PEDECIBA

ÁREA BIOLOGÍA

TESIS DE DOCTORADO

Aproximaciones de genómica estructural y funcional en tripanosomátidos

Mag. Pablo Smircich

Laboratorio de Interacciones Moleculares

Facultad de Ciencias-Universidad de la República

Orientador: Dr. Beatriz Garat

Co-orientador: Dr. Najib El-Sayed

Tribunal:

Dr. Fernando Álvarez Valín

Dr. José Sotelo Silveira

Dr. Gustavo Cerqueira

Versión electrónica disponible en:

http://lim.fcien.edu.uy/tesis/smircich/tesis.pdf

Imagen de carátula y de separadores de capítulo extraída y modificada de:

http://labspace.open.ac.uk/PUB_527_1.0

Índice

Resumen

Los tripanosomátidos patógenos *Leishmania major*, *Trypanosoma brucei* y *Trypanosoma cruzi*, usualmente denominados Tritryps, son tres protozoarios estrechamente relacionados que causan enfermedades extremadamente prevalentes en humanos. Son parásitos que divergieron tempranamente en la línea eucariota, exhibiendo procesos moleculares distintivos. En particular, los mecanismos de expresión génica son muy divergentes, habiendo un número importante de interrogantes con respecto a su funcionamiento.

Las señales que intervienen en los procesos de expresión génica están poco caracterizadas. Los esfuerzos, que se han enfocado principalmente en la búsqueda de secuencias conservadas, han resultado poco fructíferos. En este contexto nos propusimos analizar otro tipo de señales. Particularmente nos centramos en repetidos de secuencia simple y estructuras secundarias del ADN. Por un lado, ya que los repetidos de dinucleotidos han sido involucrados en varias etapas de los procesos de expresión génica en diversos organismos, decidimos describir sus patrones de localización y frecuencia en los Tritryps. Encontramos que los repetidos complementarios presentan asimetría de hebra y no se encuentran distribuidos de manera uniforme sino que tienden a encontrarse cercano a los ORFs, alejados de los límites de los DGCs. Estas características indican que los repetidos de dinucleótidos pueden estar jugando un rol global a nivel del control de la expresión génica. Por otra parte, nos propusimos analizar los patrones de curvatura intrínseca del ADN. La presencia de esta estructura secundaria en sitios funcionalmente importantes ha sido reportada, facilitando la unión de factores proteicos y otros procesos que involucran cambios conformacionales en los ácidos nucleicos. Dos aproximaciones diferentes que permitieron evaluar la importancia funcional de esta señal, fueron llevadas a cabo. Realizamos tanto una búsqueda general a nivel genómico así como una búsqueda específica centrada en los promotores de ARNPI. La búsqueda de regiones de alta curvatura en los genomas completos de tripanosomátidos mostró un claro patrón de colocalización de esta señal con las regiones que presentan marcadores de cromatina característicos de sitios de inicio transcripcional. Este resultado constituye la primer señal conservada, intrínseca al ADN, descrita para estas regiones. En *T. brucei* observamos además una coincidencia de las regiones curvadas con la base J, que ha sido involucrada en el proceso transcripcional, y una llamativa concentración en las regiones subteloméricas caracterizadas por la presencia de genes y pseudogenes de VSG. Por su parte, el análisis de curvatura intrínseca de los promotores de ADNr

mostró la conservación característica de curvatura que ha sido descrita previamente para eucariotas en general. Interesantemente, demostramos que estos patrones son compartidos por los otros promotores de ARNPI en *T. brucei*. Este hallazgo pone en evidencia que la maquinaria de ARNPI podría requerir una curvatura intrínseca determinada, independientemente del gen que transcriba.

Teniendo en cuenta que la regulación post transcripcional es el principal nivel de control de la expresión génica en tripanosomátidos, la determinación de los ARNm que se encuentran siendo traducidos (traductoma), puede resultar especialmente adecuada para analizar los perfiles de expresión génica en estos organismos. La técnica recientemente desarrollada de determinación de perfiles de huellas ribosomales, permite indagar esta pregunta a través del secuenciado masivo de fragmentos protegidos por los ribosomas. En este trabajo, presentamos el análisis *in silico* de datos obtenidos mediante la utilización de esta técnica en epimastigotas de *T. cruzi,* buscando describir el traductoma general de este estadío y ahondar en los mecanismos de modulación de la expresión génica a nivel de la traducción. Los resultados muestran que los niveles de expresión calculados a partir del traductoma son un mejor reflejo de la cantidad de proteína celular cuando se compara con los niveles determinados por proteómica cuantitativa. El análisis de la eficiencia de traducción del conjunto de los genes muestra una variabilidad importante, indicando que la traducibilidad es un paso importante de regulación. Dentro de las familias de genes con desviaciones importantes en su eficiencia, encontramos las proteínas RRM (*RNA recognition motif*) de unión al ARN. Estos importantes factores regulatorios tienen una alta ocupación ribosomal, lo que les puede permitir cambios rápidos en sus niveles de expresión. Además, fue posible la identificación de regiones que presentan altos niveles de huellas ribosomales que pueden ser asociadas con pausas del ribosoma sobre el mensajero. Varios codones que se encuentran enriquecidos en estos sitios son poco frecuentes en el genoma en general, respaldando la hipótesis de que el uso de sinónimos puede modular la cinética de la traducción.

Globalmente, los datos aquí presentados constituyen una contribución novedosa en aspectos poco explorados de la estructura y dinámica funcional del genoma de los Tritryps.

Abstract

The pathogens *Leishmania major*, *Trypanosoma brucei* and *Trypanosoma cruzi* (Tritryps), are closely related protozoans that cause extremely prevalent human diseases. These parasites are early divergent eukaryotes which exhibit peculiar molecular processes. In particular, gene expression mechanisms are unusual with many questions that are still to be answered.

Signals involved in gene expression processes are poorly characterized. Research efforts focused at describing conserved sequence signals have not been successful. In this context, we aimed to analyze other types of signals. Specifically, we focused on simple sequence repeats and DNA secondary structure. As dinucleotide repeats have been involved at various stages of gene expression in different organisms, we decided to describe their location and frequency patterns in Tritryps. We found that the complementary pairs show strand asymmetry and they are not randomly distributed, as they are found predominantly close to the CDSs and far from the DGCs boundaries. These features point out that dinucleotide repeats may be globally involved in gene expression regulation. Besides, we decided to study DNA intrinsic curvature patterns. The presence of this secondary structure at functional sites has been reported as it may facilitate protein factor binding and other processes that involve nucleic acid conformational changes. We carried out two different approaches that allowed us to evaluate the functional importance of this signal. We performed both a genome wide search and also a specific search centered at RNAPI promoters. The inspection of highly curved regions in the trypanosomatids genomes showed a clear pattern of colocalization of this signal with regions associated to chromatin markers typical of transcription initiation sites. This result represents the first instance of a DNA intrinsic signal conserved in those regions. In *T. brucei* we also observed a correlation of highly curved regions with Base J, which has been involved in the transcriptional process, and a striking concentration at subtelomeric regions, which are characterized by the presence of VSG genes and pseudogenes. On the other hand, the analysis of rDNA promoters revealed the distinctive conservation of curvature that has been described previously for eukaryotes. Remarkably, we demonstrated that these structural profiles are shared by the other RNPI promoters present in *T. brucei*. This finding may imply that the transcription factor machinery requires a specific intrinsic curvature pattern in the promoter region, which is independent of the nature of the transcribed gene.

Considering that post transcriptional regulation is the main regulatory level of gene expression in trypanosomatids, the description of the mRNAs that are being transcribed (translatome), may be particularly suitable to analyze gene expression in these organisms. The ribosome profiling method, that has been recently developed, allows addressing this issue by high throughput sequencing of ribosome protected fragments. In this work, we present the *in silico* analysis of the data gathered by applying this approach to *T. cruzi* epimastigotes, aiming to describe the general translatome of this stage and get a better understanding of the regulatory mechanisms of gene expression that operate at the translational level. The results show that the inferred expression levels using ribosome footprints better reflect the ones obtained by quantitative proteomics. Translational efficiency analysis shows a high level of variability, indicating that translatability is an important regulatory step. Among gene families with important efficiency deviations, we found the ones coding for RNA binding RRM (*RNA recognition motif*) proteins. These important regulatory factors have a high ribosomal occupancy, which may allow rapid changes of their expression levels. Besides, we could identify regions that produce high numbers of ribosomal footprints which can be associated with sites of ribosome pausing over the mRNA. Various codons that are enriched at these positions are not frequent genome wide, supporting the hypothesis that codon usage may modulate translational kinetics.

In summary, the data here presented constitute a novel contribution in barely explored aspects of the structure and functional dynamics of the Tritryps genomes.

Introducción 1

1.1 Generalidades de Tritryps

Los tripanosomátidos patógenos *Leishmania major*, *Trypanosoma brucei* y *Trypanosoma cruzi*, usualmente denominados Tritryps, son tres protozoarios estrechamente relacionados (familia Trypanosomatidae, orden Kinetoplastida). Las enfermedades causadas por estos organismos: leishmaniasis, tripanosomiasis africana y enfermedad de Chagas, respectivamente, son extremadamente prevalentes en humanos.

1.1.1 Características clínicas y epidemiológicas

Las patologías antes mencionadas tienen una incidencia global, presentándose fundamentalmente en zonas tropicales y subtropicales e infectando a millones de seres humanos y animales en áreas rurales. Actualmente no hay vacunas eficaces dirigidas a estos organismos y las drogas que se emplean, desde hace ya muchos años, son inespecíficas y presentan alta toxicidad además de fenómenos de resistencia (Barrett, M. P. *et al.* 2003). Estas parasitosis son una parte sustancial de las denominadas *neglected diseases*, enfermedades que reciben esta denominación por ser altamente prevalentes en las regiones más pobres del mundo y que, sin embargo, reciben poco apoyo financiero para investigación y programas de salud específicos con el fin de combatirlas (Hotez, P. J. *et al.* 2007; O'Connell, D. 2007).

Leishmaniasis

La leishmaniasis es una zoonosis de manifestaciones clínicas variables que es causada por más de 20 especies diferentes de parásitos del género *Leishmania* y transmitida por moscas hembras, del género *Lutzomyia* en las Américas y del género *Phlebotomus* en el resto del mundo (*sandflies* o moscas de arena) (Kamhawi, S. 2006). La patología se distribuye en 98 países habiendo 12 millones de personas infectadas, 350 millones en situación de riesgo y entre 1 y 1,5 millones de nuevos casos cada año (Desjeux, P. 2004). El número de muertes reportadas es de unos 60000 casos anuales, aunque se especula que esta cifra está muy subestimada. Considerando que el número de infectados anualmente está en aumento en varios de los países endémicos, la leishmaniasis es claramente una patología de gran importancia sanitaria a nivel mundial (Handman, E. 2001; Desjeux, P. 2004).

Clásicamente se distinguen 3 tipos de leishmaniasis: visceral, cutánea y mucocutánea. El 90% de la leishmaniasis visceral ocurre en India, Bangladesh, Nepal, Sudan y Brasil mientras que el 90% de la cutánea se da en Afganistan, Algeria, Brasil, Irán, Perú, Arabia Saudita y Siria (Figura 1.1.1.1). La leishmaniasis cutánea

generalmente se cura sin tratamiento pero puede generar cicatrices de por vida. En su forma más severa, es muy difícil de tratar y desfigurante. La leishmaniasis mucocutánea causa daños en las cavidades oral-nasal y faríngea lo que conlleva a desfiguración de la cara. Este tipo de parasitosis se observa fundamentalmente en las Américas y es causada por *L. braziliensis* y *L. guyanensis.* Sin embargo, también han habido casos en Africa, Asia y Europa causados por infección con *L. donovani*, *L major* y *L. infantum* en pacientes inmunodeprimidos. La leishmaniasis visceral es la forma más severa de la patología siendo fatal si no se realiza un tratamiento adecuado. La patología se caracteriza por fiebre, pérdida de peso, esplenomegalia, hepatomegalia y anemia. Los pacientes tratados muchas veces sufren de una forma crónica de leishmaniasis cutánea (Desjeux, P. 2004).

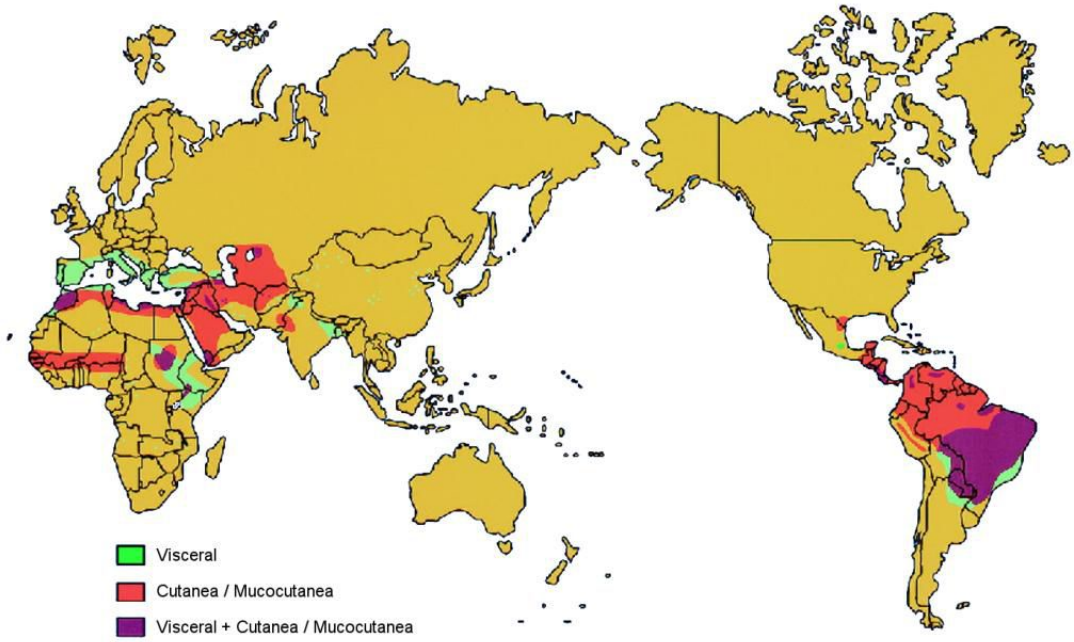


Figura 1.1.1.1 Distribución geográfica de la leishmaniasis. La referencia indica la patología presente en cada región. Extraído y modificado de (Handman, E. 2001).

Tripanosomiasis africana

Las tripanosomiasis africanas son causadas por tripanosomátidos que se transmiten mediante la mosca tsetse (*Glossinidae*). En particular, *T. brucei* causa la llamada Enfermedad del Sueño en humanos la cual, sin tratamiento, es fatal. Dado que el insecto vector se distribuye en 37 países de África Sub Sahara, las tripanosomiasis africanas son endémicas en esta extensa zona, habiéndose descrito casos de ganado afectado en todos los países de la región. Sin embargo, solo en 20 de ellos la enfermedad se presenta en humanos. La Organización Mundial de la Salud

(OMS) estima el número de casos en alrededor de 300000, estando 60 millones de personas en riesgo de contagio. Unos 100000 pacientes mueren cada año de Enfermedad del Sueño y 45000 nuevas personas son infectadas en este mismo período (Figura 1.1.1.2). El parásito no genera inmunidad por lo que las reinfecciones son comunes. En los años 60 se realizaron campañas masivas de tamizaje y tratamiento de la enfermedad lo cual hizo que las cifras de infectados decrecieran drásticamente. Sin embargo, el abandono de estas políticas llevó a que la enfermedad resurgiera a mediados de los años 90 con niveles comparables a los de mediados de siglo. En los últimos años, el número de infectados está descendiendo nuevamente gracias a la reimplantación de los programas de control (Cattand, P. *et al.* 2001).

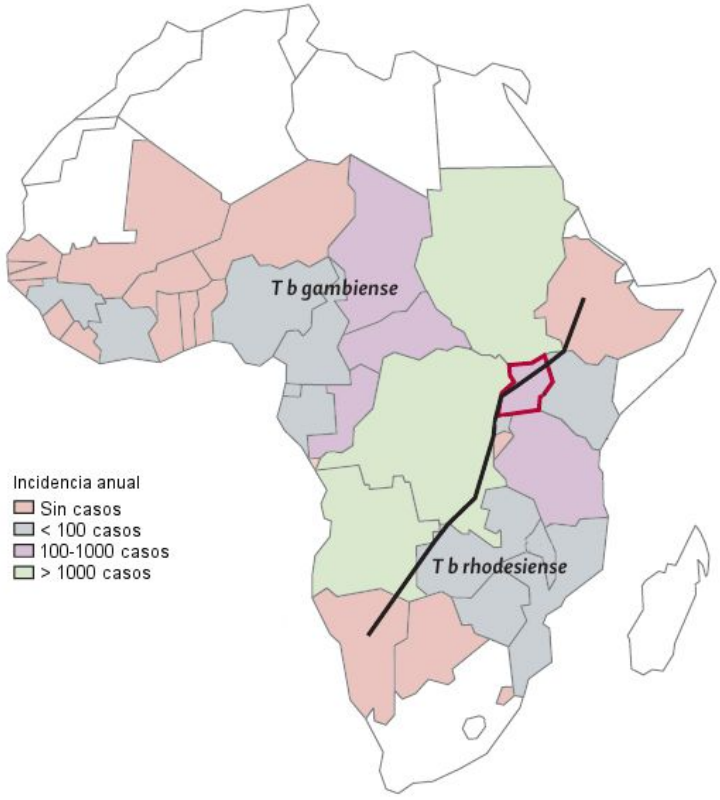


Figura 1.1.1.2 Distribución geográfica de *T. brucei*. La escala indica el número de personas infectadas. La línea negra señala el límite de las áreas de infección de *T. brucei gambiense* y *T. brucei rhodesiense*. Extraído y modificado de (Brun, R. *et al.* 2010)

Estas patologías se caracterizan por atacar el sistema nervioso y puede tener una forma aguda, en el caso de infección con *T. brucei rhodesiense,* y una presentación más crónica, en el caso de infección con *T. brucei gambiense*. La subespecie *T. brucei brucei* no es infectiva en humanos siendo los animales el reservorio natural de este parásito. Una vez que los parásitos son transmitidos por la mosca, se multiplican en el

espacio extracelular en la zona de la picadura produciendo una lesión característica. La fase hemolinfática de la enfermedad está caracterizada por una circulación de los parásitos por la sangre y la linfa produciendo olas de parasitemia que están acompañadas típicamente de episodios febriles, dolores de cabeza, picazón, linfadenopatía y circunstancialmente hepatoesplenomegalia. Estos síntomas tienden a hacerse menos severos con el desarrollo de la enfermedad y luego de un tiempo los parásitos son capaces de atravesar activamente la barrera hematoencefálica y, por lo tanto, invaden el sistema nervioso central y el fluido cerebroespinal llevando a la etapa de meningoencefalitis (Masocha, W. *et al.* 2007; Brun, R. *et al.* 2010). Esta segunda etapa está caracterizada por trastornos del sueño, los cuales dan a la enfermedad su nombre, así como desordenes neurosiquiátricos. El tiempo de evolución entre ambas etapas de la enfermedad varía con la cepa, siendo más corto (semanas o meses) con la infección por *T. brucei rhodesiense*. En este caso, el principal reservorio es el ganado y entonces la infección en humanos se da principalmente por contagio desde un animal. Sin embargo, para el caso de *T. brucei gambiense* la segunda etapa es alcanzada luego de años post infección, siendo el humano el principal reservorio del parásito y el contagio, en este caso, se da más frecuentemente de persona a persona (Odiit, M. *et al.* 1997; Checchi, F. *et al.* 2008). Otras formas de contagio son posibles, pero se estima que su frecuencia es despreciable.

Enfermedad de Chagas

El parásito *Trypanosoma cruzi* (*T. cruzi*) es el agente etiológico de la enfermedad de Chagas la cual es transmitida al huésped mamífero por diversas especies de insectos hematófagos de la familia Reduvidae (Chagas, C. 1909; De Souza, W. 2002). Esta zoonosis se ha encontrado en 150 especies de animales tanto salvajes como domésticos además de humanos. Las estimaciones indican que esta enfermedad afecta a cerca de 10 millones de personas en 21 países de América Central y del Sur, estando varios millones de personas en riesgo de contagio y se le atribuyen 13000 muertes anuales (WHO 1998; WHO 2002; Remme, J. H. F. *et al.* 2006). La infección del mamífero ocurre a través de mucosas o piel dañada y se da cuando el parásito es depositado en estos sitios junto con las heces del insecto triatomino cuando éste se alimenta. Se reconocen tres tipos de ciclos de transmisión. El ciclo doméstico es el responsable de mantener la infección en humanos y ocurre en regiones rurales que presentan construcciones precarias. En este caso, los reservorios del parásito son además de los humanos los animales domésticos. En el ciclo selvático los animales infectados son los roedores, marsupiales y otros animales salvajes. Por último, en el ciclo periselvático, los animales domésticos y ocasionalmente los humanos son

infectados por triatominos selváticos que son atraídos a las viviendas. Además de la infección vectorial, la parasitosis se puede transmitir por otras vías. La principal de éstas es la infección por transfusiones sanguíneas con sangre contaminada e incluso se han reportado casos de contagio por trasplantes de tejidos. En los últimos años, se han detectado casos de enfermedad de Chagas en países no endémicos, principalmente Estados Unidos y Europa (Bern, C. *et al.* 2011; Cantey, P. T. *et al.* 2012). El esparcimiento de la enfermedad se ha dado a causa de las migraciones de individuos infectados a estas regiones (Figura 1.1.1.3). Existe también transmisión vertical ya que en un 5% de los nacimientos con madres portadoras, los recién nacidos resultan contagiados. Por otro lado, la infección puede transmitirse por vía oral debido a la ingesta de alimentos contaminados con formas infectivas del parásitos (Yoshida, N. 2008). Desde hace varias décadas el control vectorial llevado a cabo de forma sistemática en los países del cono sur ha bajado las cifras de infección y mortalidad (de 50000 muertes al año en el 1991 a 13000 en el año 2001). Este programa de control que fue ampliamente promovido en el año 1991 en toda la región, consiguió erradicar en nuestro país la transmisión vectorial en el año 1997 (Remme, J. H. F. *et al.* 2006). De todas formas, un estudio reciente que traduce a costos monetarios las consecuencias de la enfermedad de Chagas, muestra claramente que continúa siendo un problema extremadamente serio a nivel regional y mundial. En este trabajo se ha estimado que el costo mundial de esta enfermedad es de alrededor de 7 billones de dólares anuales, superando otras enfermedades como el cáncer cervical y el rotavirus (2 billones y 4-7 billones respectivamente). Sin embargo, la inversión de dinero e investigación dirigidas al control de estas últimas patologías es mucho mayor que lo invertido en la Enfermedad de Chagas (Lee, B. Y. *et al.* 2013).

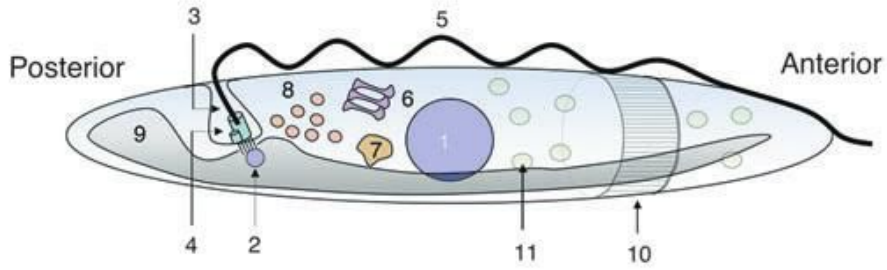
Figura 1.1.1.3 Distribución geográfica de la Enfermedad de Chagas. En rojo se muestra la zona endémica. Las flechas indican los movimientos migratorios y los números indican la cantidad de personas infectadas en países no endémicos. Extraído y modificado de (Coura, J. R. *et al.* 2010).

La enfermedad de Chagas presenta dos fases en su evolución clínica. Una fase aguda que comprende entre los dos y seis primeros meses de infección y se caracteriza por una sintomatología de intensidad y duración variables y por altas parasitemias. La mayor parte de las veces pasa desapercibida y en algunos casos, los síntomas van desde estado febril e inflamación en la zona de ingreso del parásito (Chagoma en la piel y signo de Romanha en la membrana ocular) hasta la muerte del hospedero. Estos síntomas desaparecen a los pocos meses y la enfermedad evoluciona hacia una fase crónica con parasitemia baja y asintomática. Alrededor de un 30-40% de los pacientes en esta última fase desarrolla síndrome chagásico, que toma varias formas dependiendo del sitio en el que los amastigotas se desarrollen. Las consecuencias más serias de la infección son la insuficiencia cardíaca (miocardiopatía) y la pérdida de control sobre el tracto digestivo debido a la presencia de parásitos en el sistema nervioso. Además, en algunos pacientes se observa desarrollo de megacolon y megaesófago causados por la invasión de los ganglios entéricos con células T citotóxicas y la pérdida de la inervación nerviosa del músculo (da Silveira, A. B. *et al.* 2007).

1.1.2 Características estructurales y ciclos de vida

Los Tritryps son parásitos que divergieron tempranamente en la línea eucariota (Haag, J. *et al.* 1998). Esta distancia filogenética hace que su biología sea excepcional

(Smith, D. F. *et al.* 1996). La Figura 1.1.2.1 muestra un esquema general de una célula de un tripanosomátido, destacando las características estructurales más relevantes.



1 Núcleo
2 Kinetoplasto
3 Bolsillo flagelar
4 Cuerpo Basal
5 Flagelo
6 Golgi
7 Lisosoma
8 Endosomas
9 Mitocondria
10 Citoesqueleto de microtúbulos
11 Glicosomas

Figura 1.1.2.1 Esquema de una célula de *T. brucei* mostrando sus características estructurales principales. La organización general es similar entre los diferentes tripanosomátidos, aunque las localizaciones relativas de los organelos y la morfología varían en los organismos y los estadíos particulares. Extraído y modificado de (Matthews, K. R. 2005).

En el esquema se muestra una célula de *T. brucei* la cual es elongada y tiene un citoesqueleto de microtúbulos polarizado. Este define la forma de la célula y permanece intacto durante el ciclo celular. Los microtúbulos están orientados todos con la misma polaridad, con los extremos negativos hacia el sector anterior y los positivos hacia el extremo posterior.

Entre las características más sobresalientes se encuentra la presencia de una única gran mitocondria. El ADN mitocondrial, que representa hasta un 30% del ADN celular total, conforma una estructura distintiva denominada kinetoplasto. Aunque la mitocondria ocupa gran parte del volumen celular, la localización del kinetoplasto es característica de cada etapa del ciclo de vida. Esta estructura está físicamente ligada al cuerpo basal, que se encuentra en la base del flagelo, y se localiza de forma perpendicular a su eje (Souto-Padron, T. *et al.* 1984; Ogbadoyi, E. O. *et al.* 2003). El kinetoplasto se compone de un entramado de fibras de ADN que se organizan en un conjunto de moléculas circulares topológicamente relajadas conocidas como maxicírculos y minicírculos. Existen varios miles de minicírculos, los cuales tienen un tamaño que va de 0,5 a 2,5 kb dependiendo de la especie y unas pocas decenas de maxicírculos, cuyo tamaño varía entre 20-40 kb (Shapiro, T. A. *et al.* 1995). Los maxicírculos son los encargados de codificar los ARNr y proteínas sintetizadas en la mitocondria de forma análoga a los que ocurre en otros eucariotas. Sin embargo, en

tripanosomas se da una edición generalizada de los ARNs mensajeros (*editing*), siendo ésta otra de las características distintivas de este grupo de organismos. La edición de mensajeros ocurre por adición o deleción de uridinas en lugares específicos, los cuales son definidos por los llamados ARNs guías que son codificados por los minicírculos (Shaw, J. M. *et al.* 1988; Simpson, L. *et al.* 2000; Madison-Antenucci, S. *et al.* 2002; Aphasizhev, R. *et al.* 2003; Simpson, L. *et al.* 2003).

El organelo denominado glicosoma es también característico de estos parásitos y marca una particularidad metabólica (Cazzulo, J. J. 1994). Varias etapas de la vía de la glucolisis tienen lugar dentro del mismo. Esta compartimentación de las enzimas glucolíticas es esencial para la sobrevida del parásito, tanto en estadíos que dependen exclusivamente de la glucolisis para la obtención de ATP como en estadíos que obtienen ATP por otras vías (Opperdoes, F. R. *et al.* 1977; Guerra-Giraldez, C. *et al.* 2002).

Los acidocalcisomas son organelos capaces de transportar protones y calcio y han sido identificados en todos los miembros de la familia Trypanosomatidae y muchos miembros del *phylum* Apicomplexa (Docampo, R. *et al.* 2005; Docampo, R. *et al.* 2011). Los acidocalcisomas están involucrados en varias funciones incluyendo: (i) almacenamiento de calcio, magnesio, sodio, potasio, zinc, hierro, pirofosfato inorgánico, polifosfato, (ii) homeostasis del pH y (iii) osmorregulación participando en estrecha asociación con la vacuola contráctil. Esta última está formada por varios túbulos conectados a una vacuola central localizada cerca del bolsillo flagelar (Linder, J. C. *et al.* 1977). Su función en el parásito no está del todo clara, pero se postula que podría tener un rol en la regulación del metabolismo de fosfatos (Ulrich, P. N. *et al.* 2011).

Por otro lado, la envoltura nuclear es conservada durante la división celular y el huso mitótico se forma dentro del núcleo con los microtúbulos convergiendo en polos opuestos (Vickerman, K. *et al.* 1970; Drechsler, H. *et al.* 2012). Durante la mitosis, la cromatina no se condensa en cromosomas discretos, aunque si se producen cambios de compactación en diferentes etapas del ciclo. Además, se ha demostrado la permanencia del nucleolo durante toda la división en *T. brucei* (De Souza, W. 2002; Schenkman, S. *et al.* 2011).

La vía secretoria de *T. cruzi* consiste en el retículo endoplásmico, el complejo de Golgi y un sistema de vesículas que emergen del Golgi y migran hacia el bolsillo flagelar donde se fusionan y descargan su contenido (Araripe, J. R. *et al.* 2004). En *T. brucei* existe una única agrupación del aparato de Golgi la cual está ubicada entre el

núcleo y el bolsillo flagelar. Durante la replicación celular, esta estructura se divide a partir del propio retículo, habiendo sido usado este parásito como modelo de la generación de este organelo (He, C. Y. *et al.* 2004).

Como en otros eucariotas, en estos organismos la endocitosis es el mecanismo básico de internalización de macromoléculas las cuales son degradadas en el sistema de endosomas-lisosomas. Estudios en *T. cruzi* que, como los otros tripanosomátidos, son células altamente polarizadas, han evidenciado que la actividad endocitica está restringida al bolsillo flagelar y al citosoma (invaginación profunda de la membrana plasmática que puede llegar hasta el núcleo) (de Souza, W. *et al.* 2009). Los estudios han mostrado particularidades en esta vía. Por un lado, la endocitosis ocurre únicamente en la forma epimastigota. Además, las moléculas incorporadas por lo endosomas son enviadas a unas estructuras especiales denominadas reservosomas que están localizados en la región posterior del parásito (Porto-Carreiro, I. *et al.* 2000).

Ciclo de vida de *Leishmania*

El ciclo de vida de las leishmanias tiene 2 estadíos diferenciables morfológicamente (Figura 1.1.2.2). El estadío promastigota se encuentra en el insecto vector donde se replica de forma extracelular. Durante este estadío se distinguen a su vez dos etapas. El promastigota procíclico se encuentra en el aparato digestivo del insecto, mientas que el promastigota metacíclico se encuentra en la probóscide y es el que infecta al hospedero mamífero. Por su parte, el estadío amastigota se replica de forma intracelular en el hospedero definitivo (Handman, E. 1999). Los mamíferos son infectados cuando el insecto inyecta el estadío infectivo promastigota metacíclico del parásito durante la ingesta de sangre. Los promastigotas son fagocitados por macrófagos y otros tipos de células fagocíticas. Durante el periodo de estadía en el medio extracelular, los promastigotas son combatidos por el complemento, pero existen proteínas de membrana del parásito que son capaces de detener la respuesta inmune el tiempo suficiente como para lograr que parte de los parásitos sean fagocitados. La metaloproteasa gp63 es una de las macromoléculas más relevantes en estos procesos (junto con otras como los lipofosfoglicanos, los glicosil-inositol-fosfolipidos y la cistein-proteasa B) siendo capaz de clivar el factor C3b impidiendo así la activación del complejo de ataque de membrana del complemento. Además, esta proteína interviene en la etapa de fijación a las células a ser infectadas para la posterior internalización del parásito (Brittingham, A. *et al.* 1995; Yao, C. *et al.* 2003; Olivier, M. *et al.* 2012). Dentro del fagolisosoma los parásitos se diferencian a amastigotas los cuales se dividen, lisan la célula y son capaces de reinfectar otras

células. Al ser ingeridos por un insecto, los amastigotas vuelven a diferenciarse en promastigotas procíclicos en el intestino, desde donde migran a la probóscide cerrando el ciclo.

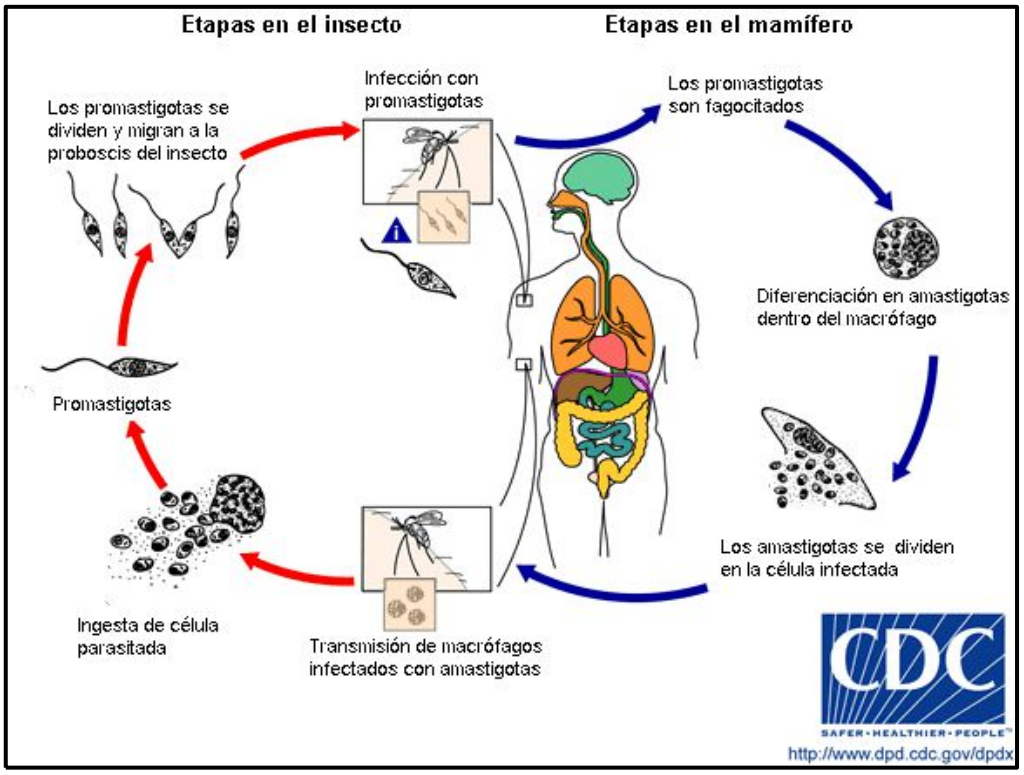


Figura 1.1.2.2 Esquema mostrando el ciclo de vida de *Leishmania*. Extraído y modificado del sitio web del *Centers for Disease Control and Prevention* (USA, www.cdc.gov).

Además del ciclo clásico, Akopyants y cols. encontraron evidencia de que los promastigotas son capaces de intercambiar información genética, lo cual es compatible con un ciclo sexual y existencia de meiosis en este estadío. En estos estudios, los investigadores consiguieron obtener parásitos resistentes a dos drogas diferentes partiendo de dos cepas, cada una resistente a una de las mismas. Aunque las tasas de intercambio sean bajas, el mecanismo puede, de todas formas, colaborar a la diversidad fenotípica en poblaciones naturales (Akopyants, N. S. *et al.* 2009).

Ciclo de vida de *T. brucei*

T. brucei tiene un ciclo de vida complejo con varios estadíos de desarrollo, todos ellos extracelulares (Matthews, K. R. 2005; Sharma, R. *et al.* 2009). Este ciclo comienza cuando una mosca tsetse infectada inyecta la forma tripomastigota metacíclica a través de la picadura para alimentarse. Los parásitos ingresan por el

sistema linfático y de ahí a la sangre en donde se diferencian a tripomastigotas sanguíneos. El parásito se replica en el torrente sanguíneo en la forma *slender* que es morfológicamente delgada. El aumento en el número de estas formas produce la acumulación de un factor soluble, denominado *stumpy-induction factor* (SIF), el cual genera un arresto en el ciclo celular en la población y el subsiguiente cambio a la forma no proliferativa denominada *stumpy* (Vassella, E. *et al.* 1997). Esto sirve para dos propósitos. En primer lugar, se limita el aumento en el número de parásitos y por lo tanto, prolonga la supervivencia del huésped y, así, la probabilidad de transmisión de la enfermedad. En segundo lugar, la detención de la población de tripanosomas en la fase G1 del ciclo celular asegura que los cambios morfológicos que se deben producir cuando el mismo ingresa en la mosca se coordinen con la reentrada en el ciclo. Cuando una mosca se alimenta de un mamífero infectado, los tripomastigotas sanguíneos son ingeridos por el insecto y en el intestino medio se diferencian a tripomastigotas procíclicos que son replicativos. Estos dejan el intestino medio para diferenciarse en epimastigotas los cuales se replican por fisión binaria en las glándulas salivares. Allí los epimastigotas se diferencian a tripomastigotas metacíclicos, los cuales serán capaces de infectar nuevamente un huésped mamífero, cerrando el ciclo (Figura 1.1.2.3). Interesantemente, hace varios años se describió que los parásitos son capaces de intercambiar información genética (Jenni, L. *et al.* 1986). Este intercambio no sería obligatorio para la compleción del ciclo de vida y se daría en condiciones de coinfección de un insecto con dos cepas distintas del parásito, aunque hay evidencia de intercambio también dentro de la misma cepa (Peacock, L. *et al.* 2009). Más recientemente, hallazgos experimentales sugieren que este intercambio se da por un proceso meiótico durante el período en que el parásito se encuentra en las glándulas salivares de la mosca (Peacock, L. *et al.* 2011).

Morfológicamente los diferentes estadíos de *T. brucei* son similares, siendo el posicionamiento del kinetoplasto relativo al extremo posterior la característica más distintiva entre ellos (Fenn, K. *et al.* 2007; Sharma, R. *et al.* 2009). En este sentido se puede observar que, en las formas sanguíneas, esta estructura se encuentra próxima al extremo posterior. En el tripomastigota procíclico, el kinetoplasto se encuentra aún posterior aunque en una ubicación más cercana al núcleo, mientras que en los epimastigotas, el kinetoplasto se encuentra anterior respecto al núcleo que tiene ubicación central en todos los estadíos. No se conoce con certeza el motivo de estos cambios aunque, como la ubicación del kinetoplasto influencia el sitio de anclaje del flagelo, se piensa que pueden estar relacionados con los diferentes requerimientos de

motilidad de los diferentes estadíos (Sherwin, T. *et al.* 1989; Garcia-Salcedo, J. A. *et al.* 2004).

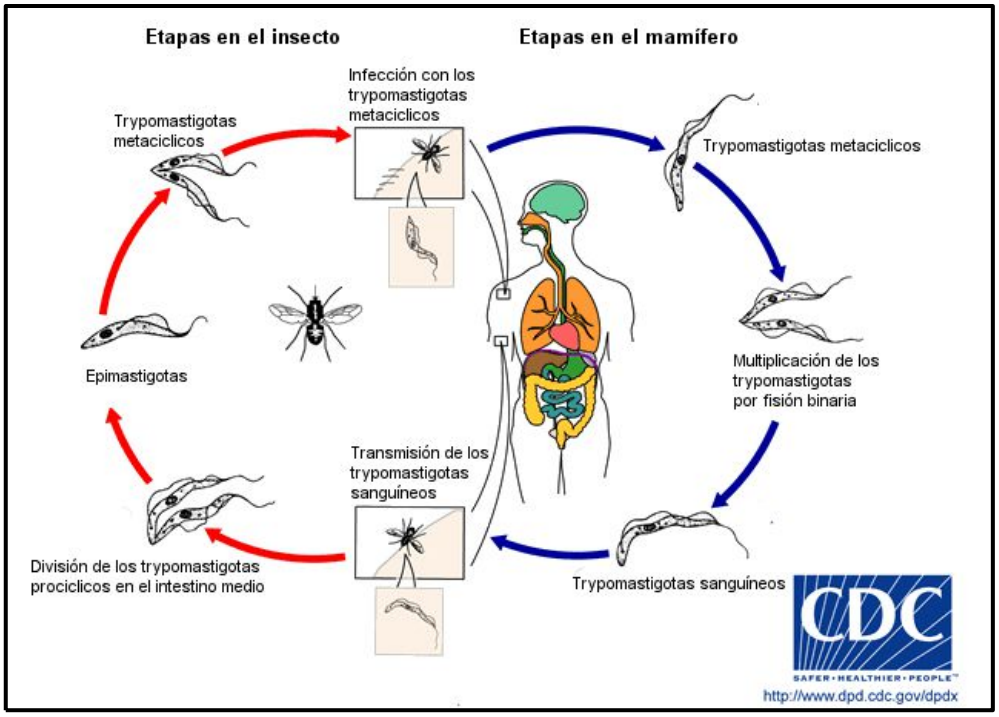


Figura 1.1.2.3 Esquema mostrando el ciclo de vida de *T. brucei*. Extraído y modificado del sitio web del *Centers for Disease Control and Prevention* (USA, www.cdc.gov).

Ciclo de vida de *T. cruzi*

El ciclo de vida de *T. cruzi* transcurre entre dos hospederos: uno invertebrado, el insecto triatomino o vinchuca (*Triatoma infestans* y *Triatoma rubrovaria* en nuestro país), y otro vertebrado, el hospedero mamífero. El parásito presenta al menos cuatro estadíos principales relacionados con sus distintos entornos:

La forma epimastigota, se desarrolla en la luz del intestino del insecto, es replicativa y característicamente presenta un flagelo anclado cerca del centro del cuerpo del parásito. En este estadío, el kinetoplasto se ubica anterior al núcleo y tiene forma de disco (De Souza, W. 2002). La forma tripomastigota metacíclica, se encuentra en la ampolla rectal del insecto, no es replicativa y tiene un flagelo libre anclado a una membrana ondulante en el cuerpo. El núcleo se encuentra cercano a la parte posterior de su cuerpo y el kinetoplasto tiene una ubicación posterior al núcleo con forma de canasto abierto (Tyler, K. M. *et al.* 2001). La forma amastigota se encuentra en las células del hospedero mamífero, siendo fundamentalmente intracelulares y son la forma replicativa en el mamífero. No tienen flagelo protuberante y su morfología es de forma oval o redondeada (de Carvalho, T. U. *et al.* 1986; De Souza, W. 2002; Yoshida,

N. *et al.* 2008; Souza, W. 2009). Los tripomastigotas sanguíneos son casi indistinguibles morfológicamente de los metacíclicos, aunque existen diferencias a nivel de su biología molecular.

En el ciclo de vida, los tripomastigotas metacíclicos presentes en las heces de un insecto infectado ingresan al hospedero mamífero a través de las mucosas o piel dañada. Una vez en el mamífero los parásitos son capaces de ingresar a las células a través de vacuolas endocíticas, llamadas vacuolas parasitóforas. Esta internalización es inducida por *T. cruzi* de forma activa (de Souza, W. *et al.* 2013). Una vez internalizados, los parásitos son liberados al citoplasma donde sufren una serie de cambios morfológicos que llevan a la diferenciación al estadío amastigota. Estos se replican activamente hasta ocupar casi la totalidad del citoplasma celular (Dvorak, J. A. *et al.* 1973; Andrews, N. W. *et al.* 1987). Luego de este período, los amastigotas se diferencian a tripomastigotas sanguíneos los cuales son capaces de producir la lisis celular (Costales, J. *et al.* 2007). Estos tripomastigotas serán capaces de infectar células cercanas, circular por el torrente sanguíneo hasta infectar células de otros tejidos o podrán ser ingeridos por un triatomino. Existen evidencias de que amastigotas presentes en la célula lisada son capaces también de infectar células de hospedero, colaborando entonces con el proceso de dispersión y persistencia del parásito dentro del mamífero (Ley, V. *et al.* 1988; Mortara, R. A. 1991; Fernandes, M. C. *et al.* 2012). Si los tripomastigotas sanguíneos son ingeridos por un vector durante su alimentación, estos diferencian a la forma epimastigota en el tracto digestivo en donde son capaces de replicar activamente. En la parte final del tracto digestivo, los epimastigotas se diferencian a tripomastigotas metacíclicos, los cuales serán expulsados del insecto junto con las heces permitiendo una nueva infección de un mamífero y cerrando el ciclo (Caradonna, K. L. *et al.* 2011) (Figura 1.1.2.4). Como comentamos anteriormente, desde hace varios años se conoce que en determinadas circunstancias *T. brucei* es capaz de reproducirse sexualmente. Sin embargo, la existencia de intercambio genético en *T. cruzi* fue sugerida más recientemente y ha sido más debatida. Aunque el intercambio ha sido demostrado en el laboratorio, el mecanismo por el cual se produce no está claro y hay poca evidencia de la frecuencia de este fenómeno en la naturaleza. Nuevos estudios han encontrado que esto podría ser más frecuente de lo que se cree actualmente, apuntando a la importancia del proceso en la generación de variabilidad de las poblaciones de parásitos (Ramirez, J. D. *et al.* 2012; Roellig, D. M. *et al.* 2013). El mecanismo por el cual tiene lugar el intercambio resulta claramente diferente al de *T. brucei*, dándose de forma no

mendeliana y a través de la generación de híbridos (Gaunt, M. W. *et al.* 2003; Minning, T. A. *et al.* 2011).

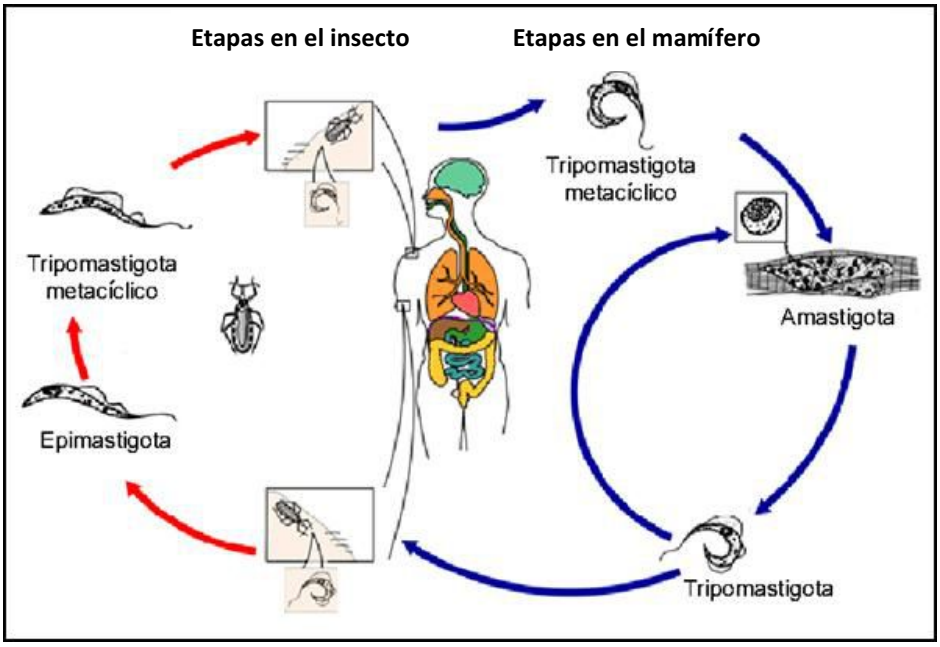


Figura 1.1.2.4 Esquema mostrando el ciclo de vida de *T. cruzi*. Extraído y modificado del sitio web del *Centers for Disease Control and Prevention* (USA, www.cdc.gov).

1.2 Genómica estructural y funcional de Tritryps

1.2.1 Genoma

Los estudios de genómica en Tritryps comenzaron a mediados de los años 90 mediante el análisis de la secuencia de grandes fragmentos de ADN clonados en librerías de cósmidos, cromosomas artificiales de bacterias y de levaduras. Estos estudios junto con análisis de electroforesis de campo pulsado y secuenciado de ESTs (*expressed sequence tags*), permitieron construir los primeros mapas físicos de los genomas de Tritryps (Blackwell, J. M. *et al.* 1999). En 1999 se secuenciaron 257 Kb del cromosoma 1 de *L. major* que lo abarcan casi completamente, siendo la primer secuencia de un cromosoma competo en estos organismos (Myler, P. J., Audleman, L., deVos, T., Hixson, G., Lemley, C., Magness, C. Rickel, E., Sisk, E., Sunkin, S., Swartzell, S., Westlake, T., Bastien P., Guoliang, F., Ivens, A., Stuart K. 1999). En el año 2005 la publicación de los genomas de los 3 tripanosomátidos modelo y los estudios derivados posteriores han incrementado nuestra comprensión de la biología de estos parásitos y dado pistas importantes de los mecanismos de interacción hospedero parásito (Berriman, M. *et al.* 2005; El-Sayed, N. M. *et al.* 2005; El-Sayed, N. M. *et al.* 2005; Ivens, A. C. *et al.* 2005).

El genoma de *L. major*

La cepa *L. major* Friedlin fue la seleccionada como cepa de referencia. El genoma haploide de este parásito resultó de aprox. 33 Mb pudiéndose ensamblar en 36 cromosomas completos que varían en tamaño de entre 0.28 a 2.8 Mb (Wincker, P. *et al.* 1996; Ivens, A. C. *et al.* 2005). El análisis reveló la presencia de 911 genes para ARN no codificantes, 39 pseudogenes y 8272 genes codificantes, gran parte de los cuales no se les pudo asignar función conocida. Además, 3083 de las secuencias codificantes se agrupan en 662 familias de genes relacionados. La mayor parte de las familias génicas de menor número de miembros se encuentran duplicadas en tándem, mientras que los miembros de las familias más grandes se encuentran distribuidos en múltiples *loci*. Poco después, los genomas de *Leishmania braziliensis* y *Leishmania infantum* fueron secuenciados, permitiendo la comparación entre estos grupos relacionados. Aunque la divergencia entre ellos se produjo entre 20 y 100 millones de años atrás (Lukes, J. *et al.* 1997), sorprendentemente los genomas resultaron ser muy similares entre si (Peacock, C. S. *et al.* 2007). Uno de los resultados más llamativos fue la observación que, a diferencia de lo que ocurre con las otras *Leishmanias* secuenciadas, *L. braziliensis* conserva los componentes de un posible sistema de interferencia de ARN. Resultados posteriores permitieron comprobar experimentalmente la existencia de este mecanismo en *L. braziliensis* así como en *L. guyanensis* y *L. panamensis* (Lye, L. F. *et al.* 2010).

El genoma de *T. brucei*

La cepa 927 de la subespecie *T. brucei brucei* fue seleccionada como referencia. Este genoma se distribuye en 11 cromosomas de gran tamaño (en el orden de las Mb) además de en un número no especificado de cromosomas de tamaño pequeño o medio (de entre 30 y 700 Kb). El genoma haploide comprende 26 Mb donde se pudieron predecir 9068 genes que incluyen 900 pseudogenes (Berriman, M. *et al.* 2005). El análisis de los subtelómeros (definidos como las regiones que se encuentran entre los telómeros y el primer gen *housekeeping*) mostró que estos contienen varios cientos de miembros de la familia de glicoproteínas variables de superficie (*variant surface glycoprotein*, VSG). Se lograron anotar en el genoma 806 miembros de esta familia, número que seguramente es una subestimación ya que no se conoce el contenido en los cromosomas pequeños e intermedios así como en regiones subteloméricas que no se han secuenciado aún. Esta familia es la responsable del fenómeno de variación antigénica, fundamental en el mecanismo de evasión del sistema inmune del hospedero mamífero en este parásito que es exclusivamente

extracelular. El mecanismo consiste en la expresión en la membrana celular de aproximadamente 10^7 copias de una única proteína VSG formando una cubierta que enmascara otros antígenos de la superficie del parásito. Esto genera que la respuesta inmunitaria sea dirigida hacia la proteína VSG que está siendo expresada en un determinado momento. Cuando esta proteína es reemplazada por otra del repertorio de VSG, la respuesta inmune deja de ser efectiva y, por lo tanto, el clon que contiene esta nueva VSG es capaz de replicarse generando una nueva ola de parasitemia (Borst, P. *et al.* 1998; Borst, P. 2002; Stockdale, C. *et al.* 2008; Weirather, J. L. *et al.* 2012). Aunque como mencionamos antes, el número de miembros de la familia de VSG es de cientos, una gran parte de los mismos son proteínas no funcionales. De hecho solo 57 genes (7%) presentan marcos de lectura abiertos con características para codificar una proteína funcional, mientras que el resto son mayormente pseudogenes. Se postula que estos pueden actuar como un reservorio de secuencias que pueden ser usados para generar variantes de VSG por recombinación (Weirather, J. L. *et al.* 2012). Además el genoma cuenta con VSG que se expresan en el estado metacíclico (MVSG), aunque su número es mucho menor. Estas proteínas son las que entran en contacto con el hospedero en las primeras etapas de la infección (Ginger, M. L. *et al.* 2002).

El genoma de *T. cruzi*

En este caso la cepa elegida para ser secuenciada fue la CL Brener, la cual resultó ser una cepa híbrida, presentando dos haplotipos diferentes (El-Sayed, N. M. *et al.* 2005). Actualmente las cepas de *T. cruzi* se asignan a 6 grupos principales: *T. cruzi* I, *T. cruzi* II, *T. cruzi* III, *T. cruzi* IV, *T. cruzi* V y *T. cruzi* VI, perteneciendo CL Brener a este último que surge a partir de los grupos II y III (Zingales, B. *et al.* 2009; Teixeira, S. M. *et al.* 2012). Este hecho sumado a la gran cantidad de secuencias repetitivas encontradas generó problemas al momento de realizar el ensamblaje del genoma (estimado en 55 Megabases (Mb) para el genoma haploide). Para atacar estos problemas, por un lado se secuenció a una cobertura de 14x que es mayor de la que generalmente se busca para secuenciar un borrador de un genoma. Así fue posible distinguir las variantes alélicas de los errores de secuenciación. Por otro lado, para ayudar a distinguir los haplotipos provenientes de cada cepa original, un miembro de uno de los grupos progenitores (cepa Esmeraldo) fue secuenciado a baja cobertura. De todas formas, a diferencia de los genomas de los otros tripanosomátidos, el genoma de *T. cruzi* fue inicialmente reportado como un conjunto de 5489 *scaffolds* construidos a partir de 8740 *contigs*. Haciendo uso de los cromosomas ensamblados de *T. brucei*, unos años después Weatherly y cols. lograron ensamblar los *contigs*

originales en 41 cromosomas, número que coincide con el obtenido por análisis de electroforesis de campo pulsado (Branche, C. *et al.* 2006).

El genoma haploide contiene aproximadamente 12000 genes que codifican para proteínas habiéndose podido asignar una posible función a la mitad. Además se identificaron 1994 genes para ARN no codificante y 3590 pseudogenes (El-Sayed, N. M. *et al.* 2005). Al menos el 50% del genoma está formado por secuencias repetidas, que consisten en retrotrasposones, repetidos subteloméricos y familias multigénicas. El gran número de familias multigénicas encontrado explica en gran medida, el hecho de que *T. cruzi* es el tripanosomátido con el mayor número de genes. En efecto, un 18% de las secuencias codificantes están presentes en más de 14 copias (Arner, E. *et al.* 2007). Las familias multigénicas más expandidas, consisten en proteínas tipo trans-sialidasas, mucinas, metaloproteasas, DGF-1 (*dispersed gene family protein 1*), proteínas RHS (*retrotransposon hot spot*) y las proteínas de superficie asociadas a mucinas (*mucin associated surface proteins,* MASP), así llamadas por su asociación en el genoma con los genes de las mucinas (De Pablos, L. M. *et al.* 2012). Esta familia, que era desconocida hasta ese momento, representa un 6% del genoma diploide del parásito (Bartholomeu, D. C. *et al.* 2009; dos Santos, S. L. *et al.* 2012). En general, en tripanosomátidos son comunes las familias multigénicas que codifican para antígenos de superficie. Particularmente en el estadío tripomastigota sanguíneo, los parásitos están expuestos a las moléculas efectoras del sistema inmune del hospedero, incluyendo anticuerpos específicos. A diferencia de los tripanosomátidos africanos, los tripomastigotas de *T. cruzi* no poseen el mecanismo de variación antigénica, sino que expresan en su superficie varias proteínas diferentes pertenecientes a familias multigénicas, siendo las más caracterizadas las mucinas, las trans-sialidasas y las MASP (Di Noia, J. M. *et al.* 2000; El-Sayed, N. M. *et al.* 2005; dos Santos, S. L. *et al.* 2012). Por lo tanto, la función de las mismas se ha asociado frecuentemente con protección y evasión del sistema inmune del hospedero (dos Santos, S. L. *et al.* 2012). Más recientemente, Franzén y cols. publicaron el borrador del genoma de la cepa Sylvio X10, perteneciente al grupo *T. cruzi* I. Este grupo es el causante principal de la enfermedad de Chagas en América Central y la región Amazónica. El tamaño de este genoma, que se considera característico del grupo I, se estimó en 44 Mb siendo menor que el de CL Brener. Aparte de esta diferencia en tamaño, los genomas revelan una organización similar. Esta cepa mostró mayor identidad de secuencia con el haplotipo no Esmeraldo de CL Brener, lo cual confirma la filogenia descrita previamente, es decir que *T. cruzi* I está más emparentado con *T. cruzi* III de donde proviene el haplotipo no Esmeraldo (Franzen, O. *et al.* 2011).

El genoma de 25Kb de los maxicírculos mitocondirales también fue secuenciado revelando que codifica para 18 proteínas mitocondriales y dos genes de ARNr. Estas secuencias mostraron que 15 de los genes codificantes tienen cambios en los marcos de lectura confirmando la necesidad de la edición de los mensajeros para su correcta expresión (Westenberger, S. J. *et al.* 2006). El secuenciado de los maxicírculos de varias cepas de *T. cruzi* ha permitido además realizar inferencias de los mecanismos de generación de los 6 subgrupos que hoy en día se reconocen para este organismo (Zingales, B. *et al.* 2009; Ruvalcaba-Trejo, L. I. *et al.* 2011). Con estos datos los autores proponen que una cepa tipo I por hibridación con cepas de tipo II, resultó en la generación de las cepas tipo III y IV. Luego una hibridación entre cepas tipo II y III resultó en los tipos V y VI.

Genómica comparativa

Aunque la distancia filogenética entre los Tritryps es grande, los análisis de búsqueda de genes ortólogos evidenciaron que los genomas comparten un núcleo de unos 6200 genes. Las secuencias proteicas deducidas de la anotación permitieron calcular 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, reflejando las relaciones filogenéticas previamente establecidas (El-Sayed, N. M. *et al.* 2005). Los transcriptomas completos han permitido mejorar la anotación existente en los genomas de los organismos modelo, detectándose casos de CDS mal anotados así como la existencia de transcritos novedosos (Siegel, T. N. *et al.* 2011). Para *T. cruzi* y *T. brucei* los genes específicos de especie corresponden, mayoritariamente, a las grandes familias expandidas de antígenos de superficie. *L. major* es el organismo con menor número de proteínas específicas y estas son fundamentalmente proteínas hipotéticas (Figura 1.2.1.1).

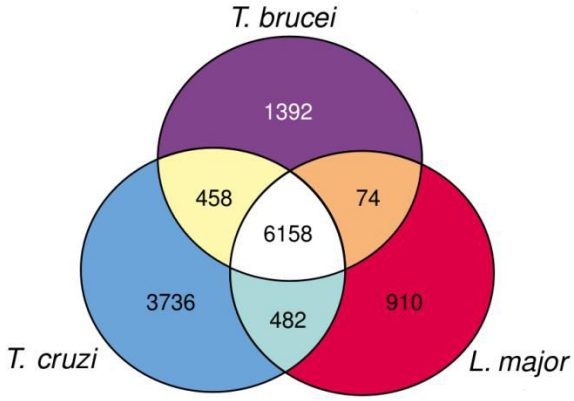


Figura 1.2.1.1 Diagrama de Venn mostrando el número de genes compartidos entre las diferentes especies secuenciadas. Extraído de (El-Sayed, N. M. *et al.* 2005).

Uno de los hallazgos más llamativos se obtuvo cuando se comenzaron a secuenciar grandes regiones genómicas y se relaciona con la organización de los genes en los cromosomas. Al observar la ubicación y orientación de éstos sobre cada hebra del ADN, se puede ver que se encuentran agrupados en grandes regiones con la misma orientación. Estos agrupamientos fueron denominados *directional gene clusters* (DGCs) y son característicos de todos los tripanosomátidos. La existencia de grandes DGCs se observó por primera vez para un cromosoma completo en *L. major* hace varios años. En efecto, el cromosoma 1 de este organismo fue el primer cromosoma entero secuenciado en tripanosomátidos y luego de su anotación sorprendió observar que los primeros 29 genes se encuentran ubicados en una hebra mientras que el resto de los 50 genes están codificados en la hebra complementaria (Figura 1.2.1.2) (Myler, P. J. *et al.* 1999; Martinez-Calvillo, S. *et al.* 2003). Luego de la compleción de los proyectos genoma, esta observación se generalizó (Berriman, M. *et al.* 2005; El-Sayed, N. M. *et al.* 2005; Ivens, A. C. *et al.* 2005).

Figura 1.2.1.2 Representación esquemática del cromosoma 1 de *L. major*. Los rectángulos de color representan los genes. Colores similares marcan el involucramiento en procesos celulares similares. Extraído y modificado de (Myler, P. J. *et al.* 1999).

Otra característica distintiva de estos genomas es la altísima conservación de la sintenia (orden de los genes) en los DGCs de los distintos genomas a pesar de la gran distancia evolutiva que los separa. De hecho un 94% de los genes que son comunes a los tres genomas se encuentran en regiones sinténicas (Figura 1.2.1.3) (El-Sayed, N. M. *et al.* 2005).

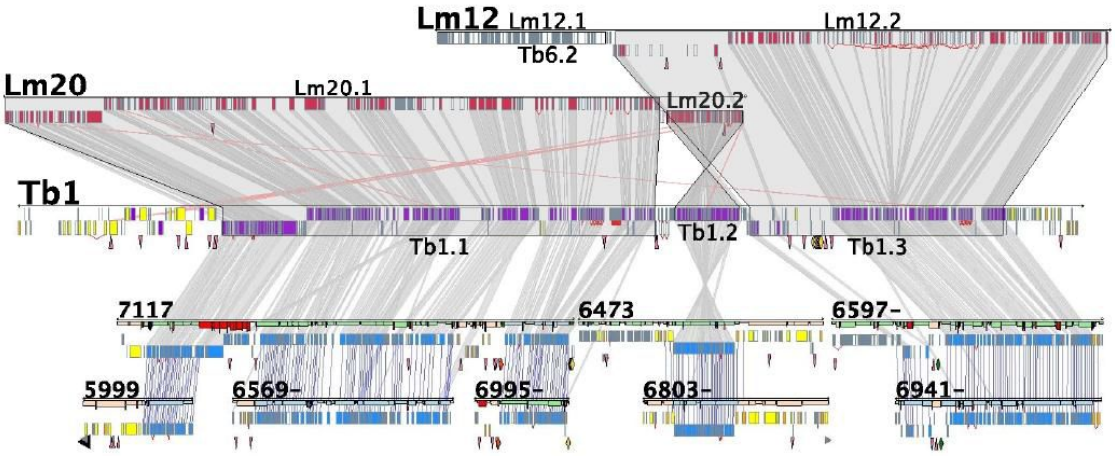


Figura 1.2.1.3 Mapa de sintenia para el cromosoma 1 de *T. brucei*. Las líneas grises unen dos genes ortólogos entre dos organismos. Se puede observar el alto grado de conservación del ordenamiento con los genes de *L. major* (arriba) y *T. cruzi* (abajo). Extraído de (El-Sayed, N. M. *et al.* 2005).

Cuando se analizan las regiones donde se pierde la sintenia se observa que estas regiones están caracterizadas por la presencia de ARNs estructurales, retroelementos y los genes específicos de especie (El-Sayed, N. M. *et al.* 2005; Padilla-Mejia, N. E. *et al.* 2009; Daniels, J. P. *et al.* 2010).

La organización en DGCs implica que existen unos pocos lugares por cromosoma en los cuales el sentido de la transcripción se invierte. Estas regiones que se encuentran entre dos DGCs consecutivos, se denominan regiones de cambio de hebra (*strand switch regions*, SSR). Los SSRs se caracterizan por tener un tamaño que va desde centenares a algunos miles de pares de bases y se clasifican como convergentes o divergentes dependiendo de la orientación de los genes de cada DGC, como muestra la Figura 1.2.1.4. Los SSRs suelen ser puntos de pérdida de la sintenia entre los organismos.

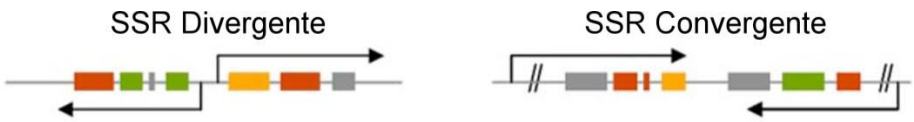


Figura 1.2.1.4 El esquema muestra los dos tipos posibles de SSRs. Las cajas de colores representan los genes mientras que las flechas indican el sentido de la transcripción. En el panel de la izquierda el SSR se define como divergente y se postula como posible inicio transcripcional. En el panel derecho el SSR se define como convergente y correspondería al sitio de finalización de la transcripción. Extraído y modificado de (Respuela, P. *et al.* 2008).

La fuerte sintenia encontrada indica una fuerte presión selectiva para preservar el orden génico de los DGCs. Sin embargo, en un principio la significancia funcional de estos agrupamientos no se conoce con certeza ya que no se encontró ninguna relación entre los genes pertenecientes al mismo DGC (ver Figura 1.2.1.2). Un estudio reciente describe que la ubicación dentro del agrupamiento no es azarosa para los genes que pertenecen a determinadas categorías funcionales de ontología génica (*gene ontology*, GO), lo cual puede estar aportando datos para la comprensión de la organización global del genoma y dando una posible explicación al mantenimiento de la sintenia entre los diferentes parásitos (Kelly, S. *et al.* 2012).

1.2.2 Transcriptoma

El transcriptoma de un organismo está determinado por el balance entre los procesos de síntesis y degradación de los ARNs celulares. En los tripanosomátidos estos procesos son extremadamente atípicos. En las siguientes secciones describiremos estos procesos haciendo foco en las particularidades que presentan.

Generalidades de las ARN Polimerasas

Los tripanosomátidos cuentan con las tres ARN polimerasas (ARNP) características de todos los eucariotas (Kelly, S. *et al.* 2005). Estas ARNP poseen alta homología con las polimerasas de otros organismos aunque han adquirido características exclusivas, tanto a nivel de sus subunidades como a nivel funcional, hechos que se reflejan en mecanismos de acción particulares (Berriman, M. *et al.* 2005; El-Sayed, N. M. *et al.* 2005; Ivens, A. C. *et al.* 2005; Brandenburg, J. *et al.* 2007; Nguyen, T. N. *et al.* 2007).

La ARNPIII transcribe los genes de ARNt y U6-snRNAs, siendo interesante recalcar que la transcripción de varios de estos últimos depende de elementos presentes en promotores para ARNt cercanos y en orientación opuesta en el cromosoma (Palenchar, J. B. *et al.* 2006; Das, A. *et al.* 2008). Los factores asociados a esta enzima no han sido caracterizados, aunque los proyectos genomas consiguieron reconocer genes homólogos a varias de las subunidades presentes en otros eucariotas. Dado que los promotores de ARNt para esta polimerasa presentan las características típicas de eucariotas, se asume que los factores TFIIIB y TFIIIC deben estar presentes para unirse a los elementos internos A y B del promotor (Das, A. *et al.* 2008).

En la mayoría de los eucariotas, la ARNPI es reclutada a promotores simples para transcribir el ARNr 45S, un precursor que es procesado a los ARNr 18S, 5.8S y 28S. Los factores de transcripción SL1 y UBF son esenciales para este reclutamiento. En tripanosomátidos, los genes para el precursor ARNr 45S son transcritos por la ARNPI como en otros eucariotas. Sin embargo, sorprendentemente esta enzima en *T. brucei* es capaz además de transcribir los ARNm de las glicoproteínas variables de superficie (VSGs) y prociclinas (Rudenko, G. *et al.* 1990; Lee, M. G. *et al.* 1997; Gunzl, A. *et al.* 2003; Pays, E. *et al.* 2004). Estas actividades están situadas en dos compartimentos subnucleares distintos, el nucleolo para los ARNr al igual que en otros eucariotas, y un sitio aparte denominado *expression site body* (ESB) donde son transcritos los ARN que codifican las VSGs. Como dijimos antes, el genoma tiene cientos de genes para estas proteínas, muchos de los cuales no son codificantes. Su transcripción solamente es posible a partir de alrededor de 20 regiones de localización telomérica que contienen un único gen de VSG. Estos sitios denominados BES (*bloodstream expression sites*) son activos en la forma sanguínea y tienen la particularidad de que uno solo de ellos se encuentra activo en un determinado momento, dando como resultado que una única molécula de VSG se expresa a la vez en el parásito. El intercambio del BES, así como el cambio del gen de VSG que se encuentra en el BES activo, es lo que produce el fenómeno de variación antigénica comentado antes. Además del gen de VSG, los BES contienen genes llamados ESAG (*expression site associated genes*) los cuales son transcritos junto con las VSG en una única unidad policistrónica (Berriman, M. *et al.* 2002; Hertz-Fowler, C. *et al.* 2008). Por otra parte, las VSG metacíclicas también son trascritas por la ARNPI, aunque estas unidades transcripcionales son monocistrónicas (*metaciclic expression sites*, MES) (Figura 1.2.2.1).

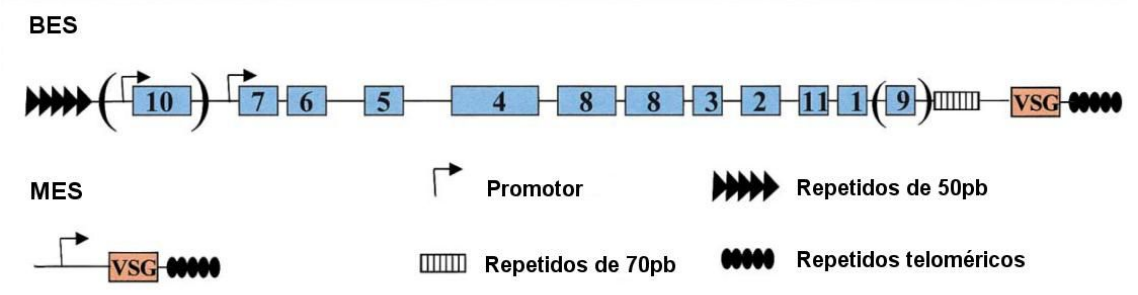


Figura 1.2.2.1 Esquema de los sitios de expresión de las proteínas variables de superficie de *T. brucei*. Entre paréntesis se muestran las ESAG que no están siempre presentes en todos los BES. Extraído y modificado de (Pays, E. *et al.* 2001).

Por su parte las prociclinas son las principales proteínas de superficie de la forma procíclica, recubriendo la superficie de este estadío de forma análoga a las VSG pero sin presentar la variabilidad de estas últimas. Las prociclinas se clasifican como EP y GPEET de acuerdo a la secuencia de los repetidos internos presentes en cada tipo (Roditi, I. *et al.* 1999). Las proteínas EP contienen repetidos de ácido glutámico (E) seguidos de prolina (P), mientras que las GPEET presentan repeticiones de este pentapeptido. Estos genes se encuentran en los cromosomas VI y X y son transcritos en unidades policistronicas con otras proteínas denominadas PAG (*procyclin associated genes*) que se encuentran río abajo (Haenni, S. *et al.* 2009).

Los promotores de las VSG y prociclinas, reconocidos por la ARNPI, son los únicos promotores clásicos descritos en tripanosomátidos para ARNs mensajeros. No presentan conservación de secuencia entre ellos ni con los promotores de los pre-ARNr, no estando claro aún si ambos reclutan a la ARNPI utilizando los mismos factores de transcripción (Palenchar, J. B. *et al.* 2006). Diferentes estudios han demostrado que los promotores de la ARNPI en eucariotas superiores se componen de un elemento UCE (*upstream control element*) que se encuentra a unos 150bp río arriba del punto de inicio de la transcripción (*transcription start site*, TSS) y un elemento *core* solapado al TSS (Paule, M. R. *et al.* 2000). Los tripanosomátidos comparten estas características generales en los promotores de los ARNr, prociclínas y VSGs, con la excepción de que en este último los dos elementos proximales son suficientes para la transcripción (Laufer, G. *et al.* 2001; Walgraffe, D. *et al.* 2005). En el caso del promotor de prociclina, encontramos tres elementos entre la posición -150 y el TSS. En el caso del ARNr los elementos básicos se centran en −57 y −27. Para las VSG tanto metacíclicas como sanguíneas los elementos están centrados en −60 y −36 (Ginger, M. L. *et al.* 2002; Palenchar, J. B. *et al.* 2006).

Es interesante señalar que a pesar de que los ARNr se caracterizan por ser secuencias altamente conservadas en eucariotas, no existe conservación de secuencia entre los promotores de diferentes organismos (Paule, M. R. *et al.* 2000). El factor responsable de interaccionar con el promotor para comenzar la formación del complejo específico de transcripción se denomina UBF (*upstream binding factor*) que es una proteína similar a las proteínas de la familia *high mobility group* (HMG). Aunque inicialmente se consideraba que cada organismo contaba con un *set* especifico de factores que era divergente al de los otros organismos, la proteína UBF es intercambiable entre varios organismos (Bell, S. P. *et al.* 1990). El reconocimiento del promotor por parte de UBF no ha mostrado especificidad de secuencia excepto por una preferencia de A y T. En *Xenopus leavis* se demostró que xUBF es capaz de

reconocer estructuras secundarias específicas provocando además cambios conformacionales en la región (Hu, C. H. *et al.* 1994). En general, las regiones de ADN que son reconocidas por proteínas específicas, tienden a poseer características estructurales dependientes de secuencia como la flexibilidad y la curvatura intrínseca adecuadas para favorecer esta unión (Crothers, D. M. 1998). Los promotores de los ARNr no son una excepción, presentando una conservación a nivel de estructura secundaria, sugiriendo que los mecanismos de transcripción de los ARNr han sido evolutivamente conservados. De hecho, las características conformacionales y termodinámicas de los promotores de ARNr muestran alrededor del TSS una estructura flexible, rodeados de una región de alta curvatura intrínseca, que podría favorecer el reconocimiento de los factores transcripcionales (Marilley, M. *et al.* 1996; Roux-Rouquie, M. *et al.* 2000).

La ARNPI de *T. brucei* comparte las 12 subunidades típicas de otros eucariotas, aunque además se ha encontrado una subunidad particular denominada RPA31 la cual se ha relacionado con la capacidad de transcripción de ARNm por parte de esta polimerasa (Kelly, S. *et al.* 2005; Walgraffe, D. *et al.* 2005; Nguyen, T. N. *et al.* 2007). Esta subunidad es compartida por otros tripanosomátidos como *T. cruzi*, por lo tanto es tentador especular que este organismo podría ser también capaz de transcribir ARNm. Una de las mayores diferencias de esta maquinaria se da en la secuencia del extremo N terminal de la subunidad RPA2. Dada su conservación en tripanosomátidos, se ha especulado que esta región explicaría las particularidades del mecanismo transcripcional de esta enzima (Daniels, J. P. *et al.* 2012).

En cuanto a la ARNPII, los análisis del genoma de los tripanosomátidos revelaron muy pocos posibles factores transcripcionales reguladores (Palenchar, J. B. *et al.* 2006). Se ha descrito un único promotor canónico que llamativamente transcribe el ARN *splice leader* (SL), ARN no codificante que interviene en el procesamiento de los pre ARN mensajeros (ver más adelante) (Gilinger, G. *et al.* 2001). En *T. brucei*, estos promotores consisten en tres elementos cortos, poco espaciados, que se localizan corriente arriba y próximos al sitio de inicio de la transcripción, mientras que en *T. cruzi* el promotor cuenta con un único elemento crítico (Campbell, D. A. *et al.* 2000). El estudio de los factores de inicio transcripcional de la ARNPII se ha restringido al análisis de los requerimientos del promotor del gen para SL. Los resultados muestran la participación de más factores de los que inicialmente se sospechaba por predicciones *in-silico* entre los que se encuentran proteínas que son

exclusivas de tripanosomátidos. Estudios genéticos y bioquímicos de los elementos del promotor del gen para SL en *Leptomonas seymouri* han llevado al aislamiento del primer factor de transcripción (PBP-1) en tripanosomátidos (Luo, H. *et al.* 1997). PBP-1 interacciona con el elemento del promotor localizado entre 60 y 80pb corriente arriba del sitio de inicio de la transcripción. Además, este factor es reclutado por los promotores de los ARN pequeños nucleares U2 y U6 transcritos por la ARNPIII, mostrando que el factor es requerido por ambas polimerasas para transcribir este tipo de ARN (Gilinger, G. *et al.* 2004). Este complejo proteico está compuesto de 3 polipeptidos que muestran homología a SNAP50, SNAPc y un tercer componente (TbSNAP42/TbSNAP2) que podría llegar a ser un ortólogo bastante divergente del SNAP190 (Luo, H. *et al.* 1997; Das, A. *et al.* 2003; Das, A. *et al.* 2005).

En *T. brucei* se caracterizó el factor de transcripción relacionado a TBP (TRF4) que además de ser reclutado por el promotor del gen para SL, también es requerido en las maquinarias transcripcionales de las ARN polimerasas I y III (Ruan, J. P. *et al.* 2004; Das, A. *et al.* 2005; Schimanski, B. *et al.* 2005). Este factor se une a regiones 3' de ciertos genes codificantes para proteínas, evidenciando un aspecto único de la transcripción por la ARN polimerasa II en estos organismos. Otros integrantes de estos complejos fueron descritos en *T. brucei* mediante experimentos de purificación por afinidad en tándem. De esta forma se describió una proteína homologa a SNAP43, un homólogo de la subunidad menor de TFIIA y una proteína de 63 kDa que corresponde a un ortólogo divergente de la subunidad más grande de TFIIA. Se encontró también un ortólogo de TFIIB que, aunque muy divergente, conserva residuos funcionalmente importantes (Das, A. *et al.* 2005; Schimanski, B. *et al.* 2005; Cribb, P. *et al.* 2009). Finalmente, se encontraron cinco subunidades de TFIIH (Lecordier, L. *et al.* 2007) y la proteína PPB1 (PSE *promoter-binding protein*) de 45kDa, que reconoce el promotor del miniexón lo que indica que sería entonces un posible factor importante en la transcripción del SL en *T. cruzi* (Wen, L. M. *et al.* 2000).

Lo discutido anteriormente nos permite concluir que la maquinaria transcripcional aparenta ser minimalista, aunque posee varios miembros exclusivos de estos organismos. Además, dentro de los factores que son compartidos con otros eucariotas, encontramos varios ejemplos que poseen diferencias significativas en dominios que han sido definidos como funcionalmente importantes (Das, A. *et al.* 2008). Esta evidencia puede sugerir varias interpretaciones. Por un lado, puede ocurrir que los factores transcripcionales hayan sufrido una gran divergencia en su secuencia, de forma de que no es posible identificarlos por métodos bioinformáticos. Otra interpretación es que proteínas exclusivas del parásito estén sustituyendo los

roles de sus contrapartes en eucariotas superiores. Por último, se ha sugerido que la maquinaria transcripcional sea realmente más sencilla. Las características de poca conservación en aminoácidos claves de factores de unión al ADN, así como el peculiar dominio carboxiterminal (CTD) de la propia polimerasa apuntan a que el mecanismo de inicio sea muy diferente a los paradigmas de eucariotas superiores. Estas particularidades podrían explicar la imposibilidad de encontrar elementos promotores clásicos. Heras y cols. han sugerido la presencia de un promotor para esta polimerasa dentro de la secuencia codificante de un retrotransposón de tipo LINE presente en tripanosomátidos. Aunque esta secuencia podría tener elementos similares a las de otros transposones que median su propia transcripción, el análisis no muestra elementos promotores clásicos. Esto estaría reforzando el concepto de que las señales que promueven la transcripción en tripanosomátidos pueden ser muy diferentes a las que se observan en otros eucariotas (Heras, S. R. *et al.* 2007).

Inicio de la transcripción mediada por la ARNPII

Si bien se han descrito algunas regiones como promotores de genes que codifican proteínas, su capacidad para incrementar el inicio de la transcripción no es muy clara (Ben Amar, M. F., Jefferies, D., Pays, A., Bakalara, N., Kendall, G., Pays, E. 1991; McAndrew, M. *et al.* 1998; Downey, N. *et al.* 1999). Varias regiones intergénicas resultan capaces de iniciar la transcripción. Llamativamente, si bien estas regiones intergénicas son indispensables para el correcto procesamiento del mensaje, no se ha encontrado ningún elemento promotor clásico en ellas (Clayton, C. E. 2002).

Por otra parte, la transcripción de los genes codificantes, por parte de la ARNPII, se da de forma policistrónica dando lugar a ARNs transcritos primarios que incluyen varios genes en una misma molécula precursora (*policistronic transcription units*, PTU). Dada la particular organización del genoma en DGCs, una hipótesis que se maneja desde hace tiempo es que los PTU pudieran abarcar un DGC completo, conteniendo los SSRs divergentes secuencias promotoras desde los cuales se daría el inicio transcripcional. Esto contrasta con la visión de que el inicio de la actividad de la polimerasa sea azaroso sobre el DGC dando lugar a policistrones con un número variable de genes. Estudios iniciales en los que se modifica la región SSR del cromosoma 1 de *L. major* sugieren que estas regiones no son esenciales para la transcripción (Dubessay, P. *et al.* 2002). Sin embargo, estudios posteriores sobre este mismo cromosoma, mediante la técnica de *run-on* específicos de hebra, demostraron que la transcripción ocurre de forma bidireccional a partir del SSR (Martinez-Calvillo, S. *et al.* 2003). Recientemente, se ha descrito que la transcripción bidireccional es una característica general de los promotores eucariotas (Chen, R. A. *et al.* 2013;

Orekhova, A. S. *et al.* 2013). Otra observación interesante que apoya que la transcripción no comienza de forma azarosa es que solo la hebra codificante produce una señal significativa. En el mismo trabajo los autores mapean los TSSs en una región de unas 100pb dentro de la región del SSR mediante 5´RACE comprobando que no es posible definirlo de forma exacta. Aunque las observaciones son compatibles con la presencia de una secuencia promotora en esa región, la secuencia allí presente no contiene motivos clásicos de promotores de eucariotas. Por otro lado, el estudio comparativo entre especies de *Leishmania* indica que este TSS no está más conservado que el resto de las regiones intergénicas. Resultados similares fueron obtenidos para el cromosoma 3 (Martinez-Calvillo, S. *et al.* 2004). Este conjunto de resultados evidencia entonces la existencia de inicios transcripcionales en los SSRs los cuales no poseen señales de secuencias específicas que dirijan este proceso (Martinez-Calvillo, S. *et al.* 2004; Monnerat, S. *et al.* 2004).

Nuevas evidencias directas e indirectas confirman la presencia de los TSS en las regiones de cambio de hebra divergentes así como, en un número menor, en regiones internas a los DGCs. Las evidencias indirectas tienen que ver con la observación de una serie de características a nivel de la cromatina que típicamente indican regiones de inicio de transcripción (Rudenko, G. 2010; Croken, M. M. *et al.* 2012). Los estudios hacen uso de la técnica de inmunoprecipitación de cromatina para detectar las regiones que están asociadas a histonas con modificaciones típicas de inicios transcripcionales. El primero de estos trabajos realizado en *T. cruzi* por Respuela y cols. demuestra que las regiones SSRs divergentes concentran histonas H3 y H4 acetiladas (H3ac y H4ac) en posiciones claves para la promoción de la transcripción y la histona H3 trimetilada en la lisina 4 (H3K4me3). Estas modificaciones de las histonas son marcas epigenéticas relacionadas universalmente con transcripción activa. El enriquecimiento de estas modificaciones en los SSRs divergentes con respecto a los SSRs convergentes es de más de 6 veces, lo cual evidencia que no es una marca que se relaciones con SSRs en general (Respuela, P. *et al.* 2008). Llamativamente, los autores encuentran un posicionamiento de los nucleosomas específico en las regiones de cambio de hebra. En otros organismos, se ha relacionado esta particularidad con la existencia de conformaciones alternativas en el ADN (Pisano, S. *et al.* 2006). Poco después, el grupo del Dr. Cross demostró que en *T. brucei* también se acumulan las histonas acetiladas (particularmente H4K10ac) en los TSS. Además en este estudio se muestra que las variantes de histonas H2AZ y H2BV tienen un patrón similar de localización en los cromosomas. Estas variantes forman nucleosomas menos estables y están correlacionadas con cromatina abierta y

promoción de la transcripción (Siegel, T. N. *et al.* 2009). Cabe señalar que en este trabajo, las variantes de histonas H3V y H4V fueron encontradas en sitios probables de terminación de la ARN Polimerasa II (Siegel, T. N. *et al.* 2009). El mismo grupo asoció poco después la presencia de H3K4me3 con la acetilación de histonas y los posibles TSS (Wright, J. R. *et al.* 2010). En este mismo sentido, Thomas y cols., trabajando en *L. major,* extienden la importancia funcional de estos sitios estudiando la correlación de los mismos con factores que participan en el inicio de la transcripción (Thomas, S. *et al.* 2009). En este estudio, los autores inmunoprecipitan ADN usando anticuerpo anti H3ac y componentes de la maquinaria de transcripción, particularmente TBP y el complejo SNAP50. Los autores encuentran histonas acetiladas en los SSRs divergentes y, como en los otros tripanosomátidos, también en regiones internas específicas. Interesantemente, en ambas localizaciones las histonas acetiladas se asocian con las proteínas de la maquinaria de transcripción basal, apoyando que estas modificaciones epigéneticas tienen relación directa con los TSS (Thomas, S. *et al.* 2009) (Figura 1.2.2.2).

Por otro lado, cabe señalar que la replicación del ADN está funcionalmente relacionada con la transcripción, habiendo evidencia de que los orígenes de replicación solapan en gran medida con los SSRs. Además, la tasa transcripcional general se ve afectada cuando se modifica la formación del complejo de inicio de la replicación (Tiengwe, C. *et al.* 2012).

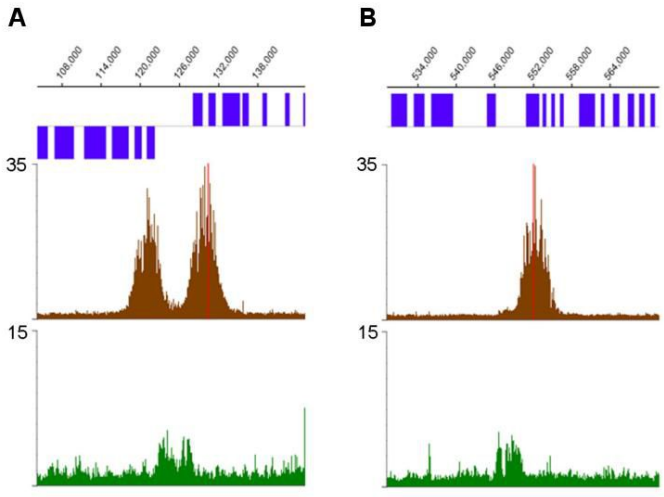


Figura 1.2.2.2 Asociación entre la localización de histonas acetiladas y la proteína TBP en dos regiones de *L. major*. Con cajas azules se representan los CDS. En marrón se observa la señal correspondiente a histona acetilada. En verde se observa la señal de TBP. A SSR divergente del cromosoma 6. B Región interna de un DGC del cromosoma 35. Extraído y modificado de (Thomas, S. *et al.* 2009).

Por otra parte, la existencia de una base nitrogenada modificada denominada base J (β-D-glucosil-hidroximetiluracilo), es una característica epigenética específica de kinetoplastídeos y otros flagelados unicelulares estrechamente relacionados (Borst, P. *et al.* 2008). Esta base es la primera hipermodificada encontrada en eucariotas y recientemente se la ha correlacionado con la dinámica transcripcional (van Luenen, H. G. *et al.* 2012). La base J fue descrita en primera instancia en *T. brucei* reemplazando alrededor de un 1% de las bases T (Gommers-Ampt, J. H. *et al.* 1993). Se forma en dos reacciones sucesivas: En primer lugar, los residuos T son hidroxilados para formar hidroximetiluracilo mientras que en el segundo paso se enlaza una glucosa. Las enzimas JBP1 y 2 están involucradas en el primer paso y una glucosil transferasa se encarga del segundo (Borst, P. *et al.* 2008). La localización y efectos de la base J tienen particularidades en cada organismo. En *T. brucei* un 50% está presente en los repetidos teloméricos, mientras que una gran parte del resto se encuentra en los subtelómeros y en los BES inactivos, por lo que se postula que puede tener un rol en el mecanismo de variación antigénica aunque no se ha reportado evidencia directa (van Leeuwen, F. *et al.* 2000; Cliffe, L. J. *et al.* 2010). En *T. cruzi* el 75% de la base J está asociada a los repetidos teloméricos, habiendo un nivel significativo también en los subtelómeros. En *L. major*, el 98% de la base J se ubica en los telómeros (Genest, P. A. *et al.* 2007). A diferencia de lo que ocurre en *T. brucei*, en este organismo la base J parece ser esencial para su sobrevivencia. La importancia de la base J en las regiones internas de los cromosomas en los Tritryps se ha comenzado a entender en los últimos años. Esta base no se localiza de forma azarosa sino que se encuentra en los SSRs (Cliffe, L. J. *et al.* 2010). En líneas celulares de *T. cruzi,* en las cuales la cantidad de base J está disminuida, se observa que aumenta la cantidad de histonas acetiladas y hay una mayor ocupación de la ARNPII en las regiones propuestas como promotores. Esto a su vez correlaciona con un aumento de la transcripción a nivel global en estas células. Con estos resultados se propone que la base J sería un factor importante para la regulación del estado de la cromatina promoviendo un estado represivo (Ekanayake, D. *et al.* 2011; Ekanayake, D. K. *et al.* 2011). Sin embargo, la función específica en cada parásito puede presentar diferencias. De hecho, estudios recientes en *L. major* demuestran que en este organismo en el cual la base J es esencial, ésta se localiza fundamentalmente en los SSRs convergentes y la ausencia de la misma provoca que la transcripción de los PTU no finalice en estos sitios al final del DGC (van Luenen, H. G. *et al.* 2012).

Además de los marcadores epigenéticos, se ha obtenido evidencia directa de que el inicio transcripcional se da en las regiones de cambio de hebra y otros sitios

internos, por secuenciado masivo de ARN (RNAseq). Esta técnica consiste en la secuenciación a alta cobertura del ADNc obtenido a partir del ARN de un organismo en un determinado momento (Mortazavi, A. *et al.* 2008). Para ello, se generan bibliotecas enriquecidas en ARNs nacientes para luego mapear estos ARNs nacientes al genoma. En *T. brucei,* los patrones de localización confirman los sitios propuestos como inicios transcripcionales basados en el estudio de las evidencias indirectas antes mencionadas (Kolev, N. G. *et al.* 2010).

Procesamiento de los ARN mensajeros

Una vez que los ARNs policistrónicos son transcritos, los ARN mensajeros individuales deben ser separados del resto. Para esto el precursor es procesado por un mecanismo intermolecular llamado *trans-splicing* que fue descrito por primera vez en los tripanosomátidos (Liang, X. H. *et al.* 2003). El mecanismo involucra la adición de un miniexón (*spliced leader*, SL) a las regiones 5' de los diferentes ARNm (Parsons, M. *et al.* 1984). El SL en los tripanosomas consiste en una secuencia de 39 a 41pb que está seguida de un intrón de tamaño variable que tiene dos estructuras del tipo *stem-loop* separadas por una región de hebra simple. La secuencia del miniexón es codificada por un ARN nuclear pequeño que se encuentra repetido en el genoma decenas de veces (el número depende de la especie), el cual es procesado en su extremo 5´ para dar lugar al SL maduro (Liang, X. H. *et al.* 2003). El miniexón aporta a cada ARN mensajero individual la estructura de CAP (caperuza), la cual cuenta con un número mayor de modificaciones con respecto a la que se encuentra en eucariotas superiores (Tschudi, C. *et al.* 2002). En tripanosomátidos, esta estructura denominada CAP 4, consiste en una 7-metilguanosina además de grupos 2'O-metilo en los cuatro primeros nucleótidos (Bangs, J. D. *et al.* 1992).

Es importante señalar que el mecanismo de *trans-splicing* es indispensable para que se generen los ARNm maduros traducibles. Esto fue demostrado también en *C. elegans* en donde se ha encontrado que el genoma presenta varios transcritos policistrónicos que contienen alrededor del 15% de los genes (Blumenthal, T. *et al.* 2002; Chen, R. A. *et al.* 2013). Estos policistrones son más cortos que en tripanosomas conteniendo típicamente 2 o 3 genes relacionados. Estos PTU son también procesados en ARN individuales por *trans-splicing*. Los autores argumentan que la existencia de este mecanismo es indispensable para el caso de policistrones en eucariotas ya que la estructura CAP es necesaria para la protección y traducción de los mensajeros (Blumenthal, T. *et al.* 2002).

El empalme del miniexón corriente arriba del sitio de inicio de la traducción, ocurre de forma coordinada y acoplada a la poliadenilación del gen adyacente. (LeBowitz, J. H. *et al.* 1993; Matthews, K. R. *et al.* 1994). La adición del miniexón requiere de señales de secuencias que se encuentran corriente arriba del codón de iniciación. Estas consisten en un sitio aceptor de *splicing* constituido por un dinucleótido AG el cual está precedido por un tracto de polipirimidinas (Hummel, H. S. *et al.* 2000; Siegel, T. N. *et al.* 2005). La poliadenilación del gen corriente arriba con respecto al sitio de adición del miniexón ocurre a una distancia cercana (142nt en promedio para *T. brucei*) y se da preferentemente en dos A consecutivas cercanas a una región rica en T (Kolev, N. G. *et al.* 2010). El *trans-splicing* ocurre co-transcripcionalmente y es fundamental para la traducción correcta de los mensajeros (LeBowitz, J. H. *et al.* 1993; Ullu, E. *et al.* 1993). La reacción ocurre de forma muy similar a la de la remoción de intrones en el corte y empalme tradicional (*cis-splicing*) consistiendo en dos reacciones de transesterificación consecutivas entre ambos exones (Gunzl, A. 2010). Interesantemente, se encontró que en algunos casos es posible que se generen ARNs intermediarios dicistrónicos que pueden ser transportados al citoplasma (Jager, A. V. *et al.* 2007). En estos casos, se postula que señales ubicadas en regiones reguladoras del pre mensajero y que son reconocidas por proteínas de unión a ARN, producirían un “salteado de exón” (*exon skipping*) por enmascaramiento de las señales de *trans-splicing* y poliadenilación habituales. Estos ARNs serían posteriormente procesados en el citoplasma dando lugar a ARNm monocistrónicos (Jager, A. V. *et al.* 2007).

La generación de datos de secuenciación masiva mediante RNAseq ha permitido el mapeo de los sitios de procesamiento a nivel global y por lo tanto, la definición de las respectivas regiones UTRs. En *T. brucei* se han reportado 3 estudios consecutivos que permitieron la descripción de más de 30000 sitios de adición del miniexón en unos 9000 genes diferentes (Siegel, T. N. *et al.* 2011). Los resultados obtenidos resultaron bastantes sorprendentes mostrando que un 85% de los genes presentan en promedio alrededor de 3 posibles sitios de procesamiento 5’. La cantidad de transcrito correspondiente a la variante principal es menos de un 60% del total, indicando que las cantidades de las formas alternativas son fisiológicamente significativas (Kolev, N. G. *et al.* 2010; Nilsson, D. *et al.* 2010; Siegel, T. N. *et al.* 2010). Con respecto a los sitios de poliadenilación, la magnitud de transcritos alternativos observada fue aún mayor, describiéndose más de 50000 sitios de poliadenilación en unos 8000 genes (Siegel, T. N. *et al.* 2011). Este fenómeno, aunque menos pronunciado, fue reportado recientemente para *L. major* y *T. vivax* (Greif, G. *et al.* 2013; Rastrojo, A. *et al.* 2013).

Al momento, no está claro si esta gran variedad de transcritos alternativos es un reflejo de una baja fidelidad de la maquinaria de procesamiento, aunque se ha sugerido que podría ser un fenómeno regulado y, por lo tanto, con consecuencias funcionales al mostrar cada variante diferentes secuencias en las regiones UTRs (Kolev, N. G. *et al.* 2010; Nilsson, D. *et al.* 2010).

Otra característica sobresaliente de los kinetoplastídeos, es la ausencia de intrones en sus genes. Sin embargo, existen excepciones a esta regla general. Mair y cols. describieron que los genes de la poliA polimerasa (PAP) de *T. cruzi* y *T. brucei* están interrumpidos. Con esta observación se pudo establecer que el mecanismo de *cis-splicing* está también presente en estos protozoarios (Mair, G. *et al.* 2000). Por otro lado, el gen que codifica para el ARNt-Tyr también contiene un intrón (Tan, T. H. *et al.* 2002; Padilla-Mejia, N. E. *et al.* 2009). Con la secuenciación de los genomas completos se puedo establecer en *T. brucei* que esta regla podría también romperse en el caso de otros genes puntuales que codifican para una helicasa de ARN y dos proteínas hipotéticas (Ivens, A. C. *et al.* 2005). Los estudios de RNAseq mencionados anteriormente permitieron la confirmación experimental del intrón presente en la PAP y en la proteína helicasa (Kolev, N. G. *et al.* 2010; Siegel, T. N. *et al.* 2010).

Regulación del estado estacionario

Por todo lo dicho anteriormente, se presume que la mayoría de los genes son transcritos de forma constitutiva a partir de los diferentes TSS dando lugar a múltiples PTUs (Andersson, B. *et al.* 1998; Worthey, E. A. *et al.* 2003). En este contexto, el aumento en el número de copia de un gen resulta un mecanismo frecuente para aumentar la expresión génica en los tripanosomátidos. De cualquier manera, a pesar de su transcripción primaria común, los genes individuales que pertenecen a la misma PTU muestran diferentes patrones de expresión, evidenciando el hecho de que la regulación en los tripanosomátidos opera principalmente a nivel post transcripcional (Clayton, C. E. 2002; Clayton, C. *et al.* 2007; Martinez-Calvillo, S. *et al.* 2010). Por lo tanto, la molécula de ARNm es uno de los principales blancos de regulación en el parásito. Los procesos de *splicing*, poliadenilación, exportación al citoplasma, compartimentación subcelular, degradación diferencial y traducibilidad, pueden ser regulados para variar la expresión de un gen determinado. Estos cambios pueden darse durante los diferentes estadíos del ciclo de vida del parásito, aunque se han descrito también genes que responden a estímulos específicos. Por ejemplo, la expresión del gen que codifica para la tubulina responde a cambios en la dinámica de

los microtúbulos (da Silva, R. A. *et al.* 2006) o la proteína HSP70 aumenta su expresión en condiciones de aumento de temperatura (Rodrigues, D. C. *et al.* 2010).

Particularmente, la estabilización o degradación del ARNm modula su vida media en diferentes estadíos y/o condiciones, siendo las regiones no traducidas que flanquean los CDS sitios claves en esta regulación (Vanhamme, L. *et al.* 1995; Furger, A. *et al.* 1997). En este sentido son fundamentales las secuencias presentes en los mismos así como las proteínas de unión a ARN reguladoras que reconocen estas regiones.

Elementos y factores reguladores

Existen numerosos elementos en *cis* que se ha visto que afectan el nivel de estado estacionario de los ARNm maduros, de los cuales la mayor parte se localizan en la región 3'UTR (Hotz, H. R. *et al.* 1997; Teixeira, S. M. 1998; Coughlin, B. C. *et al.* 2000; Brooks, D. R. *et al.* 2001; Webb, H. *et al.* 2005; Clayton, C. *et al.* 2007; Colasante, C. *et al.* 2007; Haile, S. *et al.* 2007; Robles, A. *et al.* 2008), aunque se han encontrado también en otras regiones (Mahmood, R. *et al.* 1999; Mahmood, R. *et al.* 2001; Clayton, C. E. 2002). Las señales a nivel del ARN son secuencias cortas y de baja conservación siendo en muchas ocasiones importante la estructura secundaria que adopten (Furger, A. *et al.* 1997; D'Orso, I. *et al.* 2001). La definición de un gran número de regiones UTRs conseguida por experimentos de RNAseq nos permite la búsqueda de señales conservadas en genes relacionados, sin embargo, aún no se han reportado estudios en este sentido (Siegel, T. N. *et al.* 2010).

Hasta el momento, entre los elementos reguladores más caracterizados figuran los UREs (*U rich instability elements*) los cuales han sido encontrados en varios ARN afectando su estabilidad (Haile, S. *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 (Barreau, C. *et al.* 2005; von Roretz, C. *et al.* 2011). En *T. cruzi* se han encontrado elementos ricos en AU regulando integrantes de la familia de proteínas mucinas de estos organismos (Di Noia, J. M. *et al.* 2000). La presencia de un elemento rico en U de 43 nucleótidos en un gran número de ARNm controla su abundancia en el estadío amastigota del parásito (Li, Z. H. *et al.* 2012). También se han descrito elementos ricos en G (GREs) en las regiones 3'UTR de los genes tipo mucinas TcSMUG que confieren estabilidad del ARNm en el ciclo de vida (D'Orso, I. *et al.* 2001; Vlasova-St Louis, I. *et al.* 2011). 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 estadío especifica (Bayer-Santos, E. *et al.* 2012; Pastro, L. *et al.* 2013). Con respecto a las señales en *T. brucei* los genes de la fosfoglicerato kinasa son regulados también por un elemento rico en AU presente en el 3'UTR que desestabiliza el ARNm en la forma sanguínea (Quijada, L. *et al.* 2002). La región 3'UTR de los ARNm de las proteínas EP/GPEET contiene secuencias específicas que contribuyen a la estabilidad diferencial de cada ARNm a medida que el parásito migra a través del tracto digestivo del insecto vector (Furger, A. *et al.* 1997; Hotz, H. R. *et al.* 1997; Vassella, E. *et al.* 2000). Otras secuencias regulatorias en *cis* son específicos de cada organismo. Por ejemplo, los elementos SIDER 1 y 2 fueron encontrados únicamente en el genoma de *Leishmania*. Estos elementos son retrotransposones dispersos que se ubican preferentemente en las regiones 3'UTR de los genes del parásito. SIDER 1 está relacionado con la regulación de mensajeros estadío específicos, fundamentalmente los presentes en la etapa amastigota (Wu, Y. *et al.* 2000; Boucher, N. *et al.* 2002; Rochette, A. *et al.* 2005). SIDER 2 sin embargo parece ser un elemento general de desestabilización (Bringaud, F. *et al.* 2007).

Consistentemente con la existencia de un gran número de señales diferentes, el análisis del genoma de los Tritryps ha revelado la existencia de un gran número de proteínas de unión al ARN (El-Sayed, N. M. *et al.* 2005; Kramer, S. *et al.* 2011). Sin embargo, sólo unas pocas de ellas han sido caracterizadas hasta el momento (Clayton, C. E. 2002; D'Orso, I. *et al.* 2003; Perez-Diaz, L. *et al.* 2007; Perez-Diaz, L. *et al.* 2012). Fundamentalmente se han descrito integrantes de las familias de proteínas con dominios RRM (Manger, I. D. *et al.* 1998; D'Orso, I. *et al.* 2001; D'Orso, I. *et al.* 2002; Hartmann, C. *et al.* 2007; Perez-Diaz, L. *et al.* 2007; Hartmann, C. *et al.* 2008; Guerra-Slompo, E. P. *et al.* 2012), con dominios de tipo dedos de zinc (Hendriks, E. F. *et al.* 2001; Morking, P. A. *et al.* 2004; Hendriks, E. F. *et al.* 2005; Paterou, A. *et al.* 2006; Morking, P. A. *et al.* 2012) y proteínas de la familia Pumilio (Dallagiovanna, B. *et al.* 2005; Dallagiovanna, B. *et al.* 2008). Estas proteínas están vinculadas a la regulación de la expresión génica a nivel de mensajeros interviniendo en los procesos de maduración (Matthews, K. R. *et al.* 1994; Xu, P. *et al.* 2001; Portal, D. *et al.* 2003; Portal, D. *et al.* 2003; Vazquez, M. *et al.* 2003), *editing* (Madison-Antenucci, S. *et al.* 2002), recambio o traducibilidad (D'Orso, I. *et al.* 2003; Haile, S. *et al.* 2003; Haile, S. *et al.* 2007). Llamativamente en *T. brucei,* se encontró que proteínas de *splicing* eran capaces de interaccionar con la región 3'UTR y con otras proteínas regulatorias, afectando la estabilidad de la molécula (Gupta, S. K. *et al.* 2013).

Otros factores en *trans* que intervienen en los mecanismos regulatorios son las moléculas de pequeños ARN. En tripanosomátidos, la mayor parte de la información se ha obtenido en *T. brucei*, que fue unos de los primeros organismos donde se observó el fenómeno de interferencia de ARN. Aunque se han predicho moléculas de micro ARNs, hasta el momento se han encontrado experimentalmente únicamente moléculas de ARNs pequeños interferentes (siRNAs) (Mallick, B. *et al.* 2008). La función de la maquinaria sería principalmente controlar las secuencias transponibles dando estabilidad al genoma (Atayde, V. D. *et al.* 2011) aunque la existencia de siRNAs derivados de pseudogenes abre la posibilidad de otras funciones reguladoras (Wen, Y. Z. *et al.* 2011). En *T. cruzi*, donde el mecanismo de interferencia no está presente, la población de ARNs pequeños deriva de secuencias de ARNs de transferencia no estando clara aún la funcionalidad de los mismos (Garcia-Silva, M. R. *et al.* 2010; Franzen, O. *et al.* 2011).

Degradación y traducibilidad de los ARNm

La tasa de degradación de los ARN blanco es un proceso altamente regulado en estos organismos. En eucariotas, la degradación de los mensajeros se da por la acción de exonucleasas en ambos extremos luego de la remoción de las estructuras protectoras, o sea el CAP y la cola poliA. El proceso ha sido estudiado extensivamente en eucariotas, y, de acuerdo a los datos obtenidos hasta el momento en tripanosomátidos el proceso es similar (Coller, J. M. *et al.* 2001; Coller, J. *et al.* 2004; Balagopal, V. *et al.* 2009; Chen, C. Y. *et al.* 2011). Sin embargo, existen diferencias con el modelo clásico. Por ejemplo, no se han encontrado las proteínas homólogas a las enzimas de remoción del CAP, aunque esta actividad sí se observa en extractos proteicos (Milone, J. *et al.* 2004), evidenciando nuevamente que las proteínas involucradas pueden ser muy divergentes. Se ha descrito que la maquinaria de degradación de la vía 5' y otras proteínas del metabolismo del ARN junto con ARNm específicos, se concentran formando gránulos en el citoplasma que se denominan cuerpos de procesamiento (*P bodies*), gránulos de estrés o gránulos de ARN, dependiendo de su composición y el momento en el que se forman en la célula (Cassola, A. *et al.* 2007). En tripanosomátidos se han encontrado gránulos citoplasmáticos de características análogas y que contienen proteínas con las funciones homólogas a las encontradas en eucariotas superiores (Li, C. H. *et al.* 2006; Holetz, F. B. *et al.* 2007; Kramer, S. *et al.* 2008; Schwede, A. *et al.* 2008; Kramer, S. *et al.* 2010). Se propone que este tipo de estructuras serían un reservorio de ARNs mensajeros, tanto en condiciones normales como en condiciones de estrés, pudiendo

los ARN ser degradados o volver a la población de ARN traducibles (Cassola, A. *et al.* 2007; Holetz, F. B. *et al.* 2007; Kramer, S. *et al.* 2008; Cassola, A. 2011). Se ha sugerido para tripanosomátidos, un mecanismo regulatorio que puede dar cuenta de la expresión diferencial de genes cuya concentración de ARNm en estado estacionario, no se ve modificada (Saas, J. *et al.* 2000). Este mecanismo consiste en la movilización diferencial de determinados mensajeros hacia la fracción polisomal, lo cual concuerda con la hipótesis de regulación por formación de gránulos citoplasmáticos. Consistentemente con la existencia de reservorios de ARNm no traducibles, se han encontrado evidencias experimentales de regulación traduccional (Gale, M., Jr. *et al.* 1994; Avila, A. R. *et al.* 2001; Dallagiovanna, B. *et al.* 2001; Boucher, N. *et al.* 2002; Mayho, M. *et al.* 2006; Nardelli, S. C. *et al.* 2007). 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, J. E. 1998). Este mecanismo de regulación coexiste con los antes mencionados ya que paralelamente, algunos genes están siendo regulados además a través de la estabilización de sus ARN mensajeros (Avila, A. R. *et al.* 2001). 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. 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. Se postula que de esta forma, se evitaría la presencia de reguladores negativos en esta región (Greif, G. *et al.* 2013). La tasa traduccional también puede estar influida por la selección de codones sinónimos presentes en cada gen. La hipótesis plantea que los genes de expresión alta estarían optimizados para mejorar la eficiencia de la maquinaria traduccional y/o su fidelidad (Hershberg, R. *et al.* 2008). En tripanosomátidos, el hecho de que los genes de alta expresión contienen un uso de codones particular se conoce desde hace tiempo (Parsons, M. *et al.* 1991; Alvarez, F. *et al.* 1994). La disponibilidad de los genomas completos sumado al conocimiento de los correspondientes niveles de expresión a escala global, ha permitido confirmar este hecho (Horn, D. 2008). Además, se corroboró que las moléculas de ARNt que reconocen los codones óptimos son las más abundantes, reforzando este hallazgo (Horn, D. 2008).

Relación transcriptoma-proteoma

Los mecanismos regulatorios descritos en la sección anterior definen el transcriptoma global que a su vez influencia de forma directa el proteoma de los

organismos. Aunque los primeros estudios de expresión globales en Tritryps comenzaron en los años 90 con el secuenciado de ESTs en formas sanguíneas de *T. brucei*, no fue posible obtener datos cuantitativos del transcriptoma, sirviendo estos estudios fundamentalmente para la anotación de nuevos genes (el-Sayed, N. M. *et al.* 1995; Levick, M. P. *et al.* 1996; Brandao, A. *et al.* 1997; el-Sayed, N. M. *et al.* 1997). Sin embargo, el advenimiento en un principio de la técnica de microarreglos de ADN y luego de RNAseq permitió el estudio cuantitativo de los niveles de estado estacionario de los mensajeros en diferentes fases del ciclo de vida de los parásitos. Inicialmente no estaba claro cuan variable podía llegar a ser el transcriptoma en organismos que cuentan con transcripción generalizada. De hecho, los primeros estudios de microarreglos de ARN mostraron que pocos genes presentaban expresión diferencial entre estadíos muy diferentes morfológicamente y la magnitud de esta variación era modesta (Almeida, R. *et al.* 2002; Beverley, S. M. *et al.* 2002; Diehl, S. *et al.* 2002; Minning, T. A. *et al.* 2003; Saxena, A. *et al.* 2003; Duncan, R. 2004; Duncan, R. C. *et al.* 2004). Sin embargo, estudios posteriores en *T. cruzi* detectaron un número mucho mayor de genes estadío específicos, mostrando que sí es posible observar diferencias en los niveles estacionarios de muchos de los ARNm (Minning, T. A. *et al.* 2009). Los estudios recientes de RNAseq en *T. brucei* muestran que alrededor de un 30% de los mensajeros presentan diferencias en los niveles de estado estacionario entre las formas sanguíneas y procíclicas (Siegel, T. N. *et al.* 2011).

De todos modos, en organismos que tienen una fuerte regulación traduccional, los niveles de ARNm no resultan en muchos casos un buen reflejo de la cantidad de proteína en la célula. Por lo tanto, se han hecho esfuerzos para caracterizar directamente el proteoma celular (Parodi-Talice, A. *et al.* 2004); (Atwood, J. A. *et al.* 2005). Sin embargo, hay que tener en cuenta que la potencia de las técnicas para definir los patrones globales de expresión de proteínas es menor en comparación con las técnicas que dependen de la secuenciación de ácidos nucleicos, debido a las dificultades inherentes de las técnicas. De todas formas, el proteoma de *T. brucei* fue estudiado por la técnica de SILAC (*stable isotope labeling by amino acids in culture*) que permite resultados cuantitativos. Los resultados mostraron que aunque existe una correlación positiva del proteoma con el transcriptoma los cambios observados a nivel de proteína son de mayor orden que los observados para los transcriptomas (Urbaniak, M. D. *et al.* 2012). Conclusiones similares fueron obtenidas en *T. cruzi* en donde fue analizado el cambio en el proteoma que se da durante la metaciclogénesis (de Godoy, L. M. *et al.* 2012).

Una aproximación que en principio permitiría acercarse mejor a lo observado a nivel del proteoma es el análisis de la fracción de ARNm que está siendo activamente traducida y, por lo tanto, forma parte de la fracción polisomal. Esta aproximación puede hacer uso de las nuevas técnicas de secuenciación de ácidos nucleicos, con las ventajas que eso presenta frente a las metodologías de análisis proteómico. De hecho, hace pocos años se describió, por parte del grupo de Weissman, la técnica de huellas ribosomales (*ribosomal footprinting*), en la cual se purifican los polisomas y se realiza una protección a una RNAsa para luego secuenciar masivamente por RNAseq los fragmentos obtenidos (Ingolia, N. T. *et al.* 2009). De esta manera, se puede obtener una visión tanto cualitativa como cuantitativa de la ocupación de ribosomas sobre cada mensajero particular (Ingolia, N. T. 2010; Ingolia, N. T. *et al.* 2011; Ingolia, N. T. *et al.* 2012). Los resultados obtenidos por esta técnica pueden ser de gran relevancia en los Tritryps dadas las particularidades comentadas anteriormente.

Objetivos 2

2.1 Objetivo general

Aportar a la comprensión de los mecanismos de la expresión génica en los tripanosomátidos patógenos *Leishmania major*, *Trypanosoma brucei* y *Trypanosoma cruzi* (Tritryps).

2.2 Objetivos específicos

En particular, nos planteamos contribuir en algunos aspectos estructurales y funcionales de los genomas de estos organismos:

2.2.1 Búsqueda *in silico* de señales involucradas en la expresión génica

Análisis del contenido y distribución de dinucleótidos

Análisis global de patrones de curvatura intrínseca

2.2.2 Estudio de las dinámicas de transcripción y traducción

Aproximación experimental para la identificación de sitios de inicio de la transcripción

Análisis *in silico* del perfil de huellas ribosomales

Resultados y discusión

3

3.1 Búsqueda *in silico* de señales involucradas en la expresión génica

3.1.1 Análisis del contenido y distribución de dinucleótidos en los genomas de Tritryps

Como forma de contribuir a la comprensión de las características estructurales del genoma de los tripanosomátidos, y en el contexto de los antecedentes del trabajo de investigación de nuestro grupo, nos planteamos como objetivo ahondar en el estudio de la localización y frecuencia de los repetidos de dinucleótidos en el genoma de los estos organismos. Se han descrito múltiples ejemplos en varios organismos y en diferentes procesos celulares en los cuales los repetidos de dinucleótidos son secuencias funcionalmente importantes. De hecho, cumplen un rol fundamental en los mecanismos moleculares en los que participan. En particular, existe un conjunto de resultados que sugieren el involucramiento de los repetidos de dinucleótidos en procesos como la organización de la cromatina, el metabolismo del ADN y la regulación de la expresión génica (Li, Y. C. *et al.* 2002; Plohl, M. *et al.* 2008). Además, la inestabilidad de estas secuencias las hace un sustrato ideal para la generación de diversidad genética acelerando la evolución de las regiones reguladoras (Gemayel, R. *et al.* 2010). Aunque estas secuencias son mucho menos abundantes en las regiones codificantes, cuando se encuentran presentes, se distribuyen de forma no aleatoria y han mostrado diferentes roles que han sido resumidos en Li y cols. (Li, Y. C. *et al.* 2004).

Como antecedente principal al trabajo en tripanosomátidos, contábamos con la observación de que los repetidos de dinucleótidos eran frecuentes y con una distribución particular en las regiones cercanas a las secuencias codificantes, trabajo que había sido realizado con las secuencias parciales disponibles hasta el momento en *T. cruzi* (Duhagon, M. A. *et al.* 2001). Una vez publicados los genomas completos de los Tritryps nos planteamos generalizar este trabajo.

Los resultados presentados muestran que la mayor parte de los repetidos son más frecuentes de lo que cabe esperar por azar de acuerdo a la composición de bases del genoma. Además, los repetidos complementarios presentan asimetría de hebra, confirmando la observación presentada en el trabajo publicado en 2001 para los repetidos de GT (Duhagon, M. A. *et al.* 2001). Estos repetidos además son los que presentan largos más grandes (o sea mayor número de unidades repetidas) en los tres genomas. Los repetidos de dinucleótidos no se encuentran distribuidos de manera

uniforme sino que tienden a encontrarse cercano a los ORFs y alejados de los límites de los DGCs.

Durante el desarrollo de este trabajo, participé en la generación de las bases de datos, realizando los análisis con el software TFR (Benson, G. 1999), construí e interrogué las bases de datos obtenidas en colaboración con el Dr. Hugo Naya y realicé el análisis de los mismos en el lenguaje R en colaboración con el Lic. Diego Forteza. Todo esto requirió formarme en metodologías de análisis de datos a gran escala así como en lenguajes de *scripting*, herramientas fundamentales en el resto del desarrollo de mi trabajo de Doctorado. Este trabajo se enmarcó inicialmente dentro de la tesis doctoral de la Dra. María Ana Duhagon, habiendo continuado el análisis de estos datos luego de la defensa de dicha tesis y culminando en la publicación que se incluye a continuación.

Gene 487 (2011) 29–37

Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Comparative genomic analysis of dinucleotide repeats in Tritryps

María Ana Duhagon [a,b], Pablo Smircich [a,b], Diego Forteza [a], Hugo Naya [c], Noreen Williams [d], Beatriz Garat [a,*]

[a] Laboratorio de Interacciones Moleculares, Facultad de Ciencias, 11400 Montevideo, Uruguay
[b] Departamento de Genética, Facultad de Medicina, 11800 Montevideo, Uruguay
[c] Unidad de Bioinformática, Institut Pasteur de Montevideo, 11400 Montevideo, Uruguay
[d] Department of Microbiology and Immunology, University at Buffalo, Buffalo, NY 14214 USA

ARTICLE INFO

Article history:
Accepted 14 July 2011
Available online 28 July 2011

Received by A.J. van Wijnen

Keywords:
Trypanosoma
DNA
Gene
Regulation
Microsatellite
Polycistron

ABSTRACT

The protozoans *Trypanosoma cruzi*, *Trypanosoma brucei* and *Leishmania major* (Tritryps), are evolutionarily ancient eukaryotes which cause worldwide human parasitosis. They present unique biological features. Indeed, canonical DNA/RNA *cis*-acting elements remain mostly elusive. Repetitive sequences, originally considered as selfish DNA, have been lately recognized as potentially important functional sequence elements in cell biology. In particular, the dinucleotide patterns have been related to genome compartmentalization, gene evolution and gene expression regulation. Thus, we perform a comparative analysis of the occurrence, length and location of dinucleotide repeats (DRs) in the Tritryp genomes and their putative associations with known biological processes. We observe that most types of DRs are more abundant than would be expected by chance. Complementary DRs usually display asymmetrical strand distribution, favoring TT and GT repeats in the coding strands. In addition, we find that GT repeats are among the longest DRs in the three genomes. We also show that specific DRs are non-uniformly distributed along the polycistronic unit, decreasing toward its boundaries. Distinctive non-uniform density patterns were also found in the intergenic regions, with predominance at the vicinity of the ORFs. These findings further support that DRs may control genome structure and gene expression.

1. Introduction

The protozoans *Trypanosoma cruzi*, *Trypanosoma brucei* and *Leishmania major*, the so called Tritryps (El-Sayed et al., 2005b), are the human pathogens responsible of American trypanosomiasis, African trypanosomiasis and old-world cutaneous leishmaniasis, respectively. They belong to the Family Trypanosomatidae, Order Kinetoplastida, constituting evolutionarily ancient eukaryotes that display exceptional gene expression hallmarks. For instance, protein coding genes are mainly arranged in unidirectional clusters that are transcribed as long polycistronic RNAs (Clayton, 2002). The strand-switch regions (SSRs) have been implicated in transcription and also in replication initiation and chromosome segregation (Martinez-Calvillo et al., 2004, 2003; Obado et al., 2005; Respuela et al., 2008). Nevertheless, the DNA sequence elements that mediate these processes have been largely elusive so far. The absence of canonical eukaryotic promoters, together with the lack of classical regulatory DNA signals for the transcription of protein coding genes, has led to the idea that transcription is not a major control level of gene expression in trypanosomatids. In this context, gene expression regulation is considered to be fundamentally dependent on post-transcriptional events (Palenchar and Bellofatto, 2006).

Although repetitive sequences were initially perceived as "junk" DNA, their functional importance in cell biology has been widely documented (Plohl et al., 2008; Sharma et al., 2007). Nevertheless, broader questions regarding the origin, evolution and functional significance of microsatellites remain essentially unsolved (Buschiazzo and Gemmell, 2006).

Dinucleotide repeats (DRs) constitute one type of microsatellite. The pattern of dinucleotide occurrence has been associated with genome compartmentalization, gene evolution and gene expression level (Sharma et al., 2007). Moreover, DRs have been reported as target sequences for specific protein recognition (Epplen et al., 1996; Zhang et al., 2009). We have previously analyzed DRs in the genome of *T. cruzi* using the sequences deposited in the GenBank up to the year 2001 (Duhagon et al., 2001). Our results suggested active roles of these sequences in gene expression. In further support to the latter hypothesis, some DRs were recently found at the SSRs of head to head transcriptional units (Cribb et al., 2010; Respuela et al., 2008; Siegel et al., 2009; Thomas et al., 2009), which have been involved in polymerase II transcription initiation.

Here we extend the analysis of the presence, length (nt) and location of DRs to the whole genome of *T. cruzi* and also *T. brucei* and *L. major*. The use of the three species would allow the finding of conserved patterns. The analysis of the completely assembled genomes permits the study of

Abbreviations: DR, dinucleotide repeat; (XY)n, oligodi(deoxynucleotide) of n nucleotides in length; SSR, strand-switch region; ORF, open reading frame; CDS, coding sequence; UTR, untranslated region.

* Corresponding author. Tel.: +598 2 525 86 18x7 237; fax: +598 2 525 86 17.
E-mail address: bgarat@fcien.edu.uy (B. Garat).

0378-1119/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2011.07.022

DRs in the context of the overall genomic architecture, which could shed light on their use in position dependent molecular mechanisms.

2. Materials and methods

Occurrence of DRs was examined for each of the 10 possible combinations of nucleotides (AA, TT, GG, CC, AT/TA, AG/GA, AC/CA, CT/TC, CG/GC, GT/TG) in: a) *L. major* genome, consisting *of* 33 Mb distributed in 36 chromosomes available in GenBank (http://www.ncbi.nlm.nih.gov/Genbank/) (base frequency composition: A: 0.20, C: 0.30, G: 0.30, T: 0.20); b) *T. brucei* genome, consisting of 26 Mb distributed in 11 chromosomes available in GenBank (base frequency A: 0.27, C: 0.23, G: 0.23, T: 0.27) and c) *T. cruzi* genome, consisting of 58 Mb distributed in 638 scaffolds available in TcruziDB (http://tcruzidb.org/tcruzidb/) (base frequency A: 0.25, C: 0.25, G: 0.25, T: 0.25). For each genome, strand-switch localization was determined using PERL programming language (Wall et al., 2000) scripts developed for this purpose. We defined a polycistronic unit as any region consisting of at least two annotated coding sequences (CDSs) oriented in the same direction, and intergenic region as the sequence between two consecutive CDSs. In addition, the occurrence (O), length in nucleotides (l) and mismatch (a) (as a measure of the deviation from the perfect dinucleotide repetitive sequence) were determined by the Tandem Repeat Finder program (TRF) (Benson, 1999). The TRF parameters were set to Match.Mismatch.Delta.PM.PI.Minscore.MaxPeriod = 2.7.7.80.10.7.2 to enable the detection of imperfect DRs of $l \geq 8$nt and perfect DRs of $l \geq 6$nt. Relational databases were built and tables combining strand-switch and repeat content information were created. In addition, similar tables containing the location of the DRs with respect to CDSs were constructed. The data were analyzed using R software environment for statistical computing. The randomly expected frequencies for each class of perfect DR of a defined length were set as the product of each nucleotide frequency: $((X/N)\ (Y/N))^n\ ((X/N)^{l-2n} + (Y/N)^{l-2n})$ for $n \geq 3$, where n is the copy number of the dinucleotide $(XY)_n$ and N the genome size (nt). This formula takes into account both even ($l = 2n$) and odd ($l = 2n+1$) DR lengths. The resulting frequencies were transformed to expected occurrences (E) multiplying by the corresponding N. Only integral values of E were considered. The difference between the expected frequency for a given length (l) and the expected frequency of the subsequently longer DR ($l+1$) renders the non-cumulative expected occurrence. The significance of the deviation of observed (O) from expected (E) values was measured using the χ^2 test. For each dinucleotide class, we determined the previously defined variable $D = \Sigma$ l.a (Duhagon et al., 2001), that combines the occurrence, length and mismatch factor from the TRF program (Benson, 1999).

3. Results and discussion

3.1. DRs display specific pattern of occurrence in the coding strands

In an attempt to uncover a putative functionality of the DRs, we analyzed their frequency in the genomes of *L. major*, *T. brucei* and *T. cruzi* (Berriman et al., 2005). Strand-switches were defined as indicated in Material and methods, and only the coding strands were used for all the analyses. Perfect and imperfect DRs were detected with the TRF program (Benson, 1999) as detailed in Materials and methods. The occurrence of each type of DR with $l \geq 8$ in the genome of the Tritryps is presented in Fig. 1A. The global pattern of dinucleotide abundance conforms to early reports from partial genome sequences of *T. cruzi* (Aguero et al., 2000), and interestingly, of other organisms too (Toth et al., 2000). Overall, *T. brucei* and *T. cruzi* patterns of DR occurrence are more similar than *L. major*'s.

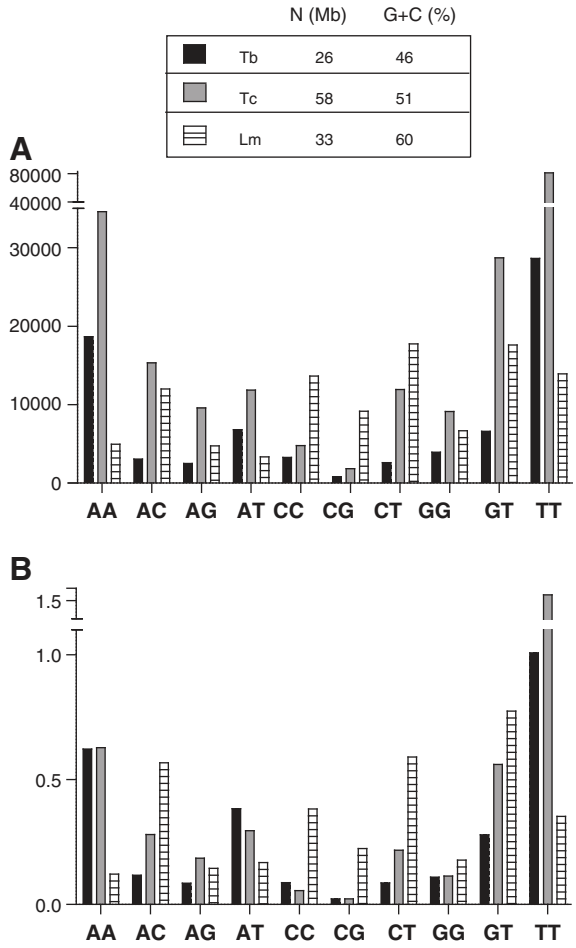


Fig. 1. Abundance of DRs in the Tritryps. A. Occurrence of the ten classes of DRs in the coding strands. Imperfect repeats of length $l \geq 8$ were used in the analysis. B. D value ($D = \Sigma$ l.a) normalized by the genome size for the ten classes of DRs with $l \geq 8$nt in the coding strands. Genome size (N) and G+C genomic content (G+C (%)) of each parasite are indicated in the upper table. Tb: *T. brucei*, Tc: *T. cruzi*, Lm: *L. major*.

Additionally, $(TT)_n$ and $(AA)_n$ are the most abundant DRs in *T. brucei* and *T. cruzi*, but only $(TT)_n$ is prominent in *L. major*. This difference can be due to the higher G + C content of *L. major* genome (60%) in comparison to the 46% for *T. brucei* and 51% for *T. cruzi*. In contrast, *L. major* presents a higher proportion of DRs containing C. In an attempt to evaluate the influence of the genomic compositional bias in the occurrence of the different DRs, we determine the deviation of the observed repeat frequencies from those expected by chance, based on the nucleotide composition of the genomes (Table 1). In accordance with our earlier report (Duhagon et al., 2001), the occurrence of the majority of DRs is strongly and significantly deviated from serendipity (χ^2 test $p < 0.001$). Above $l = 8$nt, the majority of non-cumulative observed DRs (O) are higher than the E, and this tendency becomes gradually more noticeable as the repeat length increases. Some studies have interpreted this overrepresentation as a result of polymerase mediated strand slippage above a specific repeat length threshold (reviewed in Kelkar et al., 2010). The biggest difference in O versus E values is observed for $(TT)_n$ in the three parasites, which is 6–23 times (O/E) overrepresented (i.e. $O > E$) for $l = 8$nt to more than a thousand for $l = 14$nt. Following TT, $(AA)_n$ presents the highest overrepresentation. Although $(TT)_n$ and $(AA)_n$ repeats are the most abundant in *T. cruzi* and *T. brucei* but not in *L. major*, they still present the highest O/E ratios in the three parasites. This finding suggests the existence of a conserved force that overcomes compositional bias and leads to the overrepresentation of these two types of repeats in the three genomes. The high frequency of $(AA)_n$ and $(TT)_n$ has been attributed to their involvement in putative DNA conformations and chromatin

Table 1
Observed (O) and expected (E) occurrences of perfect poly-dinucleotide repeats of length (l) $6 \leq l \leq 14$ in the coding strands of the Tritryp genomes.

	T. brucei		T. cruzi		L. major	
l=6nt	E	O	E	O	E	O
AA	7367	2011	10166	3666	1736	540
AC	9943	692	22588	1401	11249	1313
AG	9816	744	22357	1740	10979	899
AT	15127	1019	20176	1238	3642	276
CC	3357	522	12548	884	18353	1559
CG	6629	293	24839	859	35818	2247
CT	10240	829	22466	2701	11817	2441
GG	3272	586	12292	1836	17476	843
GT	10110	1148	22235	3063	11533	2046
TT	7816	2735	10056	7751	1915	1432
Total	83678	10579	179723	25139	124517	13596
l=7nt	E	O	E	O	E	O
AA	1963	3766	2499	6847	347	1196
AC	2485	290	5655	991	2847	837
AG	2449	274	5587	736	2763	438
AT	4114	457	5012	831	738	221
CC	781	423	3198	605	5503	908
CG	1539	111	6320	222	10695	1042
CT	2574	338	5620	1194	3009	1444
GG	758	526	3122	1371	5196	673
GT	2537	582	5552	2086	2920	1324
TT	2104	4927	2467	13817	389	3050
Total	21305	11694	45032	28700	34408	11133
l=8nt	E	O	E	O	E	O
AA	523	2194	614	2671	69	677
AC	616	754	1415	4175	674	1591
AG	606	642	1396	2117	652	1045
AT	1039	884	1174	1276	144	274
CC	182	159	815	161	1650	819
CG	357	268	1608	389	3193	2974
CT	641	605	1405	1868	720	3768
GG	176	230	793	355	1545	580
TG	630	1488	1385	4680	696	2617
TT	566	3133	605	5992	79	1803
Total	5337	10357	11210	23684	9422	16148
l=9nt	E	O	E	O	E	O
AA	139	1868	151	1687	14	767
AC	154	272	354	2580	171	861
AG	151	203	349	694	164	509
AT	341	313	345	600	33	119
CC	42	161	208	74	495	1464
CG	83	81	409	275	954	1249
CT	161	214	351	1374	183	1695
GG	41	259	201	184	459	712
GT	158	624	346	4237	176	1238
TT	152	3035	149	4286	16	2156
Total	1424	7030	2864	15991	2666	10770
l=10nt	E	O	E	O	E	O
AA	37	1264	37	1172	3	477
AC	38	109	89	1206	40	335
AG	37	71	87	324	39	227
AT	29	224	28	521	2	81
CC	10	184	53	61	148	1356
CG	19	27	104	27	285	552
CT	40	69	88	295	44	768
GG	9	285	51	147	136	640
GT	39	255	86	1526	42	525
TT	41	2247	36	3129	3	1440
Total	300	4735	659	8408	742	6401

	T. brucei		T. cruzi		L. major	
l=11nt	E	O	E	O	E	O
AA	10	768	9	862	1	176
AC	10	107	22	615	10	313
AG	9	50	22	203	10	133
AT	21	119	18	375	1	81
CC	2	86	14	48	45	731
CG	4	5	27	9	85	226
CT	10	52	22	453	11	508
GG	2	159	13	109	41	324
GT	10	165	22	2283	11	405
TT	11	1417	9	2287	1	641
Total	90	2928	177	7244	215	3538
l=12nt	E	O	E	O	E	O
AA	3	427	2	674	0	76
AC	2	42	6	478	2	203
AG	2	30	5	147	2	112
AT	6	127	4	331	0	55
CC	1	62	3	37	13	288
CG	1	0	7	0	25	85
CT	3	33	5	121	3	359
GG	1	76	3	109	12	136
GT	2	102	5	619	2	308
TT	3	948	2	1965	0	286
Total	23	1847	44	4481	61	1908
l=13nt	E	O	E	O	E	O
AA	1	227	1	644	0	15
AC	1	54	1	242	1	281
AG	1	28	1	132	1	122
AT	2	91	1	295	0	110
CC	0	43	1	41	4	136
CG	0	0	2	2	8	51
CT	1	37	1	376	1	386
GG	0	41	1	92	4	69
GT	1	118	2	866	1	319
TT	1	496	1	1635	0	96
Total	6	1847	11	4325	18	1585
l=14nt	E	O	E	O	E	O
AA	0	102	0	66	0	12
AC	0	27	0	27	0	169
AG	0	28	0	26	0	85
AT	0	121	0	89	0	103
CC	0	20	0	1	1	88
CG	0	0	0	0	2	8
CT	0	34	0	16	0	369
GG	0	16	0	0	1	34
GT	0	81	0	78	0	311
TT	0	209	0	217	0	38
Total	1	429	2	303	5	1179

The values that are not significantly deviated from serendipity (χ^2 test $p<0.001$) are enclosed in grey boxes.

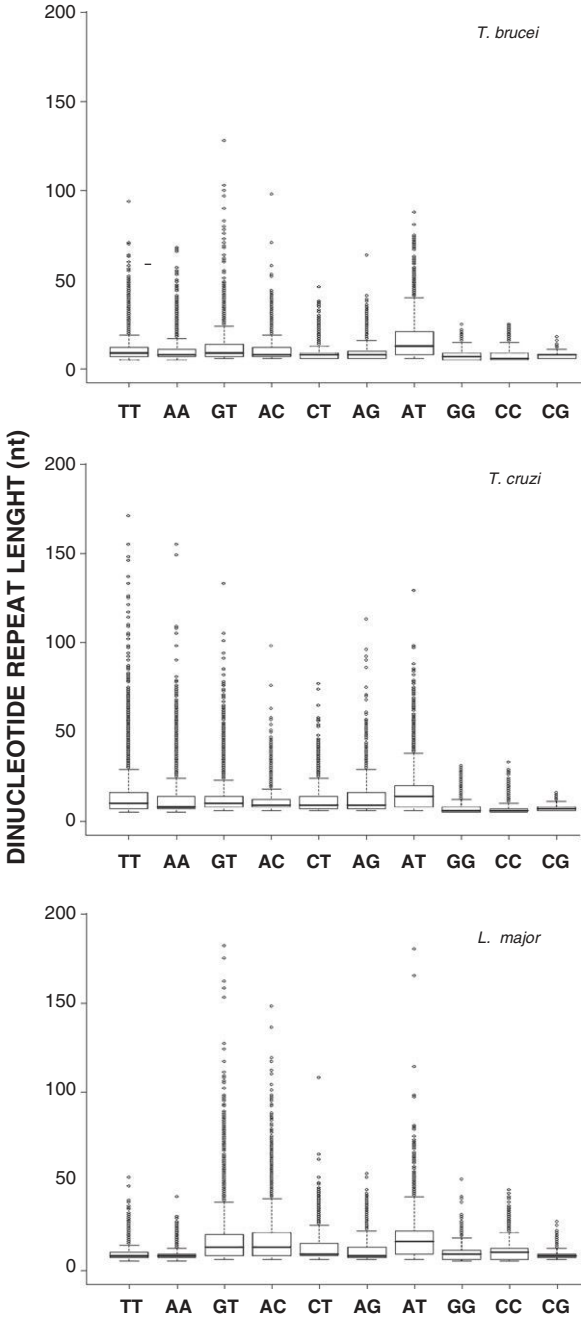


Fig. 2. Length dispersion of DRs. Length boxplots of each class of DRs of length l≥8nt in coding strands. A maximum length of 200nt was arbitrarily graphed for comparison among the organisms. Outliers are set as 98th percentile. Species name corresponding to the data is indicated above each panel.

structure (Sinden, 1994). It is worth mentioning that polyA tails from retroposons, or even processed pseudogenes, could also contribute to their abundance (Toth et al., 2000). In addition, A and T rich post-transcriptional regulatory DNA elements have been described in intergenic regions of the Tritryps and other organisms (Clayton and Shapira, 2007; Di Noia et al., 2000; D'Orso et al., 2003; Haile et al., 2008; McKee and Silver, 2007).

Additionally, $(GT)_n$ is conspicuous in the genomes of all Tritryp, ranking at the second highest position in *L. major* and third in *T. cruzi* and *T. brucei* (Fig. 1A). In *L. major*, the occurrence of this DR is even higher than the occurrences of the $(TT)_n$. Particularly, $(GT)_8$ repeat is 2–4 times more frequent than is expected by chance (χ^2 test, $p<0.0001$) in the three parasites (Table 1) reaching 744–6478 O/E ratio at l = 14nt. The abundance of poly-(GT), that we have previously communicated for *T. cruzi* (Duhagon et al., 2001), has also been reported in human (Sharma et al., 2005) and in ten other different taxa (Toth et al., 2000) pointing out to a strongly conserved mutational or selective mechanism maintaining this repeat.

Interestingly, the $(CT)_n$ is the most frequent DR in *L. major* (Fig. 1A). However, $(CT)_n$ overrepresentation is seen at l = 8nt and over, whereas $(AA)_n$ and $(TT)_n$ deviations are already detected at l = 7nt, suggesting that $(AA)_n$ and $(TT)_n$ still show an important bias in this organism. Indeed, the relative occurrence of $(CT)_n$ within the global occurrences of DRs in each Tritryp seems to accompany the A + T genomic content. $(CT)_n$ proportion could also be favored by the reported role of polypyrimidine sequence in the intergenic regions of coding strands in Tritryps (Stern et al., 2009).

Finally, $(CG)_n$ is the least frequent DR, with the exception of *L. major* (Fig. 1A). Table 1 shows that $(CG)_8$ and $(CC)_8$ are the only underrepresented (i.e. O<E) classes in all the Tritryps (O/E = 0.2–0.9). In fact, $(CG)_n$ is the only repeat that is less abundant than expected by chance at $10 \leq l \leq 14$nt (only in trypanosomes), suggesting a selective pressure against these sequences that overcomes the compositional bias of the genome. In fact, the low frequency of CG repeats has been well recognized (Jurka and Pethiyagoda, 1995; Lowenhaupt et al., 1989; Stallings, 1992; Tautz et al., 1986; Toth et al., 2000) and it has been attributed to the high C→T mutational rate in CpG dinucleotides. This phenomenon could support an increase in the TpG content, thus providing an explanation of the high occurrence of $(GT)_n$ (reviewed in Frank and Lobry, 1999). It has also been proposed that the $(CG)_n$ propensity to adopt particular conformational structures could be implicated in the negative selection of these sequences (Stallings, 1992).

In addition, we analyzed the pattern of dinucleotides using the variable "D" that combines occurrence, length and mismatch (Duhagon et al., 2001) (Fig. 1B). The result observed for D is similar to that for occurrences (Fig. 1A), further supporting the significance of the findings. Nevertheless, there are some differences between O (Fig. 1A) and D (Fig. 1B) outcomes. Indeed, considering the D value, the $(AT)_n$ becomes the fourth most represented class (instead of the fifth) in *T. cruzi* and, it rises above $(AA)_n$ and $(AG)_n$ in *L major*. In addition, differences among the three organisms were observed within a class. Note that while the occurrence of $(GT)_n$ is higher in *T. cruzi* than in the other two Tritryps, the corresponding D value turns out to be the highest in *L. major*. The difference in O and D patterns could be due to the preferential weight on length assumed by variable D. This observation prompted us to perform a more detailed analysis of the DR length distribution.

Although Tritryp multigenic families are known to be rich in diverse repetitive amino acid tracts (CITA), the low conservation of their nucleotide sequences together with the limited coding possibilities of the ten DRs and the degeneration of the genetic code, prevent these repetitive domains to skew the global DR frequency (data not shown).

3.2. The lengths of DRs are class specific and display commonalities among the parasite genomes

To further understand the putative roles of DR sequences, we studied their length distribution. A box-plot representation of each class of DR in the non-coding regions of sense strands is shown in Fig. 2. As the whole, the medians of the lengths are similar on both inter and intra-genomes. Strikingly, the AT repeat presents the longest median and the broader spread of the outliers in the three parasites, perhaps contributing to the differences between O and D discussed above. The biological significance of this signal in the three parasite genomes should be defined. In addition, the $(CG)_n$ shows the narrowest spread of outliers followed by $(GG)_n$ and $(CC)_n$, except in *L. major* where $(TT)_n$ and $(AA)_n$ are comparable. Interestingly, the longest DRs, both in *T. brucei* and *L. major*,is the $(GT)_n$. For *T. cruzi*, longer DRs are formed by the $(AA)_n$ and $(TT)_n$ stretches. In addition,

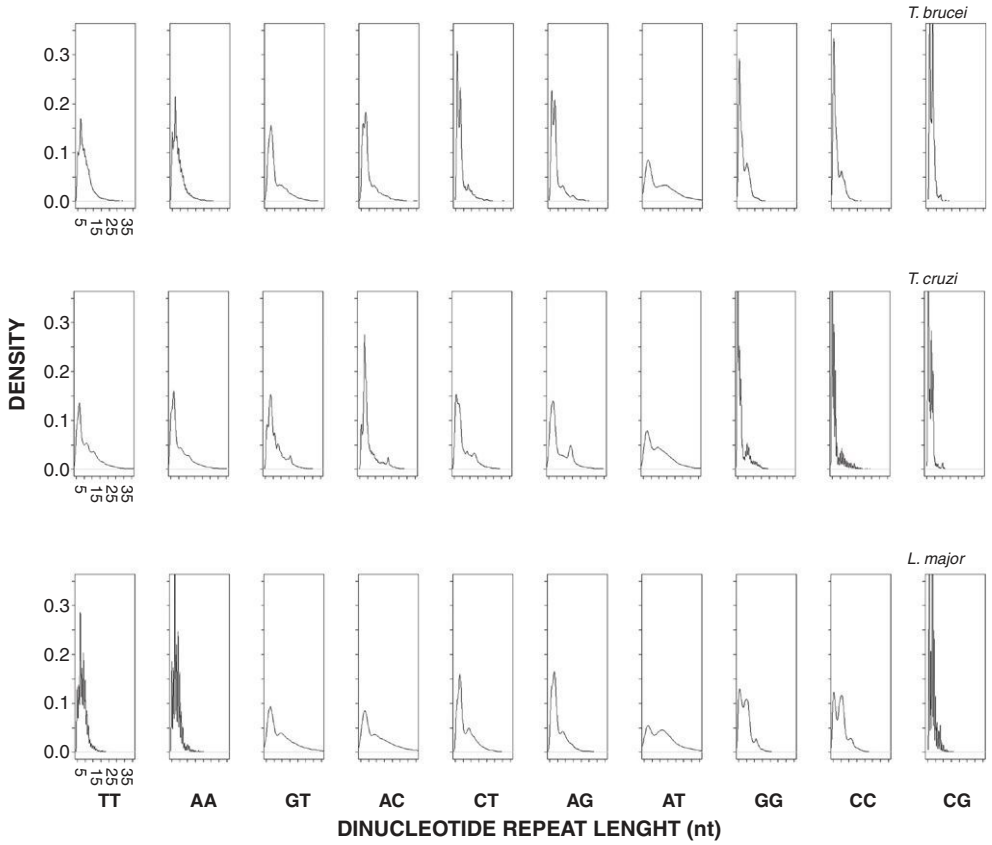


Fig. 3. Length density of DRs. The abundance of DRs in coding strands is shown as a distribution (relative frequency of a dinucleotide for a particular length range). The abscissa plots the length of the repeat (nt) with an arbitrary maximum of 35nt and the ordinate the relative frequency of repeats for a given length. Axis labels are only shown for the left most panels, being identical in the rest of them within a single species. Species name corresponding to the data is indicated above each panel.

$(GT)_n$ medians rank second after $(AT)_n$ in the three genomes. This finding is in agreement with previous reports on four mammalian organisms (*H. sapiens*, *M. musculus*, *C. elegans* and *S. cerevisiae*) by Dokholyan et al.(2000), who found that $(GT/AC)_n$ and $(AG/CT)_n$ are the longest DRs in these organisms and suggested the existence of a specific role for lengths between 20 and 60nt.

In an attempt to further study the potential importance of specific repeat lengths we looked at the length density distribution (Fig. 3). The density of DRs represents the relative frequency of DRs of a given length. Overall, DRs present a monophasic decay as expected by randomness. This is evident for $(AA)_n$ and $(TT)_n$. Nevertheless, most of the curves also display local deviations. For instance, the $(AT)_n$ shows a significant contribution of sequences ranging from 15 to 20nt that is evidenced by a secondary peak in the distribution. This is in agreement with its longest median in Fig. 2. A conspicuous secondary peak is also observed for $(GG)_n$ and $(CC)_n$ in *L. major*, together with other minor peaks. Likewise, $(CG)_n$ repeats show at least two major peaks in a narrow length range that is similar for the three genomes. Multi-peak distributions composed by several minor peaks are observed for the $(GT)_n$, $(AC)_n$, $(CT)_n$ and $(AG)_n$. Altogether, the finding of specific DR length profiles, which are sometimes conserved among the species, suggests that they hold distinctive roles.

3.3. Asymmetrical strand distribution of DRs is widespread in Tritryp genomes

It is well documented that both complementary DNA strands bear base compositional differences that may be due to either selective or mutational differences between them (Frank and Lobry, 1999). The strand asymmetry is proposed to arise from asymmetric DNA related processes such as replication, repair, transcription and chromatin packing. We have previously detected an asymmetrical strand distribution between coding and non-coding strand, for the complementary pair of DRs $(GT)_n$ and $(AC)_n$ (Duhagon et al., 2001). Here, a similar analysis of strand distribution was carried out for the complete Tritryp genomes. The analysis of perfect DRs with lengths $l \geq 8$nt shows that all the complementary pairs, except for CT/AG in *T. brucei*, are asymmetrically distributed among the strands (Fig. 4). As previously shown (Duhagon et al., 2001), the $(GT)_n$ is more frequent in the coding strand than the $(AC)_n$. In addition, $(TT)_n$ is also more frequent in the coding strand in the three genomes and $(CT)_n$ in *T. cruzi* and *L. major*. Interestingly, the $(GG)_n$ is less abundant in the coding strand of *L. major*, representing the unique reverse strand distribution of DRs among the three parasites. When imperfect DRs are included in the analysis, a similar pattern of asymmetric strand distribution between complementary repeated sequences is also observed (data not shown). In conclusion, disproportionate distribution of complementary DRs in the coding/non-coding strands is a common feature in Tritryp genomes. In addition, three out of the four pairs display the same strand orientation bias in the three genomes, pointing out to the existence of a directional mechanism for the generation or maintenance of the asymmetry. Local variations in base content (known as AT skew: $(A-T)/(A+T)$ and GC skew: $(G-C)/(G+C)$) in the coding/non-coding DNA strands have been described in most organisms (Rocha and Danchin, 2001; Mugal et al., 2010; Green et al., 2003). They have been related to the direction of replication and transcription. The G and T are usually more abundant than the C and A in the coding strands. In trypanosomatid parasites, the pattern of GC and AT skews has been previously studied by McDonagh and Nilsson (El-Sayed et al., 2005b; McDonagh et al., 2000; Nilsson and Andersson, 2005). The authors found that trypanosome base skew patterns were similar to those reported for eubacteria but different from *L. major*.

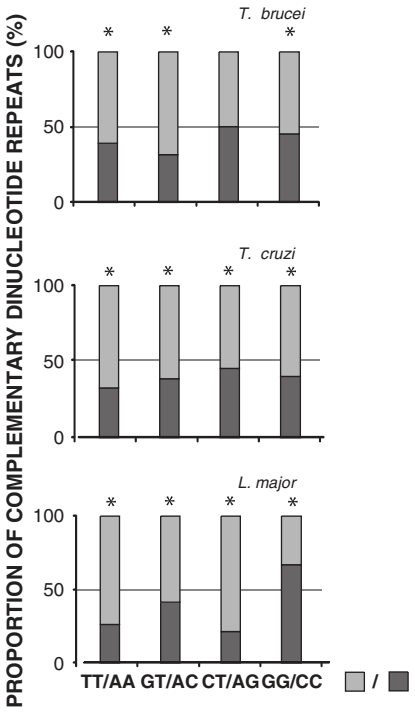


Fig. 4. Relative proportion of complementary DRs. The percentages of complementary dinucleotides repeats of length l≥8nt in the coding strands of the four parasite genomes are shown. The percentage of each DR (■) and its complementary (■) is presented (the occurrence of one member of the pair/the addition of the occurrence of both members of the pair*100). Significance of the deviation between complementary dinucleotides repeats ($p<0.0001$) yielded by the two proportions hypothesis test (Z) is indicated by *. Species name corresponding to the data is indicated above each panel.

They proposed that the gene density and inherent synonymous codon usage could explain, at least in part, the GC and AT skews in these organisms. However, the intergenic regions could be greatly influencing the base skews through the existence of strand asymmetrical specific regulatory DNA elements or even simple repeats. The GC and AT skews observed in eubacteria and trypanosomes could lead to an increase of the $(GG)_n$, $(GT)_n$ and $(TT)_n$ in the leader strand. In fact, these DRs are more frequent in the coding strands of the Tritryps (with the exception of GG in *L. major*). In the particular case of *L. major*, there is an increase in C content in the coding strand (Nilsson and Andersson, 2005), thus $(CC)_n$, $(CT)_n$ and $(TT)_n$ would be expected to be augmented. These are in fact the DRs more represented in the coding strands in *L. major*. Therefore, the abundance of DRs in the coding strands described here, accompanies the strand base compositional bias. Although strand dissymmetry has been proposed to neutrally arise from genome mutational skews (Eckert and Hile, 2009), it might also be selectively maintained because of its putative functionality (e.g. in transcription or DNA replication).

3.4. DRs are not uniformly distributed along the co-directional clusters

The Tritryp polycistronic gene arrangements are necessarily controlled by sequence elements that still remain poorly characterized. In order to study a possible role of DRs in the polycistron dynamics we determine the occurrence of DRs along the co-directional clusters ordered from head to tail (as defined in Materials and methods). When we analyzed absolute polycistronic distances, we observed a DR distribution that mainly resembles polycistron length distribution in all the Tritryps (data not shown). To eliminate the influence in the analysis of length variation among polycistrons, we used the relative distances (Fig. 5). Strikingly, non-parametric Smirnov–Kolmogorov goodness of fit test indicates that most of the DRs are non-uniformly distributed along the polycistrons (24 out of the 30, $p\leq0.05$). The uniformity displayed by $(CT)_n$, $(AG)_n$, $(GG)_n$, $(CC)_n$, $(CG)_n$ in *T. brucei* and $(AC)_n$ in *T. cruzi*, suggests that these sequences are not involved in global polycistron structuring or functionality. Among the non-uniform distributions, 16 out of 30 present a decrease of repeat density at both polycistron boundaries (e.g., $(TT)_n$ and $(AT)_n$ in *T. brucei*, $(GT)_n$ in the three parasites). This trend toward a decrease of DR density at the boundary of the polycistronic unit might be linked to the maintenance of its structural integrity. In fact, in spite of the important sequence divergence in trypanosomatid genomes, they display a strong synteny that might be sustained by the polycistronic architecture. DRs may enhance recombinational processes or polymerase inadequacy resulting in increased local mutational rates (Majewski and Ott, 2000; Jeffreys et al., 1998). Alternatively, position dependent genomic stability might arise from local variations in the efficiency of repair mechanism. Finally, the recent finding of a histone code at the SSRs of *T. cruzi* and *T. brucei* (Respuela et al., 2008; Siegel et al., 2009) strengthens the importance of chromatin configuration in polycistron dynamics (Li et al., 2002). Thus, since microsatellites have been shown to affect chromatin structure, we cannot rule out an involvement of these sequences in the establishment of polycistronic epigenetic signatures. A remarkable exception is represented by $(AA)_n$ in *L. major*, which is conversely more abundant at the ends of the polycistron. Regardless of the direction, the symmetry of these distributions (i.e., similar change at both edges of the polycistron) argues against DR involvement in DNA directional processes. Lastly, some DRs (e.g. $(AC)_n$, or $(GG)_n$) display a more complex distribution. It is worth noting that short and long polycistrons may be governed by different mechanisms; therefore, a unified analysis comprising all lengths might be masking specific roles for DRs in particular classes of polycistrons.

In conclusion, although the biological basis of the patterns described above is still unclear, they support an active role of DRs in polycistron dynamics. To our knowledge this is the first report of DR distribution along polycistronic units.

3.5. Tritryps exhibit differences in location of DRs within the intergenic region

Previous research has shown that DRs are abundant in the intergenic regions of trypanosomatid genes (Andersson et al., 1998; Duhagon et al., 2001; Nilsson and Andersson, 2005). Considering the hypothesis of these sequences bearing roles in post-transcriptional stages of gene expression, we asked about their distribution along the intergenic regions. Since, the analysis using absolute intergenic distance was strongly biased by the interCDS length distribution (data not shown), we used the relative distance of the repeat from the end of the CDS. In Fig. 6, the occurrence of DRs along the region between the 3′ end of one CDS (5′ end of the interCDS region) to the 5′ beginning of the following one (3′ end of the interCDS region) is presented. The boundaries of the interCDS are therefore enriched in UTRs. The Smirnov–Kolmogorov goodness of fit test showed that most DRs are not uniformly distributed along the intergenic regions ($P\leq0.05$). Exceptions are $(CG)_n$ in *L. major* and *T. brucei,* and $(CT)_n$ $(AG)_n$, $(AT)_n$, $(CC)_n$ in *T. brucei*. Some plots show a predominance of repeats toward one end of the intergenic region; e.g., in *T. brucei* and *L. major* $(GT)_n$ are prevalent toward the 3′UTRs and conversely, $(AC)_n$ toward the 5′UTRs in *L. major*. Interestingly, most of *T. cruzi* plots show increased DR abundance at both intergenic ends simultaneously, producing "U" shaped curves. Meanwhile, edge single peak curves are predominant in *L. major* whereas smoother patterns are detected in the genome of *T. brucei*. Finally, few DRs are under-represented near the CDSs. One example is $(AA)_n$ in *L. major*. In particular, all polypyrimidine DRs (TT; CC and TC) are under-represented at the 3′ end of the intergenic region, i.e. the beginning of the downstream coding region. It is known that the conserved pyrimidine-rich tract that precedes the SL addition site is located at an average of 53 nucleotides to the CDS (Campos et al., 2008) and that the average

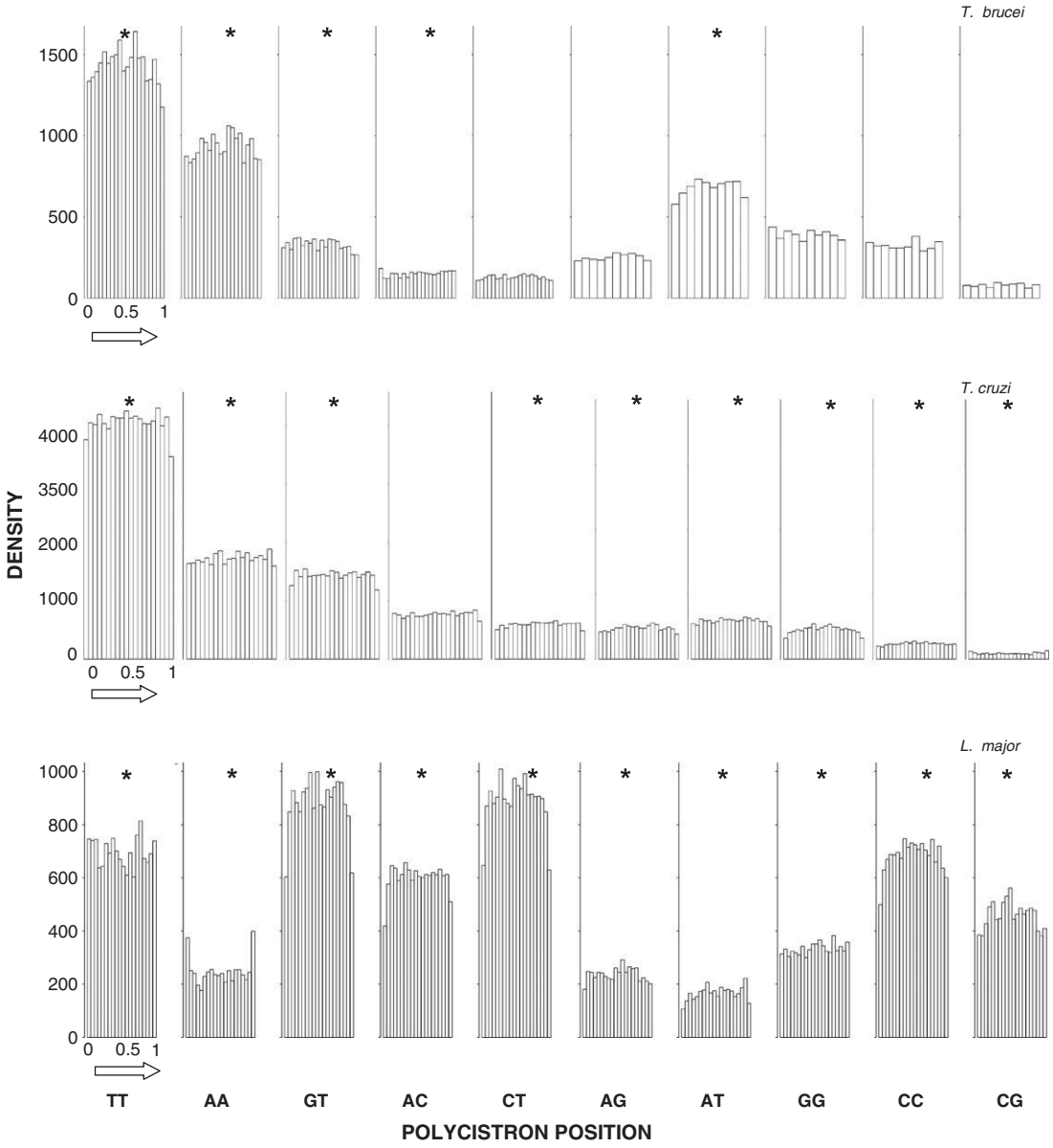


Fig. 5. DRs distribution along polycistrons. The histogram of each class of DR in the sense strand of polycistrons is shown. The relative distance from the strand-switch from head to tail (as indicated by the arrow) is depicted in the abscissa. Rejection of the null hypothesis of uniform distribution by Smirnov–Kolmogorov goodness of fit test ($p \leq 0.0001$) is indicated by *. Species name corresponding to the data is indicated above each panel.

intergenic region length has been calculated to be 1024nt in *T. cruzi* (El-Sayed et al., 2005a). Thus, in agreement with Campos et al., it is possible to speculate that the 5% 3′ most region of the intergenic region is devoid of pyrimidine-rich DRs as a strategy to achieve the positional conservation of the functional polypyrimidine tract. Similar conclusions can be withdrawn for *L. major* and *T. brucei* considering their average intergenic and UTR regions size. Besides, a reduction of polypyrimidine containing DRs is also observed at the 5′end of the intergenic region in *L. major* (for TT; CC and TC), and *T. brucei* (only for CT). This could be analogously explained, as driven by the conservation of the polypyrimidine tract that modulates the poly-adenylation process. However, the comparatively less conservation of the 5′end versus the 3′end of the intergenic region patterns could be attributed to the broader size distribution of 5′UTR lengths compared to 3′UTR (Campos et al., 2008). In addition; since there are at least two polypyrimidine motifs at the vicinity of the polyA addition site (Campos et al., 2008), the interpretation of the polypyrimidine distribution at this region becomes more complex.

Inter-specific differences in repeat distribution along the intercistronic regions might reflect a divergent use of DRs in the post-translational regulation of each parasite. Nevertheless, the general enrichment of DRs in the proximity of the CDSs seems to be a common feature in the three genomes, thus favoring their putative function as UTR sequence elements.

4. Conclusion

The origin, evolution and functionality of microsatellites in the eukaryotic genome are still scarcely understood. Various reports have uncovered the non-stochastic patterns of abundance and distribution of DRs in the genomes. Although many roles have been ascribed to them, the relative extent of neutrality and selectivity of these sequences is not yet understood. Trypanosomes are well known for the lack of canonical higher eukaryote sequence elements; thus, their molecular biology may rely on signals from non-traditional sequences. We and others have proposed a functional role for DRs in the gene expression of these

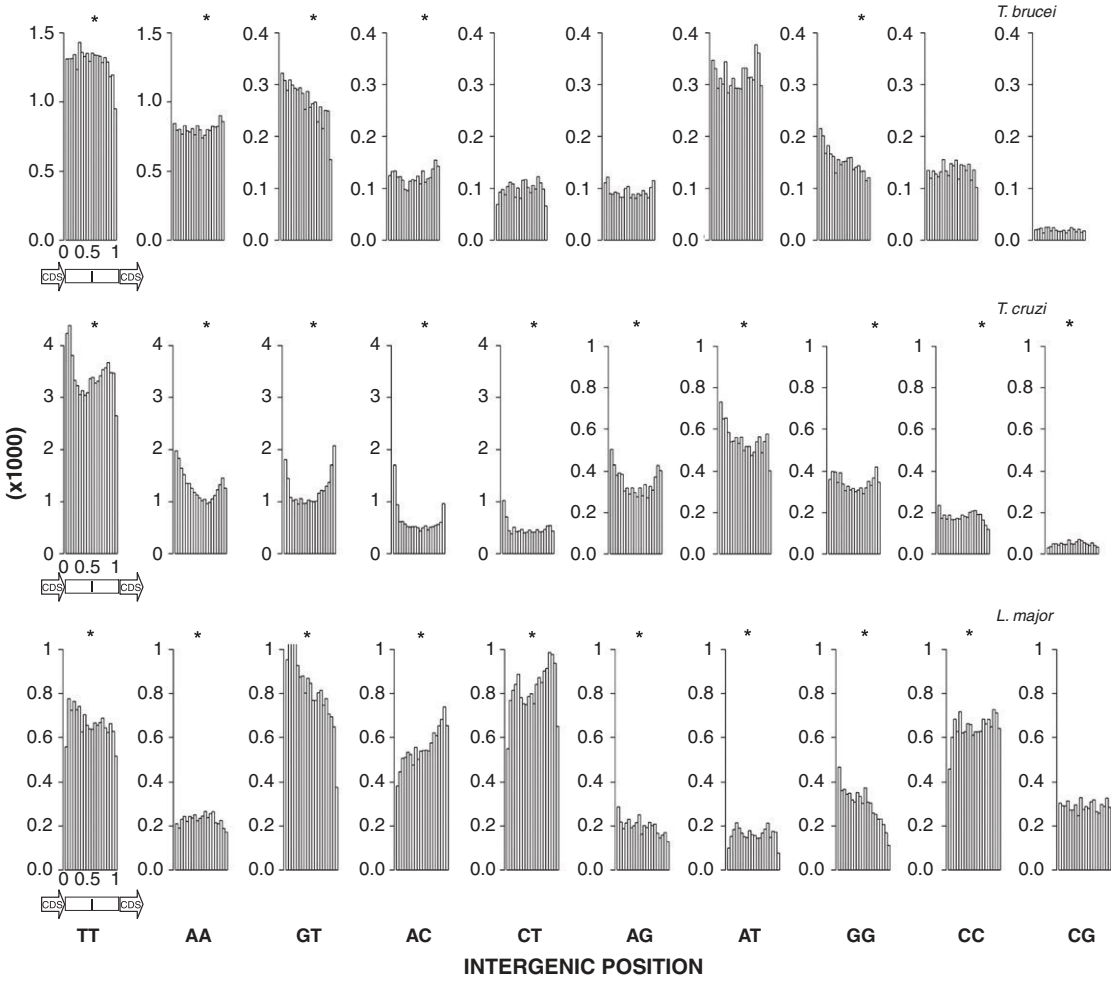


Fig. 6. DR distribution along intergenic regions. The histogram of each class of DRs in the inter-CDS regions (I) is shown. The relative distance from the 3′ end of one CDS to the 5′ beginning of the following CDS is depicted (and is indicated by the arrow). Rejection of the null hypothesis of uniform distribution by Smirnov–Kolmogorov goodness of fit test ($p \leq 0.0001$) is indicated by *. Species name corresponding to the data is indicated above each panel.

parasites. Here we address this hypothesis through the comparative bioinformatic analysis of their distribution in the coding strands of the complete genomes of the Tritryps. We found that DRs are present at frequencies that greatly differ from the stochastic expectations. In addition, they exhibit common and specific patterns of occurrence, length, strand distribution, and location both along the polycistron and intergenic regions. These associations with genomic regions actively involved in molecular processes suggest that DRs may constitute evolutionary selected sequence elements.

Acknowledgments

This work was financially supported by FIRCA n° R03 TW05665-01, PEDECIBA and CSIC, UdelaR. MAD received a PEDECIBA fellowship. We thank Dr. F. Alvarez-Valin and Dr. H. Musto for helpful discussions.

References

Aguero, F., Verdun, R.E., Frasch, A.C., Sanchez, D.O., 2000. A random sequencing approach for the analysis of the *Trypanosoma cruzi* genome: general structure, large gene and repetitive DNA families, and gene discovery. Genome Res. 10, 1996–2005.

Andersson, B., Aslund, L., Tammi, M., Tran, A.N., Hoheisel, J.D., Pettersson, U., 1998. Complete sequence of a 93.4-kb contig from chromosome 3 of *Trypanosoma cruzi* containing a strand-switch region. Genome Res. 8, 809–816.

Benson, G., 1999. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 27, 573–580.

Berriman, M., et al., 2005. The genome of the African trypanosome *Trypanosoma brucei*. Science 309, 416–422.

Buschiazzo, E., Gemmell, N.J., 2006. The rise, fall and renaissance of microsatellites in eukaryotic genomes. Bioessays 28, 1040–1050.

Campos, P.C., Bartholomeu, D.C., DaRocha, W.D., Cerqueira, G.C., Teixeira, S.M., 2008. Sequences involved in mRNA processing in *Trypanosoma cruzi*. Int. J. Parasitol. 38, 1383–1389.

Clayton, C.E., 2002. Life without transcriptional control? From fly to man and back again. EMBO J. 21, 1881–1888.

Clayton, C., Shapira, M., 2007. Post-transcriptional regulation of gene expression in trypanosomes and leishmanias. Mol. Biochem. Parasitol. 156, 93–101.

Cribb, P., Esteban, L., Trochine, A., Girardini, J., Serra, E., 2010. *Trypanosoma cruzi* TBP shows preference for C/G-rich DNA sequences *in vitro*. Exp. Parasitol. 124, 346–349.

Di Noia, J.M., D'Orso, I., Sanchez, D.O., 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.

Dokholyan, N.V., Buldyrev, S.V., Havlin, S., Stanley, H.E., 2000. Distributions of dimeric tandem repeats in non-coding and coding DNA sequences. J. Theor. Biol. 202, 273–282.

D'Orso, I., De Gaudenzi, J.G., Frasch, A.C., 2003. RNA-binding proteins and mRNA turnover in trypanosomes. Trends Parasitol. 19, 151–155.

Duhagon, M.A., Dallagiovanna, B., 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, 98–103.

Eckert, K.A., Hile, S.E., 2009. Every microsatellite is different: intrinsic DNA features dictate mutagenesis of common microsatellites present in the human genome. Mol. Carcinog. 48, 379–388.

El-Sayed, N.M., et al., 2005a. The genome sequence of *Trypanosoma cruzi*, etiologic agent of Chagas disease. Science 309, 409–415.

El-Sayed, N.M., et al., 2005b. Comparative genomics of trypanosomatid parasitic protozoa. Science 309, 404–409.

Epplen, J.T., Kyas, A., Maueler, W., 1996. Genomic simple repetitive DNAs are targets for differential binding of nuclear proteins. FEBS Lett. 389, 92–95.

Frank, A.C., Lobry, J.R., 1999. Asymmetric substitution patterns: a review of possible underlying mutational or selective mechanisms. Gene 238, 65–77.

Green, P., Ewing, B., Miller, W., Thomas, P.J., Green, E.D., 2003. Transcription-associated mutational asymmetry in mammalian evolution. Nat. Genet. 33, 514–517.
Haile, S., Dupe, A., Papadopoulou, B., 2008. Deadenylation-independent stage-specific mRNA degradation in Leishmania. Nucleic Acids Res. 36, 1634–1644.
Jeffreys, A.J., Murray, J., Neumann, R., 1998. High-resolution mapping of crossovers in human sperm defines a minisatellite-associated recombination hotspot. Mol. Cell 2, 267–273.
Jurka, J., Pethiyagoda, C., 1995. Simple repetitive DNA sequences from primates: compilation and analysis. J. Mol. Evol. 40, 120–126.
Kelkar, Y.D., Strubczewski, N., Hile, S.E., Chiaromonte, F., Eckert, K.A., Makova, K.D., 2010. What is a microsatellite: a computational and experimental definition based upon repeat mutational behavior at A/T and GT/AC repeats. Genome Biol. Evol. 2, 620–635.
Li, Y.C., Korol, A.B., Fahima, T., Beiles, A., Nevo, E., 2002. Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. Mol. Ecol. 11, 2453–2465.
Lowenhaupt, K., Rich, A., Pardue, M.L., 1989. Nonrandom distribution of long mono- and dinucleotide repeats in *Drosophila* chromosomes: correlations with dosage compensation, heterochromatin, and recombination. Mol. Cell. Biol. 9, 1173–1182.
Majewski, J., Ott, J., 2000. GT repeats are associated with recombination on human chromosome 22. Genome Res. 10, 1108–1114.
Martinez-Calvillo, S., Yan, S., Nguyen, D., Fox, M., Stuart, K., Myler, P.J., 2003. Transcription of *Leishmania major* Friedlin chromosome 1 initiates in both directions within a single region. Mol. Cell 11, 1291–1299.
Martinez-Calvillo, S., Nguyen, D., Stuart, K., Myler, P.J., 2004. Transcription initiation and termination on *Leishmania major* chromosome 3. Eukaryot. Cell 3, 506–517.
McDonagh, P.D., Myler, P.J., Stuart, K., 2000. The unusual gene organization of *Leishmania major* chromosome 1 may reflect novel transcription processes. Nucleic Acids Res. 28, 2800–2803.
McKee, A.E., Silver, P.A., 2007. Systems perspectives on mRNA processing. Cell Res. 17, 581–590.
Mugal, C.F., Wolf, J.B., von Grunberg, H.H., Ellegren, H., 2010. Conservation of neutral substitution rate and substitutional asymmetries in mammalian genes. Genome Biol. Evol. 2, 19–28.
Nilsson, D., Andersson, B., 2005. Strand asymmetry patterns in trypanosomatid parasites. Exp. Parasitol. 109, 143–149.
Obado, S.O., Taylor, M.C., Wilkinson, S.R., Bromley, E.V., Kelly, J.M., 2005. Functional mapping of a trypanosome centromere by chromosome fragmentation identifies a 16-kb GC-rich transcriptional â€oestrand-switchâ€•domain as a major feature. pp. 36–43.
Palenchar, J.B., Bellofatto, V., 2006. Gene transcription in trypanosomes. Mol. Biochem. Parasitol. 146, 135–141.
Plohl, M., Luchetti, A., Mestrovic, N., Mantovani, B., 2008. Satellite DNAs between selfishness and functionality: structure, genomics and evolution of tandem repeats in centromeric (hetero)chromatin. Gene 409, 72–82.
Respuela, P., Ferella, M., Rada-Iglesias, A., Aslund, L., 2008. Histone acetylation and methylation at sites initiating divergent polycistronic transcription in *Trypanosoma cruzi*. J. Biol. Chem. 283, 15884–15892.
Rocha, E.P., Danchin, A., 2001. Ongoing evolution of strand composition in bacterial genomes. Mol. Biol. Evol. 18, 1789–1799.
Sharma, V.K., Brahmachari, S.K., Ramachandran, S., 2005. (TG/CA)n repeats in human gene families: abundance and selective patterns of distribution according to function and gene length. BMC Genomics 6, 83.
Sharma, P.C., Grover, A., Kahl, G., 2007. Mining microsatellites in eukaryotic genomes. Trends Biotechnol. 25, 490–498.
Siegel, T.N., et al., 2009. Four histone variants mark the boundaries of polycistronic transcription units in *Trypanosoma brucei*. Genes Dev. 23, 1063–1076.
Sinden, R.R., 1994. DNA Structure and Function. Academic Press, San Diego.
Stallings, R.L., 1992. CpG suppression in vertebrate genomes does not account for the rarity of (CpG)n microsatellite repeats. Genomics 13, 890–891.
Stern, M.Z., et al., 2009. Multiple roles for polypyrimidine tract binding (PTB) proteins in trypanosome RNA metabolism. RNA 15, 648–665.
Tautz, D., Trick, M., Dover, G.A., 1986. Cryptic simplicity in DNA is a major source of genetic variation. Nature 322, 652–656.
Thomas, S., Green, A., Sturm, N.R., Campbell, D.A., Myler, P.J., 2009. Histone acetylations mark origins of polycistronic transcription in *Leishmania major*. BMC Genomics 10, 152.
Toth, G., Gaspari, Z., Jurka, J., 2000. Microsatellites in different eukaryotic genomes: survey and analysis. Genome Res. 10, 967–981.
Wall, L., Christiansen, T., Orwant, J., 2000. Programming Perl, 3rd ed. O'Reilly, Beijing; Cambridge, Mass.
Zhang, W., He, L., Liu, W., Sun, C., Ratain, M.J., 2009. Exploring the relationship between polymorphic (TG/CA)n repeats in intron 1 regions and gene expression. Hum. Genomics 3, 236–245.

3.1.2 Análisis global de patrones de curvatura intrínseca en los genomas de Tritryps

La doble hélice del ADN exhibe constantes deformaciones al momento de cumplir con sus funciones biológicas. Las proteínas son responsables muchas veces de estos cambios, aunque la propia secuencia de ADN es en muchos casos la que determina intrínsecamente distorsiones particulares de la forma B clásica (Gabrielian, A. *et al.* 1997). Aún más, existen una serie de casos caracterizados en donde se muestra que las proteínas de hecho reconocen regiones que son intrínsecamente propensas a adoptar la conformación que es inducida por la unión a la proteína (Grove, A. *et al.* 1996; Crothers, D. M. 1998). Un claro ejemplo, que ha sido estudiado en detalle, es la formación de nucleosomas. En este modelo se ha demostrado que existe una preferencia del posicionamiento de los mismos sobre secuencias que son propensas a curvarse, lo cual reduce el costo energético requerido para producir estas conformaciones (Garcia, H. G. *et al.* 2007; Miele, V. *et al.* 2008; Nair, T. M. 2010; Travers, A. A. *et al.* 2012; De Santis, P. *et al.* 2013).

En particular, la curvatura intrínseca del ADN ha sido implicada en una variedad de regiones en donde hay interacciones de proteínas con ADN, como ser los sitios de recombinación (Milot, E. *et al.* 1992; Timchenko, T. V. *et al.* 2002), orígenes de replicación (Eckdahl, T. T. *et al.* 1990; Gimenes, F. *et al.* 2008), regiones de anclaje al andamiaje nuclear (Fiorini, A. *et al.* 2006), centrómeros (Bechert, T. *et al.* 1999) y promotores transcripcionales (Nair, T. M. 1998; Gabrielian, A. E. *et al.* 1999-2000; Potaman, V. N. *et al.* 2005; Gimenes, F. *et al.* 2008).

La curvatura intrínseca está dada por la contribución local de la geometría de los nucleótidos sucesivos y se describe por una serie de parámetros (los ángulos de *roll*, *tilt* y *twist*) que permiten definir con precisión el posicionamiento de las bases con respecto al eje de la doble hélice. Estos parámetros pueden ser estimados a partir de la secuencia basándose en resultados experimentales y modelos teóricos de curvatura (Vlahovicek, K. *et al.* 2003).

Como mencionamos en la introducción, el número de señales descritas que participan en los procesos involucrados en el metabolismo del ADN en tripanosomátidos es extremadamente bajo. La bibliografía muestra que la descripción de elementos se ha focalizado en la búsqueda de secuencias consenso a nivel de la estructura primaria del ADN. Sin embargo, desde hace varios años se está acumulando evidencia de que la estructura secundaria del ADN también juega un rol

en estos procesos, hecho que no se ha tenido en cuenta al momento de realizar las búsquedas de motivos funcionales. Es así que nos propusimos hacer un análisis global de la distribución de la curvatura con el objetivo de observar si este tipo de señal tiene relevancia en los genomas de estos parásitos.

Para el desarrollo de este objetivo y utilizando una copia local del software bend.it obtuvimos la curvatura para los cromosomas completos de los Tritryps. Esto fue realizado utilizando *scripts* desarrollados en lenguajes perl y python. Con estos resultados, comprobamos en primera instancia que las regiones de cambio de hebra, de *clusters* divergentes, son sitios donde se observa una acumulación regional de curvatura que resulta mayor que la de otros lugares del genoma del mismo tamaño tomadas al azar. A partir de este resultado, realizamos la búsqueda *de novo* de regiones en donde se produjeran acumulaciones de curvatura de similares a las observadas en dichas regiones. Es así que definimos el parámetro RIIC que integra la curvatura en una región (*Regional Integrated Intrinsic Curvature*) y buscamos mediante *scripts* en lenguaje R zonas de alto RIIC en los genomas. Los resultados mostraron que las regiones de alta curvatura coinciden con marcadores de inicio transcripcional, tanto en los DSSRs como en regiones internas a los DGCs. La coincidencia con regiones internas fue comprobada en *L. major* ya que es el único organismo donde hay datos en este sentido. Además, en *L. major* y *L. infantum* se comprobó que las curvaturas permanecen en regiones sinténicas independientemente, en muchos casos, de la conservación de secuencia en la región. Los resultados en *T. cruzi* fueron similares, aunque no se pudieron hacer comparaciones con sitios internos a los DGCs debido a que no se disponen datos al respecto. En *T. brucei* observamos una coincidencia de las regiones curvadas con la base J y una llamativa concentración en las regiones subteloméricas caracterizadas por la presencia de genes y pseudogenes de VSG. Esta localización particular de las regiones de alta curvatura demuestra que la estructura secundaria del ADN sirve como señal para los procesos moleculares en los genomas de tripanosomátidos.

Por otro lado, decidimos centrarnos en la curvatura intrínseca de los promotores de ARN polimerasa I (ARNPI). Como mencionamos en la introducción estos promotores se caracterizan en eucariotas, en general, por no presentar conservación de secuencia aunque si conservan patrones de curvatura alrededor del sitio de inicio transcripcional. En este sentido comprobamos que las características conformacionales de hecho, se conservan en los promotores de ARN ribosomal (ARNr) de los tripanosomátidos. Interesantemente, demostramos también que estos patrones van más allá de los promotores de ARNr ya que son compartidos por los

otros promotores de ARNPI en *T. brucei*, los cuales no presentan conservación de secuencia entre sí. Este hallazgo pone en evidencia que la maquinaria de ARNPI, independientemente del gen que transcriba, requiere una curvatura intrínseca determinada, resultado que únicamente puede ser comprobado en *T. brucei* ya que hasta la fecha, es el único organismo para el que se ha descrito que esta polimerasa es capaz de transcribir otros genes además de los ARNr.

Estos resultados dieron lugar a la publicación de un trabajo, a la redacción de un manuscrito completo y una comunicación corta que se encuentran en las fases finales de revisión para ser enviados a publicar. Estos documentos se presentan a continuación.

Material suplementario en:

http://lim.fcien.edu.uy/tesis/smircich/Mat_suplementario/

OPEN ACCESS Freely available online

PLOS ONE

Genomic Analysis of Sequence-Dependent DNA Curvature in *Leishmania*

Pablo Smircich[1,2], Diego Forteza[1], Najib M. El-Sayed[3], Beatriz Garat[1]*

1 Laboratorio de Interacciones Moleculares, Facultad de Ciencias, Montevideo, Uruguay, **2** Departamento de Genética, Facultad de Medicina, Montevideo, Uruguay, **3** Department of Cell Biology and Molecular Genetics and Center for Bioinformatics and Computational Biology, University of Maryland College Park, Maryland, United States of America

Abstract

Leishmania major is a flagellated protozoan parasite of medical importance. Like other members of the Trypanosomatidae family, it possesses unique mechanisms of gene expression such as constitutive polycistronic transcription of directional gene clusters, gene amplification, mRNA *trans*-splicing, and extensive editing of mitochondrial transcripts. The molecular signals underlying most of these processes remain under investigation. In order to investigate the role of DNA secondary structure signals in gene expression, we carried out a genome-wide *in silico* analysis of the intrinsic DNA curvature. The *L. major* genome revealed a lower frequency of high intrinsic curvature regions as well as inter- and intra- chromosomal distribution heterogeneity, when compared to prokaryotic and eukaryotic organisms. Using a novel method aimed at detecting region-integrated intrinsic curvature (RIIC), high DNA curvature was found to be associated with regions implicated in transcription initiation. Those include divergent strand-switch regions between directional gene clusters and regions linked to markers of active transcription initiation such as acetylated H3 histone, TRF4 and SNAP50. These findings suggest a role for DNA curvature in transcription initiation in *Leishmania* supporting the relevance of DNA secondary structures signals.

Citation: Smircich P, Forteza D, El-Sayed NM, Garat B (2013) Genomic Analysis of Sequence-Dependent DNA Curvature in *Leishmania*. PLoS ONE 8(4): e63068. doi:10.1371/journal.pone.0063068

Editor: Stefan Maas, NIGMS, NIH, United States of America

Received November 28, 2012; **Accepted** March 27, 2013; **Published** April 30, 2013

Funding: This work was supported by Fondo Clemente Estable (Agencia Nacional de Investigación e Innovación); Comisión Sectorial de Investigación Científica (Universidad de la Republica) and Programa de Desarrollo de Ciencias Básicas. PS received a PhD fellowship (Agencia Nacional de Investigación e Innovación). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: bgarat@fcien.edu.uy

Introduction

Leishmania is a flagellated protozoan parasite (order Kinetoplastida) of significant medical importance in tropical and subtropical regions of the world. The parasite alternates between an intracellular amastigote form residing in vertebrate macrophages and an extracellular promastigote form living in the digestive tract of sandflies. The numerous human-infective *Leishmania* species cause a spectrum of diseases known as leishmaniasis, ranging from asymptomatic to lethal infection of internal organs. It has been estimated that more than two million new leishmaniasis cases occur each year and that 367 million people are at risk of infection [1].

Leishmania and other members of the Trypanosomatidae family possess unique mechanisms of gene expression, nonetheless, little is known about the nucleic acid signals driving them [2].

Although no precise element participating in chromosome replication, segregation and mitotic stability has been described in *Leishmania*, sequences with similarity to the yeast autonomously replicating consensus sequence and centromere DNA elements have been detected in *L. donovani* [3]. Functional characterization using unstable artificial chromosomes suggests the existence of multiple dispersed elements for mitotic stability in *L. major* [4]. Repeated sequences (direct or inverted) involved in DNA rearrangements, alteration of gene copy number (deletion or amplification), formation of extrachromosomal circular or linear amplicons and supernumerary chromosomes have been described in *Leishmania* [5,6]. In addition, retroposon traces have been reported [7,8,9] and their involvement in mRNA instability and in the control of transcription initiation have been proposed [10].

Similar to other eukaryotes, transcription initiation by RNA polymerase I and III occurs at defined core promoters in trypanosomatids [11,12,13]. Canonical signals for RNA polymerase II recruitment have been characterized for the promoters of spliced leader, (SL) genes [14,15], but not for the transcription of genes contained within the directional gene clusters (DGCs). Nuclear run-on assays performed in *L. major* indicate that RNA polymerase II transcription of genes initiates at the beginning of each divergent DGC [16,17]. ChIP-chip assays in *L. major* revealed the enrichment of acetylated H3 histone at divergent strand switch regions (SSRs), as well as the increased binding frequency of two transcription factors –TRF4 and SNAP50. Interestingly, these features also occur at other specific regions putatively related to transcription initiation [18]. Although a high G+C content of SSRs has been observed, no signals such as TATA box or other typical RNA polymerase II core promoter elements have been detected [19]. Transcription termination signals have not been clearly defined for RNA polymerase II transcription of protein-coding genes, although a tract of Ts seems to be required in the case of the spliced leader genes in *Leishmania tarentolae* [20].The

analysis of the regions between convergent DGCs, where termination signals may occur, reveals the presence of tRNA genes as well as other genes transcribed by RNA polymerase III [21,22]. More recently, the presence of the modified base J has been observed at polimerase II transcription termination sites by chromatin immunoprecipitation studies. The results show the importance of this modified base in the transcription mechanisms of these parasites [23]. In the case of RNA polymerase I, regions with the ability to form stem-loop conformations reminiscent of the prokaryotic *rho*-independent termination have been described in *L. infantum* [24] and *L. major* [25]. RNA polymerase III termination in *L. major* occurs at T runs averaging 4.87 in length [21]. The maturation of individual trypanosomatid mRNAs derived from long nascent transcripts is marked by the *trans*-splicing of a mini-exon sequence at the 5′ end and polyadenylation of the 3′end of the processed mRNA [26]. While conserved eukaryotic splicing signals (AG dinucleotide at the 3′ splice site and upstream polypyrimidine tract) have been reported for the SL addition process, no specific consensus sequence or site selection mechanism have been identified for polyadenylation. Nevertheless, *trans*-splicing and polyadenylation of adjacent genes are co-ordinated [27]. Efforts to determine the *cis*-elements responsible for post-transcriptional regulation of gene expression have led to the identification of some sequence elements, secondary structures or a combination of both (mostly in the 3′ untranslated regions (UTRs)) [28,29,30,31,32,33]. In summary, sequence elements regulating transcriptional and post-transcriptional processes, particularly for those genes transcribed by RNA polymerase II, remain largely unknown in *Leishmania*.

Difficulties in identifying DNA regulatory signals in trypanosomatids may be derived from the focus on primary structure analysis of DNA. Nevertheless, DNA conformations have been largely recognized as signals for regulation of DNA function. While evidence for the role of conformational signals in replication, DNA rearrangements and gene expression continues to accumulate in prokaryotic and eukaryotic cells [34], little is known about genomic DNA conformation in kinetoplastid parasites. Early work on *L. tarentolae* mitochondrial minicircle DNA showed the anomalous migration of restriction fragments due to the natural curvature of the DNA helix [35,36]. More recently, bioinformatic analyses of some strand switch regions in *L. major* suggested a functional role for DNA secondary structures as replication or transcription boundaries [37]. A bias in nucleotide composition [18] and poly-dinucleotides abundance [38] has also been reported for those regions.

We have carried out a detailed genome-wide analysis of intrinsic curvature (IC) in *L. major* in order to explore the enormous potential of DNA regulatory signals. When compared to other organisms, the *L. major* genome revealed a lower frequency of high intrinsic curvature regions as well as inter- and intra- chromosomal distribution heterogeneity. Using a novel method aimed at detecting region-integrated intrinsic curvature (RIIC) and based on the additive contribution of IC along regions of a given length, we identified divergent SSRs as high scoring regions. Since those regions are thought to be implicated in transcription initiation, and assuming that curvature profiles provide a relevant signal, we used those RIIC characteristics to search the rest of the *L. major* genome for other regions with similar predicted curvature output. The identified regions matched regions within the DGCs which have been reported to be associated with markers of active transcription initiation (acetylated H3 histone, TRF4 and SNAP50) [18]. A high degree of conservation of the curvature locations was observed for the two *Leishmania* species studied, that is *L. major* and *L. infantum*.

Materials and Methods

Data Sources

The genome data for L. major, L. infantum, L. mexicana, and L. braziliensis, T. cruzi and T. brucei were downloaded from TritrypDB (version 2.1 for L. major, T. brucei and T. cruzi, version 3.1 for L. infantum, L. mexicana and L. braziliensis). A fragment of approx. 7 Mb (spanning bases 136,666,200 to 143,535,568) of human chromosome 1 (Chr 1) was obtained from the NCBI reference sequence (Build 37.2) and used as an external reference group. The E. coli BW2952 strain genome sequence was downloaded from NCBI (Acc. NC_012759.1) and used as a prokaryotic reference genome.

Genome-wide Intrinsic Curvature Calculation

Individual chromosomes were split into 200 Kbases (Kb) non overlapping fragments and submitted for IC prediction using in-house scripts. The bend.it algorithm [39] was kindly provided by Dr. S. Pongor and was run locally. We used the default window size (31 bp) and bendability values (derived from nucleosome binding and DNaseI analysis) to estimate curvature. The output consisted of IC predicted angles per helical turn of the double helix (degrees/hel. turn). Results for all the fragments were parsed using in-house scripts and wiggle (WIG) files were generated for each chromosome in order to visualize the obtained results in genome browsers. Results were filtered and/or further analyzed using in-house scripts either in Python or in R programming languages [40].

Curvature Analysis of Putative Transcription Start Sites

For the statistical analysis of the IC of regions of interest, the area under the curvature plot was calculated using the Riemann sum [41]. For a given region, the value of this region integrated intrinsic curvature (RIIC) parameter was compared to a density function representing the population of RIIC scores for equal-length regions in the genome. This probability density function was estimated by Gaussian kernel density estimation based on 5×10^3 RIIC values for random windows in the parasite genome (excluding other regions of interest). A region was classified as highly curved if its RIIC score was greater than the 95% confidence interval calculated using the estimated density function ($p<0.05$) (for a graphical representation see Supplementary Figure 1). SSRs were selected for RIIC analysis. The coordinates for all the SSRs were collected using in-house perl scripts and manually curated. For each analysis, the region length was defined based on the distance between the adjacent DGCs. In addition, SSRs were considered as convergent (CSSR) or divergent (DSSR) depending on the orientation of the flanking DGCs. The number of SSRs that could be classified as highly curved regions was counted. The likelihood of counting a given number of highly curved regions was evaluated using the exact binomial test considering 0.5 as the expected probability of random category assignment.

Genome-wide Search for High RIIC Locations and Synteny

In order to predict regions of high curvature and reduce background signal generated by isolated high IC peaks, R language scripts were written to search every chromosome for 600 bp regions with a RIIC greater than the 85th percentile value for that chromosome. The number of regions of high RIIC was

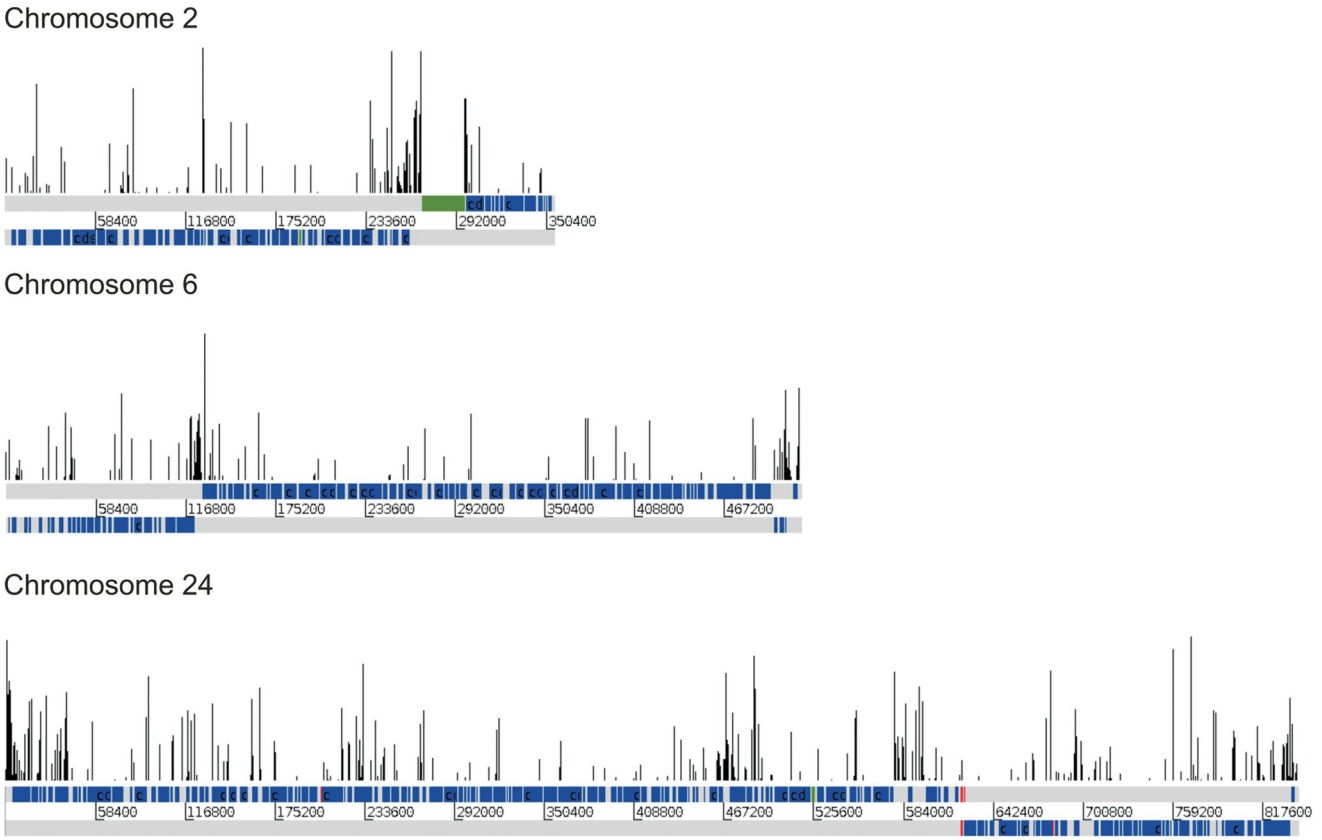


Figure 1. Graphical representation of IC peaks on selected *L. major* chromosomes. Bar plots of IC positions with an IC value greater than 9 degrees per helical turn. Both DNA strands are depicted in grey below bar plots, overlaid with CDS features shown in blue. Features labeled as ncRNA, snRNA or snoRNAs are shown in green. tRNAs are shown in red.
doi:10.1371/journal.pone.0063068.g001

computed and their location compared to the reported for TSS markers. A contingency table was built, using experimental data as the positive reference, and the Matthews correlation coefficient was calculated to determine the plausibility of using the RIIC score to predict the presence of TSS markers.

Syntenic regions were evaluated with the Artemis Comparison Tool (ACT) tool [42]. Chromosome-wide alignments were performed with BLASTN [43].

Results

Intrinsic Curvature Distribution in the *L. major* Genome

In order to characterize the secondary structure of the *Leishmania* genomic sequences, an analysis of the intrinsic curvature distribution was carried out using the bend.it algorithm. For comparison purposes, the profile obtained for a similarly sized fragment of the human chromosome 1 and the *E. coli* genome were included. All sequences assayed showed a non-symmetrical random distribution of IC (Supplementary Figure 2), a result consistent with reports for other organisms [44]. Remarkably, a clear shift towards lower values of IC was observed in *L. major*. Although less evident, a similar shift was also observed for the two other trypanosomatids, *T. brucei* and *T. cruzi*. Very similar and indistinguishable profiles were observed for all the different *Leishmania* species with available genomes (*L. major*, *L. infantum*, *L. braziliensis*, *L. mexicana*) (data not shown), probably due to their significant sequence homology [45]. In comparison to other organisms, ranging from prokaryotes to humans, including trypanosomes, the *Leishmania* IC profile showed fewer regions with high curvature and a sharper peak corresponding to a higher density of regions with lower curvature. Because of this peculiarity, we focused on the analysis of curvature distribution in *Leishmania*.

No significant differences in IC values were observed among the *L. major* chromosomes, with medians ranging from 2.37 (Chr 1) to 2.78 (Chr 36) degrees per helical turn. A similar profile was observed also for *L. infantum*, *L. braziliensis* and *L. mexicana*. Interestingly, a slight increase of the IC medians accompanying chromosome lengths could be noted in all the *Leishmania* species analyzed (Supplementary Figure 3). Indeed, when only the number of peaks with high IC (≥9 degrees/hel. turn) was plotted, a non-linear increase was observed with augmenting chromosome length. Accordingly, the average distance between high IC peaks (density of high peaks) changes from a peak every 450 bp for the smaller chromosomes to a peak every 150 bp for the larger ones (Supplementary Figure 4). In contrast to various organisms where the mean values of curvature were previously shown not to be related to the G+C genome content [44], we found the IC medians to correlate inversely with G+C content within the Tritryp genomes (Supplementary Table 1) and *Leishmania* chromosomes (Supplementary Table 2), ($R^2 = 0.8941$).

To further evaluate a potential functional role for sequence-dependent DNA curvature in specific chromosomal locations, we analyzed the location and context of peaks of high intrinsic curvature along each *L. major* chromosome. Representative profiles of some chromosomes are shown in Figure 1 (see Supplementary Figure 5 for IC profiles of all *L. major* chromosomes). To reduce the

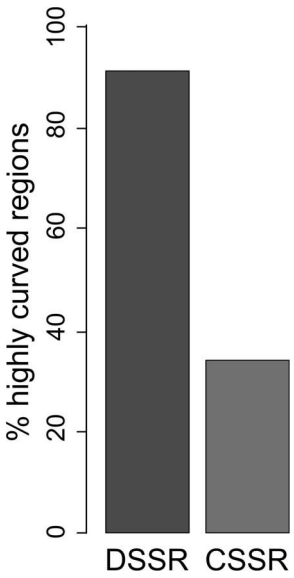


Figure 2. Regional Integrated Intrinsic Curvature analysis in *L. major*. Bar plot showing the percentage of regions with significant RIIC score ($p<0.05$), indicated as highly curved, for divergent (DSSR) and convergent (CSSR) strand switch regions.
doi:10.1371/journal.pone.0063068.g002

background, only IC peaks above 9 degrees per helical turn were considered. For short chromosomes, such as Chr 2 or Chr 6, a high density and/or intensity of IC regions can be observed close to the SSRs. At such regions, high AT content [37] and poly-dinucleotide abundance [38] have been described. Using a similar approach and the bend.it algorithm, Tosato *et al.* [37] reported the presence of peaks of high DNA curvature at SSRs after the analysis of the available chromosome data of *L. major* at the time (complete sequences of Chr 1 and 3 and partial sequences of Chrs 4, 19 and 21). While we make the same observation genome-wide, we could not assert a specific correlation between high intrinsic curvature and SSR due to the high noise levels (see for example Chr24). In addition, the absence of curvature enrichment at the convergent SSRs can be noted (see for example Chr24). These results prompted us to search for more stringent conditions or approaches for the analysis of IC profiles.

Regional Integrated Intrinsic Curvature: a New Tool for Curvature Analysis in *L. major* Chromosomes

Previous studies aiming at detecting conformational signals associated with gene regulation have generally restricted analyses to the occurrence of IC peaks with values above 15 degrees per helical turn for defined short segments with length ranging from 31 to 51 bp [44]. We were not able to make functional associations for regions characterized by high IC peaks in the *L. major* genome. In an effort to use a more sensitive tool, we developed an alternative approach based on the analysis of the area for a selected region in the curvature graph. This approach takes into account the fact that conformational signals likely derive from a combination of the effects of both frequency and intensity of IC in a given region. The scoring function was named RIIC for regional integrated intrinsic curvature.

In order to test the ability of RIIC scores to distinguish functional regions, we applied a statistical test comparing each of 98 selected regions (corresponding to 58 DSSRs and 40 CSSRs) to 5,000 control genomic regions of the same length (see Materials and Methods). A clear association of RIIC scores with divergent

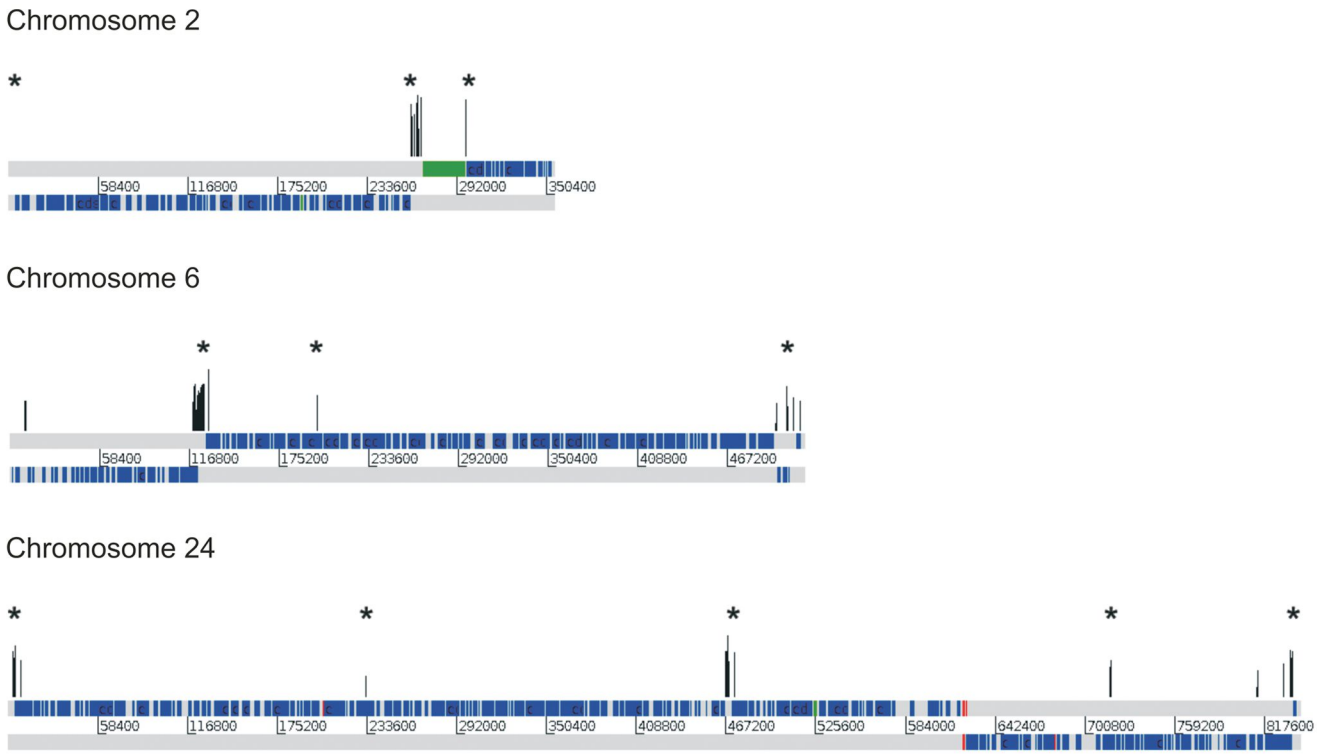


Figure 3. Graphical representation of IC for regions with high RIIC score for three *L. major* chromosomes. The graphs are the same as figure 1. IC for regions with high RIIC score are represented on top. Asterisks mark sites defined as sites associated with acetylated H3 histone [18].
doi:10.1371/journal.pone.0063068.g003

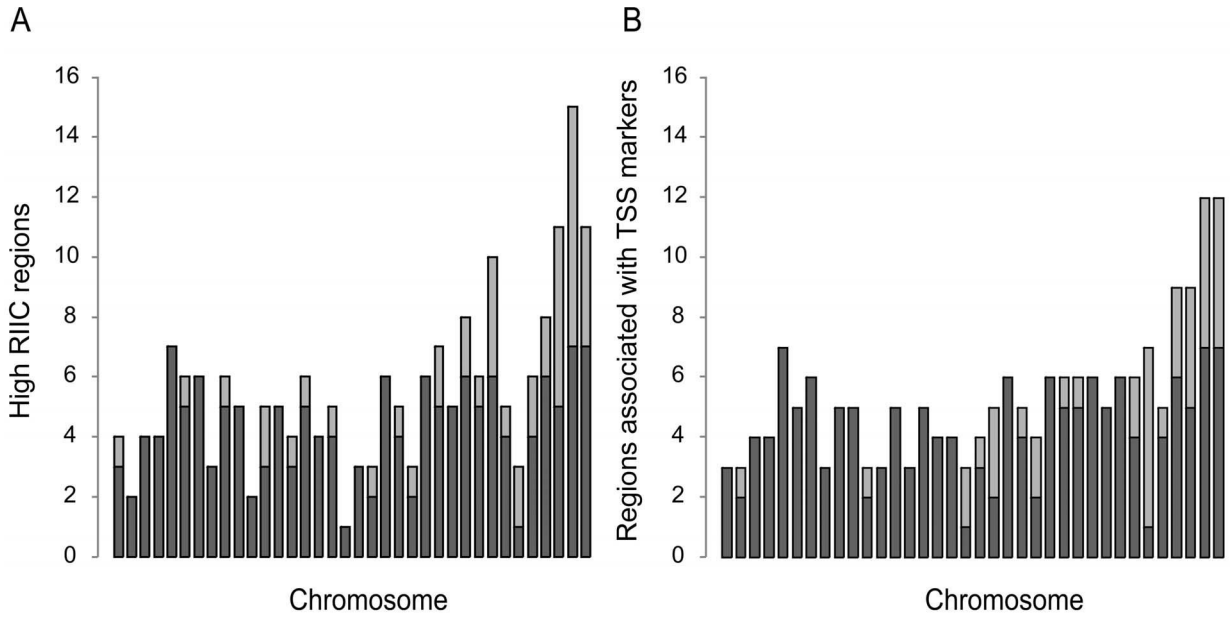


Figure 4. Comparative analysis of high RIIC regions *vs.* reported putative TSS for each *L. major* chromosome. A. The total number of high RIIC regions for each *L. major* chromosome, depicted in ascending order from left to right, is represented. Dark grey indicates high RIIC scoring regions that overlap with regions associated with TSS markers reported by Thomas *et al.* [18]. **B**. The total number of regions associated with TSS markers obtained in Thomas, *et al*. [18] for each *L. major* chromosome, depicted in ascending order from left to right, is represented. Dark grey indicates the number of regions associated with TSS markers that overlap high RIIC regions.
doi:10.1371/journal.pone.0063068.g004

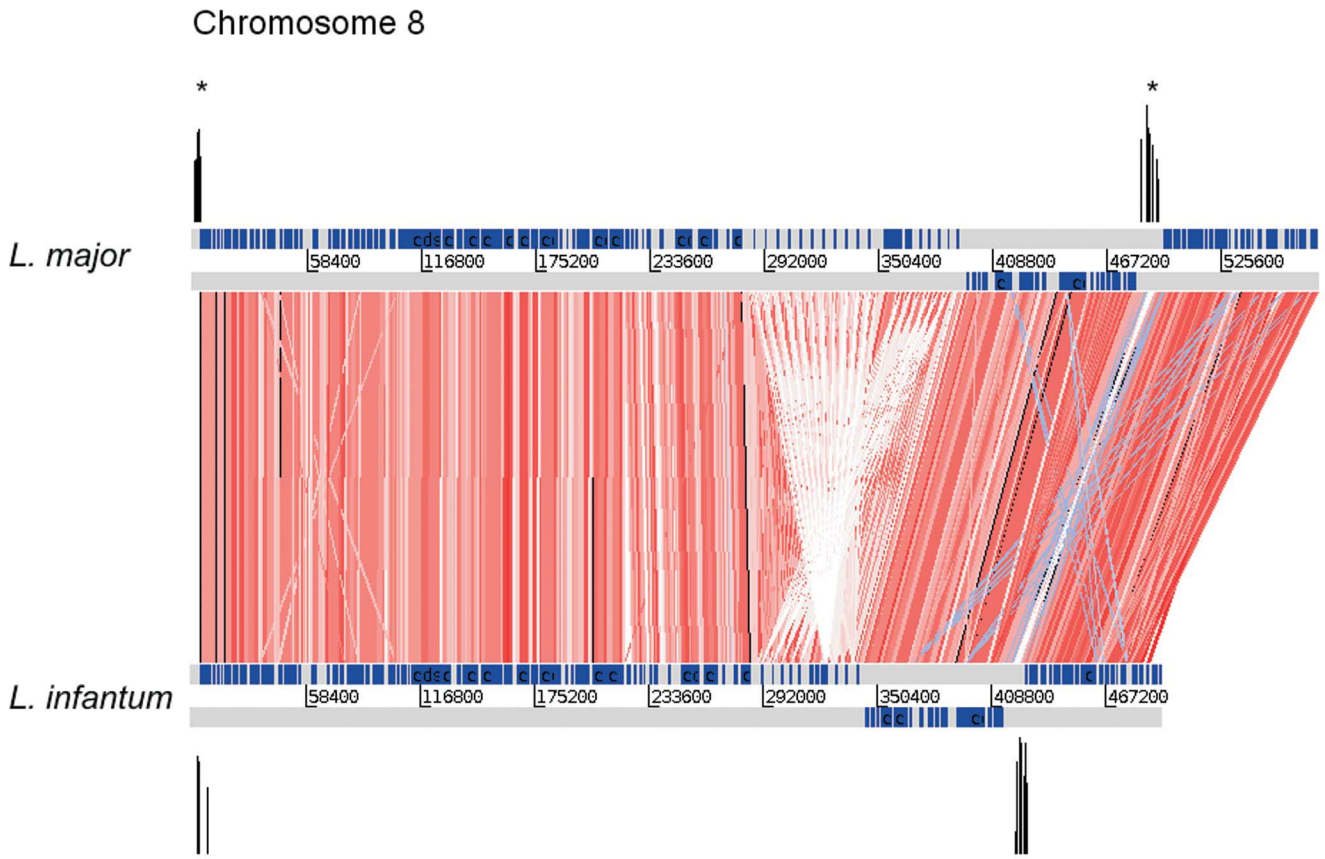


Figure 5. Location conservation of high RIIC scoring regions in *L. major* and *L. infantum* chromosome 8. The graphs are the same as figure 1. Blast HSPs longer than 100 bp and with at least 80% similarity are displayed in red scale and blue lines represent inversions.
doi:10.1371/journal.pone.0063068.g005

SSRs was found (Figure 2). Indeed, 93% of the DSSRs were found to present significantly high RIIC scores compared to random sequences ($p<0.05$), contrasting with only 33% of the CSSRs. A binomial exact test was performed to evaluate the significance of this observation, confirming that the method could clearly differentiate the DSSR as highly curved regions ($p<10^{-12}$). Divergent SSRs have been proposed as sites for transcription initiation in kinetoplastids [16,17]. This is consistent with some early work devoted to the global analysis of the association of intrinsic curvature with transcription initiation where *E. coli* promoter regions were shown to be more curved than coding sequences. Sequence-dependent DNA curvature is known to play an important role in the transcription initiation of many specific genes both in prokaryotes and eukaryotes [46]. It is therefore plausible to propose that *Leishmania* conformational curvature is associated with transcription initiation.

Location of Peaks of High RIIC Score within *L. major* Chromosomes

Since we were able to detect high RIIC scores in *L. major* divergent SSR and considering that distinguishable curvatures may constitute a signal involved in transcription initiation, we investigated the genome-wide occurrence of high RIIC scores. For that purpose, chromosomes were scanned for regions of high RIIC. Regions of 600 bp with RIIC scores higher than the 85th percentile for every particular chromosome were mapped. This approach enabled the detection of intra-DGC regions that scored as high as the distinctive DSSRs. Representative profiles of some chromosomes are shown in Figure 3 (See Supplementary Figure 6 for high RIIC profiles of all *L. major* chromosomes). Interestingly, high RIIC regions are associated with markers of active transcription initiation (acetylated H3 histone, TRF4 and SNAP50) described in [18] (indicated as asterisks in Figure 3). These three markers were reported to co-localize in the *L. major* genome with only a few exceptions. For each chromosome, a summary of the association of the *L. major* genome high RIIC regions with the location of transcription initiation markers reported by Thomas *et al.* [18] is shown in Figure 4. We identified a total of 200 regions in the *L. major* genome with high RIIC, 155 (78%) of which coincide with the location of transcription initiation markers. Among those, 148 regions co-localize with acetylated H3 histone, five co-localize only with the TF4 marker and the other two are upstream ncRNA (Figure 4A). Conversely, out of a total of 191 regions associated with TSS markers, 153 (80%) had a high RIIC score (Figure 4B). A Matthews correlation coefficient of 0.78 supports the specific association between regions of high RIIC and putative TSS. Further work would be necessary to assess if the regions characterized by high RIIC score and currently not associated with TSS markers are actually involved in *Leishmania* transcription initiation.

Location Conservation of High RIIC Scores between *L. major* and *L. infantum*

The high association of DSSRs with curvature region is well conserved when *L. major* and *L. infantum* are compared. The Chr 8 profile is shown in Figure 5 (See Supplementary Figure 7 for representations of all chromosomes). Two regions of high RIIC score associated with conserved DSSRs are present in both *L. major* and *L. infantum* homolog chromosomes.

The existence of conserved physical co-localization of genes among *Leishmanias* [45] and also in the Tritryps [47] has been reported. However no remarkable intergenic sequence conservation has been observed. Interestingly, in many cases the location of high RIIC score regions is conserved in *L. major* and *L. infantum* in spite of the absence of sequence conservation (see for example the DSSR in Figure 5). The extension of synteny to high scoring RIIC regions, here presented, strengthens their functional significance. Globally, we found that from the 200 regions of high RIIC score identified in *L. major*, 166 are conserved in *L. infantum* and 145 of them coincide with regions associated to TSS markers (87%). These results clearly suggest that the conformational signal implication in transcription initiation observed in *L. major* can be generalized to the other *Leishmania* species. For the other 21 regions with high RIIC score, which are conserved in both species, no data about association with transcription initiation or any other functional role has yet been reported. The functional relevance of this finding is worth to be further investigated.

Considering that the analysis of the conservation of the location of high RIIC scoring regions may contribute to the understanding of the role of DNA intrinsic conformational signals in transcription initiation in *Leishmania*, a detailed comparative analysis was performed between *L. major* and *L. infantum*. The search for regions of high RIIC score in *L. major* allowed the identification of 102 out of a total of the 114 DSSRs associated to the TSS markers described by Thomas [18], representing an 89% coincidence (Figure 6A). Considering the addition of regions of high RIIC score from the *L. infantum* data, the number of regions associated to the TSS markers that match with high RIIC increases to 110 (96%)(Figure 6B). An improvement is also observed when considering regions associated with TSS markers inside DGCs (20 for *L. major* and 26 for *L. major* plus *L. infantum* out of 37 associated to TSS markers in [18]) and of non coding RNAs (10 for *L. major* and 13 for *L. major* plus *L. infantum* out of 16 associated to TSS markers in [18]).

Discussion

Following the completion of the genome sequencing of the Tritryps (the three related trypanosomatid parasites: *T. brucei, T. cruzi*, and *L. major*) [47,48,49], genome wide approaches in *L. major* have been mainly aimed at the characterization of global gene and protein expression profiles during development and in response to drugs, as well as protein subcellular localization, and host-parasite interactions [18,50,51,52,53,54,55,56,57,58,59]. While nucleic acid signals acting on different steps of the gene expression process have been described, search efforts have focused on primary sequence signals and no elements involved in transcription initiation of protein coding genes have been found [2]. Nonetheless, the role of the nucleic acid conformation in molecular signaling has been largely recognized in prokaryotic and eukaryotic organisms.

The report of high IC at SSRs of some chromosomes in *L. major* over a decade ago, led to the speculation of their potential involvement at transcription or replication boundaries [37]. Using a similar approach and the bend.it algorithm, we extended the IC analysis to the entire genome. We found that *Leishmania* genomes characteristically present a high density of regions of low IC and relatively few regions of high IC. This characteristic markedly distinguishes the genome of these parasites from both the human chromosome 1 and the *E. coli* genome examined using the same tool. This distinctiveness is also evident even considering the two other Tritryp organisms. The segregation of *Leishmania* canonical behavior is not restricted to the IC density profile. *L. major* base skew pattern also differ from bacteria and even from those of *T. brucei* and *T. cruzi* [60,61]. Though the absence of relation between G+C content and mean IC genome values has been reported [44], we found an inverse correlation within Leishmania chromosomes.

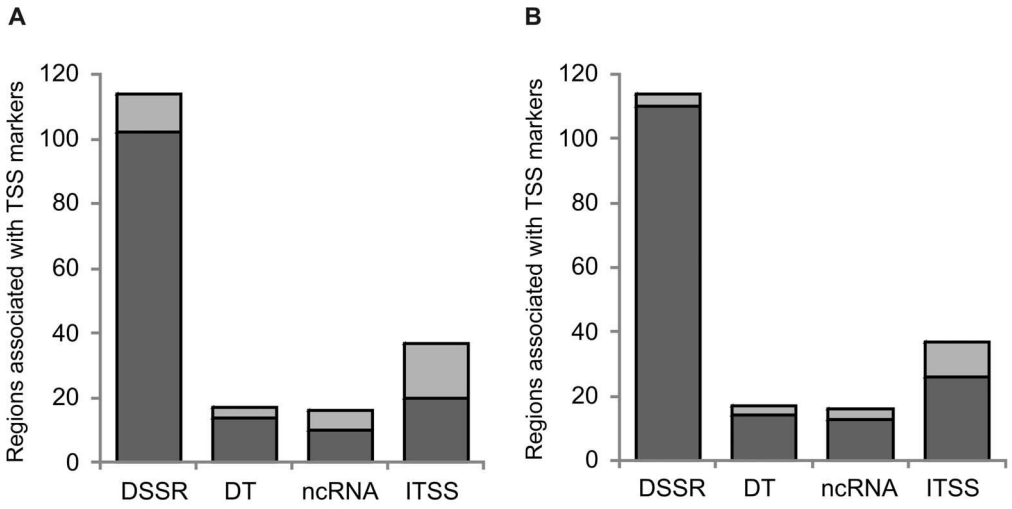


Figure 6. Comparative analysis of high RIIC regions *vs.* reported putative TSS in *Leishmania*. A. The total number of different type of regions (D SS: Divergent Strand Switch; D T: divergent telomeres; ncRNA: associated with ncRNA transcription; I TSS: internal to polycistrons TSS) associated with TSS markers obtained in Thomas, *et. al.* [18] in the *L. major* genome is represented. Dark grey indicates the number of regions associated with TSS markers that overlap with high RIIC regions. **B**. Same as in A displaying in dark grey the number of regions associated with TSS markers that overlap with high RIIC regions that are present either in *L. major* or *L. infantum*.
doi:10.1371/journal.pone.0063068.g006

Higher A+T content could increase the occurrence of A tracts that may induce DNA bends [62]. However, different sequential base orders, with variable G+C content, can also provoke DNA curvature. It has been recognized that SSRs in *L. major* are A+T rich [37], however we observe a clear difference of IC between convergent and divergent SSRs, showing that A+T content is not the only determinant of the presence of highly curved regions in DSSRs.

Taking into account the peculiar high abundance of low IC regions, we hypothesized the existence of conformational signals derived from the net effect of IC accumulation. In order to test our hypothesis, we developed a new approach using an integrative analysis of IC within a defined length region, allowing us to examine, for the first time, regions presenting a high integrated intrinsic curvature value. This approach revealed that high RIIC score regions are associated with DSSRs as well as to genome positions associated to TSS markers in *L. major*. These results support the existence of conformational signals involved in the definition of the transcription start points in *Leishmania*. It is worthwhile mentioning that a few exceptions were observed. Conversely, we also detected few regions characterized by high RIIC scores that have not been associated to TSS markers. Such regions may play a role in transcription initiation and therefore constitute interesting candidates for further investigation. Indeed, the recruitment of TSS markers to different regions may vary during cell cycle, following nutrient or heat shock and other challenges to the cell, establishing different DGCs and messenger abundance [18].

The locations of the regions with high RIIC scores appeared to be as well conserved as the gene synteny observed between *L. major* and *L. infantum*. The relevance of the conformational signal location is further outlined by the fact that this phenomenon is observed independently of sequence conservation.

It would be interesting to investigate the molecular mechanism triggered by the curvature conformation signals. Steps such as facilitated access to transcription initiation machinery directly or through accessory binding of proteins or enhanced melting of the DNA strands to assist active complex formation, may be involved. Bents, as well as other alternative DNA structures, have been involved in the regulation of transcription initiation both in prokaryotes and eukaryotes. Intrinsically bent DNA may specifically influence either chromatin folding, the binding of factors involved in basal transcription initiation and/or regulatory factors that interact with the transcription machinery. In eukaryotes, DNA curvature has been proposed as a primary nucleosome positioning signal [63,64], and low nucleosome occupancy is considered a significant feature for the binding of transcription factors [65]. In addition, it has been reported that the transcription initiation of RNA polymerase II, whose localization depends on primary sequence signals, may be facilitated by the presence and orientation of curved DNA relative to the promoter [66]. Furthermore, for transcription initiation of RNA polymerase I, that is characterized by the absence of primary sequence signals, conserved conformational signals surrounding the transcription start point have been described in eukaryotes [67].

In addition, it would be interesting to investigate whether the association of high RIIC scores with TSS markers is a peculiarity of *Leishmania* genome or eventually and/or partially conserved in related organisms. The characterization of IC genome location for the two other TriTryps, *T. cruzi* and *T. brucei*, is in progress. Nevertheless, this first genome-wide analysis allowed us to clearly identify distinct regions of intrinsic DNA curvature and to associate them with a biological function in *Leishmania*, strongly linking DNA conformational signals with transcription initiation.

Supporting Information

Figure S1 Regional Integrated Intrinsic Curvature analysis in *L. major*. The Riemann sum for the curvature plot was calculated for a SSR spanning 6176 bp in *L. major* Chr22 (from 605990 to 612166) is shown as a vertical dotted line. This score is compared to the density function representing the population of RIIC scores for equal-length regions in the genome) (solid line). In this case, the difference is statistically significant ($p<0.05$).
(PDF)

Figure S2 Genome wide curvature distribution for the Tritryps. The DNA intrinsic curvature for the Tritryps genomes *L. major* (−), *T. brucei* (–) and *T. cruzi* (…), an external reference fragment of approx. 7 Mb from human chromosome 1 spanning from 137 Mb to 144 Mb (–) and *E. coli* BW2952 strain (–), were analyzed with the bend.it algorithm, using a 31 bp window and DNaseI+nucleosome positioning data parameters.
(PDF)

Figure S3 Box plots of the intrinsic curvature distribution across *Leishmania* chromosomes. Chromosomes are depicted in ascending order from left to right. Upper panel: ***L. major*** chromosomes 1 to 36. Lower panel: ***L. infantum*** chromosomes 0 to 36; ***L. braziliensis*** chromosomes 0 to 35; ***L. mexicana*** chromosomes 0 to 34 from *L. mexicana*. For lower panels axes are as in the upper panel.
(PDF)

Figure S4 Relationship between peaks of high predicted intrinsic curvature and chromosome length. A. The number of IC peaks greater than 9 degrees per helical turn in each chromosome was plotted against the chromosome length. **B.** The average frequency of IC peaks greater than 9 degrees per helical turn (calculated as the chromosome length divided by the absolute number of peaks) is plotted against chromosome number.
(PDF)

Figure S5 Graphical representation of IC peaks on all *L. major* chromosomes. Bar plots of IC positions with an IC value greater than 9 degrees per helical turn. Both DNA strands are depicted in grey below bar plots, overlaid with CDS features shown in blue. Features labeled as ncRNA, snRNA or snoRNAs are shown in green. tRNAs are shown in red. rRNAs are shown in brown.
(PDF)

Figure S6 Graphical representation of IC for regions with high RIIC-score for all *L. major* chromosomes. The graphs are the same as figure 1. IC for regions with high RIIC score are indicated at the top. Sites associated with acetylated H3 histone [18] are indicated as small vertical lines.
(PDF)

Figure S7 Location conservation of high RIIC scoring regions in *L. major* and *L. infantum* chromosomes. The graphs are the same as figure 1. Blast HSPs longer than 100 bp and with at least 80% similarity are displayed in red scale and blue lines represent inversions.
(PDF)

Table S1 Genome intrinsic curvature in Tritryps.
(PDF)

Table S2 Chromosome intrinsic curvature in *L. major*.
(PDF)

Author Contributions

Conceived and designed the experiments: PS NMES BG. Performed the experiments: PS. Analyzed the data: PS NMES BG. Contributed reagents/materials/analysis tools: DF. Wrote the paper: PS NMES BG.

References

1. Desjeux P (2004) Leishmaniasis: current situation and new perspectives. Comp Immunol Microbiol Infect Dis 27: 305–318.
2. Martinez-Calvillo S, Vizuet-de-Rueda JC, Florencio-Martinez LE, Manning-Cela RG, Figueroa-Angulo EE (2010) Gene expression in trypanosomatid parasites. J Biomed Biotechnol 2010: 525241.
3. Boucher N, McNicoll F, Laverdiere M, Rochette A, Chou MN, et al. (2004) The ribosomal RNA gene promoter and adjacent cis-acting DNA sequences govern plasmid DNA partitioning and stable inheritance in the parasitic protozoan Leishmania. Nucleic Acids Res 32: 2925–2936.
4. Casagrande L, Ruiz JC, Beverley SM, Cruz AK (2005) Identification of a DNA fragment that increases mitotic stability of episomal linear DNAs in Leishmania major. Int J Parasitol 35: 973–980.
5. Ubeda JM, Legare D, Raymond F, Ouameur AA, Boisvert S, et al. (2008) Modulation of gene expression in drug resistant Leishmania is associated with gene amplification, gene deletion and chromosome aneuploidy. Genome Biol 9: R115.
6. Mukherjee A, Langston LD, Ouellette M (2011) Intrachromosomal tandem duplication and repeat expansion during attempts to inactivate the subtelomeric essential gene GSH1 in Leishmania. Nucleic Acids Res 39: 7499–7511.
7. Ghedin E, Bringaud F, Peterson J, Myler P, Berriman M, et al. (2004) Gene synteny and evolution of genome architecture in trypanosomatids. Mol Biochem Parasitol 134: 183–191.
8. Bringaud F, Muller M, Cerqueira GC, Smith M, Rochette A, et al. (2007) Members of a large retroposon family are determinants of post-transcriptional gene expression in Leishmania. PLoS Pathog 3: 1291–1307.
9. Requena JM, Folgueira C, Lopez MC, Thomas MC (2008) The SIDER2 elements, interspersed repeated sequences that populate the Leishmania genomes, constitute subfamilies showing chromosomal proximity relationship. BMC Genomics 9: 263.
10. Smith M, Bringaud F, Papadopoulou B (2009) Organization and evolution of two SIDER retroposon subfamilies and their impact on the Leishmania genome. BMC Genomics 10: 240.
11. de Andrade Stempliuk V, Floeter-Winter LM (2002) Functional domains of the rDNA promoter display a differential recognition in Leishmania. Int J Parasitol 32: 437–447.
12. Das A, Banday M, Bellofatto V (2008) RNA polymerase transcription machinery in trypanosomes. Eukaryot Cell 7: 429–434.
13. Rana T, Misra S, Mittal MK, Farrow AL, Wilson KT, et al. (2011) Mechanism of down-regulation of RNA polymerase III-transcribed non-coding RNA genes in macrophages by Leishmania. J Biol Chem 286: 6614–6626.
14. Agami R, Aly R, Halman S, Shapira M (1994) Functional analysis of cis-acting DNA elements required for expression of the SL RNA gene in the parasitic protozoan Leishmania amazonensis. Nucleic Acids Res 22: 1959–1965.
15. Saito RM, Elgort MG, Campbell DA (1994) A conserved upstream element is essential for transcription of the Leishmania tarentolae mini-exon gene. EMBO J 13: 5460–5469.
16. Martinez-Calvillo S, Nguyen D, Stuart K, Myler PJ (2004) Transcription initiation and termination on Leishmania major chromosome 3. Eukaryot Cell 3: 506–517.
17. Martinez-Calvillo S, Yan S, Nguyen D, Fox M, Stuart K, et al. (2003) Transcription of Leishmania major Friedlin chromosome 1 initiates in both directions within a single region. Mol Cell 11: 1291–1299.
18. Thomas S, Green A, Sturm NR, Campbell DA, Myler PJ (2009) Histone acetylations mark origins of polycistronic transcription in Leishmania major. BMC Genomics 10: 152.
19. Puechberty J, Blaineau C, Meghamla S, Crobu L, Pages M, et al. (2007) Compared genomics of the strand switch region of Leishmania chromosome 1 reveal a novel genus-specific gene and conserved structural features and sequence motifs. BMC Genomics 8: 57.
20. Sturm NR, Yu MC, Campbell DA (1999) Transcription termination and 3′-End processing of the spliced leader RNA in kinetoplastids. Mol Cell Biol 19: 1595–1604.
21. Padilla-Mejia NE, Florencio-Martinez LE, Figueroa-Angulo EE, Manning-Cela RG, Hernandez-Rivas R, et al. (2009) Gene organization and sequence analyses of transfer RNA genes in Trypanosomatid parasites. BMC Genomics 10: 232.
22. Worthey EA, Martinez-Calvillo S, Schnaufer A, Aggarwal G, Cawthra J, et al. (2003) Leishmania major chromosome 3 contains two long convergent polycistronic gene clusters separated by a tRNA gene. Nucleic Acids Res 31: 4201–4210.
23. van Luenen HG, Farris C, Jan S, Genest PA, Tripathi P, et al. (2012) Glucosylated hydroxymethyluracil, DNA base J, prevents transcriptional readthrough in Leishmania. Cell 150: 909–921.
24. Requena JM, Soto M, Quijada L, Carrillo G, Alonso C (1997) A region containing repeated elements is associated with transcriptional termination of Leishmania infantum ribosomal RNA genes. Mol Biochem Parasitol 84: 101–110.
25. Martinez-Calvillo S, Sunkin SM, Yan S, Fox M, Stuart K, et al. (2001) Genomic organization and functional characterization of the Leishmania major Friedlin ribosomal RNA gene locus. Mol Biochem Parasitol 116: 147–157.
26. Liang XH, Haritan A, Uliel S, Michaeli S (2003) trans and cis splicing in trypanosomatids: mechanism, factors, and regulation. Eukaryot Cell 2: 830–840.
27. LeBowitz JH, Smith HQ, Rusche L, Beverley SM (1993) Coupling of poly(A) site selection and trans-splicing in Leishmania. Genes Dev 7: 996–1007.
28. Haile S, Papadopoulou B (2007) Developmental regulation of gene expression in trypanosomatid parasitic protozoa. Curr Opin Microbiol 10: 569–577.

29. Clayton C, Shapira M (2007) Post-transcriptional regulation of gene expression in trypanosomes and leishmanias. Mol Biochem Parasitol 156: 93–101.
30. Boucher N, Wu Y, Dumas C, Dube M, Sereno D, et al. (2002) A common mechanism of stage-regulated gene expression in Leishmania mediated by a conserved 3′-untranslated region element. J Biol Chem 277: 19511–19520.
31. McNicoll F, Muller M, Cloutier S, Boilard N, Rochette A, et al. (2005) Distinct 3′-untranslated region elements regulate stage-specific mRNA accumulation and translation in Leishmania. J Biol Chem 280: 35238–35246.
32. Quijada L, Soto M, Alonso C, Requena JM (2000) Identification of a putative regulatory element in the 3′-untranslated region that controls expression of HSP70 in Leishmania infantum. Mol Biochem Parasitol 110: 79–91.
33. Zilka A, Garlapati S, Dahan E, Yaolsky V, Shapira M (2001) Developmental regulation of heat shock protein 83 in Leishmania. 3′ processing and mRNA stability control transcript abundance, and translation id directed by a determinant in the 3′-untranslated region. J Biol Chem 276: 47922–47929.
34. Potaman V, Sinden R (2005) DNA: Alternative Conformations and Biology. In: Ohyama T, editor. DNA conformation and transcription. Georgetown, Tex. New York, NY.: Landes Bioscience. 3–17.
35. Ntambi JM, Marini JC, Bangs JD, Hajduk SL, Jimenez HE, et al. (1984) Presence of a bent helix in fragments of kinetoplast DNA minicircles from several trypanosomatid species. Mol Biochem Parasitol 12: 273–286.
36. Wu HM, Crothers DM (1984) The locus of sequence-directed and protein-induced DNA bending. Nature 308: 509–513.
37. Tosato V, Ciarloni L, Ivens AC, Rajandream MA, Barrell BG, et al. (2001) Secondary DNA structure analysis of the coding strand switch regions of five Leishmania major Friedlin chromosomes. Curr Genet 40: 186–194.
38. Duhagon MA, Smircich P, Forteza D, Naya H, Williams N, et al. (2011) Comparative genomic analysis of dinucleotide repeats in Tritryps. Gene 487: 29–37.
39. Vlahovicek K, Kajan L, Pongor S (2003) DNA analysis servers: plot.it, bend.it, model.it and IS. Nucleic Acids Res 31: 3686–3687.
40. R_Development_Core_Team (2011) R: A Language and Environment for Statistical Computing; Computing RFfS, editor. Vienna, Austria.
41. Anton H (1999) Calculus: a new horizon. New York: Wiley. xxxii, 1130, 1129, 1113 p.
42. Carver TJ, Rutherford KM, Berriman M, Rajandream MA, Barrell BG, et al. (2005) ACT: the Artemis Comparison Tool. Bioinformatics 21: 3422–3423.
43. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215: 403–410.
44. Gabrielian A, Vlahovicek K, Pongor S (1997) Distribution of sequence-dependent curvature in genomic DNA sequences. FEBS Lett 406: 69–74.
45. Peacock CS, Seeger K, Harris D, Murphy L, Ruiz JC, et al. (2007) Comparative genomic analysis of three Leishmania species that cause diverse human disease. Nat Genet 39: 839–847.
46. Ohyama T (2005) DNA conformation and transcription. Georgetown, Tex. New York, NY.: Landes Bioscience; Springer Science Business Media. 211 p.
47. El-Sayed NM, Myler PJ, Blandin G, Berriman M, Crabtree J, et al. (2005) Comparative genomics of trypanosomatid parasitic protozoa. Science 309: 404–409.
48. Ivens AC, Peacock CS, Worthey EA, Murphy L, Aggarwal G, et al. (2005) The genome of the kinetoplastid parasite, Leishmania major. Science 309: 436–442.
49. Berriman M, Ghedin E, Hertz-Fowler C, Blandin G, Renauld H, et al. (2005) The genome of the African trypanosome Trypanosoma brucei. Science 309: 416–422.
50. Cuervo P, Domont GB, De Jesus JB (2010) Proteomics of trypanosomatids of human medical importance. J Proteomics 73: 845–867.
51. Rochette A, Raymond F, Corbeil J, Ouellette M, Papadopoulou B (2009) Whole-genome comparative RNA expression profiling of axenic and intracellular amastigote forms of Leishmania infantum. Mol Biochem Parasitol 165: 32–47.
52. Rochette A, Raymond F, Ubeda JM, Smith M, Messier N, et al. (2008) Genome-wide gene expression profiling analysis of Leishmania major and Leishmania infantum developmental stages reveals substantial differences between the two species. BMC Genomics 9: 255.
53. Adaui V, Castillo D, Zimic M, Gutierrez A, Decuypere S, et al. (2011) Comparative gene expression analysis throughout the life cycle of Leishmania braziliensis: diversity of expression profiles among clinical isolates. PLoS Negl Trop Dis 5: e1021.
54. Alcolea PJ, Alonso A, Larraga V (2011) Proteome profiling of Leishmania infantum promastigotes. J Eukaryot Microbiol 58: 352–358.
55. Alcolea PJ, Alonso A, Sanchez-Gorostiaga A, Moreno-Paz M, Gomez MJ, et al. (2009) Genome-wide analysis reveals increased levels of transcripts related with infectivity in peanut lectin non-agglutinated promastigotes of Leishmania infantum. Genomics 93: 551–564.
56. Ghedira K, Hornischer K, Konovalova T, Jenhani AZ, Benkahla A, et al. (2011) Identification of key mechanisms controlling gene expression in Leishmania infected macrophages using genome-wide promoter analysis. Infect Genet Evol 11: 769–777.
57. Depledge DP, Evans KJ, Ivens AC, Aziz N, Maroof A, et al. (2009) Comparative expression profiling of Leishmania: modulation in gene expression between species and in different host genetic backgrounds. PLoS Negl Trop Dis 3: e476.
58. Tsigankov P, Gherardini PF, Helmer-Citterich M, Zilberstein D (2012) What has proteomics taught us about Leishmania development? Parasitology 139: 1146–1157.
59. Choi J, El-Sayed NM (2012) Functional genomics of trypanosomatids. Parasite Immunol 34: 72–79.
60. McDonagh PD, Myler PJ, Stuart K (2000) The unusual gene organization of Leishmania major chromosome 1 may reflect novel transcription processes. Nucleic Acids Res 28: 2800–2803.
61. Nilsson D, Andersson B (2005) Strand asymmetry patterns in trypanosomatid parasites. Exp Parasitol 109: 143–149.
62. Haran TE, Mohanty U (2009) The unique structure of A-tracts and intrinsic DNA bending. Q Rev Biophys 42: 41–81.
63. Kiyama R, Trifonov EN (2002) What positions nucleosomes?–A model. FEBS Lett 523: 7–11.
64. Liu H, Duan X, Yu S, Sun X (2011) Analysis of nucleosome positioning determined by DNA helix curvature in the human genome. BMC Genomics 12: 72.
65. Daenen F, van Roy F, De Bleser PJ (2008) Low nucleosome occupancy is encoded around functional human transcription factor binding sites. BMC Genomics 9: 332.
66. Gimenes F, Takeda KI, Fiorini A, Gouveia FS, Fernandez MA (2008) Intrinsically bent DNA in replication origins and gene promoters. Genet Mol Res 7: 549–558.
67. Marilley M (2000) Structure-function relationships in replication origins of the yeast Saccharomyces cerevisiae: higher-order structural organization of DNA in regions flanking the ARS consensus sequence. Mol Gen Genet 263: 854–866.

Sequence-dependent DNA curvature profiles in *Trypanosomes*

Pablo Smircich[1,2], Najib M. El-Sayed[3] and Beatriz Garat[1]*

[1]Laboratorio de Interacciones Moleculares, Facultad de Ciencias, 11400 Montevideo, Uruguay
[2]Departamento de Genética, Facultad de Medicina, 11800 Montevideo, Uruguay
[3]Department of Cell Biology and Molecular Genetics and Center for Bioinformatics and Computational Biology, University of Maryland College Park, MD 20742, USA

* To whom correspondence should be addressed. Tel: +598 2 525 86 18 ext (7) 237, Fax: +598 2 525 86, Email address: bgarat@fcien.edu.uy

Abstract

Taking into account that intrinsic DNA curvature may constitute a signal for biological processes, we undertook a genome wide search of sequence-dependent curvature (SDC) in the trypanosomatid models: *Trypanosoma cruzi* and *Trypanosoma brucei*. Using the region integrated intrinsic curvature scoring (RIIC) previously developed, a non-random distribution of SDC was observed in both cases. Concordantly with the pattern we have previously described in *Leishmania*, the association of high RIIC with regions related to the transcription process both in *T. cruzi* and *T. brucei* was observed to different extents. In addition, in *T. brucei*, there is a remarkable concentration of regions with high RIIC in the subtelomeric regions of the chromosomes, where the species-specific genes for the variable surface glycoproteins (VSG) are located. These findings, while establishing particularities within trypanosomes, underscore the relevance of indirect DNA readout in these ancient eukaryotes.

Introduction

Trypanosoma cruzi and *Trypanosoma brucei* (family Trypanosomatidae, order Kinetoplastida) are flagellate protozoan parasites that cause in humans Chagas´ disease and African trypanosomiasis or sleeping sickness, respectively. These pathogens infect the poorest rural populations in developing countries in tropical and subtropical regions yielding millions of human deaths. Due to economical migrations, these diseases are now spreading world-wide [1]

In spite of the fact that trypanosomes share many characteristics, such as subcellular structures, they greatly differ on their life cycle. *T. cruzi* has a complex life cycle alternating between two extra-cellular forms in the triatomine insect: the replicative epimastigote (E) and the infective metacyclic trypomastigote (MT) and two forms in the mammalian host: the amastigote (A) which is replicative and intracellular, thus hiding from the immune system, and the non-replicative and infective bloodstream trypomastigote (T). On the other hand, *T. brucei* is exclusively extra-cellular alternating between the procyclic form (PF) in the arthropod vector (tsetse flies) and the bloodstream form (BF) in the mammalian hosts [2]. To evade the immune system, *T. brucei* forms a dense coat of variant surface glycoproteins (VSG) that are expressed from 15 telomeric expression sites, one at a time, switching from a repertoire of up to 2,000 VSG genes located in subtelomeric regions [3,4]. The high expression of a single VSG gene needed to shield the parasite is accounted by the RNA polymerase I (RNAPI) due to its high transcription initiation rate [5]. This constitutes one of the molecular characteristics that distinguish this parasite from other eukaryotic organisms, where RNAPI is unable to synthesize mRNA.

In addition to the above mentioned peculiarity, the trypanosomes have other unique features regarding gene organization and expression. Among others, it is worth to mention the organization in divergent gene clusters (DGCs), the constitutive transcription as large polycistronic gene units (PGUs), the amplification of genes in response to environmental stimuli, the mRNA *trans*-splicing, the extensive editing of mitochondrial transcripts [6,7] and the dependence on post transcriptional regulation mechanisms to coordinate gene expression [2,8]. As usual the trypanosomes´ chromosomes contain telomeric repeats at their ends. Subtelomeric regions are composed of variable repetitive elements and contain genes involved in antigenic variation in *T. brucei* or genes encoding surface antigens in *T. cruz*i, which does not undergo antigenic variation [9,10]. The hyper-modified

base J (β-D-Glucopyranosyloxymethyluracil) that is predominantly present in repetitive DNA sequences in telomeres and subtelomeres in trypanosomatids, has been more recently localized in RNA polymerase II transcription initiation and termination sites [11]. Canonical signals for RNA polymerase II promoters have only been described for the genes encoding the spliced leader (SL) a small RNA that is added by *trans*-splicing to all the protein coding genes [12]. However, transcription starting sites (TSS) [13,14] and histone variants implicated in the initiation process [15,16] have been described at the strand switch regions that separate the heads of the PGUs, named divergent SSRs (DSSRs). On the contrary, the SSRs that separate the tails of the PGUs, named convergent SSRs (CSSRs), have been shown to preferentially contain sites of transcription termination, usually presenting a polymerase III transcribed tRNA gene [14,17]. A bias in poly-dinucleotides abundance has also been reported for those regions [18]. A link between transcription and DNA replication has been recently described in *T. brucei* [19]. The observation that SSRs are responsible transcription initiation, remarks the importance of SSRs in DNA metabolism processes. In spite of the important insights achieved (recently reviewed in [20]) the molecular signals underlying these processes remain mostly under investigation [21,22].

Intrinsically curved DNA structures are often found around origin of DNA replication, DNA recombination loci and in regions that regulate transcription. In eukaryotes, this feature is common to the promoters of genes transcribed by RNA polymerase I, but there are also many reports that have described the occurrence of curved DNA at promoters of RNA polymerase II [23]. Accordingly, we have found a statistical significant association of transcription starting sites to regions of high regional integrated intrinsic curvature – RIIC- score in *Leishmanias* [24]. Though differences between trypanosomes and *Leishmania*, due to different base composition content and derived intrinsic curvature may exist, we wondered whether high RIIC regions could also be associated to a biological phenomenon in these ancient eukaryotes.

Here we present the genome wide search of sequence-dependent curvature in the 41 *T. cruzi* chromosomes and the 11 *T. brucei* megabase sized chromosomes using the region integrated intrinsic curvature scoring (RIIC) previously developed. A non-random distribution of SDC was observed in both cases. The association to different extents of high RIIC score with regions involved in the transcription process, such as DSSRs in *T. cruzi,* and location of histone variants and base J in *T. brucei*, was also found. In addition, a striking concentration of regions of

high curvature in the subtelomeric regions of the *T. brucei* chromosomes, overlapping regions of silent VSG, became evident. A MEME search for motifs within those regions discovered the presence of a canonical signal for DNA bending consisting on A tracts repeated in phase. Altogether, these findings suggest a link between DNA conformational signals and gene expression in trypanosomes.

Materials and Methods

Data source

The genome data for *T. brucei* strain Treu927and the *T. cruzi* Esmeraldo Like contigs were downloaded from TritrypDB (version 2.1). *T. brucei* chromosome regions were classified as VSG clusters, subtelomeric or core regions as in [19].

Genome-wide Intrinsic Curvature calculation

The bend.it algorithm, kindly provided by Dr. S. Pongor, was used locally to obtain the IC values (degrees per helical turn) for each base on the individual chromosomes which were sliced into 200Kb fragments. The default window size (31bp) and bendability values from nucleosome binding and DNaseI parameters were used. In-house scripts either in Python or in R programming languages were developed to obtain data and to filter and/or further analyze results. For visualization, wiggle (WIG) files were generated for each chromosome.

Genome-wide Region Integrated Instrinsic Curvature

The RIIC score was calculated as the area under the curvature plot using the Riemann sum. To assess if the *T. brucei* subtelomeric and VSG array regions have a significantly higher RIIC score than the rest of the genome, their RIIC score was compared to a density function representing the population of RIIC scores for equal-length regions in the genome as in [24]. A region was considered highly curved if the RIIC score was bigger than the 95% confidence interval for the population. The genome wide search for high RIIC regions was performed as in [24] with minor modifications. Briefly, every chromosome was scanned for 600 bp regions with a RIIC score greater than the 80^{th} percentile value for that chromosome. For the counting of *T. cruzi* SSRs associated with high RIIC, two criteria were established. Namely, a SSR was not considered if it presented internal sequencing gaps and/or was defined by DGCs of less than 6 genes.

Motif Searching

For the analysis of motif-based sequence associated to the regions of high IC the MEME suite was used. The analysis was performed on a fragment of *T. brucei*

chromosome 9 spanning bases from 1 to 319439 (the VSG array region). Sequences surrounding 30bp each high IC peak were selected for motif discovery. The randomly shuffled sequences were also submitted to the MEME suite as control.

Results and Discussion

1. Regions of high curvature are not randomly distributed along trypanosomes´ chromosomes

Since we have previously shown that sequence-dependent DNA curvature may have a biological role in *Leishmania* [24], we investigated its genomic distribution in the related tripanosomatid parasites: *T. cruzi* and *T. brucei*. The *T. cruzi* Esmeraldo like haplotype which consist in 41 chromosomes [10] was selected for the study. For *T. brucei* we focused on the 11 megabase-sized chromosomes [25]. Firstly, an analysis using the bend.it algorithm, as in [26] was carried out. Comparatively, we also analyzed the genomic curvature using the scoring RIIC function previously developed [24]. To reduce background, only peaks whose IC are above 13 degrees per helical turn, for the first approach, or high RIIC regions (as detailed in Materials and Methods), for the second one, were considered. As an example, the distribution of regions of high sequence-dependent curvature using both approaches for *T. cruzi* chromosome 9 and *T. brucei* chromosome 5 are shown in **Figure 1** (See **Supplementary Figures 1** and **2** for all the *T. cruzi* and *T. brucei* chromosomes).

Noisier outputs comparing to the ones observed for *Leishmania* [24] were found, particularly for *T. brucei*. While in *T. cruzi* the high RIIC regions are apparently associated to strand switch regions, this observation is not so plain for *T. brucei*. Nevertheless, it is evident that also in *T. brucei*, regions of high curvature are not randomly distributed. Indeed, independently of the approach used (IC –upper panel- or RIIC –middle panel-) a clear concentration of regions of high curvature can readily be observed at subtelomeric positions independently of the approach used (IC –upper panel- or RIIC –middle panel-). It is worth to note that the *Leishmania* intrinsic curvature profile, compared to other organisms, ranging from prokaryotes to humans, including trypanosomes, showed fewer regions with high curvature and a higher density of regions with lower curvature. So the high background observed could be attributed to the presence of more regions of high curvature leading to higher median IC (3.32 and 3.52 degrees/helical turn for *T. cruzi* and *T. brucei* respectively) compared to *Leishmania* (2.63 degrees/helical turn).

2. Divergent strand switch regions are associated to high RIIC in *T. cruzi* genome

Considering the profile obtained for high RIIC in the *T. cruzi* chromosomes and our previous observation in *Leishmania* which relate high RIIC with DSSR regions, we decided to further analyze if this result also holds in the *T. cruzi* genome.

In order to proceed with an unbiased correlation counting, we needed to set up conditions to reliably define SSRs in the *T. cruzi* genome. Due to the high number of sequencing gaps, currently present in the *T. cruzi* genome sequence data, all the SSRs containing gaps were not included in our analysis. Besides, we decided to focus on large DGCs and arbitrarily limited the definition of countable SSRs to those defined by DGCs containing six or more genes. Using these two criteria, we were able to count 115 SSRs that should be considered only a minimum but reliable number of SSRs in the *T. cruzi* genome. As previously reported [24], we scanned the chromosomes for high RIIC regions (see Materials and Methods). Summarized results are presented in **Figure 2.** A remarkable association of high RIIC scoring regions and DSSRs (72% of the 68 considered DSSRs) was found (Fisher's test $P < 0.0001$, Matthews' correlation coefficient of 0.33). Nonetheless, only 19% of CSSRs (9 out of 47) where associated with high RIIC scoring regions (not significant Fisher's test, Matthews' correlation coefficient of 0.005). So, a clear difference among the two types of SSR could be assessed, suggesting that DNA intrinsic curvature may be involved in transcription initiation also in this parasite. While such DNA structures may have different functions in this process, it is generally considered that they may work helping the binding of the RNA polymerase and facilitating the formation of the open DNA complex.

Taking into account that in *Leishmania* high curvature regions are also associated with internal TSSs, it would be interesting to analyze if such is also the case for *T. cruzi*. Although an enrichment of histone acetylation in the DSSR have been described in *T. cruzi* [15], currently there are no studies of the presence of TSS markers inside the DGCs. This rules out the possibility of assessing if the regions where internal high curvature is observed, also show TSS markers. In this context we can only speculate that a high percentage of the 172 internal high RIIC scoring regions found in *T. cruzi* may also serve as internal TSS. If this consideration holds true, and some of the 172 internal sites are actual markers of TSS regions, then the number of true positives will increase while the number of false positives will decrease producing an overall increase in the Mathews correlation coefficient.

3. Regions of high RIIC score are associated to the presence of base J in *T. brucei*

Probably due to the noisy output, not an obvious association of high RIIC with SSRs, can be perceived in *T. brucei*. Anyway, since we have previously found a clear association of high RIIC scoring regions with transcription initiation in *Leishmania*, a deeper analysis of the high RIIC distribution in *T. brucei* chromosomes was performed.

In spite of the fact that *T. brucei* genome (26.2Mb) is smaller than the *Leishmania´*s one (32.9Mb), the actual number of SSRs is difficult to ascertain because of the existence, as in *T. cruzi*, of putative small DGCs, mainly at the frontiers of the large ones. Nevertheless, the existence of available data of genome-wide location of epigenetic markers associated to transcription starting sites such as modified histones [16], enables the correlation analysis of putative TSS with high RIIC. We found that some of the regions associated to peaks of H4K10ac correspond to high RIIC (**Supplementary Figure 3**). The number of regions associated to this marker and the coincidence with high RIIC is shown in **Figure 3**. The significance of the association (Fisher's test $P < 0.0001$) is in accordance with our previous results in *Leishmania*. However, it is worth to note that the Matthews' correlation coefficient in this case is only 0.2, while in *Leishmania* it reaches 0.78.

Since in *T. brucei*, a role in global transcription by RNA polymerase II as well as at telomeric expression sites involved in antigenic variation has been proposed for the base J [11,27], we compared its location with the high RIIC regions profile (**Supplementary Figure 3**). The number of regions associated to this marker and the coincidence with high RIIC is shown in **Figure 3**. We found that base J location widely overlap with regions of high RIIC (Fisher's test $P < 0.0001$). The association of the presence of this modified base with high RIIC is even better than the one observed for TSS markers (Matthews' correlation coefficient of 0.46).

Though, the noisy curvature genome profile in *T. brucei* may be pointing out to the involvement of these signals in different biological processes, the existence of a link between DNA curvature and transcription process proved to be significant. However, a clear distinction between the pattern described here and the one previously described in *Leishmania* results evident. While in this later organism, the association of high RIIC with transcription initiation was clear, in *T. brucei* a better correlation is obtained with markers of transcription boundaries, that is, both initiation and

termination. Indeed, the enriched presence of base J in most of the regions associated to H4K10ac [11], questions whether the observed correlation of high RIIC with H4K10ac associated regions is actually due to the presence of the modified base. Further studies would be necessary to deeply analyze the recognition signal commonalities and differences that may be involved in the step mechanisms that are shared between this two organism and those that are not.

4. In phase A runs turn up a common motif in the subtelomeric regions in *T. brucei*

We then focused on the impressive concentration of high intrinsic curvature at the subtelomeric regions present in this organism and absent in *T. cruzi*. These regions are characterized by a high content of repetitive sequences and the presence of non active VSG genes and pseudogenes [28]. Manual inspection of the location of these regions revealed that the peaks of high IC do not correspond to -direct or inverted- repetitive DNA (data not shown).

In order to determine if any motif existed in the VSG region that would explain the high concentration of sequence-dependent curvature, a MEME analysis was performed. A pattern of two runs of 4-6 adenine tracts separated by 10bp constitutes the highest scoring motif found by this approach (**Figure 4**). The motif can be considered a common characteristic in subtelomeric regions since this conserved sequence pattern is present in 50 out of the 103 sequences used as query. Though short runs of adenines are ubiquitous along *T. brucei* genome, they are not clearly associated with high IC peaks in the core regions. Besides, not every IC peak corresponds to the presence of runs of in phase adenine tracts.

It has long been known that runs of at least 4 adenines with the 10bp phasing direct DNA bending [29,30,31,32]. The molecular structures of these DNA tracts are unusual and may vary depending on the genome context. Multiple roles for the A-tract curvature have been proposed [33]. Among them, A tracts have been involved in the DNA architecture, enhancing the recombination process and assisting chromatin structure. It is tempting to propose the putative function of conformational signal here exposed in processes such as VSG chromatin mediated silencing [34] and/or granting the vast antigenic variability needed for the efficient evasion of the immune host system that *T. brucei* has developed.

Conclusions

As expected, the genomes of Trypanosomes present a non-random distribution of sequence-dependent curvature. Regions of high intrinsic DNA curvature have been shown to have an active role in different biological functions.

Particularly, the functional role of high intrinsic DNA curvature as recognition signals has been involved in the transcription process both in prokaryotes and eukaryotes. Recently, using the regional integrated intrinsic curvature scoring, we have shown the association of regions of high intrinsic curvature with those related to the transcription initiation process in *Leishmania.*

Here we show that, the data available for *T. cruzi* have a similar behavior and may enable the extrapolation of a similar conclusion for its whole genome. Meanwhile in *T. brucei,* the association of regions of high intrinsic curvature with regions involved in transcription initiation but also in transcription termination was found to be significant. This finding points out to the existence of putative conserved process- and species specific-DNA architectural signals in kinetoplastids.

In addition, in *T. brucei*, the remarkable concentration of regions with high intrinsic curvature at the chromosome subtelomeric regions, where the species-specific genes for the gene family of the highly variable surface glycoproteins are located, may suggest a putative involvement of this structural signal facilitating the recombinational process.

Globally, the data here presented, while establishing particularities within trypanosomes, underscore the relevance of indirect DNA readout in these ancient eukaryotes.

References

1. Field V, Gautret P, Schlagenhauf P, Burchard GD, Caumes E, et al. (2010) Travel and migration associated infectious diseases morbidity in Europe, 2008. BMC Infect Dis 10: 330.
2. Kramer S (2012) Developmental regulation of gene expression in the absence of transcriptional control: the case of kinetoplastids. Mol Biochem Parasitol 181: 61-72.
3. Boothroyd JC (1985) Antigenic variation in African trypanosomes. Annu Rev Microbiol 39: 475-502.
4. Berriman M, Hall N, Sheader K, Bringaud F, Tiwari B, et al. (2002) The architecture of variant surface glycoprotein gene expression sites in Trypanosoma brucei. Mol Biochem Parasitol 122: 131-140.
5. Rudenko G, Lee MG, Van der Ploeg LH (1992) The PARP and VSG genes of Trypanosoma brucei do not resemble RNA polymerase II transcription units in sensitivity to Sarkosyl in nuclear run-on assays. Nucleic Acids Res 20: 303-306.

6. Campbell DA, Thomas S, Sturm NR (2003) Transcription in kinetoplastid protozoa: why be normal? Microbes Infect 5: 1231-1240.
7. Stuart KD, Schnaufer A, Ernst NL, Panigrahi AK (2005) Complex management: RNA editing in trypanosomes. Trends Biochem Sci 30: 97-105.
8. Martinez-Calvillo S, Vizuet-de-Rueda JC, Florencio-Martinez LE, Manning-Cela RG, Figueroa-Angulo EE (2010) Gene expression in trypanosomatid parasites. J Biomed Biotechnol 2010: 525241.
9. Dreesen O, Cross GA (2006) Consequences of telomere shortening at an active VSG expression site in telomerase-deficient Trypanosoma brucei. Eukaryot Cell 5: 2114-2119.
10. El-Sayed NM, Myler PJ, Bartholomeu DC, Nilsson D, Aggarwal G, et al. (2005) The genome sequence of Trypanosoma cruzi, etiologic agent of Chagas disease. Science 309: 409-415.
11. Cliffe LJ, Siegel TN, Marshall M, Cross GA, Sabatini R (2010) Two thymidine hydroxylases differentially regulate the formation of glucosylated DNA at regions flanking polymerase II polycistronic transcription units throughout the genome of Trypanosoma brucei. Nucleic Acids Res.
12. Gilinger G, Bellofatto V (2001) Trypanosome spliced leader RNA genes contain the first identified RNA polymerase II gene promoter in these organisms. Nucleic Acids Res 29: 1556-1564.
13. Martinez-Calvillo S, Yan S, Nguyen D, Fox M, Stuart K, et al. (2003) Transcription of Leishmania major Friedlin chromosome 1 initiates in both directions within a single region. Mol Cell 11: 1291-1299.
14. Martinez-Calvillo S, Nguyen D, Stuart K, Myler PJ (2004) Transcription initiation and termination on Leishmania major chromosome 3. Eukaryot Cell 3: 506-517.
15. Respuela P, Ferella M, Rada-Iglesias A, Aslund L (2008) Histone acetylation and methylation at sites initiating divergent polycistronic transcription in Trypanosoma cruzi. J Biol Chem 283: 15884-15892.
16. Siegel TN, Hekstra DR, Kemp LE, Figueiredo LM, Lowell JE, et al. (2009) Four histone variants mark the boundaries of polycistronic transcription units in Trypanosoma brucei. Genes Dev 23: 1063-1076.
17. Padilla-Mejia NE, Florencio-Martinez LE, Figueroa-Angulo EE, Manning-Cela RG, Hernandez-Rivas R, et al. (2009) Gene organization and sequence analyses of transfer RNA genes in Trypanosomatid parasites. BMC Genomics 10: 232.
18. Duhagon MA, Smircich P, Forteza D, Naya H, Williams N, et al. (2011) Comparative genomic analysis of dinucleotide repeats in Tritryps. Gene 487: 29-37.
19. Tiengwe C, Marcello L, Farr H, Dickens N, Kelly S, et al. (2012) Genome-wide analysis reveals extensive functional interaction between DNA replication initiation and transcription in the genome of Trypanosoma brucei. Cell Rep 2: 185-197.
20. Choi J, El-Sayed NM (2012) Functional genomics of trypanosomatids. Parasite Immunol 34: 72-79.
21. Mao Y, Najafabadi HS, Salavati R (2009) Genome-wide computational identification of functional RNA elements in Trypanosoma brucei. BMC Genomics 10: 355.
22. Kelly S, Wickstead B, Maini PK, 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.
23. Ohyama T (c2005) DNA conformation and transcription. Georgetown, Tex.
New York, NY.: Landes Bioscience; Springer Science Business Media. 211 p.

24. Smircich P, Forteza D, El-Sayed NM, Garat B (2013) Genomic analysis of sequence-dependent DNA curvature in leishmania. PLoS One 8: e63068.
25. Weatherly DB, Boehlke C, Tarleton RL (2009) Chromosome level assembly of the hybrid Trypanosoma cruzi genome. BMC Genomics 10: 255.
26. Tosato V, Ciarloni L, Ivens AC, Rajandream MA, Barrell BG, et al. (2001) Secondary DNA structure analysis of the coding strand switch regions of five Leishmania major Friedlin chromosomes. Curr Genet 40: 186-194.
27. Borst P, Sabatini R (2008) Base J: discovery, biosynthesis, and possible functions. Annu Rev Microbiol 62: 235-251.
28. Hertz-Fowler C, Figueiredo LM, Quail MA, Becker M, Jackson A, et al. (2008) Telomeric expression sites are highly conserved in Trypanosoma brucei. PLoS One 3: e3527.
29. Trifonov EN (1980) Sequence-dependent deformational anisotropy of chromatin DNA. Nucleic Acids Res 8: 4041-4053.
30. Marini JC, Levene SD, Crothers DM, Englund PT (1982) Bent helical structure in kinetoplast DNA. Proc Natl Acad Sci U S A 79: 7664-7668.
31. Nadeau JG, Crothers DM (1989) Structural basis for DNA bending. Proc Natl Acad Sci U S A 86: 2622-2626.
32. Haran TE, Mohanty U (2009) The unique structure of A-tracts and intrinsic DNA bending. Q Rev Biophys 42: 41-81.
33. De Santis P, Scipioni A (2013) Sequence-dependent collective properties of DNAs and their role in biological systems. Phys Life Rev 10: 41-67.
34. Pandya UM, Sandhu R, Li B (2013) Silencing subtelomeric VSGs by Trypanosoma brucei RAP1 at the insect stage involves chromatin structure changes. Nucleic Acids Res.

Acknowledgments

This work was supported by Fondo Clemente Estable (Agencia Nacional de Investigación e Innovación); Comisión Sectorial de Investigación Científica (Universidad de la Republica) and Programa de Desarrollo de Ciencias Básicas. PS received a Ph. D. fellowship (Agencia Nacional de Investigación e Innovación).

Figure captions

Figure 1. Graphical representation of sequence dependent curvature.
A - *T. cruzi* chromosome 9. Upper panel: Bar plots of chromosome positions with an IC value greater than 13 degrees per helical turn. Middle panel: Bar plots of chromosome positions with an RIIC value greater than the selected cutoff. Lower panel: both chromosome DNA strands are depicted in grey, overlaid with CDS features shown in blue. Features labeled as ncRNA, snRNA or snoRNAs are shown in green. tRNAs are shown in red. Assembly gaps are shown in brown.
B –. *T. brucei* chromosome 5 Panels are as in A. Subtelomeric VSG clusters are underlined.

Figure 2.Comparative analysis of high RIIC in *T. cruzi* SSRs.
The strand switch regions corresponding to gene clusters of more than 6 genes were counted and their RIIC score calculated. The bar plot shows the percentage of DSSR (dark grey) and CSSR (light grey) regions overlapping high RIIC scoring regions. Presence of internal sequencing gaps eliminates de region from the counting.

Figure 3. Comparative analysis of high RIIC in *T. brucei* regions associated to markers of the transcription process
The bar plot shows the percentage of H4K10ac associated (dark grey) and Base J containing (light grey) regions overlapping high RIIC scoring regions.

Figure 4. Logo representation of the main motif found by MEME analysis around high curvature peaks in *T. brucei* subtelomeric VSG clusters.
The sequences surrounding high IC peaks were analyzed by MEME as described in Materials and Methods, and the logo representation for the most significant motif is shown.

Figure 1

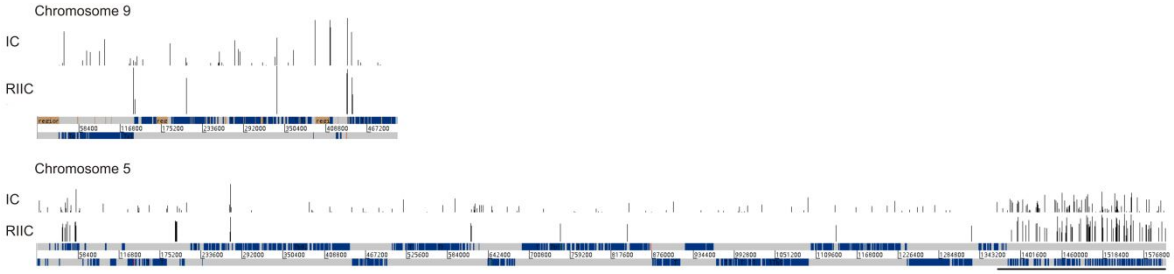


Figure 2

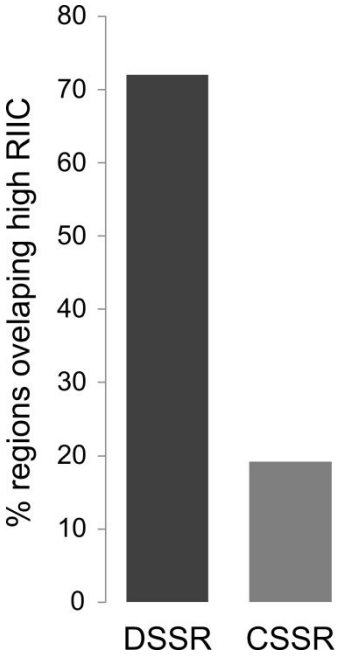


Figure 3

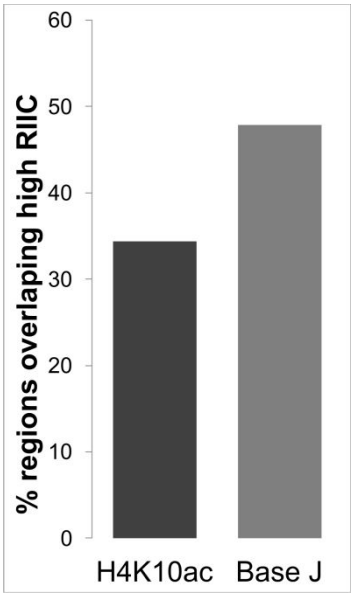
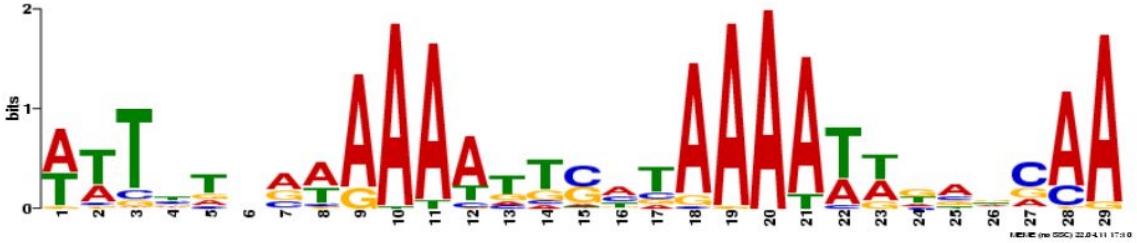

% regions overlaping high RIIC
60
50
40
30
20
10
0
H4K10ac
Base J

Figure 4


bits
2
1
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29

RNA polymerase I driven promoters for rDNA and protein coding genes in Tritryps exhibit the conserved eukaryotic conformation

Pablo Smircich[1,2], María Ana Duhagon[1,2] and Beatriz Garat[1]*

[1]Laboratorio de Interacciones Moleculares, Facultad de Ciencias, 11400 Montevideo, Uruguay
[2] Departamento de Genética, Facultad de Medicina, 11800 Montevideo, Uruguay.

for correspondence:
B. Garat: Laboratorio de Interacciones Moleculares,
Facultad de Ciencias Iguá 4225, 11400 Montevideo, Uruguay; E-mail: bgarat@fcien.edu.uy,; tel: +598 2525 86 18 ext 237; fax: +598 2 525 86 17

Abstract

Trypanosomatids are interesting models for basic research. Since they constitute a very early branch in eukaryotic evolution they present several remarkable deviations from standard paradigms. While the identification of promoters for the RNA polymerase II transcription, mainly responsible for the expression of protein coding genes has remained mostly elusive, early works in these organisms have been able to detailed define the location of RNA polymerase I (RNPI) promoters. In eukaryotes, they are well characterized by the presence of conformational rather than sequence signals directing protein recognition. To determine whether trypanosomatids would fulfill the conserved eukaryotic conformational characteristics, we analyzed the intrinsic DNA curvature in the proximity of the transcription start point of the rDNA promoter in *Trypanosoma brucei* as well as *Trypanosoma cruzi* and *Leishmania major*. We determined that these promoter regions posses intrinsic structural properties similar to other eukaryotic rDNA promoters. Furthermore, as the RNPI in *T. brucei* has the unusual property of directing the transcription of the genes coding for the highly expressed surface protein families, we extended the analysis to the regions containing their transcription start points. We found that the promoters of the genes coding for the variable surface glycoproteins (VSG), both in the metacyclic and bloodstream forms, and those coding for the procyclin proteins share the conformational characteristics of rDNA promoters. Our results indicate that rDNA curvature is conserved in ancient eukaryotes and might be a requirement of RNPI transcription independently of the class of genes being transcribed.

Keywords: Tritryps, RNA polymerase I, curvature,

Trypanosoma cruzi, *Trypanosoma brucei* and *Leishmania major*, the so called Tritryps, constitute the etiologic agents of Chagas disease, sleeping sickness and leishmaniasis respectively, causing millions of human deaths mainly in developing countries in tropical and subtropical regions. Trypanosomatids are interesting models for basic research since they constitute a very early branch in eukaryotic evolution presenting several remarkable deviations from standard eukaryotic paradigms.

Particularly, protein-coding genes are arranged in large divergent gene clusters (DGCs) with strict strand polarity (El-Sayed, et al., 2005). Initiating at still imprecisely located sites and in apparently constitutive manner, RNA polymerase II (RNPII) directs transcription of protein coding genes yielding polycistronic messengers. Individual mRNAs are generated by 5' trans-splicing, involving the addition of a small conserved RNA called spliced leader (SL) and 3' polyadenylation (Smith and Parsons, 1996). Regulation of gene expression seems to occur mainly at posttranscriptional levels (Martinez-Calvillo, et al., 2010). While in most eukaryotes RNA polymerase I (RNPI) drives rRNA gene transcription, in *T. brucei* this polymerase is also responsible for the highly developmentally regulated expression of the surface protein families, VSG and procyclins (Günzl, et al., 2003, Lee and Van der Ploeg, 1997). For these promoters, regulatory elements have been identified and their ability to increase the probability of transcription initiation has been assessed. Electrophoretic mobility shift assay experiments with the *T. brucei* VSG and procyclin promoters showed single stranded sequence specific binding suggesting that these promoters need to be in a relaxed form to allow binding of RNPI and transcription (Berberof, et al., 2000, Lee and Van der Ploeg, 1997).

In eukaryotes, the RNPI system is well characterized. rDNA transcription units are tandemly repeated been separated by an intergenic spacer region. Although the rRNA regions are among the most conserved gene sequences, the signal sequences that mediate transcription initiation show no

significant homologies. Nevertheless, the overall structural organization of rDNA promoters (*cis*-acting elements and *trans*-acting factors) of species phylogenetically as distant as protozoa, fungi, insects and mammals suggests that the general mechanisms of rDNA transcriptional regulation have been evolutionary conserved (Paule and White, 2000). The presence of conformational rather sequence signals, is responsible for directing transcription initiation. DNA structural features exhibited by ribosomal RNA gene promoters have been analyzed for several organisms (Marilley and Pasero, 1996, Roux-Rouquie and Marilley, 2000). Remarkably, a flexible structure surrounded by inherent curvature in the proximity of the transcription start point and an appropriate twist angle variation along the core promoter has been described. The reported functional exchange of transcription factors from different species may be explained by the existence a conserved structural signal (Marilley and Pasero, 1996).

In order to understand the molecular mechanisms involved in transcription initiation by RNPI in trypanosomatids, we firstly analyzed whether the Tritryp rDNA promoters would fulfill the conserved eukaryotic conformational characteristics.

For determining the sequence-dependent DNA curvature, we used bend-it, an automated prediction system based on algorithms developed using the helical asymmetry of trinucleotides (Brukner, et al., 1995, Gabrielian, et al., 1996) of the DNAtools (Munteanu, et al., 1998) at the Computer Resource for Molecular Biology of the International Centre for Genetic Engineering and Biotechnology, ICGEB (http://www.icgeb.trieste.it/dna). The sequences used for the analysis of rDNA promoters are available in Genebank and the corresponding transcription start sites (TSSs) have been already reported. That is for *T. brucei* the sequence used was AF416290.1 and the TSS has been defined (White, et al., 1986). For *T. cruzi* the sequence and TSS of the major rDNA promoter was obtained from U89779.1 (Figueroa-Angulo, et al., 2003, Martinez-Calvillo and Hernandez, 1994, Nunes, et al., 1997, Stolf, et al., 2003). Finally for *L. major*

the sequence around the TSS was extracted from AF421555.1 (de Andrade Stempliuk and Floeter-Winter, 2002). In order to compare to a previously characterized rDNA promoter, the human sequence (X01547) was used (Miesfeld and Arnheim, 1982) as previously reported (Marilley and Pasero, 1996).

We found that the Tritryp rDNA promoters conform to the eukaryotic conserved DNA structural features, in spite of the absence of significant sequence conservation among them or with the corresponding human region (data not shown). Indeed, the Tritryp rDNA promoter intrinsic curvature consensus profile presents a region of high intrinsic curvature preceding the TSS and the characteristic loss of curvature at this site (**Figure 1 A**). While *T. cruzi* and *T. brucei* rDNA promoters share an almost identical intrinsic curvature pattern that is closely similar to the human one, the pattern of curvature displayed by *L. major* is somehow different (**Figure 1 B**). The differences in base composition and intrinsic curvature median in the genome (2.63 degrees/helical turn for *L. major* while 3.36 and 3.52 degrees/helical turn for *T. cruzi* and *T. brucei* respectively)(Smircich, et al., 2013) may underlay this observation for reasons still not understood.

In *T. brucei*, RNPI transcribes also the major surface protein genes, the variant surface glycoprotein (VSG) of the bloodstream form and the acidic repetitive protein of the procyclic form.

The VSG are key actors in the evasion of the host immune system by changing the component that is exhibited at the cell surface in a process named antigenic variation. There are 10-20 telomeric bloodstream expression sites (BES) from which the VSG can be transcribed as a polycistron, including expression site associated genes, but only one among them is transcribed at a time (Becker, et al., 2004). Variation is attained through exchange of the active BES and by recombination nurtured by a massive archive of 1000 silent VSGs and VSG pseudogenes located at subtelomeres. The promoter responsible of BES transcription is located after the characteristic 50 bp repeats

and may be duplicated in specific BES (Pays, et al., 2001). The sequences of the BES promoters have been described and share a high degree of identity (Becker, et al., 2004). The *T. brucei* 427 dominant expression site (DES) promoter has been characterized in detail by Zomerdijk *et. al.* (Zomerdijk, et al., 1991, Zomerdijk, et al., 1990) and the transcription start site determined. This sequence (X17350.1) was used for the analysis here presented.

In addition, VSG are also expressed in the metacyclic infective form developed at the salivary gland of the fly vector. The genes encoding the metacyclic VSGs have been well characterized and are constituted by monocistronic messengers each one produced upon its own promoter (Pays, et al., 2001). In spite of the fact that there is an estimated number of 25 metacyclic VSG expression sites, sequence data are scarce due to their telomere location (Taylor and Rudenko, 2006). However, the transcription start site of one of the metacyclic VSG expression site has been experimentally determined (Ginger, et al., 2002) enabling the conformational analysis here proposed (AJ486955.1).

Upon entering the insect vector, the trypanosomes transform into procyclic forms and the VSG is replaced by the procyclins. These are very abundant stage-specific glycoproteins which are, like VSG, exposed at the cell surface. They present two different types of polypeptides, one of them contains a domain of tandem pentapeptides repeats (GPEET) while the other is characterized by a dipeptide repeat (EP) (Lee and Van der Ploeg, 1997). The procyclin genes are organized in two distinct loci (GPEET/PAG3 and EP1/PAG1-2) containing each one a single promoter (Brown, et al., 1992, Laufer and Gunzl, 2001, Sherman, et al., 1991). The sequences corresponding to the *GPEET* (S60066.1) and *EP1* (M38222.1) promoters were used in this work.

We analyzed the conformational characteristics of the regulatory regions of genes encoding proteins transcribed by RNA pol I in *T. brucei*. It is worth to mention that these promoters also lack conserved

sequence among them (Palenchar and Bellofatto, 2006). Nevertheless, we found that the curvature consensus profile also conforms to the pattern of intrinsic curvature conserved in eukaryotic rDNA promoters (**Figure 1 C**). While the core promoter of the genes encoding the metacyclic and bloodstream VSG as well as the *EP1*, share a curvature pattern that is closely similar to the rDNA one, the pattern of curvature displayed by the *GPEET* is different (**Figure 1 D**). The subtle differences in conformation patterns may show the existence of particular regulatory features. As a control, 35 intergenic sequences were analyzed and their curvature consensus shows no peculiarities (**Figure 1 E** and **D**).

Our findings on rDNA promoters in Tritryps, reinforce the previous consideration about intrinsic curvature being fundamental for ribosomal core promoter function (Marilley) covering these ancient eukaryotes. In addition, the data here presented further extend the promoter secondary structure characteristics beyond rDNA transcription, supporting the requirement of DNA intrinsic curvature in RNPI machinery assembly independently of the transcribed product nature. Early studies show that in spite of the absence of sequence conservation, functional RNPI hybrid promoters could be obtained depending on spacing constraints. These results could be explained by DNA architecture requirements as the ones here revealed for the core promoters. While, the requisite of specific transcription factors for the assembly of RNPI machinery at the different core promoters, could justify the absence of sequence conservation, competition experiments made evident the existence of shared transcription factors (Laufer and Gunzl, 2001). Since different sequence arrangements can lead to conserved secondary structure, these observations are compatible with the existence of conformational signals. The absence of a selective pressure to preserve sequence conservation observed in rDNA promoters of eukaryotes, is remarked in *T. brucei* where the three promoters have extensively diverged in sequence composition. However, in all these cases, the conformational organization has been strictly maintained.

Acknowledgements

This work was financially supported by

PEDECIBA and CSIC, UdelaR and by ANII.

Bibliography

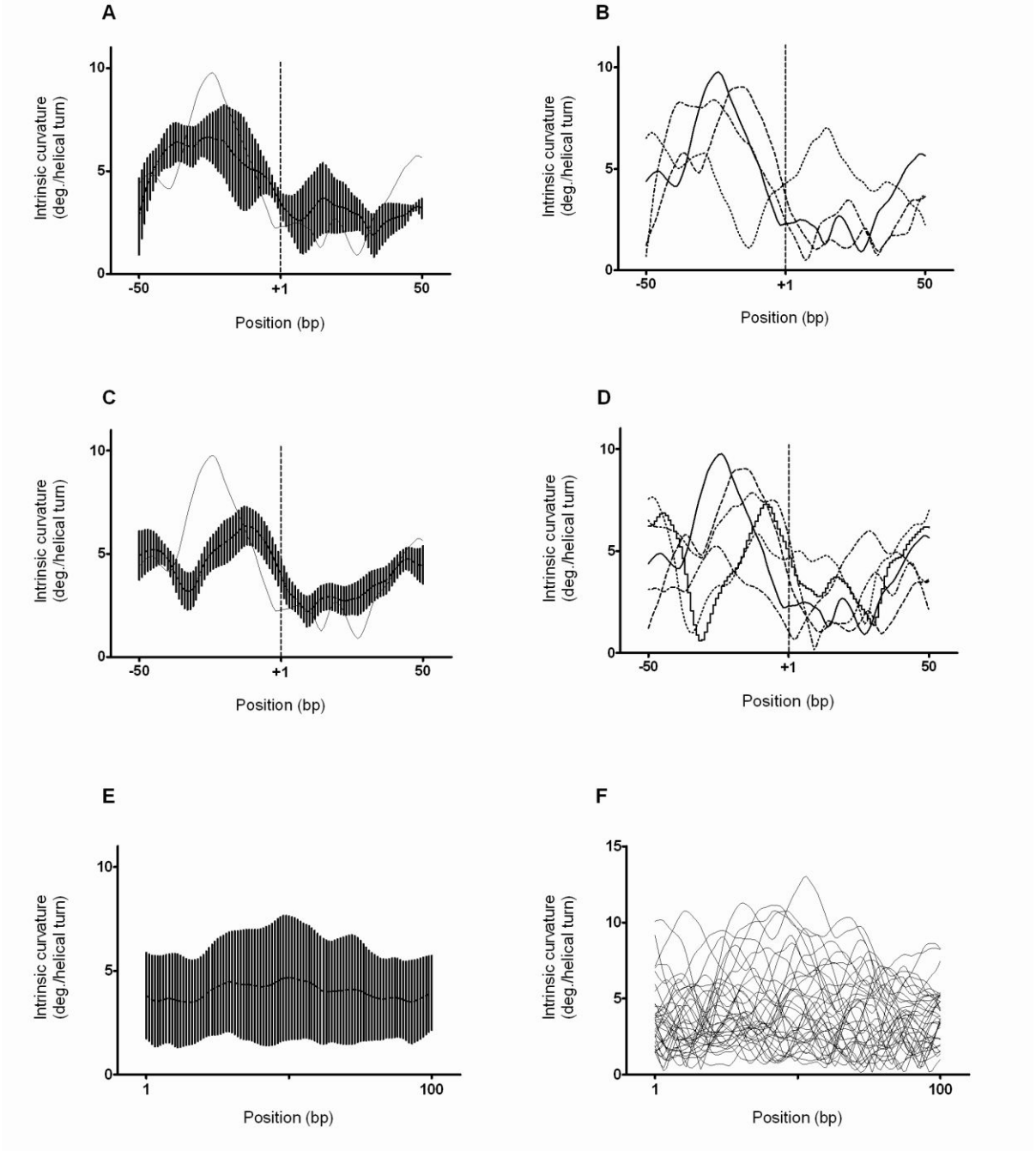
1. Becker, M., Aitcheson, N., Byles, E., Wickstead, B., Louis, E., and Rudenko, G., 2004. Isolation of the repertoire of VSG expression site containing telomeres of Trypanosoma brucei 427 using transformation-associated recombination in yeast. Genome Res 14**,** 2319-2329.
2. Berberof, M., Vanhamme, L., Alexandre, S., Lips, S., Tebabi, P., and Pays, E., 2000. A single-stranded DNA-binding protein shared by telomeric repeats, the variant surface glycoprotein transcription promoter and the procyclin transcription terminator of Trypanosoma brucei. Nucleic Acids Res 28**,** 597-604.
3. Brown, S. D., Huang, J., and Van der Ploeg, L. H., 1992. The promoter for the procyclic acidic repetitive protein (PARP) genes of Trypanosoma brucei shares features with RNA polymerase I promoters. Mol Cell Biol 12**,** 2644-2652.
4. Brukner, I., Sanchez, R., Suck, D., and Pongor, S., 1995. Sequence-dependent bending propensity of DNA as revealed by DNase I: parameters for trinucleotides. EMBO J 14**,** 1812-1818.
5. de Andrade Stempliuk, V., and Floeter-Winter, L. M., 2002. Functional domains of the rDNA promoter display a differential recognition in Leishmania. Int J Parasitol 32**,** 437-447.
6. El-Sayed, N. M., Myler, P. J., Blandin, G., Berriman, M., Crabtree, J., Aggarwal, G., Caler, E., Renauld, H., Worthey, E. A., Hertz-Fowler, C., Ghedin, E., Peacock, C., Bartholomeu, D. C., Haas, B. J., Tran, A. N., Wortman, J. R., Alsmark, U. C., Angiuoli, S., Anupama, A., Badger, J., Bringaud, F., Cadag, E., Carlton, J. M., Cerqueira, G. C., Creasy, T., Delcher, A. L., Djikeng, A., Embley, T. M., Hauser, C., Ivens, A. C., Kummerfeld, S. K., Pereira-Leal, J. B., Nilsson, D., Peterson, J., Salzberg, S. L., Shallom, J., Silva, J. C., Sundaram, J., Westenberger, S., White, O., Melville, S. E., Donelson, J. E., Andersson, B., Stuart, K. D., and Hall, N., 2005. Comparative genomics of trypanosomatid parasitic protozoa. Science 309**,** 404-409.
7. Figueroa-Angulo, E., Martinez-Calvillo, S., Lopez-Villasenor, I., and Hernandez, R., 2003. Evidence supporting a major promoter in the Trypanosoma cruzi rRNA gene. FEMS Microbiol Lett 225**,** 221-225.
8. Gabrielian, A., Simoncsits, A., and Pongor, S., 1996. Distribution of bending propensity in DNA sequences. FEBS Lett 393**,** 124-130.
9. Ginger, M. L., Blundell, P. A., Lewis, A. M., Browitt, A., Gunzl, A., and Barry, J. D., 2002. Ex vivo and in vitro identification of a consensus promoter for VSG genes expressed by metacyclic-stage trypanosomes in the tsetse fly. Eukaryot Cell 1**,** 1000-1009.
10. Günzl, A., Bruderer, T., Laufer, G., Schimanski, B., Tu, L., Chung, H., Lee, P., and Lee, M. G., 2003. RNA Polymerase I Transcribes Procyclin Genes and Variant Surface Glycoprotein Gene Expression Sites in Trypanosoma brucei. Eukaryotic Cell 2**,** 542-551.
11. Laufer, G., and Gunzl, A., 2001. In-vitro competition analysis of procyclin gene and variant surface glycoprotein gene expression site transcription in Trypanosoma brucei. Mol Biochem Parasitol 113**,** 55-65.
12. Lee, M. G., and Van der Ploeg, L. H., 1997. Transcription of protein-coding genes in trypanosomes by RNA polymerase I. Annu Rev Microbiol 51**,** 463-489.
13. Marilley, M., and Pasero, P., 1996. Common DNA structural features exhibited by eukaryotic ribosomal

gene promoters. Nucleic Acids Res 24, 2204-2211.
14. Martinez-Calvillo, S., and Hernandez, R., 1994. Trypanosoma cruzi ribosomal DNA: mapping of a putative distal promoter. Gene 142, 243-247.
15. Martinez-Calvillo, S., Vizuet-de-Rueda, J. C., Florencio-Martinez, L. E., Manning-Cela, R. G., and Figueroa-Angulo, E. E., 2010. Gene expression in trypanosomatid parasites. J Biomed Biotechnol 2010, 525241.
16. Miesfeld, R., and Arnheim, N., 1982. Identification of the in vivo and in vitro origin of transcription in human rDNA. Nucleic Acids Res 10, 3933-3949.
17. Munteanu, M. G., Vlahovicek, K., Parthasarathy, S., Simon, I., and Pongor, S., 1998. Rod models of DNA: sequence-dependent anisotropic elastic modelling of local bending phenomena. Trends Biochem Sci 23, 341-347.
18. Nunes, L. R., de Carvalho, M. R., and Buck, G. A., 1997. Trypanosoma cruzi strains partition into two groups based on the structure and function of the spliced leader RNA and rRNA gene promoters. Mol Biochem Parasitol 86, 211-224.
19. Palenchar, J. B., and Bellofatto, V., 2006. Gene transcription in trypanosomes. Mol Biochem Parasitol 146, 135-141.
20. Paule, M. R., and White, R. J., 2000. Survey and summary: transcription by RNA polymerases I and III. Nucleic Acids Res 28, 1283-1298.
21. Pays, E., Lips, S., Nolan, D., Vanhamme, L., and Perez-Morga, D., 2001. The VSG expression sites of Trypanosoma brucei: multipurpose tools for the adaptation of the parasite to mammalian hosts. Mol Biochem Parasitol 114, 1-16.
22. Roux-Rouquie, M., and Marilley, M., 2000. Modeling of DNA local parameters predicts encrypted architectural motifs in Xenopus laevis ribosomal gene promoter. Nucleic Acids Res 28, 3433-3441.
23. Sherman, D. R., Janz, L., Hug, M., and Clayton, C., 1991. Anatomy of the parp gene promoter of Trypanosoma brucei. EMBO J 10, 3379-3386.
24. 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.
25. Smith, D. F., and Parsons, M., 1996. Molecular biology of parasitic protozoa. IRL Press at Oxford University Press, Oxford ; New York.
26. Stolf, B. S., Souto, R. P., Pedroso, A., and Zingales, B., 2003. Two types of ribosomal RNA genes in hybrid Trypanosoma cruzi strains. Mol Biochem Parasitol 126, 73-80.
27. Taylor, J. E., and Rudenko, G., 2006. Switching trypanosome coats: what's in the wardrobe? Trends Genet 22, 614-620.
28. White, T. C., Rudenko, G., and Borst, P., 1986. Three small RNAs within the 10 kb trypanosome rRNA transcription unit are analogous to domain VII of other eukaryotic 28S rRNAs. Nucleic Acids Res 14, 9471-9489.
29. Zomerdijk, J. C., Kieft, R., Shiels, P. G., and Borst, P., 1991. Alpha-amanitin-resistant transcription units in trypanosomes: a comparison of promoter sequences for a VSG gene expression site and for the ribosomal RNA genes. Nucleic Acids Res 19, 5153-5158.
30. Zomerdijk, J. C., Ouellette, M., ten Asbroek, A. L., Kieft, R., Bommer, A. M., Clayton, C. E., and Borst, P., 1990. The promoter for a variant surface glycoprotein gene expression site in Trypanosoma brucei. EMBO J 9, 2791-2801.

Figure Captions

Figure 1. Intrinsic curvature maps around the TSSs for RNA polymerase I transcribed products in Tritryps. A. Intrinsic curvature average for Tritryps´rDNA promoters. Each position is plotted along with the corresponding standard error (represented as vertical bars). **B.** Individual intrinsic curvature for Tritryps´rDNA promoters. Solid line: Human. Dashed line: *T. cruzi*. Dash/Dot line: *T. brucei*. Dotted line: *L major*. **C.** Intrinsic curvature average around the TSS for all the RNPI transcribed promoters in *T. brucei* (*VSG*, *MVSG*, *GPEET* and *EP1*). **D.** Individual intrinsic curvature around the TSS for all the RNPI transcribed promoters in *T. brucei*. Solid line: Human rDNA. Dashed line: rDNA Dash/Dot line: *VSG*. Dash/Dot/Dot line: *MVSG*. Dotted line: *GPEET*. Discontinuous line: *EP1*. **E.** Intrinsic curvature average of 35 random intergenic sequences in *T. brucei.* **F.** Individual intrinsic curvature of 35 random intergenic sequences in *T. brucei.* In A and C the intrinsic curvature of human rDNA promoter was included as a reference (solid line).

Figure 1


A
B
C
D
E
F
Intrinsic curvature (deg./helical turn)
Position (bp)
-50
+1
50
1
100
0
5
10
15

3.2 Estudio de las dinámicas de transcripción y traducción

3.2.1 Aproximación experimental para la identificación de sitios de inicio de la transcripción

La transcripción en tripanosomátidos es muy distinta con respecto al paradigma eucariota. Como ya se señaló, una de las características más llamativa es la habilidad de la ARN polimerasa I (ARNPI) de transcribir genes codificantes para proteínas en *T. brucei* (genes de las VSG y prociclínas). Las regiones promotoras de la transcripción para estos casos han sido identificadas. Sin embargo, los inicios transcripcionales correspondientes a la ARNPII eran desconocidos al momento de la realización de esta parte del trabajo. Esto nos llevó a plantearnos la definición de los sitios de inicio de la transcripción para esta ARN polimerasa. Los ensayos de *run-on* (y variantes relacionadas) han sido utilizados con esta finalidad ya que permiten evaluar los ARN que están siendo transcritos en un determinado momento. Clásicamente, para la detección de estos ARN nacientes, el ARN celular es marcado de forma radioactiva mientras está siendo transcrito y los genes buscados son detectados por hibridación a sondas complementarias (Murphy, D. 1993). En este trabajo nos propusimos buscar los ARNm nacientes de manera global, lo cual requiere del marcado y purificación de estos ARNs. Para llevarlo a cabo, nos basamos en el trabajo de Patrone y cols. (Patrone, G. *et al.* 2000) en el cual las moléculas son marcadas con Biotina-16-UTP (BioUTP), para luego ser purificadas por columna de afinidad. En ese trabajo, los autores evalúan la transcripción *de novo* de genes candidatos mediante RT-PCR en tiempo real (qRT-PCR) luego de la purificación. Nosotros nos propusimos la recuperación de esta fracción para luego secuenciar de forma masiva los ARN purificados.

En general, para discriminar el transcriptoma de la ARNPII, los autores se basan en un comportamiento clásico de las ARNPs, descontando las especies transcritas por las ARNPI y ARNPIII que son bien conocidas. Sin embargo, teniendo en cuenta que en *T. brucei* existen genes que codifican proteínas transcritos por la ARNPI y que esta particularidad podría estar extendida no sólo a otros genes sino también a los otros tripanosomátidos, en primera instancia no se puede asumir la premisa de comportamiento clásico de las ARNPs. Por otra parte, el inhibidor específico de las ARNPs eucariotas (α-amanitina) bloquea secuencialmente las ARNPIII y II a medida que se aumenta su concentración. La ARNPI de tripanosomátidos comparte con en el resto de los eucariotas la característica de ser insensible a esta droga. Por lo tanto, para llevar adelante la estrategia propuesta en tripanosomátidos, nos planteamos la

necesidad de obtener el transcriptoma de la ARNPI. Una incubación con altas concentraciones de α-amanitina nos permite marcar únicamente los transcritos de ARNPI durante el ensayo de *run-on*. La aproximación de la detención de las ARNPII y ARNPIII con α-amanitina en experimentos de *run-on,* ha sido ensayada anteriormente en *T. brucei* y las condiciones necesarias reportadas (Kooter, J. M. *et al.* 1987; Rudenko, G. *et al.* 1989).

Obtención, marcado y purificación de ARNs nacientes

Si bien en nuestro laboratorio ya se habían realizado purificaciones de núcleos en *T. cruzi* para análisis bioquímico y de biología celular (Duhagon, M. A. *et al.* 2003), fue necesario implementar la obtención de núcleos transcripcionalmente activos en *T. brucei*, protocolo que se puso a punto en el laboratorio del Dr. El-Sayed en la Universidad de Maryland, EEUU.

Para la obtención de núcleos funcionales se partió de una variante del protocolo descrito en Murphy y cols. (Murphy, N. B. *et al.* 1987). El mismo consiste en el homogenado de los parásitos en condiciones suaves controlando la muestra por observación mediante microscopía de fluorescencia con tinción con DAPI. Se determinó el número de pasajes por el homogenizador que permitiera reducir a un mínimo los parásitos enteros visualizados. De esta manera, se consiguió obtener núcleos que presentaran poco daño mecánico. Luego de la purificación de estos núcleos, se realizaron los ensayos de *run-on*. Con el fin de poder purificar por afinidad las moléculas transcriptas *de novo*, se realizaron incubaciones en presencia de nucleótidos modificados. Inicialmente, se utilizó BioUTP para el marcado, método previamente utilizado por Patrone y cols. (Patrone, G. *et al.* 2000) (ver sección Materiales y Métodos).

Una vez realizados los experimentos de *run-on* y purificado el ARN (obteniéndose un rendimiento de alrededor de 5μg de ARN total a partir de los núcleos de $1x10^9$ parásitos en fase exponencial de crecimiento), la actividad de los núcleos y la incorporación del nucleótido marcado se controló mediante *dot-blot*. En la membrana se sembraron diluciones seriadas (1/10) de los ARN obtenidos de experimentos con y sin el agregado de la marca, así como un control positivo que consistió en el oligonucleótido $(dTdG)_{20}$ previamente biotinilado (Figura 3.2.1.1).

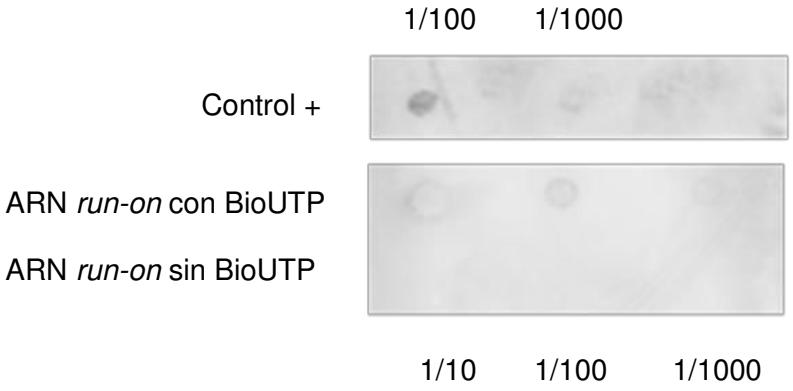


Figura 3.2.1.1 Detección de ARN sintetizado *de novo* en presencia de BioUTP mediante ensayo de *dot-blot*. Diluciones seriadas de cantidades iguales de ARN fueron sembradas y detectadas con estreptoavidina acoplada a HRP por quimioluminiscencia en una cámara CCD. En el ensayo mostrado, las concentraciones iniciales fueron de 57ng/μL para el ARN con BioUTP y 54ng/μL para el control sin BioUTP. El control + consistió en el oligonucleótido $(dTdG)_{20}$ biotinilado (600ng/μL)

Se observa que, a igual cantidad de ARN, hay señal en el caso de las muestras purificadas a partir del *run-on* en presencia de BioUTP. De hecho, la señal que se obtiene en la concentración 1/10 aparece quemada (más clara) por exceso de marca. De este ensayo podemos concluir que los núcleos preparados son funcionalmente activos y además que en las condiciones ensayadas se logra marcar los ARN que son sintetizados *de novo*.

Con este resultado nos propusimos purificar por afinidad estos ARNs sintetizados *de novo*. Para esto, se sometieron las muestras de ARN con y sin biotinilar a cromatografía de afinidad en columna de estreptoavidina-agarosa. Se preparó ADNc de las muestras purificada y sin purificar, y luego se realizaron ensayos de qRT-PCR con cebadores específicos para los genes de β-tubulina, EP1 y ARNr 5S. Se conservó el ARN no unido a la columna (*flow through*, FT) para utilizar como normalizador de las purificaciones. Las relaciones entre el ARN purificado y el FT para cada caso fueron calculadas a partir de los datos cuantitativos (Ct) obtenidos. Los resultados se resumen como la relación de las cantidades medidas por qRT-PCR que son purificadas en la columna a partir de ARN marcado y sin marcar (Figura 3.2.1.2).

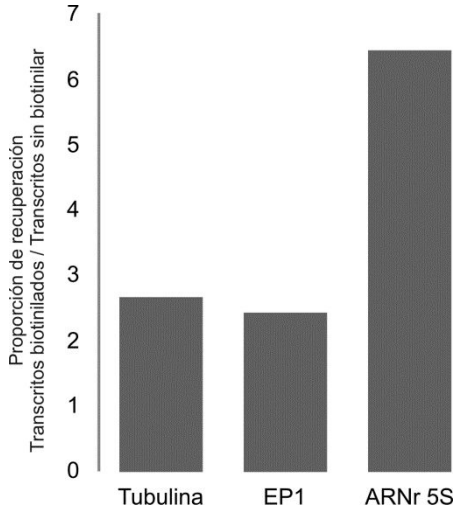


Figura 3.2.1.2 Relación entre el rendimiento de purificación de los transcritos indicados, usando la columna estreptoavidina-agarosa para los experimentos de *run-on* en presencia y en ausencia de marcado con biotina. Se toma como normalizador la cantidad de ARN obtenido en FT para cada transcrito y cada condición (marcado y sin marcar).

Por lo tanto, la cromatografía de afinidad con estreptoavidina-agarosa permitió enriquecer las muestras en ARN sintetizado *de novo* marcado con BioUTP. Sin embargo, se purifica también ARN usando las muestras de ARN sin biotinilar, lo cual puede ser atribuido a una unión inespecífica a la columna. No fue posible mejorar este resultado cambiando las condiciones de purificación, como por ejemplo realizando un mayor número de lavados de la columna. Además, notamos que el grado de purificación para la molécula de ARNr 5S es mayor que para las otras, lo cual también confirma la retención inespecífica de la columna usada en esta etapa de purificación.

Inhibición de la transcripción con α-amanitina

De cualquier manera, se procedió a estudiar si con estas condiciones de purificación éramos capaces de observar el efecto de la droga α-amanitina sobre la transcripción de las ARNP. Se usó la concentración de α-amanitina necesaria para detener la transcripción de las ARNPII y III definida por Rudenko y cols. (Rudenko, G. *et al.* 1989). Para cuantificar el grado de inhibición de las polimerasas se realizaron experimentos de qRT-PCR con los cebadores anteriormente mencionados que permiten evaluar ARNs transcritos por la ARNPI (EP1), ARNPII (β-tubulina) y ARNPIII (ARNr 5S). Dado que la transcripción de el gen de EP1 no está inhibida por la presencia de la droga, se comparó la cantidad de transcrito del gen de β-tubulina y ARNr 5S con respecto a ese gen. La Figura 3.2.1.3 muestra las cantidades relativas de los transcritos de los genes en cada condición. De esta manera, vemos que las

cantidades de transcrito de β-tubulina y ARNr 5S luego del pasaje por la columna de estreptoavidina-agarosa se ven disminuidas con respecto a la cantidad de transcrito de EP1, en presencia de α-amanitina. Sin embargo, estos ensayos muestran que la disminución de ambas transcripciones no es completa.

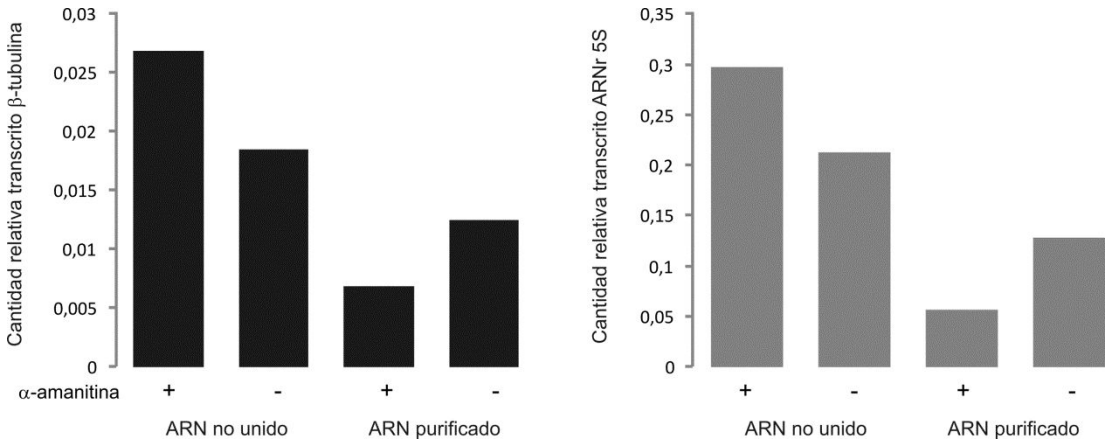


Figura 3.2.1.3 Cantidades relativas de los transcritos β-tubulina y ARNr 5S en los diferentes pasos de la purificación medidos con respecto al transcrito EP1.

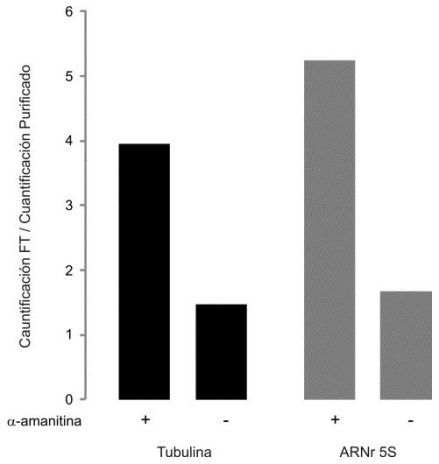


Figura 3.2.1.4 Relación entre las cantidades de transcrito de β-tubulina y ARNr 5S no retenidas (FT) y purificadas que se cuantifican en presencia y ausencia de α-amanitina (relativas al transcrito de EP1). El gráfico muestra que en presencia de α-amanitina la cantidad no retenida por la columna es mayor a la que se cuantifica en el eluído para ambos genes.

La Figura 3.2.1.4 muestra más claramente estas relaciones para ambos genes en presencia y ausencia de la droga. Esto pone en evidencia nuevamente que la

purificación por afinidad de las moléculas biotiniladas arrastra ARNs que se encontraban en la célula previo a la inhibición y marcado de los ARNs nacientes.

La presencia de este arrastre luego de la purificación desestimuló proseguir con la secuenciación masiva de las muestras para detectar el transcriptoma específico de la ARNPI. De acuerdo a los resultados obtenidos, el ruido que se obtendría de ARNs inespecíficos sería muy importante, dificultando distinguir las moléculas transcritas por la ARNPI y purificadas por la incorporación de biotina, de las moléculas de ARN del transcriptoma total purificadas inespecíficamente.

Como forma de reducir costos, y buscando una alternativa que pudiera rendir una eventual mejora en la purificación de los ARN sintetizados *de novo* ensayamos la sustitución del marcado con BioUTP por BrUTP. La purificación de ARN marcado con BrUTP se realizó mediante columna con anticuerpo comercial específico. Para esto se sometieron a cromatografía de afinidad las muestras de ARN con y sin el agregado de marca y para la cuantificación se procedió de forma análoga a lo realizado anteriormente. Sin embargo, esta estrategia no rindió resultados reproducibles.

Por lo tanto, concluimos que resultaría inconveniente realizar esta aproximación basada en marcado y purificación de ARN sintetizados *de novo* acompañados de inhibición específica de ARNPs y nos planteamos buscar estrategias alternativas. En este sentido hemos considerando la posibilidad de realizar ensayos de inmunoprecipitación de cromatina utilizando anticuerpos que reconozcan factores del complejo de transcripción basal para detectar lugares de inicio y así evaluar indirectamente los genes transcritos por esta ARN polimerasa. Otro elemento que nos desestimuló a continuar con esta aproximación fue la publicación durante el desarrollo de estos experimentos de varios trabajos en donde se abarca la pregunta de la determinación de los sitios de inicio de la ARNPII. Como comentamos en la introducción, por un lado se ha descrito la localización de modificaciones de cromatina característicos de sitios de inicio trancripcionales a nivel genómico en *T. brucei* (Respuela, P. *et al.* 2008; Siegel, T. N. *et al.* 2009; Thomas, S. *et al.* 2009; Wright, J. R. *et al.* 2010). Además, Kolev y cols. en su trabajo de caracterización del transcriptoma de *T. brucei,* realizan ensayos de enriquecimiento en ARN nacientes para su posterior secuenciación (Kolev, N. G. *et al.* 2010). Con esta estrategia análoga a la que nos planteamos, los autores determinan los sitios de inicio transcripcional, resultado que se solapa con nuestro objetivo inicial.

Materiales y Métodos

Parásitos y condiciones de cultivo

La forma procíclica de la cepa Lister 427 de *T. brucei* fue utilizada para los ensayos. Los parásitos fueron mantenidos a 28ºC en medio Cunningham (Cunningham, I. 1977), realizando pasajes a una dilución inicial de $1x10^5$ parásitos/mL.

Aislamiento de núcleos para run-on

Tampón fosfato salino (PBS)

NaCl (0,14M), KCl (2,6mM), KH2PO4 (7mM), glucosa (6mM), pH=7,4

Solución A (Sol A)

Sacarosa (0,5M), KCl (50mM), $MgCl_2$ (5mM), DTT (1mM), Tris-HCl (pH=7,4, 50mM), NP-40 (0,5%)

Solución B (Sol B)

Sacarosa (0,5M), KCl (50mM), $MgCl_2$ (5mM), DTT (1mM),Tris-HCl (pH=7,4, 50mM), Espermidina (5mM)

Solución C (Sol C)

Glicerol (25%), $MgCl_2$ (5mM), Tris-HCl (pH=7,4, 50mM), DTT (1mM)

Para la preparación de núcleos se parte de $1x10^9$ parásitos en fase exponencial de crecimiento. Los parásitos se centrifugan 5 minutos a 6000g y el *pellet* se resuspende en 10mL de PBS y 30mL de la Sol A. Se realizan 15 pasajes por el *pestle* B (0.15mm) del homogenizador *Kimble Kontes Dounce Tissue Grinder* (Fisher Scientific). Se centrifuga nuevamente 5 minutos a 6000g y el *pellet* se resuspende en 40mL de la Sol B para pasarlo 9 veces por el homogenizador utilizando el *pestle* A (0.07mm). Luego de centrifugar 5 minutos a 6000g, se resuspende el *pellet* conteniendo los núcleos en 80μL de la Sol C mezclando con la punta de la pipeta y se guarda a -80ºC hasta el momento de realizar el ensayo de *run-on*. Las homogenizaciones se realizan en hielo y las centrifugaciones a 4ºC.

Ensayo de run-on

Tampón de *run-on*

KCl (200mM), Tris-HCl (pH=7,4, 20mM), $MgCl_2$ (5mM), glicerol (20%), Sacarosa (200mM), ATP (4mM), CTP (2mM), GTP (2mM), DTT (4mM).

El ensayo comienza con la preincubación de 40μL de núcleos con 10μL de α-amanitina (5mg/mL) durante 15 minutos en hielo. Como control 40μL de núcleos se incuban en hielo con 10 μL de H_2O.

La reacción de síntesis se realiza durante 30 minutos a 29ºC agregando a los núcleos preincubados 70μL de tampón de *run-on*, 20μL de α-amanitina, 5μL de inhibidor de RNAsa (*Rnase OUT*, Invitrogen) y 5μL de Biotina-16-UTP 10mM (Roche) o 5μL de 5-Bromo Uridina tri fosfato 10mM (BrUTP, Life Technologies). Luego de pasado el tiempo se agregan 1,1mL de Trizol para parar la reacción y permitir la extracción de ARN total.

En el caso de las reacciones de *run-on* realizadas para controlar el marcado del ARN, los núcleos no se tratan con α-amanitina y el control se realiza en ausencia de BioUTP o BrUTP.

Extracción de ARN total

A los productos de la reacción de *run-on* se les agrega 200μL de Cloroformo y se centrifuga a 10000g. A la fase acuosa se le agrega 1 volumen de isopropanol y luego de 15 minutos de incubación se centrifuga el ARN a 10000g. El *pellet* es lavado con etanol 70%, se seca y se resuspende en 100μL de H_2O libre de nucleasas a 58ºC por 10 minutos. El ARN se trata con DNAsa libre de RNAsas para evitar la posible contaminación con ADN genómico (*kit DNA-free*, Ambion). Para la realización de los ensayos de *dot-blot* (ver más adelante), los ARN obtenidos fueron concentrados utilizando columnas del kit *RNA easy* (Quiagen) según las recomendaciones del proveedor.

Síntesis de ADN copia (ADNc)

La reacción de retrotranscripción se realiza con la enzima *SuperScriptIII Reverse Transcriptase* (Invitrogen) según las indicaciones del proveedor, agregando 50ng de cebadores al azar y realizando la síntesis durante 1 hora a 55ºC. 2μL de ARN sin

purificar y 11μL del ARN purificado por columna de afinidad se utilizaron como molde en las reacciones.

Ensayos de dot-blot

Las muestras se llevaron a concentraciones equivalentes y 2μL de las diluciones con ARN marcado y sin marcar se sembraron con aplicación de vacío en una membrana de nylon y luego se fijaron por exposición a UV. Luego se realizó la detección de las moléculas marcadas como se detalla más adelante.

Detección BioUTP

El protocolo se adapta de (Alegria-Schaffer, A. *et al.* 2009). Las membranas se bloquean toda la noche (ON) en TBS con 0.1% Tween 20 (TBST) y 5% de BSA. Las membranas se lavan 2 veces con TBST y se incuban una hora con 1:20000 estreptoavidina-peroxidasa (Str-HRP, Invitrogen) en el bloqueante. Se lava 3 veces en TBST y la HRP se detecta con el *kit* de quimioluminiscencia *ECL plus Western blotting* (Amersham).

Purificación de moléculas biotiniladas

Tampón de lavado

100μL de Tris-HCl (pH=7.4, 1M), 20μL EDTA (0.5M), 1.168g NaCl, H_20 csp 10mL

Tampón de unión

Tris-HCl (pH=7.4, 10mM), EDTA (1mM), NaCl (2M)

Tampón de elución

NaCl (1M), MOPS (pH=7.4, 50mM), EDTA (5.0mM), β-mercaptoetanol (2.0M)

Se lavan 100μL de la suspensión de estreptoavidina-agarosa con 10 volúmenes de tampón de lavado y se centrifuga a 500g por 3 minutos a 4ºC.

La unión de las moléculas biotiniladas se realiza en 40μL de tampón de unión mas 40 μL (aprox. 2μg) del RNA total durante 2 horas a 4ºC grados.

Luego la columna se lava 3 veces con el tampón de unión.

La elución de las moléculas biotiniladas se realiza según Jenne y cols. (Jenne, A. *et al.* 1999). La columna se incuba en 200mL de tampón de elución durante 3 min a

temperatura ambiente o 2 minutos a 95°C. Se lleva a hielo y se centrifuga a 10000g. El ARN presente en el sobrenadante se precipita en presencia de 8μL de acrilamida lineal, según las instrucciones del proveedor (Ambion) y 400μL de etanol 100%. Se incuba 15 minutos a -80ºC y se centrifuga 20 minutos a 13000g. El *pellet* se resuspende en 15μL de H_2O libre de nucleasas.

Purificación de moléculas marcadas con BrUTP

La purificación se realiza por unión de las moléculas marcadas al anticuerpo monoclonal anti-BrUTP (Sigma) para luego purificar en columna de Proteína G (Protein G Sepharose 4B Fast Flow, Sigma). El protocolo fue adaptado de Ohtsu y cols. (Ohtsu, M. *et al.* 2008).

Se lavan 20μL de columna de Proteína G con 1mL de PBS/0.1%BSA (PBS+). Se recupera la columna por centrifugación durante 2 minutos a 1000g. La columna se resuspende en 100μL de PBS+ y se incuba con 2μL de anticuerpo (en 50% glicerol) durante 1 hora a temperatura ambiente.

La columna se lava 3 veces con 500μl PBS+ y se resuspende en 100μL de PBS+. El ARN se desnaturaliza a 80ºC durante 10 minutos y 50μL (aprox. 2.5μg) son incubados ON a 4ºC en presencia de inhibidor de RNAsa (*RNase guard*, Promega). Luego la columna es lavada una vez con 0.5mL de PBS+ y 1μl de *RNase guard* y 3 veces con 0.8mL de PBS. La columna se incuba con 15μL de H_2O libre de nucleasas durante 10 minutos a 80ºC para eluir el ARN. Las cantidades de columna y anticuerpo fueron variadas durante la puesta a punto de la técnica llegando a utilizarse hasta 80μL de columna y 6μL de anticuerpo intentando aumentar la recuperación.

PCR en tiempo real

Los experimentos de *run-on* y el efecto de la α-amanitina se controlaron por qRT-PCR para cuantificar genes transcritos por las diferentes ARN polimerasas. Los genes seleccionados para evaluar cada polimerasa y sobre los cuales se diseñaron cebadores específicos fueron: para la ARNPI el gen EP1 (Tb927.10.10260) (Gunzl, A. *et al.* 2003), para la ARNPII el gen de la β-tubulina (Tb927.1.2370) (Rudenko, G. *et al.* 1989) y para la ARNPIII el ARNr 5S (Tb927.8.1387) (Das, A. *et al.* 2008).

En la reacción se hace una desnaturalización inicial de 15 minutos a 95ºC y 35 ciclos de 15 segundos a 95ºC - 60 segundos a 60ºC. La curva de desnaturalización de los productos obtenidos se realiza durante 1 minuto a 95ºC.

La reacción se detecta con la mezcla comercial *QuantiTect SYBR Green PCR Master Mix* (Qiagen) que contiene la sonda *SYBR Green*, la enzima *Hot Start* Taq ADN polimerasa, dNTPs y un tampón adecuado. Las reacciones fueron llevadas a cabo en un volumen total de 20µL conteniendo los cebadores en una concentración de 0,2µM. El cDNA fue diluido 1/10 para usar como molde en la reacción.

Cebadores

Tb_qb-Tub_f:	5’	ATGGGTACGCTGCTCATCTC	3’
Tb_qb-Tub_r:	5’	GGATGGGATGATGGAGAAAG	3’
Tb_qEP1_f:	5’	TAAGGGAGGCAAAGGCAAAG	3’
Tb_qEP1_r:	5’	TCAGTGCCATTGGTATCGTC	3’
Tb_q5S_f:	5’	CATATCCCGTCCGATTTGTG	3’
Tb_q5S_r:	5’	CATCACTGATCGCCGTACTAAC	3’

Los cebadores fueron controlados realizando una corrida con diluciones seriadas de ADN genómico de *T. brucei* como molde. Las eficiencias calculadas para todos los pares se encuentran dentro del rango aceptable para realizar las comparaciones.

3.2.2 Análisis del traductoma en Epimastigotas de *T. cruzi*

Las proteínas son generalmente los efectores últimos de las funciones celulares y, por lo tanto, el proteoma define en gran medida sus características biológicas. Sin embargo, la caracterización del mismo sigue siendo una tarea compleja. Aunque los avances en las tecnologías de espectrometría de masas a gran escala permiten actualmente aproximarnos de forma cuantitativa a los niveles de un porcentaje relativamente alto de proteínas presentes en un extracto (Brewis, I. A. *et al.* 2010; Ong, S. E. 2012), estas metodologías siguen siendo más complejas y menos poderosas (principalmente en cuanto a su sensibilidad) que los análisis a nivel de los ácidos nucleicos. Esto es, hoy en día, aún más cierto luego de la introducción de los secuenciadores de última generación, los cuales han aumentado la capacidad de producción de datos en varios ordenes de magnitud, reduciendo el precio por base analizada (Metzker, M. L. 2010). Es así que, muchos de los análisis de expresión génica que se han realizado a nivel global han sido obtenidos de experimentos de transcriptómica, en los que los ARN de un organismo o tipo celular son extraídos y retrotranscritos a ADN copia para su posterior hibridación en un microarreglo de ADN o secuenciado. Los secuenciadores de segunda generación permiten además cuantificar los niveles de ARN presentes en la muestra, asumiendo que esto es proporcional a la cantidad de lecturas que se obtienen de un determinado mensajero en el experimento (Mutz, K. O. *et al.* 2013). Los estudios demuestran que los estimativos de cantidad de ARN son muy buenos cuando los datos son correlacionados con experimentos de qRT-PCR, que es la técnica de referencia (Bullard, J. H. *et al.* 2010). Aunque los niveles de ARN son un buen estimador de la cantidad de proteína en las células, los procesos de regulación post transcripcional cada vez más se consideran puntos importantes de control y por lo tanto la correlación simple entre transcriptoma y proteoma no es fiable en todos los casos (Glisovic, T. *et al.* 2008). Entre otros motivos, esta limitación motivó hace algunos años al grupo de Wiessman (Ingolia, N. T. *et al.* 2009), a desarrollar una metodología que permite estimar el traductoma de una célula. Este nuevo término, refiere a la descripción cualitativa y cuantitativa de la fracción de los ARN mensajeros totales que están siendo traducidos en un determinado momento en la célula. Esta técnica en teoría debería permitir aproximarnos mejor a las cantidades de proteínas celulares (Ingolia, N. T. *et al.* 2009; Ingolia, N. T. 2010; Ingolia, N. T. *et al.* 2011; Ingolia, N. T. *et al.* 2012). La metodología consiste en la purificación de la fracción que se encuentra unida a polisomas para luego realizar un ensayo de protección a nucleasas. Los fragmentos obtenidos, denominados huellas ribosomales (*ribosome footprints*, RFPs),

son posteriormente secuenciados y mapeados a los transcritos, obteniéndose un perfil de huellas ribosomales sobre cada mensajero (lo cual da el nombre *ribosome profiling* a la técnica) (Ingolia, N. T. *et al.* 2009). Por lo tanto, este método permite hacer uso de las ventajas comentadas anteriormente que presenta la técnica de secuenciado masivo de ácidos nucleicos.

Si consideramos que, como discutimos en la introducción, los tripanosomátidos regulan los niveles proteicos fundamentalmente a nivel post transcripcional, la técnica de perfiles ribosomales parece especialmente buena para obtener mejores estimativos de la expresión génica con respecto a los estudios transcriptómicos y ahondar en los mecanismos de la expresión génica a nivel de la traducción.

Con esta hipótesis, desarrollé el análisis *in silico* del trabajo realizado en colaboración con el grupo del Dr. Dallagiovanna con sede en el Instituto Carlos Chagas de Curitiba, Brasil y con el grupo de Dr. Sotelo del Instituto de Investigaciones Biológicas Clemente Estable, en el cual nos propusimos estudiar el traductoma de las formas epimastigotas de *T. cruzi*.

Material suplementario en:

http://lim.fcien.edu.uy/tesis/smircich/Mat_suplementario/

Tratamiento inicial de los datos

Filtrado de las lecturas

Un total de aproximadamente 179 millones y 590 millones de lecturas se obtuvieron en 3 experimentos independientes para ARN total (transcriptoma) y RFPs (traductoma) respectivamente. Luego del filtrado por calidad y largo (Ver sección Métodos), se obtuvieron un total de 86 millones (48,32%) y 217 millones (36,87%) de lecturas para cada conjunto de datos. La mejora obtenida en la calidad, para una de las réplicas del traductoma, se muestra en la Figura 3.2.2.1, obteniéndose resultados similares en todos los casos.

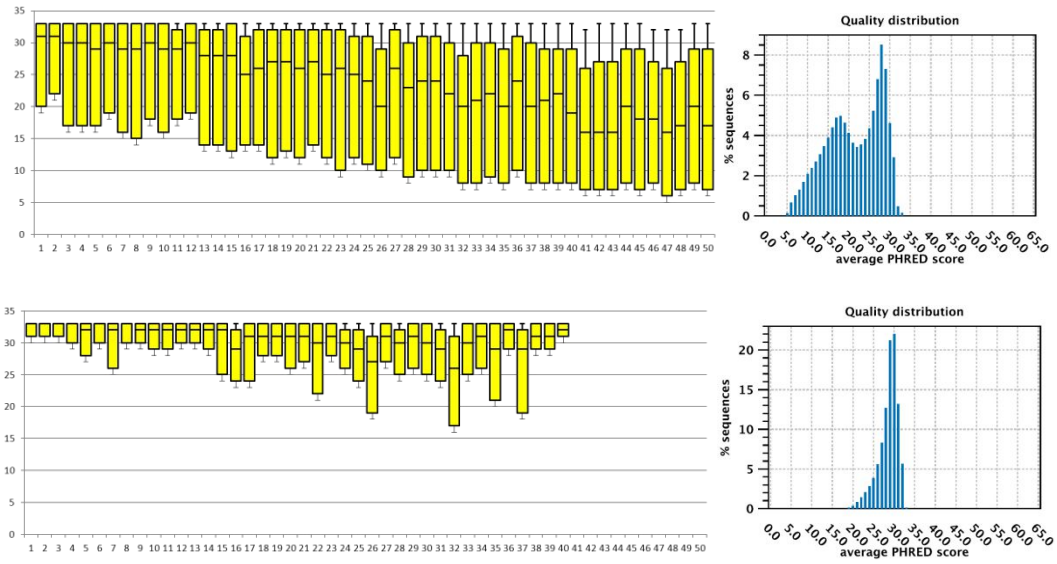


Figura 3.2.2.1 Panel izquierdo: *Boxplot* mostrando la distribución de valores *phred* de calidad en la población de bases en cada posición de las lecturas. Panel derecho: Histograma de la distribución de la calidad promedio de la población de lecturas. Panel superior: Antes del filtrado. Panel inferior: Luego del filtrado.

Los resultados de la alineación (mapeo), de las lecturas de buena calidad, a los transcritos de *T. cruzi* se resumen en la Tabla 3.2.2.1. Teniendo en cuenta que no existen diferencias significativas entre los haplotipos, se decidió continuar el trabajo con el mapeo correspondiente al haplotipo Esmeraldo.

				Secuencias Mapeadas (sin ARNr)	
	Réplica	**Número de Secuencias Iniciales**	**Filtro por calidad**	**No-Esmeraldo**	**Esmeraldo**
ARN Total	1	52467119	24759597 (47,19%)	9992192 (40,36%)	9120565 (36,84%)
	2	60509425	30568910 (50,52%)	12225673 (40,00%)	11170293 (36,54%)
	3	66452818	31375032 (47,21%)	12161617 (38,76%)	11076825 (35,30%)
	Total	**179429362**	**86703539 (48,32%)**	**34379482 (39,65%)**	**31367683 (36,18%)**
RFP	1	182288296	67791863 (37,19%)	4882181 (7,20%)	5259409 (7,76%)
	2	182345690	75657234 (41,49%)	5113385 (6,76%)	6241503 (8,25%)
	3	225383406	74092568 (32,87%)	6063146 (8,18%)	6222611 (8,40%)
	Total	**590017392**	**217541665 (36,87%)**	**16058712 (7,31%)**	**17723523 (8,15%)**

Tabla 3.2.2.1 Resumen de las cantidades de lecturas obtenidas en cada paso.

Luego de obtenidos los mapeos se realizó el cálculo de RPKM para cada gen (Ver sección Métodos). Las réplicas biológicas fueron evaluadas determinando su coeficiente de correlación (Tabla 3.2.2.2) encontrándose una buena reproducibilidad. De todas formas, observamos que el experimento de traductoma presentó una mayor variabilidad, fundamentalmente en los genes con un menor conteo de huellas (ver más adelante), seguramente reflejo de la menor cobertura obtenida en el experimento.

	Transcriptoma			Traductoma		
Réplica	1	2	3	1	2	3
1	1	0,991	0,995	1	0,993	0,963
2		1	0,994		1	0,964
3			1			1

Tabla 3.2.2.2 Matriz de correlación para las réplicas de ambos experimentos.

Los valores de RPKM promedio fueron los utilizados para estudios posteriores como estimadores de la cantidad de ARN de cada gen presente en las muestras (Mortazavi, A. *et al.* 2008).

Filtrado de los genes a analizar

Dado que las réplicas presentan un grado de variabilidad, la reproducibilidad de los datos fue estudiada específicamente para cada gen, con el fin de obtener un subconjuto que presente baja variabilidad entre las réplicas. El parámetro utilizado fue calculado dividiendo la desviación estándar del número de lecturas en cada réplica entre el número total de lecturas sobre cada gen. Este valor fue graficado en función de la cantidad de lecturas, de forma similar a lo realizado por otros autores (Ingolia, N. T. *et al.* 2009) (Figura 3.2.2.2). A partir de este gráfico se definió como criterio de aceptación: un valor de 0,2 como máxima variación aceptable y 20 lecturas mínimo mapeadas.

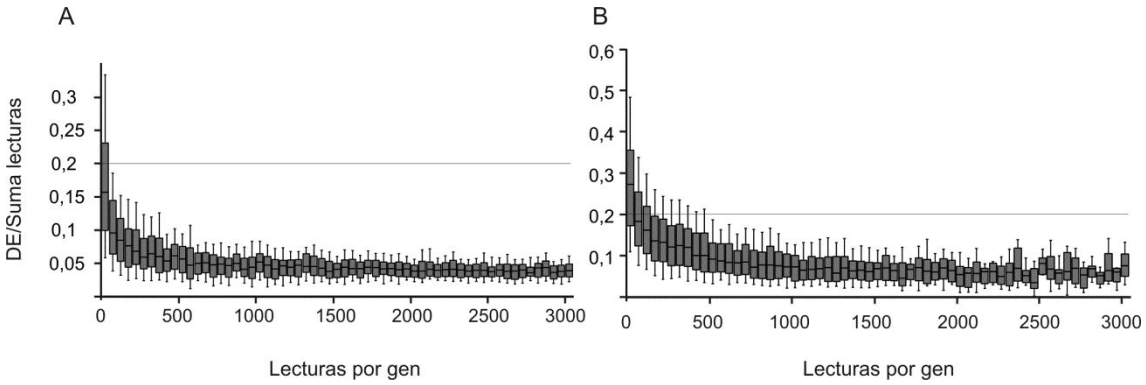


Figura 3.2.2.2 Variabilidad de las estimaciones de expresión. Para cada gen se calculó la desviación estándar (DE) de las réplicas sobre la cantidad total lecturas, para luego agruparlos en ventanas de a 50 lecturas. Los *boxplot* corresponden a la distribución de los valores de este parámetro para los genes pertenecientes a cada ventana. La línea indica el valor máximo de variación definido. A: Resultados obtenidos para el transcriptoma. B: Resultados obtenidos para el traductoma.

Los genes que pasaron los criterios de calidad en ambas muestras fueron seleccionados para continuar con la caracterización. Los números se esquematizan en la Figura 3.2.2.3.

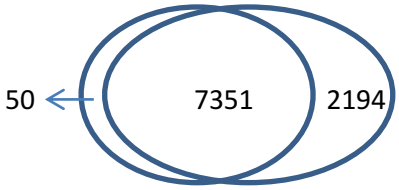


Figura 3.2.2.3 El diagrama de Venn muestra el resultado del filtrado de calidad de los genes. La intersección representa el número de genes que superan los criterios de calidad en ambos experimentos. A la derecha se representan los 2194 que superan este criterio en el experimento de transcriptómica pero no en el de traductómica. A la izquierda se representan los 50 genes que lo superan en el traductoma pero no en el transcriptoma.

Por último, dado el elevado número de familias multigénicas expandidas (DGF, GP63, MASP, RHS, TcMUCs, TS) y pseudogenes que contiene el genoma de *T. cruzi* (ver sección Introducción y (El-Sayed, N. M. *et al.* 2005)), estas secuencias fueron eliminadas del análisis general con el fin de evitar posibles sesgos en el análisis. Esto fue realizado mediante el uso de *scripts* escritos en bash y python. Luego de este último filtrado contamos con 6575 genes con los cuales fueron realizados todos los análisis que siguen a continuación, a menos que se aclare lo contrario.

Análisis de periodicidad en el mapeo de las huellas ribosomales

Se espera que las lecturas obtenidas en los experimentos de huellas ribosomales tengan un patrón de mapeo distinto al observado en el transcriptoma. Esto es así debido a que la zona que es protegida por el ribosoma mientras éste está catalizando la formación de los enlaces peptídicos, se desplaza 3 bases corriente abajo luego de la translocación. Las zonas protegidas son las mismas para todos los ribosomas ya que esto depende directamente del marco de lectura del mensajero particular. Consecuentemente, el patrón de huellas ribosomales sobre los mensajeros debería presentar un desfasaje de 3 nucleótidos entre las huellas. En cambio, en el ARN desnudo que se utiliza para secuenciar el transcriptoma, el inicio de cada lectura va a estar dado únicamente por un patrón azaroso de sitios de corte y ligado de adaptadores y, por lo tanto, las lecturas deberían estar separadas 1 nucleótido. Este fenómeno fue verificado por primera vez por Ingolia y cols. en el trabajo pionero de esta metodología (Ingolia, N. T. *et al.* 2009) y es uno de los resultados que demuestra que la técnica está realmente capturando la dinámica del ribosoma sobre el ARNm.

Los resultados del análisis de periodicidad en nuestros datos (Ver sección Métodos) se muestran en la Figura 3.2.2.4.

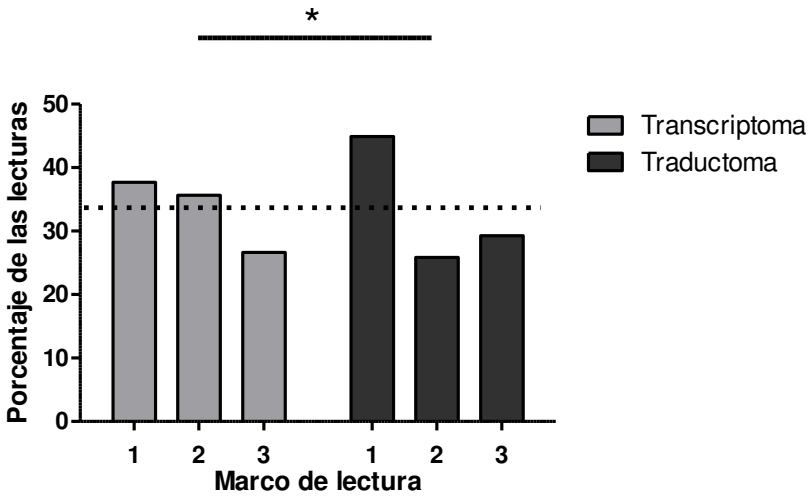


Figura 3.2.2.4 Periodicidad del mapeos de las huellas ribosomales. El gráfico de barras muestra los porcentajes de lecturas que inician en cada marco de lectura para el traductoma de *T. cruzi* y el transcriptoma usado como control. La línea punteada marca el valor esperado de 1/3 de lecturas sobre cada marco.

Para el caso del traductoma se observa que la tendencia de las lecturas es a comenzar en el primer marco de lectura, lo cual es compatible con la dinámica de traslocación del ribosoma de 3 en 3 sobre el mensajero. Interesantemente, en el marco de lectura 2 es el menos representado, fenómeno que fue observado recientemente por Michel y cols. en humanos (Michel, A. M. *et al.* 2012). En el caso del ARN total estos fenómenos no aparecen. Análisis de *chi* cuadrado muestran que ambas poblaciones son significativamente diferentes entre sí ($p < 0,05$). Este resultado nos permite concluir que nuestro protocolo experimental y de tratamiento de datos está funcionando de manera correcta y que las huellas mapeadas no son fragmentos al azar de los mensajeros, sino que reflejan el pasaje del ribosoma sobre el transcrito.

Correlación transcriptoma-traductoma-proteoma

Como comentamos al inicio de la sección, se espera que los niveles de expresión génica calculados a partir de las huellas ribosomales sean un mejor reflejo del proteoma celular. Para realizar esta comprobación estudiamos las correlaciones de los niveles de expresión calculados a partir de nuestros datos de transcriptoma y traductoma entre sí y, a su vez, con los datos disponibles de proteómica cuantitativa publicados por (de Godoy, L. M. *et al.* 2012). En la Figura 3.2.2.5 se muestra el

resultado de la correlación obtenida para los datos normalizados del transcriptoma y el traductoma de epimastigotas de *T. cruzi*.

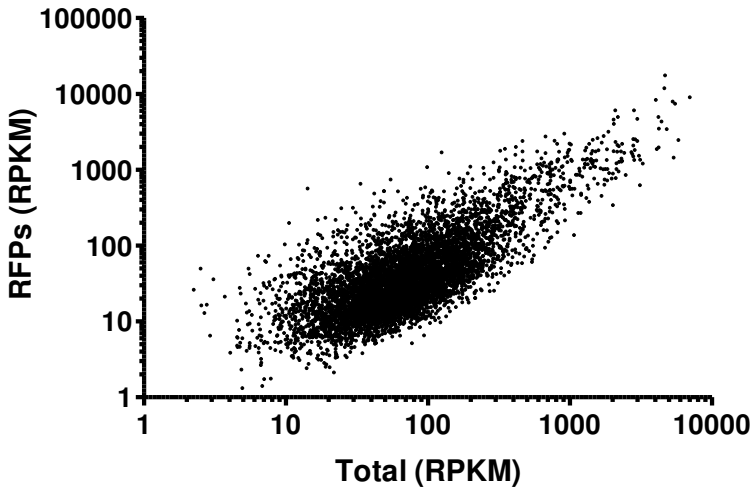


Figura 3.2.2.5 Gráfico de dispersión mostrando la correlación entre los valores de RPKM de transcriptoma y traductoma para cada gen. Los ejes se muestran en escala logarítmica para mejorar la visualización. Coeficiente de correlación de Pearson de 0,78.

Las correlaciones de ambos con el proteoma se muestran a continuación (Figura 3.2.2.6). Un total de 815 genes fueron analizados, resultantes de la intersección entre nuestros datos con los del proteoma cuantitativo.

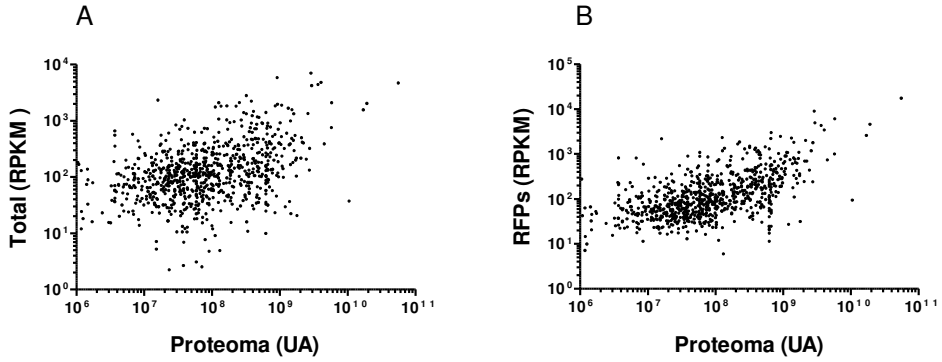


Figura 3.2.2.6 Gráfico de dispersión mostrado la correlación entre los valores de RPKM de transcriptoma y traductoma en función de los valores obtenidos en experimentos de proteómica cuantitativa. Los ejes se muestran en escala logarítmica para mejorar la visualización. A: Trasncriptoma en función del proteoma. Coeficiente de correlación de Pearson de 0,41. B: Traductoma en función del proteoma. Coeficiente de correlación de Pearson de 0,80.

Los resultados de las correlaciones muestran claramente que si bien el transcriptoma y el traductoma tienen una fuerte correlación entre sí, los valores obtenidos en el traductoma explican mejor los niveles de proteínas medidos por los experimentos proteómicos. Por lo tanto, el uso de las huellas ribosomales es una herramienta útil para realizar estudios de expresión a nivel global. Cabe señalar que para los 815 genes analizados, el coeficiente de correlación entre transcriptoma y traductoma es de 0,76 o sea que se mantiene casi incambiado con respecto a la totalidad de los genes analizados (ver Apéndice Figura 3.2.2.12).

La mediana de los valores de RPKM para la totalidad de genes y para los estudiados en el proteoma, se presenta en la Tabla 3.2.2.3

	Transcriptoma (RPKM)	Traductoma (RPKM)
Todos	66,63	33,14
Proteoma	111,21	95,87

Tabla 3.2.2.3 Mediana de los RPKM para las muestras de transcriptoma y traductoma para todos los genes analizados y para el grupo que fue cuantificado mediante proteomica.

Podemos observar que, el experimento proteómico es capaz de reportar las proteínas de alta expresión tanto a nivel de transcripción como de traducción. Nuestros datos muestran que la diferencia es más marcada cuando se considera el valor de RPKM del traductoma (ver en la siguiente sección la discusión sobre su eficiencia traduccional). Es así que los genes que se detectan en el proteoma tienen en general un nivel de expresión calculado en el traductoma alto con respecto a los que se observa en el transcriptoma (Apéndice Figura 3.2.2.12).

Regulación de la eficiencia traduccional

Uno de los puntos de control de la expresión génica de relevancia en tripanosomátidos es el inicio de la traducción (Clayton, C. E. 2002). Esto hace que el estudio de la eficiencia traduccional de los mensajeros sea una pregunta especialmente interesante en este modelo. La eficiencia fue definida como la relación entre la tasa transcripcional (en RPKM) observada en traductoma sobre la observada en el transcriptoma para cada gen de la misma manera que (Ingolia, N. T. *et al.* 2012).

En la Figura 3.2.2.7 podemos observar que los valores de eficiencia son variables para los diferentes genes, teniendo un valor medio de 0,83 y una mediana de 0,50.

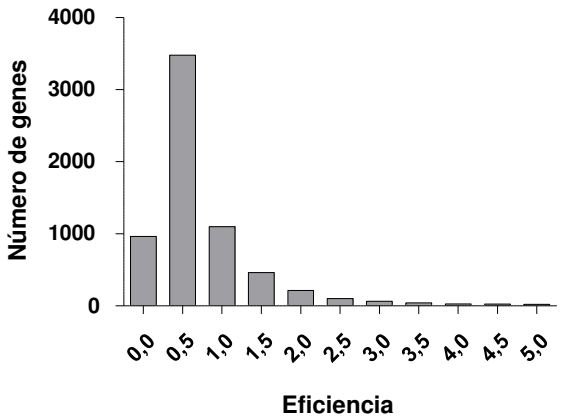


Figura 3.2.2.7 El histograma muestra la distribución de los valores de eficiencia de traducción para el conjunto de genes analizados. El eje de eficiencia fue cortado en 5 para mejorar la visualización (el parámetro presenta un valor máximo de 39).

Es interesante señalar que, como vimos antes, los valores de la cuantificación del proteoma celular tienen una distribución diferente. De hecho, si calculamos la mediana de la eficiencia para esos genes vemos que ésta es aproximadamente 0,9, lo cual está por encima de los valores más frecuentes. Esta tasa es incluso mayor que la que se observa para los genes que codifican para las proteínas ribosomales (mediana de la eficiencia 0,7), que son proteínas que no tienen regulación negativa y son de alta expresión. Esto puede ser reflejo de que las proteínas de alta expresión presentan en muchos casos, una eficiencia traduccional alta. El análisis de ontología sobre los genes de los productos descritos en el análisis de proteoma detectó una sobrerrepresentación de las familias de proteínas de respuesta a estrés general y oxidativo (ES > 1,9, parámetro que representa el enriquecimiento, ver Materiales y Métodos). En el apéndice se muestran las tablas correspondientes a los 3 primeros agrupamientos (Figura 3.2.2.13).

Para continuar con el análisis de los genes que tienen valores extremos de la distribución de eficiencias, se tomaron los genes que pertenecen al percentil 1 y 99 de la población, para baja y alta eficiencia respectivamente. Para comparar, se utilizaron también los genes que superan los percentiles 5 y 95 como criterios de corte. Los genes que caen en estas categorías fueron clasificados para detectar funciones similares según su anotación en la ontología génica.

Los genes correspondientes al 5% más alto de eficiencia traduccional (eficiencia > 2,38) fueron sometidos al análisis. Interesantemente, la categoría de genes que resultó más sobrerrepresentada fue la que corresponde a proteínas que presentan el dominio RRM (*RNA Recognition Motif*) de unión a ARN (ES = 2,71). Además, encontramos genes para proteínas de unión a ATP (ES = 3,09), relacionadas con el anabolismo de nucleótidos (ES = 2,39) y, por último, para enzimas con actividad cistein-proteasas (ES = 1,3) (Apéndice Figura 3.2.2.14). El análisis con los genes que superan el percentil 99 (65 identificadores, eficiencia > 5,54) mostró que los que codifican las proteínas con RRM son el único grupo sobrerrepresentado según nuestros criterios (Apéndice Figura 3.2.2.15). Estas proteínas son muy abundantes en tripanosomátidos ya que intervienen en los procesos de regulación post transcripcional y, por lo tanto, son fundamentales en estos parásitos (Perez-Diaz, L. *et al.* 2013).

Con este resultado, nos preguntamos cuál sería el perfil de traducibilidad de la familia en su conjunto. Para esto analizamos los 56 miembros de la familia de proteínas con RRM presentes en el haplotipo Esmeraldo (De Gaudenzi, J. *et al.* 2005). El cálculo de la mediana de la eficiencia de traducción resulto de 0,89, lo cual es significativamente mayor que el valor para el conjunto de todos los genes de este haplotipo (prueba Wilcoxon $p<0,05$). Si bien, los niveles de transcripción de estos genes no son significativamente más altos que los del resto de los mensajeros (aunque un aumento en la mediana es observable, este no es estadísticamente significativo), los niveles en la fracción ribosomal si lo son (prueba Wilcoxon $p<0,05$), lo cual explica la mayor eficiencia de traducción observada (Figura 3.2.2.8). Además, vemos como el rango intercuantil de eficiencias alcanza valores más altos en los genes para proteínas con RRM, indicando que un número importante de miembros que tienen valores altos.

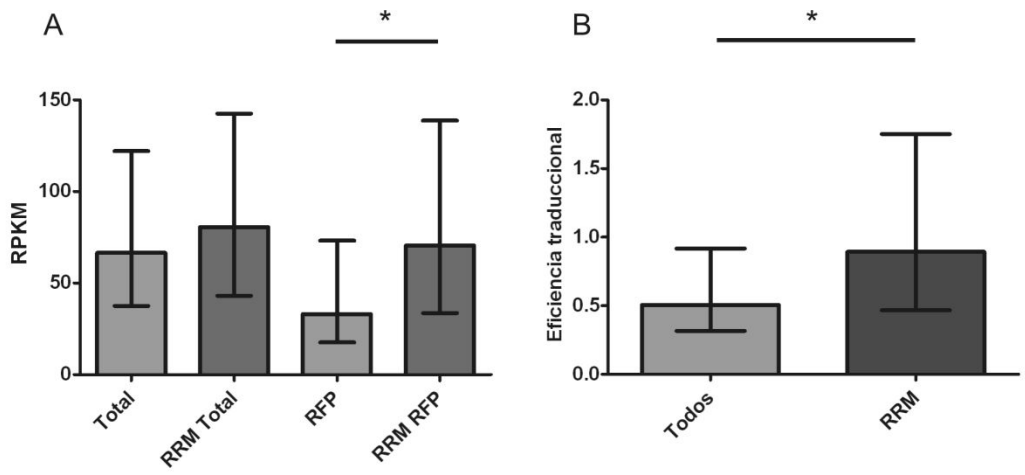


Figura 3.2.2.8 A: Comparación de los niveles de mensajeros medidos en el transcriptoma con respecto a los medidos en el traductoma para la población total de genes y las proteínas de la familia RRM. La altura de las cajas representa la mediana y las líneas marcan los valores intercuartil. B: Comparación de la eficiencia traduccional. Los asteriscos indican diferencias significativas (prueba Wilcoxon $p < 0,05$).

Estos resultados podrían indicar que los genes que codifican estas proteínas regulatorias han evolucionado mecanismos de expresión que les permiten responder rápidamente una vez que se encuentran los ARNm disponibles para ser traducidos. De hecho, esto ha sido propuesto para eucariotas superiores. En el trabajo de Mittal y cols. del 2009, los autores demuestran que los transcritos de las proteínas con RRM de levadura son poco estables y presentan una mayor ocupación de los polisomas cuando estas variables son comparadas con las del resto de los genes en ese organismo (Mittal, N. *et al.* 2009). En *T. cruzi* este resultado puede ser de especial relevancia debido a la esencialidad de los mecanismos de regulación post transcripcional y por lo tanto de las proteínas con RRM.

Otro hecho a destacar es la observación de que una alta eficiencia traduccional no implica necesariamente una elevada tasa transcripcional, por ejemplo no se encuentran en estos grupos las proteínas ribosomales que si aparecen cuando realizamos los análisis con los valores mas altos que el percentil 95 tanto en transcriptoma (ES = 37,58) (Apéndice Figura 3.2.2.16) como en traductoma (ES = 34,99) (Apéndice Figura 3.2.2.17).

Llaman la atención los 50 genes que descartamos del análisis general por su alta variabilidad en el transcriptoma, pero que sin embargo sus valores son confiables

en el traductoma. Esto deriva de que posiblemente los valores en el primero son bajos y, por lo tanto, con alta incertidumbre. Cuando analizamos la eficiencia de estos genes (aunque está claro que éste es un número aproximado ya que desconocemos con certeza el valor en el transcriptoma), ésta alcanza un valor de 12, siendo más alta incluso que el grupo del percentil 99. Dentro de este listado no se encontraron grupos de funciones comunes sobrerrepresentados (Apéndice Figura 3.2.2.18).

Con respecto a los genes que presentan una eficiencia de traducción baja, no pudimos encontrar grupos que tuvieran un ES significativo. Cuando tomamos el 5% de los genes con menor eficiencia se pueden observar algunas categorías ontológicas como las que comprenden a los genes que codifican las proteínas kinasas y proteínas de membrana con valores de $p < 0,05$ aunque el FDR no es significativo.

El otro grupo de baja eficiencia traduccional considerado, consistió en los genes descartados del análisis general por no presentar valores confiables en el traductoma, pero que sí pudieron ser cuantificadas en el transcriptoma. De estos genes se tomaron aquellos que presentaran una eficiencia de traducción menor al percentil 5 y que además su valor en el traductoma estuviera en el rango menor al percentil 10. Llamativamente, los genes que codifican las proteínas trans-sialidasas (TS) mostraron que están muy sobrerepresentadas en este grupo (ES 9,88) (Apéndice Figura 3.2.2.19). Esto es interesante ya que estas proteínas son diferencialmente expresadas en el ciclo de vida, siendo características del estadío tripomastigota, aunque hay evidencia de que miembros específicos se expresan en otros momentos del ciclo de vida. Su abundancia en los genes de baja eficiencia traduccional, puede estar indicando que uno de los mecanismos implicados en la expresión diferencial es la regulación de su traducibilidad. Existe evidencia de que esta familia posee una fuerte regulación mediada por regiones conservadas en la región 3'UTR (Jager, A. V. *et al.* 2008; Freitas, L. M. *et al.* 2011).

Esta última observación justifica además la suposición de que muchos de los genes que estamos observando con bajas eficiencias traduccionales posiblemente sean regulados a nivel de estadío y, por lo tanto, los datos generados en este trabajo pueden ser usados como referencia para futuros proyectos en donde se busquen genes de expresión diferencial.

Análisis del uso de codones sinónimos

Para varios organismos se ha demostrado que el uso de los codones sinónimos varía para los diferentes genes, habiendo codones que son preferidos sobre otros. Uno de los mecanismos que puede afectar este sesgo es la selección de los sinónimos que presentan una mejor eficiencia o una mejor fidelidad traduccional (Ikemura, T. 1981). Esta selección, se espera que actúe de forma más notoria en los genes de mayor expresión y si la hipótesis es correcta, el fenómeno debería estar relacionado directamente con la dinámica traduccional de los mensajeros. Debido a que en este trabajo contamos con las tasas transcripcionales y traduccionales decidimos estudiar la validez de estas afirmaciones en *T. cruzi*.

Varios índices han sido utilizados para medir el sesgo en el uso de codones de las secuencias codificantes. En particular, el índice de adaptación de codones (c*odon adaptation index,* CAI) mide la desviación de un gen particular con respecto a un grupo de referencia (Sharp, P. M. *et al.* 1987). Para describir el grado de apartamiento de cada gen con respecto al uso óptimo de codones del organismo, el grupo de referencia consiste en los genes para proteínas de alta expresión, ya que se espera que estos presenten el mayor grado de optimización en el uso de los sinónimos.

En nuestro caso el índice CAI para los genes de *T. cruzi* fue calculado tomando como set de referencia el 1% de los genes de mayor tasa transcripcional (RPKM mayor al percentil 99). Las correlaciones de los datos de transcriptoma y traductoma con respecto al CAI fueron estudiadas y los resultados se muestran en la Figura 3.2.2.9.

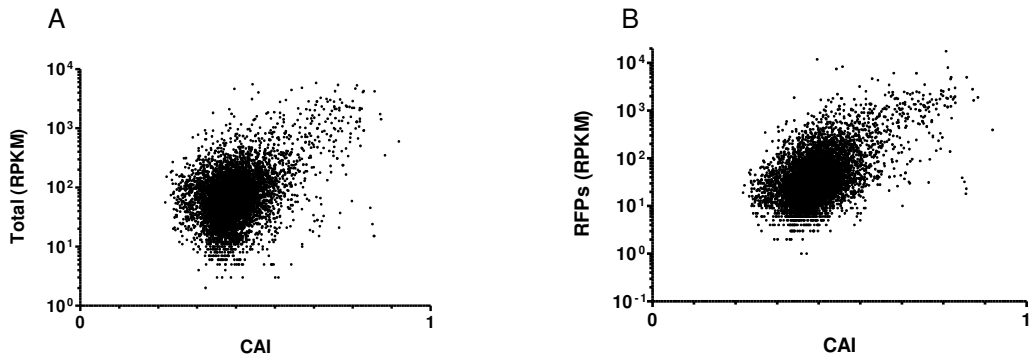


Figura 3.2.2.9 Gráficos de dispersión que muestran la correlación del índice CAI con los datos de expresión obtenidos. El eje de las Y se muestra en escala logarítmica para mejorar la visualización. A: transcriptoma (coef. Spearman=0,29) B: traductoma (coef. Spearman=0,45).

Los coeficientes de correlación muestran que los datos de traductoma tienen una mejor correlación con el CAI que los de transcriptoma (0,45 y 0,29 respectivamente), aun cuando se usaron las secuencias correspondientes a los genes de alta transcripción como referencia. Cuando las secuencias de referencia son cambiadas por las de los genes que codifican para proteínas ribosomales, los resultados son similares siendo los valores de correlación de 0,28 y 0,45 para transcriptoma y el traductoma respectivamente.

Haciendo un estudio teórico del número de ARNt en el genoma y genes repetidos en tándem, Horn propone la existencia de una optimización del uso de codones en los genes de alta expresión (Horn, D. 2008). Los datos experimentales aquí analizados muestran que efectivamente la tasa traduccional es un factor selectivo que sesga el uso de codones.

Otro punto interesante a considerar está relacionado con la influencia de los diferentes codones sinónimos sobre la dinámica de traducción. La hipótesis preponderante afirma que debido a que el uso de codones está adaptado al conjunto específico de tRNAs presentes en el organismo, los codones menos frecuentes tienden a pausar los ribosomas sobre los mensajeros (Lesnik, T. *et al.* 2000). Este fenómeno fue contradicho por Ingolia y cols. en su trabajo realizado en modelo de mamíferos (Ingolia, N. T. *et al.* 2011). Sin embargo, estudios posteriores donde se reanalizan los mismos datos, muestran que esta correlación sí es observable (Dana, A. *et al.* 2012). Para aproximarnos al análisis de esta cuestión en *T. cruzi*, estudiamos regiones donde las RFPs se encuentran sobrerrepresentadas con respecto al resto del gen. Estas regiones se pueden interpretar como sitios donde los ribosomas están pausados (o al menos enlentecidos) y, por lo tanto, generan un mayor número de huellas.

En este análisis se trabajó con un grupo de genes de buena cobertura (ver Materiales y Métodos) y se buscaron los codones donde la misma resultara ser mayor de 10 veces la mediana en el gen. La frecuencia relativa de cada sinónimo en las regiones con alta cobertura fue comparada a la frecuencia de los codones general de los genes estudiados.

Los resultados muestran que, para la mayor parte de los sinónimos (Apéndice Tabla 3.2.2.4), el codón mayor es el mismo en el grupo de los genes con pausas y en las regiones donde éstas se producen, no habiendo diferencias significativas de su

frecuencia. Sin embargo, para los codones que codifican histidina y lisina, el codón sinónimo más frecuentemente usado se invierte en las pausas (*chi* cuadrado p<0,05). En el caso de la histidina, el codón CAC es en general el más usado, mientras que el codón CAU se encuentra con mayor frecuencia en los sitios de pausas (Figura 3.2.2.10 A). Para la lisina, el codón AAA se enriquece sobre el AAG en los sitios de pausas (Figura 3.2.2.10 A). Por el contrario, cuando se estudia el uso de estos codones en los genes de alta expresión (set de referencia usado para el cálculo del CAI), se observa la tendencia opuesta (Figura 3.2.2.10 B).

Es interesante señalar el caso del glutamato, el cual fue descrito por Ingolia y cols. (Ingolia, N. T. *et al.* 2011), en su trabajo en células madre de ratón, como perteneciente al motivo que describen como más frecuente en pausas (secuencia PPE de aminoácidos y CC(A/T)CC(A/T)GAA a nivel de nucleótidos). En nuestros datos, la frecuencia relativa de los codones GAA y GAG presenta valores similares en las regiones de pausas (Figura 3.2.2.10 A), mientras que en el genoma (y más marcadamente en las proteínas de alta expresión) es el codón GAG el preferido (*chi* cuadrado p = 0,08) (Figura 3.2.2.10 B). Además, encontramos casos en los cuales la frecuencia del uso de los sinónimos no está invertida pero, de todas formas, se encuentra significativamente alterada en las regiones de pausas (*chi* cuadrado p < 0,05). Este es el caso del codón CUA que codifica leucina (Figura 3.2.2.10 A) que es además uno de los que tiene un valor más bajo de frecuencia a nivel general y está prácticamente excluido en los genes de alta expresión (Figura 3.2.2.10 B).

Otro caso interesante es el de la fenilalanina, en donde el codón UUU es preferido sobre el UUC en el genoma en general (Figura 3.2.2.10 B). Los sitios de pausas parecen seguir la misma tendencia estando el codón UUU enriquecido (*chi* cuadrado < 0,05) (Figura 3.2.2.10 A). Sin embargo, cuando observamos lo que ocurre en los genes que codifican para las proteínas de alta expresión vemos que el codón UUC es el preferido (Figura 3.2.2.10 B). Por lo tanto, la frecuencia de los sinónimos en los sitios de pausa vuelve a ser la contraria a la de las proteínas de alta expresión. Un caso análogo se observa para la isoleucina, donde el codón AUU se encuentra sobrerrepresentado en las regiones de pausas (Figura 3.2.2.10).

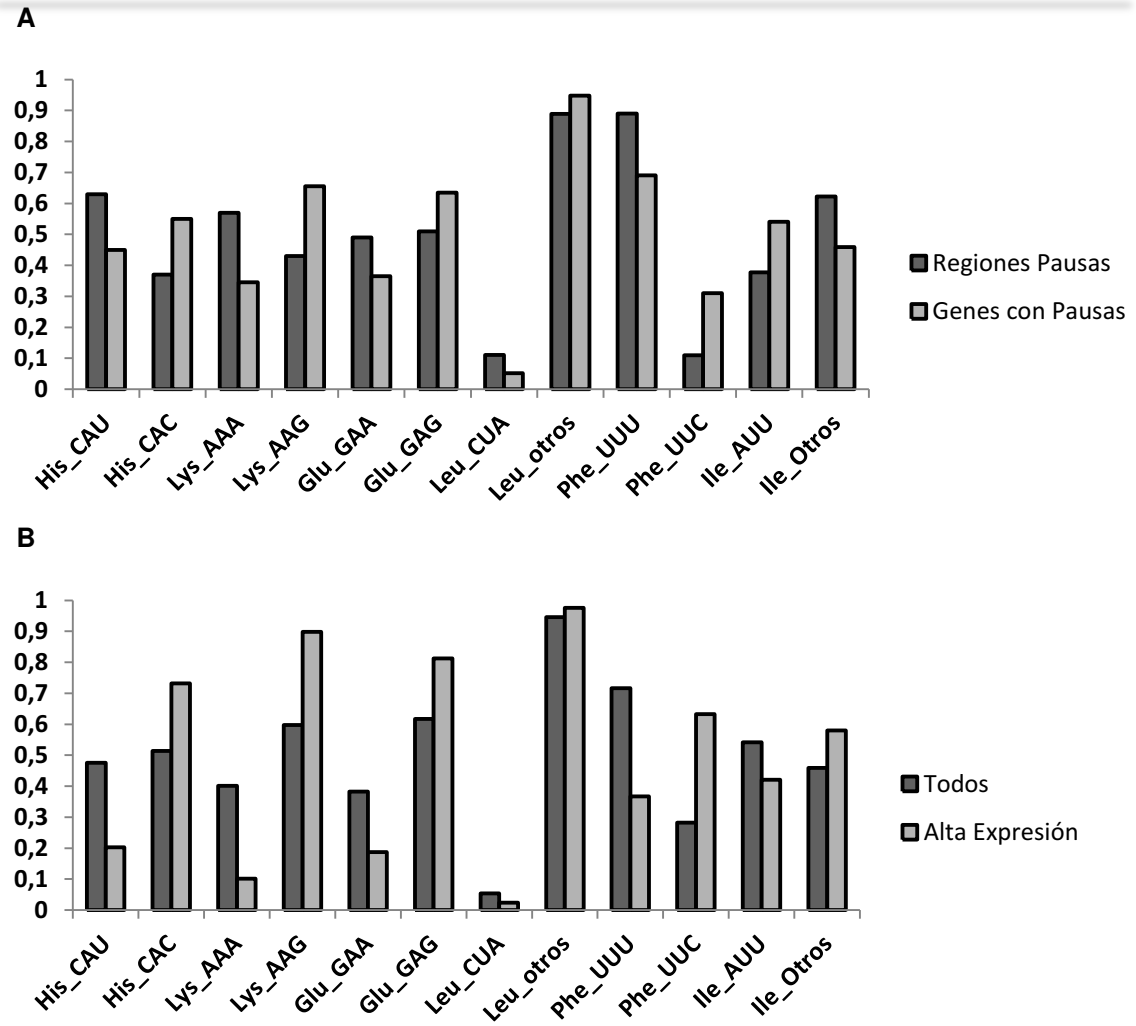


Figura 3.2.2.10 Frecuencia de cada codón sinónimo en los diferentes grupos de secuencias. Se señalan únicamente los codones en donde se observaron diferencias significativos. A: Frecuencia de los codones sinónimos en las regiones donde se observan pausas (negro) comparado con las frecuencias en general en los genes en donde se encuentran las pausas. B: Frecuencia de los codones sinónimos en la totalidad de los genes de *T. cruzi* comparada con las frecuencias en los genes de alta expresión. En el caso de la leucina y la isoleucina que están codificados por más de un sinónimo, el codón sobrerrepresentado se compara con la suma de los otros del grupo

Por consiguiente, los codones que observamos enriquecidos en los sitios de pausas, son poco frecuentes en el genoma y tienden a evitarse en proteínas de alta expresión. Como es de esperar, los cambios que se producen entre los codones son en la tercera posición, en donde los codones presentes en los genes de alta expresión son ricos en GC mientras que los enriquecidos en las pausas son ricos en AT. Esto es coherente con descripciones previas de la preferencia de codones sinónimos en *T. cruzi* (Alvarez, F. *et al.* 1994). Estos resultados, apoyan la hipótesis de que codones poco frecuentes pueden servir como sitios de detención de los ribosomas durante la traducción en *T. cruzi*.

Métodos

Obtención de huellas ribosomales

La obtención del perfil de ribosomas y su secuenciado en la plataforma Solid (Life Technologies) se llevó a cabo en Curitiba en un trabajo colaborativo entre los Dr. Dallagiovanna y Sotelo.

Brevemente, 5 x 10^9 células fueron tratadas con cicloheximida (10mg / 5x10^9 células), lisadas y la fracción de polisomas fue obtenida por fraccionamiento en un colchón de sacarosa 2M en presencia de cicloheximida. Los polisomas obtenidos en la separación fueron tratados con una endonucleasa de corte inespecífico durante 10 minutos a temperatura ambiente, para luego realizar la extracción de las moléculas de ARN (*mirVana™ miRNA Isolation Kit* de Ambion). Se seleccionaron los fragmentos de alrededor de 30pb en geles de acrilamida. Estos fragmentos, así como una fracción de ARN total extraído de epimastigotas, fueron secuenciados en la plataforma SOLiD según el protocolo estándar descrito por el proveedor (*SOLID Whole Transcriptome Analysis Kit*). Se realizaron 3 réplicas biológicas tanto de obtención de RFPs como de transcriptoma.

Filtrado de las secuencias por calidad

Como primer paso del análisis *in silico*, los adaptadores fueron removidos de las secuencias y se realizó un filtrado por calidad usando el software *CLC Genomics Workbench*. Brevemente, el algoritmo calcula la probabilidad de error de cada base según su valor *phred* de calidad reportado por el equipo. Este número comparado con un valor límite preseleccionado (0,05 en nuestro caso) para luego realizar una suma de las diferencias desde la posición 5' hacia la 3'. La región seleccionada irá desde el primer valor positivo hasta el valor máximo de esta suma. El resultado neto del proceso es la remoción de bases de baja calidad consecutivas. Luego aplicamos una restricción de largos de entre 25 y 40 nucleótidos para las RFPs y de entre 18 y 50 nucleótidos para las lecturas de ARN total.

Mapeo y cuantificación de los niveles de ARN

Las lecturas de buena calidad fueron mapeadas con el software *CLC Genomics Workbench* usando como referencia los transcritos del genoma de *T. cruzi* (haplotipos Esmeraldo y No-Esmeralo, TritrypDB versión 5.0). Para incluir una lectura en el mapeo, un 90% de la misma debe alinear con la referencia y el número de diferencias

(*mismatches*) debe ser menor a 2. Las lecturas que mapean en más de una región, son asignadas de forma proporcional a las lecturas de mapeo único sobre los genes en cuestión. Las lecturas que mapean en más de 10 lugares diferentes, son descartadas. Luego del mapeo, las secuencias correspondientes a ARN ribosomal fueron descartadas.

Una vez obtenido el número de lecturas sobre cada gen, los valores de RPKM (lecturas por kilobase de cada gen por millón de lecturas mapeadas) fueron calculados según Mortazavi y cols. (Mortazavi, A. *et al.* 2008).

El resultado de los mapeos fue exportado en formato SAM que permite obtener la información del alineamiento de cada lectura sobre la referencia (Li, H. *et al.* 2009). Los archivos SAM fueron pasados a BAM usando el paquete *SAMtools* (Li, H. *et al.* 2009) y agrupados (definiendo un *read group* para cada replica) con el paquete *picard* (http://picard.sourceforge.net) lo cual permite su visualización con el *Integrative Genome Viewer* (IGV) (Thorvaldsdottir, H. *et al.* 2013). La cobertura sobre cada base fue calculada usando el programa *bedtools* con la opción *genomecov* (Quinlan, A. R. *et al.* 2010)

Análisis de periodicidad

Este análisis fue realizado sobre un grupo de genes que tuvieran una buena cobertura a lo largo de su secuencia (ver Apéndice Figura 3.2.2.11). Este conjunto fue definido como en (Ingolia, N. T. *et al.* 2011). Para cada gen se calcula la cobertura sobre cada nucleótido para luego deslizar una ventana de 15nt no solapante. El *script* escrito en python, descarta los genes en los cuales exista una ventana (o más) en la que la mediana de cobertura sea menor a 2. Los identificadores de los genes resultantes fueron utilizados para filtrar el archivo SAM que contiene el mapeo de las lecturas sobre el genoma de referencia. Una vez obtenidos los datos de mapeo del conjunto de genes a analizar, el porcentaje de lecturas que alinean en fase con cada marco de lectura fue determinado mediante *scripts* en python en todas las réplicas.

Clasificación según la ontología génica

El proyecto de ontología génica (*gene ontology*, GO) pretende estandarizar la anotación que reciben los genes o sus productos a través de un vocabulario controlado de términos jerárquicos.

La herramienta en línea DAVID (*Database for Annotation, Visualization and Integrated Discovery*, versión 6.7), diseñada para analizar datos de genómica funcional fue usada para determinar las categorías de GO (Dennis, G., Jr. *et al.* 2003). El programa permite agrupar genes según su ontología así como comprobar si alguna de las categorías encontradas se encuentra sobrerrepresentada con respecto a lo que se espera por azar teniendo en cuenta la proporción de genes que caen dentro de esta categoría en el genoma (o una lista de referencia arbitraria). La herramienta clasifica también las proteínas según los dominios presentes en la proteína según InterPro. Consideramos un grupo sobrerrepresentado si presenta un "Puntaje de Enriquecimiento" ES>1,3 (*enrichment score,* definido como el menos logaritmo de la media geométrica de los valores p) (Huang, D. *et al.* 2009). Los valores p son calculados utilizando una variante del test exacto de Fisher (*EASE score*). Para cada categoría ontológica dentro de la agrupación, se controló el efecto del testeo múltiple sobre los valores p obteniéndose el valor de FDR (*false discovery rate*) mediante el método de Benjamini (Benjamini, Y. *et al.* 2001), aceptándose valores menores a 0,05 a menos que se especifique lo contrario. Las listas usadas como referencia fueron los genes del haplotipo Esmeraldo de *T. cruzi* o el subconjunto obtenido por los filtrados comentados anteriormente, según el caso. Dentro del paquete DAVID fueron utilizadas dos herramientas. Por un lado, se usó el "Agrupamiento por Anotación Funcional" (*functional annotation clustering*) que se centra en el análisis de las categorías de GO que están enriquecidas en la lista de genes analizados y las agrupa en función de los genes que contienen en común. Esto disminuye sustancialmente la redundancia en los reportes de categorías individuales más clásicos (Huang, D. *et al.* 2009). Además, se utilizó la herramienta de "Clasificación de Genes según Función" (*gene functional classification*) la cual se focaliza en la anotación de los genes y agrupa aquellos que tienen función similar. Ambas aproximaciones son complementarias, describiendo la primera funciones sobrerrepresentadas, mientras que la segunda detalla qué familias génicas contiene la lista de genes analizados. En este trabajo, ambos análisis dieron resultados equivalentes para las diferentes listas, por lo que sólo se presentan los resultados de la primera.

Estudio de uso de codones sinónimos

El uso relativo de codones fue estimado a partir de su frecuencia dentro del grupo de sinónimos. Los programas GCUA (*General Codon Usage Analysis*) (McInerney, J. O. 1998) e INCA (*INteractive Codon usage Analysis*) (Supek, F. *et al.* 2004) fueron utilizados para el cálculo. Este último se utilizó también para calcular el

índice CAI (*Codon Adaptation Index*) que mide la desviación del uso de codones de un gen particular con respecto a un grupo de referencia (Sharp, P. M. *et al.* 1987). La significancia estadística de las diferencias entre el uso de codones sinónimos encontrado entre las regiones con el uso de codones encontrado en los genes que las contienen, se realizó por análisis de *chi* cuadrado de contingencia. Para observar si los codones de frecuencias significativamente diferentes están sobrerrepresentados en las pausas, se calcularon las proporciones para cada sinónimo entre los sitios de pausas y las observadas en los genes.

Definición de regiones enriquecidas en RFPs en los ARNm

Para encontrar las regiones donde las huellas se encuentran sobrerrepresentadas con respecto al resto del gen, se trabajó con un grupo de genes de alta cobertura que fueron definidos de forma similar a los realizado para análisis de periodicidad. Sin embargo, el criterio fue más inclusivo y en este caso, el *script* escrito en python rechaza los genes en los cuales el promedio del los valores de cobertura en cada ventana de 15nt no fuera mayor a 2. La cobertura para cada codón fue calculada y se extrajeron los que superaran 10 veces la mediana de cobertura sobre el gen. Dada la variación en la cobertura a lo largo de los genes en los experimentos de RNAseq producida por los diferentes sesgos encontrados en la metodología, existen codones en el experimento de transcriptoma con altas coberturas. Como estos sesgos son reproducibles, estas regiones podrían encontrarse también en los mapeos de huellas y ser interpretadas como sitios de pausas de ribosomas. Por lo tanto, sólo se tuvieron en cuenta los codones provenientes de genes en los cuales el experimento de transcriptoma no presentara regiones con estas características.

Apéndice

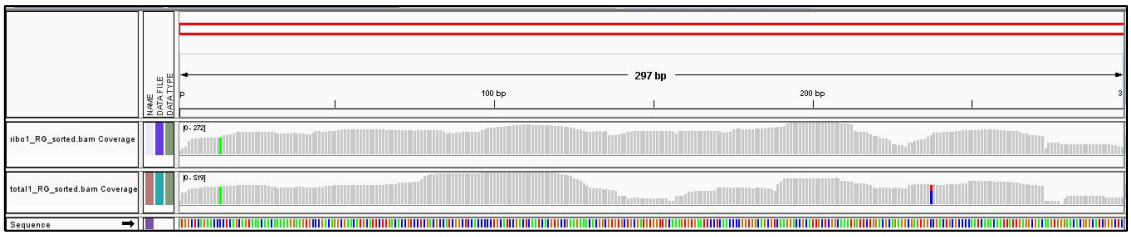


Figura 3.2.2.11 Imagen obtenida en el software IGV del gen Tc00.1047053508241.149 (cistationina β-sintasa). Este fue uno de los genes utilizados para realizar el cálculo de periodicidad de mapeo. Se aprecia la cobertura sobre toda la secuencia del mensajero, tanto en el traductoma (arriba) como en el transcriptoma (abajo).

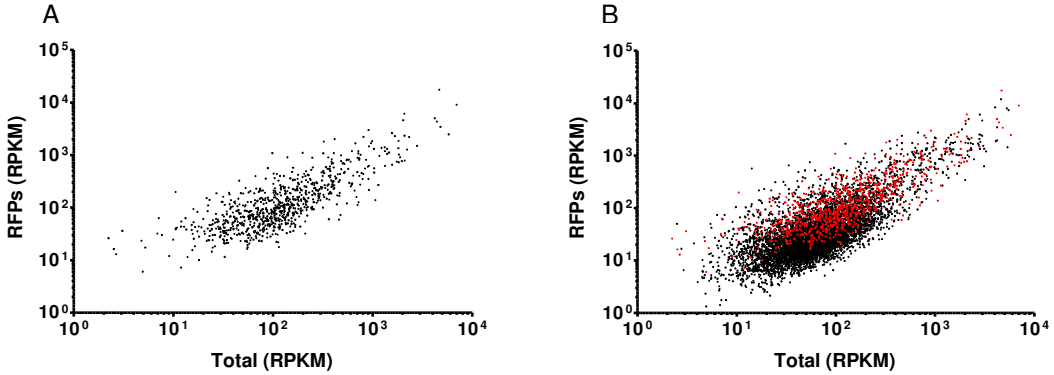
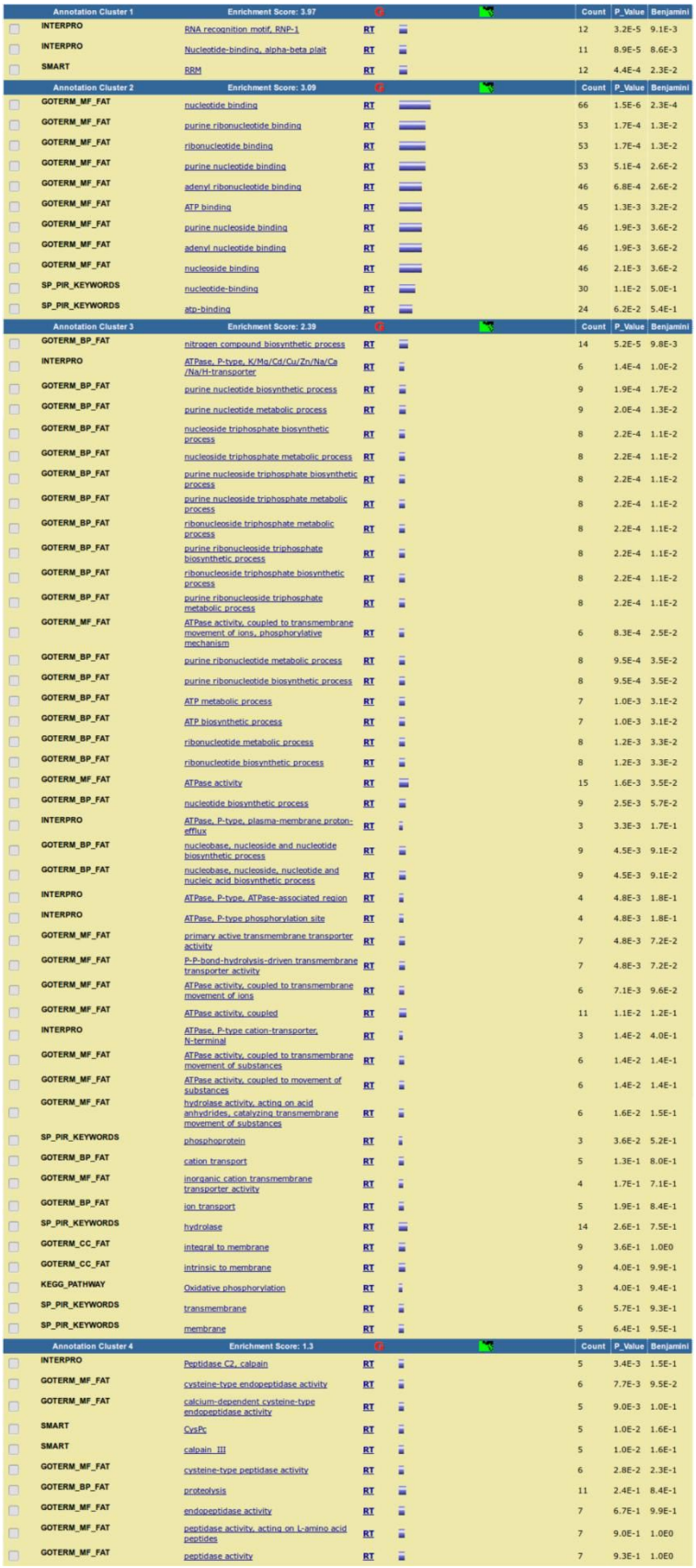


Figura 3.2.2.12 Correlación entre el nivel de RFPs y transcritos para los genes reportados en el estudio proteómico. A: Se muestran únicamente los genes reportados en el estudio proteómico. B: Se muestran todos los genes analizados y se resaltan en rojo los anteriores.

Annotation Cluster 1	Enrichment Score: 1.96			Count	P_Value	Benjamini
INTERPRO	Heat shock protein 70	RT		7	4.2E-3	5.0E-1
SP_PIR_KEYWORDS	stress response	RT		12	4.3E-3	1.4E-1
INTERPRO	Heat shock protein Hsp70	RT		7	5.0E-3	4.5E-1
PIR_SUPERFAMILY	PIRSF002581:chaperone HSP70	RT		4	3.2E-2	1.0E0
INTERPRO	Heat shock protein 70, conserved site	RT		4	5.8E-2	8.5E-1
Annotation Cluster 2	**Enrichment Score: 1.95**			**Count**	**P_Value**	**Benjamini**
INTERPRO	Thioredoxin fold	RT		13	1.5E-3	7.2E-1
GOTERM_BP_FAT	homeostatic process	RT		12	3.2E-3	2.2E-1
INTERPRO	Thioredoxin-like	RT		7	4.2E-3	5.0E-1
GOTERM_BP_FAT	cell redox homeostasis	RT		11	5.4E-3	2.5E-1
GOTERM_BP_FAT	cellular homeostasis	RT		11	6.9E-3	2.8E-1
INTERPRO	Thioredoxin domain	RT		4	1.1E-1	9.2E-1
INTERPRO	Thioredoxin, conserved site	RT		3	2.7E-1	9.9E-1
Annotation Cluster 3	**Enrichment Score: 1.9**			**Count**	**P_Value**	**Benjamini**
GOTERM_BP_FAT	cellular amino acid biosynthetic process	RT		10	1.8E-3	1.5E-1
GOTERM_BP_FAT	amine biosynthetic process	RT		10	3.4E-3	1.9E-1
GOTERM_BP_FAT	organic acid biosynthetic process	RT		11	6.9E-3	2.8E-1
GOTERM_BP_FAT	carboxylic acid biosynthetic process	RT		11	6.9E-3	2.8E-1
GOTERM_BP_FAT	glutamine family amino acid biosynthetic process	RT		4	9.2E-2	6.3E-1
GOTERM_BP_FAT	glutamine family amino acid metabolic process	RT		4	1.5E-1	7.8E-1

Figura 3.2.2.13 Resultado del agrupamiento de los 815 genes analizados con datos de proteómica cuantitativa de epimastigotas de *T. cruzi*. Datos extraídos de (de Godoy, L. M. *et al.* 2012)



Annotation Cluster 1	Enrichment Score: 3.97			Count	P_Value	Benjamini
INTERPRO	RNA recognition motif. RNP-1	RT		12	3.2E-5	9.1E-3
INTERPRO	Nucleotide-binding, alpha-beta plait	RT		11	8.9E-5	8.6E-3
SMART	RRM	RT		12	4.4E-4	2.3E-2
Annotation Cluster 2	**Enrichment Score: 3.09**			**Count**	**P_Value**	**Benjamini**
GOTERM_MF_FAT	nucleotide binding	RT		66	1.5E-6	2.3E-4
GOTERM_MF_FAT	purine ribonucleotide binding	RT		53	1.7E-4	1.3E-2
GOTERM_MF_FAT	ribonucleotide binding	RT		53	1.7E-4	1.3E-2
GOTERM_MF_FAT	purine nucleotide binding	RT		53	5.1E-4	2.6E-2
GOTERM_MF_FAT	adenyl ribonucleotide binding	RT		46	6.8E-4	2.6E-2
GOTERM_MF_FAT	ATP binding	RT		45	1.3E-3	3.2E-2
GOTERM_MF_FAT	purine nucleoside binding	RT		46	1.9E-3	3.6E-2
GOTERM_MF_FAT	adenyl nucleotide binding	RT		46	1.9E-3	3.6E-2
GOTERM_MF_FAT	nucleoside binding	RT		46	2.1E-3	3.6E-2
SP_PIR_KEYWORDS	nucleotide-binding	RT		30	1.1E-2	5.0E-1
SP_PIR_KEYWORDS	atp-binding	RT		24	6.2E-2	5.4E-1
Annotation Cluster 3	**Enrichment Score: 2.39**			**Count**	**P_Value**	**Benjamini**
GOTERM_BP_FAT	nitrogen compound biosynthetic process	RT		14	5.2E-5	9.8E-3
INTERPRO	ATPase, P-type, K/Mg/Cd/Cu/Zn/Na/Ca /Na/H-transporter	RT		6	1.4E-4	1.0E-2
GOTERM_BP_FAT	purine nucleotide biosynthetic process	RT		9	1.9E-4	1.7E-2
GOTERM_BP_FAT	purine nucleotide metabolic process	RT		9	2.0E-4	1.3E-2
GOTERM_BP_FAT	nucleoside triphosphate biosynthetic process	RT		8	2.2E-4	1.1E-2
GOTERM_BP_FAT	nucleoside triphosphate metabolic process	RT		8	2.2E-4	1.1E-2
GOTERM_BP_FAT	purine nucleoside triphosphate biosynthetic process	RT		8	2.2E-4	1.1E-2
GOTERM_BP_FAT	purine nucleoside triphosphate metabolic process	RT		8	2.2E-4	1.1E-2
GOTERM_BP_FAT	ribonucleoside triphosphate metabolic process	RT		8	2.2E-4	1.1E-2
GOTERM_BP_FAT	purine ribonucleoside triphosphate biosynthetic process	RT		8	2.2E-4	1.1E-2
GOTERM_BP_FAT	ribonucleoside triphosphate biosynthetic process	RT		8	2.2E-4	1.1E-2
GOTERM_BP_FAT	purine ribonucleoside triphosphate metabolic process	RT		8	2.2E-4	1.1E-2
GOTERM_MF_FAT	ATPase activity, coupled to transmembrane movement of ions, phosphorylative mechanism	RT		6	8.3E-4	2.5E-2
GOTERM_BP_FAT	purine ribonucleotide metabolic process	RT		8	9.5E-4	3.5E-2
GOTERM_BP_FAT	purine ribonucleotide biosynthetic process	RT		8	9.5E-4	3.5E-2
GOTERM_BP_FAT	ATP metabolic process	RT		7	1.0E-3	3.1E-2
GOTERM_BP_FAT	ATP biosynthetic process	RT		7	1.0E-3	3.1E-2
GOTERM_BP_FAT	ribonucleotide metabolic process	RT		8	1.2E-3	3.3E-2
GOTERM_BP_FAT	ribonucleotide biosynthetic process	RT		8	1.2E-3	3.3E-2
GOTERM_MF_FAT	ATPase activity	RT		15	1.6E-3	3.5E-2
GOTERM_BP_FAT	nucleotide biosynthetic process	RT		9	2.5E-3	5.7E-2
INTERPRO	ATPase, P-type, plasma-membrane proton-efflux	RT		3	3.3E-3	1.7E-1
GOTERM_BP_FAT	nucleobase, nucleoside and nucleotide biosynthetic process	RT		9	4.5E-3	9.1E-2
GOTERM_BP_FAT	nucleobase, nucleoside, nucleotide and nucleic acid biosynthetic process	RT		9	4.5E-3	9.1E-2
INTERPRO	ATPase, P-type, ATPase-associated region	RT		4	4.8E-3	1.8E-1
INTERPRO	ATPase, P-type phosphorylation site	RT		4	4.8E-3	1.8E-1
GOTERM_MF_FAT	primary active transmembrane transporter activity	RT		7	4.8E-3	7.2E-2
GOTERM_MF_FAT	P-P-bond-hydrolysis-driven transmembrane transporter activity	RT		7	4.8E-3	7.2E-2
GOTERM_MF_FAT	ATPase activity, coupled to transmembrane movement of ions	RT		6	7.1E-3	9.6E-2
GOTERM_MF_FAT	ATPase activity, coupled	RT		11	1.1E-2	1.2E-1
INTERPRO	ATPase, P-type cation-transporter, N-terminal	RT		3	1.4E-2	4.0E-1
GOTERM_MF_FAT	ATPase activity, coupled to transmembrane movement of substances	RT		6	1.4E-2	1.4E-1
GOTERM_MF_FAT	ATPase activity, coupled to movement of substances	RT		6	1.4E-2	1.4E-1
GOTERM_MF_FAT	hydrolase activity, acting on acid anhydrides, catalyzing transmembrane movement of substances	RT		6	1.6E-2	1.5E-1
SP_PIR_KEYWORDS	phosphoprotein	RT		3	3.6E-2	5.2E-1
GOTERM_BP_FAT	cation transport	RT		5	1.3E-1	8.0E-1
GOTERM_MF_FAT	inorganic cation transmembrane transporter activity	RT		4	1.7E-1	7.1E-1
GOTERM_BP_FAT	ion transport	RT		5	1.9E-1	8.4E-1
SP_PIR_KEYWORDS	hydrolase	RT		14	2.6E-1	7.5E-1
GOTERM_CC_FAT	integral to membrane	RT		9	3.6E-1	1.0E0
GOTERM_CC_FAT	intrinsic to membrane	RT		9	4.0E-1	9.9E-1
KEGG_PATHWAY	Oxidative phosphorylation	RT		3	4.0E-1	9.4E-1
SP_PIR_KEYWORDS	transmembrane	RT		6	5.7E-1	9.3E-1
SP_PIR_KEYWORDS	membrane	RT		5	6.4E-1	9.5E-1
Annotation Cluster 4	**Enrichment Score: 1.3**			**Count**	**P_Value**	**Benjamini**
INTERPRO	Peptidase C2, calpain	RT		5	3.4E-3	1.5E-1
GOTERM_MF_FAT	cysteine-type endopeptidase activity	RT		6	7.7E-3	9.5E-2
GOTERM_MF_FAT	calcium-dependent cysteine-type endopeptidase activity	RT		5	9.0E-3	1.0E-1
SMART	CysPc	RT		5	1.0E-2	1.6E-1
SMART	calpain III	RT		5	1.0E-2	1.6E-1
GOTERM_MF_FAT	cysteine-type peptidase activity	RT		6	2.8E-2	2.3E-1
GOTERM_BP_FAT	proteolysis	RT		11	2.4E-1	8.4E-1
GOTERM_MF_FAT	endopeptidase activity	RT		7	6.7E-1	9.9E-1
GOTERM_MF_FAT	peptidase activity, acting on L-amino acid peptides	RT		7	9.0E-1	1.0E0
GOTERM_MF_FAT	peptidase activity	RT		7	9.3E-1	1.0E0

Figura 3.2.2.14 Resultado del agrupamiento de los 65 genes con valores de eficiencia traduccional mayores al percentil 95 derivados del análisis de transcriptoma y traductoma en epimastigotas de *T. cruzi.*

Annotation Cluster 1	Enrichment Score: 2.71			Count	P_Value	Benjamini
INTERPRO	Nucleotide-binding, alpha-beta plait	RT		5	7.3E-4	4.4E-2
INTERPRO	RNA recognition motif, RNP-1	RT		5	9.6E-4	2.9E-2
SMART	RRM	RT		5	1.0E-2	1.8E-1
Annotation Cluster 2	**Enrichment Score: 1.14**			**Count**	**P_Value**	**Benjamini**
INTERPRO	Peptidase C2, calpain	RT		3	5.8E-3	1.1E-1
GOTERM_MF_FAT	calcium-dependent cysteine-type endopeptidase activity	RT		3	1.4E-2	5.2E-1
SMART	CysPc	RT		3	2.1E-2	1.8E-1
SMART	calpain III	RT		3	2.1E-2	1.8E-1
GOTERM_MF_FAT	cysteine-type endopeptidase activity	RT		3	3.0E-2	5.5E-1
GOTERM_MF_FAT	cysteine-type peptidase activity	RT		3	5.3E-2	6.1E-1
GOTERM_MF_FAT	endopeptidase activity	RT		3	4.4E-1	9.9E-1
GOTERM_BP_FAT	proteolysis	RT		3	4.5E-1	1.0E0
GOTERM_MF_FAT	peptidase activity, acting on L-amino acid peptides	RT		3	6.0E-1	9.6E-1
GOTERM_MF_FAT	peptidase activity	RT		3	6.4E-1	9.6E-1

Figura 3.2.2.15 Resultado del agrupamiento de los 65 genes con valores de alta eficiencia traduccional derivados del análisis de transcriptoma y traductoma en epimastigotas de *T. cruzi*.

Annotation Cluster 1	Enrichment Score: 37.58			Count	P_Value	Benjamini
GOTERM_MF_FAT	structural constituent of ribosome	RT		83	7.5E-59	1.4E-56
GOTERM_MF_FAT	structural molecule activity	RT		84	4.1E-57	3.8E-55
SP_PIR_KEYWORDS	ribosomal protein	RT		83	1.7E-52	1.4E-50
GOTERM_BP_FAT	translation	RT		88	3.0E-43	6.8E-41
GOTERM_CC_FAT	ribosome	RT		85	6.3E-34	3.7E-32
KEGG_PATHWAY	Ribosome	RT		79	5.5E-33	2.2E-31
GOTERM_CC_FAT	ribonucleoprotein complex	RT		85	9.0E-29	2.6E-27
GOTERM_CC_FAT	intracellular non-membrane-bounded organelle	RT		95	1.1E-18	2.1E-17
GOTERM_CC_FAT	non-membrane-bounded organelle	RT		95	1.1E-18	2.1E-17
Annotation Cluster 2	**Enrichment Score: 2.14**			**Count**	**P_Value**	**Benjamini**
INTERPRO	Heat shock protein 70	RT		6	4.5E-4	4.3E-2
INTERPRO	Heat shock protein Hsp70	RT		6	5.3E-4	4.1E-2
INTERPRO	Heat shock protein 70, conserved site	RT		4	5.8E-3	2.3E-1
SP_PIR_KEYWORDS	stress response	RT		8	1.4E-2	2.6E-1
SP_PIR_KEYWORDS	atp-binding	RT		15	1.0E0	1.0E0
Annotation Cluster 3	**Enrichment Score: 1.67**			**Count**	**P_Value**	**Benjamini**
INTERPRO	Thioredoxin fold	RT		8	2.6E-3	1.4E-1
GOTERM_MF_FAT	peroxidase activity	RT		5	3.6E-3	1.3E-1
GOTERM_MF_FAT	oxidoreductase activity, acting on peroxide as acceptor	RT		5	3.6E-3	1.3E-1
SP_PIR_KEYWORDS	peroxidase	RT		5	3.9E-3	1.0E-1
GOTERM_MF_FAT	antioxidant activity	RT		6	5.8E-3	1.4E-1
INTERPRO	Peroxiredoxin, C-terminal	RT		3	7.8E-3	2.7E-1
INTERPRO	Alkyl hydroperoxide reductase/ Thiol specific antioxidant/ Mal allergen	RT		3	1.4E-2	3.5E-1
KEGG_PATHWAY	Glutathione metabolism	RT		8	4.7E-2	6.3E-1
GOTERM_BP_FAT	cell redox homeostasis	RT		6	9.4E-2	8.0E-1
GOTERM_BP_FAT	cellular homeostasis	RT		6	1.0E-1	8.2E-1
GOTERM_BP_FAT	homeostatic process	RT		6	1.2E-1	8.1E-1
INTERPRO	Thioredoxin-like	RT		3	1.4E-1	9.7E-1
GOTERM_MF_FAT	oxidoreductase activity, acting on sulfur group of donors	RT		3	3.0E-1	9.6E-1
Annotation Cluster 4	**Enrichment Score: 1.63**			**Count**	**P_Value**	**Benjamini**
INTERPRO	General substrate transporter	RT		3	1.1E-2	3.2E-1
INTERPRO	Sugar/inositol transporter	RT		3	1.1E-2	3.2E-1
GOTERM_BP_FAT	transmembrane transport	RT		8	1.3E-2	6.3E-1
SP_PIR_KEYWORDS	transport	RT		8	2.0E-1	9.6E-1

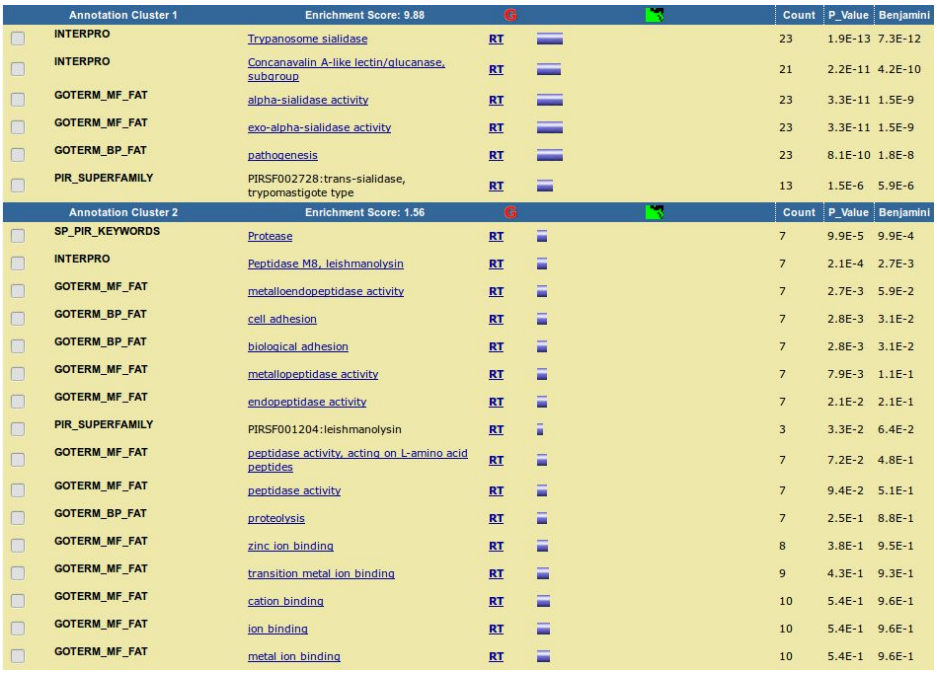
Figura 3.2.2.16 Resultado del agrupamiento de los 328 genes con valores de transcripción mayores al percentil 95 derivados del análisis de transcriptoma en epimastigotas de *T. cruzi.*

Annotation Cluster 1	Enrichment Score: 35.06			Count	P_Value	Benjamini
GOTERM_MF_FAT	structural constituent of ribosome	RT		81	1.7E-52	3.3E-50
GOTERM_MF_FAT	structural molecule activity	RT		82	8.5E-51	8.2E-49
SP_PIR_KEYWORDS	ribosomal protein	RT		82	5.3E-49	4.1E-47
GOTERM_BP_FAT	translation	RT		86	1.1E-40	2.2E-38
GOTERM_CC_FAT	ribosome	RT		84	9.4E-34	4.5E-32
KEGG_PATHWAY	Ribosome	RT		77	1.5E-30	5.1E-29
GOTERM_CC_FAT	ribonucleoprotein complex	RT		84	1.2E-28	2.8E-27
GOTERM_CC_FAT	non-membrane-bounded organelle	RT		93	4.8E-18	7.6E-17
GOTERM_CC_FAT	intracellular non-membrane-bounded organelle	RT		93	4.8E-18	7.6E-17
Annotation Cluster 2	**Enrichment Score: 3.75**			**Count**	**P_Value**	**Benjamini**
INTERPRO	Heat shock protein 70	RT		8	4.5E-6	2.0E-3
INTERPRO	Heat shock protein Hsp70	RT		8	5.8E-6	1.3E-3
SP_PIR_KEYWORDS	stress response	RT		9	5.6E-3	1.1E-1
INTERPRO	Heat shock protein 70, conserved site	RT		4	6.9E-3	2.6E-1
Annotation Cluster 3	**Enrichment Score: 2.19**			**Count**	**P_Value**	**Benjamini**
GOTERM_BP_FAT	glucose metabolic process	RT		11	3.8E-4	2.5E-2
GOTERM_BP_FAT	glucose catabolic process	RT		9	6.7E-4	3.2E-2
GOTERM_BP_FAT	monosaccharide catabolic process	RT		9	6.7E-4	3.2E-2
GOTERM_BP_FAT	hexose catabolic process	RT		9	6.7E-4	3.2E-2
GOTERM_BP_FAT	cellular carbohydrate catabolic process	RT		9	8.9E-4	3.5E-2
GOTERM_BP_FAT	alcohol catabolic process	RT		9	8.9E-4	3.5E-2
GOTERM_BP_FAT	carbohydrate catabolic process	RT		9	8.9E-4	3.5E-2
GOTERM_BP_FAT	monosaccharide metabolic process	RT		12	1.5E-3	4.9E-2
GOTERM_BP_FAT	hexose metabolic process	RT		11	4.0E-3	1.1E-1
GOTERM_BP_FAT	glycolysis	RT		6	1.2E-2	2.5E-1
GOTERM_BP_FAT	generation of precursor metabolites and energy	RT		10	1.3E-2	2.5E-1
KEGG_PATHWAY	Glyoxylate and dicarboxylate metabolism	RT		4	2.3E-1	8.9E-1
GOTERM_BP_FAT	coenzyme metabolic process	RT		5	3.0E-1	9.0E-1
GOTERM_BP_FAT	cofactor metabolic process	RT		5	4.0E-1	9.6E-1
KEGG_PATHWAY	Pentose phosphate pathway	RT		4	6.9E-1	9.9E-1
Annotation Cluster 4	**Enrichment Score: 1.43**			**Count**	**P_Value**	**Benjamini**
INTERPRO	Thioredoxin fold	RT		8	3.7E-3	2.4E-1
INTERPRO	Thioredoxin-like	RT		5	5.0E-3	2.7E-1
INTERPRO	Peroxiredoxin, C-terminal	RT		3	8.9E-3	3.0E-1
INTERPRO	Alkyl hydroperoxide reductase/ Thiol specific antioxidant/ Mal allergen	RT		3	1.6E-2	4.0E-1
SP_PIR_KEYWORDS	peroxidase	RT		4	3.3E-2	3.6E-1
GOTERM_MF_FAT	oxidoreductase activity, acting on peroxide as acceptor	RT		4	3.3E-2	6.1E-1
GOTERM_MF_FAT	peroxidase activity	RT		4	3.3E-2	6.1E-1
GOTERM_MF_FAT	antioxidant activity	RT		5	3.9E-2	6.1E-1
KEGG_PATHWAY	Glutathione metabolism	RT		7	1.2E-1	8.9E-1
GOTERM_BP_FAT	cell redox homeostasis	RT		5	2.3E-1	8.6E-1
GOTERM_BP_FAT	cellular homeostasis	RT		5	2.5E-1	8.8E-1
GOTERM_BP_FAT	homeostatic process	RT		5	2.8E-1	8.9E-1

Figura 3.2.2.17 Resultado del agrupamiento de los 328 genes con valores de traducción mayores al percentil 95 derivados del análisis de traductoma en epimastigotas de *T. cruzi.*

Annotation Cluster 1	Enrichment Score: 0.94			Count	P_Value	Benjamini
GOTERM_MF_FAT	endopeptidase activity	RT		3	6.8E-2	7.4E-1
GOTERM_MF_FAT	peptidase activity, acting on L-amino acid peptides	RT		3	1.1E-1	4.4E-1
GOTERM_MF_FAT	peptidase activity	RT		3	1.3E-1	4.1E-1
GOTERM_BP_FAT	proteolysis	RT		3	1.7E-1	6.9E-1
Annotation Cluster 2	**Enrichment Score: 0.94**			**Count**	**P_Value**	**Benjamini**
KEGG_PATHWAY	Ribosome	RT		3	3.3E-2	3.3E-2
SP_PIR_KEYWORDS	ribosomal protein	RT		3	5.4E-2	3.6E-1
GOTERM_MF_FAT	structural constituent of ribosome	RT		3	7.0E-2	5.0E-1
GOTERM_MF_FAT	structural molecule activity	RT		3	8.0E-2	4.1E-1
GOTERM_CC_FAT	ribosome	RT		3	1.2E-1	5.9E-1
GOTERM_CC_FAT	ribonucleoprotein complex	RT		3	1.5E-1	4.4E-1
GOTERM_BP_FAT	translation	RT		3	1.6E-1	8.9E-1
GOTERM_CC_FAT	non-membrane-bounded organelle	RT		3	3.7E-1	6.6E-1
GOTERM_CC_FAT	intracellular non-membrane-bounded organelle	RT		3	3.7E-1	6.6E-1

Figura 3.2.2.18 Resultado del agrupamiento de los 50 genes con valores confiables solo en el traductoma de epimastigotas de *T. cruzi.*



Annotation Cluster 1	Enrichment Score: 9.88			Count	P_Value	Benjamini
INTERPRO	Trypanosome sialidase	RT		23	1.9E-13	7.3E-12
INTERPRO	Concanavalin A-like lectin/glucanase, subgroup	RT		21	2.2E-11	4.2E-10
GOTERM_MF_FAT	alpha-sialidase activity	RT		23	3.3E-11	1.5E-9
GOTERM_MF_FAT	exo-alpha-sialidase activity	RT		23	3.3E-11	1.5E-9
GOTERM_BP_FAT	pathogenesis	RT		23	8.1E-10	1.8E-8
PIR_SUPERFAMILY	PIRSF002728:trans-sialidase, trypomastigote type	RT		13	1.5E-6	5.9E-6
Annotation Cluster 2	**Enrichment Score: 1.56**			**Count**	**P_Value**	**Benjamini**
SP_PIR_KEYWORDS	Protease	RT		7	9.9E-5	9.9E-4
INTERPRO	Peptidase M8, leishmanolysin	RT		7	2.1E-4	2.7E-3
GOTERM_MF_FAT	metalloendopeptidase activity	RT		7	2.7E-3	5.9E-2
GOTERM_BP_FAT	cell adhesion	RT		7	2.8E-3	3.1E-2
GOTERM_BP_FAT	biological adhesion	RT		7	2.8E-3	3.1E-2
GOTERM_MF_FAT	metallopeptidase activity	RT		7	7.9E-3	1.1E-1
GOTERM_MF_FAT	endopeptidase activity	RT		7	2.1E-2	2.1E-1
PIR_SUPERFAMILY	PIRSF001204:leishmanolysin	RT		3	3.3E-2	6.4E-2
GOTERM_MF_FAT	peptidase activity, acting on L-amino acid peptides	RT		7	7.2E-2	4.8E-1
GOTERM_MF_FAT	peptidase activity	RT		7	9.4E-2	5.1E-1
GOTERM_BP_FAT	proteolysis	RT		7	2.5E-1	8.8E-1
GOTERM_MF_FAT	zinc ion binding	RT		8	3.8E-1	9.5E-1
GOTERM_MF_FAT	transition metal ion binding	RT		9	4.3E-1	9.3E-1
GOTERM_MF_FAT	cation binding	RT		10	5.4E-1	9.6E-1
GOTERM_MF_FAT	ion binding	RT		10	5.4E-1	9.6E-1
GOTERM_MF_FAT	metal ion binding	RT		10	5.4E-1	9.6E-1

Figura 3.2.2.19 Clasificación funcional de las 208 genes que codifican para proteínas con valores confiables únicamente en el transcriptoma y que presentan una baja traducción (menor al percentil 10) y una baja eficiencia de traducción (menor al percentil 5) en epimastigotas de *T. cruzi.*

	Ala_GCU	Ala_GCC	Ala_GCA	Ala_GCG	Cys_UGU	Cys_UGC	Asp_GAU	Asp_GAC
Todos	0,1830	0,2619	0,2502	0,3048	0,4539	0,5189	0,5259	0,4732
Alta Expresión	0,1799	0,3696	0,1545	0,2959	0,1862	0,7318	0,3018	0,6818
Regiones Pausas	0,1775	0,3025	0,2500	0,2675	0,5000	0,5000	0,4500	0,5500
Genes con Pausas	0,1725	0,2800	0,2425	0,3050	0,4250	0,5750	0,5000	0,5000
	Glu_GAA	Glu_GAG	Phe_UUU	Phe_UUC	Gly_GGU	Gly_GGC	Gly_GGA	Gly_GGG
Todos	0,3823	0,6171	0,7164	0,2822	0,2543	0,2959	0,2125	0,2373
Alta Expresión	0,1870	0,8130	0,3670	0,6330	0,2266	0,5587	0,1366	0,0781
Regiones Pausas	0,4900	0,5100	0,8850	0,1150	0,1950	0,3900	0,2775	0,1400
Genes con Pausas	0,3650	0,6350	0,6950	0,3050	0,2625	0,3250	0,1975	0,2150
	His_CAU	His_CAC	Ile_AUU	Ile_AUC	Ile_AUA	Lys_AAA	Lys_AAG	Leu_UUA
Todos	0,4753	0,5142	0,5415	0,2871	0,1701	0,4016	0,5980	0,0829
Alta Expresión	0,2023	0,7322	0,4206	0,5426	0,0368	0,1016	0,8984	0,0147
Regiones Pausas	0,6300	0,4550	0,3767	0,4433	0,1767	0,5700	0,4300	0,1000
Genes con Pausas	0,3700	0,5500	0,5400	0,3167	0,1433	0,3450	0,6550	0,0700
	Leu_UUG	Leu_CUU	Leu_CUC	Leu_CUA	Leu_CUG	Met_AUG	Asn_AAU	Asn_AAC
Todos	0,2303	0,2486	0,1496	0,0542	0,2344	1,0000	0,5018	0,4960
Alta Expresión	0,0957	0,2372	0,2491	0,0239	0,3793	1,0000	0,2122	0,7878
Regiones Pausas	0,2000	0,2667	0,1667	0,1117	0,1550	1,0000	0,4600	0,5400
Genes con Pausas	0,2117	0,2550	0,1617	0,0517	0,2517	1,0000	0,4650	0,5350
	Pro_CCU	Pro_CCC	Pro_CCA	Pro_CCG	Gln_CAA	Gln_CAG	Arg_CGU	Arg_CGC
Todos	0,2033	0,2208	0,2690	0,3045	0,3785	0,6199	0,2484	0,2292
Alta Expresión	0,1807	0,3541	0,1451	0,3202	0,1502	0,8498	0,2835	0,5191

Regiones Pausas	0,1325	0,1875	0,2450	0,4350	0,4100	0,5900	0,2767	0,2133
Genes con Pausas	0,1900	0,2300	0,2725	0,3075	0,3600	0,6400	0,2567	0,2667
	Arg_CGA	Arg_CGG	Arg_AGA	Arg_AGG	Ser_UCU	Ser_UCC	Ser_UCA	Ser_UCG
Todos	0,1385	0,1799	0,0824	0,1216	0,1728	0,1784	0,1508	0,1621
Alta Expresión	0,0297	0,0827	0,0268	0,0581	0,1280	0,2570	0,0746	0,2173
Regiones Pausas	0,1483	0,1700	0,0217	0,1700	0,1667	0,2417	0,1667	0,0750
Genes con Pausas	0,1300	0,1733	0,0650	0,1067	0,1633	0,1867	0,1433	0,1650
	Ser_AGU	Ser_AGC	Thr_ACU	Thr_ACC	Thr_ACA	Thr_ACG	Val_GUU	Val_GUC
Todos	0,1579	0,1778	0,1639	0,2028	0,2907	0,3425	0,2566	0,1803
Alta Expresión	0,0635	0,2596	0,1160	0,2741	0,1675	0,4424	0,2164	0,2253
Regiones Pausas	0,1067	0,2417	0,1400	0,1825	0,3650	0,3100	0,2425	0,2700
Genes con Pausas	0,1500	0,1917	0,1525	0,2150	0,2850	0,3500	0,2475	0,1875
	Val_GUA	Val_GUG	Trp_UGG	Tyr_UAU	Tyr_UAC			
Todos	0,1180	0,4450	0,9391	0,3869	0,6049			
Alta Expresión	0,0294	0,5289	0,7705	0,1526	0,8474			
Regiones Pausas	0,0800	0,4050	1,0000	0,3350	0,6650			
Genes con Pausas	0,1050	0,4575	1,0000	0,3550	0,6450			

Tabla 3.2.2.4 Se muestran los valores de frecuencia de codones sinónimos en *T. cruzi* para los diferentes grupos analizados (ver Figura 3.2.2.10)

Conclusiones y perspectivas 4

En este trabajo nos propusimos aportar a la comprensión de los mecanismos de la expresión génica en los Tritryps mediante el uso de aproximaciones globales para profundizar en aspectos estructurales y funcionales de los genomas de estos organismos.

En relación a la búsqueda *in silico* de señales involucradas en la expresión génica encontramos que:

Los repetidos de dinucleótidos son más frecuentes de lo que cabe esperar por azar y presentan asimetría de hebra en los tres genomas, lo que sugiere su vinculación a procesos direccionales como la transcripción. Además, la localización no uniforme junto con la tendencia a encontrarse cerca de los ORFs y alejados de los límites de los DGCs, apoyan su posible involucramiento en la expresión génica. Los datos de transcriptoma con delimitación de UTRs permitirán la definición precisa de la ubicación con respecto a los ORFs, haciendo posible el estudio de los patrones de expresión del conjunto de genes que los contienen. Asimismo, aproximaciones con genes reporteros pueden aportar información sobre la funcionalidad de estas señales. En nuestro laboratorio se vienen desarrollando estudios en este sentido.

La curvatura intrínseca en el ADN también se distribuye en forma no uniforme en los tres genomas. Específicamente, las regiones de alta curvatura se concentran en zonas que están asociadas al inicio de la transcripción. Resultaría interesante verificar experimentalmente en *T. cruzi* si los sitios internos a los DGCs que por su nivel de curvatura se puede especular que colocalizan con TSSs presentan marcadores epigenéticos característicos. La conservación de la localización de las regiones de alta curvatura entre especies de *Leishmania*, fundamenta el estudio futuro de la posible existencia de sintenia de estas señales en los tripanosomátidos.

El estudio de curvaturas en el ADN fue posible gracias al desarrollo de una nueva herramienta que permite la integración de la curvatura intrínseca (RIIC) en regiones del genoma. Esta herramienta podrá ser aplicada para el estudio de otros modelos.

En los subtelómeros de *T. brucei*, donde se ubican los genes y pseudogenes de VSG se observa también regiones de alta curvatura. La colocalización con la base J, hace posible sugerir que esta estructura secundaria esté involucrada en procesos de expresión. A partir del reciente contacto con el Dr. Sabatini, contamos con datos que podrán proveernos de la localización precisa de la base J, permitiendo un estudio de colocalización más detallado. Por otra parte, no se puede descartar el

involucramiento de estas señales en procesos de recombinación que son frecuentes en estas regiones para ampliar el repertorio de VSG característico de la variación antigénica. Sería interesante estudiar este fenómeno en otros tripanosomátidos que presenten este mecanismo.

Los promotores de los ADNr de los Tritryps, de la misma manera que en el resto de los eucariotas, se caracterizan por no presentar una conservación de señales de secuencia primaria sino de estructura secundaria. Interesantemente, esta característica está también presente en los promotores de los genes para proteínas que en forma excepcional son transcritos por la ARNPI en *T. brucei*. Esto nos permite concluir que el reconocimiento de la estructura secundaria es una característica de la maquinaria transcripcional de la ARNPI que ha sido conservada a lo largo de la evolución. Es posible verificar la predicción de la estructura secundaria por métodos experimentales. En este sentido, hemos encarado el análisis por microscopía de fuerza atómica. Por otra parte, estamos estudiando la inclusión de un mayor número de parámetros conformacionales para intentar diferenciar estas regiones del resto del genoma. Finalmente, nos planteamos extender estos estudios al resto de los promotores descritos en estos organismos.

En cuanto a el análisis de perfiles de huellas ribosomales, los resultados obtenidos para *T. cruzi*, permiten concluir que esta metodología constituye una herramienta más adecuada que el transcriptoma para la estimación de los niveles de expresión, presentando mejores valores de correlación con el proteoma. A su vez, aporta información sobre genes que no son detectados en el proteoma por su baja expresión. Además los datos de transcriptoma aquí obtenidos amplían los disponibles actualmente que son derivados de un ensayo de microarreglos. Tanto el transcriptoma como el traductoma constituyen una referencia para estudios posteriores y como tal un insumo que podrá ser usado por el conjunto de la comunidad científica que estudia este organismo

Como era de esperar para estos organismos, se encuentran genes regulados a nivel de traducibilidad. Particularmente, se observa que los genes que codifican proteínas del tipo RRM presentan una alta eficiencia de traducción posibilitando una respuesta rápida de sus niveles de expresión. Esta eficiencia es aún mayor que la de los genes que codifican para las proteínas ribosomales en las que el nivel de expresión se ve enriquecido por la alta tasa de transcripción. Este resultado es de especial relevancia dado que las proteínas del tipo RRM están involucradas en mecanismos de regulación post transcripcional. Por otra parte, puede resultar

interesante estudiar los patrones de expresión de los genes en los cuales se observaron pausas de los ribosomas en el estadío epimastigota, ya que pueden representar ejemplos de genes regulados durante la elongación de la traducción. Asimismo, resulta de interés el estudio comparativo de perfil de huellas ribosomales a lo largo del ciclo de vida del parásito. Para esto se están analizando actualmente datos del traductoma en el estadío infectivo tripomastigota metacíclico.

El uso de codones de los genes correlaciona con el valor de tasa traduccional apoyando las teorías de adaptación para optimizar este proceso. Los sitios con alto número de huellas se encuentran enriquecidos en codones de baja frecuencia. Un estudio más profundo de las regiones donde se producen las pausas podrá proveer de información para esclarecer los mecanismos moleculares implícitos en este paso clave en la regulación traduccional. Sería interesante correlacionar la frecuencia de uso de codones con los niveles de cada ARNt que se pueden evaluar a partir de los datos obtenidos en este trabajo.

Globalmente, los datos aquí presentados constituyen una contribución novedosa en aspectos poco explorados de la estructura y dinámica funcional del genoma de los Tritryps, sirviendo de base para el estudio en profundidad de la expresión de genes o grupos de genes de interés específico.

Otros aportes

5

5.1 Avances en la caracterización de las señales de repetidos de dinucleótidos

El análisis de secuencias repetidas de dinucleótidos es una temática que continúa siendo desarrollada en nuestro laboratorio. Durante el trabajo de tesis de maestría de la Mag. Lucia Pastro, se profundizó particularmente en los repetidos de poli(CA) en *T. cruzi*. Estos repetidos son reconocidos por extractos proteicos totales y se encontró que su inserción en las regiones 3'UTR de genes reporteros lleva a una modulación dependiente del estadío del ciclo de vida. La búsqueda global identificó que un 10% de los genes presentan el repetido CA en sus regiones UTRs predichas. Además se observa una correlación con la presencia del repetido complementario poli(TG), favoreciendo la formación de estructuras secundarias en los ARNs. Las familias de los genes de las mucinas presentan este fenómeno. Los resultados en su conjunto implican que los repetidos son señales importantes en tripanosomátidos regulando la estabilidad de los mensajeros en los diferentes estadíos. En el desarrollo de este trabajo, participé en la construcción y transfección de los vectores utilizados. Además, colaboré con la construcción de las bases de datos de UTRs y los análisis de enriquecimiento de las categorías de ontología génica.

Experimental Parasitology 134 (2013) 511–518

Contents lists available at SciVerse ScienceDirect

Experimental Parasitology

journal homepage: www.elsevier.com/locate/yexpr

Implication of CA repeated tracts on post-transcriptional regulation in *Trypanosoma cruzi*

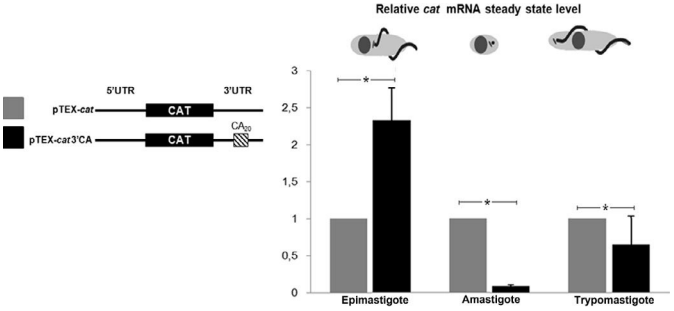
Lucía Pastro [a,b], Pablo Smircich [a,b], Leticia Pérez-Díaz [a], María Ana Duhagon [a,b,*], Beatriz Garat [a,*]

[a] Laboratorio de Interacciones Moleculares, Facultad de Ciencias, 11400 Montevideo, Uruguay
[b] Departamento de Genética, Facultad de Medicina, 11800 Montevideo, Uruguay

HIGHLIGHTS

- CA repeated tracts are present at 3′UTRs of about 10% of *T. cruzi* genes.
- Many of them are predicted to exhibit the polyCA in a single stranded conformation.
- PolyCA insertion significantly increases reporter mRNA level in epimastigote stage.
- In amastigote stage the reporter mRNA steady state is significantly diminished.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 10 August 2012
Received in revised form 13 April 2013
Accepted 19 April 2013
Available online 28 April 2013

Keywords:
Polydinucleotides
UTR
RNA conformation
Reporter assay
mRNA steady state
mRNA stability

ABSTRACT

In *Trypanosoma cruzi* gene expression regulation mainly relays on post-transcriptional events. Nevertheless, little is known about the signals which control mRNA abundance and functionality. We have previously found that CA repeated tracts (polyCA) are abundant in the vicinity of open reading frames and constitute specific targets for single stranded binding proteins from *T. cruzi* epimastigote. Given the reported examples of the involvement of polyCA motifs in gene expression regulation, we decided to further study their role in *T. cruzi*. Using an *in silico* genome-wide analysis, we identify the genes that contain polyCA within their predicted UTRs. We found that about 10% of *T. cruzi* genes carry polyCA therein. Strikingly, they are frequently concurrent with GT repeated tracts (polyGT), favoring the formation of a secondary structure exhibiting the complementary polydinucleotides in a double stranded helix. This feature is found in the species-specific family of genes coding for mucine associated proteins (MASPs) and other genes. For those polyCA-containing UTRs that lack polyGT, the polyCA is mainly predicted to adopt a single stranded structure. We further analyzed the functional role of such element using a reporter approach in *T. cruzi*. We found out that the insertion of polyCA at the 3′ UTR of a reporter gene in the pTEX vector modulates its expression along the parasite's life cycle. While no significant change of the mRNA steady state of the reporter gene could be detected at the trypomastigote stage, significant increase in the epimastigote and reduction in the amastigote stage were observed. Altogether, these results suggest the involvement of polyCA as a signal in gene expression regulation in *T. cruzi*.

1. Introduction

Trypanosoma cruzi is the etiological agent of the American trypanosomiasis or Chagas' disease, a main health problem affecting several million people in Latin America and Mexico, and currently expanding worldwide. Its life cycle is complex, involving a replicative epimastigote form at the midgut of a reduviid insect, which

* Corresponding authors. Address: Laboratorio de Interacciones Moleculares, Facultad de Ciencias Iguá 4225, 11400 Montevideo, Uruguay. Fax: +598 2 525 86 17.
E-mail addresses: mduhagon@fcien.edu.uy (M.A. Duhagon), bgarat@fcien.edu.uy (B. Garat).

0014-4894/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.exppara.2013.04.004

transforms to a non-replicative metacyclic infective form in the hindgut. The latter is transmitted through vector feces to different mammalians hosts, where it differentiates to replicative intracellular amastigotes and thereafter to non-replicative infective tripomastigotes (Chagas, 1909; Tyler and Engman, 2001).

The importance of post-transcriptional and post-translational processes in eukaryotic gene expression is being growingly recognized. *T. cruzi*, as other members of the Trypanosomatidae family, does not typically regulate transcription initiation of protein coding genes. Indeed, the genes are characteristically arranged in large polycistronic units that are constitutively transcribed, independently of their function or developmental stage. Individual mRNAs are produced by 5′ *trans*-splicing and 3′ polyadenylation. The accurate regulation of mRNA abundance and functionality, needed to adapt to the diverse environments this parasite is exposed to, are mainly defined by specific post-transcriptional mechanisms (De Gaudenzi et al., 2011; Kramer, 2012). Therefore, this protozoan has been proposed as an interesting model for the study of regulatory mechanisms of gene expression that occur after transcription initiation.

A growing number of studies have revealed the existence of RNA binding proteins, as well as *cis* signals, involved in many aspects of RNA metabolism in trypanosomatids (Araujo and Teixeira, 2011; Kramer and Carrington, 2011). Nevertheless, the signals involved in the modulation of mRNA abundance and/or translation in *T. cruzi* are still barely identified. Most studies have sought to delimit the regions in the 3′ UTRs that control mRNA abundance of specific genes in response to changes in the parasite life cycle; such is the case for the TS gene family (Abuin et al., 1999; Gentil et al., 2009; Jager et al., 2008; Nozaki and Cross, 1995; Songthamwat et al., 2007; Weston et al., 1999), the amastins (Coughlin et al., 2000; Teixeira et al., 1994), the mucins (Di Noia et al., 2000) and the alpha and beta tubulins (Bartholomeu et al., 2002; da Silva et al., 2006). Among them, the ubiquitous AU rich elements (ARE) (D'Orso and Frasch, 2001, da Silva et al., 2006, Di Noia et al., 2000; Noe et al., 2008), a G rich (GRE) (D'Orso and Frasch, 2001) and a GT repeated element (Bartholomeu et al., 2002) have been involved in mRNA stability. In addition, both an U-rich region in the 3′ UTR and the 5′ UTR of the HSP70 mRNA confer temperature-dependent regulation of mRNA stability (Rodrigues et al., 2010). On the other hand, polysomal mobilization and translatability have been described as key point of gene regulation for the expression of metacyclogenin (Avila et al., 2001) and trans-sialidase (Araujo et al., 2011). Finally, very recently, the involvement of a 43-nt U-rich element in the modulation of abundance and/or translation of subsets of transcripts in amastigotes has been described (Li et al., 2012). Thus, our current knowledge of regulatory sequence elements in *T. cruzi* is still scarce.

Considering the relevant biological roles that have been ascribed to repetitive motives in several organisms (Sinden, 1994), including parasites (Clayton, 2010) we have previously focused on the analysis of polydinucleotides in *T. cruzi* (Duhagon et al., 2011). We found an asymmetrical strand distribution of complementary polydinucleotides, favoring polyTT and polyGT in the coding strand. Interestingly, though the polyCA sequences (complementary to polyGT) are consequently not favored in the coding strand, they are highly represented in the vicinity of open reading frames, which suggests its participation in gene expression regulation. In addition, we have previously shown that single stranded polyCA tracts constitute specific targets for binding proteins from *T. cruzi* epimastigote extracts (Duhagon et al., 2001). In fact, there are numerous reports of the involvement of polyCA in gene expression regulation (Baranovskaya et al., 2009; Buisine et al., 2008; Lee et al., 2004; Lin et al., 2005; Martin-Farmer and Janssen, 1999). Besides, their function in homologue recombination has also been recognized (Buratti et al., 2001; Gabellini, 2001; Gendrel et al., 2000; Majewski and Ott, 2000; Rockman and Wray, 2002).

To further understand the role of the polyCA signals on gene expression regulation in *T. cruzi*, we performed an *in silico* genome-wide analysis to identify the genes that contain polyCA in their UTRs. We found that about 10% of *T. cruzi* genes carry polyCA on their UTRs, principally at the 3′ UTR (91.4% of them). We also detected the concurrent presence of polyCA and polyGT in approximately half of those genes. For these polyGT and polyCA containing UTRs, a secondary structure exhibiting the complementary polydinucleotides in a double stranded helix is thermodynamically favored. Loops displaying the polyCA as single stranded structure are mostly predicted for the other half of polyCA containing genes. The effect of the presence of polyCA at the UTRs was further studied using a reporter gene assay. No significant variation either on reporter mRNA steady state or half life was observed when the polyCA was located at the 5′ UTR of reporter gene on epimastigote stage. However, the insertion of polyCA at the 3′ UTR drives to a significant increase of the mRNA steady state of the reporter gene. Consistently, a significant increase of the reporter mRNA half life was also observed. Interestingly, the polyCA insertion at the 3′ UTR of reporter gene has no effect at the trypomastigote stage but a significant reduction of the mRNA steady state of the reporter gene at the amastigote stage was observed. Altogether, these findings suggest that polyCA may contribute to the regulation of life stage specific gene expression in *T. cruzi*.

2. Materials and methods

2.1. Bioinformatics analysis

A database containing *T. cruzi* Esmeraldo strain genes, including predicted UTRs was generated using the *Sequence Retrieval* tool from TriTrypDB (release 2-0). Following published data (Campos et al., 2008), sequences spanning 35 bases upstream of the start codon and 264 bases downstream from the stop codon were defined as predicted 5′ and 3′ UTRs, respectively. A set of genes (Tc.polyCA-UTR-set) containing a $(CA)_n$ or a $(AC)_n$ motifs with $n \geqslant 4$, at the predicted gene UTRs, was obtained using scripts developed for this purpose. Tc.polyCA-UTR-set orthologous genes, from TriTrypDB (*Trypanosoma brucei* and *Leishmania major*) were defined using a reciprocal blast approach (*e*-value cutoff 10^{-10}). As for *T. cruzi*, the UTR lengths for *T. brucei* (5′ UTR: 68 nt and 3′ UTR: 587 nt) and *L. major* (5′ UTR: 35 nt and 3′ UTR: 268 nt) were defined using their proposed average sizes (Campos et al., 2008).

Gene Ontology analysis was performed in the Tc.polyCA-UTR-set using the *Blast2GO* tool (<http://www.blast2go.org/start_blast2go>).

Clustering of the available data for stage specific transcript levels in *T. cruzi* (Minning et al., 2009) was done using Pearson correlation and the average linkage clustering method with Multi-Experiment Viewer (MeV) analysis software (Saeed et al., 2003).

RNA secondary structure prediction was carried out using the DINAMelt software (Server for Prediction of Melting Profiles for Nucleic Acids) using the default parameters (<http://dinamelt.bioinfo.rpi.edu/quikfold.php>).

2.2. Vector constructions

The *T. cruzi* plasmid pTEX (Kelly et al., 1992), where the chloramphenicol acetyltransferase reporter gene was cloned using *Bam*H I and *Xho* I restriction enzymes to produce pTEX-*cat*, was used as a reporter vector. Polydinucleotide double-stranded repeats were obtained by hybridization of the synthetic oligonucleotides $(dTdG)_{15}$ and $(dCdA)_{15}$ (Operon) previously treated with polynucleotide kinase enzyme (Promega). To insert the repeated sequence at the 5′ UTR and 3′ UTR of the reporter gene in pTEX-*cat*, the vector was digested with *Spe* I and with *Sal* I (New England

Biolabs) respectively, followed by filling in with Klenow fragment (New England Biolabs) and ligation to the double stranded repeat with T4 ligase enzyme (Promega). Plasmid were amplified in *Escherichia coli* and sequenced to determine the orientation and length of the CA insert. One plasmid containing $(CA)_{22}$ at the 5′ UTR (pTEX-*cat* 5′CA) and another with $(CA)_{20}$ at the 3′ UTR (pTEX-*cat* 3′CA) were selected for further analysis.

2.3. Parasites and transfections

The *T. cruzi* Dm28 clone (Contreras et al., 1988) was used. Epimastigotes were maintained at 28 °C in liver infusion tryptose (LIT) medium supplemented with 10% heat inactivated fetal bovine serum (FBS). Amastigotes and cellular trypomastigotes were prepared as previously (Chiribao et al., 2012; Duhagon et al., 2009). Briefly, epimastigotes at stationary phase were incubated in TAU medium at 28 °C for 2 h (Contreras et al., 1985). The parasites were then washed twice with PBS and immediately used to infect Vero cells (100 parasites per cell). Infected cells were incubated in RPMI at 37 °C supplemented with 10% FBS and 5% CO_2 to obtain intracellular amastigotes. Three days after infection, trypomastigote forms were obtained from medium supernatant by centrifugation at 2,000*g*. To obtain axenic amastigotes, trypomastigote forms from medium supernatant were incubated for 3 days in RPMI at 37 °C supplemented with 10% FBS and 5% CO_2.

Transfections were done as previously (Perez-Diaz et al., 2007). Epimastigotes were electroporated with plasmids pTEX-*cat*, pTEX-*cat* 5′CA and pTEX-*cat* 3′CA and then selected by culturing in presence of genetecin 500 μg/mL. The presence and maintenance of the plasmid and CA_n insertions in the selected parasites were evaluated by PCR and hybridization analysis. Briefly, 50 ng of total transfectant parasite DNA purified as described (Coderre et al., 1983) was used to amplify *cat* 5′ and 3′ UTR. The following primers: M13F: 5′-GTAAAACGACGGCCAGT-3′ and IG3R: 5′-AAAGAAAAGCAGAAAAAACTAAAAAGATGTGGC-3′, and, catF: 5′TCGTCTCAGCCAATCCCTGG-3′ and neoR: 5′-ATTCGCCGCCAAGCTCTTCAG-3′ were used to amplify the 5′ and 3′ UTRs respectively. The PCR mix contained 0.5 μM of each primer, 1x buffer STR (PROMEGA Corp.) and 0.5 U Taq polymerase (Gibco BRL Life Technology Corp.), and was incubated for 94 °C 4 min, follow by 35 cycles: 94 °C – 30 s, 50 °C – 30 s and 72 °C – 40 s and a final hold of 7 min at 72 °C. PCR products were, subjected to electrophoresis in 0.8% agarose gel, transferred to nylon membranes (Amersham) and hybridized to a γ^{32}P labeled $(dTdG)_{20}$ probe. Three replicas were analyzed.

2.4. RNA isolation and mRNA quantification

Epimastigote total RNA was extracted from 1×10^8 exponentially growing cells. For amastigote and trypomastigote, total RNA was obtained from 2×10^6 Vero cell cultures after 3 days of infection from cells or flask supernatant respectively. An RNA TRIZol (Qiagen) extraction followed by DNase I treatment with DNA-free (Ambion), according to manufacturer recommendations, was employed. cDNA was synthesized from 1 μg of total RNA using Superscript III kit first strand synthesis (Invitrogen) and oligodT primer. For quantification of retrotranscribed products (qRT-PCR), double stranded products were amplified using *cat* specific primers (CatFw: 5′- GCGTGTTACGGTGAAAACCT -3′ and CatRev: 5′-GGATTGGCTGAGACGAAAAA -3′) in a real time rotary analyzer RotorGene 6000 (Corbett). Relative amounts of target gene were calculated by normalization with the *gapdh* housekeeping gene using *gapdh* specific primers (GapdhFw: 5′-CGACAACGAGT GGGGATACT-3′ and GapdhRev: 5′-CTACAACCTTGCCGAACGAT-3′). PCR reaction mixture containing 0.9 μM of each primer, 1X QuantiTect SYBR Green PCR Master Mix (Qiagen), and 0.2 μL of cDNA template was carried out in a final volume of 10 μL. mRNA levels were compared using de $2^{-\Delta\Delta CT}$ method (Livak and Schmittgen, 2001).

Plasmid copy number was assessed measuring relative amounts of *cat* DNA over *gapdh* DNA by qPCR using the oligonucleotide primers described above. Total DNA was extracted with DNAzol (QIAGEN) under manufacturer's specifications.

2.5. mRNA decay

Actinomycin D (Sigma) (10 μg/mL) was added to pTEX-*cat*, pTEX-*cat* 5′CA and pTEX-*cat* 3′CA transfectant epimastigote cultures (2×10^7 cells/mL). 3 mL aliquots were removed at different time points (0, 4, 8 and 12 h) and RNA was extracted as described above. The fold change of *cat* mRNA was determined normalizing to *gapdh* mRNA and time zero rates were calculated using the $2^{-\Delta\Delta CT}$ method as described previously (Livak and Schmittgen, 2001).

2.6. Reporter protein measurement

CAT reporter protein levels were determined by western blot and ELISA assays. For western blot assays, protein extracts were separated by electrophoresis in 10% or 12% SDS–polyacrylamide gels and electro-transferred onto Hybond C Extra membranes (GE Helthcare) following standard procedures. Membranes were blocked by incubation in 5% skim milk powder in buffer PBS-0.1% Tween and were then incubated for 1 h at room temperature with polyclonal anti-CAT antibody (SIGMA), diluted 1:1,000. An anti-Tc38 polyclonal antibody (Duhagon et al., 2009), diluted 1:1,000, was used as internal control. Bound antibodies were detected using an IRDye 800CW Goat anti-Rabbi IgG (H + L) (Li-Cor), diluted 1:7,500 and analyzed in a G-Box (Syngene). CAT ELISA assays (ROCHE) were done following manufacturer's recommendations.

2.7. Statistical analysis

Chi square test was used to study deviation from random expectedly $(CA)_n$ occurrence at UTRs of categorized genes. mRNA statistical differences and variance were analyzed using Student test *t* and Fisher test *F* respectively.

3. Results and discussion

3.1. The CA repeated tracts are present at UTRs of many T. cruzi genes

We have previously demonstrated that *T. cruzi* genome contains polyCA (length $\geqslant$8 nt) that displayed a significant high occurrence in coding strands and are mainly located in the vicinity of open reading frames (Duhagon et al., 2011). In order to identify the genes that bear polyCA a genome-wide approach was pursued on the *T. cruzi* Esmeraldo strain genomic data (release-2.0). Since mRNA non-coding regions are barely annotated in trypanosomatids, the sizes of the UTRs were established based on previous estimation for *T. cruzi* (Campos et al., 2008). 10,005 sequences were retrieved and then analyzed searching for the presence of perfect polyCA of at least 8 nt long at the predicted UTRs. We found 1,081 genes (designated Tc.polyCA-UTR-set), from which 91.4% contain polyCA at the 3′ UTR (Table 1).

As depicted in Table 1, a great proportion of the genes containing polyCA (41.6%) code for hypothetical proteins, precluding the direct assignment of function. Interestingly, three hundred and seventy-eight genes (representing 35% of Tc.polyCA-UTR-set) belong to the *T. cruzi* exclusive mucin-associated surface protein family (MASP). Though the *T. cruzi* MASP gene family is composed by approximately 1400 genes (Bartholomeu et al., 2009), the Esmer-

Table 1
Description summary of *T. cruzi* genes containing polyCA at the UTRs.

	Total*	With $(CA)_n$&	%	$(CA)_n$ at 3′ UTR+	$(CA)_n$ at 5′ UTR+
Genes	8871	1081	12.2	91.4	9.1
Pseudogenes	1134	175	15.4	82.5	8.1
Description					
Mucin-associated surface protein (MASP)	414	378	91.3	100	
Hypothetical proteins	1295	450	34.7	90	10
Trans-sialidases	262	39	14.9	38.5	61.5
Kinases	304	15	4.9	93.3	6.7
Ribosomal proteins	132	14	10.6	50	50
Transferases	194	15	7.7	93.3	13.3
H/ACA snoRNA	69	7	10.2		
C/D snoRNA	117	7	6		
Peptidases	133	12	9	58.3	41.7
Syntaxin binding protein	2	2	100	100	
Mucins	267	2	0.7	100	
RNA binding proteins	423	6	1.4	100	
RNA helicase	348	6	1.7	100	
GTP binding proteins	23	5	21.7	60	40
Synthases	68	6	8.8	100	
DNA polymerases	23	4	17.4	100	
RNA polymerases	28	2	7.1	50	50
DNA j chaperon	30	2	6.7	100	
Mitochondrial carrier	17	3	17.6	100	
Hydrolases	38	3	7.9	100	
DNA repair	124	3	2.4	100	
Tubulin-tyrosin ligase	51	2	3.9	100	
Fatty acid elongase	57	2	3.5	100	
Translation iniciation factor	23	2	8.7	50	50
Cyclophilin	14	2	14.3	100	
Spliced leader PSE RNA transcription factor	1	1	100	100	
TSR1 splicing factor	1	1	100	100	
Peter pan protein. putative	1	1	100		100
Amino acid permease	14	2	14.3		100
Chaperonin	9	2	22.2	50	50
Neutral sphingomyelinase activ assoc. factor-like	1	1	100		100
Carboxylases	11	1	9.1		100
Tryparedoxin peroxidase. putative	4	1	25		100
Amino acid transporter	12	3	25	66.7	33.3
Transporters	59	4	6.8	75	25

* Number of genes on *Trypanosoma cruzi* genome.
& Number of genes on Tc.polyCA-UTR-data set.
\+ Percentage of genes with (CA)n at the indicated UTR. Some genes have the repeat element at both UTRs.

aldo set comprise only 414 genes annotated as MASP. Thus, a significant amount of these family's genes (91%) contain polyCA at their UTRs ($\chi 2$ $p < 0.001$). Apart from this, the representation of other gene families in the Tc.polyCA-UTR-set is not significantly different from what is randomly expected. For instance, for the trans-sialidase family, only 39 out of 262 genes carry polyCA at the UTRs ($\chi 2$ $p \geqslant 0.2$). Finally, the representations of the low copy number genes in the Tc.polyCA-UTR-set do not seem to display any bias.

Interestingly, we observed the simultaneous presence of polyGT close to polyCA in approximately half of the polyCA containing UTRs. This concurrence is significantly more frequent than expected by chance (χ^2 test $p \leqslant 0.05$). Indeed, while polyGT, of at least 8 nt, are present in 31% of the genes in *T. cruzi* genome, a higher proportion (51%) is observed in the Tc.polyCA-UTR-set (556 sequences out of the 1,081). Therefore, a tendency of these complementary polydinucleotides to co-exist in the parasite UTRs is suggested. This feature is present in all the 378 genes of the Tc.polyCA-UTR-set that encode MASP. Interestingly, 12 genes out of the 39 of the trans-sialidase family in the Tc.polyCA-UTR-set also share this characteristic. In addition, 134 genes encoding for hypothetical proteins and 32 genes distributed among the different categories listed in Table 1, also bear both polyCA and polyGT.

To search further connections among the genes in the Tc.polyCA-UTR-set, we performed two different analyses: a gene ontology, using Blast2GO, and the expression along the life cycle, using public microarray gene expression data (Minning et al., 2009). Probably due to the low conservation of *T. cruzi* protein domains in the eukaryotic gene evolution only 372 were assigned an ontology term. In agreement with our previous observation, a major group corresponding to the MASP family was identified both in the "biological process" and "molecular function" categories (198 and 185 respectively) (data not shown). The scarcity of data in the "cellular compartment" category disqualifies any assumption of relatedness. For the transcriptome analysis, no developmental gene clustering could be evidenced data are available just for 354 genes of the Tc.polyCA-UTR-set.

The high presence of polyCA at regulatory regions of the species-specific *masps* genes, together with their abundance in genes coding for hypothetical proteins, which denote proteins without known orthologues, may be pointing out a putative species-specific role for them. Concomitantly, the low abundance of this signal at regulatory regions of genes with orthologues in the TriTryps (12 out of 1,081) further supports this hypothesis (Supplementary Table 1).

3.2. The CA repeated tracts at the UTR are predicted to exhibit two main different secondary structures

In order to further investigate the function of the polyCA motifs, a preliminary study of the favored predicted secondary structure in the UTR context was undertaken.

The analysis of mRNA secondary structure for fifteen genes randomly selected from the Tc.polyCA-UTR-set shows that the thermodynamically favored conformations, for both 5′ and 3′ UTR, consist on a loop exhibiting the polyCA as a single strand (an example is shown in Fig. 1A, more examples are presented at Supplementary Figure 1). However, in ten randomly selected genes, whose UTRs also contain the polyGT, the favored secondary structures consistently display double stranded helical conformations due to the base pairing between the complementary polydinucleotides. One example is shown in Fig. 1B (for more examples see Supplementary Fig. 2).

These results suggest that polyCA in the UTRs of *T. cruzi* are present in two main different conformation involving single and double stranded structures. The extrapolation to the Tc.polyCA-UTR-set would mean that at least 50% of its genes present the CA repeated tract in a double stranded structure. Meanwhile, the other 50% of the genes of Tc.polyCA-UTR-set may exhibit the polyCA in a single stranded conformation. Caution must be taken with these results, since shorter elements, different location such as the close coding sequence, or even other elements, not as evident as the complementary polydinucleotide, also contribute to the actual conformation adopted.

3.3. The polyCA at 3′ UTR increases reporter mRNA stability in T. cruzi epimastigotes

Bearing in mind that we have previously demonstrated that polyCA constitute specific targets for single stranded nucleic acid binding proteins from *T. cruzi* epimastigotes (Duhagon et al., 2001) we studied their ability to modulate mRNA levels using a reporter gene assay. We selected the pTEX vector (Kelly et al., 1992) because it uses *gapdh* gene expression controlling regions, which allow RNA polymerase II driven transcription. This vector was modified for the expression of chloramphenicol acetyltransferase (CAT). We inserted CA repetitions either at the 5′ and 3′ UTR of the pTEX-*cat* plasmid generating pTEX-*cat* 3′CA and pTEX-*cat* 5′CA respectively (Supplementary Fig. 3A). The analysis of the secondary structure of the reporter gene UTRs indicated that the most thermodynamically favored structure for the reporter UTRs exhibit the inserted polyCA as a single stranded structure (Supplementary Fig. 3B).

We studied the reporter protein and mRNA steady state levels from *T. cruzi* epimastigotes transfected with pTEX-*cat*, pTEX-*cat* 5′CA and pTEX-*cat* 3′CA. We could not detect any change of CAT protein level using ELISA and western blot assays (data not shown). However, the transfected epimastigotes carrying the insertion of the polyCA at the 3′ UTR exhibit an approximately 2.5-fold increase of *cat* mRNA relative steady state level (t test $p \leqslant 0.05$) (Fig. 2A). The absence of correlation between the observed changes in reporter mRNA and protein levels might be due to the saturation of the translational machinery masking the mRNA variation.

In *T. cruzi*, as in kinetoplastids, genes are constitutively transcribed. However, differences of steady state mRNA levels along the life cycle stages have been observed (Minning et al., 2009). Those variations have been mostly attributed to mRNA half life differences (D'Orso et al., 2003; D'Orso and Frasch, 2001; Di Noia et al., 2000; Kramer and Carrington, 2011; Martinez-Calvillo et al., 2010). Therefore, we decided to analyze if the polyCA located at the 3′ and 5′ UTRs of the reporter gene have any effect on its mRNA stability. For this purpose, the transfected epimastigotes were incubated in the presence of Actinomycin D for transcription inhibition (Coughlin et al., 2000; da Silva et al., 2006), and total RNA was extracted and purified at various time points. The amount of *cat* mRNA relative to *gapdh* (half life approximately 7 h) (Nozaki and Cross, 1995) was quantified by qRT-PCR and then, the relative mRNA stability was calculated (Fig. 2B). No significant differences between *cat* and *gapdh* mRNA stability was observed for the transfectants carrying the control vector or the insertion of the polyCA in the 5′ UTR. However, a significant half life increase of relative reporter stability in the transfectants bearing the polyCA at the 3′ UTR is observed (t test $p \leqslant 0.05$). This finding goes in agreement with the higher mRNA steady state level previously observed. In conclusion, a significant increase of reporter mRNA steady state and half life due to the presence of single-stranded polyCA at its 3′ UTR was observed, suggesting the repeat *per se* may be responsible for the selective increase of mRNA abundance in *T. cruzi* epimastigotes.

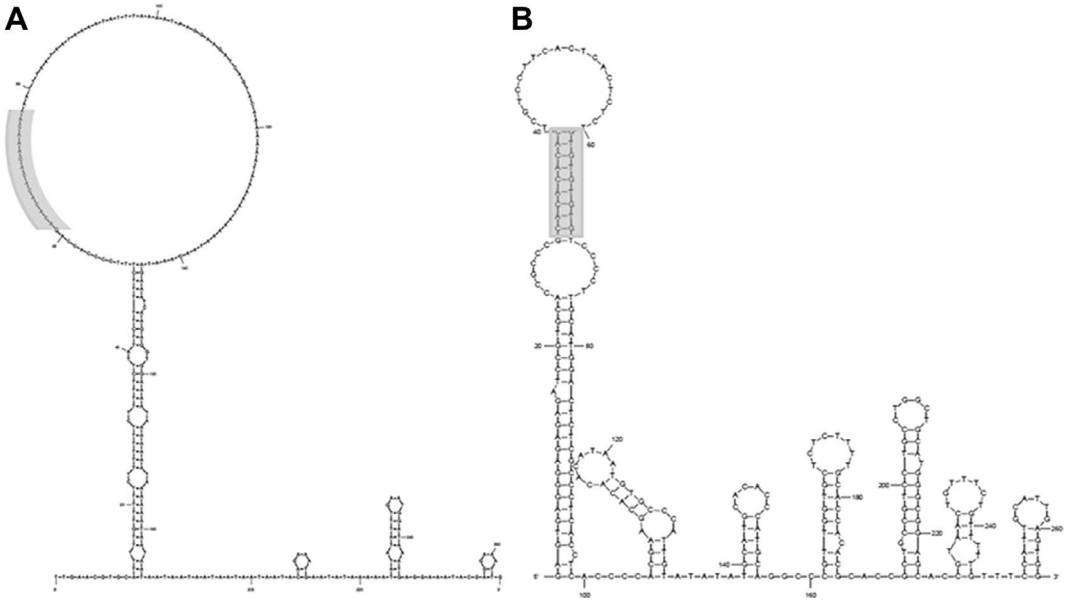


Fig. 1. Predicted secondary structure of UTRs containing polyCA A. 3′ UTR of the Tc00.1047053503841.70 gene encoding an aspartate aminotransferase, putative protein. B. 3′ UTR of the Tc00.1047053506971.10 gene encoding a MASP. The shown conformations correspond to the lowest free energy generated by The DINAMelt Server for Prediction of Melting Profiles for Nucleic Acids using default parameters. The polyCA location is highlighted.

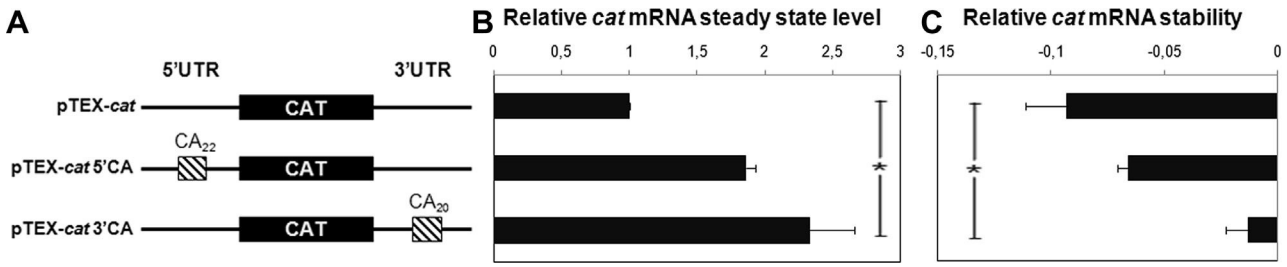


Fig. 2. Effect of the presence of polyCA at the UTRs on reporter gene expression in *T. cruzi* epimastigotes. Schematic representation of reporter vectors used: pTEX-*cat* (control) and with the insertion of polyCA at both 5′ and 3′ UTR (pTEX-*cat* 5′CA and pTEX-*cat* 3′CA, respectively). The length of the CA_n insertion is indicated. A. The mRNA steady state levels of *cat* relative to *gapdh* of *T. cruzi* epimastigotes transfected with pTEX-*cat*, pTEX-*cat* 5′CA and pTEX-*cat* 3′CA, after normalization to the vector's copy number is shown. B. Relative reporter mRNA stability, calculated as the time course slopes of *cat* relative to *gapdh* is indicated. Three biological replicas were analyzed in independent triplicates (mean values ± standard error, SE) (☆) Statistical significance difference (*t* test $p \leqslant 0.05$).

3.4. The polyCA at 3′ UTR differentially affects reporter mRNA steady state along T. cruzi life cycle

In order to determine if the stabilizing effect of the single stranded polyCA at the 3′ UTR on gene expression was a stage-specific phenomenon, we analyzed the mRNA level of the reporter gene along *T. cruzi* life cycle.

Epimastigotes transfected with the pTEX-*cat* (control) and pTEX-*cat* 3′CA were exposed to nutritional stress in order to induce the differentiation into metacyclic trypomastigotes and later used to infect Vero cells to obtain cellular amastigotes. The mRNA steady state levels of the *cat* reporter relative to *gapdh* were quantified and normalized to the respective plasmid copy number (Fig. 3). As previously shown, the transfected epimastigotes carrying the insertion of the polyCA at the 3′ UTR exhibit a significant increase of *cat* mRNA relative steady state level compared to the control parasites (*t* test $p \leqslant 0.05$). A significant variation, but in the opposite direction, is observed for the transfected amastigotes. Indeed, the transfected pTEX-*cat* 3′CA amastigotes show a lower relative steady state level of *cat* mRNA compared to the control parasites (*t* test $p \leqslant 0.05$). We have also studied the CAT protein levels in the transfected amastigotes but we have no detected any change (data not shown). At the trypomastigote stage, no significant differences between the relative steady state levels of the mRNA reporter were observed for the parasites transfected with pTEX-*cat* and pTEX-*cat* 3′CA. Our results suggest that the polyCA at the 3′ UTR has the ability to modulate the amount of the transcript in a stage specific manner. Whether the observed effect is caused by the presence of the polyCA or by the alteration of an unknown signal at the reporter 3′ UTR, remains to be determined. However, in the context of the UTR of a constitutively expressed gene, as *gapdh*, it is tempting to speculate that the polyCA is the signal accounting for the stage specific changes in mRNA abundance.

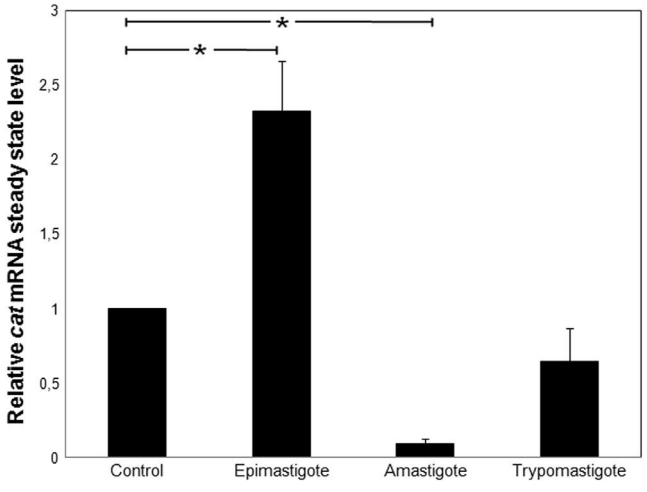


Fig. 3. Effect of the presence of polyCA at 3′UTR on reporter gene expression in *T. cruzi* life cycle stages. The mRNA steady state levels of *cat* relative to *gapdh* for pTEX-*cat* 3′CA transfected *T. cruzi*, after normalization to the vector copy number, is shown for the parasite life cycle stages: epimastigotes, amastigotes and trypomastigotes. Three independent experiments of two biological replicates were done. (∗) Statistical significance difference (*t* test $p \leqslant 0.05$).

4. Conclusions

We have previously found that polyCA are highly represented in the vicinity of open reading frames and constitute specific targets for single stranded binding proteins from *T. cruzi* epimastigote extracts (Duhagon et al., 2001, 2011). In order to investigate its putative role, we firstly identify the genes containing this sequence in their regulatory regions. We found that 10% of *T. cruzi* genes present polyCA at their UTRs, mainly at the 3′ UTR (91.4%). The high representation of this sequence in genes exclusive to *T. cruzi* indicates a putative species-specific regulatory role in gene expression.

In addition, we discovered that half of the genes that contain polyCA at the UTRs also contain polyGT, which thermodynamically favors a UTR secondary structure with the polyCA in double stranded stems. In this latter category, the vast majority (91%) of the trypomastigote stage-specific family of genes coding for mucine associated proteins (MASPs) are included. Due to the relevance of these genes in *T. cruzi* parasitism, it would be interesting to study the influence of the GT/CA motif in the expression of these particular genes.

Regarding the genes bearing a polyCA tract in the UTRs and not any polyGT tracts, alternative secondary structures exhibiting the polyCA as single stranded sequences or loops are predicted. This result indicates that this sequence might be directly accessible for the recognition by specific single stranded binding proteins, in agreement with our previous reports (Duhagon et al., 2001). This further supports a specific role of polyCA sequences and its cognate DNA binding proteins in the regulation of the gene expression of this subset of genes (approximately 500).

Using the CAT reporter assay, we found that in epimastigotes, the polyCA sequences only modulate gene expression when located at the 3′ UTR. In agreement with this finding, most mRNA stability regulators are located at the 3′ UTR (Li et al., 2012; Rattenbacher and Bohjanen, 2012; Suganuma et al., 2012; Zhao et al., 2011).

The analysis of polyCA modulation along the life cycle of the parasite, demonstrates that polyCA at 3′ UTR increases reporter mRNA stability in *T. cruzi* epimastigotes, while do not have any significant effect in the trypomastigote stage, and significantly diminishes reporter mRNA steady state in the amastigote stage. These

results suggest that the polyCA elements in the 3′ UTR would function in the stabilization of epimastigote specific genes. However, through the mining of the array gene expression data (Minning et al., 2009), we were unable to find an association between the presence of polyCA in the 3′ UTR and stage specific transcript abundance. This apparent discrepancy could be reconciled considering that gene expression profile is a consequence of the presence of various signals acting in combination, either enhancing or masking the specific element. For instance, the concurrent presence of polyCA and polyGT leading to the double stranded secondary structure predicted for the *masp* family could justify why this genes are not preferentially expressed in the epimastigote stage. The coordinated expression of set of related messengers as post-transcriptional regulons, that may involve multiple signals, has been proposed in eukaryotes (Keene, 2007) and more recently in *T. brucei* (Queiroz et al., 2009). In this context, it is tempting to speculate that alternative conformations of the polyCA may be driving to specific regulatory mechanisms. Experiments assessing the effect of native polyCA in endogenous *T. cruzi* genes would allow pursuing this hypothesis.

The data presented here support the involvement of polyCA elements in *T. cruzi* gene expression regulation, granting further studies to elucidate the specific genes and molecular mechanisms responsible for its effects.

Acknowledgments

We thank Juan Martín Marqués (UdelaR) for very helpful suggestion and laboratory use. This work was financially supported by PEDECIBA and CSIC, UdelaR. LP received an UdelaR scholarship.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.exppara.2013.04.004.

References

Abuin, G., Freitas-Junior, L.H., Colli, W., Alves, M.J., Schenkman, S., 1999. Expression of trans-sialidase and 85-kDa glycoprotein genes in *Trypanosoma cruzi* is differentially regulated at the post-transcriptional level by labile protein factors. J. Biol. Chem. 274, 13041–13047.

Araujo, P.R., Burle-Caldas, G.A., Silva, R.A., Bartolomeu, D.C., Darocha, W.D., Teixeira, S.M., 2011. Development of a dual reporter system to identify regulatory cis-acting elements in untranslated regions of Trypanosoma cruzi mRNAs. Parasitol Int. 60, 161–169.

Avila, A.R., Yamada-Ogatta, S.F., da Silva Monteiro, V., Krieger, M.A., Nakamura, C.V., de Souza, W., Goldenberg, S., 2001. Cloning and characterization of the metacyclogenin gene, which is specifically expressed during *Trypanosoma cruzi* metacyclogenesis. Mol. Biochem. Parasitol. 117, 169–177.

Baranovskaya, S., Martin, Y., Alonso, S., Pisarchuk, K.L., Falchetti, M., Dai, Y., Khaldoyanidi, S., Krajewski, S., Novikova, I., Sidorenko, Y.S., Perucho, M., Malkhosyan, S.R., 2009. Down-regulation of epidermal growth factor receptor by selective expansion of a 5′-end regulatory dinucleotide repeat in colon cancer with microsatellite instability. Clin. Cancer Res. 15, 4531–4537.

Bartholomeu, D.C., Cerqueira, G.C., Leao, A.C., daRocha, W.D., Pais, F.S., Macedo, C., Djikeng, A., Teixeira, S.M., El-Sayed, N.M., 2009. Genomic organization and expression profile of the mucin-associated surface protein (masp) family of the human pathogen *Trypanosoma cruzi*. Nucleic Acids Res. 37, 3407–3417.

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

Buisine, M.P., Wacrenier, A., Mariette, C., Leteurtre, E., Escande, F., Aissi, S., Ketele, A., Leclercq, A., Porchet, N., Lesuffleur, T., 2008. Frequent mutations of the CA simple sequence repeat in intron 1 of EGFR in mismatch repair-deficient colorectal cancers. World J. Gastroenterol. 14, 1053–1059.

Buratti, E., Dork, T., Zuccato, E., Pagani, F., Romano, M., Baralle, F.E., 2001. Nuclear factor TDP-43 and SR proteins promote in vitro and in vivo CFTR exon 9 skipping. EMBO J. 20, 1774–1784.

Campos, P.C., Bartholomeu, D.C., DaRocha, W.D., Cerqueira, G.C., Teixeira, S.M., 2008. Sequences involved in mRNA processing in *Trypanosoma cruzi*. Int. J. Parasitol. 38, 1383–1389.

Clayton, C., 2010. Repetitive elements in parasitic protozoa. BMC Biol. 8, 64.

Coderre, J.A., Beverley, S.M., Schimke, R.T., Santi, D.V., 1983. Overproduction of a bifunctional thymidylate synthetase-dihydrofolate reductase and DNA amplification in methotrexate-resistant Leishmania tropica. Proc. Natl. Acad. Sci. USA 80, 2132–2136.

Contreras, V.T., Araujo-Jorge, T.C., Bonaldo, M.C., Thomaz, N., Barbosa, H.S., Meirelles Mde, N., Goldenberg, S., 1988. Biological aspects of the Dm 28c clone of *Trypanosoma cruzi* after metacyclogenesis in chemically defined media. Mem. Inst. Oswaldo Cruz 83, 123–133.

Contreras, V.T., Morel, C.M., Goldenberg, S., 1985. Stage specific gene expression precedes morphological changes during *Trypanosoma cruzi* metacyclogenesis. Mol. Biochem. Parasitol. 14, 83–96.

Coughlin, B.C., Teixeira, S.M., Kirchhoff, L.V., 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, 12051–12060.

Chagas, C., 1909. Nova tripanozomiaze humana. Estudos sobre a morfologia e o ciclo evolutivo do Schizotrypanum cruzi, agente etiologico da nova entidade mórbida do homen. Mem. Inst. Oswaldo Cruz 1, 159–219.

Chiribao, M.L., Libisch, M.G., Osinaga, E., Parodi-Talice, A., Robello, C., 2012. Cloning, localization and differential expression of the *Trypanosoma cruzi* TcOGNT-2 glycosyl transferase. Gene 498, 147–154.

D'Orso, I., De Gaudenzi, J.G., Frasch, A.C., 2003. RNA-binding proteins and mRNA turnover in trypanosomes. Trends Parasitol. 19, 151–155.

D'Orso, I., 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, 15783–15793.

da Silva, R.A., Bartholomeu, D.C., Teixeira, S.M., 2006. Control mechanisms of tubulin gene expression in *Trypanosoma cruzi*. Int. J. Parasitol. 36, 87–96.

De Gaudenzi, J.G., Noe, G., Campo, V.A., Frasch, A.C., Cassola, A., 2011. Gene expression regulation in trypanosomatids. Essays Biochem. 51, 31–46.

Di Noia, J.M., D'Orso, I., Sanchez, D.O., 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.

Duhagon, M.A., Dallagiovanna, B., 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, 98–103.

Duhagon, M.A., Pastro, L., Sotelo-Silveira, J.R., Perez-Diaz, L., Maugeri, D., Nardelli, S.C., Schenkman, S., Williams, N., Dallagiovanna, B., Garat, B., 2009. The *Trypanosoma cruzi* nucleic acid binding protein Tc38 presents changes in the intramitochondrial distribution during the cell cycle. BMC Microbiol. 9, 34.

Duhagon, M.A., Smircich, P., Forteza, D., Naya, H., Williams, N., Garat, B., 2011. Comparative genomic analysis of dinucleotide repeats in Tritryps. Gene 487, 29–37.

Gabellini, N., 2001. A polymorphic GT repeat from the human cardiac Na+Ca2+ exchanger intron 2 activates splicing. Eur. J. Biochem. 268, 1076–1083.

Gendrel, C.G., Boulet, A., Dutreix, M., 2000. (CA/GT)(n) microsatellites affect homologous recombination during yeast meiosis. Genes Dev. 14, 1261–1268.

Gentil, L.G., Cordero, E.M., do Carmo, M.S., dos Santos, M.R., Silveira, J.F., 2009. Posttranscriptional mechanisms involved in the control of expression of the stage-specific GP82 surface glycoprotein in *Trypanosoma cruzi*. Acta Trop. 109, 152–158.

Jager, A.V., Muia, R.P., Campetella, O., 2008. Stage-specific expression of *Trypanosoma cruzi* trans-sialidase involves highly conserved 3′ untranslated regions. FEMS Microbiol. Lett. 283, 182–188.

Keene, J.D., 2007. RNA regulons: coordination of post-transcriptional events. Nat. Rev. Genet. 8, 533–543.

Kelly, J.M., Ward, H.M., Miles, M.A., Kendall, G., 1992. A shuttle vector which facilitates the expression of transfected genes in *Trypanosoma cruzi* and *Leishmania*. Nucleic Acids Res. 20, 3963–3969.

Kramer, S., 2012. Developmental regulation of gene expression in the absence of transcriptional control: the case of kinetoplastids. Mol. Biochem. Parasitol. 181, 61–72.

Kramer, S., Carrington, M., 2011. Trans-acting proteins regulating mRNA maturation, stability and translation in trypanosomatids. Trends Parasitol. 27, 23–30.

Lee, J.H., Jeon, M.H., Seo, Y.J., Lee, Y.J., Ko, J.H., Tsujimoto, Y., 2004. CA repeats in the 3′-untranslated region of bcl-2 mRNA mediate constitutive decay of bcl-2 mRNA. J. Biol. Chem. 279, 42758–42764.

Li, Z.H., De Gaudenzi, J.G., Alvarez, V.E., Mendiondo, N., Wang, H., Kissinger, J., Frasch, A.C., Docampo, R., 2012. A 43-nt U-rich element in the 3'untranslated region of a large number of Trypanosoma cruzi transcripts is important for mRNA abundance in intracellular amastigotes. J Biol Chem. 287, 19058–19069.

Lin, Z., Thomas, N.J., Wang, Y., Guo, X., Seifart, C., Shakoor, H., Floros, J., 2005. Deletions within a CA-repeat-rich region of intron 4 of the human SP-B gene affect mRNA splicing. Biochem. J. 389, 403–412.

Livak, K.J., Schmittgen, T.D., 2001. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. Methods 25, 402–408.

Majewski, J., Ott, J., 2000. GT repeats are associated with recombination on human chromosome 22. Genome Res. 10, 1108–1114.

Martin-Farmer, J., Janssen, G.R., 1999. A downstream CA repeat sequence increases translation from leadered and unleadered mRNA in *Escherichia coli*. Mol. Microbiol. 31, 1025–1038.

Martinez-Calvillo, S., Vizuet-de-Rueda, J.C., Florencio-Martinez, L.E., Manning-Cela, R.G., Figueroa-Angulo, E.E., 2010. Gene expression in trypanosomatid parasites. J. Biomed. Biotechnol. 2010, 525241.

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

Noe, G., De Gaudenzi, J.G., Frasch, A.C., 2008. Functionally related transcripts have common RNA motifs for specific RNA-binding proteins in trypanosomes. BMC Mol. Biol. 9, 107.

Nozaki, T., Cross, G.A., 1995. Effects of 3′ untranslated and intergenic regions on gene expression in *Trypanosoma cruzi*. Mol. Biochem. Parasitol. 75, 55–67.

Perez-Diaz, L., Duhagon, M.A., Smircich, P., Sotelo-Silveira, J., Robello, C., Krieger, M.A., Goldenberg, S., Williams, N., Dallagiovanna, B., Garat, B., 2007. *Trypanosoma cruzi*: molecular characterization of an RNA binding protein differentially expressed in the parasite life cycle. Exp. Parasitol. 117, 99–105.

Queiroz, R., Benz, C., Fellenberg, K., Hoheisel, J.D., Clayton, C., 2009. Transcriptome analysis of differentiating trypanosomes reveals the existence of multiple post-transcriptional regulons. BMC Genomics 10, 495.

Rattenbacher, B., Bohjanen, P.R., 2012. Evaluating posttranscriptional regulation of cytokine genes. Methods Mol. Biol. 820, 71–89.

Rockman, M.V., Wray, G.A., 2002. Abundant raw material for cis-regulatory evolution in humans. Mol. Biol. Evol. 19, 1991–2004.

Rodrigues, D.C., Silva, R., Rondinelli, E., Urmenyi, T.P., 2010. *Trypanosoma cruzi*: modulation of HSP70 mRNA stability by untranslated regions during heat shock. Exp. Parasitol. 126, 245–253.

Saeed, A.I., Sharov, V., White, J., Li, J., Liang, W., Bhagabati, N., Braisted, J., Klapa, M., Currier, T., Thiagarajan, M., Sturn, A., Snuffin, M., Rezantsev, A., Popov, D., Ryltsov, A., Kostukovich, E., Borisovsky, I., Liu, Z., Vinsavich, A., Trush, V., Quackenbush, J., 2003. TM4: a free, open-source system for microarray data management and analysis. Biotechniques 34, 374–378.

Sinden, R., 1994. DNA Structure and Function. Academic Press, San Diego.

Songthamwat, D., Kajihara, K., Kikuchi, M., Uemura, H., Tran, S.P., Yanagi, T., Higo, H., Hirayama, K., 2007. Structure and expression of three gp82 gene subfamilies of *Trypanosoma cruzi*. Parasitol. Int. 56, 273–280.

Suganuma, K., Yamasaki, S., Asada, M., Kawazu, S., Inoue, N., 2012. The epimastigote stage-specific gene expression of CESP is tightly regulated by its 3′ UTR. Mol. Biochem. Parasitol. 186, 77–80.

Teixeira, S.M., Russell, D.G., Kirchhoff, L.V., Donelson, J.E., 1994. A differentially expressed gene family encoding "amastin", a surface protein of *Trypanosoma cruzi* amastigotes. J. Biol. Chem. 269, 20509–20516.

Tyler, K.M., Engman, D.M., 2001. The life cycle of Trypanosoma cruzi revisited. Int. J. Parasitol. 31, 472–481.

Weston, D., La Flamme, A.C., Van Voorhis, W.C., 1999. Expression of *Trypanosoma cruzi* surface antigen FL-160 is controlled by elements in the 3′ untranslated, the 3′ intergenic, and the coding regions. Mol. Biochem. Parasitol. 102, 53–66.

Zhao, W., Blagev, D., Pollack, J.L., Erle, D.J., 2011. Toward a systematic understanding of mRNA 3′ untranslated regions. Proc. Am. Thorac. Soc. 8, 163–166.

Bibliografía

Akopyants, N. S. *et al.* (2009). *Demonstration of genetic exchange during cyclical development of Leishmania in the sand fly vector*. Science 324(5924): 265-8.

Alegria-Schaffer, A. *et al.* (2009). *Performing and optimizing Western blots with an emphasis on chemiluminescent detection*. Methods Enzymol 463: 573-99.

Almeida, R. *et al.* (2002). *From genomes to vaccines: Leishmania as a model*. Philos Trans R Soc Lond B Biol Sci 357(1417): 5-11.

Alvarez, F. *et al.* (1994). *Evolution of codon usage and base contents in kinetoplastid protozoans*. Mol Biol Evol 11(5): 790-802.

Andersson, B. *et al.* (1998). *Complete sequence of a 93.4-kb contig from chromosome 3 of Trypanosoma cruzi containing a strand-switch region*. Genome Res 8(8): 809-16.

Andrews, N. W. *et al.* (1987). *Stage-specific surface antigens expressed during the morphogenesis of vertebrate forms of Trypanosoma cruzi*. Exp Parasitol 64(3): 474-84.

Aphasizhev, R. *et al.* (2003). *Isolation of a U-insertion/deletion editing complex from Leishmania tarentolae mitochondria*. Embo J 22(4): 913-24.

Araripe, J. R. *et al.* (2004). *Trypanosoma cruzi: TcRAB7 protein is localized at the Golgi apparatus in epimastigotes*. Biochem Biophys Res Commun 321(2): 397-402.

Arner, E. *et al.* (2007). *Database of Trypanosoma cruzi repeated genes: 20,000 additional gene variants*. BMC Genomics 8: 391.

Atayde, V. D. *et al.* (2011). *The emerging world of small silencing RNAs in protozoan parasites*. Trends Parasitol 27(7): 321-7.

Atwood, J. A. *et al.* (2005). *The Trypanosoma cruzi proteome*. Science 309(5733): 473-6.

Avila, A. R. *et al.* (2001). *Cloning and characterization of the metacyclogenin gene, which is specifically expressed during Trypanosoma cruzi metacyclogenesis*. Mol Biochem Parasitol 117(2): 169-77.

Balagopal, V. *et al.* (2009). *Polysomes, P bodies and stress granules: states and fates of eukaryotic mRNAs*. Curr Opin Cell Biol 21(3): 403-8.

Bangs, J. D. *et al.* (1992). *Mass spectrometry of mRNA cap 4 from trypanosomatids reveals two novel nucleosides*. J Biol Chem 267(14): 9805-15.

Barreau, C. *et al.* (2005). *AU-rich elements and associated factors: are there unifying principles?* Nucleic Acids Res 33(22): 7138-50.

Barrett, M. P. *et al.* (2003). *The trypanosomiases*. Lancet 362(9394): 1469-80.

Bartholomeu, D. C. *et al.* (2009). *Genomic organization and expression profile of the mucin-associated surface protein (masp) family of the human pathogen Trypanosoma cruzi*. Nucleic Acids Res 37(10): 3407-17.

Bayer-Santos, E. *et al.* (2012). *Regulatory elements in the 3' untranslated region of the GP82 glycoprotein are responsible for its stage-specific expression in Trypanosoma cruzi metacyclic trypomastigotes*. Acta Trop 123(3): 230-3.

Bechert, T. *et al.* (1999). *All 16 centromere DNAs from Saccharomyces cerevisiae show DNA curvature*. Nucleic Acids Res 27(6): 1444-9.

Bell, S. P. *et al.* (1990). *Assembly of alternative multiprotein complexes directs rRNA promoter selectivity*. Genes Dev 4(6): 943-54.

Ben Amar, M. F., Jefferies, D., Pays, A., Bakalara, N., Kendall, G., Pays, E. (1991). *The actin gene promoter of Trypanosoma brucei.* Nucleic Acids Res. 19: 5857-5862.

Benjamini, Y. *et al.* (2001). *Controlling the false discovery rate in behavior genetics research*. Behav Brain Res 125(1-2): 279-84.

Benson, G. (1999). *Tandem repeats finder: a program to analyze DNA sequences*. Nucleic Acids Res 27(2): 573-80.

Bern, C. *et al.* (2011). *Trypanosoma cruzi and Chagas' Disease in the United States*. Clin Microbiol Rev 24(4): 655-81.

Berriman, M. *et al.* (2005). *The genome of the African trypanosome Trypanosoma brucei*. Science 309(5733): 416-22.

Berriman, M. *et al.* (2002). *The architecture of variant surface glycoprotein gene expression sites in Trypanosoma brucei*. Mol Biochem Parasitol 122(2): 131-40.

Beverley, S. M. *et al.* (2002). *Putting the Leishmania genome to work: functional genomics by transposon trapping and expression profiling*. Philos Trans R Soc Lond B Biol Sci 357(1417): 47-53.

Blackwell, J. M. *et al.* (1999). *Status of protozoan genome analysis: trypanosomatids*. Parasitology 118 Suppl: S11-4.

Blumenthal, T. *et al.* (2002). *A global analysis of Caenorhabditis elegans operons*. Nature 417(6891): 851-4.

Borst, P. (2002). *Antigenic variation and allelic exclusion*. Cell 109(1): 5-8.

Borst, P. *et al.* (1998). *Control of VSG gene expression sites in Trypanosoma brucei*. Mol Biochem Parasitol 91(1): 67-76.

Borst, P. *et al.* (2008). *Base J: discovery, biosynthesis, and possible functions*. Annu Rev Microbiol 62: 235-51.

Boucher, N. *et al.* (2002). *A common mechanism of stage-regulated gene expression in Leishmania mediated by a conserved 3'-untranslated region element*. J Biol Chem 277(22): 19511-20.

Branche, C. *et al.* (2006). *Comparative karyotyping as a tool for genome structure analysis of Trypanosoma cruzi*. Mol Biochem Parasitol 147(1): 30-8.

Brandao, A. *et al.* (1997). *Identification of transcribed sequences (ESTs) in the Trypanosoma cruzi genome project*. Mem Inst Oswaldo Cruz 92(6): 863-6.

Brandenburg, J. *et al.* (2007). *Multifunctional class I transcription in Trypanosoma brucei depends on a novel protein complex*. EMBO J 26(23): 4856-66.

Brewis, I. A. *et al.* (2010). *Proteomics technologies for the global identification and quantification of proteins*. Adv Protein Chem Struct Biol 80: 1-44.

Bringaud, F. *et al.* (2007). *Members of a large retroposon family are determinants of post-transcriptional gene expression in Leishmania*. PLoS Pathog 3(9): 1291-307.

Brittingham, A. *et al.* (1995). *Role of the Leishmania surface protease gp63 in complement fixation, cell adhesion, and resistance to complement-mediated lysis*. J Immunol 155(6): 3102-11.

Brooks, D. R. *et al.* (2001). *The stage-regulated expression of Leishmania mexicana CPB cysteine proteases is mediated by an intercistronic sequence element*. J Biol Chem 276(50): 47061-9.

Brun, R. *et al.* (2010). *Human African trypanosomiasis*. Lancet 375(9709): 148-59.

Bullard, J. H. *et al.* (2010). *Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments*. BMC Bioinformatics 11: 94.

Campbell, D. A. *et al.* (2000). *Transcription of the kinetoplastid spliced leader RNA gene*. Parasitol Today 16(2): 78-82.

Cantey, P. T. *et al.* (2012). *The United States Trypanosoma cruzi Infection Study: evidence for vector-borne transmission of the parasite that causes Chagas disease among United States blood donors*. Transfusion 52(9): 1922-30.

Caradonna, K. L. *et al.* (2011). *Mechanisms of host cell invasion by Trypanosoma cruzi*. Adv Parasitol 76: 33-61.

Cassola, A. (2011). *RNA Granules Living a Post-transcriptional Life: the Trypanosomes' Case*. Curr Chem Biol 5(2): 108-117.

Cassola, A. *et al.* (2007). *Recruitment of mRNAs to cytoplasmic ribonucleoprotein granules in trypanosomes*. Mol Microbiol 65(3): 655-70.

Cattand, P. *et al.* (2001). *Sleeping sickness surveillance: an essential step towards elimination*. Trop Med Int Health 6(5): 348-61.

Cazzulo, J. J. (1994). *Intermediate metabolism in Trypanosoma cruzi*. J Bioenerg Biomembr 26(2): 157-65.

Clayton, C. *et al.* (2007). *Post-transcriptional regulation of gene expression in trypanosomes and leishmanias*. Mol Biochem Parasitol 156(2): 93-101.

Clayton, C. E. (2002). *Life without transcriptional control? From fly to man and back again*. Embo J 21(8): 1881-8.

Cliffe, L. J. *et al.* (2010). *Two thymidine hydroxylases differentially regulate the formation of glucosylated DNA at regions flanking polymerase II polycistronic transcription units throughout the genome of Trypanosoma brucei*. Nucleic Acids Res.

Colasante, C. *et al.* (2007). *Regulated expression of glycosomal phosphoglycerate kinase in Trypanosoma brucei*. Mol Biochem Parasitol 151(2): 193-204.

Coller, J. *et al.* (2004). *Eukaryotic mRNA decapping*. Annu Rev Biochem 73: 861-90.

Coller, J. M. *et al.* (2001). *The DEAD box helicase, Dhh1p, functions in mRNA decapping and interacts with both the decapping and deadenylase complexes*. RNA 7(12): 1717-27.

Costales, J. *et al.* (2007). *A role for protease activity and host-cell permeability during the process of Trypanosoma cruzi egress from infected cells*. J Parasitol 93(6): 1350-9.

Coughlin, B. C. *et al.* (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): 12051-60.

Coura, J. R. *et al.* (2010). *Chagas disease: a new worldwide challenge*. Nature 465(n7301_supp): S6-S7.

Cribb, P. *et al.* (2009). *One- and two-hybrid analysis of the interactions between components of the Trypanosoma cruzi spliced leader RNA gene promoter binding complex*. Int J Parasitol 39(5): 525-32.

Croken, M. M. *et al.* (2012). *Chromatin modifications, epigenetics, and how protozoan parasites regulate their lives*. Trends Parasitol 28(5): 202-13.

Crothers, D. M. (1998). *DNA curvature and deformation in protein-DNA complexes: a step in the right direction*. Proc Natl Acad Sci U S A 95(26): 15163-5.

Cunningham, I. (1977). *New culture medium for maintenance of tsetse tissues and growth of trypanosomatids*. J Protozool 24(2): 325-9.

Chagas, C. (1909). *Nova tripanozomiaze humana. Estudos sobre a morfologia e o ciclo evolutivo do Schizotrypanum cruzi, agente etiologico da nova entidade mórbida do homen.* Mem. Inst. Oswaldo Cruz 1: 159-219.

Checchi, F. *et al.* (2008). *Estimates of the duration of the early and late stage of gambiense sleeping sickness*. BMC Infect Dis 8: 16.

Chen, C. Y. *et al.* (2011). *Mechanisms of deadenylation-dependent decay*. Wiley Interdiscip Rev RNA 2(2): 167-83.

Chen, R. A. *et al.* (2013). *The landscape of RNA polymerase II transcription initiation in C. elegans reveals promoter and enhancer architectures*. Genome Res.

D'Orso, I. *et al.* (2003). *RNA-binding proteins and mRNA turnover in trypanosomes*. Trends Parasitol 19(4): 151-5.

D'Orso, I. *et al.* (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): 15783-93.

D'Orso, I. *et al.* (2001). *TcUBP-1, a developmentally regulated U-rich RNA-binding protein involved in selective mRNA destabilization in trypanosomes*. J Biol Chem 276(37): 34801-9.

D'Orso, I. *et al.* (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): 50520-8.

da Silva, R. A. *et al.* (2006). *Control mechanisms of tubulin gene expression in Trypanosoma cruzi*. Int J Parasitol 36(1): 87-96.

da Silveira, A. B. *et al.* (2007). *Megacolon in Chagas disease: a study of inflammatory cells, enteric nerves, and glial cells*. Hum Pathol 38(8): 1256-64.

Dallagiovanna, B. *et al.* (2008). *Functional genomic characterization of mRNAs associated with TcPUF6, a pumilio-like protein from Trypanosoma cruzi*. J Biol Chem 283(13): 8266-73.

Dallagiovanna, B. *et al.* (2005). *Trypanosoma cruzi: Molecular characterization of TcPUF6, a Pumilio protein*. Exp Parasitol 109(4): 260-4.

Dallagiovanna, B. *et al.* (2001). *Trypanosoma cruzi: a gene family encoding chitin-binding-like proteins is posttranscriptionally regulated during metacyclogenesis*. Exp Parasitol 99(1): 7-16.

Dana, A. *et al.* (2012). *Determinants of translation elongation speed and ribosomal profiling biases in mouse embryonic stem cells*. PLoS Comput Biol 8(11): e1002755.

Daniels, J. P. *et al.* (2010). *Cell biology of the trypanosome genome*. Microbiol Mol Biol Rev 74(4): 552-69.

Daniels, J. P. *et al.* (2012). *The trypanosomatid-specific N terminus of RPA2 is required for RNA polymerase I assembly, localization, and function*. Eukaryot Cell 11(5): 662-72.

Das, A. *et al.* (2008). *RNA polymerase transcription machinery in trypanosomes*. Eukaryot Cell 7(3): 429-34.

Das, A. *et al.* (2003). *RNA polymerase II-dependent transcription in trypanosomes is associated with a SNAP complex-like transcription factor*. Proc Natl Acad Sci U S A 100(1): 80-5.

Das, A. *et al.* (2005). *Trypanosomal TBP functions with the multisubunit transcription factor tSNAP to direct spliced-leader RNA gene expression*. Mol Cell Biol 25(16): 7314-22.

de Carvalho, T. U. *et al.* (1986). *Infectivity of amastigotes of Trypanosoma cruzi*. Rev Inst Med Trop Sao Paulo 28(4): 205-12.

De Gaudenzi, J. *et al.* (2005). *RNA-Binding Domain Proteins in Kinetoplastids: a Comparative Analysis*. Eukaryot Cell 4(12): 2106-14.

de Godoy, L. M. *et al.* (2012). *Quantitative proteomics of Trypanosoma cruzi during metacyclogenesis*. Proteomics 12(17): 2694-703.

De Pablos, L. M. *et al.* (2012). *Multigene families in Trypanosoma cruzi and their role in infectivity*. Infect Immun 80(7): 2258-64.

De Santis, P. *et al.* (2013). *Sequence-dependent collective properties of DNAs and their role in biological systems*. Phys Life Rev 10(1): 41-67.

De Souza, W. (2002). *Basic cell biology of Trypanosoma cruzi*. Curr Pharm Des 8(4): 269-85.

de Souza, W. *et al.* (2013). *Active penetration of Trypanosoma cruzi into host cells: historical considerations and current concepts*. Front Immunol 4: 2.

de Souza, W. *et al.* (2009). *Electron microscopy and cytochemistry analysis of the endocytic pathway of pathogenic protozoa*. Prog Histochem Cytochem 44(2): 67-124.

Dennis, G., Jr. *et al.* (2003). *DAVID: Database for Annotation, Visualization, and Integrated Discovery*. Genome Biol 4(5): P3.

Desjeux, P. (2004). *Leishmaniasis: current situation and new perspectives*. Comp Immunol Microbiol Infect Dis 27(5): 305-18.

Di Noia, J. M. *et al.* (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(14): 10218-27.

Diehl, S. *et al.* (2002). *Analysis of stage-specific gene expression in the bloodstream and the procyclic form of Trypanosoma brucei using a genomic DNA-microarray*. Mol Biochem Parasitol 123(2): 115-23.

Docampo, R. *et al.* (2005). *Acidocalcisomes - conserved from bacteria to man*. Nat Rev Microbiol 3(3): 251-61.

Docampo, R. *et al.* (2011). *The role of acidocalcisomes in the stress response of Trypanosoma cruzi*. Adv Parasitol 75: 307-24.

dos Santos, S. L. *et al.* (2012). *The MASP family of Trypanosoma cruzi: changes in gene expression and antigenic profile during the acute phase of experimental infection*. PLoS Negl Trop Dis 6(8): e1779.

Downey, N. *et al.* (1999). *Search for promoters for the GARP and rRNA genes of Trypanosoma congolense*. Mol Biochem Parasitol 104(1): 25-38.

Drechsler, H. *et al.* (2012). *Exotic mitotic mechanisms*. Open Biol 2(12): 120140.

Dubessay, P. *et al.* (2002). *The switch region on Leishmania major chromosome 1 is not required for mitotic stability or gene expression, but appears to be essential*. Nucleic Acids Res 30(17): 3692-7.

Duhagon, M. A. *et al.* (2003). *A novel type of single-stranded nucleic acid binding protein recognizing a highly frequent motif in the intergenic regions of Trypanosoma cruzi*. Biochem Biophys Res Commun 309(1): 183-8.

Duhagon, M. A. *et al.* (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.

Duncan, R. (2004). *DNA microarray analysis of protozoan parasite gene expression: outcomes correlate with mechanisms of regulation*. Trends Parasitol 20(5): 211-5.

Duncan, R. C. *et al.* (2004). *The application of gene expression microarray technology to kinetoplastid research*. Curr Mol Med 4(6): 611-21.

Dvorak, J. A. *et al.* (1973). *Trypanosoma cruzi: interaction with vertebrate cells in vitro. 1. Individual interactions at the cellular and subcellular levels*. Exp Parasitol 34(2): 268-83.

Eckdahl, T. T. *et al.* (1990). *Conserved DNA structures in origins of replication*. Nucleic Acids Res 18(6): 1609-12.

Ekanayake, D. *et al.* (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(11): 1465-72.

Ekanayake, D. K. *et al.* (2011). *Epigenetic regulation of transcription and virulence in Trypanosoma cruzi by O-linked thymine glucosylation of DNA*. Mol Cell Biol 31(8): 1690-700.

el-Sayed, N. M. *et al.* (1995). *cDNA expressed sequence tags of Trypanosoma brucei rhodesiense provide new insights into the biology of the parasite*. Mol Biochem Parasitol 73(1-2): 75-90.

el-Sayed, N. M. *et al.* (1997). *A survey of the Trypanosoma brucei rhodesiense genome using shotgun sequencing*. Mol Biochem Parasitol 84(2): 167-78.

El-Sayed, N. M. *et al.* (2005). *The genome sequence of Trypanosoma cruzi, etiologic agent of Chagas disease*. Science 309(5733): 409-15.

El-Sayed, N. M. *et al.* (2005). *Comparative genomics of trypanosomatid parasitic protozoa*. Science 309(5733): 404-9.

Fenn, K. *et al.* (2007). *The cell biology of Trypanosoma brucei differentiation*. Curr Opin Microbiol 10(6): 539-46.

Fernandes, M. C. *et al.* (2012). *Extracellular amastigotes of Trypanosoma cruzi are potent inducers of phagocytosis in mammalian cells*. Cell Microbiol.

Fiorini, A. *et al.* (2006). *Scaffold/Matrix Attachment Regions and intrinsic DNA curvature*. Biochemistry (Mosc) 71(5): 481-8.

Franzen, O. *et al.* (2011). *The short non-coding transcriptome of the protozoan parasite Trypanosoma cruzi*. PLoS Negl Trop Dis 5(8): e1283.

Franzen, O. *et al.* (2011). *Shotgun sequencing analysis of Trypanosoma cruzi I Sylvio X10/1 and comparison with T. cruzi VI CL Brener*. PLoS Negl Trop Dis 5(3): e984.

Freitas, L. M. *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.

Furger, A. *et al.* (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): 4372-80.

Gabrielian, A. *et al.* (1997). *Distribution of sequence-dependent curvature in genomic DNA sequences*. FEBS Lett 406(1-2): 69-74.

Gabrielian, A. E. *et al.* (1999-2000). *Curved DNA in promoter sequences*. *In Silico* Biol 1(4): 183-96.

Gale, M., Jr. *et al.* (1994). *Translational control mediates the developmental regulation of the Trypanosoma brucei Nrk protein kinase*. J Biol Chem 269(50): 31659-65.

Garcia-Salcedo, J. A. *et al.* (2004). *A differential role for actin during the life cycle of Trypanosoma brucei*. EMBO J 23(4): 780-9.

Garcia-Silva, M. R. *et al.* (2010). *A population of tRNA-derived small RNAs is actively produced in Trypanosoma cruzi and recruited to specific cytoplasmic granules*. Mol Biochem Parasitol 171(2): 64-73.

Garcia, H. G. *et al.* (2007). *Biological consequences of tightly bent DNA: the other life of a macromolecular celebrity*. Biopolymers 85(2): 115-30.

Gaunt, M. W. *et al.* (2003). *Mechanism of genetic exchange in American trypanosomes*. Nature 421(6926): 936-9.

Gemayel, R. *et al.* (2010). *Variable tandem repeats accelerate evolution of coding and regulatory sequences*. Annu Rev Genet 44: 445-77.

Genest, P. A. *et al.* (2007). *Telomeric localization of the modified DNA base J in the genome of the protozoan parasite Leishmania*. Nucleic Acids Res 35(7): 2116-24.

Gilinger, G. *et al.* (2001). *Trypanosome spliced leader RNA genes contain the first identified RNA polymerase II gene promoter in these organisms*. Nucleic Acids Res 29(7): 1556-64.

Gilinger, G. *et al.* (2004). *In vivo transcription analysis utilizing chromatin immunoprecipation reveals a role for trypanosome transcription factor PBP-1 in RNA polymerase III-dependent transcription*. Mol Biochem Parasitol 134(1): 169-73.

Gimenes, F. *et al.* (2008). *Intrinsically bent DNA in replication origins and gene promoters*. Genet Mol Res 7(2): 549-58.

Ginger, M. L. *et al.* (2002). *Ex vivo and in vitro identification of a consensus promoter for VSG genes expressed by metacyclic-stage trypanosomes in the tsetse fly*. Eukaryot Cell 1(6): 1000-9.

Glisovic, T. *et al.* (2008). *RNA-binding proteins and post-transcriptional gene regulation*. FEBS Lett 582(14): 1977-86.

Gommers-Ampt, J. H. *et al.* (1993). *The identification of hydroxymethyluracil in DNA of Trypanosoma brucei*. Nucleic Acids Res 21(9): 2039-43.

Greif, G. *et al.* (2013). *Transcriptome analysis of the bloodstream stage from the parasite Trypanosoma vivax*. BMC Genomics 14(1): 149.

Grove, A. *et al.* (1996). *Localized DNA flexibility contributes to target site selection by DNA-bending proteins*. J Mol Biol 260(2): 120-5.

Guerra-Giraldez, C. *et al.* (2002). *Compartmentation of enzymes in a microbody, the glycosome, is essential in Trypanosoma brucei*. J Cell Sci 115(Pt 13): 2651-8.

Guerra-Slompo, E. P. *et al.* (2012). *Molecular characterization of the Trypanosoma cruzi specific RNA binding protein TcRBP40 and its associated mRNAs*. Biochem Biophys Res Commun 420(2): 302-7.

Gunzl, A. (2010). *The pre-mRNA splicing machinery of trypanosomes: complex or simplified?* Eukaryot Cell 9(8): 1159-70.

Gunzl, A. *et al.* (2003). *RNA polymerase I transcribes procyclin genes and variant surface glycoprotein gene expression sites in Trypanosoma brucei*. Eukaryot Cell 2(3): 542-51.

Gupta, S. K. *et al.* (2013). *Basal splicing factors regulate the stability of mature mRNAs in trypanosomes*. J Biol Chem 288(7): 4991-5006.

Haag, J. *et al.* (1998). *The molecular phylogeny of trypanosomes: evidence for an early divergence of the Salivaria*. Mol Biochem Parasitol 91(1): 37-49.

Haenni, S. *et al.* (2009). *Bidirectional silencing of RNA polymerase I transcription by a strand switch region in Trypanosoma brucei*. Nucleic Acids Res 37(15): 5007-18.

Haile, S. *et al.* (2003). *A role for the exosome in the in vivo degradation of unstable mRNAs*. RNA 9(12): 1491-501.

Haile, S. *et al.* (2007). *Developmental regulation of gene expression in trypanosomatid parasitic protozoa*. Curr Opin Microbiol 10(6): 569-77.

Handman, E. (1999). *Cell biology of Leishmania*. Adv Parasitol 44: 1-39.

Handman, E. (2001). *Leishmaniasis: current status of vaccine development*. Clin Microbiol Rev 14(2): 229-43.

Hartmann, C. *et al.* (2007). *Small trypanosome RNA-binding proteins TbUBP1 and TbUBP2 influence expression of F-box protein mRNAs in bloodstream trypanosomes*. Eukaryot Cell 6(11): 1964-78.

Hartmann, C. *et al.* (2008). *Regulation of a transmembrane protein gene family by the small RNA-binding proteins TbUBP1 and TbUBP2*. Mol Biochem Parasitol 157(1): 112-5.

He, C. Y. *et al.* (2004). *Golgi duplication in Trypanosoma brucei*. J Cell Biol 165(3): 313-21.

Hendriks, E. F. *et al.* (2005). *Disruption of the developmental programme of Trypanosoma brucei by genetic ablation of TbZFP1, a differentiation-enriched CCCH protein*. Mol Microbiol 57(3): 706-16.

Hendriks, E. F. *et al.* (2001). *A novel CCCH protein which modulates differentiation of Trypanosoma brucei to its procyclic form*. EMBO J 20(23): 6700-11.

Heras, S. R. *et al.* (2007). *The L1Tc non-LTR retrotransposon of Trypanosoma cruzi contains an internal RNA-pol II-dependent promoter that strongly activates gene transcription and generates unspliced transcripts*. Nucleic Acids Res 35(7): 2199-214.

Hershberg, R. *et al.* (2008). *Selection on codon bias*. Annu Rev Genet 42: 287-99.

Hertz-Fowler, C. *et al.* (2008). *Telomeric expression sites are highly conserved in Trypanosoma brucei*. PLoS One 3(10): e3527.

Holetz, F. B. *et al.* (2007). *Evidence of P-body-like structures in Trypanosoma cruzi*. Biochem Biophys Res Commun 356(4): 1062-7.

Horn, D. (2008). *Codon usage suggests that translational selection has a major impact on protein expression in trypanosomatids*. BMC Genomics 9: 2.

Hotez, P. J. *et al.* (2007). *Control of neglected tropical diseases*. N Engl J Med 357(10): 1018-27.

Hotz, H. R. *et al.* (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): 3017-26.

Hu, C. H. *et al.* (1994). *xUBF, an RNA polymerase I transcription factor, binds crossover DNA with low sequence specificity*. Mol Cell Biol 14(5): 2871-82.

Huang, D. *et al.* (2009). *Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources*. Nat Protoc 4(1): 44-57.

Hummel, H. S. *et al.* (2000). *Mutational analysis of 3' splice site selection during trans-splicing*. J Biol Chem 275(45): 35522-31.

Ikemura, T. (1981). *Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: a*

proposal for a synonymous codon choice that is optimal for the E. coli translational system. J Mol Biol 151(3): 389-409.
Ingolia, N. T. (2010). *Genome-wide translational profiling by ribosome footprinting*. Methods Enzymol 470: 119-42.
Ingolia, N. T. *et al.* (2012). *The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments*. Nat Protoc 7(8): 1534-50.
Ingolia, N. T. *et al.* (2009). *Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling*. Science 324(5924): 218-23.
Ingolia, N. T. *et al.* (2011). *Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes*. Cell 147(4): 789-802.
Ivens, A. C. *et al.* (2005). *The genome of the kinetoplastid parasite, Leishmania major*. Science 309(5733): 436-42.
Jager, A. V. *et al.* (2007). *mRNA maturation by two-step trans-splicing/polyadenylation processing in trypanosomes*. Proc Natl Acad Sci U S A 104(7): 2035-42.
Jager, A. V. *et al.* (2008). *Stage-specific expression of Trypanosoma cruzi trans-sialidase involves highly conserved 3' untranslated regions*. FEMS Microbiol Lett 283(2): 182-8.
Jenne, A. *et al.* (1999). *Disruption of the streptavidin interaction with biotinylated nucleic acid probes by 2-mercaptoethanol*. Biotechniques 26(2): 249-52, 254.
Jenni, L. *et al.* (1986). *Hybrid formation between African trypanosomes during cyclical transmission*. Nature 322(6075): 173-5.
Kamhawi, S. (2006). *Phlebotomine sand flies and Leishmania parasites: friends or foes?* Trends Parasitol 22(9): 439-45.
Kelly, S. *et al.* (2012). *Genome organization is a major component of gene expression control in response to stress and during the cell division cycle in trypanosomes*. Open Biol 2(4): 120033.
Kelly, S. *et al.* (2005). *An in silico analysis of trypanosomatid RNA polymerases: insights into their unusual transcription*. Biochem Soc Trans 33(Pt 6): 1435-7.
Kolev, N. G. *et al.* (2010). *The transcriptome of the human pathogen Trypanosoma brucei at single-nucleotide resolution*. PLoS Pathog 6(9): e1001090.
Kooter, J. M. *et al.* (1987). *The anatomy and transcription of a telomeric expression site for variant-specific surface antigens in T. brucei*. Cell 51(2): 261-72.
Kramer, S. *et al.* (2011). *Trans-acting proteins regulating mRNA maturation, stability and translation in trypanosomatids*. Trends Parasitol 27(1): 23-30.
Kramer, S. *et al.* (2010). *The RNA helicase DHH1 is central to the correct expression of many developmentally regulated mRNAs in trypanosomes*. J Cell Sci 123(Pt 5): 699-711.
Kramer, S. *et al.* (2008). *Heat shock causes a decrease in polysomes and the appearance of stress granules in trypanosomes independently of eIF2(alpha) phosphorylation at Thr169*. J Cell Sci 121(Pt 18): 3002-14.
Laufer, G. *et al.* (2001). *In-vitro competition analysis of procyclin gene and variant surface glycoprotein gene expression site transcription in Trypanosoma brucei*. Mol Biochem Parasitol 113(1): 55-65.
LeBowitz, J. H. *et al.* (1993). *Coupling of poly(A) site selection and trans-splicing in Leishmania*. Genes Dev 7(6): 996-1007.
Lecordier, L. *et al.* (2007). *Characterization of a TFIIH homologue from Trypanosoma brucei*. Mol Microbiol 64(5): 1164-81.
Lee, B. Y. *et al.* (2013). *Global economic burden of Chagas disease: a computational simulation model*. Lancet Infect Dis.
Lee, M. G. *et al.* (1997). *Transcription of protein-coding genes in trypanosomes by RNA polymerase I*. Annu Rev Microbiol 51: 463-89.
Lesnik, T. *et al.* (2000). *Ribosome traffic in E. coli and regulation of gene expression*. J Theor Biol 202(2): 175-85.

Levick, M. P. *et al.* (1996). *An expressed sequence tag analysis of a full-length, spliced-leader cDNA library from Leishmania major promastigotes*. Mol Biochem Parasitol 76(1-2): 345-8.

Ley, V. *et al.* (1988). *Amastigotes of Trypanosoma cruzi sustain an infective cycle in mammalian cells*. J Exp Med 168(2): 649-59.

Li, C. H. *et al.* (2006). *Roles of a Trypanosoma brucei 5'->3' exoribonuclease homolog in mRNA degradation*. RNA 12(12): 2171-86.

Li, H. *et al.* (2009). *The Sequence Alignment/Map format and SAMtools*. Bioinformatics 25(16): 2078-9.

Li, Y. C. *et al.* (2002). *Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review*. Mol Ecol 11(12): 2453-65.

Li, Y. C. *et al.* (2004). *Microsatellites within genes: structure, function, and evolution*. Mol Biol Evol 21(6): 991-1007.

Li, Z. H. *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): 19058-69.

Liang, X. H. *et al.* (2003). *trans and cis splicing in trypanosomatids: mechanism, factors, and regulation*. Eukaryot Cell 2(5): 830-40.

Linder, J. C. *et al.* (1977). *Plasma membrane specializations in a trypanosomatid flagellate*. J Ultrastruct Res 60(2): 246-62.

Lukes, J. *et al.* (1997). *Analysis of ribosomal RNA genes suggests that trypanosomes are monophyletic*. J Mol Evol 44(5): 521-7.

Luo, H. *et al.* (1997). *Characterization of two protein activities that interact at the promoter of the trypanosomatid spliced leader RNA*. J Biol Chem 272(52): 33344-52.

Lye, L. F. *et al.* (2010). *Retention and loss of RNA interference pathways in trypanosomatid protozoans*. PLoS Pathog 6(10): e1001161.

Madison-Antenucci, S. *et al.* (2002). *Editing machines: the complexities of trypanosome RNA editing*. Cell 108(4): 435-8.

Mahmood, R. *et al.* (1999). *Identification of cis and trans elements involved in the cell cycle regulation of multiple genes in Crithidia fasciculata*. Mol Cell Biol 19(9): 6174-82.

Mahmood, R. *et al.* (2001). *Characterization of the Crithidia fasciculata mRNA cycling sequence binding proteins*. Mol Cell Biol 21(14): 4453-9.

Mair, G. *et al.* (2000). *A new twist in trypanosome RNA metabolism: cis-splicing of pre-mRNA*. RNA 6(2): 163-9.

Mallick, B. *et al.* (2008). *MicroRNA switches in Trypanosoma brucei*. Biochem Biophys Res Commun 372(3): 459-63.

Manger, I. D. *et al.* (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.

Marilley, M. *et al.* (1996). *Common DNA structural features exhibited by eukaryotic ribosomal gene promoters*. Nucleic Acids Res 24(12): 2204-11.

Martinez-Calvillo, S. *et al.* (2004). *Transcription initiation and termination on Leishmania major chromosome 3*. Eukaryot Cell 3(2): 506-17.

Martinez-Calvillo, S. *et al.* (2010). *Gene expression in trypanosomatid parasites*. J Biomed Biotechnol 2010: 525241.

Martinez-Calvillo, S. *et al.* (2003). *Transcription of Leishmania major Friedlin chromosome 1 initiates in both directions within a single region.* Mol Cell 11(5): 1291-1299.

Masocha, W. *et al.* (2007). *Migration of African trypanosomes across the blood-brain barrier*. Physiol Behav 92(1-2): 110-4.

Matthews, K. R. (2005). *The developmental cell biology of Trypanosoma brucei*. J Cell Sci 118(Pt 2): 283-90.

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

Mayho, M. *et al.* (2006). *Post-transcriptional control of nuclear-encoded cytochrome oxidase subunits in Trypanosoma brucei: evidence for genome-wide conservation of life-cycle stage-specific regulatory elements*. Nucleic Acids Res 34(18): 5312-24.

McAndrew, M. *et al.* (1998). *Testing promoter activity in the trypanosome genome: isolation of a metacyclic-type VSG promoter, and unexpected insights into RNA polymerase II transcription*. Exp Parasitol 90(1): 65-76.

McCarthy, J. E. (1998). *Posttranscriptional control of gene expression in yeast*. Microbiol Mol Biol Rev 62(4): 1492-553.

McInerney, J. O. (1998). *GCUA: general codon usage analysis*. Bioinformatics 14(4): 372-3.

Metzker, M. L. (2010). *Sequencing technologies - the next generation*. Nat Rev Genet 11(1): 31-46.

Michel, A. M. *et al.* (2012). *Observation of dually decoded regions of the human genome using ribosome profiling data*. Genome Res 22(11): 2219-29.

Miele, V. *et al.* (2008). *DNA physical properties determine nucleosome occupancy from yeast to fly*. Nucleic Acids Res 36(11): 3746-56.

Milone, J. *et al.* (2004). *Characterization of deadenylation in trypanosome extracts and its inhibition by poly(A)-binding protein Pab1p*. RNA 10(3): 448-57.

Milot, E. *et al.* (1992). *Chromosomal illegitimate recombination in mammalian cells is associated with intrinsically bent DNA elements*. EMBO J 11(13): 5063-70.

Minning, T. A. *et al.* (2003). *Microarray profiling of gene expression during trypomastigote to amastigote transition in Trypanosoma cruzi*. Mol Biochem Parasitol 131(1): 55-64.

Minning, T. A. *et al.* (2009). *The steady-state transcriptome of the four major life-cycle stages of Trypanosoma cruzi*. BMC Genomics 10: 370.

Minning, T. A. *et al.* (2011). *Widespread, focal copy number variations (CNV) and whole chromosome aneuploidies in Trypanosoma cruzi strains revealed by array comparative genomic hybridization*. BMC Genomics 12: 139.

Mittal, N. *et al.* (2009). *Dissecting the expression dynamics of RNA-binding proteins in posttranscriptional regulatory networks*. Proc Natl Acad Sci U S A 106(48): 20300-5.

Monnerat, S. *et al.* (2004). *Genomic organization and gene expression in a chromosomal region of Leishmania major*. Mol Biochem Parasitol 134(2): 233-43.

Morking, P. A. *et al.* (2004). *TcZFP1: a CCCH zinc finger protein of Trypanosoma cruzi that binds poly-C oligoribonucleotides in vitro*. Biochem Biophys Res Commun 319(1): 169-77.

Morking, P. A. *et al.* (2012). *The zinc finger protein TcZFP2 binds target mRNAs enriched during Trypanosoma cruzi metacyclogenesis*. Mem Inst Oswaldo Cruz 107(6): 790-9.

Mortara, R. A. (1991). *Trypanosoma cruzi: amastigotes and trypomastigotes interact with different structures on the surface of HeLa cells*. Exp Parasitol 73(1): 1-14.

Mortazavi, A. *et al.* (2008). *Mapping and quantifying mammalian transcriptomes by RNA-Seq*. Nat Methods 5(7): 621-8.

Murphy, D. (1993). *Nuclear run-on analysis of transcription*. Methods Mol Biol 18: 355-61.

Murphy, N. B. *et al.* (1987). *Trypanosoma brucei repeated element with unusual structural and transcriptional properties*. J Mol Biol 195(4): 855-71.

Mutz, K. O. *et al.* (2013). *Transcriptome analysis using next-generation sequencing*. Curr Opin Biotechnol 24(1): 22-30.

Myler, P. J. *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): 2902-6.

Myler, P. J., Audleman, L., deVos, T., Hixson, G., Lemley, C., Magness, C. Rickel, E., Sisk, E., Sunkin, S., Swartzell, S., Westlake, T., Bastien P., Guoliang, F., Ivens, A., Stuart K. (1999). *Leishmania major Friedlin chromosome 1 has an unusual distribution of protein-coding genes.* Proc. Natl. Acad. Sci. 96(6): 2902-06.

Nair, T. M. (1998). *Evidence for intrinsic DNA bends within the human cdc2 promoter*. FEBS Lett 422(1): 94-8.

Nair, T. M. (2010). *Sequence periodicity in nucleosomal DNA and intrinsic curvature*. BMC Struct Biol 10 Suppl 1: S8.

Nardelli, S. C. *et al.* (2007). *Small-subunit rRNA processome proteins are translationally regulated during differentiation of Trypanosoma cruzi*. Eukaryot Cell 6(2): 337-45.

Nguyen, T. N. *et al.* (2007). *Active RNA polymerase I of Trypanosoma brucei harbors a novel subunit essential for transcription*. Mol Cell Biol 27(17): 6254-63.

Nilsson, D. *et al.* (2010). *Spliced leader trapping reveals widespread alternative splicing patterns in the highly dynamic transcriptome of Trypanosoma brucei*. PLoS Pathog 6(8): e1001037.

O'Connell, D. (2007). *Neglected Diseases*. Nature 449: 157.

Odiit, M. *et al.* (1997). *Duration of symptoms and case fatality of sleeping sickness caused by Trypanosoma brucei rhodesiense in Tororo, Uganda*. East Afr Med J 74(12): 792-5.

Ogbadoyi, E. O. *et al.* (2003). *A high-order trans-membrane structural linkage is responsible for mitochondrial genome positioning and segregation by flagellar basal bodies in trypanosomes*. Mol Biol Cell 14(5): 1769-79.

Ohtsu, M. *et al.* (2008). *Novel DNA microarray system for analysis of nascent mRNAs*. DNA Res 15(4): 241-51.

Olivier, M. *et al.* (2012). *Leishmania virulence factors: focus on the metalloprotease GP63*. Microbes Infect 14(15): 1377-89.

Ong, S. E. (2012). *The expanding field of SILAC*. Anal Bioanal Chem 404(4): 967-76.

Opperdoes, F. R. *et al.* (1977). *Localization of nine glycolytic enzymes in a microbody-like organelle in Trypanosoma brucei: the glycosome*. FEBS Lett 80(2): 360-4.

Orekhova, A. S. *et al.* (2013). *Bidirectional promoters in the transcription of Mammalian genomes*. Biochemistry (Mosc) 78(4): 335-41.

Padilla-Mejia, N. E. *et al.* (2009). *Gene organization and sequence analyses of transfer RNA genes in Trypanosomatid parasites*. BMC Genomics 10: 232.

Palenchar, J. B. *et al.* (2006). *Gene transcription in trypanosomes*. Mol Biochem Parasitol 146(2): 135-41.

Palenchar, J. B. *et al.* (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.

Parodi-Talice, A. *et al.* (2004). *Proteome analysis of the causative agent of Chagas disease: Trypanosoma cruzi*. Int J Parasitol 34(8): 881-6.

Parsons, M. *et al.* (1984). *Trypanosome mRNAs share a common 5' spliced leader sequence*. Cell 38(1): 309-16.

Parsons, M. *et al.* (1991). *Trypanosoma brucei: analysis of codon usage and nucleotide composition of nuclear genes*. Exp Parasitol 73(1): 101-5.

Pastro, L. *et al.* (2013). *Implication of CA repeated tracts on post-transcriptional regulation in Trypanosoma cruzi*. Exp Parasitol.

Paterou, A. *et al.* (2006). *Identification and stage-specific association with the translational apparatus of TbZFP3, a CCCH protein that promotes trypanosome life-cycle development*. J Biol Chem 281(51): 39002-13.

Patrone, G. *et al.* (2000). *Nuclear run-on assay using biotin labeling, magnetic bead capture and analysis by fluorescence-based RT-PCR*. Biotechniques 29(5): 1012-4, 1016-7.

Paule, M. R. *et al.* (2000). *Survey and summary: transcription by RNA polymerases I and III*. Nucleic Acids Res 28(6): 1283-98.

Pays, E. *et al.* (2001). *The VSG expression sites of Trypanosoma brucei: multipurpose tools for the adaptation of the parasite to mammalian hosts*. Mol Biochem Parasitol 114(1): 1-16.

Pays, E. *et al.* (2004). *Antigenic variation in Trypanosoma brucei: facts, challenges and mysteries*. Curr Opin Microbiol 7(4): 369-74.

Peacock, C. S. *et al.* (2007). *Comparative genomic analysis of three Leishmania species that cause diverse human disease*. Nat Genet 39(7): 839-47.

Peacock, L. *et al.* (2009). *Intraclonal mating occurs during tsetse transmission of Trypanosoma brucei*. Parasit Vectors 2(1): 43.

Peacock, L. *et al.* (2011). *Identification of the meiotic life cycle stage of Trypanosoma brucei in the tsetse fly*. Proc Natl Acad Sci U S A 108(9): 3671-6.

Perez-Diaz, L. *et al.* (2012). *The overexpression of the trypanosomatid-exclusive TcRBP19 RNA-binding protein affects cellular infection by Trypanosoma cruzi*. Mem Inst Oswaldo Cruz 107(8): 1076-9.

Perez-Diaz, L. *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.

Perez-Diaz, L. *et al.* (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.

Pisano, S. *et al.* (2006). *AFM imaging and theoretical modeling studies of sequence-dependent nucleosome positioning*. Biophys Chem 124(2): 81-9.

Plohl, M. *et al.* (2008). *Satellite DNAs between selfishness and functionality: structure, genomics and evolution of tandem repeats in centromeric (hetero)chromatin*. Gene 409(1-2): 72-82.

Portal, D. *et al.* (2003). *An early ancestor in the evolution of splicing: a Trypanosoma cruzi serine-arginine-rich protein (TcSR) is functional in cis-splicing*. Mol Biochem Parasitol 127(1): 37-46.

Portal, D. *et al.* (2003). *Trypanosoma cruzi TcSRPK, the first protozoan member of the SRPK family, is biochemically and functionally conserved with metazoan SR protein-specific kinases*. Mol Biochem Parasitol 127(1): 9-21.

Porto-Carreiro, I. *et al.* (2000). *Trypanosoma cruzi epimastigote endocytic pathway: cargo enters the cytostome and passes through an early endosomal network before storage in reservosomes*. Eur J Cell Biol 79(11): 858-69.

Potaman, V. N. *et al.* (2005). Alternative Conformations and Biology. DNA conformation and transcription. T. Ohyama. Georgetown, Tex. New York, NY. **Landes Bioscience:** 211.

Quijada, L. *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): 4414-24.

Quinlan, A. R. *et al.* (2010). *BEDTools: a flexible suite of utilities for comparing genomic features*. Bioinformatics 26(6): 841-2.

Ramirez, J. D. *et al.* (2012). *Contemporary cryptic sexuality in Trypanosoma cruzi*. Mol Ecol 21(17): 4216-26.

Rastrojo, A. *et al.* (2013). *The transcriptome of Leishmania major in the axenic promastigote stage: transcript annotation and relative expression levels by RNA-seq*. BMC Genomics 14(1): 223.

Remme, J. H. F. *et al.* (2006). *Tropical Diseases Targeted for Elimination: Chagas Disease, Lymphatic Filariasis, Onchocerciasis, and Leprosy*.

Respuela, P. *et al.* (2008). *Histone acetylation and methylation at sites initiating divergent polycistronic transcription in Trypanosoma cruzi*. J Biol Chem 283(23): 15884-92.

Robles, A. *et al.* (2008). *Regulation of an amino acid transporter mRNA in Trypanosoma brucei*. Mol Biochem Parasitol 157(1): 102-6.

Rochette, A. *et al.* (2005). *Characterization and developmental gene regulation of a large gene family encoding amastin surface proteins in Leishmania spp*. Mol Biochem Parasitol 140(2): 205-20.

Roditi, I. *et al.* (1999). *An unambiguous nomenclature for the major surface glycoproteins of the procyclic form of Trypanosoma brucei*. Mol Biochem Parasitol 103(1): 99-100.

Rodrigues, D. C. *et al.* (2010). *Trypanosoma cruzi: modulation of HSP70 mRNA stability by untranslated regions during heat shock*. Exp Parasitol 126(2): 245-53.

Roellig, D. M. *et al.* (2013). *Genetic Variation and Exchange in Trypanosoma cruzi Isolates from the United States*. PLoS One 8(2): e56198.

Roux-Rouquie, M. *et al.* (2000). *Modeling of DNA local parameters predicts encrypted architectural motifs in Xenopus laevis ribosomal gene promoter*. Nucleic Acids Res 28(18): 3433-41.

Ruan, J. P. *et al.* (2004). *Functional characterization of a Trypanosoma brucei TATA-binding protein-related factor points to a universal regulator of transcription in trypanosomes*. Mol Cell Biol 24(21): 9610-8.

Rudenko, G. (2010). *Epigenetics and transcriptional control in African trypanosomes*. Essays Biochem 48(1): 201-19.

Rudenko, G. *et al.* (1989). *Alpha-amanitin resistant transcription of protein coding genes in insect and bloodstream form Trypanosoma brucei*. EMBO J 8(13): 4259-63.

Rudenko, G. *et al.* (1990). *Procyclic acidic repetitive protein (PARP) genes located in an unusually small alpha-amanitin-resistant transcription unit: PARP promoter activity assayed by transient DNA transfection of Trypanosoma brucei*. Mol Cell Biol 10(7): 3492-504.

Ruvalcaba-Trejo, L. I. *et al.* (2011). *The Trypanosoma cruzi Sylvio X10 strain maxicircle sequence: the third musketeer*. BMC Genomics 12: 58.

Saas, J. *et al.* (2000). *A developmentally regulated aconitase related to iron-regulatory protein-1 is localized in the cytoplasm and in the mitochondrion of Trypanosoma brucei*. J Biol Chem 275(4): 2745-55.

Saxena, A. *et al.* (2003). *Evaluation of differential gene expression in Leishmania major Friedlin procyclics and metacyclics using DNA microarray analysis*. Mol Biochem Parasitol 129(1): 103-14.

Schenkman, S. *et al.* (2011). *Nuclear structure of Trypanosoma cruzi*. Adv Parasitol 75: 251-83.

Schimanski, B. *et al.* (2005). *Characterization of a multisubunit transcription factor complex essential for spliced-leader RNA gene transcription in Trypanosoma brucei*. Mol Cell Biol 25(16): 7303-13.

Schwede, A. *et al.* (2008). *A role for Caf1 in mRNA deadenylation and decay in trypanosomes and human cells*. Nucleic Acids Res 36(10): 3374-88.

Shapiro, T. A. *et al.* (1995). *The structure and replication of kinetoplast DNA*. Annu Rev Microbiol 49: 117-43.

Sharma, R. *et al.* (2009). *The heart of darkness: growth and form of Trypanosoma brucei in the tsetse fly*. Trends Parasitol 25(11): 517-24.

Sharp, P. M. *et al.* (1987). *The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications*. Nucleic Acids Res 15(3): 1281-95.

Shaw, J. M. *et al.* (1988). *Editing of kinetoplastid mitochondrial mRNAs by uridine addition and deletion generates conserved amino acid sequences and AUG initiation codons*. Cell 53(3): 401-11.

Sherwin, T. *et al.* (1989). *The cell division cycle of Trypanosoma brucei brucei: timing of event markers and cytoskeletal modulations*. Philos Trans R Soc Lond B Biol Sci 323(1218): 573-88.

Siegel, T. N. *et al.* (2011). *Gene expression in Trypanosoma brucei: lessons from high-throughput RNA sequencing*. Trends Parasitol 27(10): 434-41.

Siegel, T. N. *et al.* (2009). *Four histone variants mark the boundaries of polycistronic transcription units in Trypanosoma brucei*. Genes Dev 23(9): 1063-76.

Siegel, T. N. *et al.* (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.

Siegel, T. N. *et al.* (2005). *Systematic study of sequence motifs for RNA trans splicing in Trypanosoma brucei*. Mol Cell Biol 25(21): 9586-94.

Simpson, L. *et al.* (2003). *Uridine insertion/deletion RNA editing in trypanosome mitochondria: a complex business*. Rna 9(3): 265-76.

Simpson, L. *et al.* (2000). *Evolution of RNA editing in trypanosome mitochondria*. Proc Natl Acad Sci U S A 97(13): 6986-93.

Smith, D. F. *et al.* (1996). Molecular biology of parasitic protozoa. Oxford ; New York, IRL Press at Oxford University Press.

Souto-Padron, T. *et al.* (1984). *Quick-freeze, deep-etch rotary replication of Trypanosoma cruzi and Herpetomonas megaseliae*. J Cell Sci 69: 167-78.

Souza, W. (2009). *Structural organization of Trypanosoma cruzi*. Mem Inst Oswaldo Cruz 104 Suppl 1: 89-100.

Stockdale, C. *et al.* (2008). *Antigenic variation in Trypanosoma brucei: joining the DOTs*. PLoS Biol 6(7): e185.

Supek, F. *et al.* (2004). *INCA: synonymous codon usage analysis and clustering by means of self-organizing map*. Bioinformatics 20(14): 2329-30.

Tan, T. H. *et al.* (2002). *tRNAs in Trypanosoma brucei: genomic organization, expression, and mitochondrial import*. Mol Cell Biol 22(11): 3707-17.

Teixeira, S. M. (1998). *Control of gene expression in Trypanosomatidae*. Braz J Med Biol Res 31(12): 1503-16.

Teixeira, S. M. *et al.* (2012). *Trypanosomatid comparative genomics: Contributions to the study of parasite biology and different parasitic diseases*. Genet Mol Biol 35(1): 1-17.

Thomas, S. *et al.* (2009). *Histone acetylations mark origins of polycistronic transcription in Leishmania major*. BMC Genomics 10: 152.

Thorvaldsdottir, H. *et al.* (2013). *Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration*. Brief Bioinform 14(2): 178-92.

Tiengwe, C. *et al.* (2012). *Genome-wide analysis reveals extensive functional interaction between DNA replication initiation and transcription in the genome of Trypanosoma brucei*. Cell Rep 2(1): 185-97.

Timchenko, T. V. *et al.* (2002). *COMPUTATIONAL PREDICTION AND EXPERIMENTAL ANALYSIS OF THE CURVED DNAs AS THE HOT SPOTS OF RECOMBINATION*. BGRS: 40-42.

Travers, A. A. *et al.* (2012). *DNA structure, nucleosome placement and chromatin remodelling: a perspective*. Biochem Soc Trans 40(2): 335-40.

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

Tyler, K. M. *et al.* (2001). The life cycle of *Trypanosoma cruzi* revisited.

Ulrich, P. N. *et al.* (2011). *Identification of contractile vacuole proteins in Trypanosoma cruzi*. PLoS One 6(3): e18013.

Ullu, E. *et al.* (1993). *Temporal order of RNA-processing reactions in trypanosomes: rapid trans splicing precedes polyadenylation of newly synthesized tubulin transcripts*. Mol Cell Biol 13(1): 720-5.

Urbaniak, M. D. *et al.* (2012). *Comparative SILAC proteomic analysis of Trypanosoma brucei bloodstream and procyclic lifecycle stages*. PLoS One 7(5): e36619.

van Leeuwen, F. *et al.* (2000). *Tandemly repeated DNA is a target for the partial replacement of thymine by beta-D-glucosyl-hydroxymethyluracil in Trypanosoma brucei*. Mol Biochem Parasitol 109(2): 133-45.
van Luenen, H. G. *et al.* (2012). *Glucosylated hydroxymethyluracil, DNA base J, prevents transcriptional readthrough in Leishmania*. Cell 150(5): 909-21.
Vanhamme, L. *et al.* (1995). *Control of gene expression in trypanosomes*. Microbiol Rev 59(2): 223-40.
Vassella, E. *et al.* (2000). *A major surface glycoprotein of trypanosoma brucei is expressed transiently during development and can be regulated post-transcriptionally by glycerol or hypoxia*. Genes Dev 14(5): 615-26.
Vassella, E. *et al.* (1997). *Differentiation of African trypanosomes is controlled by a density sensing mechanism which signals cell cycle arrest via the cAMP pathway*. J Cell Sci 110 (Pt 21): 2661-71.
Vazquez, M. *et al.* (2003). *Unique features of the Trypanosoma cruzi U2AF35 splicing factor*. Mol Biochem Parasitol 128(1): 77-81.
Vickerman, K. *et al.* (1970). *Spindle microtubules in the dividing nuclei of trypanosomes*. J Cell Sci 6(2): 365-83.
Vlahovicek, K. *et al.* (2003). *DNA analysis servers: plot.it, bend.it, model.it and IS*. Nucleic Acids Res 31(13): 3686-7.
Vlasova-St Louis, I. *et al.* (2011). *Coordinate regulation of mRNA decay networks by GU-rich elements and CELF1*. Curr Opin Genet Dev 21(4): 444-51.
von Roretz, C. *et al.* (2011). *Turnover of AU-rich-containing mRNAs during stress: a matter of survival*. Wiley Interdiscip Rev RNA 2(3): 336-47.
Walgraffe, D. *et al.* (2005). *Characterization of subunits of the RNA polymerase I complex in Trypanosoma brucei*. Mol Biochem Parasitol 139(2): 249-60.
Webb, H. *et al.* (2005). *Developmentally regulated instability of the GPI-PLC mRNA is dependent on a short-lived protein factor*. Nucleic Acids Res 33(5): 1503-12.
Weirather, J. L. *et al.* (2012). *Mapping of VSG similarities in Trypanosoma brucei*. Mol Biochem Parasitol 181(2): 141-52.
Wen, L. M. *et al.* (2000). *PPB1, a putative spliced leader RNA gene transcription factor in Trypanosoma cruzi*. Mol Biochem Parasitol 110(2): 207-21.
Wen, Y. Z. *et al.* (2011). *Pseudogene-derived small interference RNAs regulate gene expression in African Trypanosoma brucei*. Proc Natl Acad Sci U S A 108(20): 8345-50.
Westenberger, S. J. *et al.* (2006). *Trypanosoma cruzi mitochondrial maxicircles display species- and strain-specific variation and a conserved element in the non-coding region*. BMC Genomics 7: 60.
WHO (1998). The World Health Organization.
WHO (2002). *Control of Chagas disease*. World Health Organ Tech Rep Ser 905: i-vi, 1-109, back cover.
Wincker, P. *et al.* (1996). *The Leishmania genome comprises 36 chromosomes conserved across widely divergent human pathogenic species*. Nucleic Acids Res 24(9): 1688-94.
Worthey, E. A. *et al.* (2003). *Leishmania major chromosome 3 contains two long convergent polycistronic gene clusters separated by a tRNA gene*. Nucleic Acids Res 31(14): 4201-10.
Wright, J. R. *et al.* (2010). *Histone H3 trimethylated at lysine 4 is enriched at probable transcription start sites in Trypanosoma brucei*. Mol Biochem Parasitol 172(2): 141-4.
Wu, Y. *et al.* (2000). *A new developmentally regulated gene family in Leishmania amastigotes encoding a homolog of amastin surface proteins*. Mol Biochem Parasitol 110(2): 345-57.
Xu, P. *et al.* (2001). *Identification of a spliced leader RNA binding protein from Trypanosoma cruzi*. Mol Biochem Parasitol 112(1): 39-49.

Yao, C. *et al*. (2003). *The major surface protease (MSP or GP63) of Leishmania sp. Biosynthesis, regulation of expression, and function*. Mol Biochem Parasitol 132(1): 1-16.

Yoshida, N. (2008). *Trypanosoma cruzi infection by oral route: how the interplay between parasite and host components modulates infectivity*. Parasitol Int 57(2): 105-9.

Yoshida, N. *et al.* (2008). *Trypanosoma cruzi: parasite and host cell signaling during the invasion process*. Subcell Biochem 47: 82-91.

Zingales, B. *et al.* (2009). *A new consensus for Trypanosoma cruzi intraspecific nomenclature: second revision meeting recommends TcI to TcVI*. Mem Inst Oswaldo Cruz 104(7): 1051-4.

Agradecimientos

A Verónica por los años y buenos momentos que hemos compartido y especialmente por el apoyo de los últimos meses.

Al resto de mi familia que siempre me ha incentivado a hacer lo que me gusta.

A Beatriz por abrir las puertas del laboratorio en donde he pasado tantas buenas horas. Por lo que me ha enseñado en los muchos años que fui su estudiante, por la guía y aliento en los proyectos más diversos que se han presentado durante este tiempo.

A mis amigos del LIM por las largas discusiones que tanto han enriquecido este trabajo. Gracias a todos por tirar siempre juntos para el mismo lado.

To Dr. Najib El-Sayed for receiving me at his lab as part of the team, and for the insights he provided in the intrinsic curvature work.

A los compañeros de genética por ayudarme durante estos años para cumplir con mis obligaciones compartidas.

A los miembros del tribunal que accedieron a leer y comentar este trabajo.

Este trabajo esta dedicado a Emma.