambiando en cada caso el codón iniciador.

Para comprobar si efectivamente el largo del 5' UTR de por sí es un actor influyente en la determinación del potencial represivo de los uORFs, correlacionamos el largo de los 5' UTR con la eficiencia traduccional del gen asociado para los estadios epimastigota y tripomastigota metacíclico (Figura 3.30 A y B).

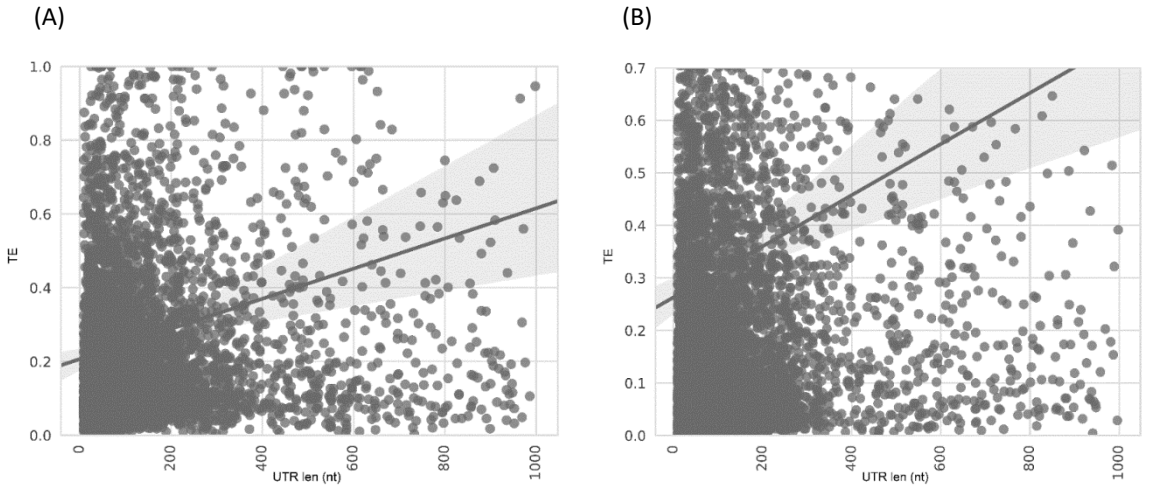


Figura 3.30. Correlación entre eficiencia traduccional y tamaño de las regiones 5' UTR determinadas para el estadio epimastigota (A) y tripomastigota metacíclico (B).

En ninguno de los estadios estudiados se observa una correlación entre el largo del 5' UTR y la eficiencia traduccional, lo que sugiere que los genes con uORF represivos tienen menor eficiencia traduccional por, posiblemente, la presencia del uORF y no así por el tamaño del 5' UTR. Finalmente, decidimos estudiar si las distintas categorías analizadas presentan diferencias en este aspecto. En la Tabla 3.13 se resumen las medianas obtenidas en cada categoría para ambos estadios.

Tabla 3.13. Tamaño medio de las regiones 5' UTR de los genes pertenecientes a las categorías *Overlap*, represivos, no represivos y total, para ambos estadios estudiados.

Estadio	Categoría	Tamaño 5' UTR (nt)
epimastigota	Overlap	44
epimastigota	Represivo	279
epimastigota	No Represivo	116
epimastigota	Total	81
tripomastigota	Overlap	76
tripomastigota	Represivo	244
tripomastigota	No Represivo	117
tripomastigota	Total	82

Por lo previamente discutido, era esperable que la categoría represiva tuviera regiones 5' UTR más grande que el resto ya que contienen uORF de mayor tamaño. Por otra parte, dada la definición de uORF no represivos (se restringe el tamaño mínimo de regiones 5'

UTR a 50 nt) era esperable que su tamaño sea mayor que la media de los genes (categoría total), y menor, por no tener la necesidad de acomodar uORF represivos, que la categoría represiva. Finalmente, la categoría *Overlap* presentó el menor tamaño de 5' UTR, lo que puede indicar que los marcos de lectura de este tipo son por sí solos buenos reguladores de la traducción, sin la necesidad de tener elementos extras que participen en la regulación. Dado que el mecanismo de control de la regulación, de esta clase, no requiere la presencia de un marco de lectura de gran tamaño, como si ocurre con los uORFs represivos, no hay necesidad de aumentar la región 5' UTR. Interesantemente, este efecto es significativamente más marcado en el estadio epimastigota, lo que apoya la idea previamente mencionada de que en el estadio tripomastigota metacíclico existe un mayor control de la regulación traduccional mediada por otros mecanismos (Figura 3.31 y Tabla 3.14).

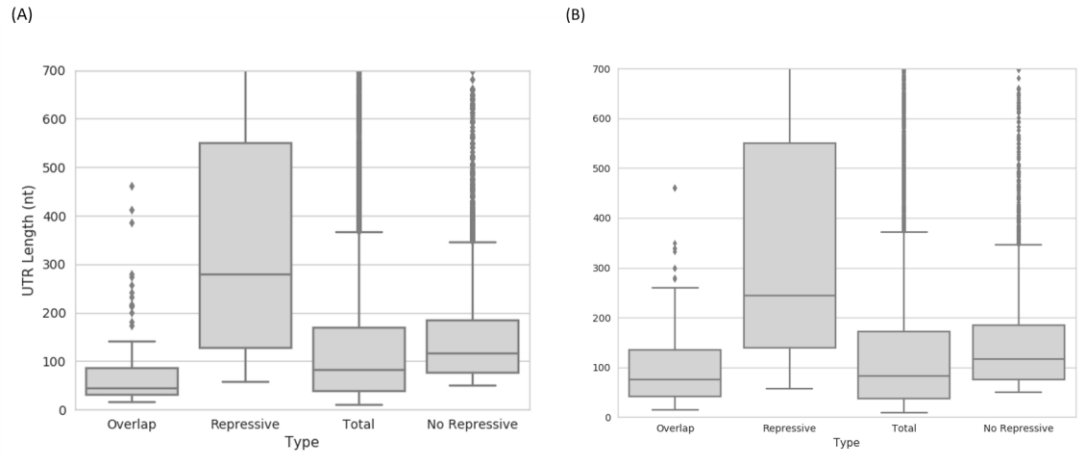


Figura 3.31. Comparación del tamaño de las regiones 5' UTR determinadas para las cuatro categorías analizadas. Resultados obtenidos en epimastigota (A) y tripomastigotas (B).

Tabla 3.14. p-valor del test no-paramétrico U de Mann-Whitney de la comparación del tamaño de las regiones 5' UTR entre las cuatro categorías analizadas (*Overlap*, Represiva, No Represiva y Total) para ambos estadios.

Estadio	Comparación	p-valor
tripomastigota metacíclico	Overlap vs Represivos	4.9e-31
tripomastigota metacíclico	Overlap vs Total	0.11
tripomastigota metacíclico	Overlap vs No Represivos	3.3e-18
tripomastigota metacíclico	Represivos vs No Represivos	7.8e-29
tripomastigota metacíclico	Represivos vs Total	5.2e-39
tripomastigota metacíclico	Total vs No Represivos	3.4e-92
epimastigota	Overlap vs Represivos	3.6e-27
epimastigota	Overlap vs Total	6.12e-6
epimastigota	Overlap vs No Represivos	8.5e-26
epimastigota	Represivos vs No Represivos	2.74e-27
epimastigota	Represivos vs Total	3.6e-38
epimastigota	Total vs No Represivos	4.1e-97

3.4.9. Determinación del contexto a nivel de secuencia primaria del codón iniciador uAUG.

Kozak propuso que la eficiencia traduccional está fuertemente determinada por el contexto del AUG iniciador (Kozak, 1978; 2002) . Sin embargo, en muchos organismos, y en T. *brucei* y *L. major* en particular, no se observa la secuencia Kozak conservada (Nakagawa et al., 2008). Para comprobar que esta secuencia tampoco se encuentra en *T. cruzi*, se obtuvo el contexto a nivel de secuencia primaria del AUG iniciador de todos los genes y de todos los uORF represivos en el estadio epimastigota. Además, esta comparación nos permite verificar la existencia de una composición nucleotídica particular en las regiones iniciadoras que impacte en la eficiencia traduccional de los genes de *T. cruzi*. Las regiones analizadas abarcan 10 nucleótidos, tanto río arriba como río abajo, de la adenina del codón iniciador. Como se puede observar en Figura 3.32 no se encontraron motivos de secuencia, ni una composición nucleotídica particular en el contexto del codón iniciador en los CDS ni en los uORF, que sugiera que en *T. cruzi* exista

un impacto del contexto iniciador a nivel de secuencia primaria como se ha reportado en organismos superiores (Kozak, 2002).

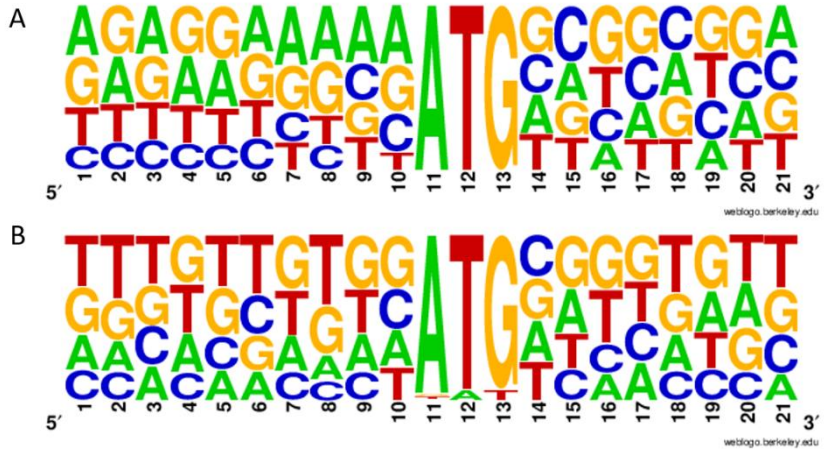


Figura 3.32. Logo de secuencia del contexto iniciador AUG generado con WebLogo 3 (Crooks et al., 2004). A) Logo centrado en el AUG del CDS principal B) Logo centrado en el AUG del uORF represivo más extenso.

3.4.10. Correlación entre densidad de uORF y potencial represivo

Se ha reportado que el número de uORFs (densidad de uORF) en la región 5' UTR de un ARNm está directamente relacionado con la eficiencia traduccional (Chew et al., 2016). A mayor densidad de uORF mayor represión traduccional. Para verificar si esa observación también se cumple, se generó un script *homemade* (material suplementario) para determinar la densidad de uORF por 5' UTR (Figura 3.33 y Figura 6.8, para epimastigota y tripomastigota metacíclico, respectivamente).

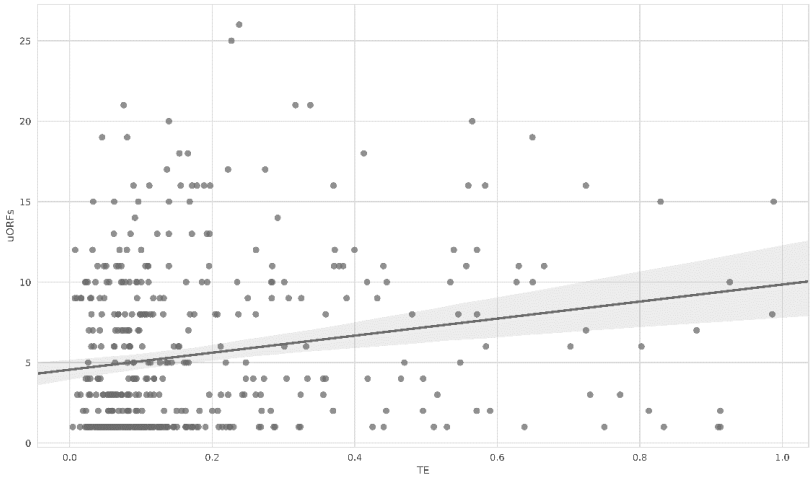


Figura 3.33. Correlación entre la densidad de uORF (uAUG) y la eficiencia traduccional del gen asociado en el estadio epimastigota.

Los resultados obtenidos indican que no existe correlación entre la cantidad de uORFs presenten en la región 5' UTR y la eficiencia traduccional del gen, al menos para los estadios estudiados.

3.4.11. Los genes asociados a 5' UTR contenedores de marcos de lectura represivos son de baja expresión

Hemos demostrado que la presencia de marcos de lectura represivos lleva a la disminución de la eficiencia traduccional de ARNm contenedor. Dado la disponibilidad de datos proteómicos en los estadios analizadps nos preguntamos si esa disminución se vería reflejada en la abundancia de proteínas (de Godoy et al., 2012). Mediante un *script* de elaboración propia (material suplementario) se correlacionó la abundancia proteica de grupos ARNm contenedores de uORF de uAUG (características represivas) contra ARNm portadores de uORF no-AUG. Tanto para el estadio epimastigota como para el tripomastigota metacíclico se puede observar que la abundancia proteica no cambia según el codón iniciador (Figura 3.34 y Figura 3.35)

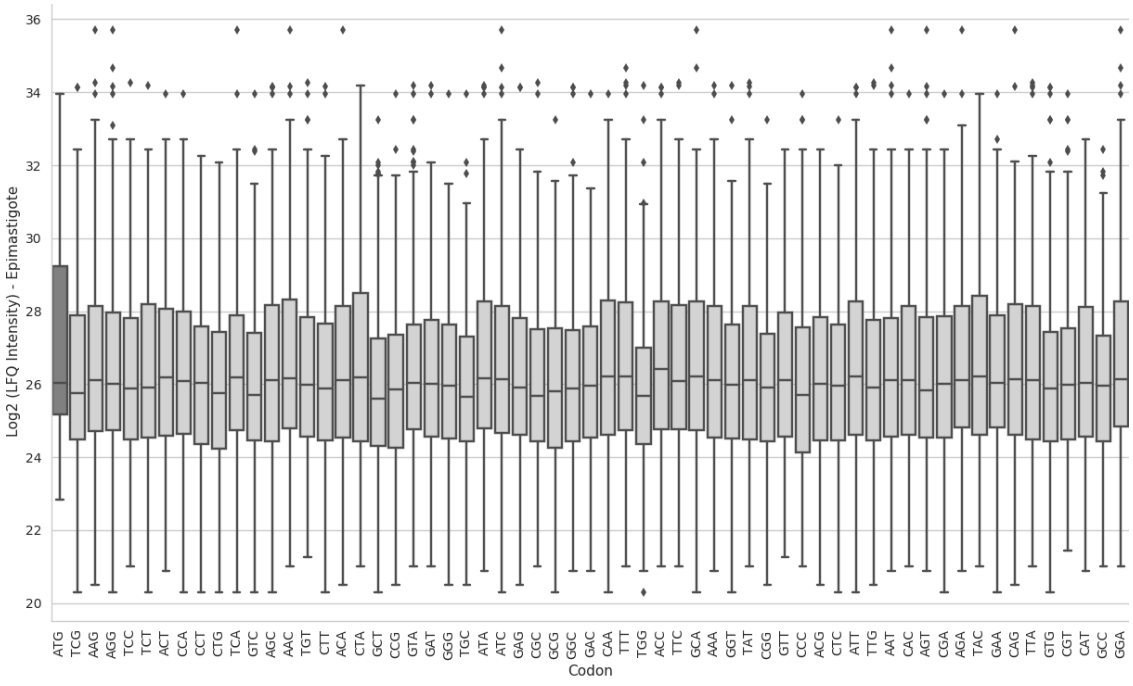


Figura 3.34. Comparación de los niveles de proteína asociados a genes cuyas regiones 5' UTR presentar uORFs asociados los 61 posibles codones iniciadores. Los datos de proteómica fueron obteniedos de (de Godoy et al., 2012) y representan el estadio epimastigota.

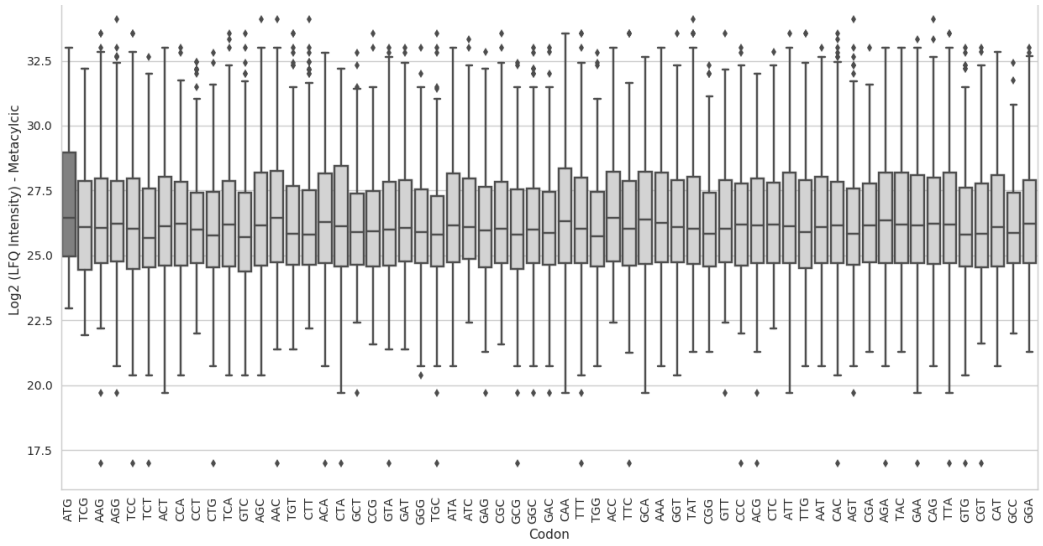


Figura 3.35. Comparación de los niveles de proteína asociados a genes cuyas regiones 5' UTR presentar uORFs asociados los 61 posibles codones iniciadores. Los datos de proteómica fueron obteniedos de (de Godoy et al., 2012) y representan el estadio tripomastigota metacíclico.

Sin embargo, de los aproximadamente 600 ARNm que contienen uORF, menos de 40 fueron detectados. Por ende, estos resultados no son representativos de la situación real y posiblemente reflejen los genes de alta expresión que si son detectados en los estudios proteómicos. Para testear esta hipótesis y dado que la abundancia proteica se relaciona directamente con la cantidad de huellas ribosomales, decidimos calcular los valores de RPKM (*script* de elaboración propia; material suplementario) de todos los genes de *T. cruzi* y compararlos con aquellos ARNm que portan uORF no detectados en el proteoma.

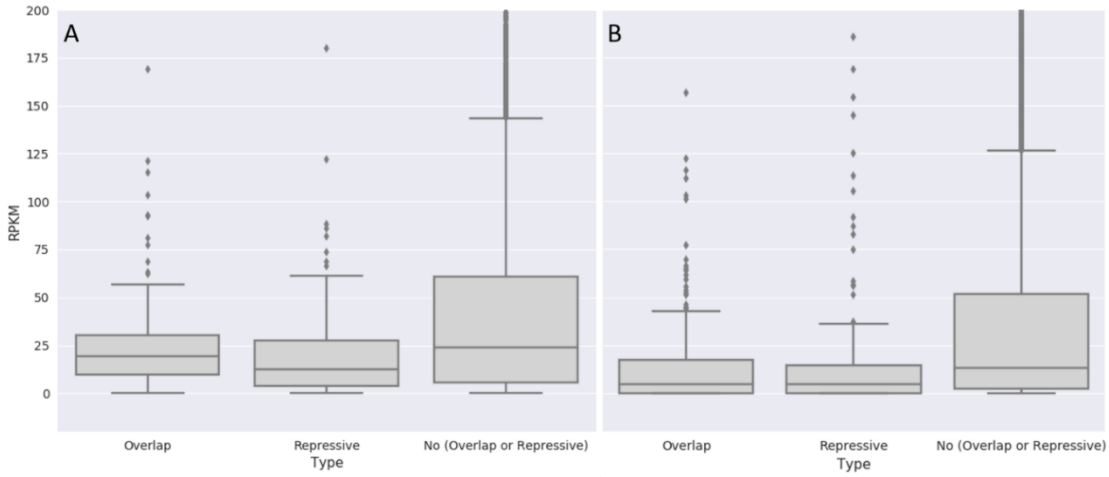


Figura 3.36. Comparación entre los valores de RPKM de los ARNm con uORF contra el resto de los ARNm. A) Valores determinados para el estadio epimastigota. B) Valores determinados para el estadio tripomastigota metacíclico.

En la Figura 3.36 A y B se puede evidenciar que los ARNm cuyas regiones 5' UTR contienen uORF represivos y que no fueron detectados en el proteoma, presentan en promedio un menor nivel de traducción que el resto de los genes (test U de Mann Whitney, material suplementario). Nos preguntamos entonces si existe una representación diferencial de genes en estos datos, dependiente de la categoría en la que fueron clasificados (*Overlap*, uORF represivo, no represivo y total). El enriquecimiento fue evaluado mediante un test de Fisher (Tabla 3.15) y nos permitió comprobar que existe una menor representación en el proteoma de las proteínas cuyos ARNm contienen marcos represivos mientras que lo opuesto se observa para las proteínas que no los contienen. Dado que, la eficiencia traduccional está estrechamente correlacionada con la abundancia proteica, los resultados obtenidos están dentro de los esperados, y permiten explicar lo observado (Schwanhausser et al., 2011). Los análisis fueron realizados para los estadios epimastigota (Figura 3.37 A) y tripomastigota metacíclico (Figura 3.37 B).

Tabla 3.15. Estudio de la representación de ARNm pertenecientes a distintas categorías, con respecto a la totalidad de los ARNm presentes en datos proteómicos obtenidos de (de Godoy et al., 2012). Analizado mediante test de Fisher.

Categoría	Estadio	Fisher p-valor
Overlap - Total	Epimastigota	1,07E-04
Repressive - Total	Epimastigota	2,95E-05
No Repressive - Total	Epimastigota	< 2.2E-16
Overlap - Total	Metacíclico	2,25E-04
Repressive - Total	Metacíclico	8,70E-03
No Repressive - Total	Metacíclico	< 2.2E-16

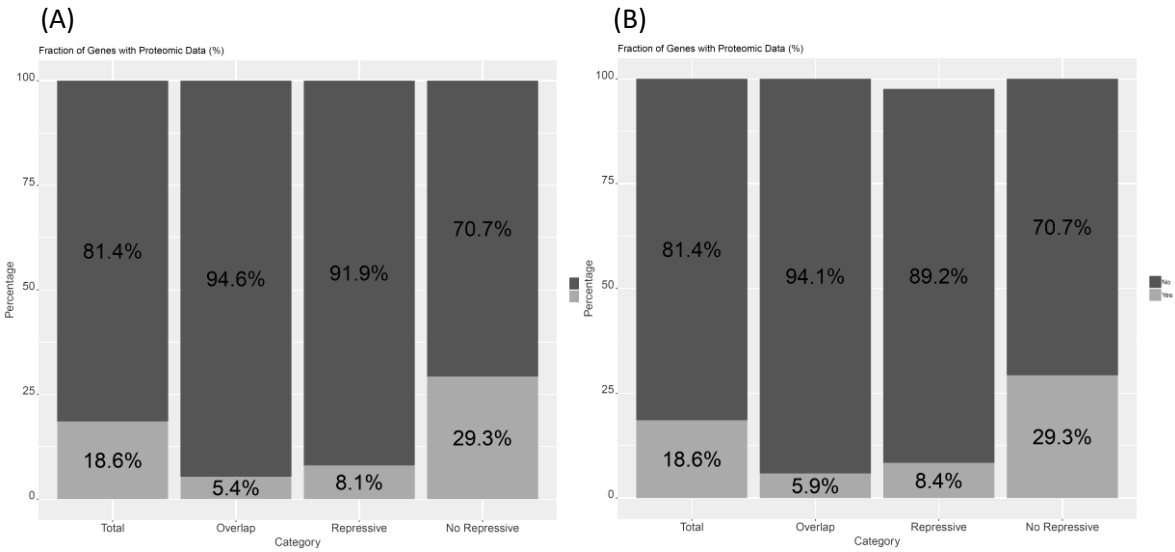


Figura 3.37. Visualización de la fracción de los genes presentes en los datos proteómicos de (de Godoy et al., 2012) en el estadio epimastigota (A) y tripomastigota metacíclico (B).

Este fenómeno se correlaciona con una disminución en la cantidad de huellas ribosomales observados en los genes con marcos represivos en comparación a los genes con marcos no represivos. Interesantemente, estos últimos son los que presentan mayor cantidad de huellas ribosomales, evidenciando la buena TE observada anteriormente (Figura 3.38 y Tabla 3.16).

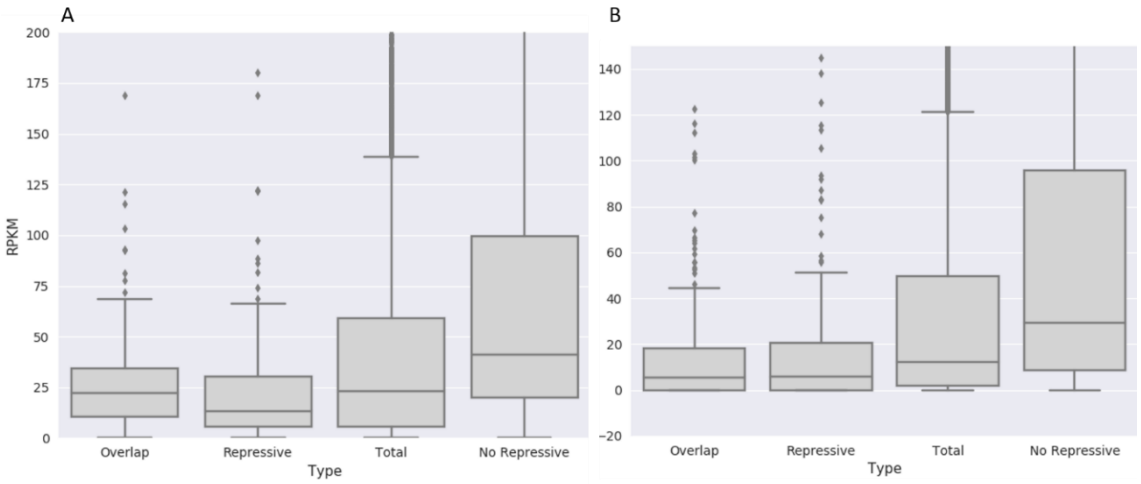


Figura 3.38. Comparación de los niveles de huellas ribosomales (RPKM) entre las cuatro categorías analizadas en los estadios epimastigota (A) y tripomastigota metacíclico (B).

Tabla 3.16. p-valor del test no-paramétrico U de Mann-Whitney de la comparación del nivel traduccional (RPKM) entre los genes pertenecientes a las cuatro categorías analizadas (Overlap, Represiva, No Represiva y Total) para ambos estadios.

Estadio	Comparación	p-valor
tripomastigota metacíclico	Overlap vs Represivos	0.30
tripomastigota metacíclico	Overlap vs Total	7.9e-8
tripomastigota metacíclico	Overlap vs No Represivos	1.4e-31
tripomastigota metacíclico	Represivos vs No Represivos	2.2e-22
tripomastigota metacíclico	Represivos vs Total	3.8e-05
tripomastigota metacíclico	Total vs No Represivos	1.2e-108
epimastigota	Overlap vs Represivos	0.014
epimastigota	Overlap vs Total	0.13
epimastigota	Overlap vs No Represivos	1.7e-13
epimastigota	Represivos vs No Represivos	5.6e-27
epimastigota	Represivos vs Total	6.6e-05
epimastigota	Total vs No Represivos	2.2e-137

3.4.12. Análisis de categorías génicas de genes con 5’ UTR con uORFs represivos y no represivos.

El análisis de las categorías génicas para los genes con 5’ UTR categorizados como no represivos, muestra un enriquecimiento en genes que participan en procesos catabólicos, movimientos celulares, transporte , entre otras funciones *house keeping*, generalmente asociados a niveles de expresión altos (Figura 3.39 y Tabla 3.17).

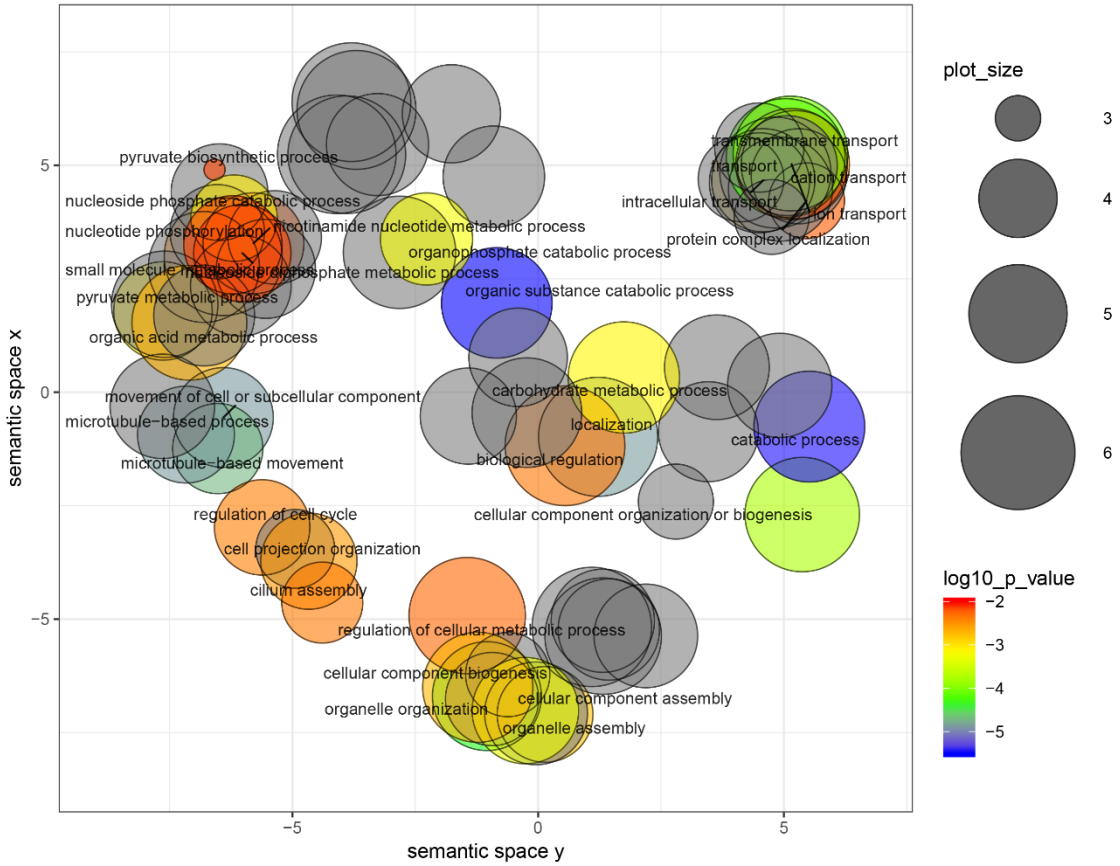


Figura 3.39. Categorías génicas de procesos biológicos sobrerrepresentadas en genes con 5’ UTR, detectadas en el estadio epimastigota, los cuales poseen uORFs de características no represiva. Los círculos coloreados indican un p-valor < 0.01, su agrupamiento revela categorías relacionadas funcionalmente y el tamaño muestra la frecuencia de la categoría de la base de datos UniProtKB. La visualización y sumarización de la ontología fue realizada con REVIGO (Supek et al., 2011).

Tabla 3.17. Categoría génicas (procesos biológicos) sobrerrepresentadas (p-valor < 0.01) en los genes cuyas 5' UTR fueron categorizadas como no represivas en el estadio epimastigota.

GO ID	Description	log10 p-value
GO:1901575	organic substance catabolic process	-5,50
GO:0009056	catabolic process	-5,48
GO:0044248	cellular catabolic process	-4,88
GO:0006928	movement of cell or subcellular component	-4,88
GO:0007017	microtubule-based process	-4,87
GO:0051179	localization	-4,87
GO:1901565	organonitrogen compound catabolic process	-4,85
GO:0007018	microtubule-based movement	-4,71
GO:0006996	organelle organization	-4,35
GO:0006810	transport	-4,25
GO:0051234	establishment of localization	-4,17
GO:0055085	transmembrane transport	-4,08
GO:0016043	cellular component organization	-4,03
GO:0071840	cellular component organization or biogenesis	-3,76
GO:0030163	protein catabolic process	-3,68
GO:0022607	cellular component assembly	-3,30
GO:0046434	organophosphate catabolic process	-3,22
GO:0005975	carbohydrate metabolic process	-3,11
GO:0006090	pyruvate metabolic process	-3,05
GO:1901292	nucleoside phosphate catabolic process	-3,05
GO:0046907	intracellular transport	-2,86
GO:0070925	organelle assembly	-2,79
GO:0051649	establishment of localization in cell	-2,71
GO:0044085	cellular component biogenesis	-2,71
GO:0009057	macromolecule catabolic process	-2,71
GO:0006082	organic acid metabolic process	-2,68
GO:0008150	biological_process	-2,54
GO:0044257	cellular protein catabolic process	-2,52
GO:0051603	proteolysis involved in cellular protein catabolic process	-2,52
GO:0030030	cell projection organization	-2,52
GO:0044782	cilium organization	-2,52
GO:0019752	carboxylic acid metabolic process	-2,50
GO:0044281	small molecule metabolic process	-2,45
GO:0009166	nucleotide catabolic process	-2,43
GO:0043436	oxoacid metabolic process	-2,40
GO:0051726	regulation of cell cycle	-2,39
GO:0006812	cation transport	-2,38
GO:0060271	cilium assembly	-2,36
GO:0030031	cell projection assembly	-2,36
GO:0046700	heterocycle catabolic process	-2,34
GO:0044270	cellular nitrogen compound catabolic process	-2,34
GO:0019439	aromatic compound catabolic process	-2,34
GO:1901361	organic cyclic compound catabolic process	-2,34
GO:0065007	biological regulation	-2,33
GO:0031323	regulation of cellular metabolic process	-2,27
GO:0006605	protein targeting	-2,26
GO:0010564	regulation of cell cycle process	-2,24
GO:0031503	protein complex localization	-2,21
GO:0006520	cellular amino acid metabolic process	-2,19
GO:0046496	nicotinamide nucleotide metabolic process	-2,17
GO:0019362	pyridine nucleotide metabolic process	-2,17

GO:0006811	ion transport	-2,12
GO:0006096	glycolytic process	-2,12
GO:0006165	nucleoside diphosphate phosphorylation	-2,12
GO:0009135	purine nucleoside diphosphate metabolic process	-2,12
GO:0009179	purine ribonucleoside diphosphate metabolic process	-2,12
GO:0009185	ribonucleoside diphosphate metabolic process	-2,12
GO:0006757	ATP generation from ADP	-2,12
GO:0046031	ADP metabolic process	-2,12
GO:0042866	pyruvate biosynthetic process	-2,12
GO:0009132	nucleoside diphosphate metabolic process	-2,12
GO:0046939	nucleotide phosphorylation	-2,12
GO:0030150	protein import into mitochondrial matrix	-2,08
GO:0007346	regulation of mitotic cell cycle	-2,08
GO:0032465	regulation of cytokinesis	-2,00
GO:0051302	regulation of cell division	-2,00
GO:0050793	regulation of developmental process	-2,00
GO:0051783	regulation of nuclear division	-2,00

Por otra parte, el análisis de las categorías génicas de los genes que portan en su 5' UTR uORFs con características represivas, muestra un leve enriquecimiento en la categoría elongación traduccional y fosforilación (Figura 3.40 y Tabla 3.18).

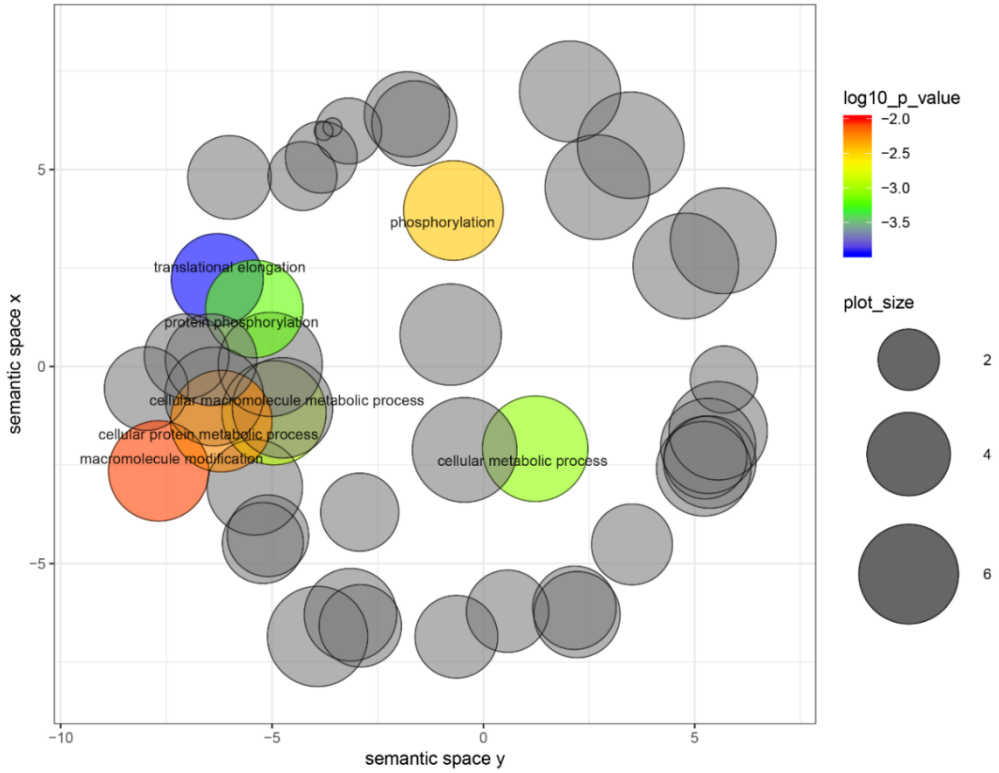


Figura 3.40. Categorías génicas de procesos biológicos sobrrepresentadas en genes con 5' UTR, detectadas en el estadio epimastigota, los cuales poseen uORFs de características no represiva. Los círculos coloreados indican un p-valor < 0.01, su agrupamiento revela categorías relacionadas funcionalmente y el tamaño muestra la frecuencia de la categoría de la base de datos UniProtKB. La visualización y sumarización de la ontología fue realizada con REVIGO.

Tabla 3.18. Categoría génicas (procesos biológicos) sobrerrepresentadas (p-valor < 0.01) en los genes cuyas 5' UTR fueron categorizadas como represivas en el estadio epimastigota.

GO ID	Description	log10 p-value
GO:0006414	translational elongation	-3,954
GO:0009987	cellular process	-3,5076
GO:0006468	protein phosphorylation	-3,199
GO:0044237	cellular metabolic process	-3,0789
GO:0044260	cellular macromolecule metabolic process	-2,9067
GO:0008150	biological_process	-2,8345
GO:0016310	phosphorylation	-2,4525
GO:0036211	protein modification process	-2,4296
GO:0006464	cellular protein modification process	-2,4296
GO:0044267	cellular protein metabolic process	-2,2434
GO:0043412	macromolecule modification	-2,0675
GO:0006796	phosphate-containing compound metabolic process	-2,0181

Las observaciones realizadas en el estadio tripomastigota metacíclico produjeron resultados similares, sugiriendo que no existen grandes diferencias entre los procesos controlados por uORF entre los estadios estudiados. De forma general se puede concluir que los genes de alta expresión, como por ejemplo *house keeping*, están enriquecidos en la categoría no represiva, mientras que categorías más específicas están sometidas a un control traduccional mediado por marcos de lecturas represivos.

3.4.13. Determinación de uso de uORF diferencial

Para poder evaluar si existe una regulación estadio específica mediado por uORF, comparamos el tamaño de las regiones 5' UTRs entre los estadios estudiados. En total se obtuvieron las regiones 5' UTR de 6245 genes. Para la mayoría de los genes el sitio de trans-empalme mayoritario detectado fue el mismo en ambos estadios (83%, Figura 3.41), además la mayor parte de las UTR diferenciales tuvieron una diferencia menor a 10 nucleótidos (~60%, Figura 3.42).

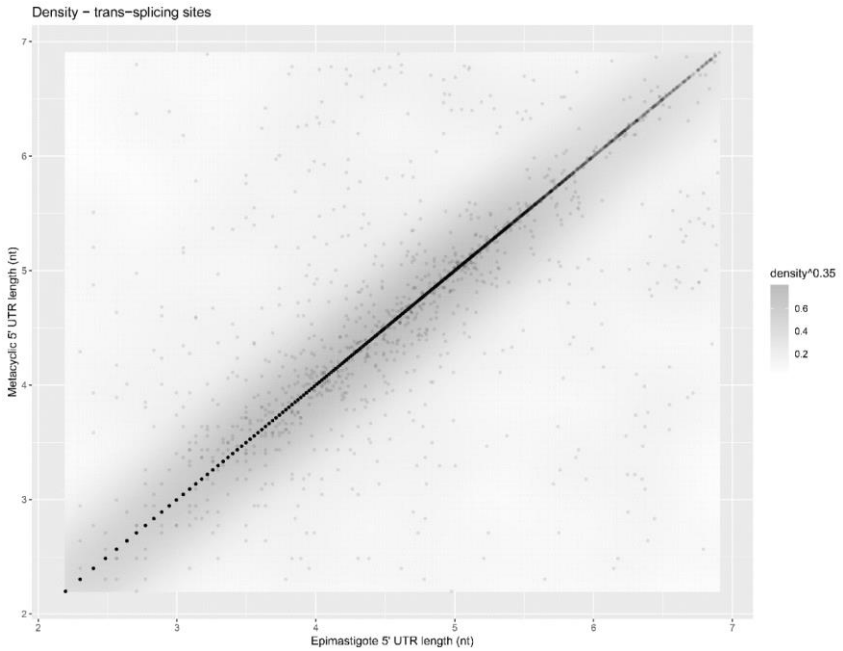


Figura 3.41. Comparación de tamaño entre las regiones 5’ UTR de los ARNm presentes en los estadios epimastigota y tripomastigota metacíclico. Las regiones 5’ UTRs fueron definidas utilizando UTRme.

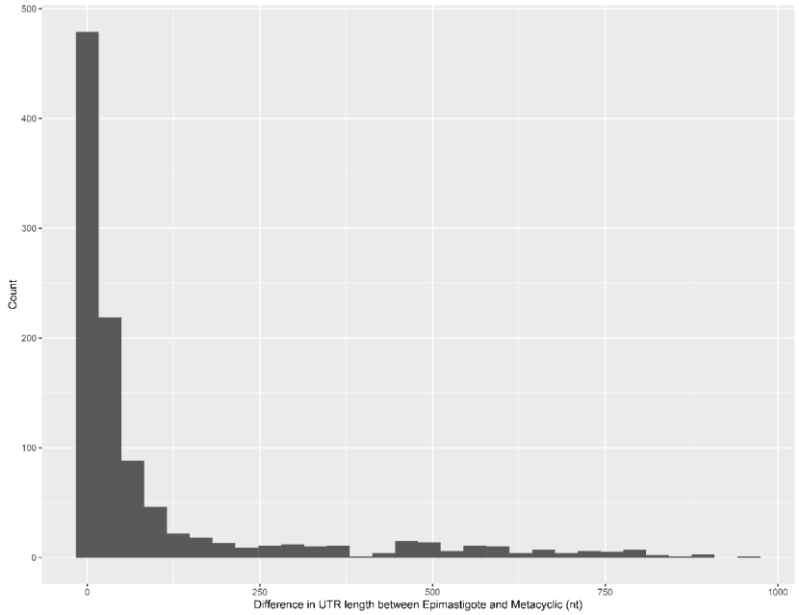


Figura 3.42. Distribución de las diferencia de tamaño entre las regiones 5’ UTR de epimastigota y tripomastigota metacíclico.

En total (considerando ambos estadios) se detectaron 572 genes con 5’ UTR portadores de uORF represivos, mientras que 492 genes presentaron uAUG solapantes (no en fase) con el CDS principal. En la primer categoría se lograron detectar las regiones 5’ UTRs de 454 genes de los cuales 160 presentaron diferencias en cuanto al tamaño del 5’UTR (35%) en 22 el cambio de tamaño del 5’ UTR generó la aparición de un uORF represivo en el

estadio epimastigota, y coincidentemente, en otros 22 genes el cambio provocó la aparición de un uORF represivo en el estadio tripomastigota metacíclico .De los 492 genes cuyas regiones 5' UTR fueron clasificadas como *Overlap* para 390 fue posible definir regiones 5' UTRs en ambos estadios, de los cuales 116 (~30%) difirieron en tamaño. 14 de los 116 genes fueron únicos de epimastigota mientras que 25 fueron únicos de tripomastigotas metacíclicos.

En la Tabla 3.19, se muestra únicamente aquellos genes que el cambio de UTR determinó la aparición de un marco represor (*Overlap* y/o uORF) y que además presentaron cambios significativos en TE durante la metaciclogénesis.

Tabla 3.19. Genes que regulan diferencialmente su eficiencia traduccional (padj < 0.05 y FC de 2) y que poseen marcos represivo en un solo estadio. Se nombra el estadio donde se encuentra el elemento represor.

Estadio	Gen	padj	Log2FC	productos
epimastigota	TcCLB.510755.129	0	2,393	60S ribosomal protein L12, putative
epimastigota	TcCLB.507093.170	0	3,556	hypothetical protein, conserved
epimastigota	TcCLB.509455.40	0,003	4,339	hypothetical protein, conserved
epimastigota	TcCLB.510517.100	0,013	1,945	hypothetical protein, conserved
epimastigota	TcCLB.504149.150	0,013	2,497	hypothetical protein, conserved
epimastigota	TcCLB.507009.40	0,032	-1,72	galactokinase, putative
tripomastigota metacíclico	TcCLB.507929.10	0	4,3	kinesin, putative (fragment)
tripomastigota metacíclico	TcCLB.508307.130	0	2,256	hypothetical protein, conserved
tripomastigota metacíclico	TcCLB.510055.150	0,007	2,461	hypothetical protein, conserved
tripomastigota metacíclico	TcCLB.511557.40	0,013	2,015	hypothetical protein, conserved
tripomastigota metacíclico	TcCLB.511555.50	0,02	2,596	hypothetical protein, conserved
tripomastigota metacíclico	TcCLB.510607.80	0,022	-2,832	DGF-1
tripomastigota metacíclico	TcCLB.506473.20	0,032	1,184	DnaJ chaperone protein, putative
tripomastigota metacíclico	TcCLB.510751.4	0,037	-1,854	aminopeptidase P, putative
tripomastigota metacíclico	TcCLB.505807.160	0,048	2,247	hypothetical protein, conserved

Los resultados aquí presentados permiten concluir que *T. cruzi* utiliza uORFs como un mecanismo general para regular la eficiencia traduccional. La eficacia para regularla se

asoció a características intrínsecas al uORF como su posición en el 5' UTR, tamaño y codón iniciador. Se evidenció como los genes con uORF represivos presentan bajos niveles de huellas ribosomales estando subrepresentados en datos proteómicos, caso opuesto a los genes cuya región 5' UTR fue catalogada como no represiva. Finalmente, si bien se observan marcos represivos específicos de estadio, estos no parecen influir de forma general en los cambios de eficiencia traduccional observados en la metaciclogénesis. Cabe destacar que el *abstract* de un manuscrito conteniendo estos resultados ha sido aceptado para ser considerado en la publicación *Frontiers in Genetics*.

3.4.14. Estrategia

3.4.14.1. Datos genómicos

Los datos genómicos de *T. cruzi* y su anotación génica fueron obtenidos de la base de datos TriTrypDB (versión 32).

3.4.14.2. Determinación de regiones UTRs

Las secuencias UTR fueron determinadas para los estadios epimastigota y tripomastigota mediante la utilización del programa UTRme (Radio et al., 2018), a partir de los datos transcriptómicos generados en (Li et al., 2016). Para la determinación de los marcos de lectura se utilizaron únicamente los 5' UTR determinados por el sitio de trans-empalme de mayor puntaje y que fueron mayores a 5 nucleótidos. De esta manera se obtuvieron 8206 regiones en el estadio epimastigota y 8217 en tripomastigota.

A su vez, debido a la dificultad en determinar los extremos 5' UTR de las familias multigénicas, ya sea por el gran contenido de regiones repetidas o por problemas de ensamblado, se decidió removerlas del análisis. Las familias multigénicas removidas incluyen las grandes familias de proteínas de superficie como las MASPs, GP63, Mucinas y TS. Finalmente, se decidió eliminar las 5' UTRs que presenten una región codificante para proteínas dentro de ella. Para esto, se utilizó la herramienta BLASTX (Altschul et al.,

1990), eliminando UTRs que presentaron un *hit* contra las CDS de *T. cruzi* con un e-valor menor a 0.005.

3.4.14.3. Generación de logos de secuencia.

Los logos de secuencia fueron generados con la herramienta online WebLogo 3, utilizando parámetros por defecto.

3.4.14.4. Análisis de enriquecimiento de ontología génica

Los análisis de ontología génica fueron realizados mediante la base de datos TriTrypDB. La visualización y reducción de las categorías fue realizado mediante REVIGO en conjunto con el entorno gráfico del lenguaje R.

3.5. Búsqueda de motivos de secuencia primaria y secundaria en regiones UTR en genes co-modulados de la familia de trans-sialidasas y de proteínas ribosomales.

La optimización en el análisis de datos de *Ribosome Profiling* hechos en esta tesina confirmaron los resultados previamente obtenidos por el grupo (Smircich et al., 2015), detectando diferencias significativas en la TE de dos grupos de genes: los genes que codifican las proteínas ribosomales (PR) y los genes pertenecientes a la superfamilia de las trans-sialidasas (TS).

En el presente capítulo profundizaremos en los mecanismos de regulación post-transcripcional de estos grupos, estudiando específicamente la presencia de motivos presentes en regiones UTR (objetivo específico 5). Dado que la regulación génica post-transcripcional no parece deberse exclusivamente a motivos lineales (Goodarzi et al., 2012; De Gaudenzi et al., 2013), estudiaremos la presencia de regiones regulatorios a niveles de secuencia primaria y estructura secundaria.

3.5.1. Identificación de PR y de TS

3.5.1.1. Proteínas Ribosomales

El ribosoma de tripanosomátidos ha sido muy estudiado ya que posee características únicas que incluyen: ARNr fragmentado, expansión inusual de segmentos de ARNr, y variaciones de PR en cuanto estructura, estequiometría y composición en comparación con la mayoría de los ribosomas eucarióticos (Gao et al., 2005; Liu et al., 2016; Shalev-Benami et al., 2016). Sin embargo, si bien existe una anotación de origen automático de los genes ribosomales en los genomas de referencia de tripanosomátidos, no existe un análisis fino de sus integrantes. Por lo tanto procedimos a identificar las proteínas ribosomales de *T. cruzi* CL Brener basados en la estructura cristalográfica del ribosoma en kinetoplástidos (Gao et al., 2005; Hashem et al., 2013; Liu et al., 2016; Shalev-Benami et al., 2016) y nombrarlas basadas en (Ban et al., 2014). Determinamos los números de

copias mediante el TriTrypDB y mediante búsqueda bibliográfica anotamos descripciones y particularidades (Gao et al., 2005; Ayub et al., 2009; Hashem et al., 2013; Liu et al., 2016; Shalev-Benami et al., 2016). Los resultados se exponen en la Tabla 3.20. En total se identificaron 146 PR no mitocondriales (22 de ellos pertenecen a la cepa CL Brener haplotipo No-Esmeraldo, dado que no se encontraban presentes en los cromosomas del haplotipo CL Brener Esmeraldo-Like)

Tabla 3.20. Identificación de proteínas ribosomales de *T. cruzi*.

Subunidad	Nombre	Id TriTrypDB	# copias	Función	Comentarios	Anclaje a srRNA
Mayor	uL1	TcCLB.506963.10	1	60S ribosomal protein L10a	-	-
Mayor	uL2	TcCLB.511181.100 / TcCLB.511527.34	2	60S ribosomal protein L8	Muy conservada	-
Mayor	uL3	TcCLB.510879.110 / TcCLB.510879.120	2	60S ribosomal protein L3	Inserción de a.a 209-211 y extensión C-terminal, contacto con sRNA4	srRNA2
Mayor	uL4	TcCLB.503643.3	1	60S ribosomal protein L4	Extensión C-terminal	-
Mayor	uL5	TcCLB.508197.10 / TcCLB.508197.39	2	60S ribosomal protein L11	Extensión N y C-terminal	-
Mayor	uL6	TcCLB.504181.10 / TcCLB.511729.40	2	60S ribosomal protein L9	-	-
Mayor	eL6	TcCLB.505843.20 / TcCLB.507709.50	2	60S ribosomal protein L6	Residuos insertados en 128-143 estan en la region menos ordenada	srRNA3
Mayor	eL8	TcCLB.506401.320	1	60S ribosomal protein L7a	Extensión N-terminal	-
Mayor	uL11	TcCLB.511071.171 / TcCLB.510755.129	2	60S ribosomal protein L12	-	-
Mayor	uL13	TcCLB.506315.50 / TcCLB.506739.150	2	60S ribosomal protein L13a	Contacta srRNA3	-
Mayor	eL13	TcCLB.510323.30	1	60S ribosomal protein L13	Conservado	-
Mayor	uL14	TcCLB.508461.480 / TcCLB.508461.490	2	60S ribosomal protein L23	Conservado	-
Mayor	eL14	TcCLB.506861.30 / TcCLB.506937.30	2	60S ribosomal protein L14	Acortamiento N-terminal y Extensión C-terminal (vs levadura)	-
Mayor	uL15	TcCLB.508461.500 / TcCLB.508461.510	2	60S ribosomal protein L27A/L29	Conservado	-
Mayor	eL15	TcCLB.510767.10 / TcCLB.506945.290	2	60S ribosomal protein L15	Conservado	-
Mayor	uL16	TcCLB.510241.50 / TcCLB.510243.40	2	60S ribosomal protein L10	-	-
Mayor	uL18	TcCLB.510765.60 / TcCLB.510765.70	2	60S ribosomal protein L5	Inserción de a.a 136-141, 222-2224, deleción 163-165 (levadura) y extensión C-terminal, contacto con sRNA4	-
Mayor	eL18	TcCLB.506181.50 / TcCLB.504147.20	2	60S ribosomal protein L18	Inserción de a.a 72-77	-
Mayor	eL19	TcCLB.509149.60 / TcCLB.509149.40	2	60S ribosomal protein L19	Conservado, extensión C-terminal de 168	srRNA1

					a.a, contacta SSU(forma parte del canal de salida)	
Mayor	eL20	TcCLB.511001.120	1	60S ribosomal protein L18a	Conservado	-
Mayor	eL21	TcCLB.506405.149	1	60S ribosomal protein L21e	Conservado	-
Mayor	uL22	TcCLB.503449.10 / TcCLB.506213.80	2	60S ribosomal protein L17	Acortamiento C-terminal (20 a.a)	-
Mayor	eL22	TcCLB.504147.120 / TcCLB.510719.160	2	60S ribosomal protein L22	Conservado	-
Mayor	uL23	TcCLB.509151.140	1	60S ribosomal protein L23a	Extensión N-terminal (90 aa)	-
Mayor	uL24	TcCLB.511067.20 / TcCLB.510761.14	2	60S ribosomal protein L26	Extensión N-terminal (21 aa)	-
Mayor	eL24	TcCLB.503611.20 / TcCLB.503611.40	2	60S ribosomal protein L24	Acortamiento C-terminal	-
Mayor	eL27	TcCLB.511545.20 / TcCLB.511545.40	2	60S ribosomal protein L27	Conservado	-
Mayor	eL28	TcCLB.510101.30 / TcCLB.510101.40	2	60S ribosomal protein L28	Inserción de a.a 118-132	-
Mayor	uL29	TcCLB.509979.90 / TcCLB.509979.95	2	60S ribosomal protein L35	Inserción de a.a 81-84	-
Mayor	eL29	TcCLB.510719.30 / TcCLB.510719.35	2	60S ribosomal protein L29	Extensión C-terminal (vs Levadura)	-
Mayor	uL30	TcCLB.508207.100 / TcCLB.508207.110	2	60S ribosomal protein L7	Conservado	-
Mayor	eL30	TcCLB.503453.30 / TcCLB.503453.40	2	60S ribosomal protein L30	Conservado	-
Mayor	eL31	TcCLB.510737.70 / TcCLB.510737.79	2	60S ribosomal protein L31	Extensión C-terminal contacta KSD	srRNA4
Mayor	eL32	TcCLB.510769.90 / TcCLB.511145.20	2	60S ribosomal protein L32	Conservado	-
Mayor	eL33	TcCLB.506559.470	1	60S ribosomal protein L35a	Extensión N-terminal, inserción 89-98	srRNA3
Mayor	eL34	TcCLB.506963.20 / TcCLB.507831.90	2	60S ribosomal protein L34	Extensión N-terminal, inserción 45-50, contacta srRNA1	srRNA1
Mayor	eL36	TcCLB.510767.20	1	60S ribosomal protein L36	Conservado	-
Mayor	eL37	TcCLB.506885.14 / TcCLB.510431.274	2	60S ribosomal protein L37	Conservado	-
Mayor	eL38	TcCLB.503881.39 / TcCLB.503575.34 / TcCLB.509351.4	3	60S ribosomal protein L38	Contacta srRNA1	-
Mayor	eL39	TcCLB.511217.145	1	60S ribosomal protein L39	Conservado	-
Mayor	eL40	TcCLB.507483.4 / TcCLB.506655.20	2	60S ribosomal protein L40 (Polyubiquitin)	-	-
Mayor	eL41	-	0	-	-	-
Mayor	eL42	TcCLB.507105.40 / TcCLB.509583.4	2	60S ribosomal protein L44	Conservado	-
Mayor	eL43	TcCLB.511145.46 / TcCLB.507641.233	2	60S ribosomal protein L37a	Conservado	-
Mayor	P0	TcCLB.508355.250	1	60S acidic ribosomal protein P0	-	-
Mayor	P1A	TcCLB.510309.40	1	60S acidic ribosomal protein	-	-

Mayor	P1B	TcCLB.510309.50	1	60S acidic ribosomal protein P2	-	-
Mayor	P2A	TcCLB.505977.26	1	60S acidic ribosomal protein P2	-	-
Mayor	P2B	TcCLB.509165.40	1	60S acidic ribosomal protein P2 beta	-	-
Menor	eS1	TcCLB.510999.39 / TcCLB.511001.18 / TcCLB.511001.9	3	40S ribosomal protein S3A	-	-
Menor	uS2	TcCLB.442383.9 / TcCLB.503757.10 / TcCLB.509825.14	3	40S ribosomal protein SA	-	-
Menor	uS3	TcCLB.507677.39 / TcCLB.430605.29	2	40S ribosomal protein S3	-	-
Menor	uS4	TcCLB.506401.120 / TcCLB.504163.30	2	40S ribosomal protein S9	-	-
Menor	eS4	TcCLB.511051.50 / TcCLB.511051.39	2	40S ribosomal protein S4	-	-
Menor	uS5	TcCLB.503833.40 / TcCLB.506213.60	2	40S ribosomal protein S2	-	-
Menor	eS6	TcCLB.510769.49	1	40S ribosomal protein S6	-	-
Menor	uS7	TcCLB.510101.170 / TcCLB.510101.180	2	40S ribosomal protein S5	-	-
Menor	eS7	TcCLB.506829.39	1	40S ribosomal protein S7	-	-
Menor	uS8	TcCLB.509381.20 / TcCLB.508041.30	2	40S ribosomal protein S15A	-	-
Menor	eS8	TcCLB.511069.10 / TcCLB.511069.20	2	40S ribosomal protein S8	-	-
Menor	uS9	TcCLB.503899.20 / TcCLB.503899.30	2	40S ribosomal protein S16	-	-
Menor	uS10	TcCLB.508823.120 / TcCLB.508823.140	2	40S ribosomal protein S20	-	-
Menor	eS10	TcCLB.506679.150 / TcCLB.506679.140	2	40S ribosomal protein S10	-	-
Menor	uS11	TcCLB.506945.230 / TcCLB.409117.20	2	40S ribosomal protein S14	-	-
Menor	uS12	TcCLB.504181.30 / TcCLB.504181.20	2	40S ribosomal protein S23	-	-
Menor	eS12	TcCLB.508551.20 / TcCLB.506181.59	2	40S ribosomal protein S12	-	-
Menor	uS13	TcCLB.506679.94 / TcCLB.506679.100	2	40S ribosomal protein S18	-	-
Menor	uS14	TcCLB.506025.14 / TcCLB.509201.15 / TcCLB.511805.15	3	40S ribosomal protein S29	-	-
Menor	uS15	TcCLB.510029.70 / TcCLB.511189.30	2	40S ribosomal protein S13	-	-
Menor	uS17	TcCLB.507837.50 / TcCLB.511139.20	2	40S ribosomal protein S11	-	-
Menor	eS17	TcCLB.508827.70 / TcCLB.508827.79	2	40S ribosomal protein S17	-	-
Menor	uS19	TcCLB.511809.99 / TcCLB.511809.130 / TcCLB.511811.10	3	40S ribosomal protein S15	-	-
Menor	eS19	TcCLB.510879.20	1	40S ribosomal protein S19	-	-

Menor	eS21	TcCLB.510101.430 / TcCLB.510101.420	2	40S ribosomal protein S21	Extensión C-terminal de 164 a.a	-
Menor	eS24	TcCLB.507681.160 / TcCLB.507681.150	2	40S ribosomal protein S24	-	-
Menor	eS25	TcCLB.509233.190	1	40S ribosomal protein S25	-	-
Menor	eS26	TcCLB.503801.20	1	40S ribosomal protein S26	-	-
Menor	eS27	TcCLB.506963.14	1	40S ribosomal protein S27	-	-
Menor	eS28	TcCLB.506413.20 / TcCLB.506413.30	2	40S ribosomal protein S33	-	-
Menor	eS30	TcCLB.507019.86 / TcCLB.507019.83	2	40S ribosomal protein S30	-	-
Menor	eS31	TcCLB.510409.39 / TcCLB.510293.40	2	40S ubiquitin/ribosomal protein S27a	-	-
Menor	RACK1	TcCLB.511211.120 / TcCLB.511211.130	2	RACK1	-	-

3.5.1.2. Proteínas Trans-sialidasas

Las proteínas TS fueron identificadas según el trabajo previo del grupo de Bartholomeu (Freitas et al., 2011), en donde definen 8 grupos basados en agrupamientos a nivel de secuencia de 505 miembros, de los cuales 174 pertenecen a la cepa CL Brener Esmeraldo-Like en la cual se realizó el presente análisis. Además, se incluyeron miembros que no pertenecían a ningún agrupamiento definido por Bartholomeu, pero que no estaban descritas como pseudogenes.

3.5.2. Determinación y caracterización de regiones UTR

Las secuencias de las regiones UTR fueron obtenidas en el capítulo anterior (ver sección 3.4.14.2).

En epimastigotas la mediana de las regiones 5’ UTR fue de ~70 nucleótidos mientras que para las regiones 3’ UTR fue de ~250 (70 y 270 en tripomastigota sanguíneo). Esta observación indica que las regiones UTR de *T. cruzi* son menores que las informadas en *T. brucei* donde la mediana reportada es de ~90 nucleótidos (sin incluir secuencia SL) para las regiones 5’ UTR y 400 las 3’ UTR (Kolev et al., 2010; Nilsson et al., 2010; Siegel et al., 2010). Para determinar si esta diferencia era real o si principalmente estaba basada en diferencias a nivel de metodologías, utilizamos UTRme con los datos generados en (Kolev

et al., 2010), para analizar las regiones UTR de *T. brucei*. Las regiones 5' UTR obtenidas fueron de ~80 nucleótidos, mientras que las 3' UTR fueron de ~440. Si bien se observan diferencias en los resultados obtenidos, estos sugieren que las regiones UTR de *T. brucei* son levemente más grandes que las de *T. cruzi.*

Para los grupos proteicos en los que se hace foco en este capítulo, se obtuvieron datos de regiones UTR para 112 PR y 198 TS. En el caso particular de las PR, donde también se excluyeron pseudogenes, se observó que sus regiones UTR son más chicas que la media (Tabla 3.21). De hecho, ya se ha descrito que los genes que codifican las proteínas ribosomales poseen regiones 5'UTR muy cortas (22 nt en *T. brucei*, (Antwi et al., 2016)) y se postula que de esta forma se evitaría la presencia de reguladores negativos en esta región (Greif et al., 2013; Jensen et al., 2014). Interesantemente, las vías que controlan la síntesis de proteínas ribosomales descriptas para mamíferos dependen de las secuencias TOP ubicadas en la región 5' UTR. Los miembros de la familia de ARNm TOP se caracterizan por varias características estructurales: un residuo de C invariable en el inicio de la región 5' UTR, seguido de un tramo ininterrumpido de 4 a 15 pirimidinas; una proporción similar de residuos C y U en el tramo de pirimidina de la mayoría de los miembros; una región rica en GC inmediatamente después del motivo 5' TOP (Meyuhas and Kahan, 2015). Sin embargo, hasta el momento no se ha reportado la conservación del motivo para organismos modelos como *C. elegans* y levadura (Meyuhas and Kahan, 2015) y se ha sugerido que no estaría presente en tripanosomátidos (Parsons and Myler, 2016).

Se observa el caso opuesto para TS, observándose un aumento notorio del tamaño de las regiones 3' UTR (Tabla 3.21), esto posibilitaría la existencia de motivos regulatorios a nivel de secuencia primaria y secundaria. Las regiones UTR de TS son difíciles de determinar ya que estas regiones repetitivas sufren de problemas de ensamblado y colapsado de secuencias. Se han determinado las regiones 3' UTR de algunas decenas de TS y los valores reportados fueron de ~700 para proteínas expresadas en el estadio tripomastigota y ~400 para las expresados en epimastigotas (Jager et al., 2008).

Tabla 3.21. Resumen de tamaño y número de secuencias UTR analizadas del grupo de proteínas ribosomales (PR) y la superfamilia de trans-sialidasas (TS) para los estadios epimastigota y tripomastiogta sanguíneos. Las regiones UTR fueron obtenidas mediante el programa UTRme utilizando datos provenientes de (Li et al., 2016).

	Mediana largo - 5	Mediana largo - 3	5 #	3 #
Epimastigota - PR	21	103	94	109
Epimastigota - TS	100	973	164	97
Tripomastigota - PR	21	96	91	109
Tripopamstigota- TS	97	862	164	158

Para las UTRs obtenidas, nos preguntamos si la composición nucleotídica presenta sesgos composicionales. Procedimos a comparar las regiones UTR de ambos grupos de genes contra un conjunto de UTRs al azar, estableciendo en ambos casos largos de UTR próximos a los valores medios exhibidos en Tabla 3.21. Los análisis aquí mostrados corresponden a los estadios donde se observa mayor eficiencia traduccional, epimastigota y tripomastigota para PR y TS respectivamente. Examinando la Figura 3.43, se puede observar como para las PR la región 5’ (A) no difiere significativamente con las regiones al azar (B) observándose en ambos casos una subrepresentación de citosinas y descartándose claramente la presencia de motivos TOP en estas regiones. La región 3’ (C) presenta un claro enriquecimiento en Timina en la región posterior con respecto al azar (D). También se observa aquí la menor cantidad de citosina con respecto al resto de las bases nucleotídicas.

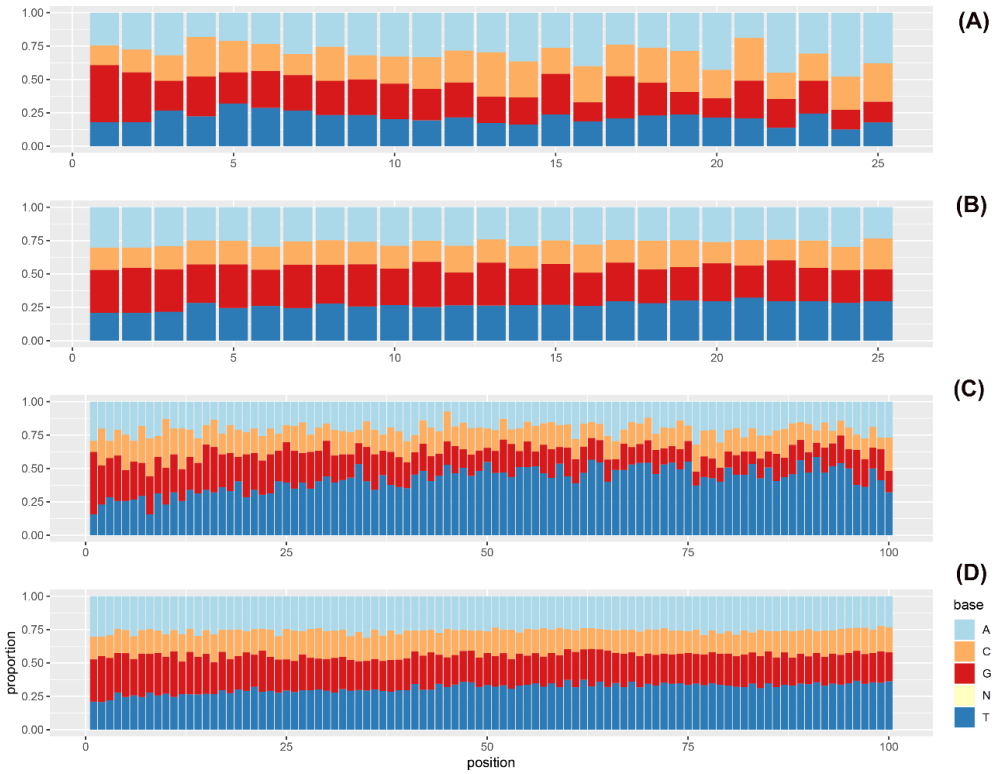


Figura 3.43. Análisis de composición de bases de regiones UTR. A) Composición de las 25 primeras bases de 94 regiones 5' UTR de PR detectadas en el estadio epimastigota. B) Composición 25 bases de 94 regiones 5' UTR obtenidas al azar. C) Composición de las 100 primeras bases de 109 regiones 3' UTR de PR detectadas en el estadio epimastigota. D) Composición de 100 bases de 109 regiones 3' UTR obtenidas al azar.

De forma análoga se estudiaron las regiones UTRs de TS. La Figura 3.44, muestra un resultado distinto a lo obtenido para PR. Tanto las regiones 5' como 3' (A) y (C) evidencian la existencia de sesgos conservados en estas regiones. En la región 5', en la proximidad al AUG del CDS principal, existe una región rica en CA que podría tener roles regulatorios (Duhagon et al., 2001; Pastro et al., 2013). La región 3' UTR presenta una casi depleción de timina (uracilo) al inicio de la región seguido por un tracto muy rico en esta base hasta la posición 250 aproximadamente. También se puede observar como a partir de aproximadamente la posición 500 surge un patrón que se asemeja al exhibido en regiones 3' UTR seleccionadas al azar.

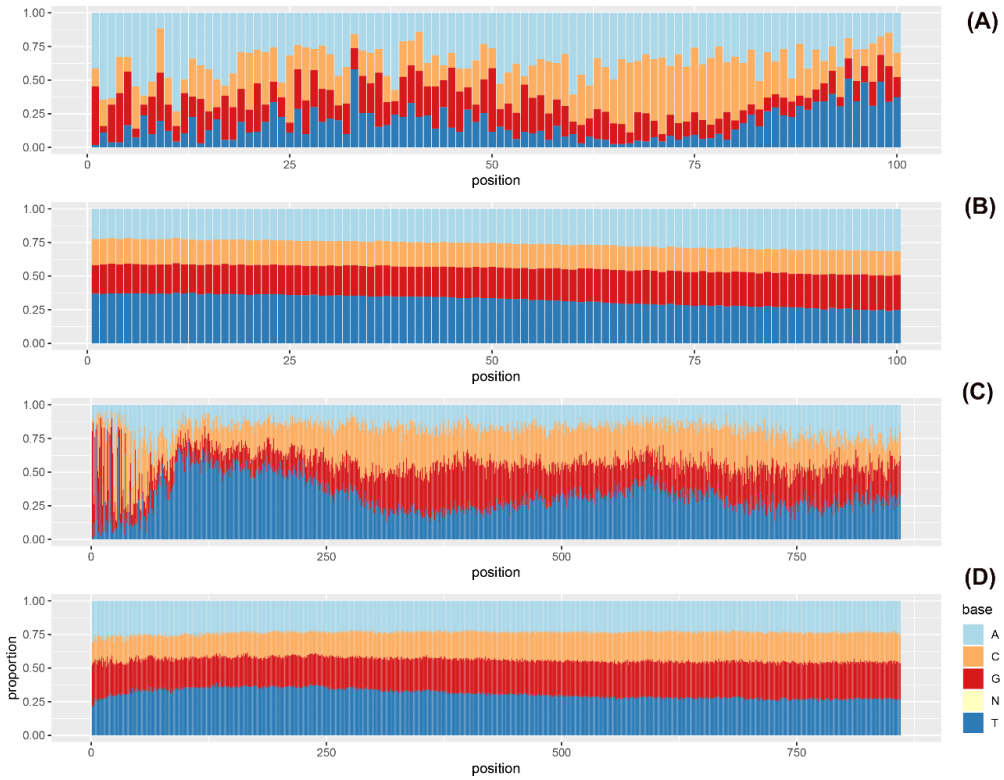


Figura 3.44. Análisis de composición de bases de regiones UTR. A) Composición de las 100 primeras bases de 164 regiones 5' UTR de TS detectadas en el estadio tripomastigota sanguíneo. B) Composición 100 bases de 164 regiones 5' UTR obtenidas al azar. C) Composición de las 860 primeras bases de 158 regiones 3' UTR de TS detectadas en el estadio tripomastigota sanguíneo. D) Composición de 860 bases de 158 regiones 3' UTR obtenidas al azar.

Los patrones no al azar de ocurrencia de bases sugieren la existencia de motivos en las regiones UTR de los grupos analizados.

3.5.3. Análisis de co-modulación de los diferentes miembros de la familia trans-sialidasas y proteínas ribosomales

Para establecer si existen motivos comunes en las regiones UTR que influencien los niveles de estado estacionario, traducción y por ende eficiencia traduccional de genes co-regulados, procedimos a generar agrupamientos de proteínas a partir de los datos de conteos obtenidos en el capítulo 3.1. Para las proteínas ribosomales se establecieron 3 grupos (material suplementario) que interesantemente dependen del tipo de estudio y no del estadio (Figura 3.45). Esto evidencia que los patrones de huellas ribosomales son distintivos, diferenciándose significativamente de los de estado estacionario de ARNm.

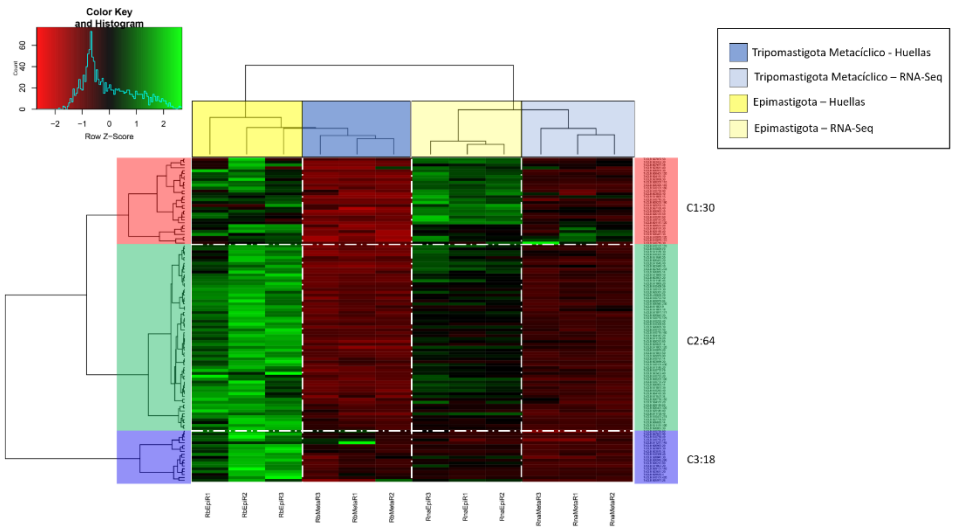


Figura 3.45. Mapa de calor que muestra la variación de expresión a nivel de ARNm (RNA-Seq) y traducción (huellas) para los genes codificantes de proteínas ribosomales a partir de datos provenientes de (Smircich et al., 2015). Con líneas punteadas se marca la separación entre los distintos agrupamientos en las distintas muestras. En total se establecieron 3 agrupamientos, C1 (en rojo) tiene 30 miembros; C2 (en verde) 64 y C3 (en azul) 18 miembros.

Para poder obtener una visión más clara de cómo se comporta cada uno de los *cluster* con respecto a los niveles de estado estacionario/traducción, calculamos el medioide de cada uno de ellos y graficamos de forma independiente (Figura 3.46).

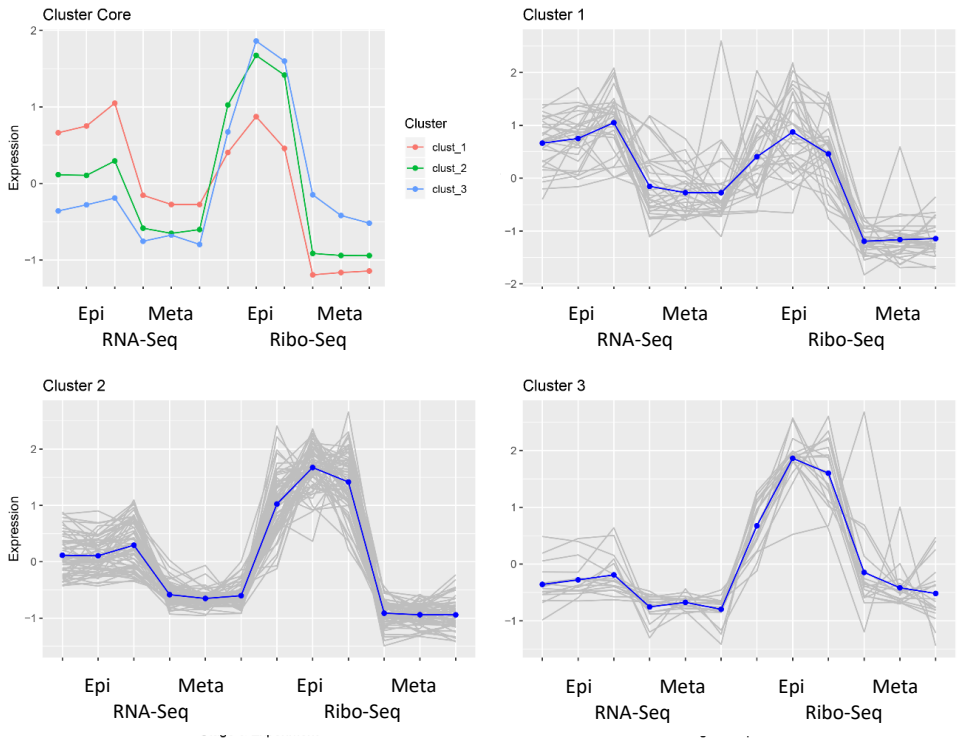


Figura 3.46. Arreglos individuales de niveles de expresión de los agrupamientos de proteínas ribosomales. Las líneas azules representan la expresión del medioide del *cluster*, mientras que las grises corresponden al resto de los integrantes.

En ambos estadios C1 exhibe altos niveles de estado estacionario y bajos de traducción implicando una baja TE (calculados en el capítulo 3.1); menor que el resto de los *cluster* (material suplementario). El agrupamiento C3 presenta los valores más altos de huellas ribosomales, determinando la mayor eficiencia traduccional en el estadio tripomastigota metacíclico. La mayor parte de las proteínas de este agrupamiento logran escapar del fenómeno general de apagado traduccional. Finalmente, el *cluster* 2 presenta una situación intermedia.

Semejantemente, se establecieron 3 agrupamientos para las proteínas TS (material suplementario), los cuales también se agrupan por estudio (Figura 3.47).

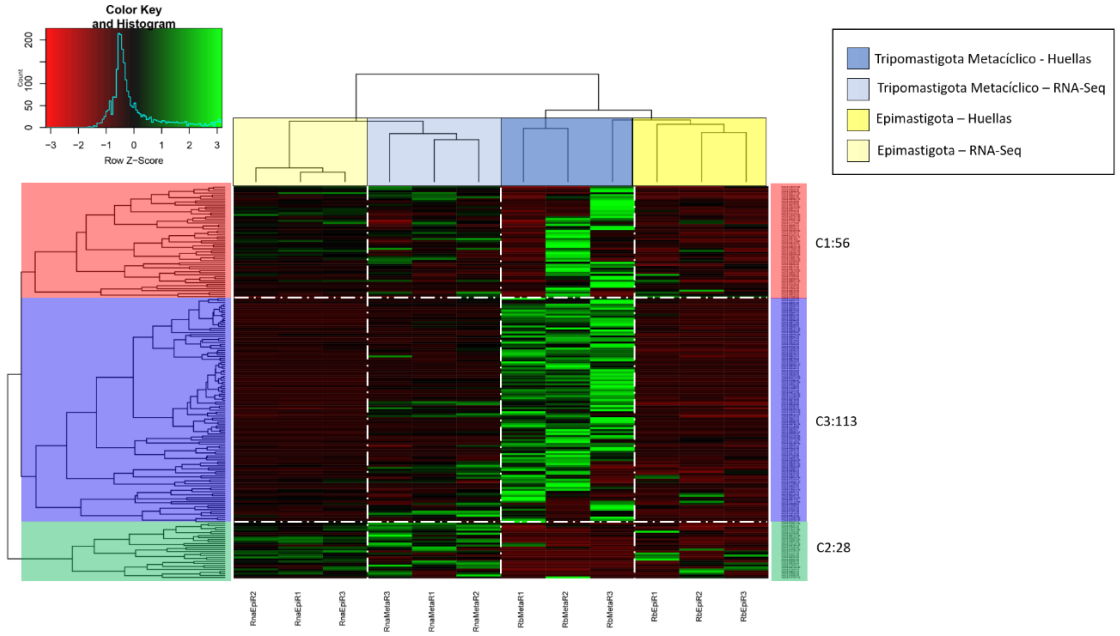


Figura 3.47. Mapa de calor que muestra la variación de expresión a nivel de ARNm (RNA-Seq) y traducción (huellas) para los genes codificantes de proteínas de la superfamilia trans-sialidasa a partir de datos provenientes de (Smircich et al., 2015). Con líneas punteadas se marca la separación entre los distintos agrupamientos en las distintas muestras. En total se establecieron 3 agrupamientos, C1 (en rojo) tiene 56 miembros; C2 (en verde) 28; C3 (en azul) 113 miembros.

Análogamente, calculamos los medioides de cada *cluster* (Figura 3.48). Aquí se evidencian dos comportamientos distintos. Los agrupamientos C1 y C3 presentan elevada actividad traduccional en el estadio metacíclico como es esperado. Sin embargo, el agrupamiento C2 presenta mayor actividad en el estadio epimastigota, y casi nula en tripomastigotas metacíclicos siendo posiblemente miembros de la familia característicos del estadio epimastigota. Llamativamente no logramos observar una correlación con los grupos de TS descritos por (Freitas et al., 2011).

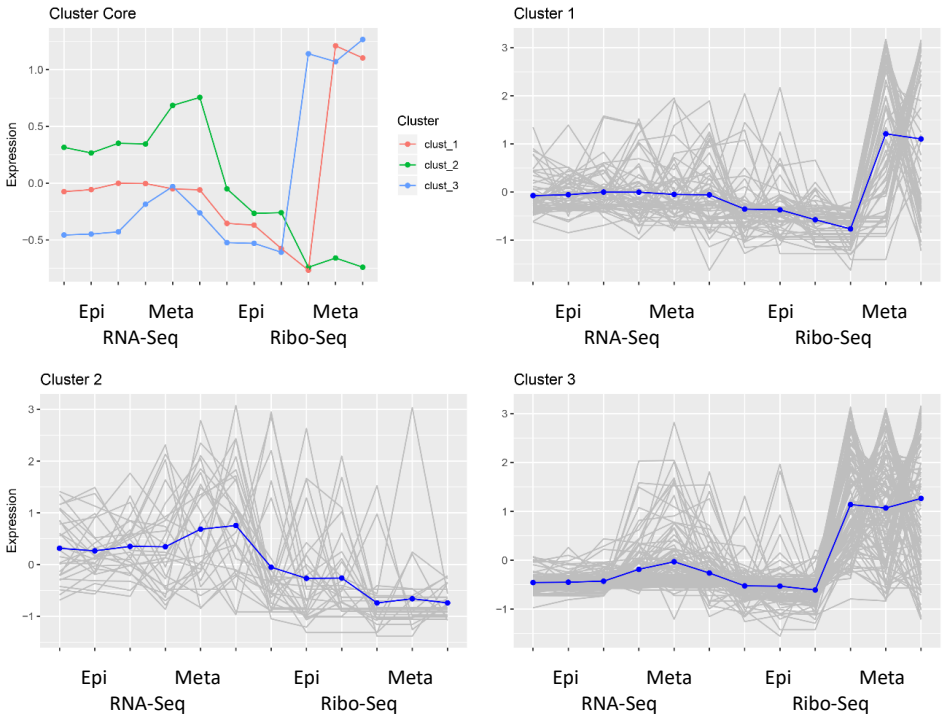


Figura 3.48. Arreglos individuales de la expresión de los agrupamientos de proteínas TS. Las líneas azules representan la expresión del medioide del *cluster*, mientras que las grises corresponden al resto de los integrantes.

La información de los agrupamientos generados, asociados a sus valores de eficiencia traduccional se muestran en el material suplementario.

3.5.4. Búsqueda de motivos lineales y de estructura secundaria de genes co-regulados.

Los motivos de estructura primaria y secundaria fueron determinados mediante la utilización de MEME (Bailey et al., 2015) y BEAM (Pietrosanto et al., 2016), respectivamente (ver sección 3.5.5 Estrategia), obteniéndose resultados equivalentes a los obtenidos durante una pasantía en el laboratorio de Dr. De Gaudenzi utilizando algoritmos similares. La búsqueda de motivos se realizó tanto para todos los miembros de PR y TS de forma conjunta como para cada uno de los agrupamientos predichos de forma individual. Si bien se utilizaron las regiones UTR de ambos estadios, las diferencias no fueron significativas, presentándose únicamente los análisis realizados en los estadios donde se observa mayor expresión (epimastigota y tripomastigota para las PR Y TS, respectivamente). Es importante destacar que los motivos lineales son difíciles de definir, especialmente en genomas como los que analizamos donde se observa un alto

porcentaje de repetidos y las regiones intergénicas contienen elementos ricos en pirimidina (El-Sayed et al., 2005a; De Gaudenzi et al., 2013).

3.5.4.1. Proteínas ribosomales

Los análisis realizados en las regiones 5' UTR de las PR no revelaron motivos como preveíamos dada la falta de sesgos descrita en la sección anterior (3.5.2). Interesantemente, el tamaño medio de las 5' UTRs pertenecientes al agrupamiento C3 fue de 16 nucleótidos en comparación con los 24 de C1 y C2. Si bien no hemos determinado la significancia de esta situación, resulta intrigante ya que gran parte de los miembros de C3 escapan la represión traduccional en tripomastigotas metacíclicos.

En la región 3' UTR la situación es opuesta. El agrupamiento C3 presenta 3' UTR mayores (126) que C1 y C2 (102 y 96 respectivamente). El análisis de motivos lineales en los tres agrupamientos reveló la presencia de un motivo rico en U (Figura 3.49). En particular observamos enriquecido el motivo UUUXUUU (siendo X cualquier base) con un valor esperado de 3e-012, identificado por MEME en 101 de 109 PR. Mediante el programa MEGA (Kumar et al., 2018), inspeccionamos y editamos (ver sección 3.5.5 Estrategia) manualmente las secuencias 3' UTRs de las PR determinando el número de motivos UUUVUUU (V representa las bases A, C y G). En 52 de 109 PR se observa el motivo UUUGUUU, en 38 UUUCUUU y EN 44 UUUAUUU (Tabla 3.22).

Tabla 3.22. Presencia de motivos UUUVUUU en las regiones 3' UTR de proteínas ribosomales de *T. cruzi*.

	Presente	Ausente
UUUGUUU	52	57
UUUCUUU	38	71
UUUAUUU	44	65

El motivo más representado fue UUUGUUU, una versión similar (UUGUU) fue identificada por De Gaudenzi (De Gaudenzi et al., 2013), en donde se discute la presencia del mismo motivo en las PR de Nemátodos (Hajarnavis and Durbin, 2006). Además, los motivos ARE (*AU rich elements*) y los tractos poli(U) hacen que las regiones 3' UTR de las PR sean blancos posibles de regulación por unión de RBPs (Noe et al., 2008). Interesantemente,

el motivo UAUUUUUU reconocido por la previamente mencionada RBP10 de *T. brucei* (la cual interacciona con proteínas ribosomales) se encuentra en 14 PR (material suplementario).

Mediante el programa FIMO de MEME-SUITE determinamos si este motivo esta enriquecido en las regiones 3' UTR de PR con respecto a todas las regiones 3' UTR de *T. cruzi*. Si bien encontramos que el motivo se encuentra enriquecido, confirmamos la observación realizada en (De Gaudenzi et al., 2013) que se trata de un motivo que se encuentra presente en un gran porcentaje de genes. Interesantemente, la Figura 3.43 sugiere que los tractos de U se ubican preferentemente en la región más distal del codón de parada del CDS principal hecho que pudimos confirmar (Figura 3.49).

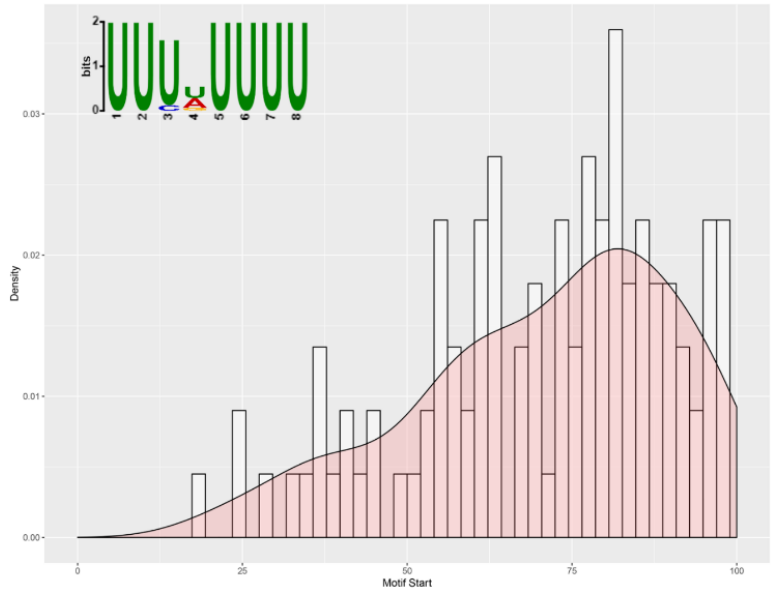


Figura 3.49. Gráfico de densidad e histograma que muestra la posición relativa de motivo UUUXUUU con respecto a la secuencia UTR de los genes codificantes para las proteínas ribosomales de *T. cruzi*.

Esta observación es interesante ya que en tripanosomátidos los motivos de regulación de las regiones 3' UTR suele posicionarse en las regiones próximas al codón de parada (De Gaudenzi et al., 2013).

Dado que las regiones 3’ UTR del agrupamiento C3 son de mayor tamaño, nos preguntamos si existen diferencias con respecto a la ubicación del motivo dentro de la región UTR. La Figura 3.50 sugiere que el motivo en los miembros del agrupamiento C3 se ubican en regiones más próximas al CDS que C1 y C2. La implicancia de este fenómeno aún no ha sido determinada.

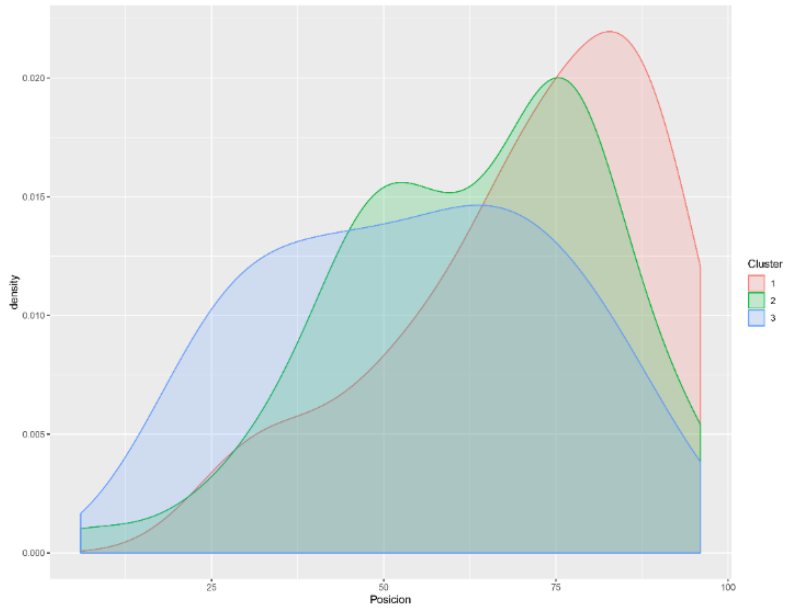


Figura 3.50. Posición relativa del motivo UUUXUUU en las proteínas ribosomales de *T. cruzi*, individualizada por agrupamiento.

3.5.4.2. Proteínas trans-sialidasas

Las 5’ UTR de TS no revelaron diferencias significativas a nivel de secuencia primaria que expliquen los agrupamientos observados. Sin embargo, se pudo observar un alto nivel de conservación entre los miembros de cada agrupamiento. Para caracterizar las regiones conservadas, se alinearon todas las secuencias 5’ UTR de TS mediante ClustalW (Larkin et al., 2007) y se extrajeron los bloques conservados con MEGA. Posteriormente, se realinearon las secuencias con T-Coffee y se generó el logo de secuencia con WebLogo 3. El logo representa 110 pb (Figura 3.51 A),cubriendo en totalidad la región 5’ de muchas TS (Tabla 3.21), y confirma las observaciones realizadas previamente (Figura 3.44). En particular se evidencia una región próxima al codón AUG del CDS principal rica en AU (elemento ARE) muy conservada y una región inmediatamente anterior poli(CA). Además,

se observan otras dos regiones muy conservadas, una entre las bases 21-41 (región A) y la otra entre 55-70 (región B).

Los resultados obtenidos a nivel de estructura secundaria muestran en los agrupamientos C1 y C3 la formación de un pequeño tallo en la región A previamente remarcada (Figura 3.51 B y C). La estructura presente en al agrupamiento C1 se encuentra en el 65% de las secuencias mientras que la estructura de C3 en el 40%, ambas con un p-valor > 0.001 reportado por BEAM. En particular, el agrupamiento C3 presentó una pequeña extensión con respecto a la estructura reportada para C1 que se evidencia en la Figura 3.51 A como la región enmarcada sin colorear. Interesantemente, el agrupamiento C2, el cual presenta un patrón de expresión distinto al resto de los *cluster* no presentó estructuras secundarias significativas. La significancia de este motivo queda por ser determinada.

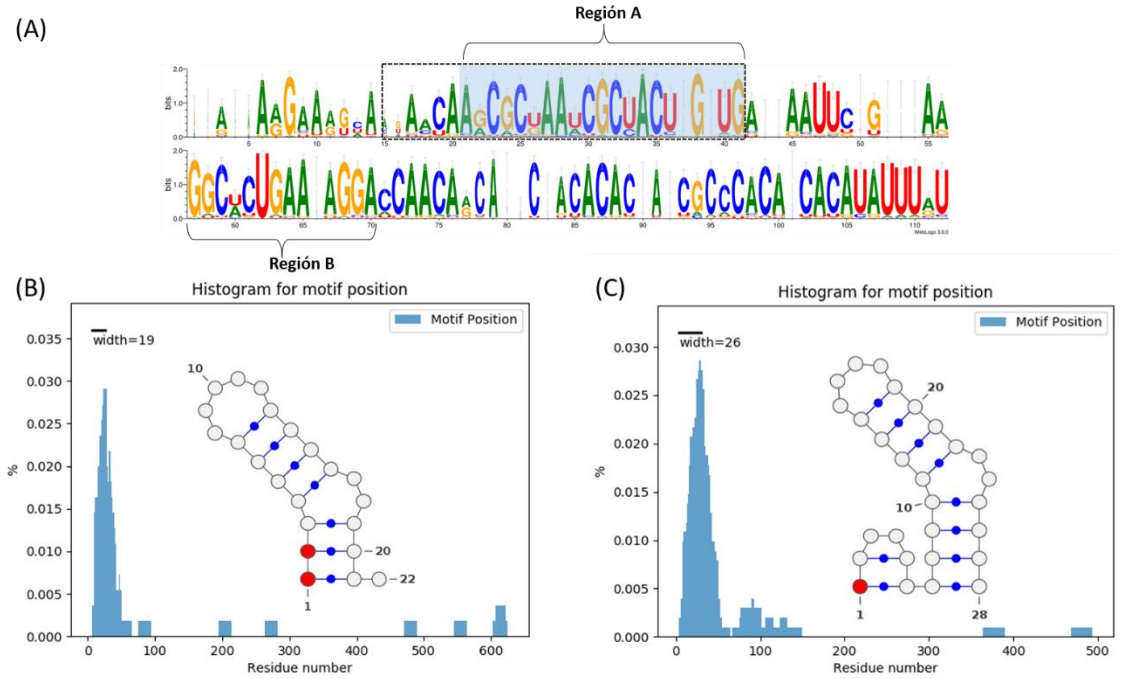


Figura 3.51. Análisis de motivos lineales y de estructura secundaria de las regiones 5' UTR de genes codificantes para proteínas de la superfamilia trans-sialidas. A) Logo de secuencia del bloque conservado entre los miembros de la superfamilia realizado con WebLogo 3. En azul se marca la región de la secuencias que representa la estructura secundaria en B y en punteado la extensión de esta estructura exhibida en C. B) Estructura secundaria conservada entre los miembros de C1, se muestra la estructura y la posición relativa de la estructura. C) Estructura secundaria conservada entre los miembros de C1, se muestra la estructura y la posición relativa de la estructura. Las estructuras secundarias fueron realizadas con BEAM el cual utiliza VaRNA (Darty et al., 2009) para producir la representación gráfica.

Las regiones 3' UTRs de las proteínas TS también son muy conservadas, principalmente en la región próxima al CDS no observándose diferencias significativas entre los distintos grupos. A modo representativo, mostramos los motivos de secuencia primarios observados en el *cluster* 1, Figura 3.52. Se pueden observar 5 motivos (ver sección 3.5.5 Estrategia) 4 de ellos pertenecientes a una misma región (bloque verde, amarillo, violeta y celeste) y uno más próximo al CDS (bloque rojo). Interesantemente el bloque verde contiene una región muy rica en U. Esta región rica en U presente en regiones 3' UTR de TS y otras proteínas (MASP, mucinas, GP63, proteínas quinasas) ya ha sido caracterizada por De Gaudenzi donde reportan una regulación preferencial en el estadio amastigota a través de la interacción con la proteína TcUBP1, mediando también un desestabilización del transcripto en epimastigota (Li et al., 2012).

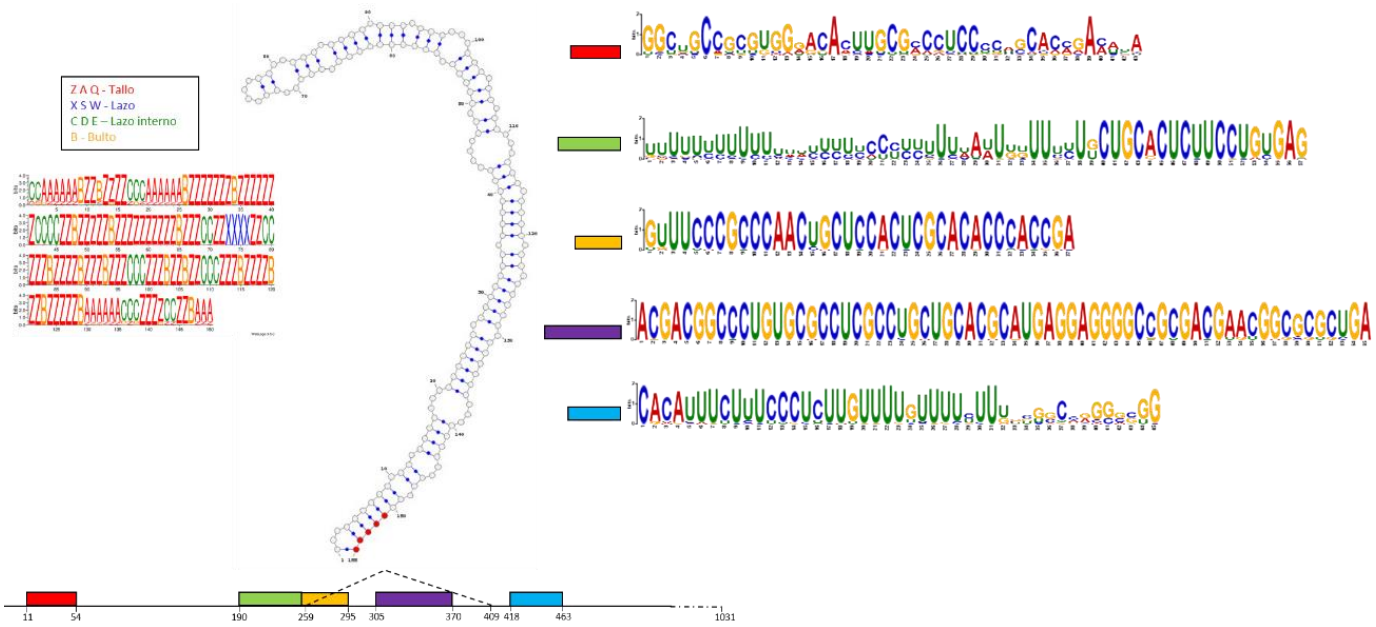


Figura 3.52. Motivos lineales y estructurales sobrerrepresentados en los miembros del cluster C1 de la superfamilia TS. Las líneas punteadas representan la región donde se forma la estructura predicha. Los motivos lineales fueron determinados mediante MEME (p-valor < 0.01) y la estructura secundaria mediante BEAM (p-valor < 0.01).

De forma análoga a lo realizado para las regiones 5', se obtuvieron los bloques conservados de proteínas TS que presentaron un patrón similar al descrito en la Figura 3.52, y se generó un motivo consenso (Figura 3.53). Se observa una gran conservación en la región correspondiente a los bloques amarillo y especialmente el violeta (Figura 3.53). Este último se utilizó para determinar la presencia del motivo en las regiones 3' UTR

predichas para todos los genes expresados en el estadio en tripomastigota. Se utilizó FIMO con un *cutoff* de 1e-9 resultando 264 genes en cuya 3' UTR está presente el motivo. De estos, 161 son TS, 67 son MASP, 9 son mucinas, 14 son GP63 y 8 son SAP (*serine-alanine-and proline-rich protein*) y por lo tanto la gran mayoría corresponden a familias multigénicas de superficie, por lo que constituye un motivo interesante en el cual profundizar. Es posible que este motivo sea parte del motivo rico en U detectado en (Li et al., 2012). La presencia del motivo lineal conservado ya ha sido observada por De Gaudenzi (comunicación personal). Finalmente, se evaluó si el motivo estaba asociado a un agrupamiento particular pero no se obtuvieron resultados concluyentes, pero si una presencia casi total en el agrupamiento 2 (Tabla 3.23).

Tabla 3.23. Presencia del motivo conservado en los agrupamientos de TS.

Cluster	Presente	Ausente
1	34	12
2	21	3
3	51	37

Con respecto a la eficiencia traduccional de los genes que contienen este motivo se encontraron 41 miembros (todas TS) que presentaban diferencialmente mayor eficiencia en tripomastigotas metacíclico (FDR < 0.01 Log2FC < -1) y ninguno aumentado diferencialmente en epimastigota (material suplementario).

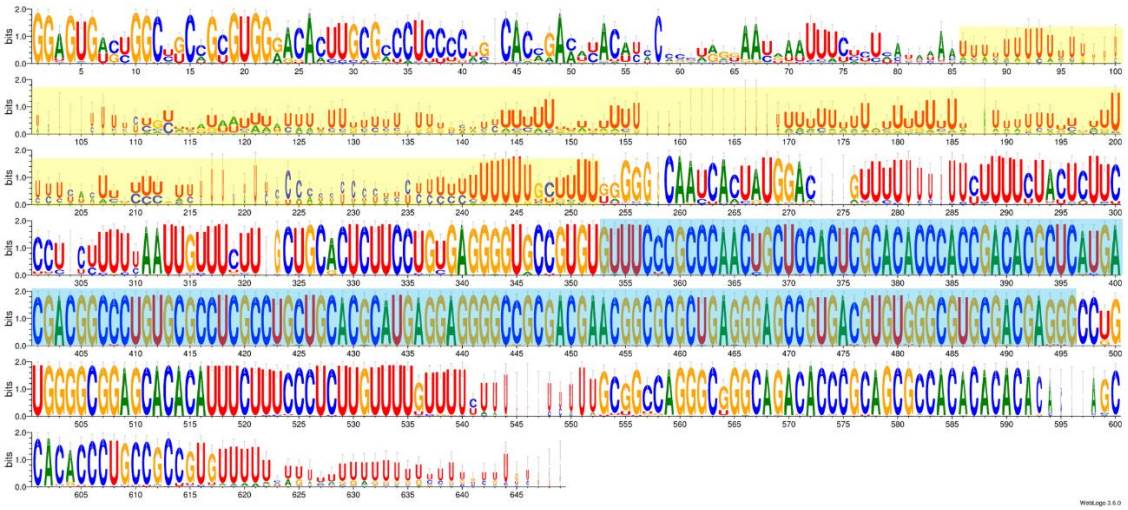


Figura 3.53. Logo de secuencia de regiones 3' UTR de proteína TS que presentan motivos lineales y estructurales conservados. En amarillo se resalta la región rica en U y en azul la región correspondiente a la estructura secundaria predicha (bloques amarillo y violeta). El logo fue realizado con WebLogo 3 (Crooks et al., 2004).

El análisis de estructura secundarias reveló la presencia de un gran tallo muy estructurado de unos 150 pares de base (máximo permitido por BEAM) en la región conservada previamente descrita (Figura 3.52 y Figura 3.53). La conservación del motivo en ambos niveles estudiados lo posiciona como un motivo regulatorio prometedor. Este motivo estructural se ubica próximo al codón de parada del CDS principal (Figura 3.54).

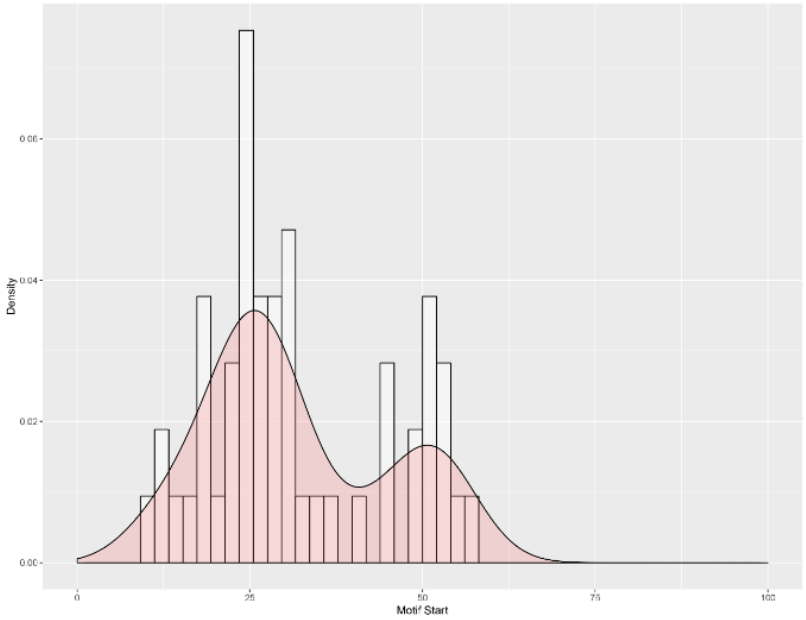


Figura 3.54. Gráfico de densidad e histograma que muestra la posición relativa de motivo de estructura secundaria con respecto a la secuencia UTR de los genes codificantes para la superfamilia de proteínas trans sialidadas de *T. cruzi*.

Los resultados expuestos en este capítulo sugieren la existencia de motivos regulatorios en las regiones no traducidas de las proteínas ribosomales y de la superfamilia trans-sialidasas, para los cuales se planea seguir profundizando a nivel experimental y computacional.

3.5.5. Estrategia

3.5.5.1. Generación de regiones *background*

Para determinar si los motivos encontrados en las regiones UTR son representativos de un conjunto de genes o sí por el contrario surgen simplemente por azar, es necesario tener un conjunto de secuencias control. Se obtuvieron todas las regiones aledañas al

CDS de cada uno de los genes (+- 1000 pb) utilizando TriTrypDB. Para cada agrupamiento, mediante un *script homemade*, se seleccionaron al azar 5000 regiones del tamaño de la mediana de cada agrupamiento. De esa forma se generó una secuencia control por agrupamiento formado.

3.5.5.2. Generación de agrupamientos de genes co-regulados

A partir de los datos de RNA-Seq y Ribo-Seq realizamos agrupamientos jerárquicos. La normalización de los conteos se realizó con el paquete EdgeR. A continuación, escalamos y construimos los árboles (dendrogramas). Dado que estamos comparando patrones de expresión génica, necesitamos escalar los datos, de lo contrario, todos los genes altamente expresados se agruparán, incluso si tienen patrones diferentes entre las muestras. Después, calculamos la distancia entre los genes (mediante el método de Spearman) y las muestras como 1 menos la correlación de un gen/muestra con otro (que tan diferente se comporta un gen de otro en cada muestra). Esta distancia es utilizada por realizar el agrupamiento jerárquico y construir los dendrogramas para los genes y las muestras.

3.5.5.3. Determinación de motivos lineales

Los motivos lineales fueron determinados mediante la utilización de la herramienta MEME dentro del paquete MEME-suite (versión 5.0.2). Se utilizaron parámetros por defecto y las secuencias control generadas anteriormente. Se definió en el caso de las regiones 5’ UTR la búsqueda de 3 motivos y en las regiones 3’ UTR 5 motivos. El tamaño mínimo del motivo fue establecido en 6 y el máximo en 150. El resto de los parámetros fueron establecidos por defecto.

3.5.5.4. Determinación de motivos de estructura secundaria

Para determinar la presencia de motivos de estructura secundaria en las regiones UTR utilizamos BEAM (BEArMotif finder). BEAM es un método que explora conjuntos de ARN no alineados que comparten una propiedad biológica buscando los mejores motivos de

estructura secundaria local y evaluando su importancia con respecto a un conjunto de regiones al azar. En particular, seleccionamos el algoritmo RNAFold de paquete Vienna RNA 2.0 (Lorenz et al., 2011) que determina el mínimo de energía libre para la estructura secundaria de un transcripto o región del mismo. BEAM utiliza VaRNA para visualizar la estructura secundaria de los motivos encontrados. Se definió en el caso de las regiones 5' UTR la búsqueda de 5 motivos y en las regiones 3' UTR 10 motivos. El tamaño mínimo del motivo fue establecido en 5 y el máximo en 150. El resto de los parámetros fueron establecidos por defecto.

3.5.5.5. Determinación de tractos UUUXUUU en las regiones 3' UTR de PR

Utilizando el programa MEGA editamos la secuencia en busca de tractos poli(U). Un tracto poli(U) fue definido como una secuencia de al menos 6 uridinas interrumpidas por, como máximo, 1 base. Además, en caso de existir interrupción por una base, esta debe tener al menos 3 U de cada lado. Los resultados de esta búsqueda se presentan en material suplementario.

4. Conclusiones

I. Diseñamos una metodología capaz de mejorar sensiblemente el análisis de eficiencia traduccional diferencial. La metodología fue aplicada para reanalizar datos de *Ribosome Profiling* (Smircich et al. 2015), confirmando los resultados obtenidos previamente y mejorando ampliamente el análisis de los mismos revelando nuevas proteínas y procesos diferenciales (objetivo específico 1).

II. Desarrollamos dos herramientas de interfaz gráfica que facilitan el análisis de listas de genes producto de análisis de expresión diferencial (objetivo específico 2). DARK representa una mejora ostensible en la anotación de proteínas de tripanosomátidos y consideramos que será una herramienta de gran utilidad para los investigadores de estos organismos. Actualmente, nos encontramos escribiendo un manuscrito con estos resultados. Por otra parte, IdMiner permite retratar rápidamente los términos biológicos que se encuentran con más frecuencia en la literatura en relación con los genes de interés, y ofrece una manera rápida e intuitiva de obtener estos documentos relevantes. IdMiner es organismo independiente, por lo que los investigadores de cualquier área lo pueden utilizar. Contamos con un borrador asociado a la herramienta que se presenta en la tesis.

III. Definimos las regiones UTR de *T. cruzi* a través del desarrollo de UTRme. UTRme facilita la identificación de sitios de procesamiento a partir de datos RNA-Seq e informa sobre su confianza asociada. UTRme puede ser utilizado en otros organismos (como mostramos para *E. granulosus*). UTRme cuenta con interfaz gráfica y no requiere experiencia previa en el análisis de datos RNA-seq. Los resultados de este capítulo fueron publicados en el *journal Frontiers in Genetics* (Radio et al., 2018), (objetivo específico 3).

IV. Obtuvimos resultados que sugieren fuertemente que la presencia de uORFs en los ARNm de *T. cruzi* reducen su eficiencia traduccional, con lo que estos

elementos podrían actuar como un mecanismo regulatorio general. El *abstract* de un manuscrito conteniendo estos resultados ha sido aceptado para ser considerado en la publicación *Frontiers in Genetics* (objetivo específico 4).

V. Identificamos motivos regulatorios en las regiones no traducidas de las proteínas ribosomales y de la superfamilia trans-sialidasas, para los cuales se planea seguir profundizando su caracterización a nivel experimental y computacional (objetivo específico 5).

Por lo anteriormente expuesto los resultados de esta tesis permiten profundizar en los conocimientos de la regulación de la expresión génica en *T. cruzi* y generan herramientas de uso general para contestar estas y otras interrogantes acerca de la biología del parásito.

5. Bibliografía

Agabian, N. (1990). Trans splicing of nuclear pre-mRNAs. *Cell* 61(7), 1157-1160.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. *J Mol Biol* 215(3), 403-410. doi: 10.1016/S0022-2836(05)80360-2.

Alvarez, P., Buscaglia, C.A., and Campetella, O. (2004). Improving protein pharmacokinetics by genetic fusion to simple amino acid sequences. *J Biol Chem* 279(5), 3375-3381. doi: 10.1074/jbc.M311356200.

Amorim, J.C., Batista, M., da Cunha, E.S., Lucena, A.C.R., Lima, C.V.P., Sousa, K., et al. (2017). Quantitative proteome and phosphoproteome analyses highlight the adherent population during Trypanosoma cruzi metacyclogenesis. *Sci Rep* 7(1), 9899. doi: 10.1038/s41598-017-10292-3.

Andreeva, A., Howorth, D., Chothia, C., Kulesha, E., and Murzin, A.G. (2014). SCOP2 prototype: a new approach to protein structure mining. *Nucleic Acids Res* 42(Database issue), D310-314. doi: 10.1093/nar/gkt1242.

Antwi, E.B., Haanstra, J.R., Ramasamy, G., Jensen, B., Droll, D., Rojas, F., et al. (2016). Integrative analysis of the Trypanosoma brucei gene expression cascade predicts differential regulation of mRNA processing and unusual control of ribosomal protein expression. *BMC Genomics* 17, 306. doi: 10.1186/s12864-016-2624-3.

Aphasizhev, R., Aphasizheva, I., Nelson, R.E., Gao, G., Simpson, A.M., Kang, X., et al. (2003). Isolation of a U-insertion/deletion editing complex from Leishmania tarentolae mitochondria. *EMBO J* 22(4), 913-924. doi: 10.1093/emboj/cdg083.

Aslett, M., Aurrecoechea, C., Berriman, M., Brestelli, J., Brunk, B.P., Carrington, M., et al. (2010). TriTrypDB: a functional genomic resource for the Trypanosomatidae. *Nucleic Acids Res* 38(Database issue), D457-462. doi: 10.1093/nar/gkp851.

Ayub, M.J., Atwood, J., Nuccio, A., Tarleton, R., and Levin, M.J. (2009). Proteomic analysis of the Trypanosoma cruzi ribosomal proteins. *Biochem Biophys Res Commun* 382(1), 30-34. doi: 10.1016/j.bbrc.2009.02.095.

Bailey, T.L., Johnson, J., Grant, C.E., and Noble, W.S. (2015). The MEME Suite. *Nucleic Acids Res* 43(W1), W39-49. doi: 10.1093/nar/gkv416.

Bairoch, A., and Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res* 28(1), 45-48.

Ban, N., Beckmann, R., Cate, J.H., Dinman, J.D., Dragon, F., Ellis, S.R., et al. (2014). A new system for naming ribosomal proteins. *Curr Opin Struct Biol* 24, 165-169. doi: 10.1016/j.sbi.2014.01.002.

Bangs, J.D., Crain, P.F., Hashizume, T., McCloskey, J.A., and Boothroyd, J.C. (1992). Mass spectrometry of mRNA cap 4 from trypanosomatids reveals two novel nucleosides. *J Biol Chem* 267(14), 9805-9815.

Beaupere, C., Chen, R.B., Pelosi, W., and Labunskyy, V.M. (2017). Genome-wide Quantification of Translation in Budding Yeast by Ribosome Profiling. *J Vis Exp* (130). doi: 10.3791/56820.

Becco, L., Smircich, P., and Garat, B. (2019). Conserved motifs in nuclear genes encoding predicted mitochondrial proteins in Trypanosoma cruzi. *PLoS One* 14(4), e0215160. doi: 10.1371/journal.pone.0215160.

Benne, R., Van den Burg, J., Brakenhoff, J.P., Sloof, P., Van Boom, J.H., and Tromp, M.C. (1986). Major transcript of the frameshifted coxII gene from trypanosome mitochondria contains four nucleotides that are not encoded in the DNA. *Cell* 46(6), 819-826.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., et al. (2000). The Protein Data Bank. *Nucleic Acids Res* 28(1), 235-242.

Berna, L., Chiribao, M.L., Greif, G., Rodriguez, M., Alvarez-Valin, F., and Robello, C. (2017). Transcriptomic analysis reveals metabolic switches and surface remodeling as key processes for stage transition in Trypanosoma cruzi. *PeerJ* 5, e3017. doi: 10.7717/peerj.3017.

Berna, L., Rodriguez, M., Chiribao, M.L., Parodi-Talice, A., Pita, S., Rijo, G., et al. (2018). Expanding an expanded genome: long-read sequencing of Trypanosoma cruzi. *Microb Genom* 4(5). doi: 10.1099/mgen.0.000177.

Bernabo, G., Levy, G., Ziliani, M., Caeiro, L.D., Sanchez, D.O., and Tekiel, V. (2013). TcTASV-C, a protein family in Trypanosoma cruzi that is predominantly trypomastigote-stage specific and secreted to the medium. *PLoS One* 8(7), e71192. doi: 10.1371/journal.pone.0071192.

Blumenthal, T., Evans, D., Link, C.D., Guffanti, A., Lawson, D., Thierry-Mieg, J., et al. (2002). A global analysis of Caenorhabditis elegans operons. *Nature* 417(6891), 851-854. doi: 10.1038/nature00831.

Burgos, J.M., Risso, M.G., Breniere, S.F., Barnabe, C., Campetella, O., and Leguizamon, M.S. (2013). Differential distribution of genes encoding the virulence factor trans-sialidase along Trypanosoma cruzi Discrete typing units. *PLoS One* 8(3), e58967. doi: 10.1371/journal.pone.0058967.

Buscaglia, C.A., Campo, V.A., Frasch, A.C., and Di Noia, J.M. (2006). Trypanosoma cruzi surface mucins: host-dependent coat diversity. *Nat Rev Microbiol* 4(3), 229-236. doi: 10.1038/nrmicro1351.

Buschiazzo, A., Muia, R., Larrieux, N., Pitcovsky, T., Mucci, J., and Campetella, O. (2012). Trypanosoma cruzi trans-sialidase in complex with a neutralizing antibody: structure/function studies towards the rational design of inhibitors. *PLoS Pathog* 8(1), e1002474. doi: 10.1371/journal.ppat.1002474.

Caeiro, L.D., Alba-Soto, C.D., Rizzi, M., Solana, M.E., Rodriguez, G., Chidichimo, A.M., et al. (2018). The protein family TcTASV-C is a novel Trypanosoma cruzi virulence factor secreted in extracellular vesicles by trypomastigotes and highly expressed in bloodstream forms. *PLoS Negl Trop Dis* 12(5), e0006475. doi: 10.1371/journal.pntd.0006475.

Calvo, S.E., Pagliarini, D.J., and Mootha, V.K. (2009). Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. *Proc Natl Acad Sci U S A* 106(18), 7507-7512. doi: 10.1073/pnas.0810916106.

Cazzulo, J.J. (1994). Intermediate metabolism in Trypanosoma cruzi. *J Bioenerg Biomembr* 26(2), 157-165.

Clayton, C. (2013). The regulation of trypanosome gene expression by RNA-binding proteins. *PLoS Pathog* 9(11), e1003680. doi: 10.1371/journal.ppat.1003680.

Clayton, C.E. (2002). Life without transcriptional control? From fly to man and back again. *EMBO J* 21(8), 1881-1888. doi: 10.1093/emboj/21.8.1881.

Clements, J.M., Laz, T.M., and Sherman, F. (1988). Efficiency of translation initiation by non-AUG codons in Saccharomyces cerevisiae. *Mol Cell Biol* 8(10), 4533-4536.

Cohen, M., and Varki, A. (2010). The sialome--far more than the sum of its parts. *OMICS* 14(4), 455-464. doi: 10.1089/omi.2009.0148.

Cosentino, R.O., and Aguero, F. (2012). A simple strain typing assay for Trypanosoma cruzi: discrimination of major evolutionary lineages from a single amplification product. *PLoS Negl Trop Dis* 6(7), e1777. doi: 10.1371/journal.pntd.0001777.

Coughlin, B.C., Teixeira, S.M., Kirchhoff, L.V., and Donelson, J.E. (2000). Amastin mRNA abundance in Trypanosoma cruzi is controlled by a 3'-untranslated region position-dependent cis-element and an untranslated region-binding protein. *J Biol Chem* 275(16), 12051-12060.

Coura, J.R., and Borges-Pereira, J. (2010). Chagas disease: 100 years after its discovery. A systemic review. *Acta tropica* 115(1-2), 5-13.

Cremona, M.L., Sanchez, D.O., Frasch, A.C., and Campetella, O. (1995). A single tyrosine differentiates active and inactive Trypanosoma cruzi trans-sialidases. *Gene* 160(1), 123-128.

Crooks, G.E., Hon, G., Chandonia, J.M., and Brenner, S.E. (2004). WebLogo: a sequence logo generator. *Genome Res* 14(6), 1188-1190. doi: 10.1101/gr.849004.

Chagas, C. (1909). Nova tripanozomiaze humana: estudos sobre a morfolojia e o ciclo evolutivo do Schizotrypanum cruzi n. gen., n. sp., ajente etiolojico de nova entidade morbida do homem. *Memórias do Instituto Oswaldo Cruz* 1(2), 159-218.

Chavez, S., Eastman, G., Smircich, P., Becco, L.L., Oliveira-Rizzo, C., Fort, R., et al. (2017). Transcriptome-wide analysis of the Trypanosoma cruzi proliferative cycle identifies the periodically expressed mRNAs and their multiple levels of control. *PLoS One* 12(11), e0188441. doi: 10.1371/journal.pone.0188441.

Chen, R.A., Down, T.A., Stempor, P., Chen, Q.B., Egelhofer, T.A., Hillier, L.W., et al. (2013). The landscape of RNA polymerase II transcription initiation in C. elegans reveals promoter and enhancer architectures. *Genome Res* 23(8), 1339-1347. doi: 10.1101/gr.153668.112.

Chen, Y.J., Tan, B.C., Cheng, Y.Y., Chen, J.S., and Lee, S.C. (2010). Differential regulation of CHOP translation by phosphorylated eIF4E under stress conditions. *Nucleic Acids Res* 38(3), 764-777. doi: 10.1093/nar/gkp1034.

Chew, G.L., Pauli, A., and Schier, A.F. (2016). Conservation of uORF repressiveness and sequence features in mouse, human and zebrafish. *Nat Commun* 7, 11663. doi: 10.1038/ncomms11663.

Choi, J., and El-Sayed, N.M. (2012). Functional genomics of trypanosomatids. *Parasite Immunol* 34(2-3), 72-79. doi: 10.1111/j.1365-3024.2011.01347.x.

Chung, B.Y., Hardcastle, T.J., Jones, J.D., Irigoyen, N., Firth, A.E., Baulcombe, D.C., et al. (2015). The use of duplex-specific nuclease in ribosome profiling and a user-friendly software package for Ribo-seq data analysis. *RNA* 21(10), 1731-1745. doi: 10.1261/rna.052548.115.

d'Avila-Levy, C.M., Boucinha, C., Kostygov, A., Santos, H.L., Morelli, K.A., Grybchuk-Ieremenko, A., et al. (2015). Exploring the environmental diversity of kinetoplastid flagellates in the high-throughput DNA sequencing era. *Mem Inst Oswaldo Cruz* 110(8), 956-965. doi: 10.1590/0074-02760150253.

D'Orso, I., and Frasch, A.C. (2001). Functionally different AU- and G-rich cis-elements confer developmentally regulated mRNA stability in Trypanosoma cruzi by interaction with specific RNA-binding proteins. *J Biol Chem* 276(19), 15783-15793. doi: 10.1074/jbc.M010959200.

D'Orso, I., and Frasch, A.C. (2002). TcUBP-1, an mRNA destabilizing factor from trypanosomes, homodimerizes and interacts with novel AU-rich element- and Poly(A)-binding proteins forming a ribonucleoprotein complex. *J Biol Chem* 277(52), 50520-50528. doi: 10.1074/jbc.M209092200.

da Silva Augusto, L., Moretti, N.S., Ramos, T.C., de Jesus, T.C., Zhang, M., Castilho, B.A., et al. (2015). A membrane-bound eIF2 alpha kinase located in endosomes is regulated by heme and controls differentiation and ROS levels in Trypanosoma cruzi. *PLoS Pathog* 11(2), e1004618. doi: 10.1371/journal.ppat.1004618.

Dan-Goor, M., Nasereddin, A., Jaber, H., and Jaffe, C.L. (2013). Identification of a secreted casein kinase 1 in Leishmania donovani: effect of protein over expression on parasite growth and virulence. *PLoS One* 8(11), e79287. doi: 10.1371/journal.pone.0079287.

Darty, K., Denise, A., and Ponty, Y. (2009). VARNA: Interactive drawing and editing of the RNA secondary structure. *Bioinformatics* 25(15), 1974-1975. doi: 10.1093/bioinformatics/btp250.

De Gaudenzi, J., Frasch, A.C., and Clayton, C. (2005). RNA-binding domain proteins in Kinetoplastids: a comparative analysis. *Eukaryot Cell* 4(12), 2106-2114. doi: 10.1128/EC.4.12.2106-2114.2005.

De Gaudenzi, J.G., Carmona, S.J., Aguero, F., and Frasch, A.C. (2013). Genome-wide analysis of 3'-untranslated regions supports the existence of post-transcriptional regulons controlling gene expression in trypanosomes. *PeerJ* 1, e118. doi: 10.7717/peerj.118.

De Gaudenzi, J.G., D'Orso, I., and Frasch, A.C. (2003). RNA recognition motif-type RNA-binding proteins in Trypanosoma cruzi form a family involved in the interaction with specific transcripts in vivo. *J Biol Chem* 278(21), 18884-18894. doi: 10.1074/jbc.M301756200.

de Godoy, L.M., Marchini, F.K., Pavoni, D.P., Rampazzo Rde, C., Probst, C.M., Goldenberg, S., et al. (2012). Quantitative proteomics of Trypanosoma cruzi during metacyclogenesis. *Proteomics* 12(17), 2694-2703. doi: 10.1002/pmic.201200078.

de Souza, W., Attias, M., and Rodrigues, J.C. (2009). Particularities of mitochondrial structure in parasitic protists (Apicomplexa and Kinetoplastida). *Int J Biochem Cell Biol* 41(10), 2069-2080. doi: 10.1016/j.biocel.2009.04.007.

Dean, S., Sunter, J.D., and Wheeler, R.J. (2017). TrypTag.org: A Trypanosome Genome-wide Protein Localisation Resource. *Trends Parasitol* 33(2), 80-82. doi: 10.1016/j.pt.2016.10.009.

Docampo, R., de Souza, W., Miranda, K., Rohloff, P., and Moreno, S.N. (2005). Acidocalcisomes - conserved from bacteria to man. *Nat Rev Microbiol* 3(3), 251-261. doi: 10.1038/nrmicro1097.

Docampo, R., and Moreno, S.N. (2011). Acidocalcisomes. *Cell Calcium* 50(2), 113-119. doi: 10.1016/j.ceca.2011.05.012.

DosReis, G.A. (2011). Evasion of immune responses by Trypanosoma cruzi, the etiological agent of Chagas disease. *Braz J Med Biol Res* 44(2), 84-90.

Duhagon, M.A., Dallagiovanna, B., and Garat, B. (2001). Unusual features of poly[dT-dG].[dC-dA] stretches in CDS-flanking regions of Trypanosoma cruzi genome. *Biochem Biophys Res Commun* 287(1), 98-103. doi: 10.1006/bbrc.2001.5545.

Eastman, G., Smircich, P., and Sotelo-Silveira, J.R. (2018). Following Ribosome Footprints to Understand Translation at a Genome Wide Level. *Comput Struct Biotechnol J* 16, 167-176. doi: 10.1016/j.csbj.2018.04.001.

Eddy, S.R. (2011). Accelerated Profile HMM Searches. *PLoS Comput Biol* 7(10), e1002195. doi: 10.1371/journal.pcbi.1002195.

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S.R., Luciani, A., Potter, S.C., et al. (2019). The Pfam protein families database in 2019. *Nucleic Acids Res* 47(D1), D427-D432. doi: 10.1093/nar/gky995.

El-Sayed, N.M., Myler, P.J., Bartholomeu, D.C., Nilsson, D., Aggarwal, G., Tran, A.N., et al. (2005a). The genome sequence of Trypanosoma cruzi, etiologic agent of Chagas disease. *Science* 309(5733), 409-415. doi: 10.1126/science.1112631.

El-Sayed, N.M., Myler, P.J., Blandin, G., Berriman, M., Crabtree, J., Aggarwal, G., et al. (2005b). Comparative genomics of trypanosomatid parasitic protozoa. *Science* 309(5733), 404-409. doi: 10.1126/science.1112181.

Elias, M.C., Marques-Porto, R., Freymuller, E., and Schenkman, S. (2001). Transcription rate modulation through the Trypanosoma cruzi life cycle occurs in parallel with changes in nuclear organisation. *Mol Biochem Parasitol* 112(1), 79-90.

Els, M.C., Laver, W.G., and Air, G.M. (1989). Sialic acid is cleaved from glycoconjugates at the cell surface when influenza virus neuraminidases are expressed from recombinant vaccinia viruses. *Virology* 170(1), 346-351.

Europe, P.M.C.C. (2015). Europe PMC: a full-text literature database for the life sciences and platform for innovation. *Nucleic Acids Res* 43(Database issue), D1042-1048. doi: 10.1093/nar/gku1061.

Fervers, P., Fervers, F., Makalowski, W., and Jakalski, M. (2018). Life cycle adapted upstream open reading frames (uORFs) in Trypanosoma congolense: A post-transcriptional approach to accurate gene regulation. *PLoS One* 13(8), e0201461. doi: 10.1371/journal.pone.0201461.

Fiebig, M., Kelly, S., and Gluenz, E. (2015). Comparative Life Cycle Transcriptomics Revises Leishmania mexicana Genome Annotation and Links a Chromosome Duplication with Parasitism of Vertebrates. *PLoS Pathog* 11(10), e1005186. doi: 10.1371/journal.ppat.1005186.

Fleming, I.M., Paris, Z., Gaston, K.W., Balakrishnan, R., Fredrick, K., Rubio, M.A., et al. (2016). A tRNA methyltransferase paralog is important for ribosome stability and cell division in Trypanosoma brucei. *Sci Rep* 6, 21438. doi: 10.1038/srep21438.

Frasch, A.C. (2000). Functional diversity in the trans-sialidase and mucin families in Trypanosoma cruzi. *Parasitol Today* 16(7), 282-286.

Freire-de-Lima, L., Fonseca, L.M., Oeltmann, T., Mendonca-Previato, L., and Previato, J.O. (2015). The trans-sialidase, the major Trypanosoma cruzi virulence factor: Three decades of studies. *Glycobiology* 25(11), 1142-1149. doi: 10.1093/glycob/cwv057.

Freitas, L.M., dos Santos, S.L., Rodrigues-Luiz, G.F., Mendes, T.A., Rodrigues, T.S., Gazzinelli, R.T., et al. (2011). Genomic analyses, gene expression and antigenic profile of the trans-sialidase superfamily of Trypanosoma cruzi reveal an undetected level of complexity. *PLoS One* 6(10), e25914. doi: 10.1371/journal.pone.0025914.

Fritsch, C., Herrmann, A., Nothnagel, M., Szafranski, K., Huse, K., Schumann, F., et al. (2012). Genome-wide search for novel human uORFs and N-terminal protein extensions using ribosomal footprinting. *Genome Res* 22(11), 2208-2218. doi: 10.1101/gr.139568.112.

Furger, A., Schurch, N., Kurath, U., and Roditi, I. (1997). Elements in the 3' untranslated region of procyclin mRNA regulate expression in insect forms of Trypanosoma brucei by modulating RNA stability and translation. *Mol Cell Biol* 17(8), 4372-4380.

Gao, H., Ayub, M.J., Levin, M.J., and Frank, J. (2005). The structure of the 80S ribosome from Trypanosoma cruzi reveals unique rRNA components. *Proc Natl Acad Sci U S A* 102(29), 10206-10211. doi: 10.1073/pnas.0500926102.

Genois, M.M., Paquet, E.R., Laffitte, M.C., Maity, R., Rodrigue, A., Ouellette, M., et al. (2014). DNA repair pathways in trypanosomatids: from DNA repair to drug resistance. *Microbiol Mol Biol Rev* 78(1), 40-73. doi: 10.1128/MMBR.00045-13.

Godoy, P.D., Nogueira-Junior, L.A., Paes, L.S., Cornejo, A., Martins, R.M., Silber, A.M., et al. (2009). Trypanosome prereplication machinery contains a single functional orc1/cdc6 protein, which is typical of archaea. *Eukaryot Cell* 8(10), 1592-1603. doi: 10.1128/EC.00161-09.

Goodarzi, H., Najafabadi, H.S., Oikonomou, P., Greco, T.M., Fish, L., Salavati, R., et al. (2012). Systematic discovery of structural elements governing stability of mammalian messenger RNAs. *Nature* 485(7397), 264-268. doi: 10.1038/nature11013.

Greif, G., Ponce de Leon, M., Lamolle, G., Rodriguez, M., Pineyro, D., Tavares-Marques, L.M., et al. (2013). Transcriptome analysis of the bloodstream stage from the parasite Trypanosoma vivax. *BMC Genomics* 14, 149. doi: 10.1186/1471-2164-14-149.

Griffin, E., Re, A., Hamel, N., Fu, C., Bush, H., McCaffrey, T., et al. (2001). A link between diabetes and atherosclerosis: Glucose regulates expression of CD36 at the level of translation. *Nat Med* 7(7), 840-846. doi: 10.1038/89969.

Guerra-Slompo, E.P., Probst, C.M., Pavoni, D.P., Goldenberg, S., Krieger, M.A., and Dallagiovanna, B. (2012). Molecular characterization of the Trypanosoma cruzi specific RNA binding protein TcRBP40 and its associated mRNAs. *Biochem Biophys Res Commun* 420(2), 302-307. doi: 10.1016/j.bbrc.2012.02.154.

Haile, S., Cristodero, M., Clayton, C., and Estevez, A.M. (2007). The subcellular localisation of trypanosome RRP6 and its association with the exosome. *Mol Biochem Parasitol* 151(1), 52-58. doi: 10.1016/j.molbiopara.2006.10.005.

Hajarnavis, A., and Durbin, R. (2006). A conserved sequence motif in 3' untranslated regions of ribosomal protein mRNAs in nematodes. *RNA* 12(10), 1786-1789. doi: 10.1261/rna.51306.

Hannaert, V., Bringaud, F., Opperdoes, F.R., and Michels, P.A. (2003). Evolution of energy metabolism and its compartmentation in Kinetoplastida. *Kinetoplastid Biol Dis* 2(1), 11. doi: 10.1186/1475-9292-2-11.

Hannaert, V., and Michels, P.A. (1994). Structure, function, and biogenesis of glycosomes in kinetoplastida. *J Bioenerg Biomembr* 26(2), 205-212.

Harmer, J., Towers, K., Addison, M., Vaughan, S., Ginger, M.L., and McKean, P.G. (2018). A centriolar FGR1 oncogene partner-like protein required for paraflagellar rod assembly, but not axoneme assembly in African trypanosomes. *Open Biol* 8(7). doi: 10.1098/rsob.170218.

Hashem, Y., des Georges, A., Fu, J., Buss, S.N., Jossinet, F., Jobe, A., et al. (2013). High-resolution cryo-electron microscopy structure of the Trypanosoma brucei ribosome. *Nature* 494(7437), 385-389. doi: 10.1038/nature11872.

Hassan, M.A., Vasquez, J.J., Guo-Liang, C., Meissner, M., and Nicolai Siegel, T. (2017). Comparative ribosome profiling uncovers a dominant role for translational control in Toxoplasma gondii. *BMC Genomics* 18(1), 961. doi: 10.1186/s12864-017-4362-6.

Hendriks, E.F., Robinson, D.R., Hinkins, M., and Matthews, K.R. (2001). A novel CCCH protein which modulates differentiation of Trypanosoma brucei to its procyclic form. *EMBO J* 20(23), 6700-6711. doi: 10.1093/emboj/20.23.6700.

Hinnebusch, A.G. (2011). Molecular mechanism of scanning and start codon selection in eukaryotes. *Microbiol Mol Biol Rev* 75(3), 434-467, first page of table of contents. doi: 10.1128/MMBR.00008-11.

Hoffmann, A., Jakob, M., and Ochsenreiter, T. (2016). A novel component of the mitochondrial genome segregation machinery in trypanosomes. *Microb Cell* 3(8), 352-354. doi: 10.15698/mic2016.08.519.

Hotz, H.R., Hartmann, C., Huober, K., Hug, M., and Clayton, C. (1997). Mechanisms of developmental regulation in Trypanosoma brucei: a polypyrimidine tract in the 3'-untranslated region of a surface protein mRNA affects RNA abundance and translation. *Nucleic Acids Res* 25(15), 3017-3026.

Huang da, W., Sherman, B.T., and Lempicki, R.A. (2009). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* 37(1), 1-13. doi: 10.1093/nar/gkn923.

Huntley, R.P., Sawford, T., Mutowo-Meullenet, P., Shypitsyna, A., Bonilla, C., Martin, M.J., et al. (2015). The GOA database: gene Ontology annotation updates for 2015. *Nucleic Acids Res* 43(Database issue), D1057-1063. doi: 10.1093/nar/gku1113.

Ingolia, N.T., Ghaemmaghami, S., Newman, J.R., and Weissman, J.S. (2009). Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* 324(5924), 218-223. doi: 10.1126/science.1168978.

Ingolia, N.T., Lareau, L.F., and Weissman, J.S. (2011). Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* 147(4), 789-802. doi: 10.1016/j.cell.2011.10.002.

Jager, A.V., Muia, R.P., and Campetella, O. (2008). Stage-specific expression of Trypanosoma cruzi trans-sialidase involves highly conserved 3' untranslated regions. *FEMS Microbiol Lett* 283(2), 182-188. doi: 10.1111/j.1574-6968.2008.01170.x.

Jaskowska, E., Butler, C., Preston, G., and Kelly, S. (2015). Phytomonas: trypanosomatids adapted to plant environments. *PLoS Pathog* 11(1), e1004484. doi: 10.1371/journal.ppat.1004484.

Jensen, B.C., Ramasamy, G., Vasconcelos, E.J., Ingolia, N.T., Myler, P.J., and Parsons, M. (2014). Extensive stage-regulation of translation revealed by ribosome profiling of Trypanosoma brucei. *BMC Genomics* 15, 911. doi: 10.1186/1471-2164-15-911.

Johnson, N.R., Yeoh, J.M., Coruh, C., and Axtell, M.J. (2016). Improved Placement of Multi-mapping Small RNAs. *G3 (Bethesda)* 6(7), 2103-2111. doi: 10.1534/g3.116.030452.

Keene, J.D. (2007). RNA regulons: coordination of post-transcriptional events. *Nat Rev Genet* 8(7), 533-543. doi: 10.1038/nrg2111.

Khare, R., Leaman, R., and Lu, Z. (2014). Accessing biomedical literature in the current information landscape. *Methods Mol Biol* 1159, 11-31. doi: 10.1007/978-1-4939-0709-0_2.

Kiss, D.L., Baez, W., Huebner, K., Bundschuh, R., and Schoenberg, D.R. (2017). Impact of FHIT loss on the translation of cancer-associated mRNAs. *Mol Cancer* 16(1), 179. doi: 10.1186/s12943-017-0749-x.

Kolev, N.G., Franklin, J.B., Carmi, S., Shi, H., Michaeli, S., and Tschudi, C. (2010). The transcriptome of the human pathogen Trypanosoma brucei at single-nucleotide resolution. *PLoS Pathog* 6(9), e1001090. doi: 10.1371/journal.ppat.1001090.

Kolev, N.G., Ramey-Butler, K., Cross, G.A., Ullu, E., and Tschudi, C. (2012). Developmental progression to infectivity in Trypanosoma brucei triggered by an RNA-binding protein. *Science* 338(6112), 1352-1353. doi: 10.1126/science.1229641.

Kozak, M. (1978). How do eucaryotic ribosomes select initiation regions in messenger RNA? *Cell* 15(4), 1109-1123.

Kozak, M. (1987). Effects of intercistronic length on the efficiency of reinitiation by eucaryotic ribosomes. *Mol Cell Biol* 7(10), 3438-3445.

Kozak, M. (2001). Constraints on reinitiation of translation in mammals. *Nucleic Acids Res* 29(24), 5226-5232.

Kozak, M. (2002). Pushing the limits of the scanning mechanism for initiation of translation. *Gene* 299(1-2), 1-34.

Kriventseva, E.V., Kuznetsov, D., Tegenfeldt, F., Manni, M., Dias, R., Simao, F.A., et al. (2019). OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Res* 47(D1), D807-D811. doi: 10.1093/nar/gky1053.

Kumar, S., Stecher, G., Li, M., Knyaz, C., and Tamura, K. (2018). MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Mol Biol Evol* 35(6), 1547-1549. doi: 10.1093/molbev/msy096.

Laird, P.W. (1989). Trans splicing in trypanosomes—archaism or adaptation? *Trends in Genetics* 5, 204-208.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10(3), R25. doi: 10.1186/gb-2009-10-3-r25.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., et al. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics* 23(21), 2947-2948. doi: 10.1093/bioinformatics/btm404.

Law, C.W., Chen, Y., Shi, W., and Smyth, G.K. (2014). voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome biology* 15(2), R29.

Lawless, C., Pearson, R.D., Selley, J.N., Smirnova, J.B., Grant, C.M., Ashe, M.P., et al. (2009). Upstream sequence elements direct post-transcriptional regulation of gene expression under stress conditions in yeast. *BMC Genomics* 10, 7. doi: 10.1186/1471-2164-10-7.

LeBowitz, J.H., Smith, H.Q., Rusche, L., and Beverley, S.M. (1993). Coupling of poly(A) site selection and trans-splicing in Leishmania. *Genes Dev* 7(6), 996-1007.

Lee, S., Liu, B., Lee, S., Huang, S.X., Shen, B., and Qian, S.B. (2012). Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. *Proc Natl Acad Sci U S A* 109(37), E2424-2432. doi: 10.1073/pnas.1207846109.

Letunic, I., and Bork, P. (2018). 20 years of the SMART protein domain annotation resource. *Nucleic Acids Res* 46(D1), D493-D496. doi: 10.1093/nar/gkx922.

Levin, M.J., Vazquez, M., Kaplan, D., and Schijman, A.G. (1993). The Trypanosoma cruzi ribosomal P protein family: classification and antigenicity. *Parasitol Today* 9(10), 381-384.

Li, W., Wang, W., Uren, P.J., Penalva, L.O.F., and Smith, A.D. (2017). Riborex: fast and flexible identification of differential translation from Ribo-seq data. *Bioinformatics* 33(11), 1735-1737. doi: 10.1093/bioinformatics/btx047.

Li, Y., Shah-Simpson, S., Okrah, K., Belew, A.T., Choi, J., Caradonna, K.L., et al. (2016). Transcriptome Remodeling in Trypanosoma cruzi and Human Cells during Intracellular Infection. *PLoS Pathog* 12(4), e1005511. doi: 10.1371/journal.ppat.1005511.

Li, Z.H., De Gaudenzi, J.G., Alvarez, V.E., Mendiondo, N., Wang, H., Kissinger, J.C., et al. (2012). A 43-nucleotide U-rich element in 3'-untranslated region of large number of Trypanosoma cruzi transcripts is important for mRNA abundance in intracellular amastigotes. *J Biol Chem* 287(23), 19058-19069. doi: 10.1074/jbc.M111.338699.

Liao, Y., Smyth, G.K., and Shi, W. (2013). The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Res* 41(10), e108. doi: 10.1093/nar/gkt214.

Liu, Z., Gutierrez-Vargas, C., Wei, J., Grassucci, R.A., Ramesh, M., Espina, N., et al. (2016). Structure and assembly model for the Trypanosoma cruzi 60S ribosomal subunit. *Proc Natl Acad Sci U S A* 113(43), 12174-12179. doi: 10.1073/pnas.1614594113.

Loper, E., and Bird, S. (2002). NLTK: the natural language toolkit. *arXiv preprint cs/0205028*.

Lorenz, R., Bernhart, S.H., Honer Zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P.F., et al. (2011). ViennaRNA Package 2.0. *Algorithms Mol Biol* 6, 26. doi: 10.1186/1748-7188-6-26.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology* 15(12), 550.

Lukes, J., Butenko, A., Hashimi, H., Maslov, D.A., Votypka, J., and Yurchenko, V. (2018). Trypanosomatids Are Much More than Just Trypanosomes: Clues from the Expanded Family Tree. *Trends Parasitol* 34(6), 466-480. doi: 10.1016/j.pt.2018.03.002.

Lukes, J., Skalicky, T., Tyc, J., Votypka, J., and Yurchenko, V. (2014). Evolution of parasitism in kinetoplastid flagellates. *Mol Biochem Parasitol* 195(2), 115-122. doi: 10.1016/j.molbiopara.2014.05.007.

Manger, I.D., and Boothroyd, J.C. (1998). Identification of a nuclear protein in Trypanosoma brucei with homology to RNA-binding proteins from cis-splicing systems. *Mol Biochem Parasitol* 97(1-2), 1-11.

Mani, J., Guttinger, A., Schimanski, B., Heller, M., Acosta-Serrano, A., Pescher, P., et al. (2011). Alba-domain proteins of Trypanosoma brucei are cytoplasmic RNA-binding proteins that interact with the translation machinery. *PLoS One* 6(7), e22463. doi: 10.1371/journal.pone.0022463.

Maria, T.A., Tafuri, W., and Brener, Z. (1972). The fine structure of different bloodstream forms of Trypanosoma cruzi. *Ann Trop Med Parasitol* 66(4), 423-431.

Maris, C., Dominguez, C., and Allain, F.H. (2005). The RNA recognition motif, a plastic RNA-binding platform to regulate post-transcriptional gene expression. *FEBS J* 272(9), 2118-2131. doi: 10.1111/j.1742-4658.2005.04653.x.

Marques, C.A., Tiengwe, C., Lemgruber, L., Damasceno, J.D., Scott, A., Paape, D., et al. (2016). Diverged composition and regulation of the Trypanosoma brucei origin recognition complex that mediates DNA replication initiation. *Nucleic Acids Res* 44(10), 4763-4784. doi: 10.1093/nar/gkw147.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal* 17(1), 10-12.

Matrosovich, M., Herrler, G., and Klenk, H.D. (2015). Sialic Acid Receptors of Viruses. *Top Curr Chem* 367, 1-28. doi: 10.1007/128_2013_466.

Matthews, K.R., Tschudi, C., and Ullu, E. (1994). A common pyrimidine-rich motif governs trans-splicing and polyadenylation of tubulin polycistronic pre-mRNA in trypanosomes. *Genes Dev* 8(4), 491-501.

McCarthy, D.J., Smyth, G.K., and Robinson, M.D. (2009). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26(1), 139-140. doi: 10.1093/bioinformatics/btp616.

McCarthy, J.E. (1998). Posttranscriptional control of gene expression in yeast. *Microbiol Mol Biol Rev* 62(4), 1492-1553.

Mendonca-Previato, L., Penha, L., Garcez, T.C., Jones, C., and Previato, J.O. (2013). Addition of alpha-O-GlcNAc to threonine residues define the post-translational modification of mucin-like molecules in Trypanosoma cruzi. *Glycoconj J* 30(7), 659-666. doi: 10.1007/s10719-013-9469-7.

Meyuhas, O., and Kahan, T. (2015). The race to decipher the top secrets of TOP mRNAs. *Biochim Biophys Acta* 1849(7), 801-811. doi: 10.1016/j.bbagrm.2014.08.015.

Michel, A.M., Andreev, D.E., and Baranov, P.V. (2014). Computational approach for calculating the probability of eukaryotic translation initiation from ribo-seq data that takes into account leaky scanning. *BMC Bioinformatics* 15, 380. doi: 10.1186/s12859-014-0380-4.

Minning, T.A., Weatherly, D.B., Atwood, J., 3rd, Orlando, R., and Tarleton, R.L. (2009). The steady-state transcriptome of the four major life-cycle stages of Trypanosoma cruzi. *BMC Genomics* 10, 370. doi: 10.1186/1471-2164-10-370.

Mirdita, M., von den Driesch, L., Galiez, C., Martin, M.J., Soding, J., and Steinegger, M. (2017). Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic Acids Res* 45(D1), D170-D176. doi: 10.1093/nar/gkw1081.

Mitchell, A.L., Attwood, T.K., Babbitt, P.C., Blum, M., Bork, P., Bridge, A., et al. (2019). InterPro in 2019: improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Res* 47(D1), D351-D360. doi: 10.1093/nar/gky1100.

Molyneux, D.H. (2014). Neglected tropical diseases: now more than just 'other diseases'—the post-2015 agenda. *International health* 6(3), 172-180.

Moreira, D., Lopez-Garcia, P., and Vickerman, K. (2004). An updated view of kinetoplastid phylogeny using environmental sequences and a closer outgroup: proposal for a new classification of the class Kinetoplastea. *Int J Syst Evol Microbiol* 54(Pt 5), 1861-1875. doi: 10.1099/ijs.0.63081-0.

Mortara, R.A., Andreoli, W.K., Taniwaki, N.N., Fernandes, A.B., Silva, C.V., Fernandes, M.C., et al. (2005). Mammalian cell invasion and intracellular trafficking by Trypanosoma cruzi infective forms. *An Acad Bras Cienc* 77(1), 77-94. doi: /S0001-37652005000100006.

Mucci, J., Risso, M.G., Leguizamon, M.S., Frasch, A.C., and Campetella, O. (2006). The trans-sialidase from Trypanosoma cruzi triggers apoptosis by target cell sialylation. *Cell Microbiol* 8(7), 1086-1095. doi: 10.1111/j.1462-5822.2006.00689.x.

Mugo, E., and Clayton, C. (2017). Expression of the RNA-binding protein RBP10 promotes the bloodstream-form differentiation state in Trypanosoma brucei. *PLoS Pathog* 13(8), e1006560. doi: 10.1371/journal.ppat.1006560.

Myler, P.J., Audleman, L., deVos, T., Hixson, G., Kiser, P., Lemley, C., et al. (1999). Leishmania major Friedlin chromosome 1 has an unusual distribution of protein-coding genes. *Proc Natl Acad Sci U S A* 96(6), 2902-2906.

Nakagawa, S., Niimura, Y., Gojobori, T., Tanaka, H., and Miura, K. (2008). Diversity of preferred nucleotide sequences around the translation initiation codon in eukaryote genomes. *Nucleic Acids Res* 36(3), 861-871. doi: 10.1093/nar/gkm1102.

Newberry, L.B., and Paulin, J.J. (1989). Reconstruction of the chondriome of the amastigote form of Trypanosoma cruzi. *J Parasitol* 75(4), 649-652.

Nilsson, D., Gunasekera, K., Mani, J., Osteras, M., Farinelli, L., Baerlocher, L., et al. (2010). Spliced leader trapping reveals widespread alternative splicing patterns in the highly dynamic transcriptome of Trypanosoma brucei. *PLoS Pathog* 6(8), e1001037. doi: 10.1371/journal.ppat.1001037.

Noe, G., De Gaudenzi, J.G., and Frasch, A.C. (2008). Functionally related transcripts have common RNA motifs for specific RNA-binding proteins in trypanosomes. *BMC Mol Biol* 9, 107. doi: 10.1186/1471-2199-9-107.

Notredame, C., Higgins, D.G., and Heringa, J. (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol* 302(1), 205-217. doi: 10.1006/jmbi.2000.4042.

Ouellette, M., and Papadopoulou, B. (2009). Coordinated gene expression by post-transcriptional regulons in African trypanosomes. *J Biol* 8(11), 100. doi: 10.1186/jbiol203.

Palenchar, J.B., and Bellofatto, V. (2006). Gene transcription in trypanosomes. *Mol Biochem Parasitol* 146(2), 135-141. doi: 10.1016/j.molbiopara.2005.12.008.

Palenchar, J.B., Liu, W., Palenchar, P.M., and Bellofatto, V. (2006). A divergent transcription factor TFIIB in trypanosomes is required for RNA polymerase II-dependent spliced leader RNA transcription and cell viability. *Eukaryot Cell* 5(2), 293-300. doi: 10.1128/EC.5.2.293-300.2006.

Pantano, L., Estivill, X., and Marti, E. (2011). A non-biased framework for the annotation and classification of the non-miRNA small RNA transcriptome. *Bioinformatics* 27(22), 3202-3203. doi: 10.1093/bioinformatics/btr527.

Parodi, A.J., Pollevick, G.D., Mautner, M., Buschiazzo, A., Sanchez, D.O., and Frasch, A.C. (1992). Identification of the gene(s) coding for the trans-sialidase of Trypanosoma cruzi. *EMBO J* 11(5), 1705-1710.

Parsons, M., and Myler, P.J. (2016). Illuminating Parasite Protein Production by Ribosome Profiling. *Trends Parasitol* 32(6), 446-457. doi: 10.1016/j.pt.2016.03.005.

Pastro, L., Smircich, P., Perez-Diaz, L., Duhagon, M.A., and Garat, B. (2013). Implication of CA repeated tracts on post-transcriptional regulation in Trypanosoma cruzi. *Exp Parasitol* 134(4), 511-518. doi: 10.1016/j.exppara.2013.04.004.

Peikert, C.D., Mani, J., Morgenstern, M., Kaser, S., Knapp, B., Wenger, C., et al. (2017). Charting organellar importomes by quantitative mass spectrometry. *Nat Commun* 8, 15272. doi: 10.1038/ncomms15272.

Perez-Diaz, L., Duhagon, M.A., Smircich, P., Sotelo-Silveira, J., Robello, C., Krieger, M.A., et al. (2007). Trypanosoma cruzi: molecular characterization of an RNA binding protein differentially expressed in the parasite life cycle. *Exp Parasitol* 117(1), 99-105. doi: 10.1016/j.exppara.2007.03.010.

Perez-Diaz, L., Pastro, L., Smircich, P., Dallagiovanna, B., and Garat, B. (2013). Evidence for a negative feedback control mediated by the 3' untranslated region assuring the low expression level of the RNA binding protein TcRBP19 in T. cruzi epimastigotes. *Biochem Biophys Res Commun* 436(2), 295-299. doi: 10.1016/j.bbrc.2013.05.096.

Pietrosanto, M., Mattei, E., Helmer-Citterich, M., and Ferre, F. (2016). A novel method for the identification of conserved structural patterns in RNA: From small scale to high-throughput applications. *Nucleic Acids Res* 44(18), 8600-8609. doi: 10.1093/nar/gkw750.

Price, M.N., and Arkin, A.P. (2017). PaperBLAST: Text Mining Papers for Information about Homologs. *mSystems* 2(4). doi: 10.1128/mSystems.00039-17.

Quijada, L., Guerra-Giraldez, C., Drozdz, M., Hartmann, C., Irmer, H., Ben-Dov, C., et al. (2002). Expression of the human RNA-binding protein HuR in Trypanosoma brucei increases the abundance of mRNAs containing AU-rich regulatory elements. *Nucleic Acids Res* 30(20), 4414-4424.

Rachidi, N., Taly, J.F., Durieu, E., Leclercq, O., Aulner, N., Prina, E., et al. (2014). Pharmacological assessment defines Leishmania donovani casein kinase 1 as a drug target and reveals important functions in parasite viability and intracellular infection. *Antimicrob Agents Chemother* 58(3), 1501-1515. doi: 10.1128/AAC.02022-13.

Radio, S., Fort, R.S., Garat, B., Sotelo-Silveira, J., and Smircich, P. (2018). UTRme: A Scoring-Based Tool to Annotate Untranslated Regions in Trypanosomatid Genomes. *Front Genet* 9, 671. doi: 10.3389/fgene.2018.00671.

Rajkowitsch, L., Vilela, C., Berthelot, K., Ramirez, C.V., and McCarthy, J.E. (2004). Reinitiation and recycling are distinct processes occurring downstream of translation termination in yeast. *J Mol Biol* 335(1), 71-85.

Rassi, A., Jr., Rassi, A., and Marin-Neto, J.A. (2010). Chagas disease. *Lancet* 375(9723), 1388-1402. doi: 10.1016/S0140-6736(10)60061-X.

Read, L.K., Lukes, J., and Hashimi, H. (2016). Trypanosome RNA editing: the complexity of getting U in and taking U out. *Wiley Interdiscip Rev RNA* 7(1), 33-51. doi: 10.1002/wrna.1313.

Remmert, M., Biegert, A., Hauser, A., and Soding, J. (2011). HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat Methods* 9(2), 173-175. doi: 10.1038/nmeth.1818.

Ribeirao, M., Pereira-Chioccola, V.L., Eichinger, D., Rodrigues, M.M., and Schenkman, S. (1997). Temperature differences for trans-glycosylation and hydrolysis reaction reveal an acceptor binding site in the catalytic mechanism of Trypanosoma cruzi trans-sialidase. *Glycobiology* 7(8), 1237-1246.

Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* 16(6), 276-277.

Rocklin, M. (Year). "Dask: Parallel computation with blocked algorithms and task scheduling", in: *Proceedings of the 14th Python in Science Conference*: Citeseer).

Romaniuk, M.A., Frasch, A.C., and Cassola, A. (2018). Translational repression by an RNA-binding protein promotes differentiation to infective forms in Trypanosoma cruzi. *PLoS Pathog* 14(6), e1007059. doi: 10.1371/journal.ppat.1007059.

Sakyiama, J., Zimmer, S.L., Ciganda, M., Williams, N., and Read, L.K. (2013). Ribosome biogenesis requires a highly diverged XRN family 5'->3' exoribonuclease for rRNA processing in Trypanosoma brucei. *RNA* 19(10), 1419-1431. doi: 10.1261/rna.038547.113.

Santos, C., Ludwig, A., Kessler, R.L., Rampazzo, R.C.P., Inoue, A.H., Krieger, M.A., et al. (2018). Trypanosoma cruzi transcriptome during axenic epimastigote growth curve. *Mem Inst Oswaldo Cruz* 113(5), e170404. doi: 10.1590/0074-02760170404.

Schenkman, S., Jiang, M.S., Hart, G.W., and Nussenzweig, V. (1991). A novel cell surface trans-sialidase of Trypanosoma cruzi generates a stage-specific epitope required for invasion of mammalian cells. *Cell* 65(7), 1117-1125.

Schenkman, S., Pascoalino Bdos, S., and Nardelli, S.C. (2011). Nuclear structure of Trypanosoma cruzi. *Adv Parasitol* 75, 251-283. doi: 10.1016/B978-0-12-385863-4.00012-5.

Schwanhausser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., et al. (2011). Global quantification of mammalian gene expression control. *Nature* 473(7347), 337-342. doi: 10.1038/nature10098.

Sengupta, J., Nilsson, J., Gursky, R., Spahn, C.M., Nissen, P., and Frank, J. (2004). Identification of the versatile scaffold protein RACK1 on the eukaryotic ribosome by cryo-EM. *Nat Struct Mol Biol* 11(10), 957-962. doi: 10.1038/nsmb822.

Shalev-Benami, M., Zhang, Y., Matzov, D., Halfon, Y., Zackay, A., Rozenberg, H., et al. (2016). 2.8-A Cryo-EM Structure of the Large Ribosomal Subunit from the Eukaryotic Parasite Leishmania. *Cell Rep* 16(2), 288-294. doi: 10.1016/j.celrep.2016.06.014.

Siegel, T.N., Hekstra, D.R., Wang, X., Dewell, S., and Cross, G.A. (2010). Genome-wide analysis of mRNA abundance in two life-cycle stages of Trypanosoma brucei and identification of splicing and polyadenylation sites. *Nucleic Acids Res* 38(15), 4946-4957. doi: 10.1093/nar/gkq237.

Siegel, T.N., Tan, K.S., and Cross, G.A. (2005). Systematic study of sequence motifs for RNA trans splicing in Trypanosoma brucei. *Mol Cell Biol* 25(21), 9586-9594. doi: 10.1128/MCB.25.21.9586-9594.2005.

Sigrist, C.J., Cerutti, L., de Castro, E., Langendijk-Genevaux, P.S., Bulliard, V., Bairoch, A., et al. (2010). PROSITE, a protein domain database for functional characterization and annotation. *Nucleic Acids Res* 38(Database issue), D161-166. doi: 10.1093/nar/gkp885.

Simpson, A.G., Stevens, J.R., and Lukes, J. (2006). The evolution and diversity of kinetoplastid flagellates. *Trends Parasitol* 22(4), 168-174. doi: 10.1016/j.pt.2006.02.006.

Smircich, P., Eastman, G., Bispo, S., Duhagon, M.A., Guerra-Slompo, E.P., Garat, B., et al. (2015). Ribosome profiling reveals translation control as a key mechanism generating differential gene expression in Trypanosoma cruzi. *BMC Genomics* 16, 443. doi: 10.1186/s12864-015-1563-8.

Smircich, P., El-Sayed, N.M., and Garat, B. (2017). Intrinsic DNA curvature in trypanosomes. *BMC Res Notes* 10(1), 585. doi: 10.1186/s13104-017-2908-y.

Smith, D.F., and Parsons, M. (1996). *Molecular biology of parasitic protozoa.* Oxford ; New York: IRL Press at Oxford University Press.

Solari, A.J. (1995). Mitosis and genome partition in trypanosomes. *Biocell* 19(2), 65-84.

Souto-Padron, T., de Souza, W., and Heuser, J.E. (1984). Quick-freeze, deep-etch rotary replication of Trypanosoma cruzi and Herpetomonas megaseliae. *J Cell Sci* 69, 167-178.

Steinegger, M., and Soding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol* 35(11), 1026-1028. doi: 10.1038/nbt.3988.

Supek, F., Bosnjak, M., Skunca, N., and Smuc, T. (2011). REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS One* 6(7), e21800. doi: 10.1371/journal.pone.0021800.

Team, R. (2015). RStudio: integrated development for R. *RStudio, Inc., Boston, MA URL http://www. rstudio. com* 42, 14.

Team, R.C. (2013). R: A language and environment for statistical computing.

Tschudi, C., and Ullut, E. (2002). Unconventional rules of small nuclear RNA transcription and cap modification in trypanosomatids. *Gene Expr* 10(1-2), 3-16.

Tuorto, F., Legrand, C., Cirzi, C., Federico, G., Liebers, R., Muller, M., et al. (2018). Queuosine-modified tRNAs confer nutritional control of protein translation. *EMBO J* 37(18). doi: 10.15252/embj.201899777.

Tyler, K.M., and Engman, D.M. (2001). The life cycle of Trypanosoma cruzi revisited. *Int J Parasitol* 31(5-6), 472-481.

UniProt, C. (2019). UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res* 47(D1), D506-D515. doi: 10.1093/nar/gky1049.

Urbaniak, M.D. (2009). Casein kinase 1 isoform 2 is essential for bloodstream form Trypanosoma brucei. *Mol Biochem Parasitol* 166(2), 183-185. doi: 10.1016/j.molbiopara.2009.03.001.

Vanhamme, L., and Pays, E. (1995). Control of gene expression in trypanosomes. *Microbiol Rev* 59(2), 223-240.

Vasquez, J.J., Hon, C.C., Vanselow, J.T., Schlosser, A., and Siegel, T.N. (2014). Comparative ribosome profiling reveals extensive translational complexity in different Trypanosoma brucei life cycle stages. *Nucleic Acids Res* 42(6), 3623-3637. doi: 10.1093/nar/gkt1386.

Vattem, K.M., and Wek, R.C. (2004). Reinitiation involving upstream ORFs regulates ATF4 mRNA translation in mammalian cells. *Proc Natl Acad Sci U S A* 101(31), 11269-11274. doi: 10.1073/pnas.0400541101.

Vercelli, C.A., Hidalgo, A.M., Hyon, S.H., and Argibay, P.F. (2005). Trypanosoma cruzi trans-sialidase inhibits human lymphocyte proliferation by nonapoptotic mechanisms: implications in pathogenesis and transplant immunology. *Transplant Proc* 37(10), 4594-4597. doi: 10.1016/j.transproceed.2005.10.013.

Vilela, C., and McCarthy, J.E. (2003). Regulation of fungal gene expression via short open reading frames in the mRNA 5'untranslated region. *Mol Microbiol* 49(4), 859-867.

von Roretz, C., Di Marco, S., Mazroui, R., and Gallouzi, I.E. (2011). Turnover of AU-rich-containing mRNAs during stress: a matter of survival. *Wiley Interdiscip Rev RNA* 2(3), 336-347. doi: 10.1002/wrna.55.

Waterhouse, R.M., Seppey, M., Simao, F.A., Manni, M., Ioannidis, P., Klioutchnikov, G., et al. (2017). BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol Biol Evol*. doi: 10.1093/molbev/msx319.

Wethmar, K. (2014). The regulatory potential of upstream open reading frames in eukaryotic gene expression. *Wiley Interdiscip Rev RNA* 5(6), 765-778. doi: 10.1002/wrna.1245.

Wethmar, K., Barbosa-Silva, A., Andrade-Navarro, M.A., and Leutz, A. (2014). uORFdb--a comprehensive literature database on eukaryotic uORF biology. *Nucleic Acids Res* 42(Database issue), D60-67. doi: 10.1093/nar/gkt952.

Wurst, M., Seliger, B., Jha, B.A., Klein, C., Queiroz, R., and Clayton, C. (2012). Expression of the RNA recognition motif protein RBP10 promotes a bloodstream-form transcript pattern in Trypanosoma brucei. *Mol Microbiol* 83(5), 1048-1063. doi: 10.1111/j.1365-2958.2012.07988.x.

Yamada-Ogatta, S.F., Motta, M.C., Toma, H.K., Monteiro-Goes, V., Avila, A.R., Muniz, B.D., et al. (2004). Trypanosoma cruzi: cloning and characterization of two genes whose expression is up-regulated in metacyclic trypomastigotes. *Acta Trop* 90(2), 171-179. doi: 10.1016/j.actatropica.2003.10.018.

Ye, J., Zhang, Y., Cui, H., Liu, J., Wu, Y., Cheng, Y., et al. (2018). WEGO 2.0: a web tool for analyzing and plotting GO annotations, 2018 update. *Nucleic Acids Res* 46(W1), W71-W75. doi: 10.1093/nar/gky400.

Zhong, Y., Karaletsos, T., Drewe, P., Sreedharan, V.T., Kuo, D., Singh, K., et al. (2017). RiboDiff: detecting changes of mRNA translation efficiency from ribosome footprints. *Bioinformatics* 33(1), 139-141. doi: 10.1093/bioinformatics/btw585.

Zingales, B., Andrade, S.G., Briones, M.R., Campbell, D.A., Chiari, E., Fernandes, O., et al. (2009). A new consensus for Trypanosoma cruzi intraspecific nomenclature: second revision meeting recommends TcI to TcVI. *Mem Inst Oswaldo Cruz* 104(7), 1051-1054.

6. Anexo

6.1. Optimización del análisis de datos de Ribosome Profiling

Tabla 6.1. Anotación de genes que bajan su eficiencia traduccional en el pasaje de epimastigota a tripomastigota metacíclico.

ID	Annotation
TcCLB.511139.20	40S ribosomal protein S11, putative
TcCLB.506181.59	40S ribosomal protein S12, putative
TcCLB.510029.70	40S ribosomal protein S13, putative
TcCLB.509381.20	40S ribosomal protein S15A, putative
TcCLB.503899.20	40S ribosomal protein S16, putative
TcCLB.506213.60	40S ribosomal protein S2, putative
TcCLB.504181.30	40S ribosomal protein S23, putative
TcCLB.506963.14	40S ribosomal protein S27, putative
TcCLB.430605.29	40S ribosomal protein S3, putative
TcCLB.506413.30	40S ribosomal protein S33, putative
TcCLB.511001.18	40S ribosomal protein S3A, putative
TcCLB.511001.9	40S ribosomal protein S3A, putative
TcCLB.511069.10	40S ribosomal protein S8, putative
TcCLB.511069.20	40S ribosomal protein S8, putative
TcCLB.506401.120	40S ribosomal protein S9, putative
TcCLB.509825.14	40S ribosomal protein SA, putative
TcCLB.509165.40	60S acidic ribosomal protein P2 beta (H6.4), putative
TcCLB.510309.50	60S acidic ribosomal protein P2, putative
TcCLB.510243.40	60S ribosomal protein L10, putative
TcCLB.506963.10	60S ribosomal protein L10a, putative
TcCLB.510755.129	60S ribosomal protein L12, putative
TcCLB.510323.30	60S ribosomal protein L13, putative
TcCLB.506937.30	60S ribosomal protein L14, putative
TcCLB.503449.10	60S ribosomal protein L17, putative
TcCLB.504147.20	60S ribosomal protein L18, putative
TcCLB.511001.120	60S ribosomal protein L18a, putative
TcCLB.504147.120	60S ribosomal protein L22, putative
TcCLB.510101.30	60S ribosomal protein L28, putative
TcCLB.503643.3	60S ribosomal protein L4, putative (fragment)
TcCLB.508207.110	60S ribosomal protein L7, putative
TcCLB.504181.10	60S ribosomal protein L9, putative
TcCLB.510737.70	60S ribosomal subunit protein L31, putative
TcCLB.510737.79	60S ribosomal subunit protein L31, putative
TcCLB.508369.50	acetyl-CoA carboxylase (fragment)

TcCLB.504137.80	adenylyl cyclase-associated protein, putative
TcCLB.511733.10	aldehyde dehydrogenase, putative
TcCLB.509967.30	aspartyl-tRNA synthetase, putative
TcCLB.507993.10	ATP12 chaperone protein, putative
TcCLB.506941.90	ATP-dependent DEAD/H DNA helicase recQ family, putative
TcCLB.508973.50	ATP-dependent RNA helicase DBP2B, putative
TcCLB.511731.50	ATP-dependent RNA helicase, putative
TcCLB.511067.14	BolA-like protein, putative
TcCLB.508277.10	condensin subunit 1, putative
TcCLB.506493.90	cysteine peptidase, Clan CA, family C2, putative
TcCLB.507711.320	DNA mismatch repair protein MSH2, putative
TcCLB.510859.50	Domain of unknown function (DUF3437), putative
TcCLB.506599.10	elongation factor 1-gamma (EF-1-gamma), putative
TcCLB.506677.39	Elongation factor G1, mitochondrial, putative
TcCLB.506925.120	eukaryotic translation initiation factor 5A
TcCLB.506563.20	exosome-associated protein 1, putative
TcCLB.507641.280	heat shock protein 60
TcCLB.508551.30	hexose transporter, putative
TcCLB.508551.39	hexose transporter, putative
TcCLB.511041.40	hexose transporter, putative
TcCLB.503887.79	HIT zinc finger, putative
TcCLB.510533.150	hypothetical protein
TcCLB.511407.30	hypothetical protein
TcCLB.503813.20	hypothetical protein, conserved
TcCLB.504111.40	hypothetical protein, conserved
TcCLB.506205.30	hypothetical protein, conserved
TcCLB.506729.50	hypothetical protein, conserved
TcCLB.506739.120	hypothetical protein, conserved
TcCLB.506885.60	hypothetical protein, conserved
TcCLB.507093.170	hypothetical protein, conserved
TcCLB.507275.20	hypothetical protein, conserved
TcCLB.507467.94	hypothetical protein, conserved
TcCLB.507611.270	hypothetical protein, conserved
TcCLB.507641.90	hypothetical protein, conserved
TcCLB.508137.40	hypothetical protein, conserved
TcCLB.508569.70	hypothetical protein, conserved
TcCLB.508721.20	hypothetical protein, conserved
TcCLB.508951.70	hypothetical protein, conserved
TcCLB.509245.29	hypothetical protein, conserved
TcCLB.509455.20	hypothetical protein, conserved
TcCLB.509455.50	hypothetical protein, conserved
TcCLB.509801.20	hypothetical protein, conserved
TcCLB.509965.170	hypothetical protein, conserved
TcCLB.510745.10	hypothetical protein, conserved
TcCLB.510761.22	hypothetical protein, conserved
TcCLB.511071.150	hypothetical protein, conserved
TcCLB.511215.10	hypothetical protein, conserved

TcCLB.511467.50	hypothetical protein, conserved
TcCLB.511577.30	hypothetical protein, conserved
TcCLB.511745.50	hypothetical protein, conserved
TcCLB.509871.39	hypothetical protein, conserved (pseudogene)
TcCLB.510769.60	hypothetical protein, conserved (pseudogene)
TcCLB.510751.40	Importin 1 (fragment)
TcCLB.506625.60	inositol 5-phosphatase 1, putative
TcCLB.507929.10	kinesin, putative (fragment)
TcCLB.504075.10	kinetoplast poly(A) polymerase 1
TcCLB.506743.170	Met-10+ like-protein, putative
TcCLB.509761.10	Metallopeptidase family M24/FACT complex subunit (SPT16/CDC68)/Histone chaperone Rttp106-like, putative
TcCLB.509805.190	mitochondrial carrier protein
TcCLB.504125.50	mitochondrial carrier protein, putative
TcCLB.508277.370	mitochondrial DNA topoisomerase II, putative
TcCLB.507927.20	mitochondrial oligo_U binding protein TBRGG1, putative
TcCLB.506869.70	mitochondrial structure specific endonuclease I (SSE-1), putative
TcCLB.506401.240	Molybdopterin guanine dinucleotide synthesis protein B, putative
TcCLB.506947.110	myosin heavy chain, putative, frameshift
TcCLB.511047.40	NADH-cytochrome b5 reductase, putative
TcCLB.506591.69	Nucleoporin NUP53b
TcCLB.504769.80	Nucleoporin NUP92
TcCLB.509965.290	p22 protein precursor, putative
TcCLB.503905.50	PSP1 C-terminal conserved region, putative
TcCLB.506773.130	pumilio-repeat, RNA-binding protein, putative
TcCLB.508277.290	Rab-GTPase-TBC domain containing protein, putative
TcCLB.511211.130	receptor for activated C kinase 1, putative
TcCLB.506365.20	retrotransposon hot spot protein (RHS, pseudogene), putative
TcCLB.503611.20	ribosomal protein L24, putative
TcCLB.511545.40	ribosomal protein L27, putative
TcCLB.510879.110	ribosomal protein L3, putative
TcCLB.503801.20	ribosomal protein S26, putative
TcCLB.511805.15	ribosomal protein S29, putative
TcCLB.506829.39	ribosomal protein S7, putative
TcCLB.506925.40	RNA polymerase I second largest subunit, putative
TcCLB.510143.80	RNA-binding protein, putative
TcCLB.508241.120	rrp44p homologue, putative
TcCLB.510565.20	sphingolipid delta 4 desaturase, putative
TcCLB.511133.20	thymidine kinase, putative
TcCLB.509835.20	translation initiation factor eIF2B delta subunit, putative
TcCLB.505807.270	ubiquitin hydrolase, putative
TcCLB.510293.40	ubiquitin/ribosomal protein S27a, putative
TcCLB.506691.80	Ubiquitin-like domain containing protein, putative
TcCLB.506529.130	WD repeat and HMG-box DNA-binding protein, putative
TcCLB.507485.90	Wee1-like protein kinase, putative

Tabla 6.2. Anotación de genes que suben su eficiencia traduccional en el pasaje de epimastigota a tripomastigota metacíclico.

ID	Product Description
TcCLB.507811.100	amino acid permease, putative
TcCLB.511107.41	ATP-dependent Clp protease subunit heat shock protein 100 (HSP100), putative (fragment)
TcCLB.506775.190	Calcium/calmodulin-dependent protein kinase kinase, putative
TcCLB.506563.130	calpain-like cysteine peptidase (pseudogene), putative
TcCLB.506623.60	dispersed gene family protein 1 (DGF-1), putative
TcCLB.507257.60	Fumarate hydratase class I, cytosolic
TcCLB.506529.508	glucose-6-phosphate isomerase, glycosomal, putative
TcCLB.510289.40	GPI inositol deacylase precursor, putative
TcCLB.511309.20	hydrolase-like protein, putative
TcCLB.511751.140	hypothetical protein
TcCLB.503865.70	hypothetical protein, conserved
TcCLB.504213.100	hypothetical protein, conserved
TcCLB.505789.10	hypothetical protein, conserved
TcCLB.505789.20	hypothetical protein, conserved
TcCLB.506289.70	hypothetical protein, conserved
TcCLB.506401.24	hypothetical protein, conserved
TcCLB.506625.120	hypothetical protein, conserved
TcCLB.509805.210	hypothetical protein, conserved
TcCLB.511365.90	hypothetical protein, conserved
TcCLB.511529.200	hypothetical protein, conserved
TcCLB.506529.600	metacyclin II, putative
TcCLB.509767.10	neutral sphingomyelinase activation associated factor-like protein
TcCLB.511431.10	protein kinase, putative
TcCLB.506717.20	receptor-type adenylate cyclase, putative
TcCLB.511043.60	receptor-type adenylate cyclase, putative
TcCLB.510257.30	ribulose-5-phosphate 3-epimerase, putative
TcCLB.510263.30	surface protease GP63, putative
TcCLB.400945.10	trans-sialidase (pseudogene), putative
TcCLB.404711.10	trans-sialidase (pseudogene), putative
TcCLB.503501.50	trans-sialidase (pseudogene), putative
TcCLB.503957.60	trans-sialidase (pseudogene), putative
TcCLB.504229.50	trans-sialidase (pseudogene), putative
TcCLB.504769.190	trans-sialidase (pseudogene), putative
TcCLB.506487.50	trans-sialidase (pseudogene), putative
TcCLB.506667.110	trans-sialidase (pseudogene), putative
TcCLB.507179.7	trans-sialidase (pseudogene), putative
TcCLB.507445.30	trans-sialidase (pseudogene), putative
TcCLB.507611.70	trans-sialidase (pseudogene), putative
TcCLB.507905.30	trans-sialidase (pseudogene), putative
TcCLB.508121.30	trans-sialidase (pseudogene), putative

TcCLB.508163.90	trans-sialidase (pseudogene), putative
TcCLB.508539.169	trans-sialidase (pseudogene), putative
TcCLB.508607.70	trans-sialidase (pseudogene), putative
TcCLB.508755.11	trans-sialidase (pseudogene), putative
TcCLB.508835.30	trans-sialidase (pseudogene), putative
TcCLB.508991.30	trans-sialidase (pseudogene), putative
TcCLB.509349.10	trans-sialidase (pseudogene), putative
TcCLB.509921.20	trans-sialidase (pseudogene), putative
TcCLB.510275.160	trans-sialidase (pseudogene), putative
TcCLB.510307.250	trans-sialidase (pseudogene), putative
TcCLB.510441.30	trans-sialidase (pseudogene), putative
TcCLB.510451.10	trans-sialidase (pseudogene), putative
TcCLB.510849.20	trans-sialidase (pseudogene), putative
TcCLB.511569.30	trans-sialidase (pseudogene), putative
TcCLB.511797.210	trans-sialidase (pseudogene), putative
TcCLB.511875.110	trans-sialidase (pseudogene), putative
TcCLB.420293.20	trans-sialidase, Group II, putative
TcCLB.503447.20	trans-sialidase, Group II, putative
TcCLB.503759.10	trans-sialidase, Group II, putative
TcCLB.503767.10	trans-sialidase, Group II, putative
TcCLB.503955.40	trans-sialidase, Group II, putative
TcCLB.504099.50	trans-sialidase, Group II, putative
TcCLB.504343.10	trans-sialidase, Group II, putative
TcCLB.504769.100	trans-sialidase, Group II, putative
TcCLB.506021.20	trans-sialidase, Group II, putative
TcCLB.507047.40	trans-sialidase, Group II, putative
TcCLB.507213.40	trans-sialidase, Group II, putative
TcCLB.507611.170	trans-sialidase, Group II, putative
TcCLB.507875.220	trans-sialidase, Group II, putative
TcCLB.508285.60	trans-sialidase, Group II, putative
TcCLB.508563.20	trans-sialidase, Group II, putative
TcCLB.509777.30	trans-sialidase, Group II, putative
TcCLB.510005.20	trans-sialidase, Group II, putative
TcCLB.510307.230	trans-sialidase, Group II, putative
TcCLB.511311.20	trans-sialidase, Group II, putative
TcCLB.511349.100	trans-sialidase, Group II, putative
TcCLB.511585.230	trans-sialidase, Group II, putative
TcCLB.506499.170	trans-sialidase, Group V, putative
TcCLB.506737.90	trans-sialidase, Group V, putative
TcCLB.508109.60	trans-sialidase, Group V, putative
TcCLB.508977.40	trans-sialidase, Group V, putative
TcCLB.509979.320	trans-sialidase, Group V, putative
TcCLB.510377.10	trans-sialidase, Group V, putative
TcCLB.511173.370	trans-sialidase, Group VI, putative
TcCLB.506683.110	trans-sialidase, Group VIII, putative
TcCLB.503667.10	trans-sialidase, putative
TcCLB.503957.10	trans-sialidase, putative

TcCLB.505363.19	trans-sialidase, putative
TcCLB.506413.89	trans-sialidase, putative
TcCLB.433673.10	trans-sialidase, putative (fragment)
TcCLB.435601.10	trans-sialidase, putative (fragment)
TcCLB.505155.4	trans-sialidase, putative (fragment)
TcCLB.510171.10	trans-sialidase, putative (fragment)
TcCLB.511877.10	Trypomastigote, Alanine, Serine and Valine rich protein (TASV), subfamily B

```
my $bowtie_cl = "bowtie $format -v $$options{'mismatches'} -p $$options{'bowtie_cores'} -S";
    if($$options{'mmap'} eq 'n') {
        $bowtie_cl .= " -a -m 1"; # Agrego opcion -m 1
    } elsif ($$options{'bowtie_m'} =~ /^\d+$/) {
        $bowtie_cl .= " -a -m $$options{'bowtie_m'}";
    } else {
        $bowtie_cl .= " -a";
    }

    if($$options{'mismatches'} > 0) {
        $bowtie_cl .= " --best --strata";
    }
    if ($$cqual) {
        $bowtie_cl .= " -Q $$cqual";
    }
    if($format =~ /C/) {
        $bowtie_cl .= " --col-keepends";
    }
    $bowtie_cl .= " -C --sam-RG ID:$base"; # Se agrega opción -C. Mo viene por defecto
    $bowtie_cl .= " $ebwt_base.cs -"; # Se agrego .cs no viene por defecto.
    if ($gz) {
        $bowtie_cl = "gzip -d -c $$reads | " . $bowtie_cl;
    } else {
        $bowtie_cl .= " < $$reads";
    }
```

Figura 6.1. Sección del código modificada para la utilización de ShortStack con datos provenientes de tecnología SOLid.

6.2. Desarrollo de herramientas bioinformáticas que permitan mejorar el análisis de genes diferencialmente expresados

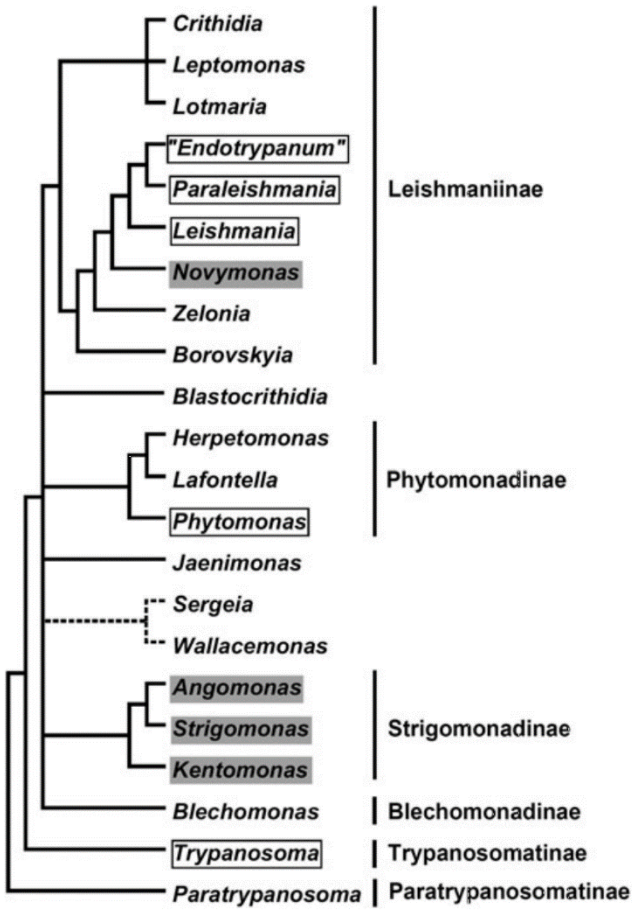


Figura 6.2. Árbol filogenético de la familia de tripanosomátidos. Adaptado de (Simpson et al., 2006).

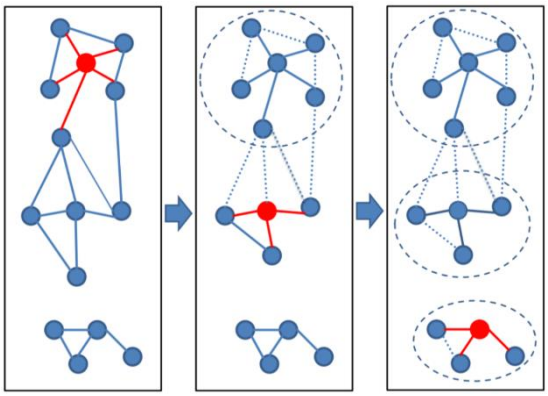


Figura 6.3. Algoritmo *Greedy set cover* utilizado por MMseqs2 para realizar los agrupamientos. Obtenido de https://github.com/soedinglab/mmseqs2/wiki.

Tabla 6.3. Determinación de eficiencia traduccional en los estadios tripomastigota metacíclicos y epimastigota.

Estadio	Comparación	p-valor
tripomastigota metacíclico	Overlap vs Represivos	0.00051
tripomastigota metacíclico	Overlap vs Total	6.85e-11
tripomastigota metacíclico	Overlap vs No Represivos	2.79e-29
tripomastigota metacíclico	Represivos vs No Represivos	4.61e-6
tripomastigota metacíclico	Represivos vs Total	0.15
tripomastigota metacíclico	Total vs No Represivos	1.62e-38
epimastigota	Overlap vs Represivos	3.56e-27
epimastigota	Overlap vs Total	6.15e-06
epimastigota	Overlap vs No Represivos	8.5e-26
epimastigota	Represivos vs No Represivos	2.75e-27
epimastigota	Represivos vs Total	3.58e-38
epimastigota	Total vs No Represivos	4.07e-97

6.3. Determinación de la regulación mediada por la presencia de uORFs en las regiones 5’ UTRs

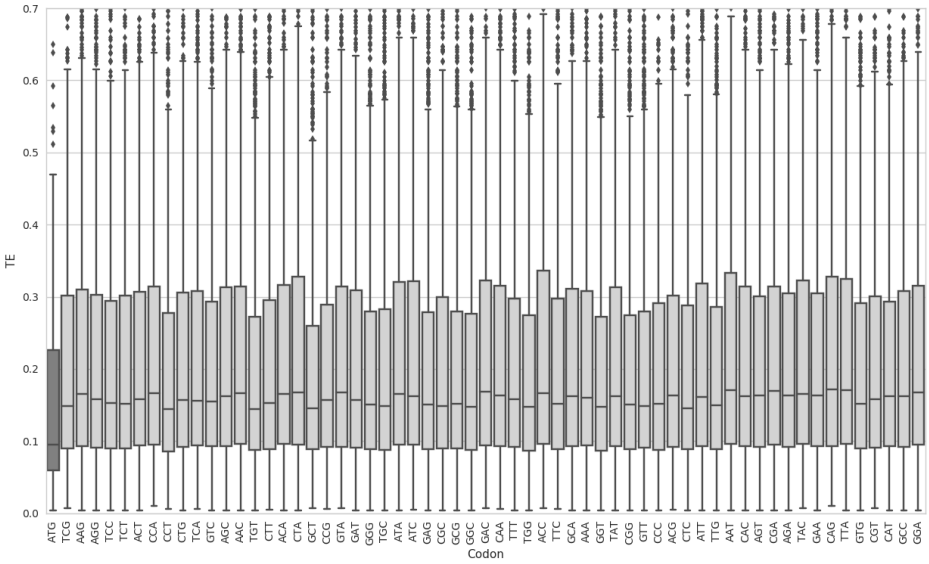


Figura 6.4. Comparación de la eficiencia traduccional de genes expresados en el estadio epimastigota, que en las regiones 5’ UTR contienen uORFs con potencial represivo (sin Overlaps). Cada caja representa la determinación de uORFs variando únicamente el codón iniciador.

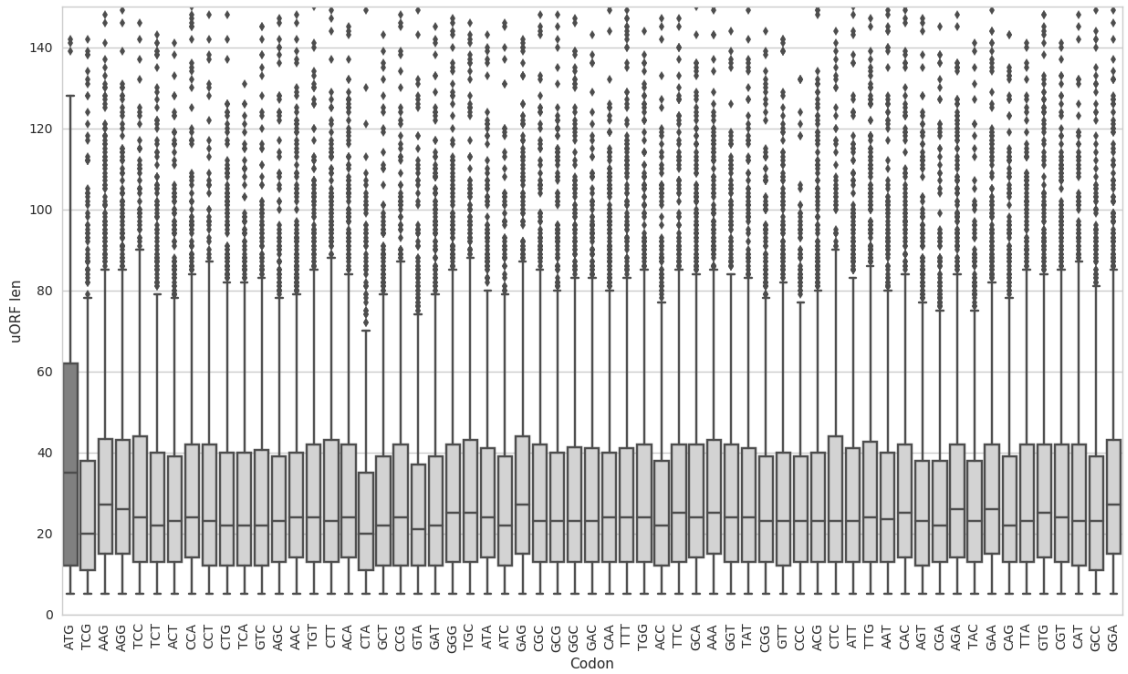


Figura 6.5. Comparación del tamaño en aminoácidos para uORF rerpesivos del estadio tripomastigota sanguíneo, cambiando en cada caso únicamente el codón iniciador.

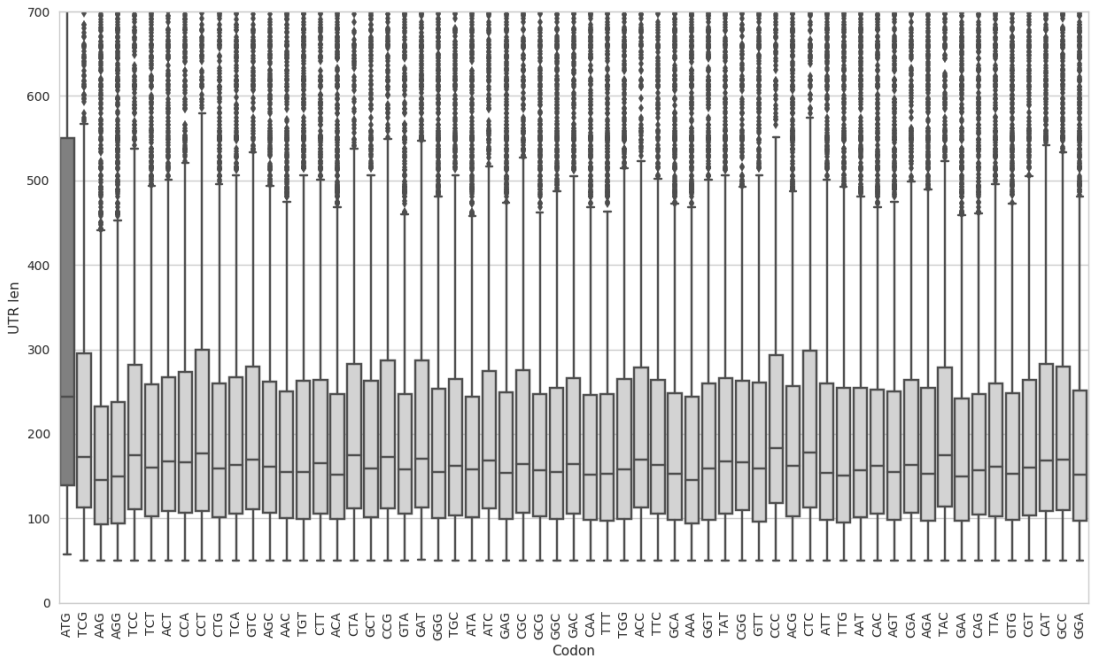


Figura 6.6. Comparación del tamaño de las regiones 5’ UTR determinada en el estadio tripomastigota de genes que contienen al menos un uORF con potencial represivos, cambiando en cada caso únicamente el codón iniciador.

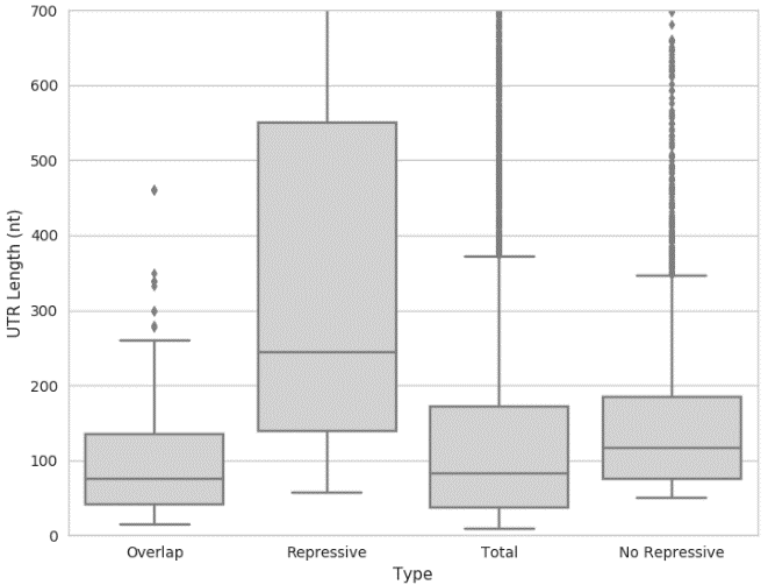


Figura 6.7. Comparación del tamaño de las regiones 5' UTR determinadas en el estadio tripomastigota, para las cuatro categorías analizadas.

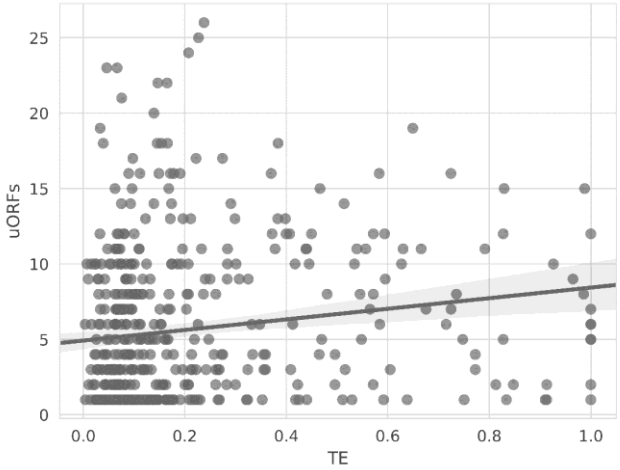


Figura 6.8. Correlación entre la densidad de uORF (uAUG) y la eficiencia traduccional del gen asociado en el estadio tripomastigota.