Genómica comparativa y biología molecular de tripanosomas analizando datos de RNAseq

Matías Rodríguez

Orientador: Fernando Alvarez

Indice

Genómica comparativa y biología molecular de tripanosomas analizando datos de RNAseq

Generalidades

La tripanosomiasis es un conjunto de enfermedades provocadas por diferentes especies del género Trypanosoma con un área de incidencia global, que se da principalmente en regiones tropicales y subtropicales. Los tripanosomas son protozoarios parásitos obligatorios de importancia médica y veterinaria capaces de infectar humanos y animales tanto domésticos como salvajes.

Los miembros de los géneros Trypanosoma, Leishmania, Crithidia y Phytomonas entre otros conforman la familia Trypanosomatidae, perteneciente a la clase Kinetoplastida. Se caracterizan por poseer un único flagelo similar a una membrana ondulante que utilizan para desplazarse y adherirse. Otra de sus peculiaridades morfológicas es que poseen un kinetoplasto, que se trata de un organelo que contiene el DNA mitocondrial y es una extensión de su única mitocondria; además su citoesqueleto está compuesto principalmente por microtúbulos en arreglos muy regulares.

Las diferentes especies de tripanosomas que infectan mamíferos y tienen importancia en la salud humana se agrupan en dos categorías; los Stercoraria que se desarrollan en el tubo digestivo de un insecto vector y son transmitidos a través de las heces y los Salivaria que se desarrollan en la parte anterior del tracto digestivo (algunos dentro de las glándulas salivales) del insecto vector e infecta cuando este se alimenta.

Los tripanosomas del grupo Salivaria son de origen africano y el principal vector transmisor son varias especies de dípteros del género Glossina constituido por las moscas tse-tse. Los principales animales afectados son bovinos por *T. congolense* (enfermedad llamada nagana), ovinos por *T. vivax*, cabras por *T. brucei brucei*, suinos por *T. simiae*; también pueden verse afectados equinos y camélidos. *T. vivax* puede encontrarse en áreas sin el vector tse-tse, en Africa subsahariana. Los tripanosomas que causan enfermedades en humanos son *T. brucei rhodiense* y *T. brucei gambiense* muy similares al *T. brucei brucei* que infecta animales. (Kahn et al. 2005)

La mayor parte de las transmisiones por tse-tse es cíclica, es decir el parásito realiza parte de su ciclo en el vector. Esta comienza cuando la sangre de un animal infectado es ingerida por la mosca; el tripanosoma experimenta una serie de cambios importantes particularmente en su membrana celular la que pierde su cobertura de glicoproteínas variables de superficie (VSG) siendo sustituidas por otras (prociclinas). Luego se multiplica por fisión binaria, es la llamada forma procíclica. Posteriormente el parásito migra hacia la parte anterior del tubo digestivo, siendo la ubicación específica variable dependiendo de la especie de Trypanosoma. En *T. brucei spp* migra del intestino a las glándulas salivales; en *T. congolense* a la hipofaringe, en *T. vivax* todo el ciclo ocurre en la probóscide. (Kahn et al. 2005) Una vez en la parte anterior readquiere su cobertura de VSG y se vuelve infectivo bajo la forma de tripomastigota metacíclico.

En la tripanosomiasis africana los parásitos inoculados en la piel crecen y causan hinchazón localizada. Luego migran a los nódulos linfáticos, y luego a la sangre donde se multiplican. *T. brucei* y *T. vivax* invaden tejidos y puede causar daños en varios órganos. El ciclo de los tripanosomas africanos (Figura 1) comprende solamente etapas extracelulares. En humanos la enfermedad provocada por los estos tripanosomas es conocida como “enfermedad del sueño” y los síntomas son dolor de cabeza, fiebre, fatiga, hinchazón de nódulos linfáticos, dolor en músculos y articulaciones. Puede provocar problemas neurológicos luego de la invasión del sistema nervioso central y si no es tratada puede ser mortal. En el centro y oeste de África *T.brucei gambiense* provoca una forma crónica de enfermedad del sueño. En el este y sur de África *T. brucei rhodesiense* causa una forma aguda. (Barret et al. 2003)

La transmisión mecánica (en la cual el vector solo actúa como una “jeringa infectada”, es decir sin haber completado el ciclo) puede ocurrir también a través de tse-tse u otros vectores.

En América la transmisión es solamente mecánica; en el caso de *T. vivax* el principal vector en América del Sur y Central son especies de *Tabanus spp*.

El agente causante de la tripanosomiasis americana, la llamada enfermedad de Chagas, es *T. cruzi* (Figura 1) cuyo vector son insectos hemípteros de la sub-familia Triatominae; del cual todos sus miembros son hematófagos y vectores potenciales. La mayoría están distribuidos a lo largo de América donde suelen ser conocidos bajo el nombre de vinchuca.

Cuando un triatomino infectado se alimenta de la sangre de un mamífero, libera los parásitos en sus heces, cerca del sitio de la herida. Lo parásitos entran a través de la herida o a través de una membrana mucosa como la conjuntiva bajo su forma infectiva de tripomastigotas. Otra forma de infección,

probablemente la ancestral es por vía de las mucosas gástricas, la cual ocurre generalmente en mamíferos insectívoros o en animales que ingieren frutas contaminadas con *T. cruzi*.
Dentro del hospedero, los tripanosomas americanos invaden las células cercanas al sitio de inoculación donde se diferencian en amastigotas intracelulares que se multiplican por fisión binaria y luego se diferencian a tripomastigotas liberándose al torrente sanguíneo. Infectan células en una variedad de tejidos y vuelven a repetir el ciclo; los tripomastigotas sanguíneos no se replican, a diferencia de lo que ocurre con lo tripanosomas africanos. El triatomino se infecta al ingerir sangre infectada, y luego en el intestino medio del vector el parásito se transforma en epimastigota donde se multiplica y diferencia a su forma infectiva. (Barret et al. 2003)
La infección por *T. cruzi* se presenta como una lesión nodular e hinchazón en el sitio de infección y luego puede ser asintomática o con algunas manifestaciones como fiebre, anorexia, linfoadenopatía y miocarditis. La forma crónica sintomática puede no ocurrir hasta años o décadas más tarde y manifestarse como problemas en el tracto digestivo y pérdida de peso o como una miocardiopatía que puede llegar a ser mortal. (Kahn et al. 2005)

Evasión de la respuesta inmune
Todo patógeno debe evadir la respuesta inmune de su hospedero para establecer la infección.
Los tripanosomas africanos son exclusivamente extracelulares por lo cual están siempre expuestos al sistema inmune, pero una densa y altamente inmunogénica capa de glicoproteínas los protege contra la lisis mediada por complemento. Una vez que han madurado anticuerpos específicos las inmunoglobulinas lisan los tripanosomas que tengan la misma cubierta de glicoproteínas. Sin embargo un pequeño porcentaje en cada nueva generación de parásitos cambia a una glicoproteína antigénicamente diferente. En *T. brucei* hay más de mil genes diferentes que codifican para estas glicoproteínas variantes de superficie (VSG) y la expresión secuencial de estos genes produce poblaciones de parásitos antigénicamente diferentes lo que le permite la supervivencia en el huésped mamífero. (Barret et al. 2003)
T. cruzi adopta una vida intracelular que lo protege de la inmunidad humoral. La superficie de *T. cruzi* está cubierta por glicoproteínas de tipo mucina para la cual codifican centenares de genes y expone una región amino-terminal hipervariable que le brinda variabilidad antigénica. La respuesta inmune genera anticuerpos contra las glicoproteínas que cubren su superficie y los eliminan; sin embargo los tripanosomas tienen múltiples genes que codifican para diferentes glicoproteínas que no son vulnerables a la respuesta inmune; esta variación antigénica resulta en la persistencia del parásito.
El elevado número de tipos de glicoproteínas antigénicas es lo que dificulta el desarrollo de una vacuna y permite las reinfecciones.

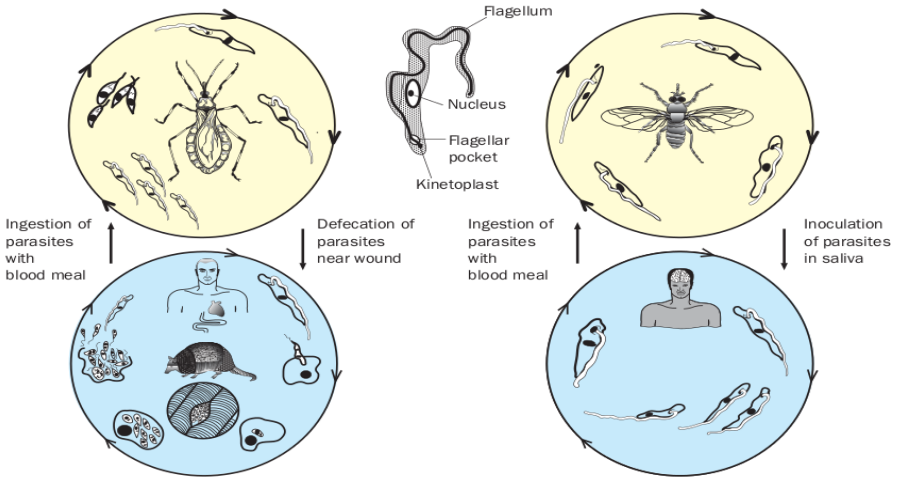


Figura 1: Ciclo de vida de *T. cruzi* (a la izq.) y *T.brucei* (a la der.) y aspectos particulares de su morfología.(Barret et al. 2003)

Biología molecular

Genomas de tripanosomas

Existe una marcada diferencia en el tamaño y densidad de los genes entre las diferentes especies de tripanosomas. Aunque la mayoría de los genes de los tripanosomas en el mismo contexto genómico se encuentran conservados, hay importantes diferencias que seguramente reflejan adaptaciones específicas a una presión selectiva especie-específica, a diferentes patofisiologías y estrategias de supervivencia de cada organismo (ver Tabla 1).

T. brucei tiene solo 11 cromosomas diploides junto a numerosos cromosomas pequeños que contienen extensas regiones de secuencias repetitivas; *T. cruzi* tiene 28 pares de cromosomas, mientras que el cariotipo de *T.vivax* es incierto.

T. brucei, *T. cruzi* y *L. major* son denominados como el grupo Tritryp y la comparación de su contenido génico muestra la presencia de un núcleo conservado del proteoma de unos 6200 genes formando grandes clústers. (El-Sayed et al. 2005)

	T. brucei	T. vivax	T.cruzi
Tamaño del genoma haploide (Mbp)	25	48	55
Nº de cromosomas (por genoma haploide)	11*	**	28**
Nº de K genes (por genoma haploide)	9	11	12
*Excluyendo unos 100 minicromosomas que suman 10Mbp **No se conoce con exactitud			

Tabla 1: Tabla comparativa de los genomas de las tres especies consideradas en este trabajo.

La característica fundamental de los genomas de tripanosomas es la organización de sus genes en grupos que son transcriptos juntos como un solo pre-mRNA que luego es procesado.

Este pre-mRNA lleva la información génica para la síntesis de varias cadenas polipeptídicas por lo que se le denomina policistrón, en cierta forma similar a los operones bacterianos; aunque a diferencia de estos son de mucho mayor tamaño y los genes no suelen estar funcionalmente relacionados entre si.

La mayoría de los genes conservados entre las especies de tripanosomas también conservan el orden y la orientación a lo largo de los cromosomas; a esto se le denomina conservación de la sintenia y es un indicador de la proximidad filogenética. Muchos genes especie-específico especialmente las grandes familias de antígenos de superficie se encuentran en zonas donde no hay conservación de la sintenia como las regiones internas de los cromosomas y las regiones subteloméricas. Las secuencias subteloméricas quedan definidas como aquellas que se encuentra entre el telómero y el primer gen constitutivo.

La mayoría de los rearreglos de bloques donde hay conservación de la sintenia representan inversiones y translocaciones, pero parece haber varios casos de fusión cromosómica en *T. brucei*.

La variación y diversidad antigénica son características de *T. brucei* y *T. cruzi*, y la localización de grupos de genes que codifican para proteínas de superficie en regiones subteloméricas junto a la presencia de numerosos retroelementos dentro de estas regiones puede hacer aumentar la frecuencia de recombinación y proveer de un mecanismo para generar variabilidad de secuencia rápidamente.

La ubicación de genes como MASPs (mucin associated surface proteins) y DGF-1 (Dispersed Gene Family) de *T. cruzi*, VSG (variant suface glycoprotein) en *T. brucei* y RHS (Retroposon Hot Spot) en ambos en estas regiones hace suponer que se encuentran involucrados en la evasión a la respuesta inmune y a la supervivencia en diferentes huéspedes.

La recombinación frecuente en estas regiones resulta en grandes polimorfismos de tamaño, de hasta 2Mb entre cromosomas homólogos de *T. brucei* y *T.cruzi*.

Los retroelementos, RNAs estructurales y la expansión de familias génicas con frecuencia se encuentran asociados a discontinuidades en la conservación de la sintenia que junto con rearreglos, pérdidas, adquisiciones y divergencias en los genes dan características únicas al genoma de cada parásito. También hay pérdida de la conservación de la sintenia en las cercanías a regiones de cambio de hebra (strand-switch regions lo cual puede que sea un reflejo de elevadas tasas de recombinación en estos sitios. Parece existir una fuerte presión selectiva para mantener el orden génico y los clústers de genes intactos a pesar de la extensa divergencia de secuencia entre los propios genes. (El-Sayed et al. 2005)

En *T.brucei* más de 20% del genoma codifica para genes en regiones subteloméricas la mayoría de los cuales son especie-específico y están relacionados con la capacidad del parásito de experimentar variación antigénica en las vías sanguíneas de su huésped mamífero.
Las regiones subteloméricas de *T.cruzi* son extensas y consisten mayormente en arreglos de genes intercalados de la superfamilia trans-sialidasa, DGF-1 y RHS junto a vestigios interrumpidos de retroelementos de la familia VIPER, elementos repetitivos SIRE, retrotransposones no-LTR L1Tc y los no autónomos NARTc, y retroelementos DIRE. Otra característica de *T. cruzi* es la presencia de grandes islas de genes que codifican para proteínas de superficie como trans-sialidasa, mucina, proteínas de superficie asociadas a mucinas (MASP), retrotransposones y genes RHS. (El-Sayed et al. 2005)

Transcripción en tripanosomas
Una característica de todos los genes codificantes en tripanosomas es su disposición cromosómica.
Los genes se encuentran agrupados en varios clústers del siguiente modo, un grupo de genes se encuentran en una hebra, luego le continua una región de 1 a 13 Kb no transcripta llamada región de cambio de hebra o 'strand switch' (STS), y luego de esta región comienza un segundo conjunto de genes codificado en la hebra complementaria, y así sucesivamente (ver Figura 2). Por lo tanto los grupos de genes policistrónicos radian de forma bidireccional desde una región STS. (Palenchar et al. 2006)

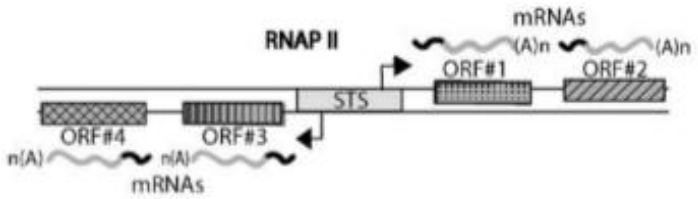


Figura 2: Ejemplo de una de las regiones strand switch (STS) desde cuyos extremos radian dos grupos de genes. (Liang et al 2003.)

De forma similar a otros mRNAs de eucariotas, los mRNAs de tripanosomas poseen modificaciones en los extremos 5' y 3'. Integral para la resolución de mRNAs individuales a partir de los ORFs policistrónicos co-transcriptos hay un conjunto reacciones de poliadenilación y capping. Para resolver los pre-mRNAs policistrónicos en unidades traducibles estables los tripanosomas tienen siempre pronto prefabricado un cap altamente modificado derivado de transcriptos del gen SL-RNA presente en un alto número de copias. Un abundante suministro de SL-RNAs pre-capeados prontos para usar le permite al parásito sintetizar mensajeros maduros estables.
Los genes de SL-RNA están presentes en el genoma de *T. brucei* en aproximadamente unas 200 copias, formando clústers en tándem.
En los tripanosomas, los pre-mRNA contienen varios cistrones en tándem, separados por un espacio intergénico donde el evento de poliadenilacion en 3' del primer gen tiene lugar luego de que el segundo gen de ese pre-mRNA recibe un cap conteniendo el SL-RNA en el extremo 5'. Este singular conjunto de eventos de procesado del pre-mRNA permite a una única ronda de transcripción de la RNA polimerasa II (RNAP II) producir numerosos mRNAs funcionales.
Con esto los tripanosomas evitan el laborioso proceso de capping co-transcripcional presente en muchos otros eucariotas que requieren una actividad pausada de RNAP II, el reclutamiento de enzimas con capacidad de capping y la formación de un dominio carboxi-terminal especialmente modificado dentro de la subunidad mayor de RNAP II. La transcripción génica policistrónica también permite a los tripanosomas evitar el repetitivo proceso de reinicio de la transcripción. (Palenchar et al. 2006)
Un único SL RNA es trans spliced a cada cistrón y el sitio de clivaje asociado con la poliadenilación no es un sitio consenso sino que está determinado por su distancia en relación a el tracto de polipirimidinas corriente abajo del cistrón vecino. (Mayer et al. 2005)
Se han encontrado eventos de trans-splicing en varias especies que no se encuentran filogenéticamente relacionadas. La hipótesis de que el trans-splicing surgió de forma independiente en muchos grupos requiere de un mecanismo que pueda evolucionar fácilmente. Actualmente se sabe que el trans-splicing usa la misma maquinaria molecular que el cis-splicing excepto por pequeñas modificaciones (la U1snRNP del spliceosoma está ausente y es reemplazada por una snRNP (small nuclear

ribonucleoprotein) formada por el precursor SL RNA).
Es posible transferir la capacidad de trans-splicing a organismos que no la tienen proveyéndolos del precursor SL RNA lo cual sugiere que es el único requerimiento para adquirir esta capacidad. Los SL aparte del sitio donante de splicing tienen una estructura secundaria que permite el ensamblado de una snRNP. Estas últimas dos características son compartidas con los snRNA del spliceosoma U1, U2, U4 y U5. Algunos experimentos han demostrado que es posible convertir un U1 snRNA en un SL simplemente agregando un sitio de splicing y cambiando unos pocos nucleótidos. Por lo cual la evolución repetida de trans-splicing parece ser un escenario plausible. (Douris et al. 2010)
El trans-splicing de SL está presente en el filo Euglenezoa, que abarca kinetoplástidos y euglenoides. Debido a que la divergencia entre estas clases es muy antigua se le puede considerar como una característica ancestral de este filum. Existe muy poca similitud de secuencia entre los SL de euglenoides y los de kinetoplástidos; sin embargo hay poca divergencia de secuencias entre los diferentes kinetoplástidos donde 23 de los 39-41 nts se conservan entre varios géneros. (Hastings 2005)

Regulación de la producción de mRNAs
En eucariotas y eubacteria el inicio de la transcripción es un punto regulatorio clave para controlar el nivel de expresión génica.
La maquinaria transcripcional de RNAP II de los tripanosomas resulta en una versión minimalista de lo encontrado en eucariotas superiores, junto a factores específicos de tripanosomas. Mientras que homólogos de todas las subunidades de la RNAP II de eucariotas están presentes, muchos de los factores basales de transcripción no han sido encontrados, y otros difieren mucho en estructura y actividad respecto a los de eucariotas superiores.
A pesar de esta peculiaridad, las RNA polimerasas de tripanosomas se asemejan a sus homólogas eucariotas en cuanto a actividad transcripcional y estructura de múltiples subunidades que claramente han asumido las demandas específicas del parásito.
En *T. brucei* la RNA polimerasa I transcribe los pre-rRNA del clúster de los genes ribosomales 18S, 5.8S y 28S, además de mRNAs ciclo específicos de las regiones llamadas 'expression sites', donde se encuentran la forma sanguínea de las glicoproteínas variantes de superficie (VSG), las prociclinas de la forma procíclica, los ESAGs (expression site associated genes) y los PAGs (procyclin associated genes) entre otros. La RNA polimerasa II transcribe mRNAs así como el gen SL con su cap específico de tripanosomas. La RNA polimerasa III transcribe tRNAs, 5S RNA y snRNAs. (Douris et al. 2010)
La transcripción policistrónica acoplada con la ausencia de promotores de RNAP II clásicos indica que la iniciación de la transcripción no es un factor limitante en la producción de mRNAs en tripanosomas donde parece ocurrir una expresión génica constitutiva. Estas características propias de los tripanosomas deben ser consideradas teniendo en cuenta que la principal actividad del parásito es la supervivencia en el ambiente del huésped. Factores inherentes a la supervivencia del parásito incluyen la replicación exitosa en ambientes dispares, la habilidad de adaptarse rápidamente de un huésped como el insecto vector a un mamífero, cierto grado de autonomía respecto al huésped y además evadir la respuesta del sistema inmune.
El constante flujo de transcriptos a su vez puede promover una rápida respuesta a cambios y al estrés ambiental. Mientras que esta transcripción a gran escala parece un despilfarro energético podría en realidad asegurar la supervivencia del parásito de cara a la fluctuación ambiental. En lugar de promotores diferenciales produciendo genes dependientes de RNAP II los tripanosomas tienen un genoma plástico que les permite utilizar la amplificación génica para afectar la producción de mRNAs y de ese modo la abundancia proteica. (Palenchar et al. 2006)

Transcriptómica

El transcriptoma es el conjunto completo de transcriptos en un tipo celular, en un estadio de desarrollo específico o condición fisiológica. Comprender el transcriptoma es esencial para interpretar los elementos funcionales del genoma y revelar constituyentes moleculares de células y tejidos, y también para comprender el desarrollo y las enfermedades.

El objetivo de la transcriptómica es catalogar todas las especies de transcriptos; incluyendo mRNAs, ncRNA, y snRNA para determinar la estructura transcripcional de los genes como sitios de inicio, extremos 5' y 3', patrones de splicing y modificaciones post-transcripcionales. En los últimos años el desarrollo de tecnologías de secuenciado de DNA de alto rendimiento ha provisto de un nuevo método para el mapeado y la cuantificación de transcriptomas, llamado RNASeq (RNA sequencing). (Wang et al. 2009)

RNAseq

Los experimentos de RNAseq (Figura 3) utilizan tecnologías de secuenciado de alto rendimiento para obtener información de los RNAs presentes en una muestra permitiendo determinar el nivel de expresión de los genes, eventos de splicing, ediciones del RNA, identificación de regiones no codificantes, variaciones en secuencias, SNPs, etc.

El RNA extraído de una muestra es fragmentado típicamente por hidrólisis, luego es, retrotranscripto a cDNA (en algunos casos el paso de fragmentación se da luego de esta etapa por DNAsa I o sonicación). A partir de este cDNA se genera una librería con adaptadores para posteriormente realizar una amplificación por PCR con primers de hexámeros aleatorios u oligo(dT) y luego secuenciar estos amplicones utilizando plataformas de tecnologías de secuenciado de alto rendimiento como Solid, Roche454 o Illumina. Este secuenciado final refleja la composición y cantidad de moléculas de RNA presentes en la muestra original, bajo la forma de fragmentos de cDNA de una longitud de entre 30 y 300 pb (reads) dependiendo de la tecnología utilizada.

El número de reads que se obtienen de un transcripto de RNA depende de la abundancia de ese transcripto, el largo y la densidad de reads puede ser usada para cuantificar la expresión génica con una precisión superior a los microarrays de DNA ya que tiene una menor variabilidad técnica lo cual facilita su reproducibilidad. Debido a que los costos de secuenciado se han ido reduciendo y a una mayor fiabilidad sobre los microarrays los experimentos de secuenciado masivo se han popularizado. A pesar de que en términos generales RNAseq es muy preciso, al igual que otros ensayos de expresión génica es necesario considerar los errores debido a la variabilidad inherente al proceso y a las propias fuentes. Estos dependen de los protocolos de preparación de librerías, el secuenciado y la variabilidad biológica entre réplicas del mismo experimento. (Trapnell et al. 2012)

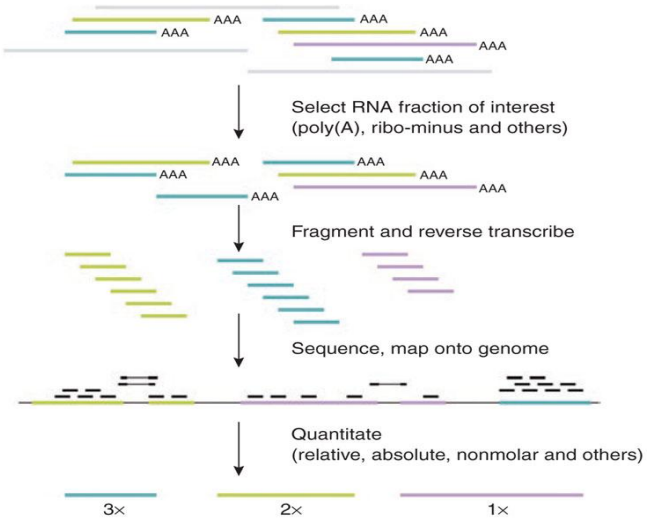


Figura 3: Pasos típicos de un experimento de RNAseq. (Mortazavi et al. 2008)

Fuentes de variabilidad en RNAseq

Obtención de muestras

Las muestras, aunque sean del mismo tejido presentan variabilidad inherente a la diversidad del material que estamos estudiando, por lo que dos secuenciaciones del mismo tejido pueden arrojar resultados diferentes al analizar su RNA. A estas muestras se las conoce como réplicas biológicas.

La expresión génica es un proceso estocástico y varía entre unidades consideradas parte de la misma población. La variabilidad biológica tiene importantes implicaciones en el diseño, análisis e interpretación de los resultados de experimentos de RNAseq. Por ejemplo una gran diferencia de expresión en cierto gen puede ser de gran importancia si este gen tiende a variar muy poco su nivel de expresión; sin embargo la misma diferencia en otro gen puede carecer de significado en el caso de que su expresión sea muy variable. Si hay disponibles solo unas pocas réplicas biológicas, será imposible determinar con precisión el nivel de variabilidad biológica en la expresión para cada gen estudiado. (Hansen et al. 2011)

Creación de librerías de cDNA

El protocolo estándar para la preparación de librerías para secuenciado de transcriptoma utilizando la plataforma Illumina Genome Analyzer comienza con la extracción total de RNA, seguido de un enriquecimiento de poly(A) utilizando beads de oligo(dT) si queremos concentrarnos en los ARNm (de lo contrario podemos realizar otra estrategia de enriquecimiento), fragmentación de RNA y retrotranscripción en DNA doble hebra complementario (dscDNA) con primers de hexámeros aleatorios. De ser necesario, podemos obtener también librerías hebra específicas utilizando distintos adaptadores en los extremos 5'y 3'' de los ARNs. El dscDNA es luego secuenciado comenzando con reparación de extremos, adición de un nucleótido de A y ligación de un adaptador. El priming con hexámeros aleatorios es utilizado para generar reads a lo largo de todo el transcrito expresado, aunque la cobertura de la secuencia resultante no es uniforme. La fragmentación de RNA antes de la retrotranscripción logra coberturas más uniformes dentro de un transcripto.

El uso de hexámeros aleatorios como primers resulta en un sesgo en la composición de nucleótidos y esto influye en la uniformidad de la ubicación de los reads a lo largo de los transcriptos expresados. Hay un fuerte patrón distintivo en la frecuencia de los primeras 13 posiciones en el extremo 5' de los reads de RNAseq. Luego de las primeras 13 posiciones, las frecuencias de nucleótidos se tornan independientes de la posición. Se han observado sesgos en la composición de nucleótidos al inicio de los reads; este efecto no puede ser atribuido a un sesgo en la amplificación por PCR, y en cierta forma es producto de los hexámeros de primers aleatorios utilizados durante la transcripción reversa. Para experimentos de RNAseq 4/5 de los reads comienzan por G o C. (Schwartz et al. 2011)

Amplificación de librerías

La amplificación por PCR de las librerías en el protocolo estándar de Illumina se eliminan muchos loci con un contenido GC mayor al 65% disminuyendo su representación a una centésima parte. Por su parte los amplicones con un contenido GC menor a 12% se ven disminuidos a una décima parte de valores estandarizados por qPCR. Entre ambos porcentajes de contenido GC no se encuentra variación notable. (Aird et al. 2011)

Secuenciado del cDNA

El error al asignar bases en el secuenciado no es uniforme, siendo mínimo al inicio del read, con un 0,3% de error y siendo máximo al final del mismo llegando a un 3,8% de errores en las últimas bases. A su vez no todos los tipos de error tienen igual probabilidad de ocurrir; siendo las transversiones (purina-pirimidina) (en particular A>C) las más frecuentes. (Dohm et al. 2008)

La aleatoriedad inherente en muchos de los pasos de preparación de RNAseq lleva a fragmentos cuyos puntos de inicio parecen ser elegidos aproximadamente de forma aleatoria. Sin embargo análisis cuidadosos revelan un sesgo tanto posicional como secuencia-específico. El sesgo posicional se refiere al efecto local por el cual ciertos fragmentos están preferentemente ubicados hacia el inicio o fin del transcripto. Lo sesgos secuencia-específico son efectos globales donde las secuencias que rodean el inicio o fin del fragmento potencial afectan la posibilidad de ser seleccionado para el secuenciado. Estos sesgos pueden afectar las estimaciones de expresión y por ello es importante corregirlos durante un análisis de un experimento de RNAseq. (Roberts et al. 2011)

Normalización de los datos

La normalización de los datos de un experimento de RNAseq permite comparar de forma precisa los datos de niveles de expresión tanto dentro de una misma librería como entre diferentes librerías.

Los métodos de normalización pueden diferir dependiendo de si se comparan datos de una misma muestra o de muestras diferentes.

La normalización de los datos dentro de una misma librería permite cuantificar los niveles de expresión de cada gen en relación a otros genes de la misma muestra. Debido a que los transcriptos de mayor longitud están representados por un mayor número de reads aunque el nivel de expresión sea el mismo que el de otro gen menos expresado; la forma de normalizar estos datos dentro de una librería es dividiendo el número de reads por la longitud de cada gen.

Cuando se desea comparar los resultados obtenidos en diferentes librerías se debe tener en cuenta la profundidad con que esa librería ha sido secuenciada, es decir si el número de veces que un transcripto es secuenciado difiere entre librerías los resultados de expresión no podrían compararse. Por ello para comparar los niveles de expresión entre librerías se debe normalizar por el número total de reads que pueden ser mapeados contra el genoma.

La forma de normalizar los datos de forma que sean comparables tanto dentro de una misma librería como entre diferentes librerías es utilizando los niveles como RPKM (reads por kilobase por millón de reads mapeados). RPKM está definido como el número de reads que mapean un gen sobre el total de reads de la librería que mapean (en millones) multiplicado por la longitud del gen (en kilobases). (Mortazavi et al. 2008)

Tecnologías de secuenciado

La principal tecnología de secuenciado de los experimentos de RNAseq utilizados en el presente trabajo para la resolución del transcriptoma de los tripanosomas fue la de Illumina. (Figura 4)

La tecnología Illumina utiliza amplificación de puente, esto es una amplificación de fase sólida en donde las moléculas de cDNA son pegadas a una superficie y amplificadas in situ generando clusters de moléculas de DNA clonalmente idénticas.

Luego se realiza un secuenciado por síntesis donde 4 análogos de dNTPS marcados con fluorescencia son terminadores reversibles. En cada paso el nucleótido correcto es incorporado y su identidad revelada por el color de su marca fluorescente. El grupo 3'OH está bloqueado para prevenir otra incorporación. Luego de obtener la información de la base incorporada, con lavado se quita la marca, se revierte el bloqueo y el proceso de síntesis continúa. Las reacciones de secuenciado ocurren en forma masiva en paralelo en celdas sobre una superficie de vidrio que contiene millones de clusters de moléculas de ADN idénticas. (Morozova et al. 2009).

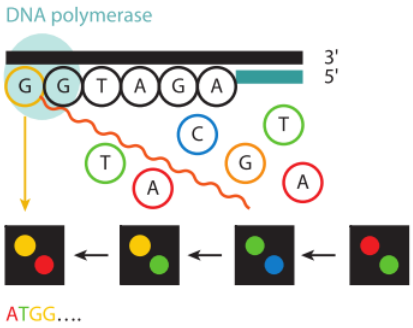


Figura 4: Illumina utiliza secuenciado por síntesis donde análogos de nucleótidos marcados con fluorescencia sirven como terminadores de reacción reversibles. (Morozova et al. 2009)

Objetivos

Los objetivos del presente trabajo fueron divididos en dos áreas; por un lado comprobar las características sobre la composición nucleotídica y su relación con los niveles de expresión en los genomas de tripanosomátidos y por otra parte estudiar la dinámica de la transcripción en estos organismos utilizando datos de RNAseq obtenidos de bases de datos públicas y de recientes secuenciados realizados en el Institut Pasteur de Montevideo.

El primer objetivo fue comparar el contenido de GC3 de los genomas de tres tripanosomas, *T. brucei*, *T. vivax* y *T. cruzi* evaluando la relación entre el contenido de GC3 y los niveles de expresión. Para ello se utilizan datos de secuenciado obtenidos de las bases de datos SRA del NCBI para *T. brucei* y de experimentos de RNAseq realizados en el Institut Pasteur de Montevideo para *T. cruzi* y *T. vivax.*

Se mapean los reads obtenidos de los experimentos de RNAseq contra los genomas de cada tripanosoma utilizando el software de alineamiento de secuencias Blast y luego se utilizan scripts en R y Python con el fin de obtener los valores sobre niveles de expresión y contenido GC3

El otro objetivo es determinar las características de los eventos de trans-splicing en las especies antes mencionadas; utilizando las mismas fuentes de datos. Se pretende estudiar las características de los eventos de trans-splicing de SL agrupándose los genes según la distancia antes del sitio de inicio del gen en donde tiene lugar el splicing. El fin de esto es establecer si distintos grupos de genes presentan patrones diferentes de trans-splicing. La estrategia utilizada para corroborarlo es realizando análisis de la ontología génica de estos grupos utilizando el test exacto de Fischer para determinar el enriquecimiento de determinados términos de ontología.

Sesgo de uso de codones

El código genético determina cual de los 61 tripletes o codones corresponden a cada uno de los 20 aminoácidos. Debido a que hay más codones que aminoácidos, el código genético es redundante o degenerado. Mientras que solo dos aminoácidos son codificados por un único codón (Met, Trp), la mayoría de los aminoácidos están codificados por dos, cuatro y seis codones diferentes.

Diferentes codones que codifican para el mismo aminoácido son conocidos como codones sinónimos. Cambios en la secuencia de DNA de una proteína entre dos codones sinónimos se asume que en general no tienen efecto y por ello los cambios sinónimos son llamados cambios silenciosos. Sin embargo a pesar de que los codones sinónimos codifican para el mismo aminoácido, se ha demostrado que en una gran variedad de organismos diferentes codones sinónimos son utilizados con frecuencias diferentes, y en general existe un sesgo hacia el uso de determinados tipos de codones. Este fenómeno ha sido llamado sesgo de codones y no es el comportamiento esperado en la distribución de codones si estos fueran asignados de forma aleatoria.

En una gran variedad de organismos, los codones sinónimos son utilizados en frecuencias diferentes. Los estudios en el uso de codones sinónimos muestran que caracteres con efectos fenotípicos minúsculos pueden ser sujetos a selección natural, aunque las variaciones en los patrones de mutación hacen difícil distinguir entre procesos evolutivos selectivos y neutrales. (Sharp et al. 1995)

Clásicamente el sesgo en la utilización de codones sinónimos ha sido asociado a un balance entre las fuerzas de selección y mutación en una población finita, con un mayor sesgo en los genes altamente expresados reflejando una fuerte selección para la eficiencia traduccional. Entre los grupos de codones sinónimos reconocidos por varios tRNAs, aquellos codones mejor reconocidos por los tRNAs más abundantes son utilizados más frecuentemente que aquellos reconocidos por especies de tRNAs poco frecuentes. Entre los codones reconocidos por el mismo tRNA, aquellos que hacen un par Watson-Crick natural con el anticodón en la posición oscilante "wobble" son generalmente los utilizados, aunque hay excepciones a esta regla. El sesgo es más extremo en los genes altamente expresados lo cuales producen grandes cantidades de proteínas que en aquellos genes débilmente expresados. (Bulmer 1991)

Estos dos modelos, es decir el de mutación y el de selección, hacen predicciones diferentes. Por ejemplo, si hay una fuerte presión selectiva para mejorar la eficiencia traduccional, entonces esta presión será mayor para los genes altamente expresados que para aquellos con menor expresión. Entonces el modelo de selección traduccional predice una correlación entre el sesgo de uso de codones y niveles de expresión génica. Los codones sinónimos que son utilizados preferentemente en genes altamente expresados (codones óptimos) deberían también corresponder a los tRNAs más abundantes. Además, la acción de la selección debería resultar en una disminución en la probabilidad de fijar mutaciones hacia codones no óptimos. Por otra parte, el modelo de sesgo mutacional a priori no predice ninguna relación

entre el sesgo de uso de codones y la expresión génica. Tal presión mutacional debería afectar todas las posiciones en un genoma, no solo los sitios sinónimos sino también los sitios silenciosos como intrones o regiones intragénicas. Otros procesos aparte de la selección y la mutación pueden afectar el uso de codones. Notablemente se ha mostrado que la transcripción puede ser mutagénica pudiendo inducir una correlación entre la expresión génica y la composición de bases, y por ende el uso de codones. (Duret 2002)

Se ha demostrado que la variación en el sesgo de codones entre genes del mismo organismo depende de varios parámetros, tales como los niveles de expresión, composición de aminoácidos, longitud del gen, estructura de los mRNA y consideraciones de nivel de ruido de proteínas. En la mayoría de estos casos, existe la evidencia de que la selección actuando en diferentes pasos durante la expresión de proteínas da forma al sesgo de codones. Además fuerzas globales diferencian el sesgo de codones de genes entre diferentes organismos. El sesgo de codones especie específico está fuertemente correlacionado con el porcentaje de contenido GC promedio del organismo; genes de organismos filogenéticamente relacionados o con un contenido de tRNAs similar tienen un sesgo de uso de codones similar. El parámetro más significativo que explica la diferencia en el sesgo de codones entre diferentes organismos es el contenido GC, el cual está determinado por procesos que involucran a todo el genoma. En algunos casos el sesgo de uso de codones en procariotas puede ser predicha utilizando información obtenida únicamente de secuencias intragénicas. (Chen et al. 2004).

Las explicaciones selectivas y neutrales sobre el sesgo de uso de codones no son mutuamente excluyentes, y ambos tipos de mecanismos seguramente tienen un papel en dar forma al patrón de variación dentro y entre genomas. Los codones adaptados a los pools de tRNA son generalmente utilizados en los genes de alta expresión, porque estos genes experimentan gran presión para la eficiencia traduccional, la precisión o ambas. La elongación eficiente de un transcripto puede aumentar el número de proteínas resultantes o puede proveer de un beneficio global para la célula aumentando el número de ribosomas disponibles para traducir otros mensajeros, aun si no mejora la eficiencia del propio transcrito. Una elongación precisa beneficia la célula reduciendo el costo de productos inútiles de traducciones erróneas. Existen evidencias que apuntan a ambos casos, y ambas hipótesis no son mutuamente excluyentes. (Plotkin et al. 2011) Existen algunos casos en los cuales la presión mutacional y la selección van en direcciones contrapuestas, por ejemplo podemos tener sesgo mutacional hacia AT, pero codones óptimos con en G o C en sus terceras posiciones. Esto genera un gradiente en el cual los genes de menor expresión, y por tanto con menor presión selectiva, son ricos en AT en su tercera posición, mientras que los genes de alta expresión, donde domina la selección traduccional, usan codones terminados en G o C. De esta forma se genera un gradiente en el contenido GC3 de los genes que es función del nivel de expresión

Debido a que la mayoría de los aminoácidos permiten sustituciones sinónimas que cambian el contenido de G+C en la tercera posición del codón (GC3), éste representa la principal fuente de variación en el uso de codones en las especies de tripanosomátidos. Por otra parte el estudio de las secuencias no codificantes de tripanosomas muestra un sesgo mutacional hacia A+T lo cual afecta de forma diferente las secuencias codificantes y a las no codificantes. El efecto de estas presiones en la tercera posición del codón parece ser inversamente proporcional a los niveles de expresión génica. (Alvarez et al. 1994)

Materiales y métodos

Debido a las características propias de los genomas de los tripanosomas, como la organización de los genes en policistrones donde grupos de genes son co-transcriptos y dado que no poseen una maquinaria que permita la regulación del inicio de la transcripción; la mayor parte de la regulación de la expresión génica debe darse a nivel post-transcripcional.

Por ello el nivel de expresión no se puede correlacionar exactamente con la abundancia de los transcriptos; sin embargo es la mejor aproximación disponible.

Para analizar las preferencias de uso de codones sinónimos y verificar su sesgo en tripanosomátidos se estudiaron la abundancia de transcriptos de los genes de *T. brucei*, *T. cruzi* y *T. vivax* con la finalidad de evaluar el uso de codones en los genes de mayor y menor expresión.

En los tres casos los datos del transcriptoma (ver Tabla 2) fueron obtenidos por secuenciado masivo de mRNA (RNAseq) utilizando tecnología Illumina. Los datos de secuenciado de *T. cruzi* y *T. vivax* se obtuvieron del Institut Pasteur Montevideo y los de *T. brucei* fueron descargados desde la base de datos

Sequence Read Archive del NCBI (www.ncbi.nlm.nih.gov/sra), submission number SRA012290, datos obtenidos por Kolev et al.
En el caso de *T. cruzi* se utilizaron datos de secuenciado de información de tres estadios 10.111.341 reads para el estadio amastigota, 15.107.022 reads para el estadio epimastigota y 14.304.207 reads para el estadio tripomastigota, mapeados contra el genoma de 23.216 genes.
En *T. vivax*, los reads utilizados fueron 28.896.031 para el estadio sanguíneo, mapeados contra los 11.865 genes
Los datos de secuenciados de *T. brucei* obtenidos de la base SRA consta de 33.338.202 reads del estadio procíclico y el genoma cuenta con 8.529 genes.
Los reads del transcriptoma de los tres organismos fueron mapeados contra sus respectivos genomas utilizando BLAST.

Especie	Estadio	N de reads	N de genes
T. cruzi	Amastigota	10.111.341	23.216
	Epimastigota	15.107.022	
	Tripomastigota	14.304.207	
T. brucei	Procíclico	33.338.202	8.529
T. vivax	Sanguíneo	28.896.031	11.865

Tabla 2: Tabla con el número de reads y genes para cada especie y estadio utilizados para este trabajo.

Los niveles de expresión dentro de cada librería fueron normalizados dividiendo el número de reads que pueden ser alineados por la longitud del gen. Esto es debido a que los transcriptos de mayor longitud están representados por un mayor número de reads aunque el nivel de expresión sea el mismo que el de otro gen menos expresado.
Con los resultados de los niveles de expresión de cada gen, se procedió a analizar el contenido en G y C en la tercera posición del codón (GC3) para cada uno de ellos, utilizando scripts en Python y realizando histogramas para diferentes porcentajes de los genes de mayor y menor expresión utilizando el lenguaje R.

Resultados y discusión

GC3 para genes de mayor y menor expresión

Contenido de GC3 en *T. vivax*

Para *T. vivax* en el estadio sanguíneo, hay una marcada diferencia en el contenido GC3 para los genes de mayor y menor expresión. (Figura 5) El promedio de GC3 para el 20% de los genes de menor expresión es de 0,5345 y para el 20% de los genes de mayor expresión es de 0,6347.

En el histograma el pico para los genes de menor expresión se da alrededor de 0,51, donde más de la mitad son proteínas hipotéticas conservadas. En los genes de mayor expresión hay una concentración de genes entre los valores de GC3 de 0,6 y 0,7 donde predominan genes de proteínas hipotéticas y también se encuentran genes de RNA polimerasa DNA dependiente, proteínas ribosomales 40S y 60S y RHS. Cerca de 0,9 también aumenta el número de genes, siendo mayormente genes de las histonas H2A, H2B y H4.

Los resultados en *T. vivax* son los que muestran una distinción más clara entre los genes de alta y baja expresión en relación a su contenido de GC3, tal cual es esperable.

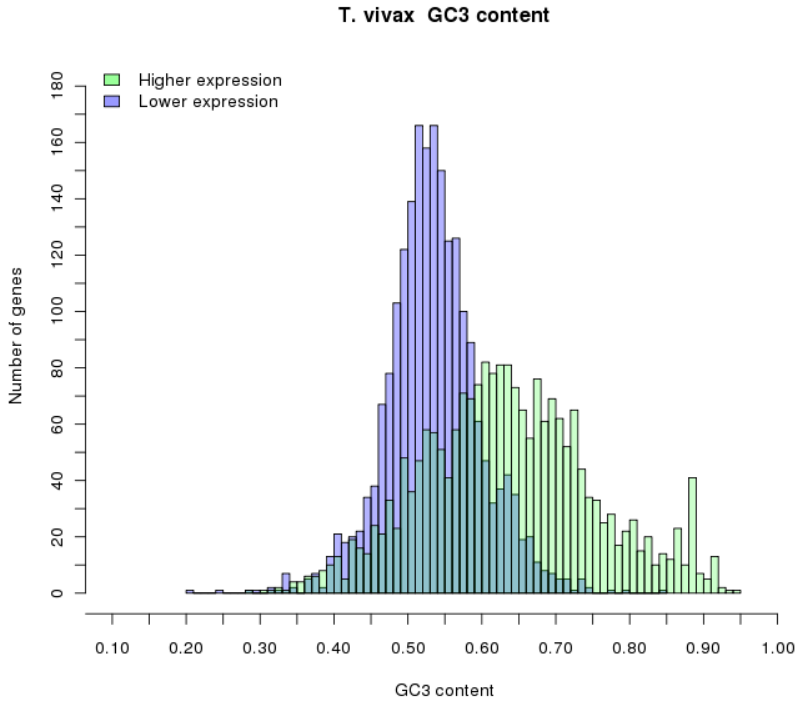


Figura 5: Histograma de uso de codones en *T. vivax* para el 20% de los genes de mayor y menor expresión

Contenido de GC3 en *T. brucei*

En *T. brucei* en el estadio procíclico también puede notarse una diferencia en el contenido GC3 en para el 20% de los genes de mayor expresión y el 20% de los de menor expresión. (Figura 6)

El promedio de GC3 para los genes de mayor expresión es de 0,4825 y para los genes de menor expresión es de 0,5497.

La concentración de genes de alta expresión a partir del valor de 0,7 de GC3 son genes de histonas H2A, H2B, proteínas ribosomales 40S y 60S y algunas HSP.

La diferencia entre los valores de GC3 para los genes de mayor y menor expresión no es tan notable como lo es en el caso de *T. vivax*, pero aún así se puede observar la tendencia a un mayor contenido de GC3 por parte de los genes mas expresados.

Los genes de alta expresión pero de bajo contenido en GC3 son en su mayoría proteínas hipotéticas y varios RHS; los de baja expresión pero con un contenido de GC3 alto son en su mayoría proteínas hipotéticas.

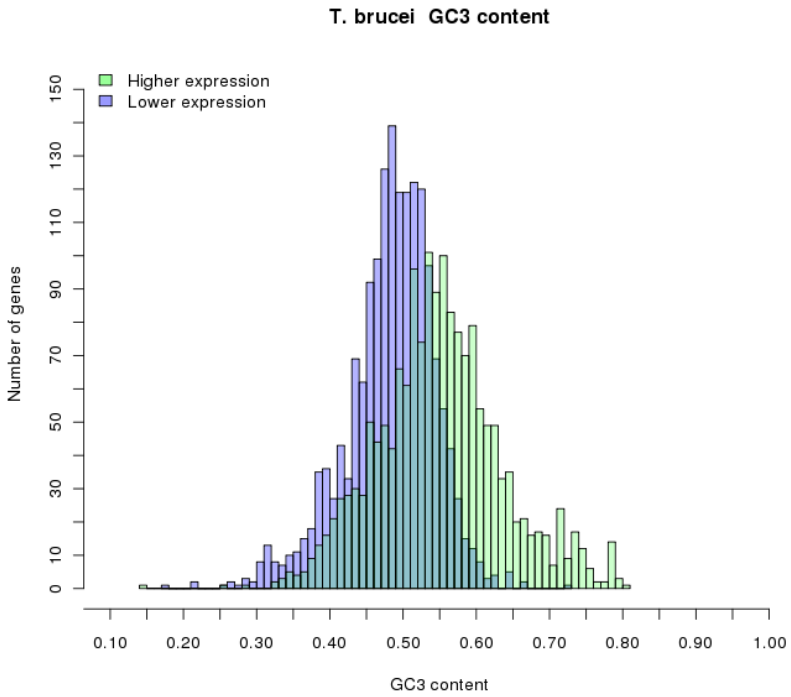


Figura 6: Histograma de uso de codones en *T. brucei* en el estadio procíclico, para el 20% de los genes de mayor y menor expresión

Contenido de GC3 en *T. cruzi*

En el caso de *T. cruzi* para el estadio amastigota no se observa el mismo patrón de uso de codones, siendo el contenido GC3 algo mayor en los genes de menor expresión. (Figura 7)

En la curva de los genes de mayor expresión las salientes entre niveles de contenido GC3 de 0,9 a 0,95 son casi en su totalidad histonas, alrededor de 0,65 son genes de mucinas TcMUCI, TcMUCII, TcSMUGS y TcSMUGL.

El pico principal de la gráfica que concentra la mayor cantidad de genes se encuentra en 0,55 y se trata en su mayoría de trans-sialidasas y retroposon hot spots. La saliente a la izquierda de bajo contenido en GC3, de la que inclina la tendencia a hacia genes altamente expresados con bajo GC3, son en su mayor parte genes de MASP (mucine-associated suface proteins) y podría ser un artefacto ya que por tratarse de una familia multigénica es posible que aparezca sobrerepresentada en el histograma. Si se considera un contenido de GC3 de 0,35 a 0,45 más de la mitad son genes de MASP.

En la curva de los genes de menor expresión en general, unas tres cuartas partes, se tratan de proteínas hipotéticas conservadas y ningún grupo de genes en particular sobresale.

El promedio de GC3 para los genes de mayor expresión es de 0,5586 y para los genes de menor expresión es de 0,5933.

Para el estadio epimastigota de *T. cruzi* no hay mayores diferencias con el estadio amastigota, solo una menor expresión en general de los genes. En el estadio tripomastigota dentro de los genes de mayor expresión se encuentra un incremento en los niveles de los genes de MASP.

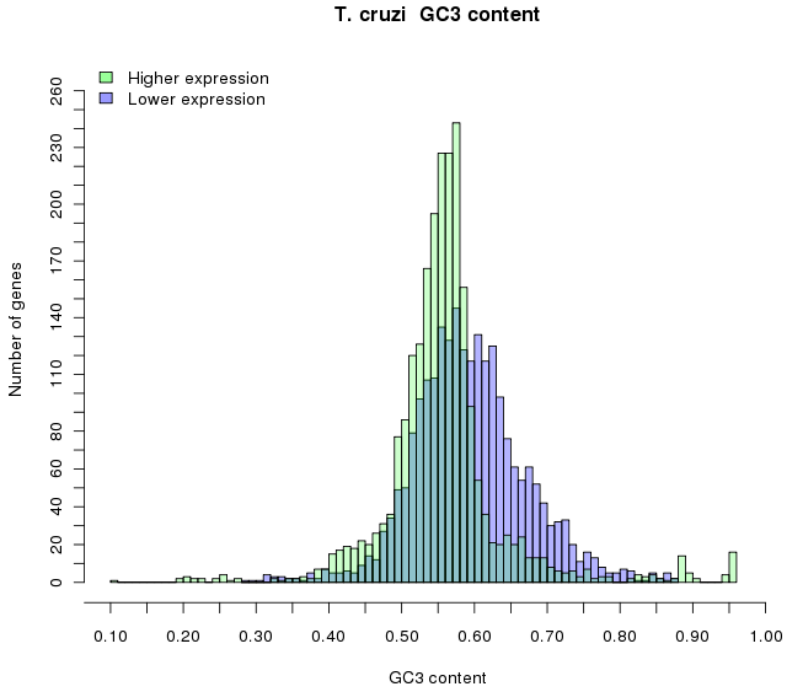


Figura 7: Histograma de uso de codones en *T. cruzi* en el estadio amastigota, para el 20% de los genes de mayor y menor expresión

GC3 y conservación de genes

Debido a que los genes de *T. cruzi* no presentan el comportamiento esperado en cuanto a la relación entre los niveles de expresión y GC3 similar al descrito para los restantes tripanosomas, se decide verificar si el grado de conservación de los genes afecta esta relación. Para ello todos los grupos de genes ortólogos son alineados utilizando ClustalW y en base a las distancias obtenidas para cada alineamiento, se dividen los genes en dos grupos, los conservados y los no conservados. Los valores de distancia calculados por ClustalW reflejan el número de sustituciones en relación al largo del alineamiento sin contar los gaps. Son considerados como conservados aquellos genes cuya distancia al comparar las secuencias entre las especies es menor a un valor de 0.2 cambios por sitio.

A su vez los genes considerados como conservados y no conservados fueron separados en base a su contenido de GC3 dividiéndolos como GC3 alto y bajo asignando aproximadamente la mitad de los genes para cada grupo. Este procedimiento se realizó en las tres especies que estamos estudiando en este trabajo

Cuando se promedian los niveles de expresión y el contenido de GC3 para cada uno de estos grupos se puede observar que el comportamiento de los genes se aproxima al esperado en todos los casos. En el caso de *T. brucei* y de *T. viva* las diferencias son más notables y suceden tanto para los genes conservados como para los no conservados, siendo las diferencias de expresión siempre mayores en el grupo de genes conservados.

En *T. brucei* en los genes conservados el nivel de expresión se ve quintuplicado cuando el contenido GC3 aumenta en un 8%; en *T. vivax* la diferencia es algo menor, pero aun así los niveles de expresión son triplicados para los genes con un GC3 un 11% mayor. En los genes no conservados las diferencias entre los niveles de expresión y el contenido GC3 son mucho menores, pero aun así se mantiene la relación esperada.

En *T. cruzi* en los genes conservados la diferencia es mucho menor, la expresión aumenta un 20%, cuando el contenido GC3 aumenta un 23% y en los genes no conservados no se encuentra esta relación entre expresión y GC3.

Cuando se compara el contenido de GC3 y nivel de expresión utilizando solamente aquellos genes que se encuentran conservados entre los tres tripanosomas, en todos los casos se encuentra una relación entre contenido GC3 y nivel de expresión, tal cual se esperaba.

Anteriormente cuando se consideraron todos los genes, *T.vivax* y *T.brucei* mostraban esta relación, pero no así *T.cruzi*. La diferencia en el contenido GC3 es notablemente mayor en todos los tripanosomas, si considero solamente los genes conservados, en lugar de tomar solo el 20% de los más y menos expresados (Tabla 3).

		Tbrucei		Tcruzi		Tvivax	
		expresión	GC3	expresión	GC3	expresión	GC3
conservados	GC3 alto	2.369	0.581	0.053	0.670	73.801	0.610
	GC3 bajo	0.525	0.500	0.044	0.521	26.084	0.505
no conservados	GC3 alto	0.360	0.557	0.038	0.642	30.072	0.578
	GC3 bajo	0.302	0.468	0.041	0.517	19.081	0.482

Tabla 3: Valores de expresión de los tres tripanosomas y el contenido GC3 para genes conservados y no conservados.

Preferencia de uso de GC3 en cada uno de los grupos de codones sinónimos

Cuando se estudia la preferencia de codones sinónimos para cada uno de los aminoácidos entre los genes de mayor y menor expresión se puede notar de clara forma la preferencia de codones finalizados en GC en los genes de mayor expresión. Estas diferencias son notables en el caso de *T. vivax* (Figura 8) y *T. brucei* (Figura 9), no habiendo tal sesgo en *T. cruzi* (Figura 10), coincidiendo con los histogramas presentados anteriormente.
Esto indica claramente que la preferencia por codones terminados en C o G es coherente en todos los aminoácidos

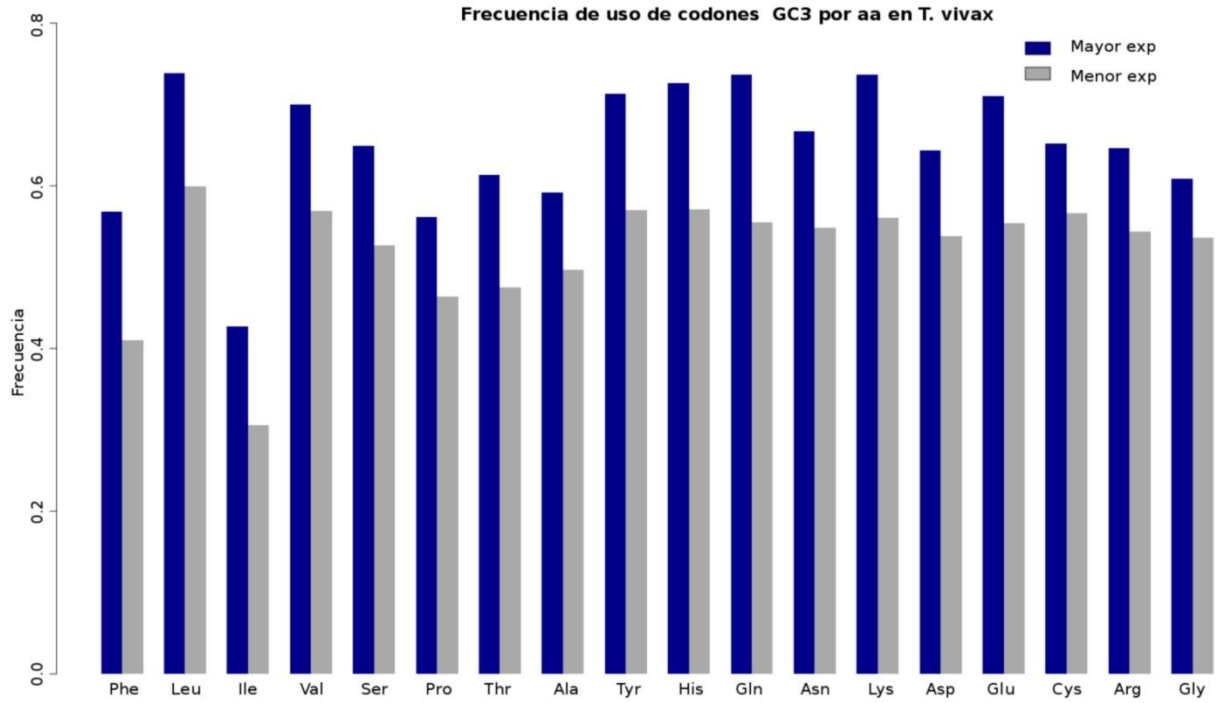


Figura 8: Preferencia en el uso de codones sinónimos en *T. vivax*, para el 20% de los genes de mayor y menor expresión

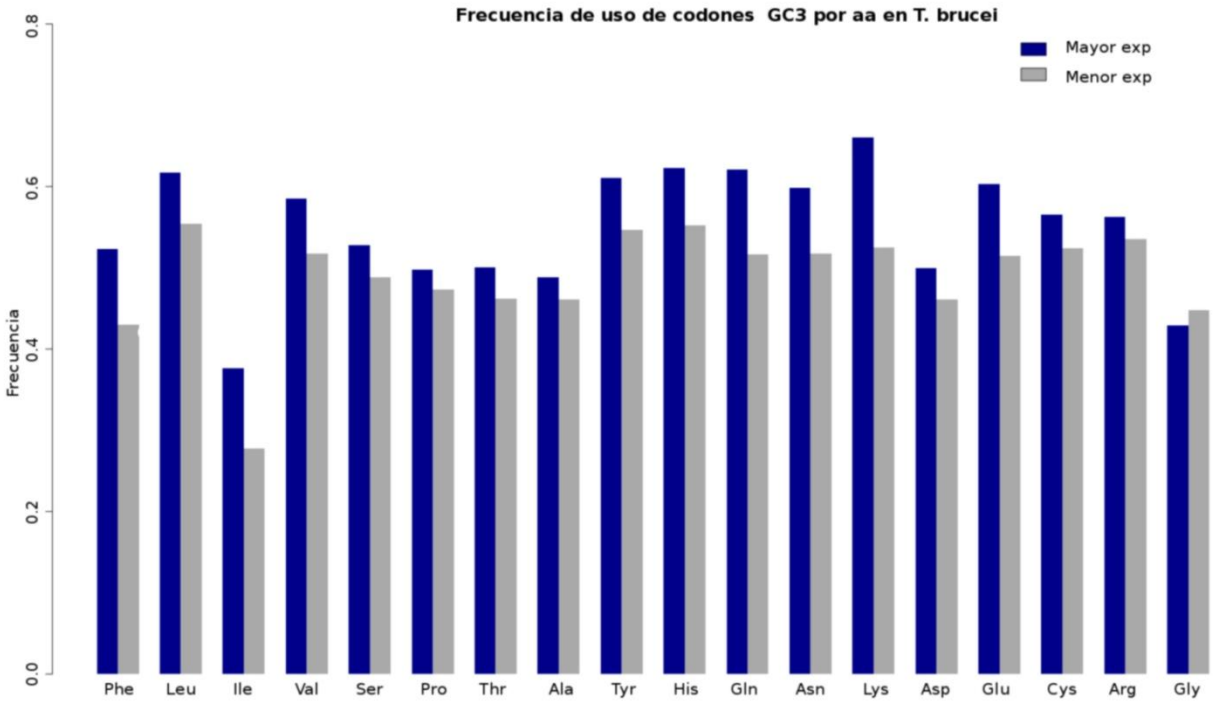


Figura 9: Preferencia en el uso de codones sinónimos en *T. brucei*, para el 20% de los genes de mayor y menor expresión

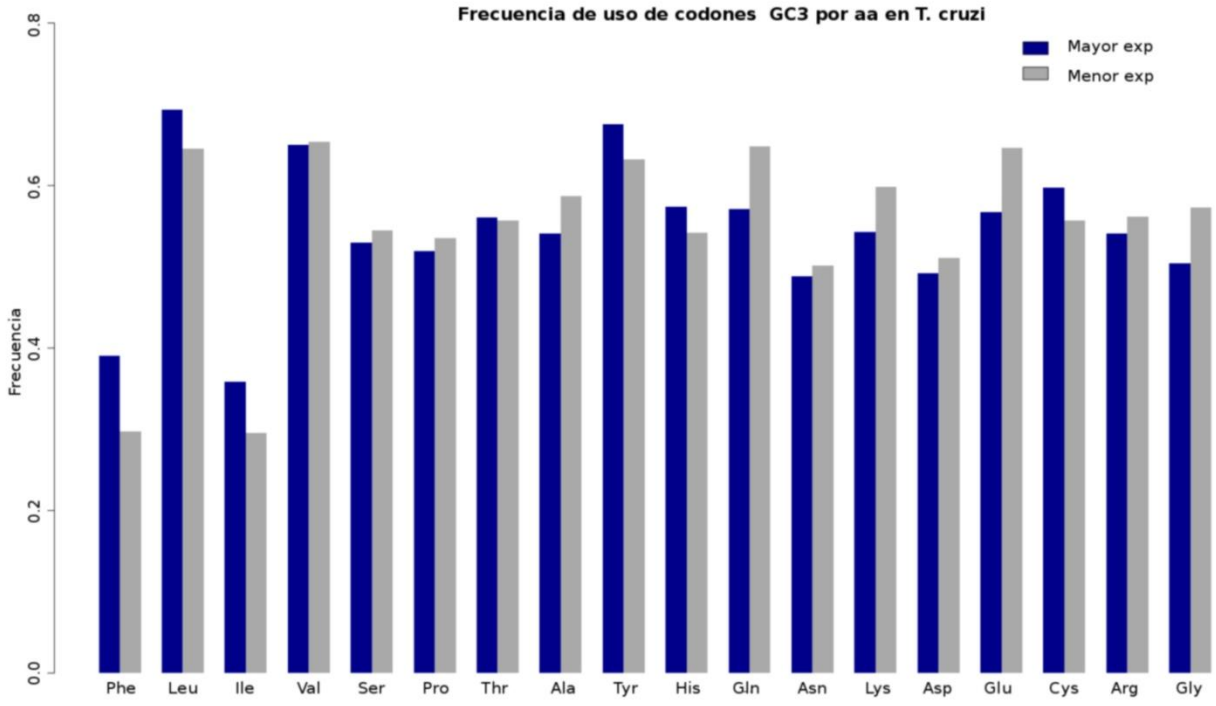


Figura 10: Preferencia en el uso de codones sinónimos en *T. cruzi*, para el 20% de los genes de mayor y menor expresión

Trans-splicing y el spliced leader

Debido a la peculiar organización del genoma de los tripanosomátidos en grupos de genes co-transcriptos como policistrones los sitios y mecanismos de iniciación de la transcripción y terminación para genes codificantes de proteínas son en su mayoría desconocidos.

Debe realizarse un proceso de maduración de los RNA policistrónicos para producir mRNAs individuales que maduran durante el proceso acoplado de trans-splicing y poliadenilación.

Durante el trans-splicing, el RNA del Spliced Leader (SL) dona una secuencia de 39 nucleótidos que luego encontraremos unida al extremo 5' de todos los mRNAs y provee de la estructura 5' para el cap del mRNA (Figura 11). El mecanismo de splicing intramolecular es una característica antigua de los eucariotas y se encuentra en muchos protozoarios, nematodos, platelmintos, algunos cordados (tunicados) y otros organismos, e involucra mecanismos similares al cis-splicing. El spliceosoma es capaz de efectuar cis-splicing de pre-mRNA como la eliminación de intrones y además de realizar trans-splicing, esto es, unir dos moléculas de RNA separadas para formar una molécula quimérica.

El proceso de trans-splicing en tripanosomas es muy similar al cis-splicing canónico. El RNA de SL es uno de los sustratos y presenta el dinucleótido GU en el extremo 5' del intrón y un AG en 3', en el sitio donde ocurre el splicing. (Mayer et al. 2005)

Existen dos formas en las que puede ocurrir el trans-splicing; cuando el sitio del pre-mRNA donante se une a un sitio aceptor el cual en la mayoría de los casos es transcrito desde el mismo gen o región genómica y por trans-splicing de SL. En el trans-splicing por SL el donante es una molécula de SL RNA cuya única función es donar una corta secuencia líder la cual es transferida al sitio aceptor de trans-splicing del pre-mRNA, el cual se transforma en el extremo 5' del mRNA maduro. Los RNAs de SL son secuencias cortas que contiene sitios donantes pero ningún sitio aceptor y una estructura de cap 5' hipermodificada. El sitio donante divide funcionalmente la molécula en dos segmentos. Durante el splicing el segmento 5' se comporta como el primer exón de un gen y el segmento 3' como la parte 5' de un intrón convencional. (Hastings 2005)

La secuencia señal que determina el sitio aceptor de trans-splicing en general suele consistir del dinucleótido AG en el sitio de unión del exón precedido por una secuencia de polipirimidinas de longitud variable. Nucleótidos adicionales corriente abajo del sitio de trans-splicing parecen modular la eficiencia del trans-splicing, aunque solo unos pocos casos han sido estudiados. El trans-splicing se encuentra espacialmente y temporalmente acoplado a la poliadenilación del mRNA que se encuentra corriente arriba en el policistrón, es decir en el gen anterior. (Kolev et al 2010)

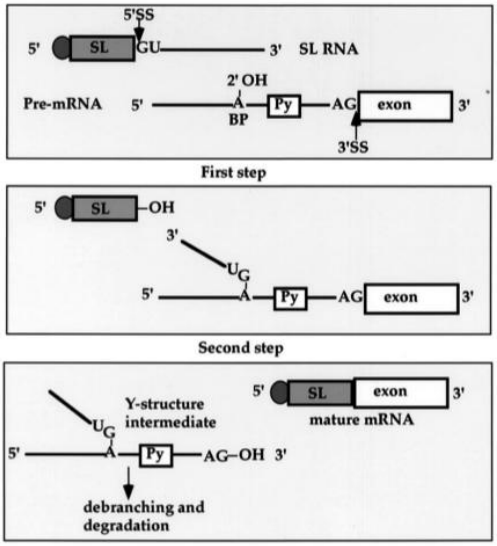


Figura 11: Serie de pasos que involucra el evento de trans-splicing de SL. Están señalados el sitio de trans-splicing 5' GU en el SL RNA y el sitio 3' AG en el pre-mRNA, BP (branching point) el sitio de ramificación y Py, el tracto de polipirimidinas. (Liang et al. 2003)

Para los tripanosomátidos el SL RNA es un RNA estructural esencial y necesario en grandes cantidades de forma continua, para asegurar la expresión de genes codificantes de proteínas. La adición de SL por trans-splicing es un paso de procesado obligatorio para los mRNAs de tripanosomátidos y requiere de una molécula de SL RNA para la maduración de una de mRNA. Por ello el patógeno depende en una fuerte expresión constitutiva de los genes de SL RNA a lo largo de todo su ciclo de vida. Todos los organismos con trans-splicing de SL tienen una o unas pocas especies de RNA de SL; en general los tripanosomátidos tienen aproximadamente unos 200 genes sin intrones transcritos por la RNAP II los cuales se encuentran organizados en repeticiones en tándem de aproximadamente 1,35 kb las que permiten mantener la elevada tasa de síntesis de SL RNA. (Schimanski et al. 2004)
Entre los tripanosomas estudiados y en general en todos los miembros de la clase Kinetoplastida, las secuencias de SL muestran un elevado grado de conservación (Figura 12).

```
AACTAAAGCTTTTATTAGAACAGTTTCTGTACTATATTG   T.vivax
AACTAACGCTATTATTAGAACAGTTTCTGTACTATATTG   T.brucei
AACTAACGCTATTATTGATACAGTTTCTGTACTATATTG   T.cruzi
```

Figura 12: Conservación de las secuencias del spliced leader en diferentes especies de tripanosomas
.

Materiales y métodos
Se utilizaron los 33,3 millones de reads de RNAseq de *T. brucei* del estadio procíclico obtenidos de la base de datos SRA del NCBI del trabajo de Kolev et al. y los 28,9 millones de reads de *T. vivax* para el estadio sanguíneo secuenciados en el Institut Pasteur de Montevideo, cedidos por G. Greif.
El trabajo de Kolev et al. fue realizado utilizando *T. brucei rhodesiense,* aislando los poly(A) y generando los cDNA doble hebra con primers de hexadeoxinucleótidos o oligo(dT). Esto genera un enriquecimiento en reads generados hacia el extremo 3' de los transcriptos, por lo cual además utilizaron la secuencia del SL como primers para así capturar el extremo 5' de los transcriptos. (Kolev et al. 2005)
Está técnica conocida como SL trapping enriquece la muestra con las secuencias de los SL y facilita el estudio del trans-splicing, por lo cual el número de reads utilizados para mapear los SL en *T. brucei* fue mucho más elevado que lo utilizado en *T. vivax,* el cual se hizo con polyA enriqueciendo de esta forma en mRNAs.
Para estudiar los eventos de trans-splicing de SL se realizó un alineamiento utilizando BLAST con las secuencias de los SL contra los reads con el fin de identificar aquellos que contuvieran secuencias SL y por tanto derivados de esta parte del mensajero.
Luego se eliminó de cada read la secuencia desde la primer base hasta el final del SL y estos reads trimeados son mapeados nuevamente contra los genes de cada tripanosoma considerando una secuencia de 500 bases antes del sitio de inicio de cada gen, para de esta forma poder determinar el sitio de inserción de la secuencia del SL en el 5' UTR del gen. De este alineamiento se tomaron solamente aquellos reads que mapean desde la posición de inicio y no se admitieron mismatches para reducir posible matcheos ruidosos.
Los sitios de trans-splicing fueron determinados utilizando un script en Python en donde se consideraron sitios promedio en alrededor de 3 bases de ls posición de inserción del SL. Es decir todos aquellos reads que mapearon en torno a las 3 bases de un sitio fueron asignados al mismo sitio de trans-splicing y se consideró la posición promedio.
A partir de estas posiciones se determinó la ubicación respecto al inicio del gen de cada sitio de splicing.
Se observó que en algunos casos ciertas posiciones de los sitios de trans-splicing caían en el interior de la región codificante de los genes, por lo cual se decidió realizar alineamientos de genes ortólogos agregando otra especie de tripanosoma, *T. congolense*, con el fin de verificar la precisión de la anotación de los codones de iniciación. Este alineamiento fue realizado utilizando el software ClustalW.
Por otra parte se agruparon los genes según la distancia del sitio de trans-splicing respecto al inicio del gen con el fin de estudiar si existe alguna relación entre la distancia AUG al sitio de trans-splicing y ciertos grupos de genes. Para ello se realizaron análisis de ontología génica con el fin de evaluar si existe un enriquecimiento en ciertos tipos de genes para determinadas distancias de trans-splicing, utilizando el software Blast2GO.

Resultados y discusión

Genes con spliced leader

Se realizaron alineamientos de los SL contra los reads para *T.brucei* y *T.vivax* teniendo en cuenta una distancia de 500 nucleótidos antes del inicio del gen.
En *T.brucei* se consideraron solamente aquellas secuencias donde mapearon más de 5 reads; de los 8094 genes expresados en 6263 se identificaron los sitios aceptores de SL. Para *T. vivax*, sin embargo debido al menor tamaño de la muestra, se consideró el SL solamente cuando mapearon más de 2 reads y se encontraron sitios SL en 4262 de los 10632 genes expresados.
La diferencia respecto al número de reads requeridos busca reflejar el hecho de que en los experimentos de RNAseq fueron diferentes. *T.* brucei fue realizado utilizando SL-trapping y *T. vivax* con polyA que enriquece la muestra en mRNAs por lo cual para este último los reads que mapearon contra SL fueron muchos menos. Un aspecto a remarcar y sobre el cual volveremos más adelante es que alrededor de 1/3 de los genes de *T. brucei* y 1/4 de los de *T.vivax* tienen sitios aceptores de SL a menos de 35 nucleótidos de distancia del inicio del gen.

Número de spliced leaders por gen

Muchos de los genes analizados tienen más de un sitio aceptor de trans-splicing en la región estudiada (Tabla 4). En los casos más extremos ciertos genes tienen hasta 5 sitios de trans-splicing y en algunos de ellos pueden haber sitios internos al gen (Tabla 5); en general estos sitios de trans-splicing son cercanos a la región de inicio del gen (de -35 al inicio del gen), pero en algunos casos se observan sitios de trans-splicing internos al gen en centenares o miles de bases.
Más de un sitio aceptor dentro de la misma región intergénica puede ser utilizado para la generación de mRNAs maduros que codifican para el mismo ORF pero con diferentes extensiones para su 5' UTR.
El papel de la presencia de más de un sitio aceptor es incierto sin embargo podría estar relacionado con cierto impulso evolutivo que permite la acumulación de sitios aceptores corriente arriba del ORF garantizando el procesado correcto del pre-mRNA por trans-splicing. (Mayer et al. 2005)
En kinetoplástidos, alteraciones en el uso del sitio aceptor de un gen puede implicar en sitios de poliadenilación alternativa en el gen corriente arriba. Debido a que no hay secuencias señal para la poliadenilación, la cual ocurre a unos 100 nt de distancia del sitio de trans-splicing. Los sitios aceptores alternativos pueden ser utilizados como un mecanismo regulador de la expresión de los genes corriente arriba. Cabe destacar que la regulación de la expresión génica en tripanosomas es principalmente post-transcripcional y diferentes 5' y 3' UTR pueden contribuir con esta regulación generando diferentes estructuras en los UTR o siendo blanco de diferentes factores de traducción. (Mayer et al. 2005)

Gen	Producto	read	pos	read	pos	read	pos	read	pos	read	pos	read	pos	read	pos
Tb427tmp.01.3676	40S ribosomal protein S17 putative	2560	**-18**	316	**-14**	7	**-11**	79	**-5**	6	**-2**	9	**9**	7	**21**
Tb427.07.1040	40S ribosomal protein S16 putative	40	**-35**	8	**-30**	3287	**-14**	451	**-10**	14	**-6**	27	**15**	6	**18**
Tb427tmp.211.3280	60S ribosomal subunit protein L31 putative	6	**-32**	3690	**-21**	502	**-17**	51	**-4**	5	**0**				
Tb427tmp.211.2640	60S ribosomal protein L23 putative	4985	**-15**	514	**-11**	118	**-5**	17	**-1**	9	**9**				
TvY486_1001030	40S ribosomal protein S23 putative	266	**-13**	4	**-8**	2	**7**								
TvY486_1115850	chaperonin HSP60 mitochondrial precursor	4	**-30**	2	**-26**	3	**-13**								
TvY486_1111010	60S ribosomal protein L22 putative	289	**-19**	2	**-15**	2	**-10**								
TvY486_1103060	60S ribosomal protein L37 putative	257	**-34**	11	**-20**	2	**-16**								

Tabla 4: Genes con varios sitios de splicing en *T. brucei* (genes Tb427) y en *T. vivax* (genes TvY486). Read indica el número de reads que mapean y pos, la posición en donde ocurre el trans-splicing. Se consideraron las posiciones donde mapean más de 2 reads en ambas especies. Los números negativos indican las posiciones anteriores al sitio de inicio y los positivos las posiciones internas al gen.

Los eventos de trans-splicing muy internos al gen podrían considerarse como sitios de trans-splicing del gen siguiente; sin embargo aquellos a pocas bases del inicio del gen, tras descartar problemas de anotación, requieren de mayor investigación.

Gen	Producto	read	pos	read	pos	read	pos	read	pos	read	pos	read	pos
TvY486_1111000	hypothetical protein	2	**5**	289	**9**	2	**203**	2	**243**	289	**437**	2	**441**
TvY486_1111270	hypothetical protein	2	**15**	2	**35**	2	**297**	2	**317**				
Tb427.02.1860	hypothetical protein conserved	7	**1159**	9	**1230**	30	**1255**	5	**1258**				
Tb427.05.330	receptor-type adenylate cyclase GRESAG 4	26	**76**	13	**1462**	33	**1493**	10	**2589**				
Tb427.08.5750	hypothetical protein conserved	7	**927**	7	**2189**								

Tabla 5: Genes con varios sitios de trans-splicing internos en *T. brucei* (genes Tb427) y en T. vivax (genes TvY486). Reads indica el número de reads que mapean y pos, la posición en donde ocurre el trans-splicing en relación al AUG de inicio. Los números negativos indican las posiciones anteriores al sitio de inicio y los positivos las posiciones internas al gen. En el caso de los genes de *T. vivax* los únicos sitios de trans-splicing de estos genes son internos; mientras que en *T. brucei*, tienen además de los mostrados un par de sitios externos prioritarios para el trans-splicing, donde mapean un mayor número de reads.

La utilización de sitios aceptores internos a los ORFs podría contribuir a una regulación negativa de la expresión génica post-transcripcional. Otra explicación posible del significado de los sitios aceptores dentro de los ORFs deriva de estudios de la proteína LYT1 de *T. cruzi* la cual tiene tres transcriptos diferentes, uno con trans-splicing en la región intergénica (en la posición -46) y dos dentro del ORF (uno en +10 y otro en +85). Cuando se expresa el transcripto intergénico (en -46) la proteína producida resultante tendrá una señal de exportación a la membrana, en el caso del transcripto con el trans-splicing en +10 el producto tendrá localización nuclear y en el otro caso, en +85, el transcripto produce una proteína trunca. Estos tres productos son regulados por lo cual tienen una expresión diferencial durante el ciclo de vida del parásito. (Manning-Cela et al. 2002)

Conservación de los sitios de trans-splicing

Para estudiar el grado de conservación de los sitios de inserción del spliced leader en *T. brucei* y *T. vivax*, se estudiaron los genes ortólogos de ambas especies. Los genes ortólogos fueron obtenidos por best reciprocal hit de alineamientos realizados con BLAST y luego se alinearon con ClustalW.

De los ortólogos se tomo la secuencia aminoacídica y nucleotídica, para ésta última con un margen de 50 nucleótidos hacia 5' del inicio del gen.

En la mayoría de los alineamientos de los genes ortólogos muestran cierta conservación en las posiciones de los sitios de splicing, pero muy poca en las secuencias. Se estudiaron 200 genes ortólogos de *T.brucei* y *T.vivax* de los cuales 37 mostraban una ubicación de su sitio de trans-splicing en una región menor o igual a 3 pb.

Corrección de la anotación

Algunos de los genes con trans-splicing registran estos eventos en regiones internas de los genes. Con el fin de descartar que esto pueda deberse a problemas en la anotación del codón de iniciación, se realizó el alineamiento de los genes ortólogos, obtenidos por best reciprocal hit para las especies *T.brucei, T. vivax* y *T.congolense*.

En varios casos se detectaron problemas en la anotación de los genes, tanto de *T. brucei* como de *T. vivax*, donde los ortólogos comienzan a alinear bien a partir de un codón de inicio más interno que el anotado.

En muchos de estos casos si considero este segundo AUG como sitio de inicio de la transcripción, entonces muchos de los sitios de trans-splicing previamente asignados a regiones internas de los genes pasan a encontrase en regiones externas (Figura 13).

Esto también afecta la conservación de la distancia del sitio de splicing, como se muestra en los ejemplos de la tabla siguiente donde un supuesto sitio de splicing interno tras ser corregido pasa a ser externo y además su nueva posición muestra conservación de la distancia, aunque no así de la secuencia.

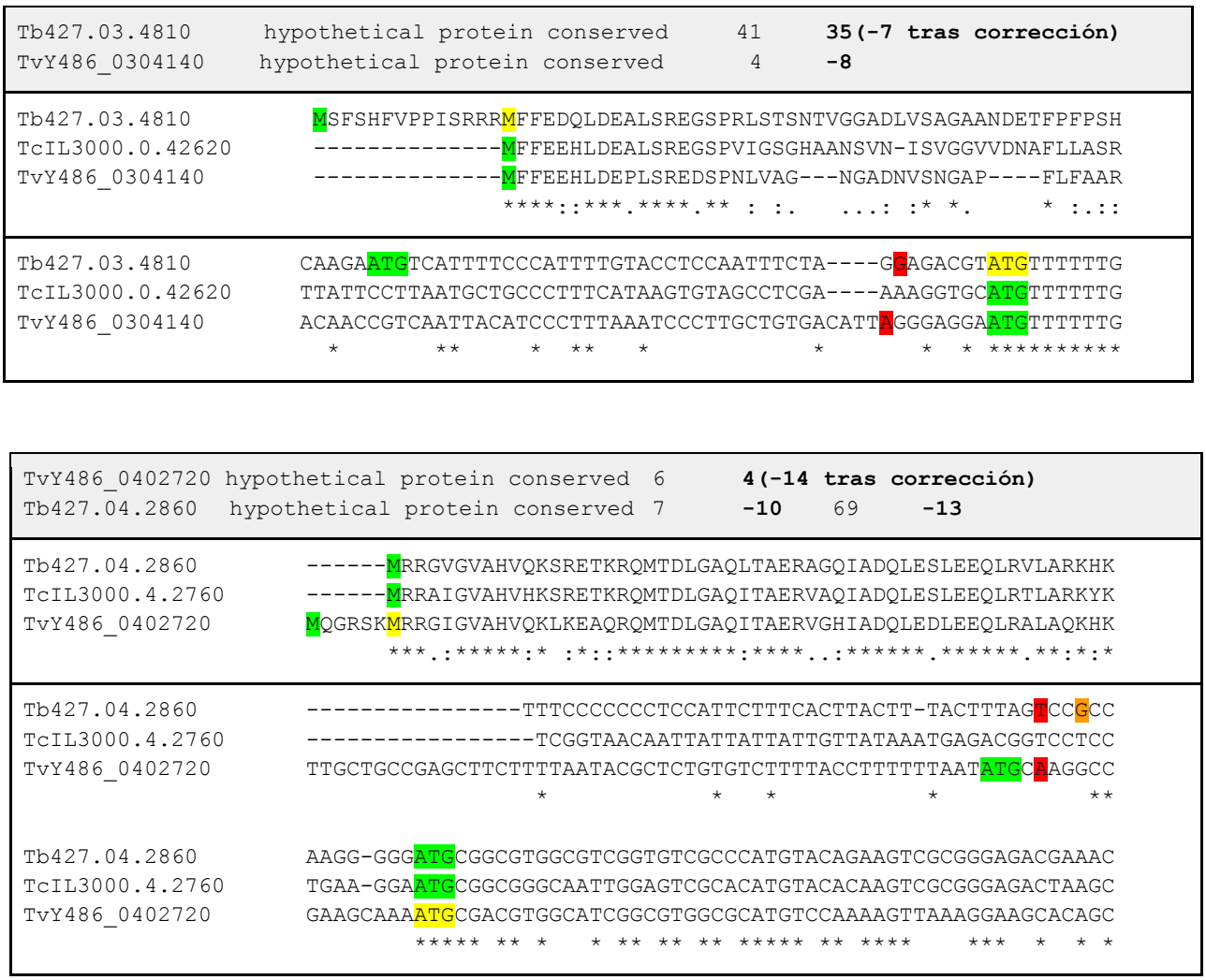


Figura 13: Ejemplos de genes mal anotados en *T. brucei* y *T.vivax*. En verde los sitios de inicio del gen, en amarillo el sitio de inicio candidato, en rojo el punto de splicing mayoritario y en naranja el secundario. Al lado de la descripción del gen se indica el número de reads que mapean, y en negrita la posición del sitio de splicing.

Reconocimiento del sitio de trans-splicing

Con el reciente secuenciado del transcriptoma de tres estadios de *T. cruzi* en el Institut Pasteur de Montevideo, se aprovechó esta información para estudiar eventos de trans-splicing de SL también en este organismo.
Sin embargo debido a la estrategia usada para realizar el secuenciado de RNA (full RNA stranded, con pérdida de parte 5' de ARN), se obtuvieron muy pocos reads en los que se pudiera afirmar con valores confiables que los mismos corresponden al sitio de trans-splicing (ver más adelante)como para realizar un estudio más a fondo de los eventos de trans-splicing.
Otro de los posibles problemas es que se utilizó la cepa JR de *T. cruzi* cuya anotación es incompleta. Lo que implicó que tuviéramos que realizar una anotación de la misma para poder llevar adelante este análisis (dicha anotación es parte de otro trabajo, ver más abajo).

```
>tcruziJRcl4_Contig5_961
CTTCTCCAATTACACACCCCGGTCAAAACAACAACAACAACAAAAAAGGAAAATAATTATCTGTATTTGGCTCTTATGCTTGGGAT
>HWUSI-EAS1675R:7:651h8aaxx:3:32:18808:12304
------------------TATTATTGATACAGTTTCTGTACTATATTGGAAAATAATTATCTGTATTTGGCTCTTATGCTTG
>Tcruzi_SL
---------AACTAACGCTATTATTGATACAGTTTCTGTACTATATTG
```

Figura 14: Alineamiento de la región de trans-splicing centrada en AG, el read que alinea con el inicio del gen y la secuencia del SL. En rojo los nucleótidos del SL y del read que alinean y en azul los del gen y el read. En fondo amarillo se indica la zona donde ocurrió el trans-splicing.

El procedimiento para obtener las posiciones del sitio de trans-splicing es similar a lo realizado en *T. brucei* y *T.vivax* pero al obtener un número tan bajo de reads no es necesario agrupar los reads en bloques dependiendo de cómo mapean. Tras alinear los reads con el SL se recorta de los reads hasta la posición final de mapeo con el SL y la secuencia restante, en azul en el read del ejemplo anterior (Fig. 14), se mapea contra los genes extraídos de los contigs con un margen de unos 300 nucleótidos corriente arriba del sitio de inicio del gen.
Tras aplicar ciertos filtros para asegurar la calidad de los alineamientos se obtuvieron unos 218 hits correspondientes a 218 reads que mapean en 214 genes.
Sin embargo esta información es suficiente como para realizar un logo de secuencias en la zona del trans-splicing de SL.
Un logo de secuencias es una representación gráfica de la conservación de la secuencia de nucleótidos o aminoácidos. Es creado a partir de un conjunto de secuencias alineadas y muestra el grado de consenso y la diversidad de las secuencias. Diferentes residuos en la misma posición son escalados de acuerdo a su frecuencia.
El logo a continuación (Figura 15) fue realizado utilizando la cepa JR de *T. cruzi* tomando 50 nucleótidos antes del sitio de trans-splicing y los 15 siguientes; cabe destacar la presencia resaltada del dinucleótido AG en las posiciones -2 y -1, además del tracto de polipirimidinas (múltiples Ts) que le precede.

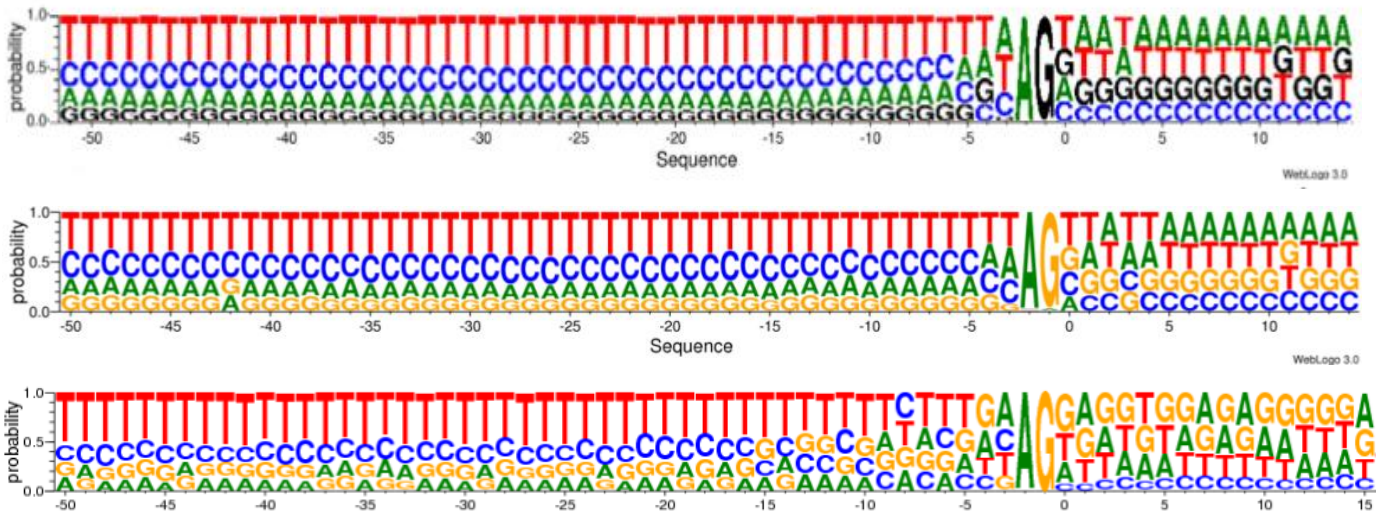


Figura 15: Logo de secuencias de 50 nucleótidos antes del sitio de trans-splicing y 15 después en *T. brucei*, (Kolev et al. 2005) *T. vivax*, (Greif et al. in press) y *T. cruzi*. Las posiciones de los nucleótidos son relativas al sitio de trans-splicing.

Estos resultados son consecuentes con los obtenidos previamente en *T. vivax* y en *T. brucei* en donde el dinucleótido AG es el punto clave de la secuencia para que ocurra el trans-splicing, aunque la secuencia que sigue a continuación de este punto no muestra una similitud con las encontradas en los otros tripanosomas en donde se encuentra un tracto polyA. (Siegel 2005)

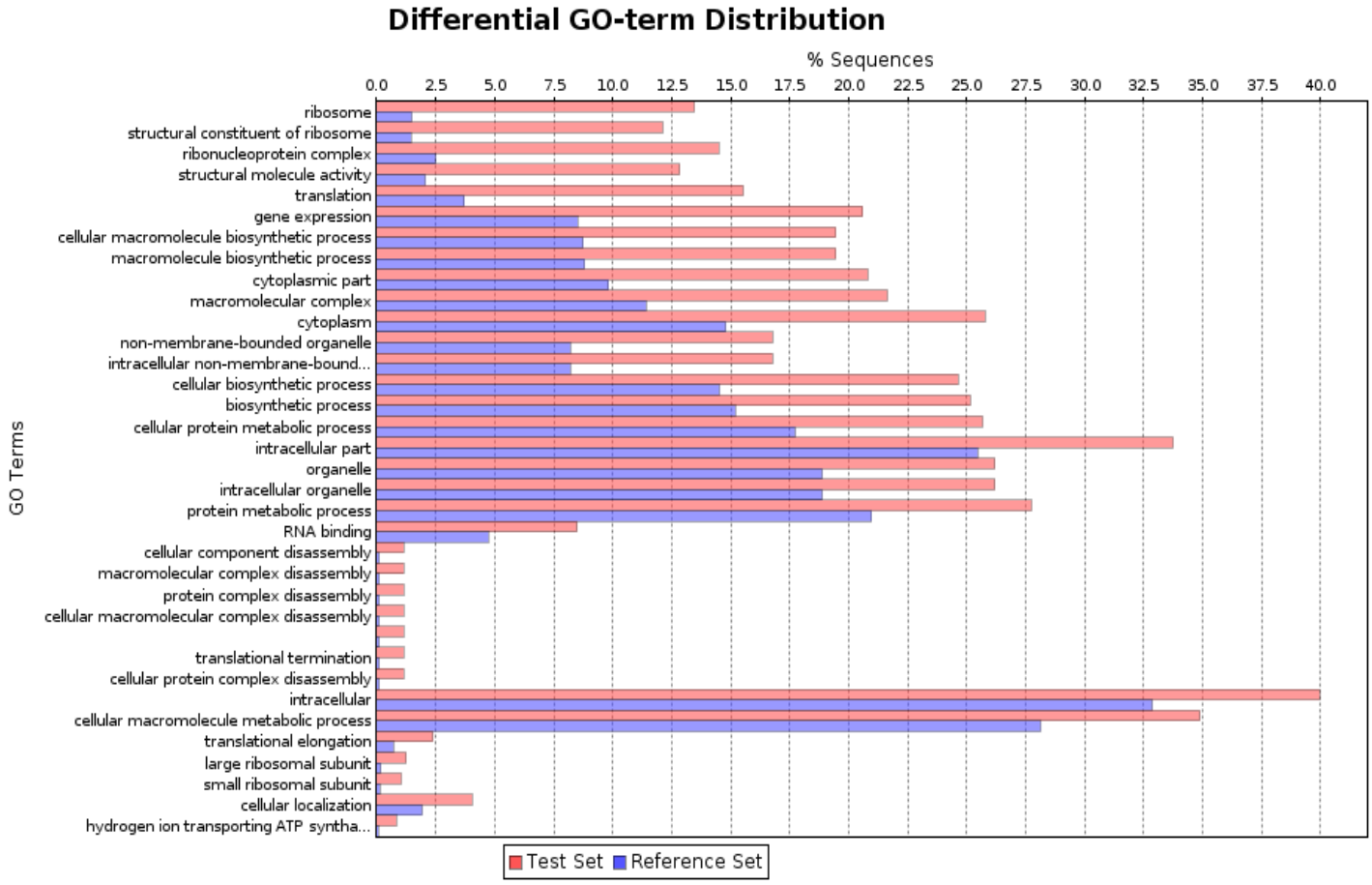
Análisis de términos de Gene Ontology utilizando Blast2GO

Blast2GO es una herramienta bioinformática para la anotación y análisis de secuencias génicas o de proteínas. Permite mapear secuencias con BLAST contra secuencias ingresadas en el GenBank del NCBI, obtener una descripción del mejor hit de la secuencia mapeada, y obtener los términos GO , EC e InterPro asociados, utilizando las herramientas propias de cada base de datos.
El proyecto Gene Ontology (GO) busca estandarizar la representación de genes y los atributos de los productos génicos entre especies y bases de datos; con un glosario de términos para describir las características de los productos génicos y datos de anotación; utiliza tres tipos de datos, información de Componente Celular, Función Molecular y Proceso Biológico asociado.
Además Blast2GO realiza búsquedas en la base de datos de Enzyme Comission, una clasificación de enzimas basada en el tipo de reacción química que catalizan y términos InterPro su paquete de buscadores asociados que permite buscar patrones o motivos proteicos; una colección de base de datos de familias de proteínas, dominios y sitios funcionales como ProDom, PRINTS, PIR-PSD, Pfam, SMART, TIGR, PROSITE, PANTHER, SUPERFAMILY, Gene3D, en donde se buscan características identificables encontradas en proteínas conocidas. (Quevillon et al. 2005)
Además otros programas integrados al paquete InterProScan (Hunter et al 2011) permiten permiten hacer predicciones de novo de motivos tales como péptidos señal con el software SignalP (Nordahl et al 2011) y dominios trans-membrana con TM-HMM (Krogh et al 2001).
El primer paso en la utilización de Blast2GO es mapear los genes contra la base de datos de proteínas no redundantes (nr) del NCBI.
Luego se continua con la obtención de los términos de ontología génica (GO) asociados a los hits obtenidos de BLAST, haciendo diferentes mapeos; ya sea directamente sobre la base de datos de GO u

obteniendo información adicional complementaria de bases de datos del NCBI o UniProt.
A continuación se selecciona del pool de términos GO obtenidos aplicando ciertas reglas de anotación en la ontología de términos encontrados en donde se pueden utilizar diferentes criterios más o menos estrictos.
Con este procedimiento se puede asociar un grupo de genes de acuerdo a sus términos de ontología génica asociados y así poder identificar patrones fácilmente.
Se buscó comprobar si existía enriquecimiento de un tipo en particular de genes (categoría de GO) en relación a las distancias de SL respecto al sitio de inicio del gen; los genes fueron agrupados según esta distancia y dichos grupos comparados con la totalidad de los genes que mapearon con secuencias de SL.
Para validar si el grupo de genes de interés tenía una diferencia estadísticamente significativa con el grupo de referencia se utilizó el test exacto de Fisher provisto por el programa.
El test exacto de Fisher es un test estadístico para tablas de contingencia, utilizado para determinar si las proporciones de los datos obtenidos poseen diferencias significativas en dos grupos diferentes.
La hipótesis nula tiene lugar cuando las proporciones relativas de una variable son independientes de una segunda variable.
En este caso interesa determinar si las proporciones de los términos de ontología génica difieren significativamente en el conjunto de genes de interés respecto a la totalidad de los genes. La hipótesis nula de este test seria la no existencia de una diferencia en el grupo de genes considerados en ambos grupos.
El problema de testear simultáneamente múltiples hipótesis radica en que a mayor número de tests, aumenta la proporción de la posibilidad de rechazar erróneamente las hipótesis nulas descartadas tras el cálculo del valor p. Por esto es que es necesario corregir el valor de p para tests múltiples.
El control de FDR (False Discovery Rate) es un método utilizado en estos casos para corregir los valores p de las estimaciones para comparaciones múltiples. FDR controla la proporción de falsos positivos entre todas las hipótesis significativas. Por ejemplo si se predice que 200 observaciones serán falsos positivos, y el FDR máximo para estas es de 0.2, entonces 20 de ellas se esperan que sean falsos positivos. El valor de q de FDR es análogo al valor p, es decir para cada hipótesis el valor de q es el mínimo FDR para el cual el test puede ser considerado significativo.
Para encontrar aquellos términos GO enriquecidos en un conjunto de genes en particular, se realizó un análisis de enriquecimiento utilizando el test exacto de Fisher de Blast2GO, con un valor de filtro de hasta 0,05 bajo el modo de filtrado FDR 'False DiscoveryRate'.

Resultados de análisis de Blast2GO

Se realizó un estudio de enriquecimiento en términos de ontología génica de los genes que contienen sitios aceptores de trans-splicing en la proximidad del AUG, este estudio se llevó a cabo tanto *T. brucei* como en *T. vivax*.
A continuación se muestran los resultados de estos análisis para *T.brucei* y *T.vivax* (Figuras 16 y 17):



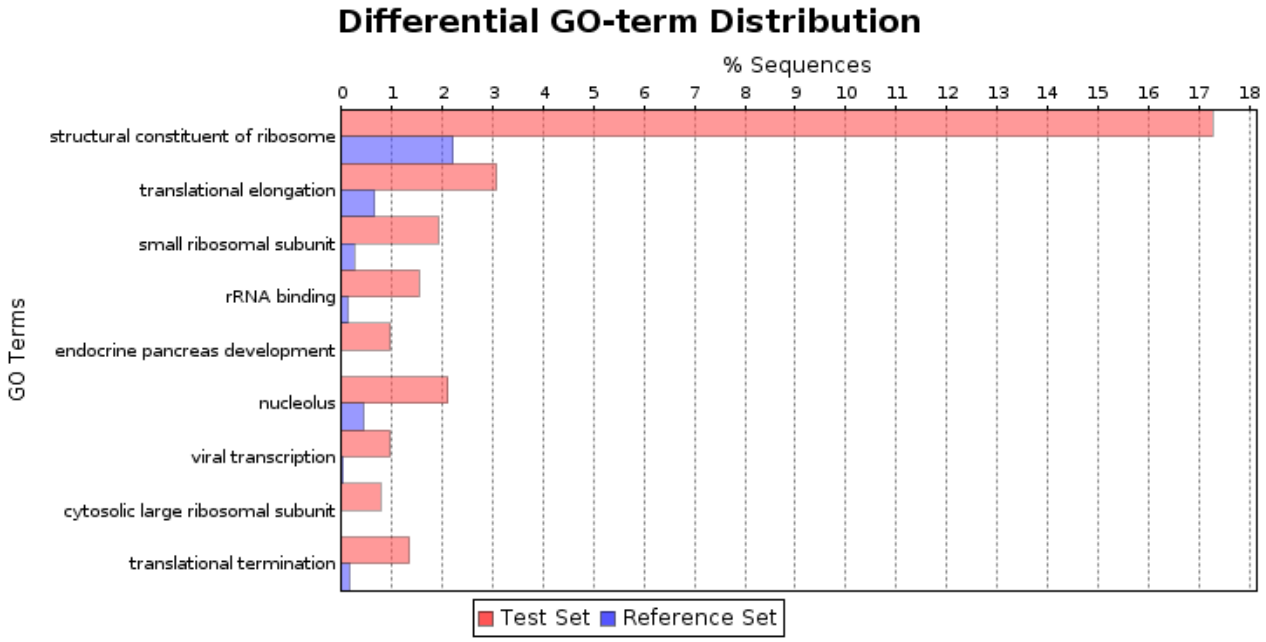


Figura 16: Gráfica de los resultados del análisis de enriquecimiento de términos de ontología génica utilizando el test exacto de Fisher, para los genes de *T. brucei* con spliced leader entre -1 y -35 nucleótidos antes del inicio del gen. En la gráfica superior se utilizaron todos los términos de ontología génica y en la inferior tras realizar una agrupamiento de términos y filtrado provisto por el programa.

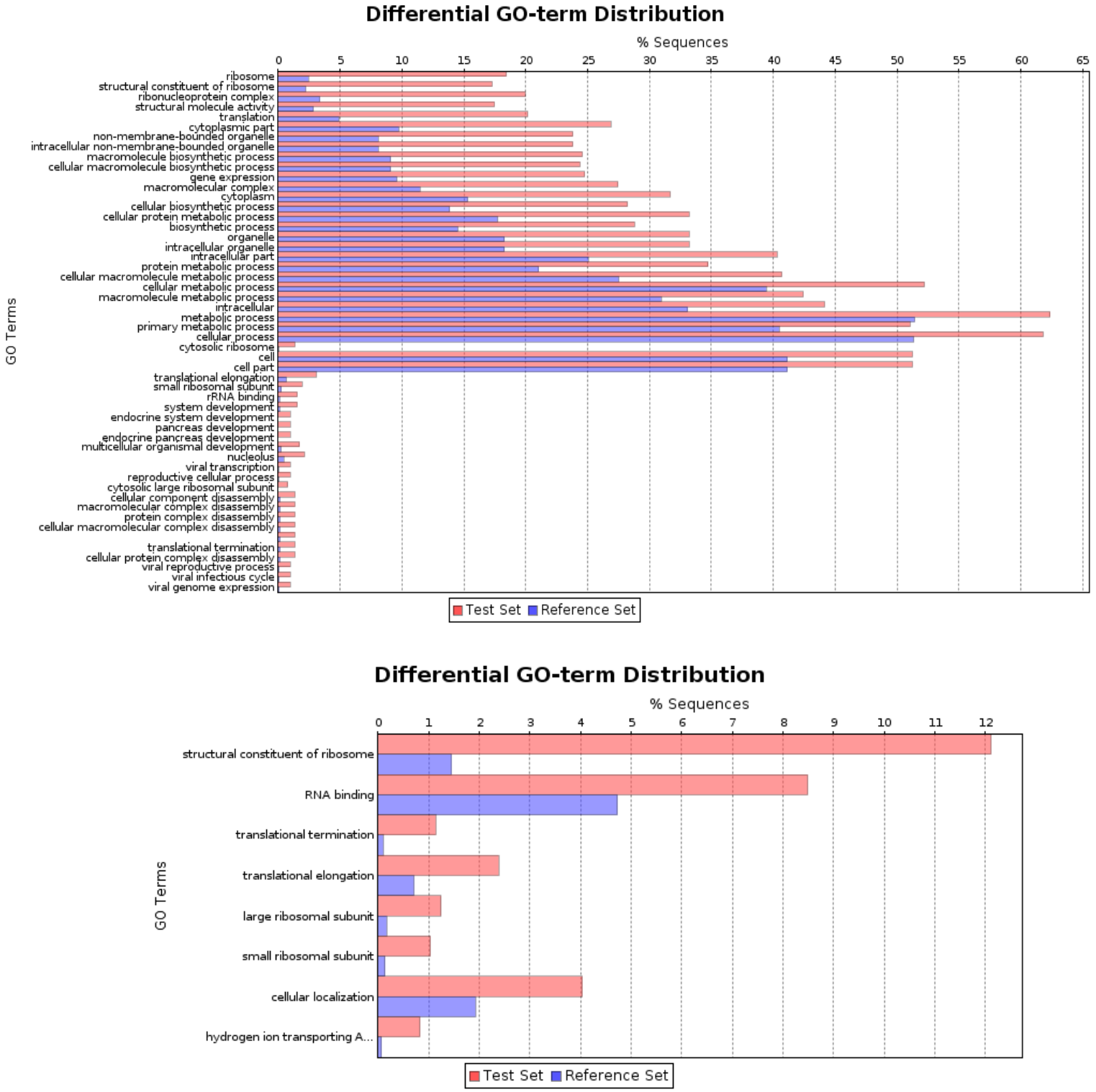


Figura 17: Gráfica de los resultados del análisis de enriquecimiento de Ontología de Genes para los genes de *T. vivax* con spliced leader entre -1 y -35 nucleótidos antes del sitio de inicio del gen. En la gráfica superior se utilizaron todos los términos de ontología génica y en la inferior tras realizar una agrupamiento de términos y filtrado provisto por el programa.

Los resultados del análisis de enriquecimiento con términos de ontología génica para esta región de unos 35 nucleótidos antes del inicio de la parte codificante del gen (AUG) muestran una clara sobreabundancia de genes pertenecientes a componentes estructurales de los ribosomas, factores de elongación, pequeñas subunidades ribosomales y proteínas de interacción con el rRNA; en general diferentes grupos de genes asociados con el proceso de la traducción.
Además de considerar está región en su totalidad; se tomaron regiones cada vez menores en distancia respecto al sitio de inicio del gen y las regiones inmediatamente internas.

Distancias	**T. brucei**		**T. vivax**	
	Genes	**SL**	**Genes**	**SL**
-35 a -1	1885	2954	1029	1205
-20 a -1	1196	1721	536	590
-10 a -1	653	814	190	204
-5 a -1	356	385	53	53
1 a 5	41	41	33	34
1 a 10	96	103	69	72
1 a 20	152	187	108	111
1 a 35	209	268	147	156

Tabla 6: Tabla con número de genes con spliced leader y cantidad de spliced leader por regiones, respecto a la distancia del sitio inicio del gen. Estos genes fueron luego utilizados para el análisis de ontología génica del Blast2GO.

Tomando distancias cada vez menores respecto al inicio del gen (Tabla 6), resulta en un aumento de la proporción de genes asociados con la traducción que se ven representados en el conjunto de datos.
Considerandos solamente los primeros 5 nucleótidos antes del inicio del gen, la proporción de genes asociados a la traducción en *T. brucei* aumenta considerablemente como se muestra a continuación (Figura 18).

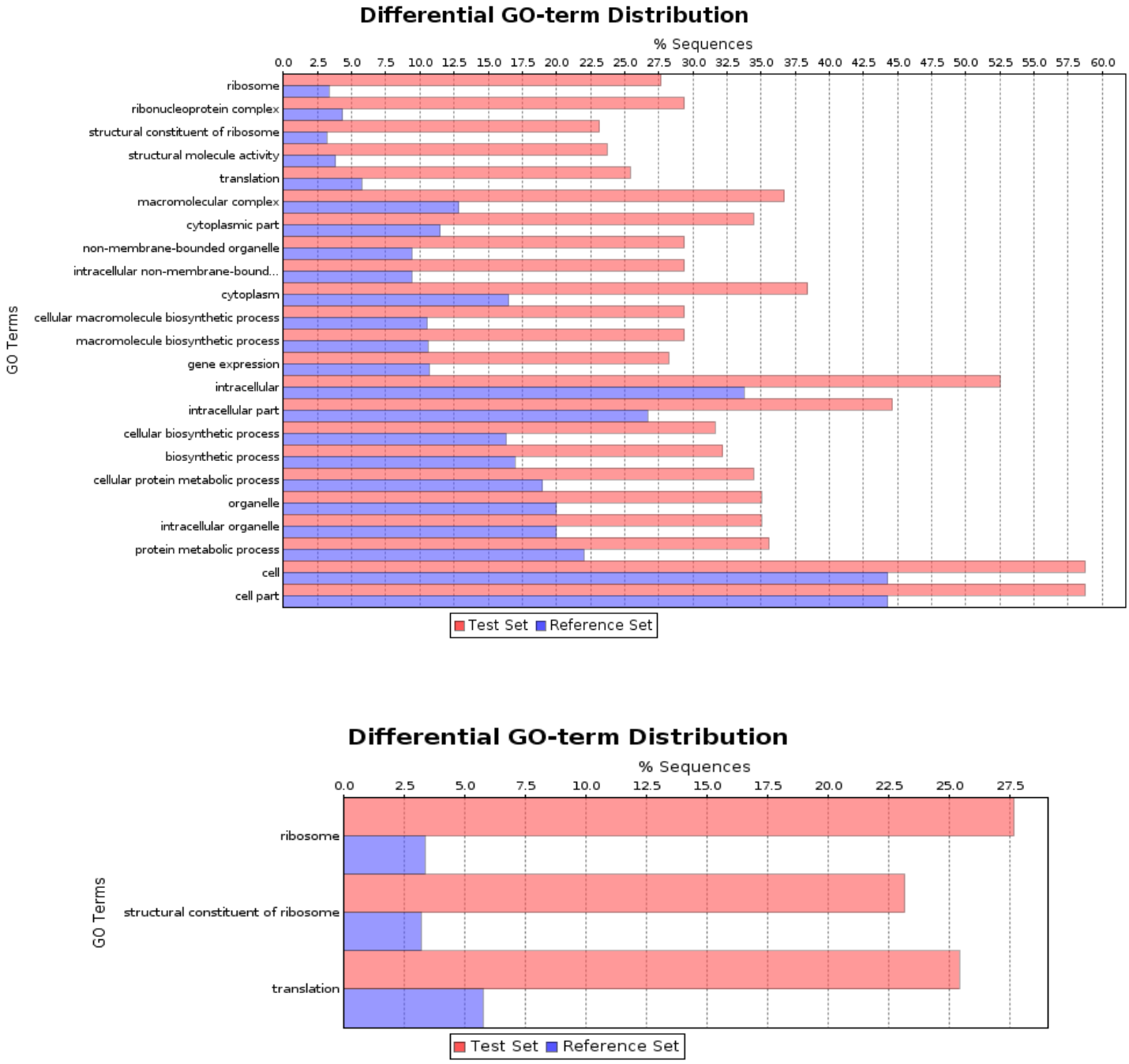


Figura 18: Gráfica de los resultados del análisis de enriquecimiento de ontología génica del Blast2GO para los genes de *T. brucei* con spliced leader entre -1 y -5 nucleótidos antes del inicio del gen. En la gráfica superior se utilizaron todos los términos de ontología génica y en la inferior tras realizar una agrupamiento de términos y filtrado provisto por el programa.

En el caso de *T. vivax* debido al menor número de SL identificados y en la región -1 a -5 solo se encuentran 53 genes, no se producen resultados estadísticamente significativos para esta distancia. Sin embargo en el caso de *T. brucei* aún considerando distancias menores, por ejemplo tomando solamente aquellos genes cuyo sitio de splicing coincide con el sitio de inicio de del gen, se encuentran 75 genes y un análisis de enriquecimiento en términos de ontología da resultados aún más sesgado hacia genes asociados a la traducción. (Figura 19)

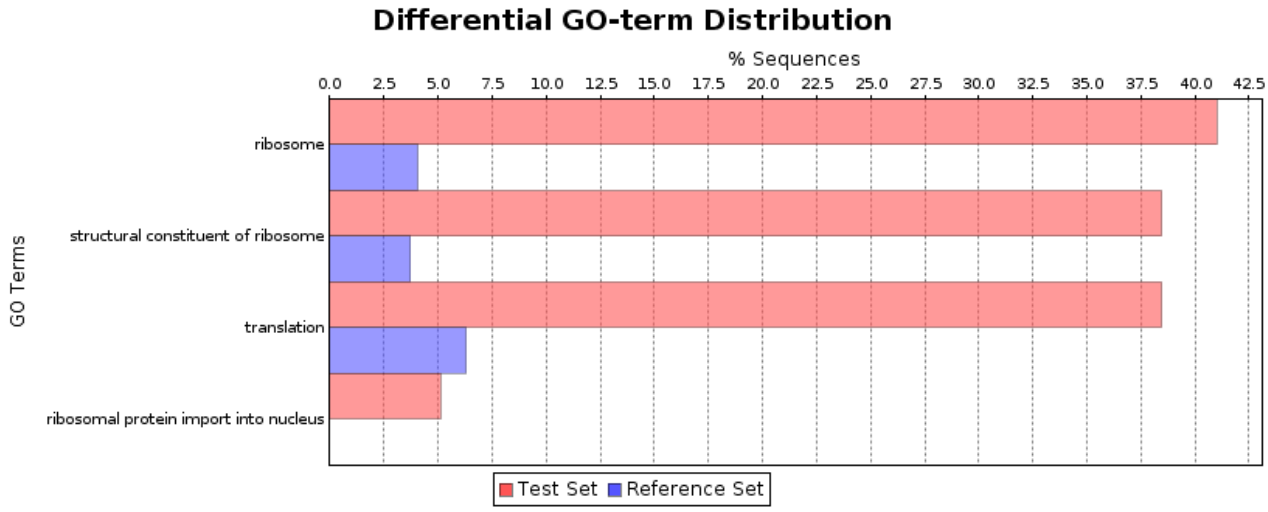


Figura 19: Gráfica de los resultados del análisis de enriquecimiento de ontología génica del Blast2GO para los genes de *T. brucei* cuyo sitio de inserción del spliced leader coincide con el sitio de inicio del gen. En la gráfica se muestran los términos luego de realizar un agrupamiento y filtrado.

Cuando se consideran los genes con sitios de splicing en las regiones internas de los genes hasta los 35 nucleótidos los resultados obtenidos son similares, siempre hay una mayor proporción de genes ribosomales y genes asociados con la traducción de lo que se esperaría según el conjunto total de genes tomado como referencia. Es importante considerar que también parte de estos genes pueden tener problemas de anotación y que por lo tanto los sitios de splicing no caigan en regiones internas sino en realidad en sitios externos, cercanos al sitio de inicio.

Conclusiones

En el presente trabajo se estudiaron dos aspectos sobre los genomas de tripanosomas; por un lado el contenido de GC3 y utilización de codones sinónimos en relación a los niveles de expresión génica, y por otra parte las características de los eventos de trans-splicing de SL.
Los resultados sobre la relación entre GC3 y el nivel de expresión génica fueron compatibles con estudios previos. Concretamente en *T.brucei* y *T. vivax* se observa claramente que los genes mas expresados son aquellos con un contenido de GC3 mas elevado; sin embargo en *T. cruzi* esto no se observa claramente cuando se consideran todos lo genes. En cambio cuando se restringe a aquellos genes mas conservados entre las tres especies de tripanosomas, la relación entre GC3 elevado y nivel de expresión se hace más notable aún, e incluso esta diferencia puede observarse en *T. cruzi*.
Un aspecto interesante a estudiar es la relación entre la abundancia relativa de los tRNAs y la preferencia por el uso de codones, sin embargo algunas de las complicaciones radican en la propia estructura de los tRNAs y el uso de bases modificadas, lo cual dificulta su resolución por RNAseq.
Los eventos de trans-splicing de SL fueron estudiados en *T. brucei* y *T.vivax*. En la mayoría de los genes se identificaron sitios de trans-splicing al estudiar una región de 500 pb antes del AUG de inicio. En

algunos casos se detectaron sitios de trans-splicing internos a la región codificante; debido a ello se estudiaron los ortólogos entre *T.vivax*, *T.brucei* y *T.congolense*, encontrándose en varios casos errores en la anotación. Tras corregir estos errores muchos de los sitios de trans-splicing que previamente habían sido identificados como internos, pasan a ocupar región externas en el 5' UTR del gen.
Se estudió en *T.cruzi* la composición nucleotídica de la región donde tiene lugar el trans-splicing, observándose el mismo patrón conocido en *T.brucei* y *T.vivax*, donde la preferencia por el sitio de trans-splicing es en el dinucleótido AG precedido por una larga secuencia de polipirimidinas.
Además se consideraron los genes agrupados según la distancia del evento de trans-splicing al sitio AUG, y tras realizar un análisis de términos de ontología génica para estos genes agrupados, se observó que aquellos donde el trans-splicing tiene lugar a 35 pb o menos estan enriquecidos en genes relacionados al proceso traduccional, tal como genes ribosomales o constituyentes estructurales de los ribosomas. Cuando se consideran grupos de genes con trans-splicing más cercanos al sitio de inicio del gen, aumenta aún mas la proporción de genes ribosomales, respecto al total de genes.
La ausencia de 5’ UTR en los transcriptos de los genes relacionados con el aparato traduccional sugeriría que no podrían estar sujetos a un control post-transcripcional. La ausencia de un segmento entre el spliced leader y el codón de inicio, en donde eventualmente se podrían unir elementos reguladores o formar estructuras secundarias del RNA como stem-loops, en cierta forma similares a los termómetros de RNA bacterianos, implicaría que una vez que el complejo de iniciación ribosomal se ha ensamblado no hay posibilidades de que se bloquee el inicio de la traducción.
Recientemente se ha propuesto que los tripanosomas podrían contener elementos reguladores en cis sobre el RNA, ubicados en el 5' UTR, lo cual podría ser parte de un mecanismo para sensar cambios ambientales como la temperatura. (Kramer 2011)

Referencias

Aird, D. et al., 2011. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biology*, 12(2)

Alsford, S. et al., 2011. High-throughput phenotyping using parallel sequencing of RNA interference targets in the African trypanosome. *Genome research*, 21(6), pp.915-24.

Alvarez, F., Robello, C. & Vignali, M., 1994.Evolution of codon usage and base contents in kinetoplastid protozoans. *Molecular biology and evolution*, 11(5), pp.790-802.

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

Bulmer, M., 1991.The selection-mutation-drift theory of synonymous codon usage. *Genetics*, 129(3), pp.897-907.

Chen, S.L. et al., 2004. Codon usage between genomes is constrained by genome-wide mutational processes. *Proceedings of the National Academy of Sciences of the United States of America*, 101(10), pp.3480-5.

Dohm, J.C. et al., 2008. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Research*, 36(16), p.e105.

Douris, V., Telford, M.J. & Averof, M., 2010.Evidence for multiple independent origins of trans-splicing in Metazoa. *Molecular biology and evolution*, 27(3), pp.684-93.

Duret, L., 2002. Evolution of synonymous codon usage in metazoans. *Current Opinion in Genetics & Development*, 12(6), pp.640-9.

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

Greif, G. et al. Transcriptome analysis of the bloodstream stage from the parasite Trypanosoma vivax (in press)

Hansen, K.D. et al., 2011. Sequencing technology does not eliminate biological variability. *Nature Biotechnology*, 29(7), pp.572-3.

Hastings, K.E.M., 2005. SL trans-splicing: easy come or easy go? *Trends in genetics* 21(4), pp.240-7.

Hunter et al., 2011. InterPro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Research* doi: 10.1093/nar/gkr948

Kahn, C., & Scott, L., 2005. The Merck Veterinary Manual .9th edition. Merck.

Kolev, N.G. et al., 2010. The transcriptome of the human pathogen Trypanosoma brucei at single-nucleotide resolution. *PLoS pathogens*, 6(9), e1001090.

Kramer, S., 2012. Developmental regulation of gene expression in the absence of transcriptional control: the case of kinetoplastids. *Mol Biochem Parasitol*, 181(2):61-72

Krogh, A et al., 2001. Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. Journal of Molecular Biology, 305(3):567-580

Liang, X. et al 2003. trans and cis splicing in Trypanosomatids: mechanisms, factors and regulation. Eukaryotic Cell 2(5):830

Manning-Cela, R; Gonzales, A & Swindle, J. 2002.Alternative splicing of LYT1 transcripts in Trypanosoma cruzi. *Infection and Immunity*, 70(8), pp.4726-4728.

Mayer, M.G. & Floeter-Winter, L.M., 2005. Pre-mRNA trans-splicing: from kinetoplastids to mammals, an easy language for life diversity. *Memórias do Instituto Oswaldo Cruz*, 100(5), pp.501-13.

Morozova, O., Hirst, M. & Marra, M.A., 2009. Applications of new sequencing technologies for transcriptome analysis. Annual Review of Genomics and Human Genetics, 10, pp.135–51.

Mortazavi, A., et al., 2008.Mapping and quantifying mammalian transcriptomes by RNA-Seq.*Nature Methods,* 5(7), pp.621-628.

Nordahl, P. et al., 2011. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nature Methods*, 8, pp785-786

Osório, A.L. et al., 2008. Trypanosoma (Duttonella) vivax: its biology, epidemiology, pathogenesis, and introduction in the New World - a review. *Memórias do Instituto Oswaldo Cruz*, 103(1), pp.1–13.

Palenchar, J. & Bellofatto, V. 2006. Gene transcription in trypanosomes. *Molecular & Biochemical Parasitology*, 146, pp.135-141.

Plotkin, J.B. & Kudla, G., 2011. Synonymous but not the same: the causes and consequences of codon bias. *Nature Reviews. Genetics*, 12(1), pp.32-42.

Quevillon E., et al 2005. InterProScan: protein domains identifier. *Nucleic Acids Research.* 33 (Web Server issue): W116-W120

Roberts, A. et al., 2011. Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome Biology*, 12(3), p.R22.

Schimanski, B. et al., 2004. The Trypanosoma brucei spliced leader RNA and rRNA gene promoters have interchangeable TbSNAP50-binding elements. *Nucleic acids research*, 32(2), pp.700-9.

Schwartz, S., Oren, R. & Ast, G., 2011. Detection and removal of biases in the analysis of next-generation sequencing reads. *PloS One*, 6(1), p.e16685.

Sharp, P.M. et al., 1995. DNA sequence evolution: the sounds of silence. *Philosophical Transactions of the Royal Society of London.Series B, Biological sciences*, 349(1329), pp.241-7.

Siegel, N et al. 2005. Systematic Study of Sequence Motifs for RNA trans Splicing in Trypanosoma brucei. M*olecular and Cell Biology* 25(21), pp.9586-9594

Trapnell, C. et al., 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols*, 7(3), pp.562-578.

Wang, Z., Gerstein, M. & Snyder, M., 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews. Genetics*, 10(1), pp.57–63.