Tesis de Doctorado en Ciencias Biológicas
Pedeciba

Genómica funcional y evolutiva de tripanosomas africanos: el modelo *Trypanosoma vivax*

Gonzalo Greif
Unidad de Biología Molecular, Institut Pasteur Montevideo
Orientadores: Fernando Alvarez-Valín y Carlos Robello
2015

Indice

Resumen.

Los tripanosomas africanos presentan características particulares de gran interés, entre las que resaltan un sistema de evasión de la respuesta inmune único en la naturaleza, así como interesantes mecanismos de regulación de la expresión génica. Asimismo debido a sus complejos ciclos de vida, en los que alternan entre dos hospederos, estos parásitos presentan grandes cambios a nivel metabólico, particularmente en la actividad mitocondrial.

Con el fin de entender el surgimiento y la evolución de estas peculiaridades es que decidimos trabajar con *Trypanosoma vivax,* una especie de gran potencial como modelo de estudio de los tripanosomas africanos, debido a su ubicación evolutiva en la rama más ancestral de estos parásitos. Por otra parte, su introducción en América, donde logró expandirse sin la presencia del insecto vector (independiente de la transmisión cíclica) y usando otros vectores que actúan meramente como agentes mecánicos, resulta de gran importancia para investigar los procesos adaptativos y evolutivos que tuvieron como respuesta a este cambio del modo de transmisión.

En esta tesis, se presentan los resultados obtenidos mediante análisis de RNA-seq en la etapa sanguínea de aislados americanos de *T. vivax*. Los datos de la secuenciación, ensamblaje de transcriptos, anotación funcional y otras características derivadas de este análisis se encuentran disponibles a través de una base de datos disponible online (bioinformática.fcien.edu.uy/Tvivax). Estos datos permitieron por un lado el estudio de las glicoproteínas variables de superficie (VSGs) asociadas con el mecanismo de evasión de la respuesta inmune. Los resultados de la comparación de la composición de membrana de estos parásitos contra *Trypanosoma brucei*, muestran diferencias notables a nivel de la expresión de ARN mensajeros asociados a proteínas de membrana, dejando abierta la pregunta respecto al rol protector de estas proteínas en *T. vivax* y por ende su rol ancestral en la evasión de la respuesta inmune.

Respecto a los patrones de expresión génica, mostramos datos relacionados al uso diferencial de sitios de *trans-splicing*, el que proponemos como posible mecanismo alternativo de regulación de la expresión génica.

Finalmente, respecto al estudio de la adaptación a la transmisión mecánica y su relación con la actividad mitocondrial, en este trabajo realizamos un estudio comparativo del genoma mitocondrial en tres cepas de *Trypanosoma vivax* (una de origen africano que realiza el ciclo de vida completo en el insecto vector, y dos cepas americanas que son mecánicamente transmitidas y permanecen únicamente como formas sanguíneas). La cepa africana presenta un genoma mitocondrial completo y

totalmente funcional, mientras que las cepas americanas muestran el inicio de un proceso de degradación de su genoma mitocondrial. Por otra parte el *editing* postranscripcional (necesario para obtener proteínas funcionales a partir de los genes mitocondriales) ocurre únicamente en dos genes en las cepas americanas, mientras que en la cepa africana ocurre normalmente en todos los genes que lo requieren. Los genes que sí son editados en las cepas americanas (A6-ATPase y RPS12) juegan un rol esencial también durante la etapa sanguínea de estos parásitos. El análisis de la población de minicírculos (necesarios para el *editing*) muestra una diversidad reducida en las cepas americanas, y se encuentran principalmente aquellos que contienen los ARNg necesarios para el *editing* de los dos genes correctamente editados. El hecho de que estos dos genes permanezcan completamente funcionales difiere a lo reportado en las cepas *Trypanosoma brucei*-like: *T. evansi* y *T. equiperdum* que restringen su ciclo de vida a la fase sanguínea. Esto junto con otras diferencias, es indicativo que las cepas americanas de *T. vivax* están siguiendo un camino evolutivo, en su adaptación a la transmisión mecánica, diferente al que ocurrió en las cepas derivadas de *T. brucei*. Cabe resaltar además, que estos acontecimientos ocurrieron en un período de tiempo relativamente corto, si consideramos que el ingreso de estos parásitos en América no lleva más de 500 años.

En suma, en esta tesis se presentan los resultados obtenidos mediante el análisis del parásito *Trypanosoma vivax* respecto a dos temas fundamentales de su biología. Por un lado, se aportan elementos para elucidar el mecanismo de variación antigénica en su estado ancestral, aportando información sobre la evolución de este sistema, así como otros interesantes aspectos relacionados con posibles mecanismos de regulación de la expresión génica. Por otra parte se presentan los resultados de la comparación de los genomas mitocondriales de aislados americanos y africanos de estos parásitos, mostrando evidencia sobre cambios en el genoma mitocondrial que habrían tenido lugar durante la adaptación a la pérdida del ciclo completo observado en los parásitos aislados en América.

Introducción.

a. Tripanosomátidos

a.1 Generalidades

El orden kinetoplastida está constituido por un grupo de protozoarios que incluye organismos de vida libre, y parásitos de invertebrados, vertebrados y también de plantas. Constituye una rama temprana en el árbol evolutivo eucariota y posee características morfológicas, bioquímicas y genéticas que son únicas de este grupo. El orden debe su nombre a la presencia de una intrincada estructura formada por ADN y proteínas de su única mitocondria, localizada en una región especializada de la matriz mitocondrial, llamado kinetoplasto [1].

Dentro de este grupo, la familia Trypanosomatidae (organismos exclusivamente parásitos) es causante de importantes enfermedades en humanos (leishmaniasis –*Leishmania spp*-, enfermedad del sueño –*Trypanosoma brucei*- y enfermedad de Chagas –*Trypanosoma cruzi*-), así como también en animales silvestres y domesticados (provocada por *T. brucei brucei*, *Trypanosoma vivax, Trypanosoma evansi y Trypanosoma congonlense*).

La enfermedad de Chagas o tripanosomiasis americana causada por *T. cruzi*, fue descrita por el médico brasileño Carlos Chagas en 1909 [2], siendo la primera vez en la historia que una misma persona describe la enfermedad, el agente causante y el vector [2, 3]. Esta enfermedad es una de las endemias más expandidas de América Latina, aunque en las últimas décadas se ha observado con mayor frecuencia en los Estados Unidos de América, Canadá, muchos países europeos y algunos países del Pacífico Occidental. Se estima que en el mundo hay entre 7 y 8 millones de personas infectadas por este parásito [4].

La leishmaniasis, causada por *Leishmania spp* comprende un amplio espectro de enfermedades, tanto en humanos como en animales cuyas manifestaciones clínicas, dependiendo de la especie de *Leishmania* abarcan desde úlceras de piel hasta infecciones letales de órganos internos. De amplia distribución en América, Europa y Asia, es considerada una de las más importantes enfermedades desatendidas, causando aproximadamente 150.000 muertes anuales [5]. Se estima que en el mundo hay aproximadamente 2 millones de casos anuales con 367 millones de personas con riesgo de contraer la enfermedad [6].

Los tripanosomas africanos, también llamados Salivaria (debido a que completan su ciclo de vida en las glándulas salivales o en las piezas bucales del insecto vector) son causantes de una serie de enfermedades de importancia sanitaria y veterinaria conocidas como tripanosomiasis africanas. Esta enfermedad en humanos es llamada enfermedad del sueño y afecta a 36 países de la región sub-sahariana donde existen las moscas del género Glossina (mosca *tse-tsé*) que pueden trasmitirla. Dos subespecies de *T. brucei*, *T. brucei gambiense* y *T. brucei rhodesiense*, dan cuenta de la enfermedad en humanos, mientras que una tercer subespecie *T. brucei brucei* sólo infecta animales. *T. b. gambiense* es causante de la forma crónica de la enfermedad del sueño en África central y oriental, responsable del 98% de los casos notificados, mientras que *T. b. rhodesiense* da lugar a la forma aguda de la enfermedad en África del sur y occidental. La enfermedad tiene dos etapas, la primera conocida como la fase hemolinfática donde los parásitos se restringen a la sangre y al sistema linfático, mientras que la segunda corresponde a una etapa posterior de la enfermedad conocida como la fase neurológica, caracterizada por la presencia del parásito en el líquido cerebroespinal. Si no es tratada, la enfermedad del sueño causa la muerte en meses (*T. b. rhodesiense*) o años (*T. b. gambiense*) [7]. En 2004, se estimaba en 50.000 a 70.000 el número de personas infectadas anualmente, y desde entonces el número ha ido disminuyendo, reportándose 9.878 casos en 2009 y 3.679 nuevos casos en 2014. Según estimaciones de la Organización Mundial de la Salud, el número real de casos es actualmente de 30.000. Se estima que la población expuesta al riesgo de esta enfermedad es de unos 70 millones de personas [4].

En animales, una de las tripanosomiasis africanas más conocidas es la enfermedad llamada Nagana, ocasionada por *T. vivax*, *T. congolense* y *T. brucei*. Esta enfermedad tiene un fuerte impacto en la economía de los países que la padecen debido a que produce pérdidas importantes en la ganadería. Se estima, que sólo en la región de África sub-sahariana habría pérdidas económicas cercanas a los 5 mil millones de dólares anuales por esta enfermedad [8]. A su vez el ganado y los animales domésticos son el principal reservorio de los tripanosomas causantes de la enfermedad en humanos. Si bien los parásitos causantes de la Nagana son de origen africano y se trasmiten por la mosca *tse-tsé*, algunos se han dispersado a zonas libres de este insecto vector. Particularmente, en Sudamérica dos especies de tripanosomas africanos (*T. vivax* y *T. evansi*) que pudieron haber sido introducidos a mediados del siglo diecinueve en la Guayana francesa a través de ganado tipo cebú importado desde Senegal según algunos autores [9, 10], y se expandieron rápidamente por transmisión mecánica a través de vectores de varias especies de moscas hematófagas (Tabánidos y Stomoxys). La introducción no puede datarse exactamente, ya que también podría haber ocurrido desde 1733 directamente desde Africa [11], o desde los viajes de la colonización de América que trajeron ganado desde Europa o Africa [12].

La introducción de vacas, cabras, ovejas y equinos por los colonizadores pudo haber sido responsable de la introducción en América en diferentes lugares y momentos [12]. Hoy afectan áreas dedicadas a la ganadería extensiva, desde Venezuela hasta Paraguay y el norte de Argentina [13]. Se prevé que esta enfermedad siga expandiéndose hacia zonas más sureñas siendo el nuestro uno de los países que probablemente se vea afectado en el futuro próximo. En este sentido cabe resaltar que tanto *Trypanosoma vivax* como *Trypanosoma evansi* han sido detectados incluso en ganado en Rio Grande do Sul, a sólo pocos kilómetros de la frontera con nuestro país [14, 15]. Esto ubicaría a la tripanosomiasis africana en la categoría de parasitosis emergente en la región.

a.2 Ciclo de vida

Los tripanosomátidos presentan complejos ciclos de vida con diferentes etapas de desarrollo que alternan, en general, entre hospederos vertebrados e invertebrados. Durante sus ciclos de vida sufren importantes cambios tanto morfológicos como metabólicos y muestran gran plasticidad para adaptarse a entornos muy diferentes que requieren un alto grado de regulación diferencial en la expresión de proteínas [16].

En la Figura 1 se esquematiza las diferentes formas por las que alterna el parásito durante su ciclo de vida. Las mismas se clasifican en función de la ubicación del flagelo (anterior, media o posterior), y la ubicación del kinetoplasto con respecto al núcleo. No todas las especies pasan por todas las formas, ni éstas son totalmente equivalentes, biológicamente, entre distintas especies.

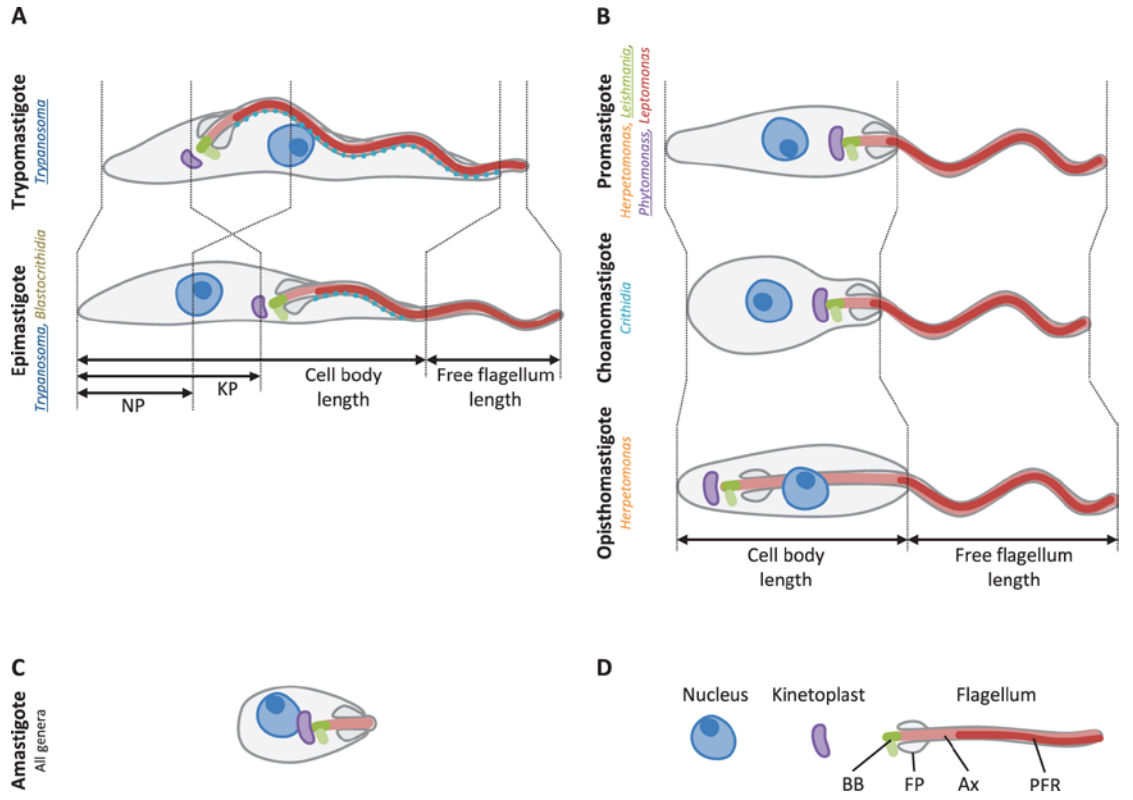


Figura 1. Diagrama de diferentes formas morfológicas observadas en tripanosomátidos. A. Morfología con flagelo lateralmente unido al cuerpo celular (epimastigotas y tripomastigotas). B. Morfología con flagelo libre (se extiende desde el bolsillo flagelar). C. Morfología sin flagelo móvil (Amastigota). D. Clave de estructuras: Núcleo (azul), kinetoplasto (violeta), y flagelo (cuerpo basal –BB-, bolsillo flagelar –FP-, axonema –Ax- y *paraflagellar rod protein* –PFR-). Métricas: KP, distancia kinetoplasto-posterior, NP, distancia núcleo-posterior. El género en el que cada morfología ocurre (género monofilético se indica subrayado). Tomado de [17].

Una característica a destacar es que los parásitos *Leishmania spp*. y *T. cruzi* son intracelulares obligados en el mamífero, mientras que los tripanosomas africanos se desarrollan extracelularmente en este hospedero. En la Figura 2 se esquematiza las principales formas y etapas del ciclo de vida de los tripanosomátidos intracelulares (ejemplificado por el ciclo de *T. cruzi*) y de aquellos que permanecen extracelulares en el hospedero mamífero (ejemplificado por *T. brucei*).

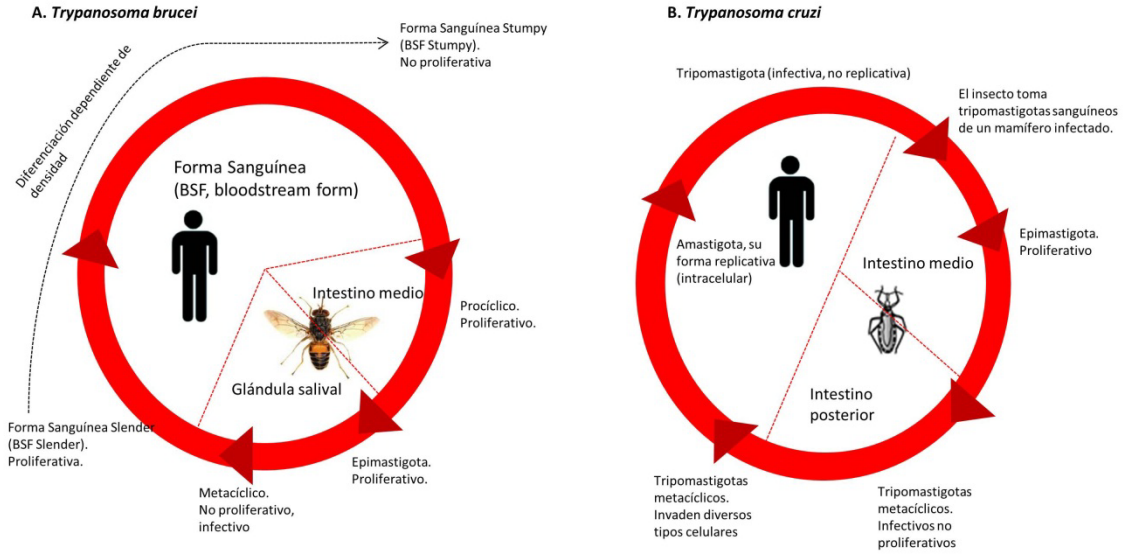


Figura 2. Esquema del ciclo de vida de los parásitos *T. brucei* y *T. cruzi*. A. Ciclo de vida resumido de *T. brucei*. B. Ciclo de vida resumido de *T. cruzi*.

El ciclo de vida en *Trypanosoma cruzi* involucra dos hospederos, un mamífero que es considerado como hospedero definitivo y un insecto triatomino, el cual funciona como vector. En el insecto triatomino el parásito se replica en su forma epimastigota en el intestino medio y se diferencia a su forma infectiva (tripomastigota metacíclico, no replicativo) cuando alcanza el intestino posterior y es expulsado en las heces, accediendo al torrente sanguíneo del hospedero mamífero a través de heridas de la piel o atravesando membranas mucosas. Los tripomastigotas de *T. cruzi* invaden una gran variedad de tipos celulares, y son capaces de escapar de la vacuola fagolisosómica y multiplicarse en el citoplasma celular (forma amastigota). Luego de multiplicarse por división celular, los amastigotas vuelven a diferenciarse en la forma tripomastigota (forma infectiva, no replicativa) que puede volver a infectar otras células o comenzar nuevamente el ciclo cuando la sangre es ingerida por un insecto [5].

En *T. brucei*, durante la fase del ciclo de vida correspondiente al insecto vector (mosca *tse-tsé*) los parásitos se mueven desde el intestino medio (forma llamada procíclica) hasta las piezas bucales del insecto antes de ser inoculado en el hospedero mamífero en su forma infectiva. Los epimastigotas de *T. brucei* no permanecen en las piezas bucales, ya que antes deben migrar desde el proventrículo hacia las glándulas salivales,

donde se transforman en epimastigotas metacíclicos (la forma infectiva) para luego ser expulsados con la saliva para infectar al hospedero mamífero (Figura 3).

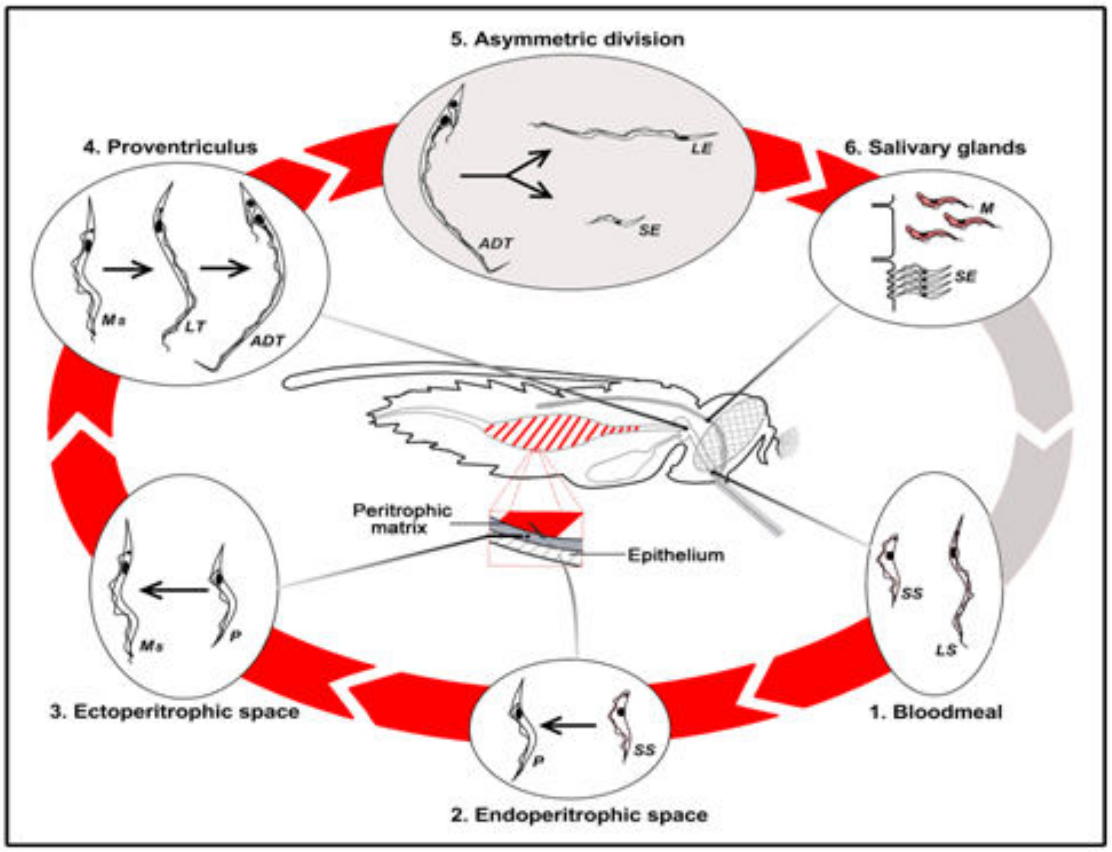


Figura 3. Ciclo de vida de *T. brucei* dentro del insecto vector. Luego de la ingesta de sangre infectada (1), los parásitos sanguíneos se transforman rápidamente en la forma procíclica en el lumen del intestino medio de la mosca. Las formas procíclicas atraviesan la Matriz Peritópica (PM) (2) y se establece y prolifera en el espacio ectoperitópico (ES) (3). Luego de varios ciclos de división, el parásito vuelve a cruzar la matriz y migra al proventrículo (4) donde se transforma inicialmente en forma mesocíclica y luego en tripomastigota largos (LT). Luego se produce una división asimétrica, para dar lugar a las formas epimastigotas largas y cortas (5).Por último, las formas epimastigotas cortas se unen a la glándula salival para proliferar y diferenciarse en las formas infectivas (epimastigota metacíclico) (6). Tomado de [18].

Una vez en el mamífero los epimastigotas metacíclicos sufren diversos cambios (morfológicos y fisiológicos) dando lugar a la forma sanguínea (*bloodstream form*). *Trypanosoma brucei* permanece completamente extracelular durante toda su estadía en el mamífero, donden se puede encontrar en dos formas: *slender* o *stumpy*. La primera es replicativa mientras que la segunda es quiescente. El pasaje de una forma a la otra depende de la densidad de la población parasitaria. La forma *slender* se encuentra durante la fase exponencial de crecimiento del parásito, mientras que la forma *stumpy* se encuentra en los picos de parasitemia (cuando la densidad de parásitos es elevada). El pasaje de *slender* a *stumpy* es un mecanismo de autocontrol poblacional similar al *quorum sensing* de las bacterias [19]. El ciclo vuelve a comenzar

cuando el insecto vector toma sangre infectada, y los parásitos pasan a su forma procíclica en el intestino medio del invertebrado [20].

Como ya ha sido mencionando en los tripanosomas africanos la transmisión cíclica involucra como vectores a diversas especies del genero *Glossina* (mosca *tse-tsé*), cuya distribución se restringe al África sub-sahariana. Sin embargo, la transmisión también puede ser mecánica (sin completar el ciclo), lo que ha permitido una distribución mundial más allá de la presencia del vector. Este es el caso de la presencia de tripanosomas africanos (particularmente en *T. vivax* y *T. evansi*) en América, como ya fue mencionado, donde la transmisión es exclusivamente mecánica, perpetuada por tábanos [21], otros insectos hematófagos, vampiros [22, 23] y jeringas contaminadas (debido a la práctica común de utilizar una única jeringa para, por ejemplo, la vacunación de ganado) [24].

Una de las particularidades del ciclo de los tripanosomas africanos es la capacidad de permanecer constantemente expuestos al sistema inmune del hospedero mamífero durante la infección, ya que permanecen en el torrente sanguíneo. Para poder perpetuarse en este entorno altamente hostil, estos parásitos desarrollaron uno de los sistemas más sofisticados de evasión de la respuesta inmune, denominado variación antigénica, que será desarrollado más adelante.

a.3 Organización genómica de tripanosomátidos

El genoma de los tripanosomátidos es diploide, aunque un trabajo reciente, propone en el caso de *Leishmania* la aneuploidía como la norma más que la excepción tanto para cepas de laboratorio como para aislados naturales[25]. Las poblaciones son en general de origen monoclonal, aunque existe evidencia de intercambio sexual en *Leishmania spp*. [25], *T.cruzi* [26] y *T. brucei* [27].

El tamaño (haploide) de los genomas secuenciados se encuentra entre 14 Mb (*T. rangeli*) y 47,5 Mb (*T. vivax*) distribuidos en un número variables de cromosomas dependiendo de la especie. Por ejemplo, se encuentran tan sólo 11 grandes cromosomas diploides en *T. brucei*, y numerosos cromosomas pequeños que contienen regiones altamente repetitivas, mientras que *T. cruzi* y *L. major* contienen 28 y 36 pares de cromosomas más pequeños, respectivamente [28].

El genoma de los tripanosomátidos se organiza en grupos de diez hasta cientos de genes ordenados consecutivamente en una misma hebra del ADN. Esta organización fue observada por primera vez en el cromosoma 1 de *Leishmania major,* el primer cromosoma secuenciado de un tripanosomátido [29], que contiene 85 genes organizados en dos grupos, 32 genes (el primero) en la hebra complementaria y los restantes 53 en la hebra molde [29, 30], como se esquematiza en la Figura 4. Por otra parte estos grupos presentan un patrón de transcripción policistrónico al estilo operón bacteriano, lo cual será abordado con mayor detalle en la siguiente sección.

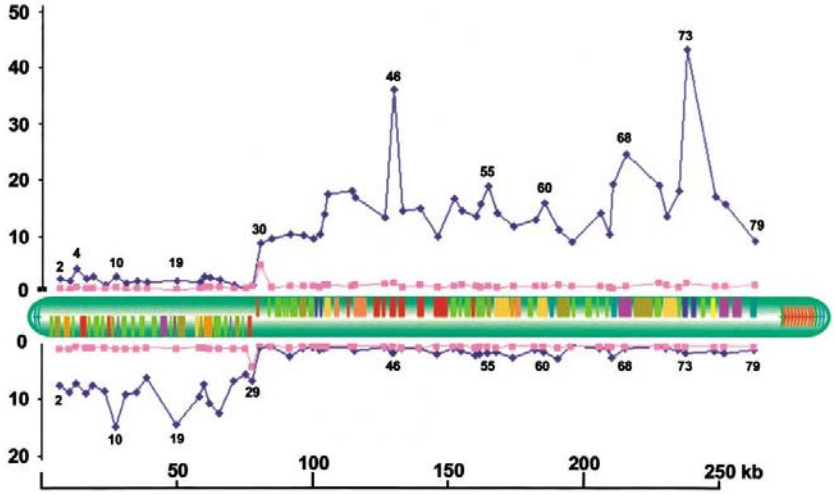


Figura 4. Demostración de transcripción bi-direccional en el cromosoma 1 de *Leishmania major* en ensayos de ARN *run-on*. Se grafica la señal de hibridación de ARN nuclear en diferentes sondas que representan la hebra superior (sobre el mapa del cromosoma 1) o las de la hebra inferior (se muestran debajo). Los números indican la posición de los genes de cada sonda. En las abscisas se indica la señal para cada sonda (línea azul). Se observa cómo hasta el gen 29, la transcripción es desde la hebra inferior, y a partir de la posición 30, la transcripción es desde la hebra superior. Tomado de [30].

La región donde sucede el cambio de hebra se denomina de *switch* transcripcional, los cuales pueden ser convergentes o divergentes según se muestra en la Figura 5:

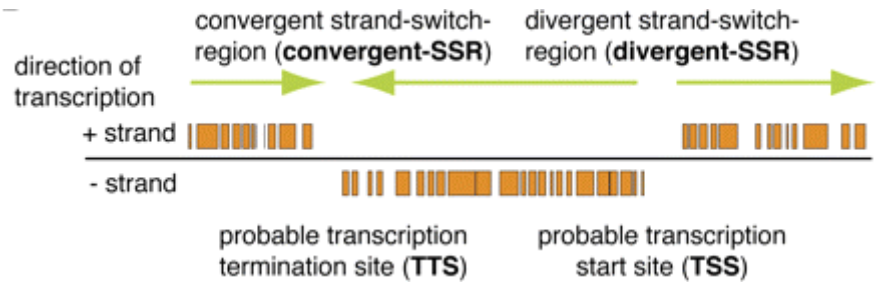


Figura 5. Sitios de *switch* de transcripción (cambio de hebra y de cistrón) convergentes y divergentes. Las regiones convergentes contienen sitios de finalización de transcripción de la ARN polimerasa II, mientras que las regiones divergentes contienen sitios de comienzo de transcripción de la ARN polimerasa II. Los bloques naranjas representan marcos de lectura abiertos y las flechas verdes indican la dirección de la transcripción (tomado de [31]).

El proyecto genoma de tripanosomátidos "Tritryp" publicó en julio de 2005 el genoma completo de *Trypanosoma cruzi* [28] junto con los genomas de *Trypanosoma brucei* [32] y *Leishmania major* [33], mostrando que la mayoría de los genes de estos organismos se organizan del modo antes mencionado.

Los genomas de los tripanosomátidos secuenciados mostraron altos niveles de conservación de la sintenía (conservación del orden de sus genes) [34], aunque la divergencia entre estas especies se calcula en 200 a 500 millones de años [35]. Prácticamente todos los *clusters* de genes ortólogos (94%) entre los tres genomas están en regiones de sintenía conservada. Asimismo, alineamientos de secuencias aminoacídicas muestran una identidad promedio de 57% entre secuencias codificantes (CDS) de *T.brucei* y *T. cruzi*, y 44% entre este último y *L. major*.

La mayoría de los genes especie-específicos (de los cuales *T. cruzi* y *T. brucei* presentan una mayor proporción -32% y 26% respectivamente- respecto a *L. major* -12%-) ocurren en regiones cromosómicas internas no sinténicas y en regiones subteloméricas. En general se trata de familias de genes que codifican para proteínas de superficie. Estas familias, junto con ARNs estructurales y retroelementos, se asocian con interrupciones de sitios de sintenía. Una característica remarcable en los genomas de *T. brucei* y *T. cruzi* es la expansión de estas familias génicas (como las Glico-proteínas variables de superficie –VSGs- en *T. brucei* y las trans-sialidasas y mucinas y proteínas asociadas en *T. cruzi*) [36].

Otra de las características de los tripanosomátidos es que los genes, salvo 2 excepciones caracterizadas en *T. brucei*, no presentan intrones.

Desde la publicación en 2005 de los genomas de estos tres parásitos del orden kinetoplastida, y gracias a las nuevas tecnologías de secuenciación, se han incorporado decenas de nuevos genomas (en versión borrador o completos) a las bases públicas. Entre estos genomas secuenciados encontramos los de *L. infantum* y *L. braziliensis* [37], *L. donovani* [38, 39], *L. mexicana* [40], *L. tarentolae* [41], *T. brucei gambiense* [42], *T. congolense* y *T. vivax* [43], *L. amazonensis* [44], *T. rangeli* [45] y *T. evansi* [46]. En la Tabla 1 se resumen los genomas de tripanosomátidos disponibles en la base de datos tritrypdb (www.tritrypdb.org).

Tabla 1. Genomas disponibles en tritrypdb (www.tritrypdb.org, versión 24). Se indica organismo, número anotado de genes, versión del genoma, tamaño y si se encuentra publicado.

[Organismo]	[Número de genes]	[Datos]	[Version del Genoma]	[Mb]	[Publicado?]
Crithidia fasciculata strain Cf-Cl	11950	BeverleyLab	21/01/2014	40.29	no
Endotrypanum monterogeii strain LV88	nd	GenBank	05/02/2013	32.52	no
Leishmania aethiopica L147	nd	GenBank	07/08/2013	31.99	no
Leishmania amazonensis MHOM/BR/71973/M2269	nd	GenBank	25/07/2013	29.03	si
Leishmania arabica strain LEM1108	nd	GenBank	12/06/2013	31.44	no
Leishmania braziliensis MHOM/BR/75/M2903	8966	BeverleyLab	03/03/2014	35.21	no
Leishmania donovani BPK282A1	8195	GeneDB	16/01/2013	32.44	si
Leishmania enriettii strain LEM3045	nd	GenBank	13/06/2013	30.78	no
Leishmania gerbilli strain LEM452	nd	GenBank	05/08/2013	31.40	no
Leishmania infantum JPCM5	8381	GeneDB	16/01/2013	32.13	si
Leishmania major strain Friedlin	9378	GeneDB	16/01/2013	32.86	si
Leishmania mexicana MHOM/GT/2001/U1103	9063	GeneDB	16/01/2013	32.11	si
Leishmania panamensis MHOM/COL/81/L13	nd	GenBank	21/02/2013	31.26	no
Leishmania sp. MAR LEM2494	nd	GenBank	10/06/2013	30.87	no
Leishmania tarentolae Parrot-TarII	8530	ULAVAL	22/06/2011	31.63	si
Leishmania tropica L590	nd	GenBank	12/06/2013	32.99	si
Leishmania turanica strain LEM423	nd	GenBank	29/07/2013	32.32	no
Trypanosoma brucei gambiense DAL972	10000	GeneDB	16/01/2013	22.15	si
Trypanosoma brucei Lister strain 427	9302	GeneDB	20/10/2010	26.75	no
Trypanosoma brucei TREU927	12094	GeneDB	28/08/2013	35.83	si
Trypanosoma congolense IL3000	13358	GeneDB	16/01/2013	41.37	si
Trypanosoma cruzi CL Brener Esmeraldo-like	10597	GeneDB	16/09/2014	32.53	si
Trypanosoma cruzi CL Brener Non-Esmeraldo-like	11106	GeneDB	16/09/2014	32.53	si
Trypanosoma cruzi Dm28c	11398	GenBank	13/11/2013	27.35	si
Trypanosoma cruzi JR cl. 4	nd	GenBank	17/01/2013	41.48	no
Trypanosoma cruzi marinkellei strain B7	10282	Franzen	1.0	38.65	si
Trypanosoma cruzi Sylvio X10/1	10947	GenBank	02/10/2012	38.59	si
Trypanosoma cruzi Tula cl2	nd	GenBank	26/06/2013	83.51*	si
Trypanosoma evansi strain STIB 805	10174	Schnaufer	03/06/2014	25.43	si
Trypanosoma grayi ANR4	10686	FieldLab	21/03/2014	20.95	si
Trypanosoma rangeli SC58	7479	GenBank	30/10/2013	14.02	si
Trypanosoma vivax Y486	12581	GeneDB	16/01/2013	47.50	si

* Probablemente el tamaño indicado no se corresponda con el tamaño real, puede deberse a problemas de ensamblajes o que se refiera al tamaño del genoma diploide. nd: no determinado.

a.4 Expresión génica en tripanosomátidos

Como se dijo antes, los genes se encuentran organizados en grupos a nivel del ADN, y éstos se transcriben como precursores policistrónicos que pueden tener hasta 600 kb [47]. Esta molécula precursora sufre un proceso de maduración consistente en la adición en el extremo 5' de una secuencia denominada miniexón -o *spliced-leader*- y la poliadenilación del extremo 3' (Figura 6). El miniexón es adicionado por un mecanismo peculiar llamado *trans-splicing*, ya que la secuencia del miniexón es codificada de forma independiente en otra región del genoma del parásito (Figura 6).

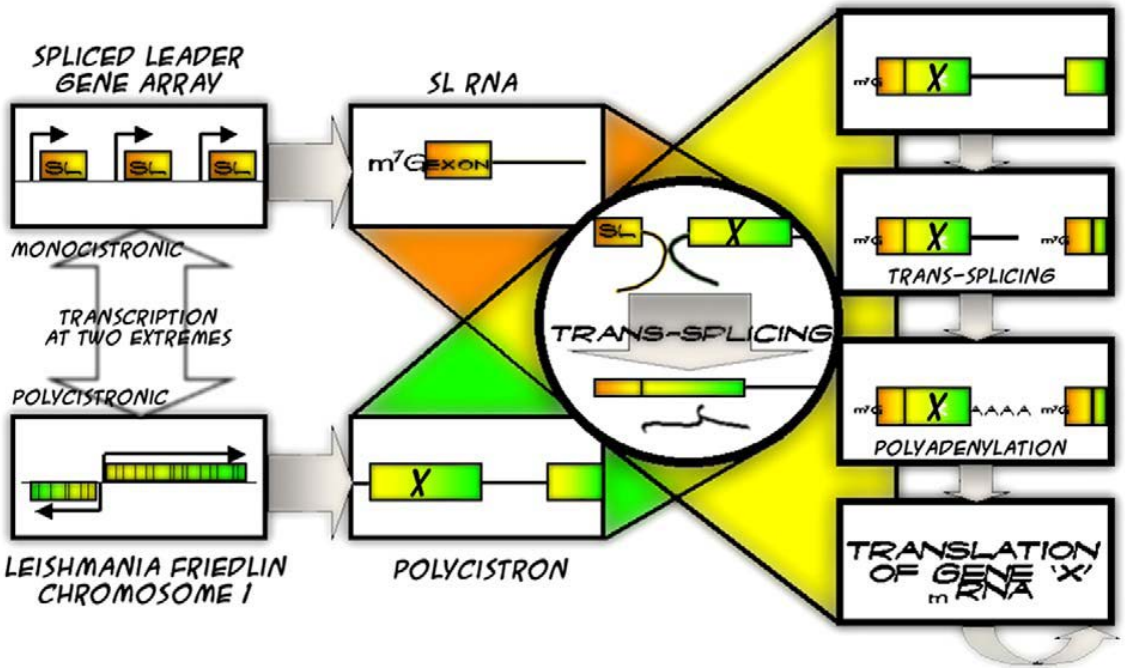


Figura 6. Esquema de transcripción discontinua y producción de ARNm maduro mediante *trans-splicing*. A la izquierda se muestra la transcripción por la ARNpolII de genes repetidos en tándem (SL ARN) y una región policistrónica (ej. cromosoma 1 de *L. major*). En el centro de muestran los transcriptos primarios y el resultado del mecanismo del *trans-splicing* y la maduración final del mensajero (incorporación de cola poliA) se muestra a la derecha (Figura tomada de [16]).

Una vez sintetizado un ARN policistrónico, los genes para los que codifica pueden ser expresados: (1) con similares niveles, como las calmodulinas o tubulinas [48, 49], (2) a muy diferentes niveles, como los genes de la glicoproteína variable de superficie (VSG) [49, 50], o (3) a niveles que varían enormemente dependiendo del estadio de desarrollo del parásito, como en el caso de la aldolasa [51].

La biosíntesis de grandes moléculas de ARN precursoras de ARNm conlleva la pregunta de cómo regulan la expresión génica los tripanosomátidos. Existen varias evidencias para afirmar que, a diferencia de la mayor parte de los eucariotas, donde el principal

punto de regulación es a nivel del inicio de la transcripción, en los tripanosomátidos la regulación es co- o post-transcripcional [52].

Consecuentemente, el determinante primario de la regulación de la expresión génica en tripanosomátidos es a nivel de las reacciones de procesamiento que ocurren en las regiones intergénicas del pre ARNm, en la estabilidad y degradación diferencial de los diferentes mensajeros [16] así como a nivel traduccional, como lo muestran trabajos recientes en *T. brucei* y *T. cruzi* utilizando "ribosome profiling" [53-55]. Esto es una diferencia sustancial con la mayoría de los eucariotas, donde el determinante primario de la regulación es a nivel del inicio de la transcripción [56].

Los tripanosomas poseen además copias altamente conservadas de las tres ARN polimerasas eucariotas [6]. El rol de cada una de ellas se esquematiza en la Figura 7.

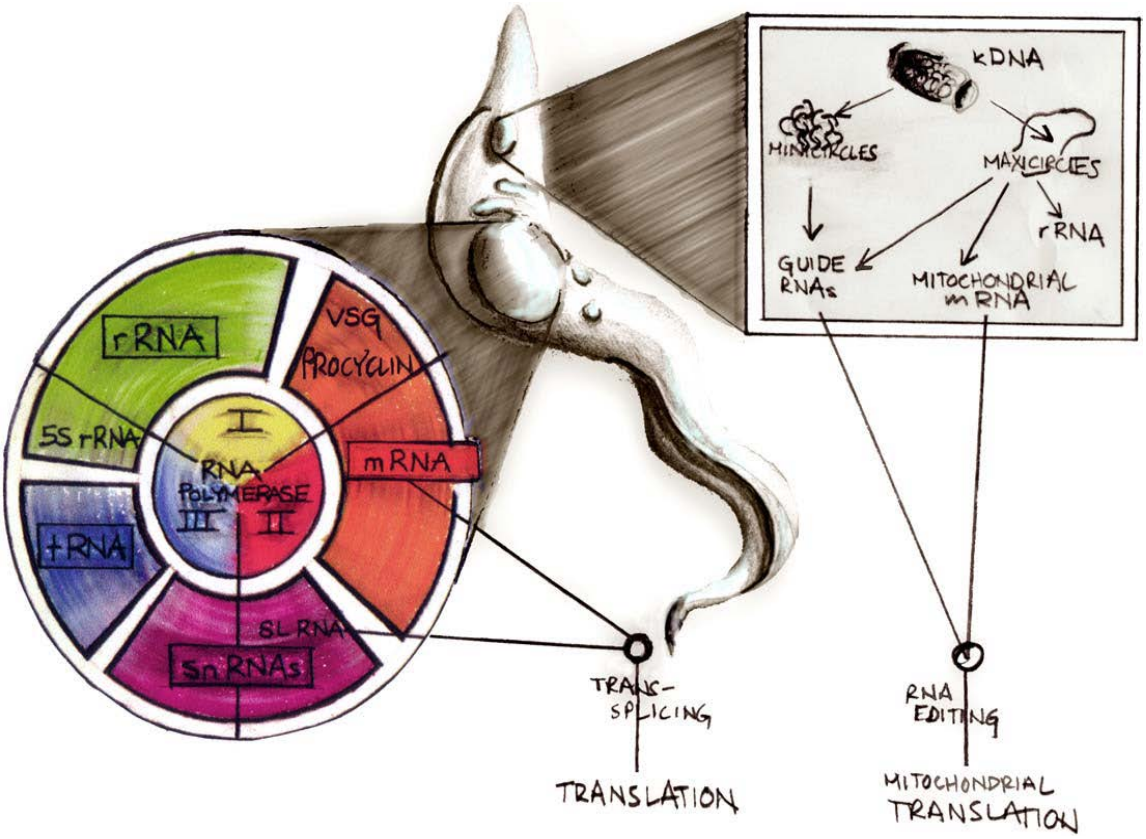


Figura 7. Transcripción en tripanosomátidos. La transcripción de diferente clase de genes nucleares y qué ARN polimerasa (I, II o III) es responsable se indica en el círculo a la izquierda. El ARNt es sintetizado en el núcleo y transportado a la mitocondria a través del citosol (Figura tomada de [16]).

La ARN polimerasa II es la responsable de la transcripción de la mayoría de los ARN codificantes para proteínas. Los policistrones transcriptos por esta polimerasa no necesariamente codifican para proteínas funcionalmente relacionadas (como ocurre en los operones de bacterias y nematodos) [16]. La regulación, entonces, de estos ARNm está determinada a nivel post-transcripcional, fundamentalmente mediada por

la degradación del ARNm controlada por la región 3' UTR [57] o por diferencias en la eficiencia traduccional [53, 54].

Una excepción a este patrón de regulación lo constituye, en los tripanosomas africanos, el arreglo de genes que codifican para las glicoproteínas variantes de superficie (VSG) y prociclinas en formas sanguíneas y procíclicas, respectivamente, que son co-reguladas y co-transcriptas por la ARN polimerasa I.

Si bien la regulación de la expresión es mayoritariamente post-transcripcional, indicando que la abundancia de los transcriptos no variaría significantemente entre las diferentes etapas de vida de estos parásitos, un trabajo reciente mostró mediante Digital Gene Expression (DGE) que existen varios grupos de genes en el genoma de *T. b. gambiense* pertenecientes a la misma unidad policistrónica que son expresados diferencialmente, y son co-regulados tanto a nivel de la transcripción como de su estabilidad [57].

En este trabajo se obtuvo información sobre 7360 genes (correspondiente a 81% de los genes anotados de *T. brucei*) y se observó que 73 genes se encontraban sobre-expresados en las formas sanguíneas y 25 genes sobre-expresados en las formas procíclicas respecto a las sanguíneas (utilizando un umbral de *Fold Change* –FC- de 2,5 y un *False Discovery Rate* -FDR- de 0,1).

Como se esperaba una gran proporción de los genes sobre-expresados en las formas sanguíneas (22/72) están asociados a la variación antigénica de superficie, los genes VSG y ESAG (*expression site associated genes*), ambos expresados en altos niveles en esta etapa del ciclo del parásito [57].

Asimismo, las diferencias de metabolismo entre las formas procíclica y sanguínea en cuanto a la fuente de energía (glucosa en las formas sanguíneas y prolina en las procíclicas) se ve reflejada en la sobre-regulación en las formas sanguíneas del gen del transportador de glucosa THT1, el gen ALD que codifica para la fructosa-bi-fosfato aldolasa, entre otros [57].

Este trabajo, así como trabajos de *microarrays*, tanto en *T. brucei* [58] como en *T. cruzi* [59], revelan la existencia de regulación de la abundancia de ARN mensajeros asociadas a diferentes etapas del ciclo de vida de los parásitos. Muchos de los cambios observados en estos trabajos se pueden asociar fácilmente a los cambios que sufre el parásito durante los diferentes estadios: la transición entre formas que se dividen y que no se dividen afectan mayormente funciones como transcripción y traducción, mientras que las transiciones entre hospedero afectan principalmente el metabolismo, las proteínas de superficie y el transporte intracelular [58].

Siegel y colaboradores [60], utilizando la herramienta de RNAseq mostraron no sólo la expresión diferencial de algunos genes (incluso dentro de la misma unidad policistrónica), sino también el uso de sitios de *trans-splicing* alternativo como mecanismo de regulación de la transcripción entre dos estadios del parásito *T. brucei* (forma procíclica y sanguínea). Asimismo Nilsson y colaboradores [61] mediante secuenciación profunda con la metodología de *spliced-leader trapping*, mostraron que el uso de sitios de *trans-splicing* alternativo es un mecanismo importante de regulación de la expresión génica en los tripanosomátidos.

Los trabajos recientes de Vasquez y colaboradores [54], Jensen y colaboradores [53] y Smircich y colaboradores [55] mostraron, mediante la metodología de huella ribosomal (*ribosome profiling)* [62], los grandes cambios que se producen a nivel traduccional y de eficiencia traduccional en los diferentes estadios del parásito *T. brucei* [53, 54] y *T. cruzi* [55]. Esta metodología permite determinar la tasas de traducción, y permite además la cuantificación de transcriptos que están siendo traducidos (la correlación entre estos datos y los niveles de proteínas se ajusta mucho mejor que cuando se utilizan sólo los valores de ARNm obtenidos mediante RNAseq).

Por otra parte, no sólo se obtienen valores cuantitativos, sino también información sobre la posición específica que ocupan los ribosomas en el ARNm. Esta información es importante, ya que la asociación de un ribosoma con un ARNm no significa que éste esté siendo traducido (por ejemplo puede estar asociado a la región 5'UTR) [54]. Asimismo evidencia la presencia de marcos de lectura abiertos pequeños (uORF, por *upstream* ORF) [63] en el 5' UTR de algunos genes con implicancia en la eficiencia traduccional de los CDS corriente abajo de éstos.

En el trabajo de Vasquez y colaboradores se observaron diferencias de la eficiencia traduccional de hasta 100 veces entre genes, así como diferencias estadio-específicas para cientos de genes comparando entre las formas sanguíneas y procíclicas del parásito *T. brucei* [54].

Asimismo, en el caso de *T. cruzi*, se observaron diferencias en la cantidad de genes traducidos entre la forma replicativa – epimastigota - (74 % del transcriptoma) y la forma infectiva - tripomastigota metacíclico - (58% del transcriptoma). Esta diferencia es explicada fundamentalmente por inhibición específica de la traducción en un número importante de ARNs mensajeros [55]. Este trabajo, así como los realizados en el parásito africano modelo *T. brucei*, muestran la importancia del control de la traducción como mecanismo de regulación génica en los tripanosomátidos.

a.5 Genoma mitocondrial

Una de las principales características de los kinetoplástidos, a lo cual deben su nombre es la organización del ADN mitocondrial (kADN o kinetoplasto) en una intrincada estructura que consiste en una gran de red de moléculas circulares de ADN concatenadas como se muestra en la clásica microscopía electrónica del ADN mitocondrial de *Crithidia fasciculata* (

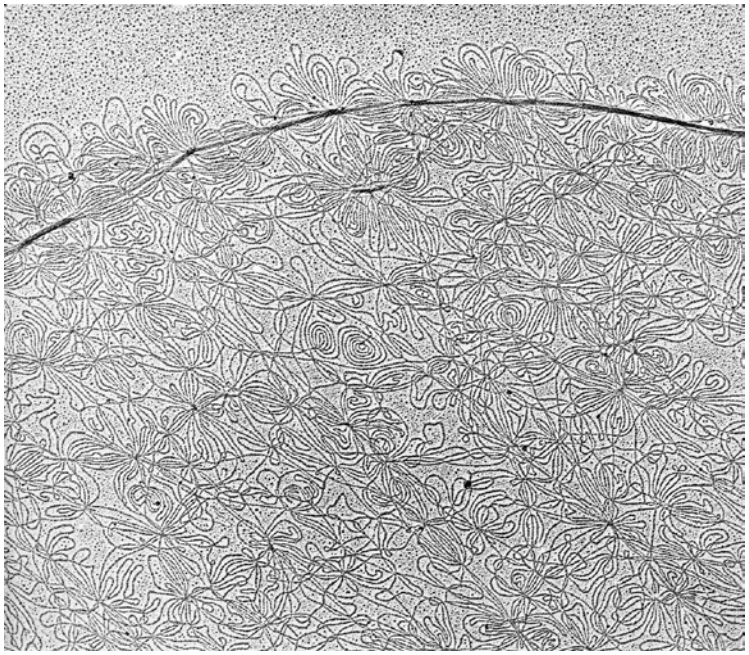


Figura 8). El kADN se encuentra, además organizado en dos tipos de moléculas circulares denominadas maxicírculo y minicírculo.

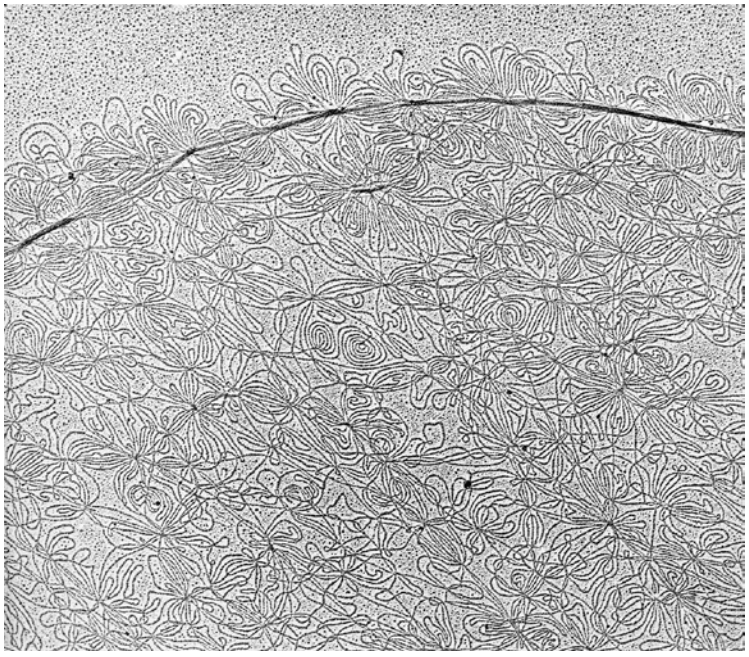
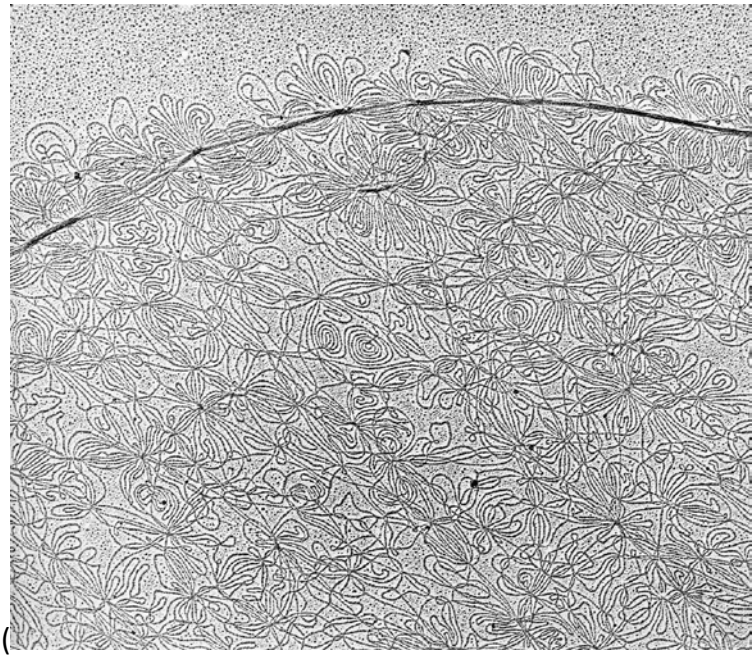


Figura 8. Sección de microscopía electrónica del kADN de *C. fasciculata*. Los pequeños *loops* son minicírculos (2,5 Kb); las largas hebras enrolladas son probablemente fragmentos de maxicírculo (38 Kb). Tomado de [64].



La estructura relajada (Figura 8) es comparable en tamaño a la totalidad de la célula, pero cuando la red se encuentra en la matriz de la única mitocondria celular, se encuentra condensada en una estructura altamente organizada en forma de disco [64].

El kADN se replica una única vez por ciclo celular y se encuentra sincronizado con la replicación y división del ADN nuclear [65].

En los tripanosomas africanos, la función mitocondrial es cambiante a lo largo del ciclo de vida de los parásitos y puede ser completamente funcional y activa (observada en las formas procíclicas en el insecto vector), hasta funcional y morfológicamente reprimida en las formas sanguíneas. Además, en las sub-especies de *T. brucei*, *T. b. evansi* y *T. b. equiperdum*, (parásitos que han perdido el ciclo completo, permaneciendo únicamente en forma sanguínea), se observa la pérdida parcial (diskinetoplástidos) o total de su ADN mitocondrial (akinetoplástidos), respectivamente [66].

Los cambios que sufre el parásito durante su ciclo de vida, se acompasan con grandes cambios en la actividad y morfología mitocondrial, y es por ello un modelo atractivo para estudiar procesos de activación y represión de vías específicas, así como las consecuencias de la pérdida del ADNk.

a.5.1 Maxicírculos:

Los maxicírculos se encuentran en decenas de moléculas por mitocondria [64], y su tamaño varía entre 20 y 40 Kb dependiendo de cada especie [67]. En los maxicírculos se encuentran los principales genes asociados a la función mitocondrial (principalmente subunidades de la cadena respiratoria) así como 2 genes de ARN ribosomal. La región codificante (cantidad de genes, tipo y orden) se encuentra conservada en todos los tripanosomátidos [68].

Una de las principales características de los genes que se encuentran en los maxicírculos es que algunos de ellos sufren grandes procesos de edición post-transcripcional de su secuencia para poder ser traducidos. Este proceso denominado *editing* de genes mitocondriales, consiste en la adición y eliminación de uridinas, y es detallado en la siguiente sección (a.6).

a.5.2 Minicírculos:

Las moléculas de kADN denominadas minicírculos, varían en tamaño entre 400 pares de bases (*T. vivax*) hasta 10 Kb (*T. avium*) [67].

Cada minicírculo esta concatenado con 2 o 3 minicírculos vecinos (Figura 9). Los minicírculos se encuentran covalentemente cerrados, excepto cuando la red se replica. Además a diferencia de otros ADNs circulares en otros tipos celulares, éstos no se encuentran superenrollados [69].

El número de moléculas de minicírculos varía entre 5 y 50 mil copias, y aunque en general son uniformes en cuanto a su tamaño, presentan gran variación a nivel de secuencia [67]. En el caso de *T. brucei* por ejemplo, el set de minicírculos incluye entre 300 y 400 clases de secuencias diferentes [70].

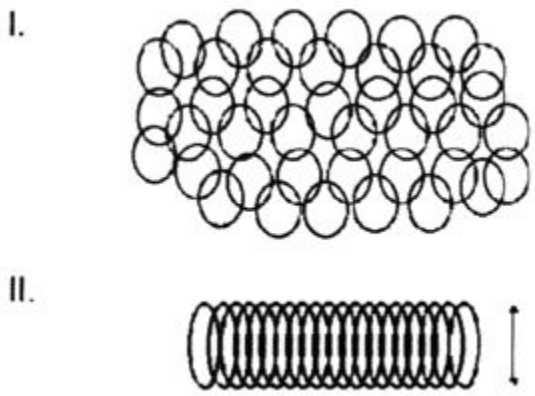


Figura 9. Diagrama mostrando la organización de los minicírculos. I. Representación de un segmento aislado de una red mostrando minicírculos interconectados. **II.** Diagrama de una sección de la red condensada en un disco *in vivo*. La flecha indica el espesor del disco que es aproximadamente la mitad de la circunferencia de un minicírculo. Tomado de [64].

Aunque presentan gran variación a nivel de secuencia, los minicírculos presentan una región de 120 nucleótidos que contiene 3 bloques (CSB por *conserved sequences block*) con diferentes grados de conservación, particularmente la región CSB-3 (o UMS por *Universal Minicircle Sequence*) de 12 nucleótidos se encuentra altamente conservada en todos los tripanosomátidos [71]. Este nivel de conservación ha sido asociado a su función, ya que funcionaría como origen de replicación [72].

Como se mencionó más arriba, algunos de los genes codificados por el maxicírculo deben ser altamente editados para ser traducidos. La información genética para el *editing* está contenida en moléculas denominadas ARN guía (ARNg), y los mismos se encuentran mayormente codificados en los minicírculos (aunque algunos ARNg son codificados en los maxicírculos). Hasta el momento, la codificación de los ARNg es la única función conocida de los minicírculos. El número de clases de secuencias de minicírculos diferentes es tal que permite la edición completa de los genes mitocondriales que lo requieren.

a.6 *Editing* del ARN mitocondrial

a.6.1. Introducción

El *editing*, es decir la edición del ARN, es un fenómeno post-transcripcional que se encuentra en todos los dominios de la naturaleza, y fue descripto por primera vez en el gen mitocondrial COII de *T. brucei* [73].

El *editing* del ARN mensajero mitocondrial en tripanosomas es un proceso post-transcripcional por el cual se insertan o quitan residuos de uridina (Us), en el ARNm precursor. La inserción/deleción de estos residuos es capaz de corregir desplazamientos en el marco de lectura (*frameshifts*), introducir codones de comienzo y fin, e incluso agregar la mayor parte de la secuencia codificante. El *editing* recupera marcos de lectura abiertos de los ARNm mitocondriales, lo que puede darse por la adición de pocos residuos de uridina (el caso de COXII se adicionan 4 residuos de uridina para recuperar el marco de lectura) hasta cientos de inserciones en una única molécula de pre-ARNm. La inserción de uridinas es cerca de 10 veces más frecuente que la deleción [74].

En la Figura 10, se muestra como ejemplo la secuencia pre y post-editada del gen de la proteína Ribosomal S12 (RPS12), un gen extensamente editado, donde se deletan 28 residuos de timidina, y se incorporan 114 residuos de uridina.

```
No Ed   A-G-GATTTTTTG-A-GG--G--GTTTGCTTTTTG----G----G---G----A-G--A-
Ed      AuGuGATTTTT-GuAuGGuuGuuGTTTGC-----GuuuuGuuuuGuuuGuuuuAuGuuAu

No Ed   -A-A-GAGTCCTTTTCCGA--GCCCAG--CCGG-AA-CGACG-G-G--G-A-GCCTTG-G
Ed      uAuAuGAGTCC----CCGAuuGCCCAGuuCCGGuAAuCGACGuGuGuuGuAuGCC--GuG

No Ed   ----A---G-A-AA----G-G-GG--G--GCG--G---------G--G-G-GTTTT--GG-
Ed      uuuuAuuuGuAuAAuuuuGuGuGGuuGuuGCGuuGuuuuuuuuGuuGuGuGTTTTuuGGu

No Ed   --GCATTTGTCG--A---A--A-AGATTTTTGTTTTGG-GG-GG----G--GA---ACCC
Ed      uuGCATTTGTCGuuAuuuAuuAuAGA-----G----GGuGGuGGuuuuGuuGAuuuACCC

No Ed   TTTTTGTTG-ATAAAG-A--A-ACACG-A-G---A--AA--AA
Ed      -----G--GuATAAAGuAuuAuACACGuAuGuuuAuuAAuuAA
```

Figura 10. Edición del gen mitocondrial RPS12. Alineamiento de secuencia sin editar y editada del gen RPS12**.** En rojo se indican los residuos de uridina que se agregan postranscripcionalmente, la deleción de residuos de timidina es indicado con **-.** No Ed: Secuencia no editada. Ed: Secuencia editada postranscripcionalmente.

En la Tabla 2 se muestran los genes presentes en el ADNk, el tipo de edición y el rol que cumplen.

Tabla 2. Genes presentes en el maxicirulo de tripanosomátidos. Los genes se encuentran ordenados de acuerdo a su posición en el genoma mitocondrial, se indica el grado de edición, la descripción y su función. CTE. Cadena de transporte de electrones.

Gen	Editing	Descripción	Función
12S	No editado	ARN Ribosomal 12S	ARN ribosomal
9S	No editado	ARN Ribosomal 9S	ARN ribosomal
ND8	Pan-editing	subunidad 8 NADH deshidrogenase ?	Compejo I (CTE)
ND9	Pan-editing	subunidad9 NADH deshidrogenase ?	Compejo I (CTE)
MURF5	No editado	maxicircle unidentified reading frame 5	Sin función conocida
ND7	Pan-editing	subunidad 7 NADH deshidrogenase ?	Compejo I (CTE)
COIII	Pan-editing	Citocromo oxidasa III	Complejo IV (CTE)
Cyb	5' editing	Citocromo b	Complejo III (CTE)
A6-ATPase (MURF4)	Pan-editing	ATP sintasa	Complejo V (CTE)
ND2 (MURF1)	No editado	subunidad 2 NADH deshidrogenase ?	Compejo I (CTE)
CR3	Pan-editing	función no conocida	Sin función conocida
ND1	No editado	?ubunidad 1 NADH deshidrogenase ?	Compejo I (CTE)
COII	editing parcial	Citocromo oxidasa II	Complejo IV (CTE)
MURF2	5' editing	maxicircle unidentified reading frame 2	Sin función conocida
COI	No editado	Citocromo oxidasa I	Complejo IV (CTE)
CR4	Pan-editing	función no conocida	Sin función conocida
ND4	No editado	subunidad 4 NADH deshidrogenase ?	Compejo I (CTE)
ND3	Pan-editing	subunidad 3 NADH deshidrogenase ?	Compejo I (CTE)
RPS12	Pan-editing	Proteína ribosomal S12	Ribosoma
ND5	No editado	subunidad 5 NADH deshidrogenase ?	Compejo I (CTE)

Los ARN guía, moléculas de entre 40 y 60 nucleótidos de largo, codificadas principalmente en los minicírculos, especifican la posición y el número de residuos de uridinas a ser insertados o eliminados. La reacción de *editing* requiere tres pasos: el corte del ARNm, la inserción o deleción de residuos de uridina y por último la re-ligación de la molécula de ARNm editada. Estas reacciones son catalizadas por un complejo de edición de ARN (RECC, RNA *editing core complex*). En *T. brucei*, por ejemplo, el proceso requiere más de 100 proteínas para la expresión de las 18 proteínas codificadas en el genoma mitocondrial [75]. El procesamiento, incluye además, el corte endonucleotídico del transcripto primario (policistrónico), la adición de una cola poli-A pequeña (30 nt), el *editing* propiamente dicho en el caso de los genes que lo requieren y por último la adición de colas largas poli-A para generar el ARNm totalmente editado [66]. Estos mecanismos se resumen en la Figura 11.

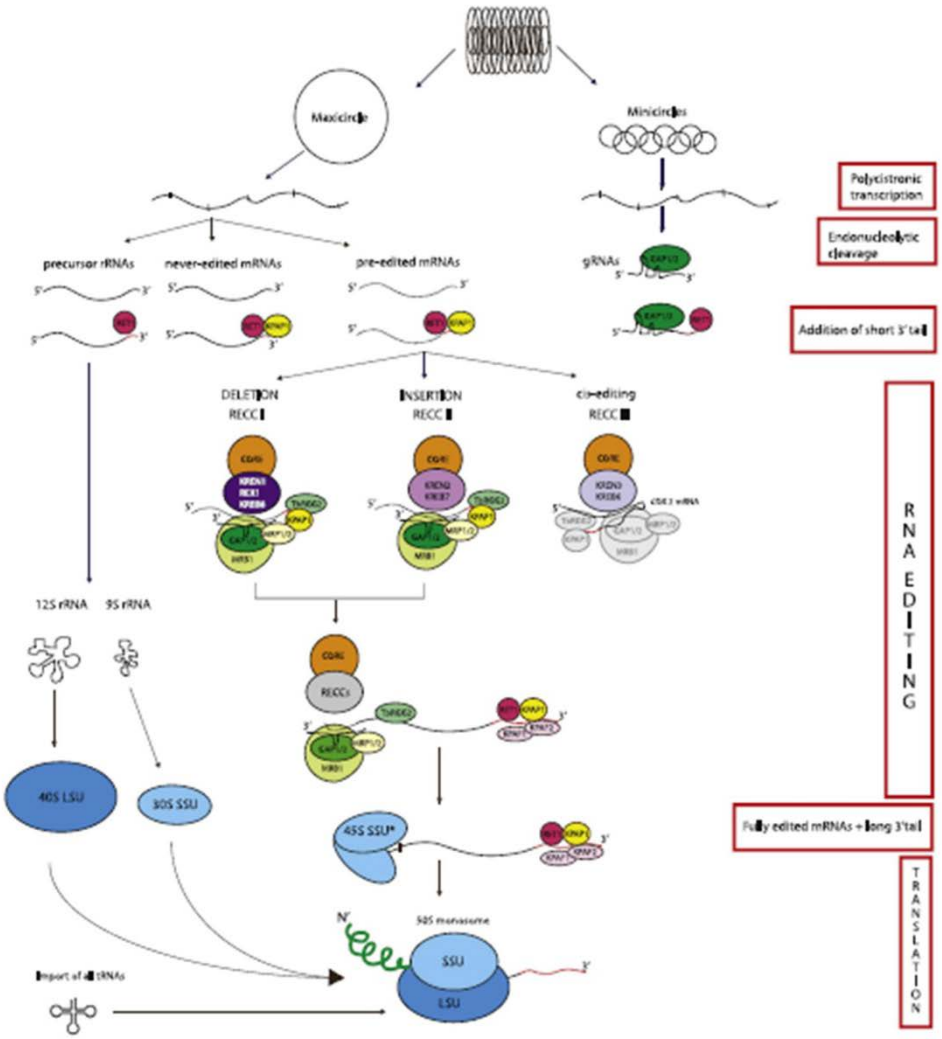


Figura 11. Esquema de los principales procesamientos de ARNm mitocondrial en kinetoplástidos. En la figura se esquematizan los principales procesamientos que sufre el ARNm desde la transcripción a partir del maxicírculo. Los ARNm transcriptos se pueden clasificar en: ribosomales, nunca editados y editados. Los ARNm que serán editados se transcriben y luego son procesados por el complejo RECC y finalmente son poliadenilados. La traducción requiere la importación de ARNt del citoplasma ya que ningún ARNt se encuentra codificado en el ADN mitocondrial. Tomado de [66].

a.6.2. Mecanismo de editing

El *editing* del ARN refiere a cualquier proceso de modificación post-transcripcional que introduce cambios en la secuencia transcripta relativa al gen correspondiente, cambiando entonces la información del ARN. Estos procesos de edición no incluyen el *splicing* ni el procesamiento terminal del ARN [76].

Cuando el *editing* de ARN fue descubierta por Benne y colaboradores [73] fue considerado como un extraño mecanismo primitivo de los protozoarios [77]. En este trabajo, Benne describió la inserción de cuatro residuos de uridina en el gen COII que no se encontraban en el genoma mitocondrial, generando un marco de lectura abierto. Los mecanismos de inserción y la fidelidad de la inserción eran desconocidos.

Blum y colaboradores, descubrieron luego los ARNg, que codificaban la información necesaria para insertar las uridinas en la posición correcta [78, 79].

El ARNg se une al ARNm pre-editado correspondiente mediante una región anchor de 8 a 12 nucleótidos situada en el extremo 5’ del ARNg. La interacción es facilitada mediante apareamientos de Watson y Crick, así como emparejamientos G:U. Las zonas de *editing* son marcadas por *mismatches* entre el ARNg y el ARNm a ser editado. Estos *mismatches* son blancos de las endonucleasas que cortan el ARNm pre-editado. Durante la inserción, un residuo de uridina se agrega al 3’ hidroxilo del extremo 5’ del producto cortado por la enzima terminal-uridil transferasa (TUTase). En la deleción, una exonucleasa remueve un residuo de uridina del extremo 3’ del producto 5’ clivado. Por último una ligasa, utilizando la hidrólisis de ATP une los productos 5’ y 3’ completando una ronda de *editing* [77]. El mecanismo se esquematiza en la siguiente figura (Figura 12).

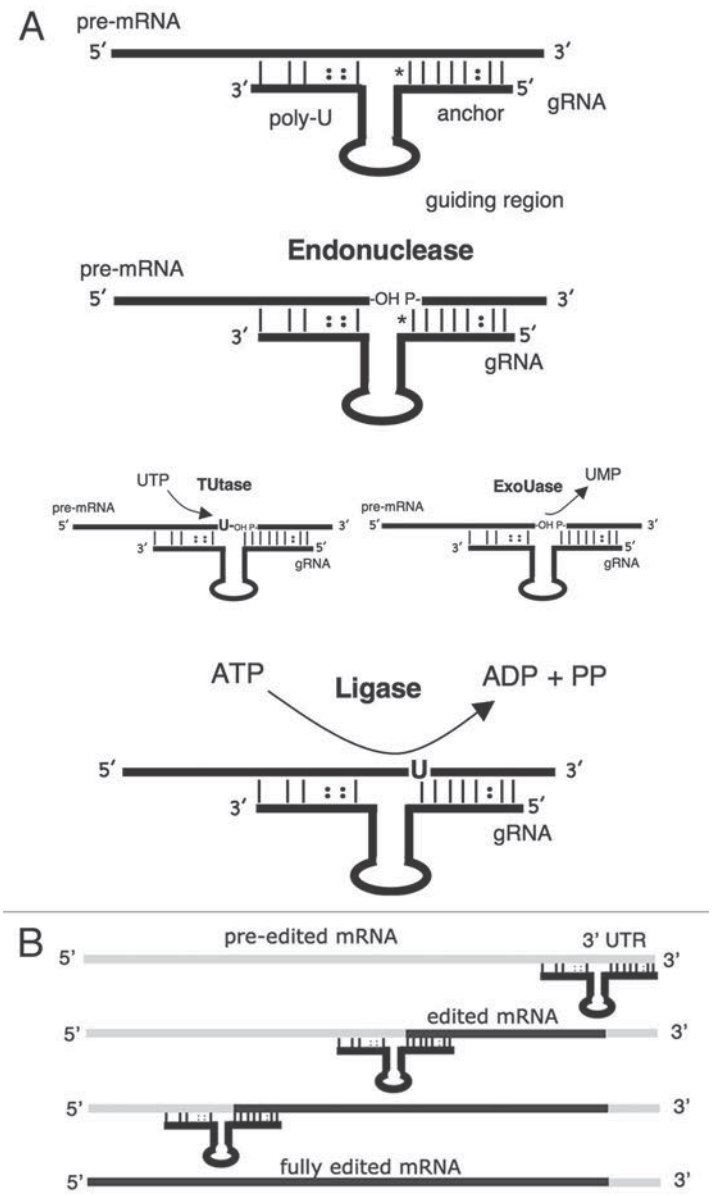


Figura 12. **Esquema del mecanismo general de inserción y deleción durante el proceso de *editing* del ARNm en Tripanosomas. A.** Mecanismo general de inserción y deleción en el ARNm mitocondrial. El mecanismo se explica en el texto. **B.** El progreso de la edición transcurre del extremo 3′ del pre-ARNm al extremo 5′. Tomado de [77].

b. Los tripanosomas Africanos

b.1 Variación Antigénica de Superficie

Como fue señalado antes, una de las particularidades de los tripanosomas africanos es que no presentan estado intracelular en el mamífero (son exclusivamente sanguíneos), y por esta razón están expuestos permanentemente al sistema inmune. Muy probablemente como consecuencia de lo anterior han desarrollado una forma notable y eficiente de evadir la respuesta inmune del mamífero, conocida como variación antigénica. Esta, que es sin duda la característica más distintiva de los tripanosomas africanos, consiste en la expresión secuencial de genes extremadamente variables, los cuales codifican para diferentes copias de glicoproteínas variables de superficie (VSGs). Este familia de proteínas caracterizada bioquímicamente por primera vez por Cross en 1975 [80] se corresponde en la microscopía electrónica con una cubierta electrón-densa en la superficie de las formas infectivas sanguíneas de estos tripanosomas [81]. Esta densa capa homogénea representa una barrera prácticamente infranqueable para los anticuerpos [82].

Se trata de una glicoproteína de 58 kDa, con un anclaje GPI (Glico-fosfatidil-inositol), con estructuras predominantemente de tipo α-hélices y con secuencias altamente variables en el N-terminal que se expresa monoalélicamente [83].

Es decir que la cubierta de superficie está compuesta por una sola especie de VSG, conteniendo alrededor de 5 millones de homodímeros idénticos de esta glicoproteína. Se expresa un único gen VSG por vez a partir de un repertorio estimado en más de 1500 genes diferentes [32]. Estos parásitos logran evadir la respuesta inmune cambiando periódicamente el gen VSG que es expresado, siendo este un evento que ocurre con una frecuencia de 10^{-3} por célula [84]. Sin embargo, varias evidencias indican que el patrón típico observado en la variación antigénica (es decir, la presencia de ondas de parasitemia en el hospedero mamífero, ver Figura 19) no puede ser únicamente explicada por la interacción entre los antígenos producidos y la respuesta inmune generada contra ellos. Sería necesario que existan mecanismos propios de control poblacional (similar al *quorum sensing*, ya descritos para bacterias), que implicara pasar de la forma replicativa (*slender*) a la forma quiescente (*stumpy*) [19, 85].

El sistema genético de la variación antigénica consiste en el ya mencionado repertorio de copias silenciosas y el sitio de expresión (ES). Este último de localización telomérica, contiene a la copia de VSG que está siendo expresada por un parásito en particular. En el caso de *T. brucei* el repertorio de copias silenciosas (llamadas también copias

básicas) se encuentra formando *clusters* intracromosómicos que contienen un número variable de genes distintos. En esta especie gran parte del repertorio lo componen fragmentos génicos o pseudogenes, la mayoría de los cuales presentan o bien codones de terminación (codones *stop*) "*in frame*", o *indels* (inserciones o deleciones) que cambian el marco de lectura. De hecho sólo el 7% de las copias silenciosas de VSGs codifican para proteínas totalmente funcionales [32].

Llama la atención que las distintas proteínas de VSG presentan una identidad aminoacídica muy baja, inferior al 20%, a pesar de lo cual su estructura tridimensional es sorprendentemente similar, como lo revela el análisis mediante cristalografía de alta resolución de dos variantes de VSG, MITat1.2 e ILTat1.24, que presentan divergencia extrema a nivel de sus estructuras primarias [82, 86].

Otro de los roles importantes de estas proteínas relacionados con la evasión del sistema inmune es el de "limpieza de anticuerpos" (*antibody clearence* en inglés), ya que las VSG de la superficie están siendo continuamente recicladas. Cuando un anticuerpo específico se une a estas proteínas, la dinámica de movimiento del parásito lleva a estos complejos hacia el polo posterior desde dónde son internalizadas para su reciclaje (en el bolsillo flagelar). Esta internalización provoca la degradación del anticuerpo (vía lisosoma) y el reciclaje de la proteína VSG [83].

b.2 *Trypanosoma vivax*

El parásito *T. vivax* es un tripanosoma Africano (Salivaria) que pertenece al subgénero Dutonella. Las principales características morfológicas de las formas sanguíneas de este parásito incluyen la disposición terminal del kinetoplasto, un desarrollo medio de la membrana ondulante y la presencia de un flagelo libre [13] como se muestra en la figura siguiente (Figura 13).

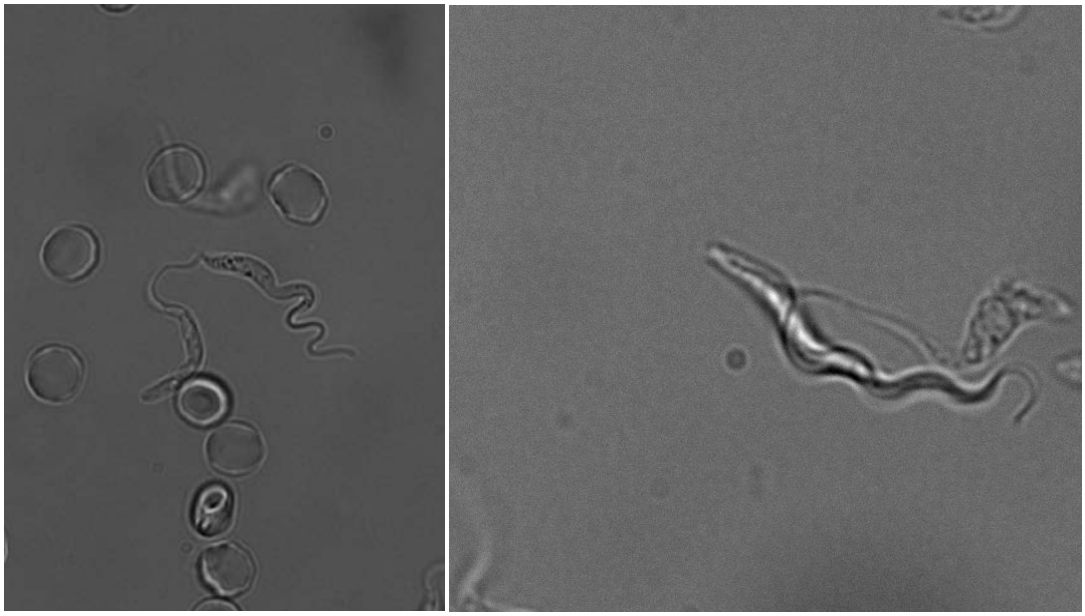


Figura 13. Observación microscópica de sangre de un ratón C57BL/6 infectado con la capa Y486 de *T. vivax.* Fotografías tomadas con microscopio Olympus IX81 en dos aumentos. Unidad de Microscopía, Institut Pasteur Montevideo).

A pesar de su importancia como patógeno animal, *T. vivax* no ha sido estudiado en profundidad, siendo una de las razones la dificultad de cultivarlo en el laboratorio. A diferencia de *T. brucei*, *T. vivax* no infecta, en general, ratones de laboratorio aunque puede ser adaptado. Sin embargo las cepas de *T. vivax* Y486, Y58 y V953 (aisladas de ganado infectado en Nigeria) fueron aisladas directamente en ratones [87]. Estas cepas alcanzan altos picos de parasitemia en ratones de laboratorio y han sido extensamente utilizadas.

La cepa mejor caracterizada de las tres, que cuenta con el genoma secuenciado es la cepa Y486. Recientemente D'Archivio y colaboradores [88] lograron sistemas de cultivo y diferenciación *in vitro* generando una poderosa herramienta para continuar los estudios de caracterización de este modelo.

Respecto a las patologías producidas por este parásito, la principal manifestación de la enfermedad en el ganado (conocida como Nagana), en su fase aguda, es una gran anemia y compromiso de la función cardíaca, seguida de la invasión al sistema nervioso central, abortos, daños a nivel testicular y de ovarios. En general los aislados de África occidental son más patogénicos que los provenientes del oriente del continente. Brotes altamente mortales han sido reportados en ganado vacuno, cabras, ovejas y caballos en regiones no endémicas de Brasil [12].

b.2.1 Posición evolutiva de Trypanosoma vivax

La filogenia de los kinetoplástidos (y en particular del genero *Trypanosoma*) ha sido un tema controversial en algunos aspectos. En ausencia de evidencia paleontológica, la historia evolutiva de los kinetoplástidos fue reconstruida a través de estudios comparativos usando varias fuentes informativas: morfología, ciclo de vida y distribución de sus hospederos.

Los primeros estudios de filogenia molecular utilizando genes ribosomales sugirieron que los tripanosomas constituían un grupo parafilético. Sin embargo, posteriormente se demostró que estos resultados contenían sesgos sistemáticos causados por las peculiaridades del ARNr, y el posterior análisis de genes codificantes de proteínas permitió inferir que se trata de un grupo monofilético [89].

La mayor parte de las especies del género que fueron estudiadas por métodos moleculares se agrupan en un número pequeño de clados, correlacionados con factores como el taxón del hospedero, su ecología y especialmente el taxón del vector [90].

De acuerdo a lo postulado por Hoare, C. [91], *T. vivax* representa un fósil viviente dentro de los tripanosomas africanos (Salivaria), donde observamos características que permanecen detenidas en determinadas etapas de su adaptación a la mosca *tse-tsé* (su hospedero intermediario) [91].

Un ejemplo, respecto al estado ancestral de *T. vivax*, lo constituye el hecho de que este parásito desarrolla su ciclo enteramente en la probóscide y no sobrevive en el intestino del vector (a diferencia de *T. brucei* que puede sobrevivir en el intestino e invade las glándulas salivales), mostrando una etapa inicial de la evolución de estos parásitos que comenzaron no sólo a sobrevivir en la probóscide, sino a desarrollar parte de su ciclo en el vector [91]. Más recientemente, Haag y colaboradores [92], apoyaron con datos moleculares la divergencia temprana de *T. vivax* dentro de los Salivaria.

Es por esto que decidimos tomar como modelo de estudio el parásito *Trypanosoma vivax*. Como se ha dicho, se trata de un tripanosoma de origen africano que se ha extendido por América. Por otra parte, como mencionamos en el párrafo anterior, la ubicación filogenética de *T. vivax* (ver Figura 14), en la base del árbol de los tripanosomas africanos (o sea, habiéndose escindido del tronco principal antes que el resto de las especies), lo transforman en un modelo clave para entender varios fenómenos y procesos de esta familia de tripanosomátidos y cómo se encontraban en su estado ancestral. Un ejemplo, que abordamos en este trabajo, lo constituye el estudio de la variación antigénica de superficie en estos parásitos, que probablemente representa una etapa inicial en el desarrollo del mismo, como veremos en los resultados de esta tesis.

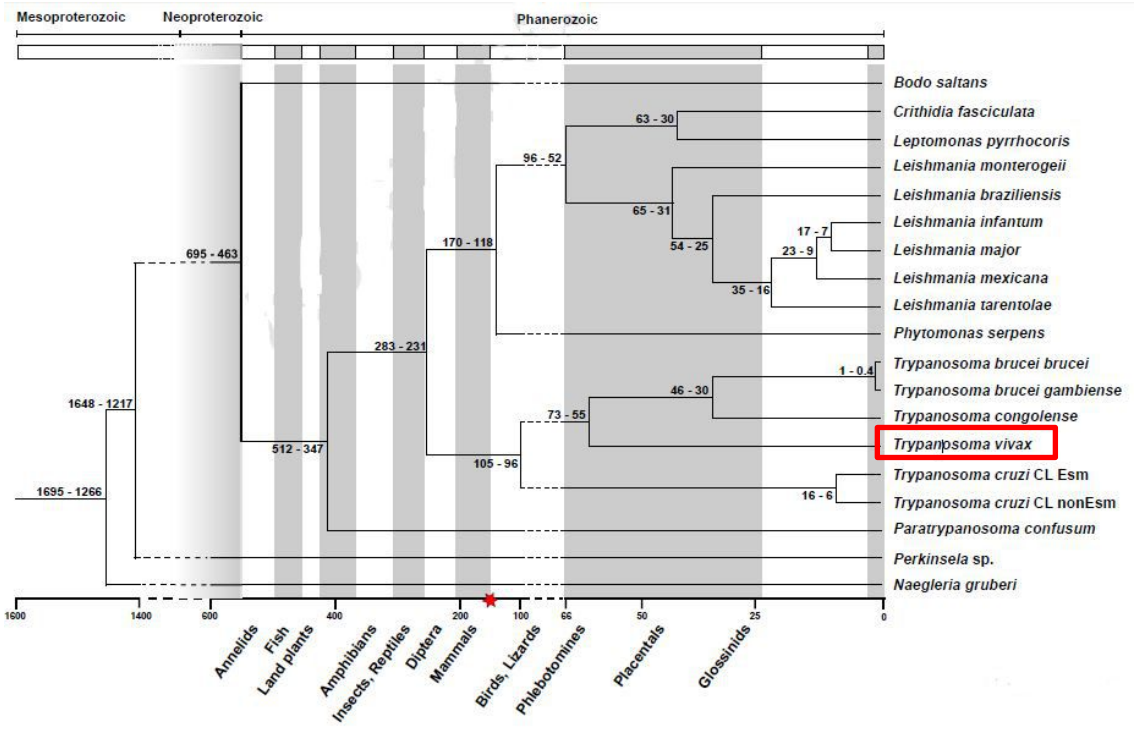


Figura 14. Árbol filogenético de los kinetoplastidos flagelados. El árbol fue construido utilizando un set de datos de 42 proteínas concatenadas de 18 especies de kinetoplástidos y utilizando *Naegleria* como grupo externo. En el recuadro rojo se muestra la ubicación de *Trypanosoma vivax*. Modificado de [93].

b.2.2 La introducción en América

De las tres especies principales de tripanosomas africanos trasmitidos por la mosca *tse-tsé* [94] que afectan a los rumiantes en África Sub-Sahariana, *T. vivax* y *T. evansi* se han dispersado a zonas dónde no está presente el insecto vector, y son trasmitidos de forma mecánica tanto en África como en América del Sur [95, 96].

La introducción en América no puede datarse exactamente, aunque se sugiere que la vía de ingreso fueron ganados contaminados provenientes de África occidental introducidos a partir de la colonización 500 años atrás y que continuaron en los siglos siguientes siguiendo las rutas del tráfico de esclavos desde África [12]. El ingreso de cabras, ovejas, vacas y equinos por los colonizadores pudo ser la vía de ingreso de *T. vivax* en América y pudo haber ocurrido en diferentes momentos y lugares [12].

En América fue descrito por primera vez en la Guyana Francesa por Leger y Viene en 1919 [97] y nombrado como *Trypanosoma guyanense* (luego se mostró que se trataba de *T. vivax* [98]).

Sin la presencia del insecto vector en América, la expansión de los parásitos se debió a la transmisión mecánica mediada por varias especies de moscas hematófagas (Tabánidos y Stomoxys), las cuales funcionan como vectores. Asimismo se ha descrito la transmisión por medio de otros insectos hematófagos, vampiros [22, 23] y jeringas contaminadas [24]. Hoy afectan áreas dedicadas a la ganadería, desde México hasta Paraguay [13]. La Nagana ha sido recientemente reportada incluso en el sur de Rio Grande do Sul, a sólo pocos kilómetros de la frontera con Uruguay [99], lo que indica que el riesgo de que nuestro país se vea afectado en el futuro próximo por esta enfermedad es alto.

En la Figura 15 se representa la distribución del parásito a nivel mundial y la distribución de los diferentes genotipos que muestran que las variantes americanas son más cercanas genéticamente a las cepas provenientes de África occidental.

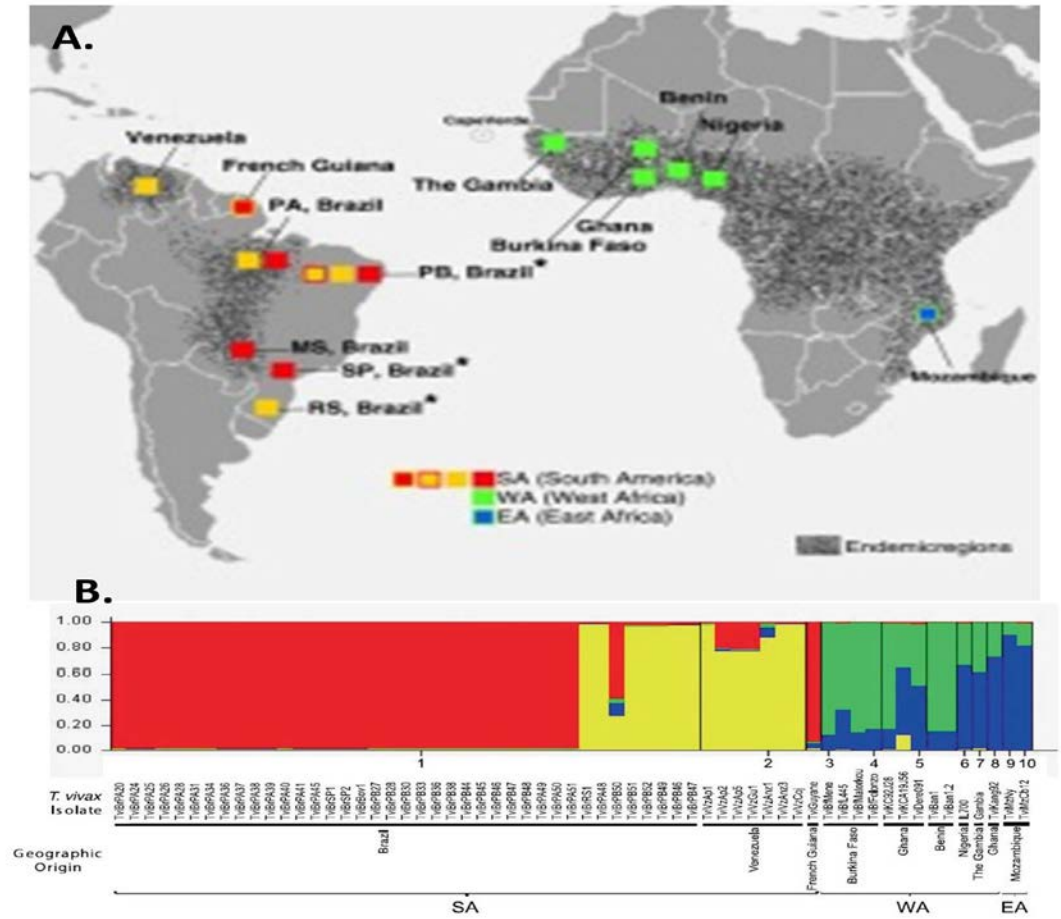


Figura 15. **Distribución geográfica de los diferentes genotipos de *Trypanosoma vivax* a nivel mundial**. **A.** En colores se indican los diferentes aislados analizados en el trabajo de García y colaboradores. Los puntos negros indican zonas endémicas. **B.** Relación entre los microsatélites analizados de los diferentes aislados estudiados. Tomado de [12].

Justificación

Los tripanosomátidos son organismos que presentan características particulares de gran interés. Específicamente, los tripanosomas africanos presentan un sistema de evasión de la respuesta inmune único en la naturaleza. Asimismo debido a sus complejos ciclos de vida, alternando entre dos hospederos, presentan grandes cambios a nivel metabólico, y particularmente en la actividad mitocondrial.

Es, en este contexto que decidimos trabajar con *Trypanosoma vivax,* una especie de gran potencial como modelo de estudio de los tripanosomas africanos, debido a su ubicación evolutiva en la rama más ancestral de estos parásitos. Este hecho nos permite estudiar e intentar dilucidar diversos procesos y mecanismos de interés en su estado ancestral, más precisamente el mecanismo de variación antigénica y los procesos de regulación de la expresión génica.

Por otra parte, su introducción en América, donde logró expandirse sin la presencia del insecto vector (independiente de la transmisión cíclica) usando otros vectores que actúan meramente como agentes mecánicos, resulta de gran importancia para investigar los procesos adaptativos y evolutivos que tuvieron como respuesta a este cambio del modo de transmisión. Además, este acontecimiento ocurrió en un período de tiempo relativamente corto, si consideramos que el ingreso de estos parásitos en América no lleva más de 500 años.

En particular, en este trabajo nos planteamos el estudio comparativo del genoma mitocondrial en la cepa africana Y486 y dos aislados americanos: MT1 y Liem-176 a los efectos de estudiar posibles cambios de nivel genómico en este proceso de adaptación a la transmisión mecánica. Debido al rol cambiante de la mitocondria durante el ciclo de vida de los parásitos en correlación principalmente con las necesidades energéticas en cada caso (altamente funcional en los parásitos presentes en el insecto vector natural y con una funcionalidad basal en la etapa sanguínea en el mamífero).

En suma, planteamos abordar el estudio de diversos aspectos de la biología de *T. vivax* debido a la importancia de este parásito como modelo en la aparición de diversas características de los tripanosomas africanos. Esto se realizará mediante análisis genómicos y transcriptómicos comparativos de diferentes cepas del parásito *Trypanosoma vivax* y otros parásitos relacionados. Se pretende aportar elementos para, por un lado elucidar el mecanismo de variación antigénica en su estado ancestral aportando información sobre la evolución de este sistema, y por otra parte estudiar si la pérdida del ciclo completo en los parásitos aislados en América fue acompañada de cambios a nivel del genoma mitocondrial.

Objetivos generales y específicos

A continuación se plantean los objetivos generales y específicos de este trabajo:

1- Realizar estudios sobre el transcriptoma de *Trypanosoma vivax* para contribuir al conocimiento de la genómica funcional y evolutiva de los tripanosomas africanos, con especial énfasis en los mecanismos de evasión de la respuesta inmune mediada por VSGs. Para ello se plantearon los siguientes objetivos específicos:

a. Obtención del transcriptoma completo de la cepa americana Liem-176 de *T. vivax*.
b. Anotación de genes encontrados.
c. Reconstrucción *in silico* del estado metabólico (reconstrucción de vías en actividad).
d. Cuantificación de transcriptos.
e. Búsqueda de mecanismos de regulación génica -por ejemplo, evidencias de degradación diferencial transcriptos, transcripción de regiones no codificantes de proteínas-.
f. Búsqueda de genes VSG que se expresen y análisis de expresión de proteínas de la superficie celular.

2- Comparación del genoma mitocondrial de diferentes cepas de *T. vivax* con el fin de realizar un análisis comparativo que permita investigar los cambios producidos durante el proceso de adaptación a la transmisión mecánica en las cepas americanas. En este sentido nos planteamos los siguientes objetivos específicos:

a. Determinar la secuencia completa de los genomas mitocondriales: maxicírculos y minicírculos de cepas americanas y africanas de *T. vivax*.
b. Comparación de los genomas mitocondriales secuenciados.
c. Análisis de mecanismos de *editing* mitocondrial mediante la comparación de los datos genómicos y transcriptómicos mitocondriales.

Materiales y Métodos

Infección experimental y purificación de parásitos.

Para este trabajo se utilizaron parásitos *T. vivax* de las cepas americanas (Liem-176 y MT1, aislados naturales de bovinos venezolanos) y la cepa africana Y486 (cedida por Philippe Büscher del Institute of Tropical Medicine Antwerp de Bélgica).

En el caso de las cepas americanas, se procedió a la infección de ovejas inmuosuprimidas a través de la inoculación intravenosa de sangre criopreservada de ovinos infectados conteniendo parásitos. Se realizó el conteo parasitario, hasta que se alcanzó una parasitemia de $2x10^7$ parásitos/mL, se extrajo sangre y se realizó la purificación de los parásitos según se detalla en [100]. Las infecciones en ovinos fueron realizadas en la Universidad Simón Bolívar (Caracas, Venezuela) bajo supervisión veterinaria, se realizó control diario de temperatura y hematocrito (el cual nunca descendió por debajo del 30%).

Los parásitos de la cepa Y486 fueron propagados en ratones C57BL/6 de acuerdo a Chamond y colaboradores [101]. Brevemente, ratones de la cepa C57BL/6, de 7 a 10 semanas de edad, fueron inoculados intraperitonealmente con 100 µL de sangre de ratón (criopreservada) conteniendo aproximadamente 10^6 parásitos. Se realizó el conteo de parásitos cada 2-4 días post-inyección a partir de sangrado de 5 µL de sangre del seno submandibular. En el pico de parasitemia (10^8-10^9 parásitos/mL) se procedió al sangrado final y eutanasia de los animales (Protocolo CEUA 013-11). Estos experimentos fueron llevados a cabo en el bioterio de la Unidad de Animales Transgénicos y de Experimentación del Institut Pasteur de Montevideo. Los parásitos fueron purificados a partir de la sangre mediante centrifugación durante 5 minutos a 300 g, luego de 5 minutos se observa enriquecimiento en el sobrenadante de parásitos viables.

Purificación de ácidos nucleicos.

El ARN total fue aislado a partir de parásitos purificados utilizando columnas Illustra RNAspin Mini Kit (GE Healthcare) o Tri-Reagent (Invitrogen) seguido de purificación de ARN total con las columnas Direct-Zol RNA™ Miniprep kit (Zymo Research). El ARN fue cuantificado espectrofotométricamente (Nanodrop) y la calidad chequeada mediante Bioanalyzer (Agilent).

El ADN fue extraído a partir de parásitos purificados utilizando el kit Quick-gDNA™ MiniPrep kit (Zymo Research). El ADN fue cuantificado espectrofotométricamente y chequeada su calidad mediante visualización en gel de agarosa 0,8% teñidos con Bromuro de Etidio.

Construcción de bibliotecas y secuenciación

Para la realización de las bibliotecas de RNAseq, se realizó ADNc a partir de 1-5 µg de ARN total con la enzima SuperScript II (Invitrogen), utilizando el cebador 5'-CTGGAG(T)16VN-3' para la síntesis de la primera hebra. Luego fue sintetizada la segunda hebra utilizando el kit Double-Stranded cDNA Synthesis Kit (Invitrogen) de acuerdo a las instrucciones del fabricante. Por último, se realizó la digestión con la enzima *GsuI* (Fermentas) que permite el corte de las colas poliA (secuencia de corte enzimático incluida en el cebador de síntesis de la primera hebra).

Las bibliotecas para la secuenciación 454 se realizaron con el kit GS Titanium DNA Library preparation kit (Roche) de acuerdo a las instrucciones del fabricante. La secuenciación se realizó en Life Sequencing S.L, Biopolis (Valencia, España).

La secuenciación de Illumina para los experimentos de RNAseq se realizó en la Universidad de Washington, en un equipo GaIIx (Illumina), a partir de las mismas bibliotecas de ADNc, las cuales fueron re-fragmentadas y se le agregaron los adaptadores universales de Illumina.

Por último, las secuenciaciones de ADN genómico y RNAseq de la cepa africana, fueron realizadas en la Unidad de Biología Molecular del Insitut Pasteur Montevideo, en los secuenciadores MiSeq (Illumina) y GaIIx (Illumina). Las bibliotecas se realizaron con los kit Nextera o Nextera XT (Illumina) a partir de ADN total o ADNc (en el caso de las bibliotecas de RNAseq).

PCR y Secuenciación Sanger.

Los cebadores y condiciones utilizadas en el PCR y secuenciación Sanger para el ensamblado final del maxicírculo, la confirmación de minicírculos y confirmación del gen VSG expresado se encuentran en el material suplementario de [102] y [103] (Figura suplementaria 3 de Anexo 1 y Archivo suplementario 1 de Anexo 2).

Análisis bioinformáticos.

Para la realización de los análisis bioinformáticos (ensamblajes, anotación funcional, determinación de niveles de transcripción, etc.) se utilizaron diversos paquetes y programas informáticos que se muestran en la Tabla 3, indicando en cada caso la referencia bibliográfica (si corresponde), y sobre qué clase de datos fue utilizado.

Detalles sobre el uso de los mismos se encuentran en cada sección de resultados y en [102] y [103].

Tabla 3. En la tabla se indican los paquetes bioinformáticos utilizados para el análisis de datos de secuenciado masivo.

Herramienta	Datos	Observaciones	Cita
FastQC	General	Control de calidad de secuenciado profundo	http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
RNAseqQC	RNAseq Liem-176	Control de calidad de ensamblado de datos de RNAseq	[104]
Scythe	DNAseq Liem-176/MT1	Filtrado de secuencias	https://github.com/najoshi/sickle
Mira	RNAseq Liem-176	Ensamblajes de secuencias (454)	[105]
Newbler	RNAseq Liem-176	Ensamblajes de secuencias (454)	[106]
Blast	RNAseq/DNAseq	Control de contaminación, chequeo de datos, etc.	[107]
Bowtie	RNAseq Liem-176	Mapeo de secuencias (RNAseq)	[108]
Bowtie2	DNAseq mitocondrial	Mapeo de secuencias (DNAseq mitocondrial/RNAseq)	[109]
ESTscan	RNAseq Liem-176	Anotación funcional de secuencias	[110]
InterProScan	RNAseq Liem-176	Análisis de dominios funcionales	[111]
AnEnPi	RNAseq Liem-176	Asignación de vías metabólicas	[112]
Blast2GO	RNAseq Liem-176	Asignación de términos GO	[113]
Erange	RNAseq Liem-176	Determinación de niveles de transcripción	[114]
ABySS	DNAseq MT1	Ensamblaje *de novo* del genoma mitocondrial MT1	[115]
Spades	DNAseq Liem-176	Ensamblaje *de novo* del genoma mitocondrial Liem 176	[116]
SAMtools	General	Manejo general de datos	[117]

Resultados y Discusión

Esta sección se presenta en dos partes, la primera de ellas relacionada con los análisis de genómica funcional de la cepa americana Liem-176 de *T. vivax*. Los datos de la parte A de esta sección se encuentran relacionados al Objetivo Nº1.

En la segunda parte de esta sección (Parte B) se presentan los resultados obtenidos respecto al Objetivo general Nº2, relacionados con los cambios observados en el genoma mitocondrial entre las cepas americanas (Liem-176 y MT1) y la africana (Y486) de *T. vivax*.

Parte A. Genómica evolutiva y funcional en *Trypanosoma vivax*

Algunos de los aspectos más interesantes que queremos resaltar en esta sección son los siguientes:

1. Obtención del transcriptoma de la cepa americana Liem-176 de *T. vivax*, su anotación funcional y la creación de una base de datos pública, así como la cuantificación de los transcriptos identificados.

2. Variación antigénica de superficie y análisis comparativo de la composición de la membrana celular de *T. brucei* y *T. vivax*.

3. Patrones de expresión génica:
    a. Identificación de regiones no transcriptas por el parásito.
    b. Uso diferencial de sitios de *trans-splicing* como posible mecanismo de regulación de la traducción de algunos genes.

Análisis transcriptómicos de la cepa *T. vivax* Liem-176.

Ensamblaje y contenido génico.

Se realizó el ensamblaje de secuencias obtenidas a partir de ARN total de parásitos aislados en el primer pico de parasitemia de ovejas infectadas experimentalmente. Estas secuencias fueron obtenidas con la metodología 454 FLX (Roche).

Se obtuvieron varios ensamblajes con distintas herramientas computacionales los cuales fueron comparados, concluyéndose que aquel obtenido con la herramienta Mira [105] era el de mayor calidad. Este ensamblaje contiene 67850 RNAseq *contigs* sobre los cuales se realizó una anotación funcional. Un primer aspecto que se tuvo en cuenta fue la obtención de traducciones virtuales con baja tasa de error. Esto es un aspecto critico en la anotación de transcriptomas, particularmente aquellos obtenidos con la tecnología 454 (Roche). El problema consiste en traducir usando el marco de lectura correcto, pues los *contigs* resultantes del ensamblaje de RNAseq suelen tener algunos "*indels*" cortos, los cuales destruyen la continuidad del marco de lectura abierto (ORF), resultando por tanto en una traducción correcta sólo de segmentos de *contig*. Esta complicación es particularmente importante en secuencias generadas con la tecnología 454, dado que presenta una alta tasa de error en la estimación del largo de homopolímeros [118]. Es por ello que se utilizó la herramienta ESTscan [110], la cual detecta el marco de lectura más probable utilizando propiedades estadísticas (extraídas previamente de un set de secuencias de entrenamiento) y realiza "correcciones" mediante *indels* que permiten restablecer el marco de lectura más probable.

Con esta herramienta se logró la identificación de 13385 *contigs* (transcriptos) que pudieron ser traducidos. De éstos, a 3834 se le asignaron términos de ontología de genes. Además, 7796 transcriptos (o *contigs* ensamblados) presentan homólogos en otras especies de tripanosomátidos (usando como criterio un e-*value* < $1e^{-10}$ de Blastp de las secuencias traducidas virtualmente), lográndose también para éstos la anotación funcional mediante transferencia de anotación por homología. Realizamos también la reconstrucción de las vías metabólicas del parásito usando la herramienta AnEnPi [112].

Los datos tanto crudos como los ensamblajes, así como la anotación funcional (términos de ontologías, hits de blast con sus respectivos alineamientos) y las vías metabólicas se encuentran disponibles en la base de datos pública desarrollada por Matías Rodríguez y Miguel Ponce de León que se encuentra disponible en la siguiente dirección web: http://bioinformatica.fcien.edu.uy/Tvivax/. Los mismos también han sido depositados en el repositorio SRA del NCBI (ACC Nº SRX170781).

Otro aspecto a resaltar es que cerca de 1000 *contigs*, transcriptos en diferentes niveles, no presentan homólogos en otras especies de tripanosomátidos (*Leishmania spp*., *T. cruzi* o *T. brucei*), es decir se trata de secuencias especie específicas. A partir de 570 de estos *contigs*, para los cuales fue posible obtener la secuencia completa del transcripto, realizamos una caracterización primaria utilizando una batería de herramientas informáticas para identificar señales de localización subcelular, así como análisis de dominios. En la siguiente tabla (Tabla 4) se resumen las principales características de estos transcriptos.

Tabla 4. Principales características de los *contigs* identificados como especie-específicos en la cepa americana Liem-176. TMH: Dominio transmembrana, GPI: Anclaje Glicofosfatidil-inositol. Signal P: secuencias que presentan péptido señal P.

***Contigs* (transcriptos) especie-específicos**	TMH+ GPI +Signal P	TMH+GPI	TMH+ Signal P	GPI + Signal P	Solo TMH	Solo GPI	Solo Signal P	No TMH+ GPI+ SignalP	**Total**
Presente en Genoma Y486	2	14	20	9	110	3	36	335	529
De origen Americano (no presente en genoma de referencia Y486)	0	2	1	0	19	0	0	19	41
Total	2	16	21	9	129	3	36	354	570

Un detalle de estos transcriptos se encuentra en el material suplementario de [103] y parte de la tabla se muestra en el Anexo 1 (Tabla Suplementaria 4).

Además, dentro de este grupo encontramos 41 contigs que no se encuentran en el genoma de *T. vivax* disponible (cepa Y486, genbank AC Nº CAEX00000000.1), indicando que podría tratarse de genes específicos de la cepa americana Liem-176 (Tabla 4). Posteriormente a la publicación de estos datos, Jackson y colaboradores [119] presentaron datos confirmatorios sobre la presencia de genes especie-específicos sobre todo relacionados a la membrana celular.

En análisis más recientes, utilizando el genoma de cepas americanas de *T. vivax*, pudimos corroborar estos observaciones iniciales: cerca de un 10% de los genes de la cepa africana Y486 no están presentes en las cepas americanas, y alrededor de 200 genes presentes en las cepas americanas no se encuentran en la cepa africana (manuscrito en preparación).

Cuantificación de transcriptos.

Para tener una mejor aproximación al cálculo de los niveles de transcripción, se realizó una secuenciación con la tecnología Illumina para obtener una mayor profundidad de secuenciado. La correlación entre los datos de cuantificación de las dos tecnologías utilizadas es alta (R^2=0,83) como se observa en la siguiente figura (Figura 16).

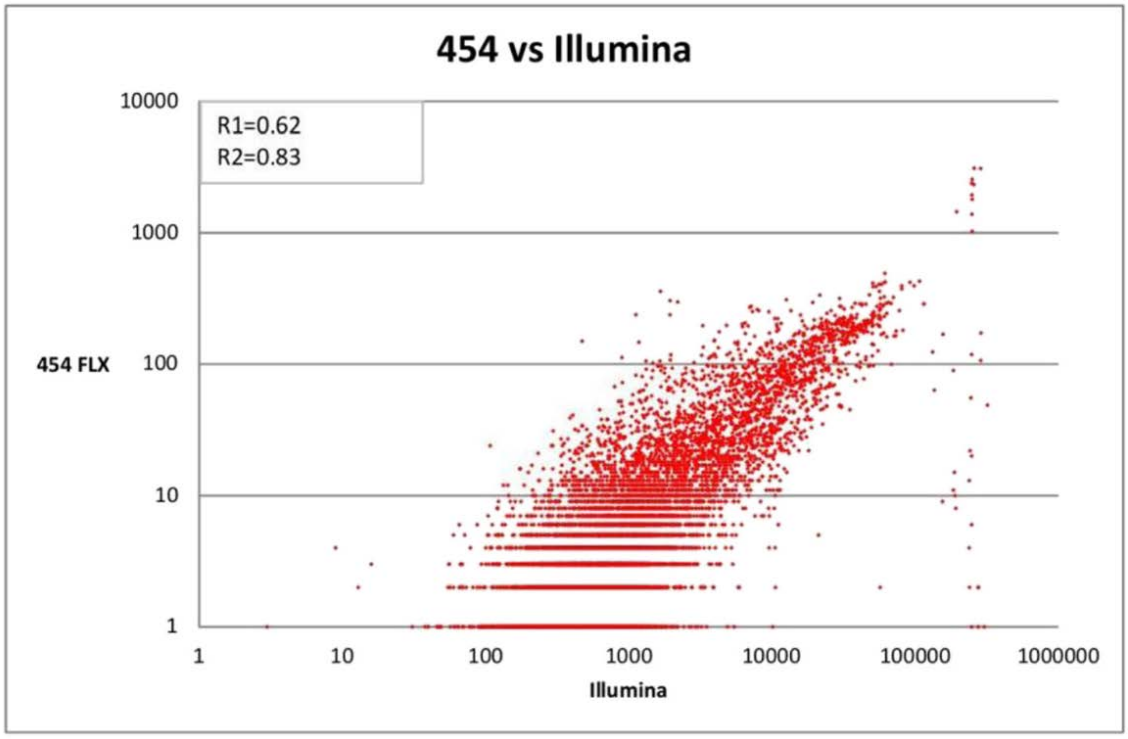


Figura 16. Correlación entre los datos de cuantificación de transcriptos con la tecnología Illumina y 454. En la figura se grafican los resultados de cuantificación para cada *contig* (punto rojo) para ambas tecnologías (Illumina y 454). R1 y R2 son los coeficientes de correlación antes (R1) y después (R2) de eliminar los puntos con una discrepancia extrema entre las dos tecnologías (en general sobre-representados con la tecnología Illumina).

Para realizar la cuantificación de los transcriptos identificados se realizó la adaptación del paquete informático Erange [114] para los requerimientos de las características genómicas de nuestro modelo. En el caso de estos parásitos que presentan muchos genes repetidos en tándem o dispersos en el genoma, la aproximación de Erange es la más adecuada ya que permite asignar aquellos *reads* (secuencias) que mapean en más de un lugar del genoma a una posición determinada (Anexo 1, Tabla Suplemenetaria 3). Para usar esta herramienta fue necesario el desarrollarlo de scripts en *python* y *perl* para modificar los archivos de anotación, incluyendo las regiones 5' y 3' UTR para una mejor aproximación al cálculo de los valores de transcripción de cada gen.

Variación antigénica de superficie y composición de la membrana celular en *T. vivax.*

Respecto al estudio de la variación antigénica y la composición de la membrana celular, se lograron obtener resultados sumamente interesantes. El primer paso para comenzar el estudio del patrón de variación antigénica de esta especie consistió en la identificación del gen VSG expresado por estos parásitos. Debido a las restricciones de concentración de ARN para realizar los experimentos de secuenciado en 454, se partió de una mezcla de ARN total obtenido de diferentes infecciones (correspondientes al primer pico de parasitemia de animales infectados con un mismo stock de parásitos).

Para la identificación del gen VSG activo, se realizó primero la búsqueda de candidatos putativos dentro de los genes de mayor expresión (debido a que el ARNm de los genes VSG representa en el caso de *T. brucei* entre el 5% y 10% del ARNm total).

Sobre los *contigs* candidatos, se realizó un BLAST contra el total de secuencias depositadas en genbank (nr). La secuencia identificada mostró una identidad elevada (90,4%) a la primer secuencia de VSG de *T. vivax* reportada por Gardiner y colaboradores [120] correspondiente a un aislado de África occidental y denominada Ildat 2.1 (ver Figura Suplementaria 2, Anexo 1). Resulta interesante el alto nivel de similitud entre la secuencia identificada y la secuencia reportada previamente. Este alto nivel de identidad, considerando la divergencia entre los aislados americanos y africanos, es mucho mayor al esperado para genes que normalmente divergen rápido [121].

Por un lado, este resultado confirma datos previos respecto al origen de las especies americanas de *T. vivax* más relacionadas con las especies provenientes de África occidental [12, 122]. Por otra parte, resultó llamativo que el gen identificado en nuestra cepa no se encuentra en el genoma de la cepa de laboratorio Y486 (también proveniente de África occidental).

Para confirmar la ausencia de este gen VSG en la cepa africana Y486 y confirmar la presencia en el aislado americano, se sintetizaron cebadores específicos y se realizó la amplificación por PCR utilizando como molde el ADN genómico del aislado americano (Liem-176) y la cepa africana (Y486) (Figura 17 y Figura Suplementaria 3, Anexo 1).

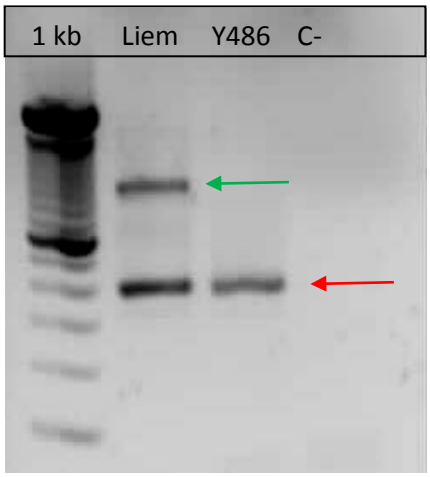


Figura 17. Resultado de amplificación del gen VSG identificado. En la figura se muestra un gel de agarosa teñido con bromuro de etidio con la amplificación de una región común a ambos genomas (Liem-176 e Y486, flecha roja) y la amplificación del gen VSG identificado, únicamente presente en el aislado americano (flecha verde). 1 kb, marcador de peso molecular (1Kb DNA Ladder, Invitrogen). C-. Control negativo de amplificación.

En la Figura 17 se observa la ausencia de amplificación en el genoma de la cepa africana Y486 del gen VSG identificado en la cepa americana. Como control positivo de amplificación se utilizaron cebadores diseñados contra una región común a ambos genomas. Este resultado indica que la ausencia de este gen en el genoma de la cepa africana no se debe a un problema de ensamblaje genómico. Asimismo, el gen que expresa la variante de VSG expresada por la cepa Y486 (Ildat1.2) no se detecta en el transcriptoma de la cepa Liem-176.

Un aspecto a destacar sobre las proteínas VSG en *T. vivax* se relaciona con los niveles de expresión del gen VSG identificado en el transcriptoma en relación a los niveles de expresión de los genes VSG en *T. brucei*. Resulta sorprendente que si bien en *T. vivax* la concentración del ARNm de esta proteína es alta (casi el doble que los niveles de expresión de los genes altamente expresados de alfa y beta tubulina, Tabla Suplementaria 5 Anexo 1), la cantidad de *reads* de Illumina que mapean en dicho gen, no se corresponden con los elevados valores reportados para el gen VSG expresado en el caso de *T. brucei* [60, 61, 123].

Para profundizar en este análisis, nos preguntamos si en *T. vivax* la composición de proteínas de membrana era similar a la reportada en el parásito *T. brucei*. Para ello, se realizó la comparación en los niveles de expresión (abundancia de ARNm) de todos aquellos genes anotados con localización en la superficie celular. En la Figura 18 se muestra este análisis, dónde se desprende que, al menos a nivel de ARNm, la composición de los transcriptos de proteínas de superficie es muy diferente cuando se comparan ambas especies.

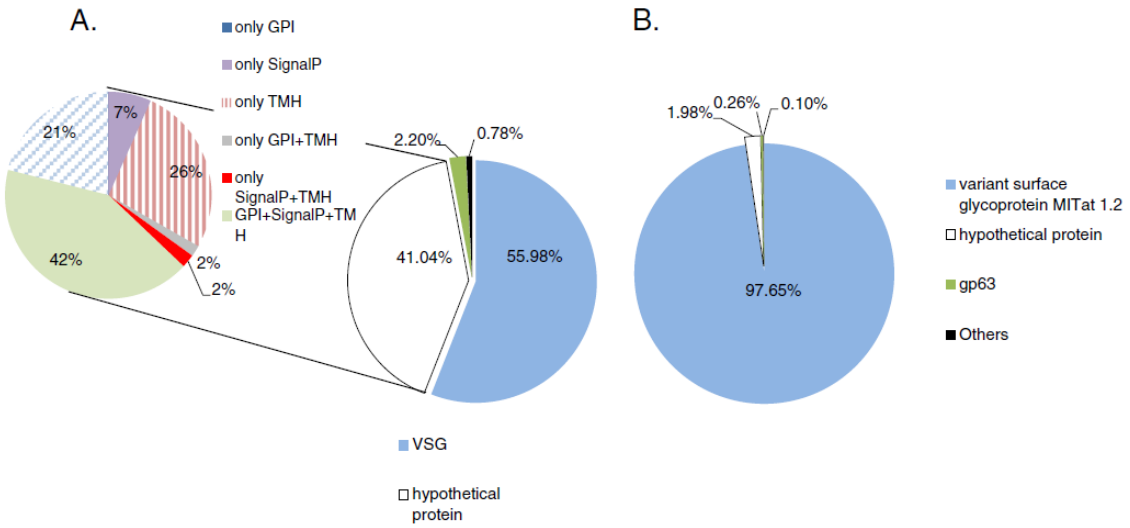


Figura 18. Composición de proteínas de membrana inferida de los niveles de expresión. **A**. Análisis de datos de expresión de transcriptos de proteínas de membrana de *T. vivax*. **B.** Mismo análisis que en A realizado sobre datos de *T. brucei*.

Por un lado en el parásito *T. brucei*, el transcripto del gen VSG es altamente predominante, representando aproximadamente el 98% de las secuencias (*reads*) de genes que codifican proteínas de superficie, sin embargo cuando observamos el parásito *T. vivax*, si bien el gen VSG es predominante no alcanza más del 56% de las secuencias (Figura 18). Estos datos hacen pensar que la composición de las membranas es muy diferente entre estas especies. La abundancia del ARNm codificante de la proteína VSG en *T. vivax* es concordante con datos de microscopía electrónica reportados previamente, que muestran que en *T. vivax* la cubierta de VSG es mucho menos densa que la observada en *T. brucei* [124].

Un primer aspecto de interés está relacionado con los niveles de expresión de los genes VSGs y su eficiencia protectora como se mencionó más arriba. Estas observaciones llevan a plantearnos acerca de cuál sería el rol ancestral de la proteína VSG, dado que con una abundancia comparativamente menor (en relación a *T. brucei*), no podría funcionar eficientemente como cubierta protectora.

Es de destacar, que al igual a lo que sucede en *T. brucei*, la infección provoca picos de parasitemia (relacionadas en el caso de *T. brucei*, al cambio del gen VSG que se expresa) como puede verse en la siguiente figura (Figura 19).

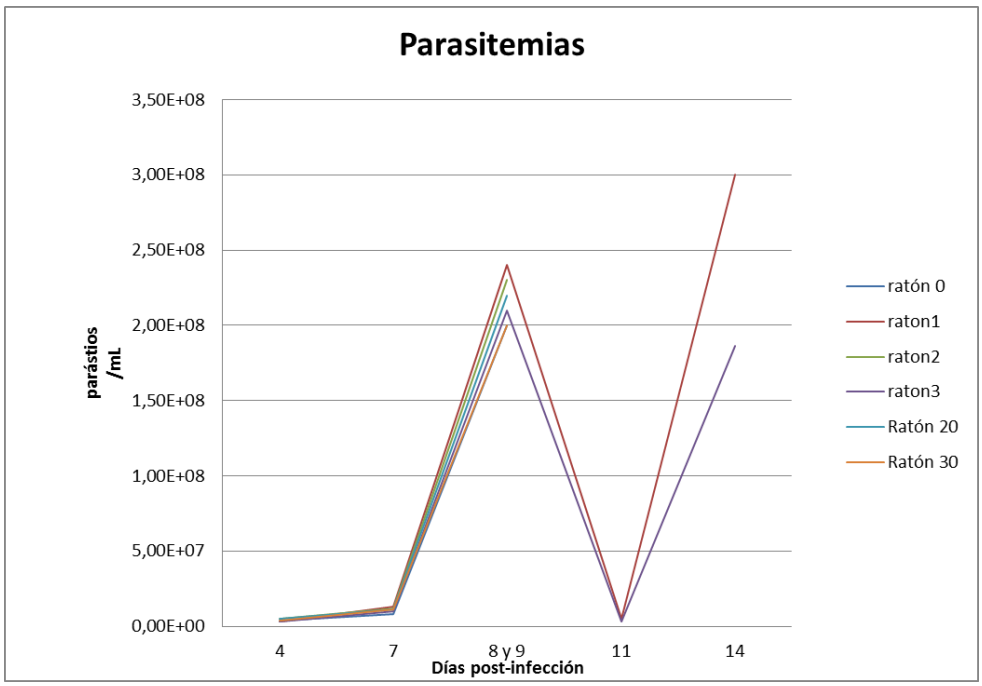


Figura 19. Infección de *T. vivax* Y486 en ratones C57BL/6. En la figura se muestra la cantidad de parásitos/mL de sangre en ratones C57BL/6 infectados con parásitos de la cepa Y486 en diferentes días post-infección.

Resulta interesante entonces, determinar cómo es el mecanismo de evasión de la respuesta inmune en el parásito *T. vivax,* para ello estamos realizando experimentos de RNAseq-DNAseq en los diferentes picos de parasitemia para, por un lado obtener información respecto a la dinámica de expresión de genes VSG y por otro determinar la organización genómica del sitio de expresión del gen VSG activo.

Patrones de expresión génica.

Respecto al último tópico que resaltamos de esta primera sección, sobre los mecanismos de regulación de la expresión génica, inicialmente se realizó la asignación de niveles de expresión en los CDS anotados en genbank de *T. vivax* (11866 CDS) (los datos están disponibles en [103] y una parte se muestra en la Tabla Suplementaria 3 del Anexo 1). Una primera observación muestra regiones genómicas (que contienen CDS) que no presentan mapeo de secuencias (es decir no serían transcriptas por la forma sanguíneas de *T. vivax*) como se observa en la Figura 20.

La ausencia de secuencias que mapeen en determinadas regiones del genoma disponible (Y486) podría deberse a diferencias entre las cepas (recordemos que el transcriptoma analizado corresponde a la cepa Liem-176). Cómo se muestra en la siguiente figura (Figura 20), al menos para las zonas analizadas no parece ser el caso,

ya que se diseñaron cebadores específicos para estas regiones y se mostró la presencia de estas regiones en ambos genomas.

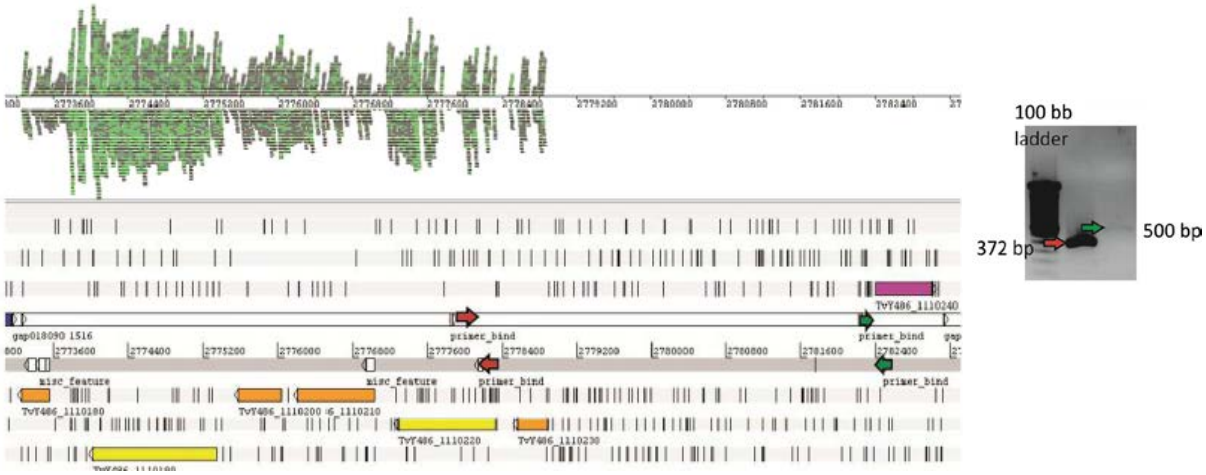


Figura 20. Mapeo de *reads* de Illumina sobre región genómica de *T. vivax*. Se observa una región donde no hay mapeo de *reads*. Las flechas (rojas y verdes) indican las secuencias sobre las que se diseñaron los cebadores. A la derecha se muestra un gel de agarosa 1,5% teñido con bromuro de etidio con los productos de amplificación de los cebadores antes mencionados.

Este resultado sugiere la existencia de algún mecanismo de silenciamiento de algunas regiones enteras del genoma. Si bien para la mayor parte de los genes anotados encontramos *reads* de secuenciación indicando que están siendo transcriptos (en concordancia con la visión aceptada que los tripanosomátidos transcriben la totalidad de sus genes y el control de la expresión es mayormente a nivel post-transcripcional), algunas regiones del genoma con CDS no muestran actividad transcripcional (Figura 20). Este hecho, podría indicar la presencia de mecanismos de control de la expresión génica, por ejemplo mediante el silenciamiento de determinadas regiones. Nuevos experimentos deben ser realizados para confirmar estos datos, por ejemplo determinar si existen regiones genómicas de alta compactación que eviten la transcripción de algunas zonas específicas.

Por último queremos hacer referencia en esta sección al uso diferencial de sitios de *trans-splicing* como posible mecanismo de regulación de la traducción de algunos genes. El primer paso para realizar este estudio consistió en la identificación de *reads* que presentaran la secuencia del *spliced-leader* o miniexon. Se recuperaron 159395 secuencias de Illumina que presentaban el miniexon de *T. vivax*. Con esta estrategia logramos la identificación de los sitios de *trans-splicing* de 5959 genes, de los cuales 3350 presentan un sitio único y 2609 presentaron 2 o más sitios posibles de adición del miniexon.

En la Figura 21 se muestra, a modo de ejemplo la visualización en Artemis para una región genómica dónde se mapean el total de las secuencias de Illumina y las secuencias que presentan el *Spliced-Leader.* En esta figura se observan dos genes, uno de ellos (TvY486_00300750) presenta dos sitios de *trans-splicng* mientras que el segundo (TvY486_00300760) presenta sólo un sitio. Puede apreciarse además que la utilización de los sitios de *splicing* no es homogénea, en el gen TvY486_00300750 el sitio con ubicación 5' presenta una profundidad de mapeo (la cual es proporcional a su utilización) tres veces superior al sitio de *trans-splicing* localizado corriente abajo. Patrones de uso diferencial de sitios de *trans-splicing* se observan en la mayoría de los genes que presentan más de un sitio. Estos resultados son concordantes con observaciones realizadas en *T. brucei* [61].

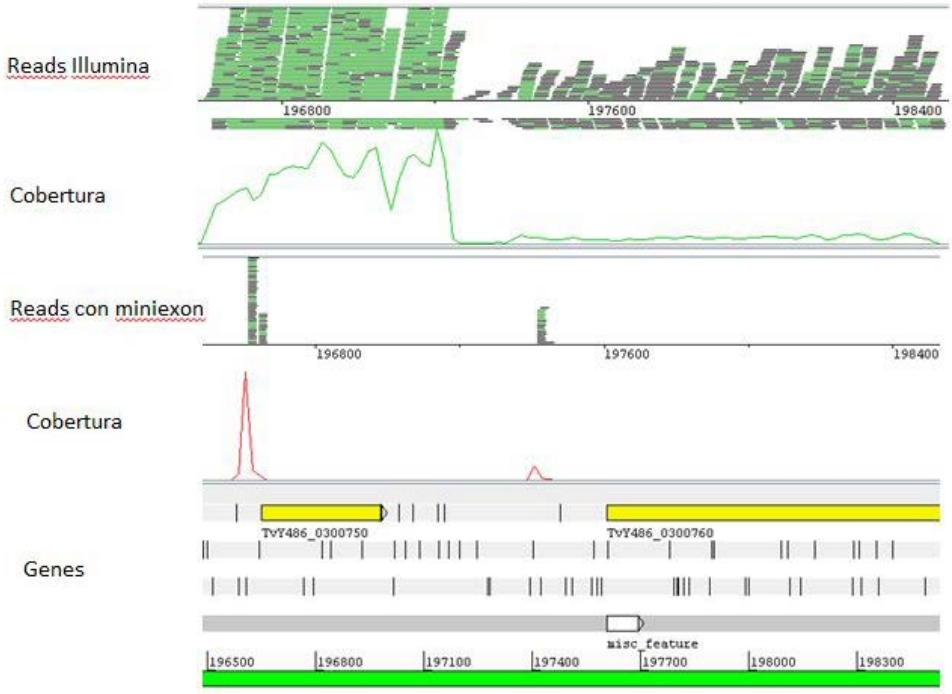


Figura 21. Visualización en Artemis de mapeo de secuencias totales y con miniexon en una región del genoma de *T. vivax*. En la figura se observa, arriba el mapeo de reads totales y su cobertura, en el medio el mapeo de reads con miniexon y el diagrama de cobertura y debajo la región genómica sobre la cual mapean. Se observa como el gen TvY486_0300750 presenta dos sitios de *trans-splicing* y el gen TvY486_0300760 sólo tiene un sitio de adición del miniexon.

En la siguiente figura (Figura 22) se muestra la distribución de los sitios de *splicing* por gen, el número máximo de sitios posibles de *splicing* para un gen resultó en 9 posiciones posibles y el promedio es de 1,48. Este último número es sensiblemente menor al reportado en el genoma de *T. brucei* (2,4-2,9 sitios por gen) [123]. Asimismo se determinó el largo de la región 5'UTR para los genes en los cuales se identificó el sitio de *trans-splicing*, resultando en una media de 132 bases para el primer sitio (siendo considerado el primer sitio aquel que presenta mayor cantidad de secuencias)

y 164 bases para el segundo sitio, que se encuentra en el rango de lo anteriormente reportado para *T. brucei* [60, 61, 123].

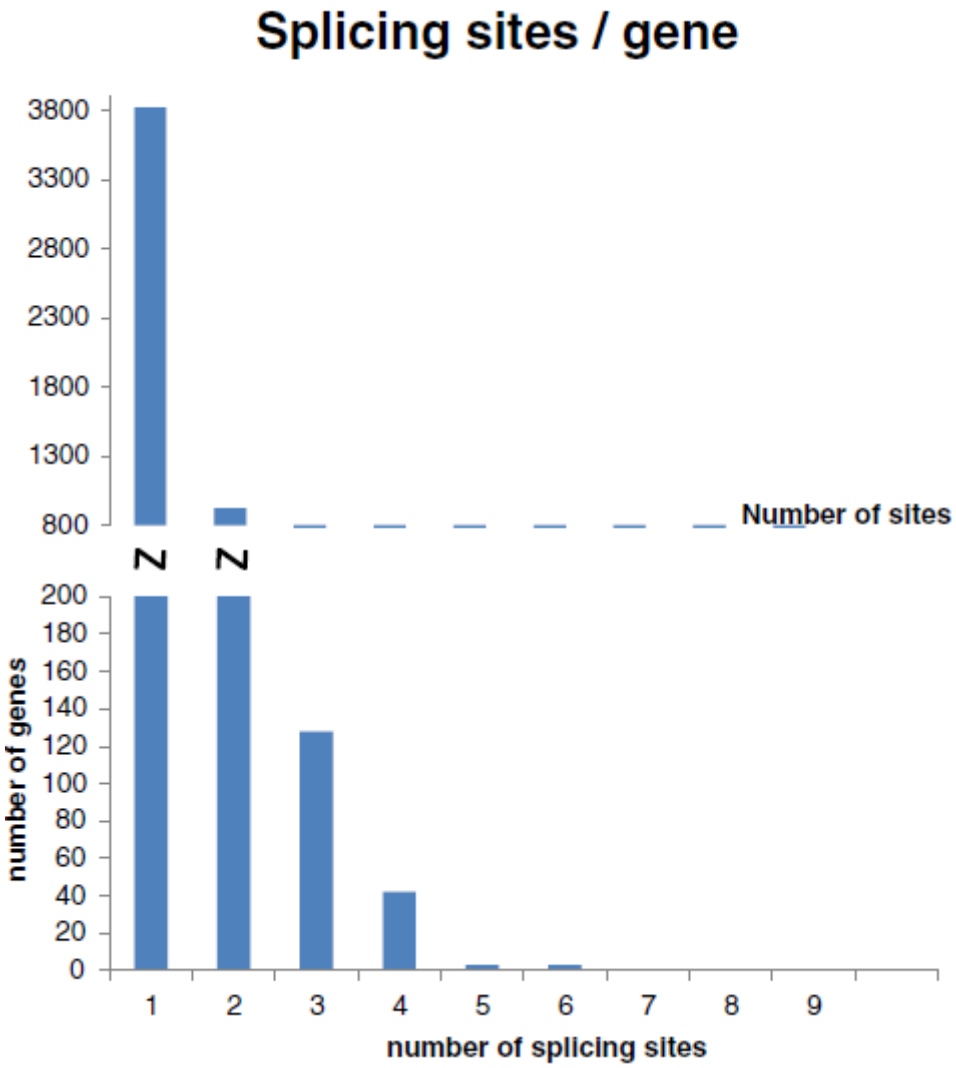


Figura 22. Número de sitios de *trans-splicing* para cada gen. En la gráfica se muestra la cantidad de genes que presentan 1, 2 y hasta 9 sitios de *trans-splicing*.

También se determinaron las secuencias consenso de los sitios de trans-splicing, observándose un patrón muy similar al reportado para *T. brucei* (Figura 23).

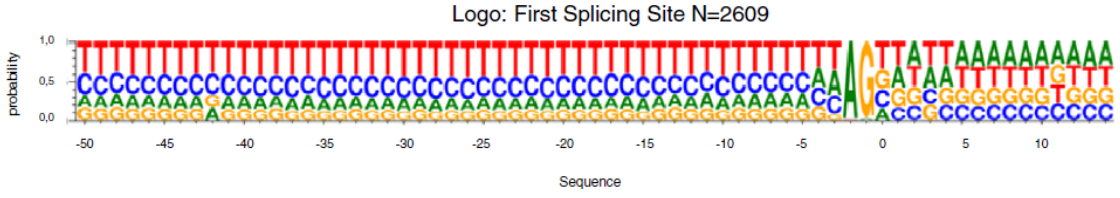


Figura 23. Secuencia consenso de sitios de *trans-splicing* en *T. vivax*. Se muestra la secuencia consenso para el primer sitio de *trans-splicing* (considerando 50 bases antes del sitio de *trans-splicing*, y 15 bases corriente abajo).

Además se realizó la comparación con datos de *Spliced-leader trapping* publicados en el parásito *T. cruzi*, mostrando que el patrón también es muy similar en los parásitos tripanosomátidos americanos (Figura 24).

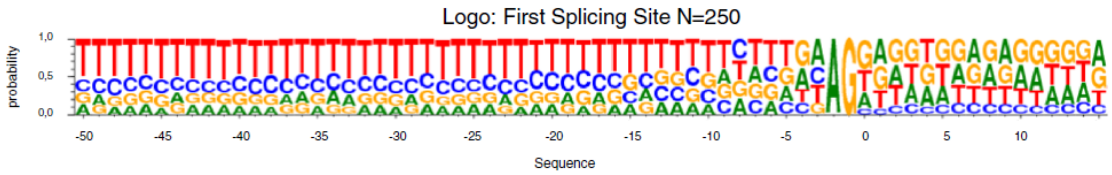


Figura 24. Secuencia consenso de sitios de *trans-splicing* en *T. cruzi*. Se muestra la secuencia consenso para el primer sitio de *trans-splicing* (considerando 50 bases antes del sitio de *trans-splicing*, y 15 bases corriente abajo).

También observamos que no sólo se conservan las secuencias consenso de *trans-splicing*, sino que además se conservan las posiciones y cantidad de sitios de *trans-splicing* al menos entre los parásitos *T. brucei* y *T. vivax* (Figura 25).

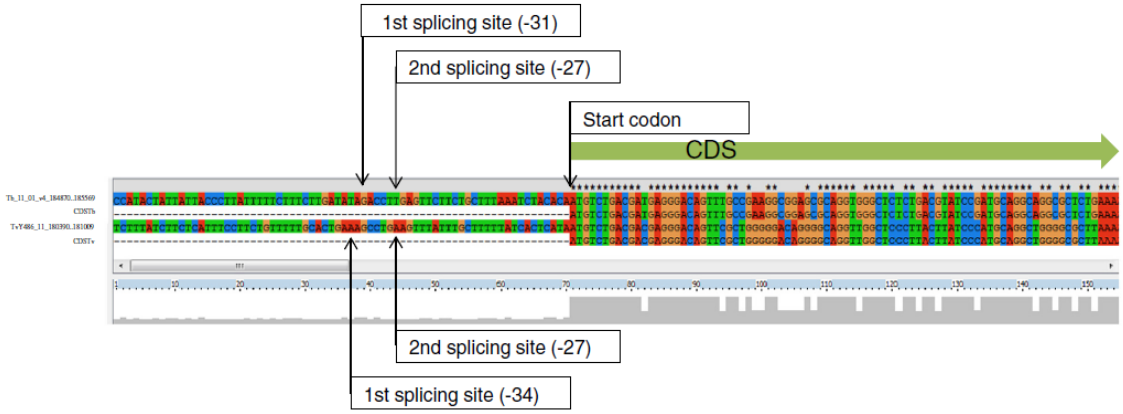


Figura 25. Ejemplo representativo de la conservación de la distancia a los sitios de *trans-splicing* alternativo en *T. brucei* y *T. vivax*. El alineamiento muestra la conservación elevada en la región codificante (CDS) y una baja conservación en la región 5' no traducida (5' UTR). La secuencia superior correponde a *T. brucei* y la inferior a *T. vivax.*

Respecto a este último punto es interesante destacar que aunque las posiciones y cantidad de sitios de *trans-splicing* es conservada entre estas dos especies, la conservación a nivel de secuencia del 5' no traducida (5' UTR) es mínima, lo que implica que el determinante más importante en la localización de los sitios de *trans-splicing* es la distancia hasta el codón de inicio y no la secuencia.

Por último, destacamos que un número importante de genes presentaban el sitio de *trans-splicing* (único o el principal en el caso de tener más de un sitio) muy cercano al codón de inicio de la traducción. Además, muchos de estos genes no presentaban región 5'UTR ya que el sitio de *trans-splicing* estaba inmediatamente antes del codón de inicio de la traducción. Para investigar si estos genes presentaban algún patrón

particular (es decir si codificaban para proteínas que cumplieran roles específicos), se realizó una análisis de enriquecimiento de ontologías (*Gene Ontology enrichment*) utilizando la herramienta Blast2GO [113].

El resultado mostró una sobre-representación de proteínas ribosomales, factores de elongación y otras proteínas relacionadas con la maquinaria traduccional. Asimismo, se encontraron proteínas de respuesta a estrés y relacionadas con la interacción con ARN (Figura 26). También se realizó este análisis para los datos disponibles de *T. brucei* observándose el mismo fenómeno.

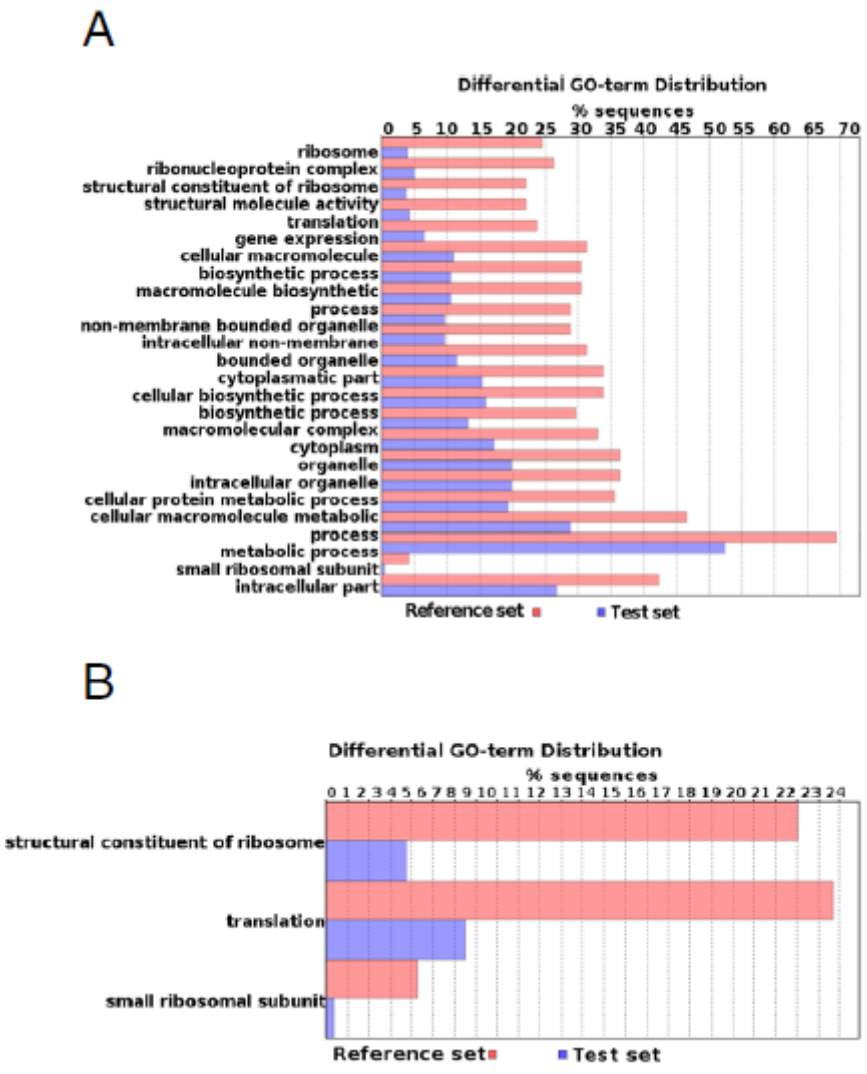


Figura 26. Análisis de ontología de genes. Distribución de términos de GO que exhiben diferencias estadísticamente significativas ($p < 0,05$). Se observa el enriquecimiento de términos de GO en los genes conteniendo sitios de *trans-splicing* cercanos al codón de inicio (distancia <= 10 nucleótidos). El análisis fue realizado a partir de 196 genes. Los paneles A y B, corresponden a dos niveles de GO.

Se ha propuesto, que genes altamente expresados presentan un acortamiento en las regiones 5' y 3' UTR para reducir el costo energético en la síntesis proteica [125]. Sin embargo nuestros datos de niveles de expresión indican que estos genes no son transcriptos a tasas más elevadas que el promedio general del genoma (T-test, $p < 0,05$). Otra alternativa es que este acortamiento sea un indicador de genes que se deben expresar constitutivamente. En este sentido se ha demostrado en *Leishmania* que la presencia del *spliced-leader* asegura el reclutamiento del complejo ribosomal 40S, entonces la ausencia de una región 5'UTR evitaría la unión de posibles proteínas reguladores, es decir, que una vez unido el complejo ribosomal, no sería posible bloquear el inicio de la traducción. Los trabajos recientes de huella ribosomal sobre el transcriptoma de *T. brucei* [53, 54] indican, en el mismo sentido, como el largo de las regiones 5'UTR puede permitir o no la inclusión de elementos regulatorios (por ejemplo, la presencia de uORFs) que afecten la eficiencia de traducción. Si bien en el estudio no hacen referencia a las secuencias 5' UTR, la protección del ribosoma (huella ribosomal) les permite secuenciar 12 bases upstream el codón de inicio y 15 bases downstream, obteniendo un perfil de 28 bases. Sería interesante evaluar en estos datos, si para los genes que encontramos con 5'UTR inexistente, las 12 bases upstream corresponden a la secuencia del 3' terminal del *spliced-leader*, lo cual confirmaría nuestros datos.

A continuación se presenta el trabajo publicado con los datos que se resumieron antes.

Transcriptome analysis of the bloodstream stage from the parasite *Trypanosoma vivax*. Greif G, Ponce de Leon M, Lamolle G, Rodriguez M, Piñeyro D, Tavares-Marques LM, Reyna-Bello A, Robello C, Alvarez-Valin F.

BMC Genomics. 2013 Mar 5;14:149. doi: 10.1186/1471-2164-14-149.

Greif et al. BMC Genomics 2013, 14:149
http://www.biomedcentral.com/1471-2164/14/149

BMC Genomics

RESEARCH ARTICLE Open Access

Transcriptome analysis of the bloodstream stage from the parasite *Trypanosoma vivax*

Gonzalo Greif[1†], Miguel Ponce de Leon[2†], Guillermo Lamolle[2], Matías Rodriguez[2], Dolores Piñeyro[1,3], Lucinda M Tavares-Marques[4], Armando Reyna-Bello[4], Carlos Robello[1,3] and Fernando Alvarez-Valin[2*]

Abstract

Background: *Trypanosoma vivax* is the earliest branching African trypanosome. This crucial phylogenetic position makes *T. vivax* a fascinating model to tackle fundamental questions concerning the origin and evolution of several features that characterize African trypanosomes, such as the Variant Surface Glycoproteins (VSGs) upon which antibody clearing and antigenic variation are based. Other features like gene content and trans-splicing patterns are worth analyzing in this species for comparative purposes.

Results: We present a RNA-seq analysis of the bloodstream stage of *T. vivax* from data obtained using two complementary sequencing technologies (454 Titanium and Illumina).
Assembly of 454 reads yielded 13385 contigs corresponding to proteins coding genes (7800 of which were identified). These sequences, their annotation and other features are available through an online database presented herein. Among these sequences, about 1000 were found to be species specific and 50 exclusive of the *T. vivax* strain analyzed here. Expression patterns and levels were determined for VSGs and the remaining genes. Interestingly, VSG expression level, although being high, is considerably lower than in *Trypanosoma brucei*. Indeed, the comparison of surface protein composition between both African trypanosomes (as inferred from RNA-seq data), shows that they are substantially different, being VSG absolutely predominant in *T. brucei*, while in *T. vivax* it represents only about 55%. This raises the question concerning the protective role of VSGs in *T. vivax*, hence their ancestral role in immune evasion.
It was also found that around 600 genes have their unique (or main) trans-splice site very close (sometimes immediately before) the start codon. Gene Ontology analysis shows that this group is enriched in proteins related to the translation machinery (e.g. ribosomal proteins, elongation factors).

Conclusions: This is the first RNA-seq data study in trypanosomes outside the model species *T. brucei*, hence it provides the possibility to conduct comparisons that allow drawing evolutionary and functional inferences. This analysis also provides several insights on the expression patterns and levels of protein coding sequences (such as VSG gene expression), trans-splicing, codon patterns and regulatory mechanisms. An online *T. vivax* RNA-seq database described herein could be a useful tool for parasitologists working with trypanosomes.

Background

African trypanosomes, also known as Salivaria (acquiring this name because they complete the life cycle in the mouthparts or in salivary glands of the insect vector), are the causative agents of disease in humans, domestic and wild mammals. Some sub-species of *Trypanosoma brucei* species complex are responsible for producing the so called sleeping sickness in humans that affects thousands of persons each year in sub-Saharan African countries. *T. brucei*, along with other species of salivarian trypanosomes are the aetiological agents of a variety of livestock diseases not only in Africa, but also South America and Asia are affected by some species [1]. These cattle diseases, generally referred to as nagana, are accountable for important economic losses in the affected countries. Salivarian trypanosomes also infect wild animals (mostly ungulates), which may operate as natural reservoirs.

* Correspondence: falvarez@fcien.edu.uy
†Equal contributors
[2]Sección Biomatemática, Facultad de Ciencias, Universidad de la Republica Uruguay, Montevideo, Uruguay
Full list of author information is available at the end of the article

Apart from their African origin, other two distinguishing features of this group of trypanosomes are that they are mammalian parasites only and that their vectors are several species of the genus *Glossina* (tsetse flies). While in Africa Salivaria trypanosomes are transmitted both cyclically by tsetse flies and mechanically (i.e. without completing the cycle), in America only mechanical transmission by tabanids [2], other hematophagous fly species and even vampire bats has been observed [1,3]. It is not clear whether the ability to be transmitted mechanically by blood sucking insects other than tsetse flies is the ancestral transmission mode or a secondary adaptation to the particular environments that these parasites were exposed when they invaded African regions where *Glossina* was not present or new continents such as America or Asia. In this regard it is worth mentioning that early branching salivarians (like *T. vivax*) complete their cycle entirely in the proboscis of the fly (they cannot survive in the gut). This has been interpreted by Hoare (1972) [4] as a relict form, representing an intermediary stage in the evolutionary pathway from the ancestral mechanical transmission to full adaption to the salivary glands of tsetse fly.

However, the most remarkable adaptation of Salivaria trypanosomes is related to the fact that they remain exclusively extracellular in the mammalian host (in the bloodstream or in connective tissues), and hence permanently exposed to the immune system during infection. In all likelihood, such adverse condition is the reason (i.e. selective force) why in these parasites has evolved their most distinctive trait: a sophisticated strategy, called antigenic variation, to evade the host immune response. This strategy consists in periodically changing a dense protective coat composed by an extremely abundant (10^7 copies) and immunogenic protein, the so-termed Variant Surface Glycoprotein (VSG). These parasites express only one VSG gene at a time, from a repertoire of silent copies that in the case of *T. b. brucei* contains more than 1500 different genes [5]. This mechanism allows transient immune evasion, since after changing the variable glycoprotein that was being expressed, an entirely new parasite population arises that is not recognized by the host's immune system which has developed an antibody response directed against the previous VSG. By repeating this cycle, the parasites are able to maintain the infection.

Reconstructions of evolutionary relationships using sequence data have shown that Salivaria trypanosomes are an indisputable monophyletic clade composed by three main groups [6]. These groups are basically in agreement with the traditional classification based on morphological and life cycle data proposed by Hoare (1972) [4]. The first group; which is fully coincident with Trypanozoon subgenus, contains the model species *T. brucei brucei*, the human pathogens *T. brucei gambiense* and *T. brucei rhodesiense* and two species of veterinary importance, *Trypanosoma evansi* and *Trypanosoma equiperdum*. A second group, the subgenus Nanomonas, includes two small sized nagana causing species, *Trypanosoma congolense* and *Trypanosoma simiae* (which are far more divergent to each other than is the divergence inside Trypanozoon). Finally, the Dutonella subgenus contains *Trypanosoma vivax*, another nagana causing species with economic importance, both in Africa and America. The use of suitable molecular markers on samples taken from the wild have recently disclosed that *T. vivax* also exhibits substantial intragroup diversity, comparable to that observed in *T. congolense* [7,8]. A relevant biological/evolutionary aspect of this last group, is that it occupies a crucial phylogenetic position because besides being the earliest branching Salivaria, its divergence predates that of the remaining ones by a big amount. This key evolutionary position, sometimes incorrectly referred to as being "the most primitive", makes *T. vivax* a fascinating model to study fundamental questions concerning the origin and evolution of several features that characterize African trypanosomes. Indeed, the availability of data from *T. vivax* brings the possibility of making evolutionary inferences concerning the ancestral or derived state of relevant biological features by means of comparisons with *T. brucei* (or/and other salivarians) and consequently provides the opportunity of analyzing these traits in different stages of their evolution (as it has been mentioned before for the mode of transmission).

A recent evolutionary genomic analysis has been conducted in *T. vivax* and other representative species of African trypanosomes, comparing their repertoires of silent VSG genes, how they are organized and diverge aiming to understand the evolution of these proteins and how they gave rise to novel functions [9]. It was found that species differ in the organization of their silent VSG archive, something that may result in different mechanisms for generating antigenic diversity. Besides, these authors suggest that while in *T. brucei* and *T. congolense* there is a high rate of recombination between silent VSG copies, this phenomenon is much less pronounced in *T. vivax*. This analysis, however, barely addresses the topic of the expression of this fundamental group of proteins. In fact no previous genome wide studies on gene expression have been published on *T. vivax*. To tackle this and other important questions, we have conducted RNA-seq analyses of the bloodstream stage in *T. vivax* using different and complementary ultra-high throughput sequencing technologies. Deep sequencing in trypanosome species other than *T. brucei* may contribute to understand several topics concerning the biology of trypanosomatids (notably regulation of gene expression) by giving the possibility of conducting

comparative analyses and providing an evolutionary perspective. Surprisingly, this technology has been scarcely used in trypanosomatids, being restricted to the model species *T. b. brucei* and more recently RNAseq has been used in *Leishmania tarentolae* to explore the role of the nucleotide *J* (β-D-glucosyl-hydroxymethyluracil) in transcription regulation [10].

Methods

Parasites

Experimental infection and parasite purification

T. vivax from the bovine Venezuelan isolate (LIEM-176) were used in this work.

Purification of trypanosomes was done as follows: immunosupressed six-months-old cross-bred sheep were inoculated intravenously with cryopreserved blood containing *T. vivax*. When parasitemia reached values of 2 x 10^7 trypanosomes/ml, blood was extracted and mixed with an equal amount of Percoll (Sigma) containing 8.55% sucrose, 2.0% glucose, pH 7.4 and then centrifuged at 17000 g, 20 minutes at 4°C. Parasites were recovered from top and middle layer of Percoll gradient, resuspended in PBS (sodium phosphate 40 mM, pH 7.5, NaCl 150 mM) containing 1% glucose (PBSG) and subsequently centrifuged at 6000 g for 15 minutes at 4°C. The pellet containing parasites was washed twice with PBSG to remove residual Percoll. Partial purified parasites were resuspended in PBSG and applied to a DEAD-cellulose anion exchange column. Purified parasites were eluted free from red cells, examined by microscopy and counted in a Neubauer chamber. Further details can be found in [11]. *T. vivax* Y486 was grown on mice as described by Chamond et al. [12]. Briefly, 7 to 10-weeks-old male C57BL/6 mice were used. RNA and DNA samples for downstream analysis were obtained from 10^1–10^5 bloodstream forms obtained at the peak of parasitemia (day 8–10 post infection). Parasites were maintained by weekly passages in mice and new stabilates were appropriately and regularly frozen. All animal work was conducted in accordance with relevant national and international guidelines. Mice were housed in the animal care facilities from Institut Pasteur of Montevideo (Uruguay). Animal housing conditions and protocols used in the present work were previously approved by the CEUA (Ethical Committee for Laboratory Animal Use) under the number 013–11 according to the Ethics Chart of animal experimentation which includes appropriate procedures to minimize pain and animal suffering. Infections in sheep were conducted under veterinary supervision with daily control of temperature and hematocrit which never descended below 30%.

RNA purification and quality control

Total RNA was isolated from 10^9 parasites using Illustra RNAspin Mini Kit (GE Healthcare, USA) according to manufacturer's protocol. Obtained RNA was quantified in a Nanodrop (Thermo Scientific, USA) and its integrity was checked by Bioanalyzer (Agilent, USA).

Library construction and sequencing

Double-stranded cDNA was generated from 25 μg of total RNA using a SuperScript II Double-Stranded cDNA Synthesis Kit (Invitrogen) according to the manufacturer's instructions, except for oligonucleotide used for first strand synthesis and 5-methyl-dCTP (Jena Biosciences) instead of dCTP. The primer used was 5′ CTGGAG$(T)_{16}$VN 3′, the 5′ end of the primer contain the restriction site for the enzyme *GsuI*. After the synthesis of the second strand, the cDNA was precipitated with 1/10 volumes of Sodium Acetate (3 M, pH =5.2), 2 μL of glycogen (15 μg/mL) and 3 volumes of absolute ethanol and resuspended in 70 μL of RnaseFree water. 65 μL of cDNA was digested with *GsuI* (Fermentas) for 4 hours at 30°C to cleave the poliA tails. The digested cDNA was used to prepare the 454 and Illumina libraries.

454 library preparation and sequencing

Library was prepared using the GS Titanium DNA library preparation kit (Roche) according to the manufacturer's protocol starting with 2.5 μg of cDNA. The emPCR was done with GS Titanium SV empPCR kit (Roche) according to manufacturer's instructions. We used GS Titanium Sequencing Kit XLR70 (Roche) to sequence 1/2 GS Titanium PicoTiterPlate kit 70 × 75 in 454 Genome Sequencer FLX System (Roche). Illumina sequencing was carried out in a GAIIx on the same cDNA library which was re-fragmented and universal Illumina adaptors were added. Raw data were deposited in the NCBI database under submission number SRA056332.

Bioinformatics and data analysis

Data quality analysis

The details of sequence data obtained by 454 and Illumina sequencing are presented in Additional file 1: Table S1. For the first technology 187491 reads, with an average length of 295 nt. were obtained. This corresponds to 54 Mb of sequence data. For the second technology, 37 million of reads (36 nt), corresponding to 1332 megabases were obtained. Several quality tests were carried out. In the first place, the percentage of contaminating reads present in the sample (i.e. corresponding to the host) was determined. For this purpose the reads were mapped into the sheep genome using Blastn. By doing this it was possible to establish that only 433 reads (i.e. 0.20%) were of host origin. A similar figure was obtained for Illumina reads (in this case mapped using Bowtie [13]). The same procedure was followed for other possible contaminating sources (such as human) and only traces were detected (e.g. 12 reads

from 454 technology corresponding to human). This indicates that the quality of the starting material was high.

In the second place the number of artificially repeated reads (i.e. those corresponding to the cases when the same cDNA segment is sequenced more than once) were identified. This distortion (common in 454 sequencing) is introduced during the emulsion PCR step because a single cDNA molecule, but multiple beads are located in the same micro reactor. For genomic sequences these are customary identified as "same-start reads" provided that it is unlikely that by chance alone multiple DNA segments obtained by random fragmentation of a genome start at exactly the same position. However, it is obvious that for RNA derived DNA (cDNA), sharing the "same-start" is not uncommon. For this reason the candidates of artificially duplicated reads were identified as those ones that start and end at exactly the same nucleotide. About 15000 reads (9%) fall in this category (Additional file 1: Table S1), such proportion of repetitions is low when compared with other studies, where this kind of reads can be as abundant as 25%. These repeated reads were collapsed for further analysis.

In the next step, reads corresponding to ribosomal RNA were identified, totaling 2267 (1.21%) in the case of 454 FLX. The percentage of rRNA reads was significantly higher for Illumina (more than 2 million, which corresponds to slightly more than 6%). Such a small number is unusual considering that this type of RNA normally represents more than 70% of the RNA population, thus indicating that the filtering strategy of using an oligo-dT containing primer turned out to be very effective in order to get rid of ribosomal RNA. In addition, this methodology does not restrict the isolated RNA to mature mRNA either, as it can be inferred from the fact that other types of RNA molecules are quite abundant in the sample. In effect, transcripts derived from the kinetoplast genome are relatively abundant (Additional file 1: Table S1). For the case of maxicircle, it was possible to identify them using simple homology search, given that these genomes are relatively conserved among trypanosomatids. But, such strategy was not suitable for minicircle derived RNA identification because of their lack significant conservation. Therefore the incidence of this latter group was not determined.

To evaluate the genome coverage of RNA-seq data produced in this work, 454 and Illumina reads were mapped to genomic sequences (retrieved form GenBank) in order to estimate the sequencing depths of the top, middle and botton 1000 expressed genes. This was done using RNA-SeQC program [14], detailed results are presented in Additional file 2: Figure S1.

Assembling and functional annotation

Assembling of 454 reads was conducted using two different computer programs Mira [15] and Newbler (Roche, Switzerland). The two resulting assemblies were compared to each other, in order to assess their qualities and determine which one was more appropriate for subsequent analyses. The quality of the assembly was assessed by comparing the assembled contigs with a reference set containing well defined mRNA sequences. To assess quality, two variables were measured, the proportion of the reference mRNA that is well reconstructed (P) and the number of contigs falling in each mRNA reconstruction fraction (N), so that the overall quality of the assembly is given by: $Q = \sum_i N_i P_i$. This comparison was done using a reference set consisting of protein coding sequences available in GenBank that are putatively expressed in the bloodstream stage. These were identified on the basis that their *T. brucei* orthologs are unambiguously expressed in this stage. In turn, the latter condition was determined by testing which *T. brucei* protein coding genes are observed in the bloodstream EST collection. It should be noted that this collection was built using traditional Sanger sequencing from poly A + RNA, which due to the low sensitivity of the method, contains mainly unequivocally transcribed genes. The results obtained allowed us to draw two useful conclusions. In the first place Mira outperforms Newbler, yet by a narrow margin; provided that the contigs built by Mira reconstruct better the mRNA (i.e. the Q statistics is higher). Secondly, more than 92% of the putatively expressed mRNAs are tagged (either by contigs containing several reads, or by individual reads), hence indicating the 454 derived sequence dataset is a good picture of the transcriptional state of the parasite (Additional file 3: Table S2).

Functional annotation of RNA derived contigs was carried out using a set of complementary tools: ESTscan [16], Blast2Go [17], InterProScan [18] and AnEnPi [19]. In the first place, to identify *T. vivax* genes encoding proteins with a known or unknown function, it was necessary to obtain high quality virtual translation of contigs. This translation is not always the straightforward exercise of mechanically applying the genetic code to possible ORFs. Instead, contigs often contain serious translation problems derived from sequencing errors that may change the reading frame (frameshifts). To handle this complication the ESTscan program was used. This application employs a Hidden Markov Model (that uses the distribution of codons) to restore the correct frame by introducing indels. The program needs to be calibrated (trained) in such a way that it recognizes possible alteration in the ORFs on the basis of their statistical properties [16]. For training the ESTscan HMM, *T. vivax* coding and intergenic sequences were retrieved from public databases. After this step, functional annotation of the translated contigs was done combining the results of Blastp against nr NCBI (all non-redundant GenBank CDS translations plus other well curated

databases) and a domain analysis based on Interproscan. Results of both sources were integrated using Blast2Go, which allows assigning GO terms to the entries by using simple annotation rules. Because B2G is quite conservative to assign ontology terms, the analysis was complemented with a simple Blastp search against translated nr NCBI. Besides the AnEnPi pipeline [19] was used on KEGG in order to predict possible metabolic pathway that are active in the bloodstream stage of *Trypanosoma vivax*.

Determination of transcription levels

To determine the transcription levels we decided to use Erange [20] software on Illumina data. After cleaning low quality reads, the remaining reads were mapped to the *T. vivax* genome (retrieved from GenBank) using Bowtie [13] and allowing up to 1000 multimatches and up to 1 mismatch. RNA-seq Erange pipeline was used with minor modifications. It is important to take into account that in genomes like this one, which contain several related paralogous genes, the use of computer applications that consider the unique regions of the genes to re-normalize the assignment of multimatching reads (like Erange), is essential. This approach permits also determining which ones of the paralogous genes from a multigene family are really expressed at a given time. For 454 data (where the problem of multimatching reads is mitigated or simply eliminated, because of reads' lengths) transcription levels were computed using in house Bash and Perl scripts to parse Blast or Bowtie outputs. rpkm estimates are presented on Additional file 4: Table S3.

We note that in these analyses it was not possible to use biological replicates. Because of the limited amounts of RNA isolated in each individual infection, it was necessary to pool all samples together. Although this is not optimal because variability is not assessed, for a couple of reasons such limitation is not critical for this study. First, the main focus of our study is not compare different moments of the parasite life cycle aiming to determine which genes are up or down regulated. Moreover, since our starting material is a pool of different biological independent samples, large variance that might especially affect low expression genes (and yield a distorted picture) is largely alleviated. Transcript levels for *T. brucei* genes were also estimated as described above using published RNA-seq data [21] retrieved from the SRA archive.

Identification of splice-acceptor sites

cDNA sequence tags (36 bp) that contained terminal Spliced Leader sequence (SL) were extracted from the Illumina output. The SL sequence was found in a 0.5% (171200) of the reads and in the majority (94.8%) of them in the sense direction, as expected because of constraints imposed by the cDNA size-fractionation and sequencing protocols. The Spliced Leader segment was trimmed from these sequence tags with a homemade python script. Sequences greater than 19 bases were used in downstream pipeline. Genomic matches were identified by mapping these reads with Bowtie against the *T. vivax* genome sequence. No mismatches were allowed.

We used output of Bowtie (sam file), genomic information as given by the gff files and blockbuster software [22] to cluster the mapped reads in order to detect trans-splicing sites in the chromosomes and other genomic sequences from *T. vivax*. Cluster information was parsed with gff information with homemade Perl script and a final table with gene information and trans-splicing sites associated were obtained. A similar pipeline was used to analyze trans-splicing patterns in *T. brucei* and *Trypanosoma cruzi*.

In the case of *T. cruzi*, the RNAseq data from three stages of the life cycle of the parasite (epimastigote, trypomastigote and amastigote) were obtained in our laboratory (further analysis on this data will be published elsewhere). Due to the sequencing strategy used in *T. cruzi* (stranded) the number of Spliced Leader containing reads was modest; this restricted the type comparisons that could be conducted in this species to only the determination the splicing motifs.

Results and discussion

Assembling 454 reads and functional annotation of resulting contigs.

Because 454 FLX sequencer yields long reads, it is possible conduct “de novo” assembling to obtain good quality contigs. This was done with two different computer programs, Mira and Newbler (Roche) using optimized parameters for RNA-seq assembling.

The results obtained allowed us to conclude that Mira outperforms Newbler, since the contigs obtained represent better reconstructions of full length mRNA (i.e. the Q statistics is higher). Besides, more than 92% of the putatively expressed mRNAs are tagged (either by contigs containing several reads, or by individual reads), hence indicating that the 454 derived sequence dataset is a good picture of the transcriptional state of the parasite (Additional file 3: Table S2).

As mentioned before high quality virtual translations of contigs were obtained using ESTscan. A total of 13385 translatable sequences were identified by ESTscan among which 6583 contained more than one read. Functional annotation, carried out using Blast2GO, enabled us to identify 3834 contigs for which it was possible to assign one or more Gene Ontology terms. However, the number of contigs whose virtual translation have homologs in other species (blast e-value $<1e^{-10}$) was 2 times as much (7796), and hence it was possible to make a relatively reliable primary functional assignment for these contigs as well. In addition, we could determine a tentative

enzymatic function using KEGG search for a substantial number of virtual translations totalizing 327 EC numbers assigned to 1281 contigs. Additional results on the functional annotation are available in the web page (see next section).

Finally, it is worth mentioning that more than 1000 contigs that are transcribed at different levels and unequivocally correspond to protein coding genes, do not have homologs in other species, including other trypanosomatids such as *Leishmania sp*, *T. cruzi* and the African trypanosomes for which genome sequence is available (*T.b. brucei*, *T.b. gambiense* and *T. congolense*). This means that in all likelihood they are species specific. Among these *T. vivax* specific contigs, around 50 genes have not even been reported in the *T. vivax* genome available in GenBank, indicating that very probably they are specific of the strain LIEM-176. 564 species specific contigs for which it was possible to build a full cDNA were chosen for additional analysis. A preliminary characterization of these genes was carried out using a battery of informatics tools such as those that identify signals for sub-cellular localization and domain analysis. These results are presented in Additional file 5: Table S4. Database web interface.

A relational database (MySQL) was built to store and browse the data and results produced in this work. In fact the database contains raw as well as processed and annotated data as described in the previous section. A Pyhton web application was developed using the Django programming framework. This application provides user-friendly data querying, browsing and visualization through a web interface (http://bioinformatica.fcien.edu.uy/Tvivax/). In this web interface it is possible to search for, and retrieve reads, contigs as well as virtual translations. Besides, the database can be searched using different criteria such as length, depth of the contigs (i.e. expression level), GO terms, Enzyme Commission numbers, Blast e-values, keywords, etc. or a combination of these criteria. Moreover any sequence can be blasted against the dataset. The annotated entries are linked to the reference databases used for their annotation, namely Amigo Gene Ontology [23], KEGG repository at EBI and NCBI. In addition the database offers the possibility to highlight *on-the-fly* the enzymes in the pathway image files downloaded from the KEGG FTP site. Expression of Variant Surface Glycoproteins in T. vivax, and the protein composition of the cellular surface.

Because of the strategic evolutionary position of *T. vivax*, as the earliest branching African trypanosome, it is important to analyze in this species the expression patterns of Variant Surface Glycoproteins, as well as the organization of this gene family to help shedding light on diverse questions concerning the origin and evolution of antigenic variation.

To this aim we first tried to identify the VSGs that were present in our RNA sample by using a simple strategy, which consisted in searching putative candidates among the most abundant mRNAs (namely those contigs built with the highest number of 454 reads), provided the high expression levels that these proteins exhibit. By doing this, only one candidate VSG was found. Surprisingly, the mRNA identified was highly similar (DNA sequence identity of 90.4%, see Additional file 6: Figure S2) to the only VSG sequence already reported for *T. vivax* that was derived from a West Africa isolate called Ildat 2.1 [24]. Even if this finding confirms previous reports that suggest that the American *T. vivax* (or more correctly *T.vivax*- like given the great intra taxon diversity inside Dutonella) is closely related to West African strains [25] some remarks should be made. It should be taken into account that *T. vivax* was introduced in America around 1850, in the French Guiana, by infected Zebus imported from Africa [26-29]. Since its introduction, *T. vivax* has been disseminated by horse flies (Tabanidae) [4] and stable flies (*Stomoxys* spp.) [30], and it was rapidly dispersed throughout South America. However, the degree of sequence similarity seems to be much higher than what we would have expected if account is taken to the fact that these genes normally diverges extremely fast. Indeed, the comparison of VSG genes among *T. brucei* strains reveals that even closely related subspecies (like *T. brucei brucei* and *T. brucei gambiense* and the so called Tororo isolates) have very divergent silent repertoires [31]. In addition, it should be noted that this VSG gene was not identified in the draft genome deposited in GenBank corresponding to the Y486 strain. We tested this absence by PCR using two sets of primers specifically designed to amplify this gene. Both primers sets failed to amplify, thus confirming that this VSG copy is really not present in the Y486 strain (Additional file 7: Figure S3). Conversely, the gene encoding the VSG protein expressed by Y486 (Ildat 1.2) is not detected in Liem-176 transcriptome. Considering that Y486 also belongs to the same group of West African *T.vivax*-like strains [7], these two results seem to be conflicting. Alternatively they indicate that the two processes of genetic differentiation of their silent archives, sequence divergence (involving single nucleotide changes) and genome plasticity (gene gain, loss and reshuffling) are not necessarily correlated, especially in this initial phase of taxa differentiation.

As far as the expression level of the main VSG is concerned, it is interesting to note that although its transcript abundance is very high (twice as much as the already highly expressed alpha and beta tubulins, see Additional file 8: Table S5), the number of Illumina reads mapping on this contig corresponding to the VSG gene, is not nearly as high as those reported for VSGs in

T. brucei, where they represent between 5% and 11% of all sequenced reads [21,32,33]. This is an interesting aspect and raises several questions concerning the membrane protein composition and diversity (in terms of relative abundance of their constituent proteins) in this and other African trypanosomes.

These questions can be answered by assessing the expression levels (as indicated by RNAseq data, see further details in next section) of genes encoding proteins predicted to have surface location, thus allowing us to compare the surface protein composition from both African parasites. As it emerges from Figure 1, it is evident that while in *T. brucei* the VSG is absolutely predominant (representing approximately 98%), in *T. vivax* it only represents about 55%. Other very common trypanosomatidae membrane proteins, like GP63, are almost absent in *T. brucei*, while they exhibit appreciable frequencies in *T. vivax*. These results thus indicate that the cellular surface of *T. vivax* is substantially different from that of *T. brucei* (and very likely from other African trypanosomes). In turn, these results concerning the much lower membrane concentration of VSG proteins are in keeping with previous ones from electron microscopy [34], which indicate that in *T. vixax* the VSG surface coat is noticeably less dense than in *T. brucei*. In addition, these results taken together raise the question of what would be the role of VSGs in *T. vivax* (and perhaps its ancestral role) in immune system evasion, provided that such relative lower concentrations cast some doubts on how efficiently it could act as a fully protective coat as it happens in *T. brucei*. Needless to say, proteomic analysis will provide more substantial data to help gaining additional insight on this fundamental point. In particular it would be important to analyze the efficiency of antibodies targeted against invariant membrane epitopes. Assessing transcription levels.

One of most useful features of RNA-seq analysis is that it allows direct and quite accurate estimations of transcript levels. Given that the number of reads matching with the transcripts of a given gene is expected to be proportional with the concentration of the mRNA molecule as well as with its length. Then the normalized (by length) numbers of 454 reads used for assembling of a given contig, or the number of Illumina reads mapping on the same contig (or on the corresponding genomic CDS) can be used as a measure of expression level.

In the first place we compared how congruent are the two sequencing technologies used in this work for estimating transcript levels. Specifically, we compared the number of 454 FLX reads used in the assembly of a given contig versus the number of Illumina reads mapping on the same contig. As it can be observed in Figure 2, even though for some points (contigs) the estimation differs, the agreement is quite remarkable ($r = 0.83$). The genes (contigs) exhibiting estimations that are inconsistent between the two technologies were further analyzed to understand the reasons why these two technologies yield contradictory estimations. Indeed, for several genes very few 454 reads contributed to their assembly, while many of the same genes were tagged by considerable number of Illumina reads. Even if it is reasonable that many low expression genes that are tagged by Illumina reads will be not detected by 454 FLX sequencing technology,

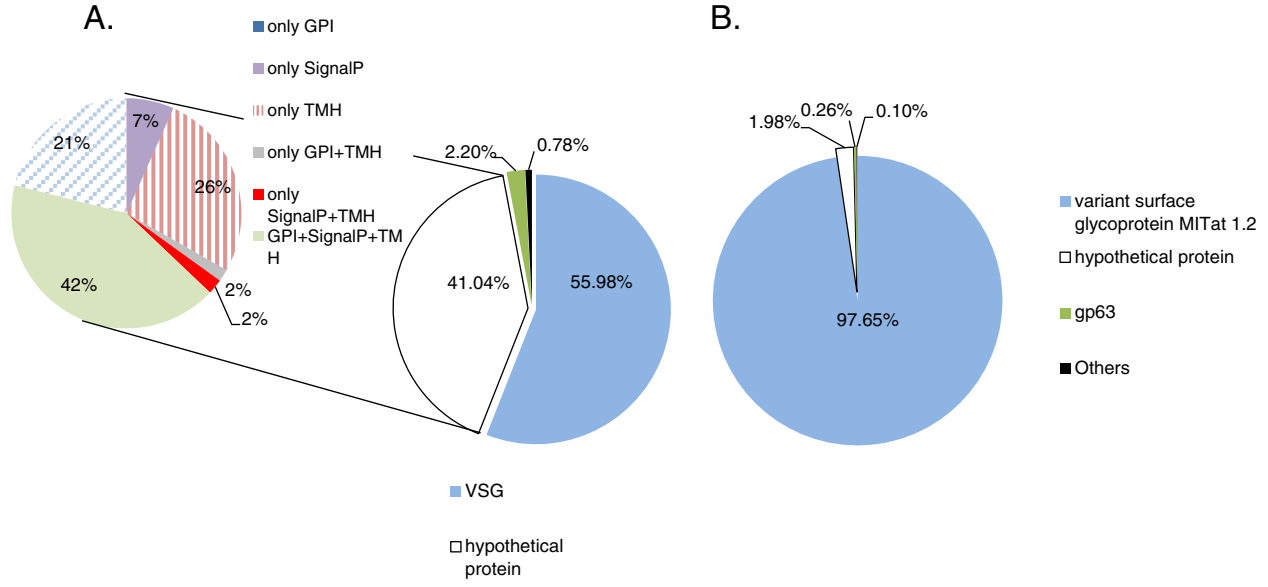


Figure 1 Protein Membrane composition as inferred from expression levels. A. *T. vivax*. **B**. *T. brucei*.

given the comparatively small number of reads the latter technology yields, in few cases the disagreement between the two technologies goes far beyond than what would be expected by random variability. In effect, since the ratio in the numbers of reads between the two technologies is 181 (see Additional file 1: Table S1), which is close to the regression coefficient in Figure 2A, it follows that several genes on which map many thousands of Illumina reads (>10 thousands) are not expected to be tagged by none or so few 454 reads. The visual inspection of these troublesome points shows that they correspond to DNA segments having extreme compositional biases. On the other hand the comparison between the two technologies was also conducted by mapping their reads on genomic regions to assess the variability in sequencing depths estimated by each method. Again it is possible to observe that there is a good agreement between both methods (Figure 2B).

Estimation of transcription levels for 11886 CDS annotated in GenBank was done using the Erange software that corrects multiple matching reads considering unique parts of genes for their assignation. The gene expression levels (read count and RPKMs) are available in Additional file 4: Table S3.

An unexpected observation is that several genes and genomic regions appear to be non-transcribed at all in the bloodstream stage of *T. vivax*, as it can be

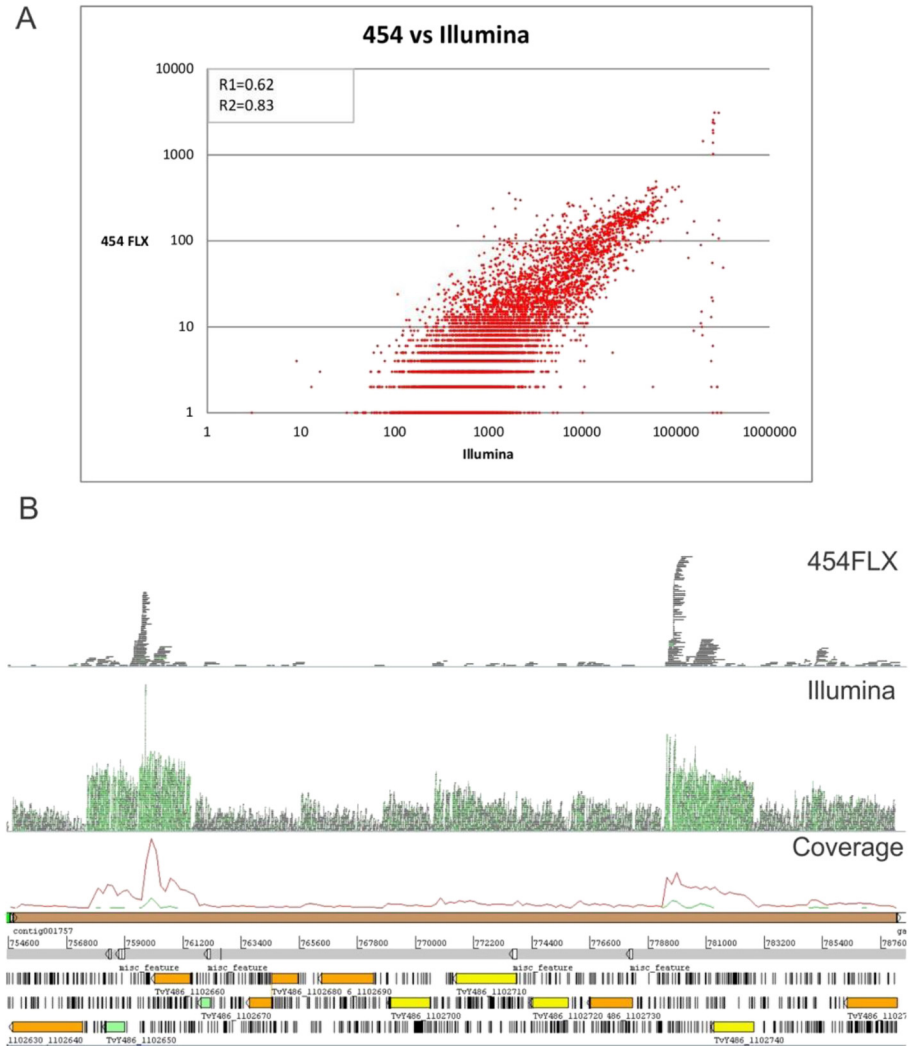


Figure 2 Comparison of transcription levels estimation. A. Scaterplot of the number of 454 FLX reads used in the assembly of a given contig versus the number of Illumina reads mapping on the same contig. R1 and R2 stand for the correlation coefficients before and after disregarding the points that exhibit an extreme discrepancy between the two technologies (those ones forming an almost vertical line on the rightmost part of the figure). **B**. This figure depicts a depth profile showing the reads that map on a given genomic region. The upper part corresponds to 454 FLX, Illumina reads appear in the middle and the last part corresponds to the graphical representation of the corresponding genomic region.

appreciated in Figure 3 panels A and C, which shows that some of these regions are devoid of reads. We note that this result could be attributed to the fact that the reference genome used to map reads and the RNA used in this work come from different *T. vivax* strains (Y486 and LIEM-176 respectively), thus the absence of reads in some regions could be simply the result that the regions in question are not present in Liem-176. To control this possibility, we decided to test if the genomic DNA segments without reads mapping on them are also present in the genome of the strain from where the RNA comes. Primers specific for these regions were designed (indicated in red and green in Figure 3B). The PCR results presented in Figure 3C indicate that the regions with no reads are definitively present in the LIEM-176 strain, and hence the absence of reads mapping on them is in all likelihood the result of their lack of transcription. These results have implications on the long standing questions concerning the mechanisms of gene expression regulation in trypanosomatids. Indeed, the current accepted view is that in trypanosomatids everything (or almost everything) is promiscuously transcribed, and they regulate their gene expression mainly post-transcriptionally, either by differential RNA maturation and degradation, or by controlling translation initiation or even post-translationally [35]. Hence, the results presented in Figure 3 showing that certain genes and genomic regions are not transcribed, strongly suggest that regulation of transcription initiation might also play an important role in gene regulation. Gene expression levels and codon biases in trypa-nosomes.

It is well established that in most organisms synonymous codons are not randomly used [36,37]. Biased codon usage may result from a diversity of factors, among which translational efficiency (translational selection) is one of the most important, being related to the fact that the preferred codons in highly expressed genes are recognized by the most abundant tRNAs. More than fifteen years ago, we have shown that in trypanosomatids there is enormous intragenomic variability in codon biases, and this was essentially the result of the interplay between mutational biases and translational selection. In this analysis it was also shown that, in the African trypanosome *T. brucei*, the putatively highly expressed genes exhibit essentially the same kind, but with lesser strength, of codon biases as in *T. cruzi* (towards G and C ending codons) [38]. One of the main drawbacks of these analyses, is that the data on expression levels were very fragmentary or simply assumptions (for instance we assumed that proteins like ribosomal proteins, elongation factors, and enzymes from glycolysis were highly expressed). Interestingly, some of these results were confirmed more recently, yet no analysis was carried out so far comparing codon preferences using robust data on gene expression [39]. The availability of NGS data gives the opportunity to re-address this topic from a more reliable perspective.

Figure 4A, shows the frequencies of G + C ending codons in the 20% most and least expressed genes in *T. vivax*. Even if it is possible to see that there is substantial variability inside each group, it is also clear that there is a very strong preference for G- and/or C-ending codons in the majority of genes that are more actively transcribed, and this preference also holds when each synonymous codon group is considered separately (Additional file 9: Figure S4). In agreement with previous results, it is possible to observe that also in *T. brucei* the highly expressed genes exhibit clear preference for G- and C-ending codons. However two differences should be pointed out. First, the overall distribution in GC_3 values is shifted towards the left (namely lower values), and second the difference in GC_3 preference between low and high expression genes is less pronounced than that observed in the other trypanosome. Based on these results it is possible to conclude that the process of weakening of codon biases observed nowadays in the high expression genes from *T. brucei* only affected the branch leading to the Trypanozoon subgenus, and not all Salivaria trypanosomes, provided that *T. vivax* did not undergo such a process.

An interesting observation is that in both African *Trypanosoma* species there is a group of genes that exhibit an atypical behavior in the sense that they are expressed at high or very high levels, yet they display weak or none GC_3 biases. Furthermore, in both species the respective groups of unusually behaving genes include many species specific proteins and also proteins like many ribosomal proteins and translation factor 5a (well known to be highly expressed in most species). In addition, the two groups contain many genes that are coincident (i.e. orthologs) between *T. vivax* and *T. brucei* (Additional file 10: Table S6). It should be noted that the very existence of several orthologous that are highly expressed and lack codon biases in both species suggests that this unusual behavior cannot be attributed to natural variability in codon preferences that could eventually display high expression genes. Instead, this very likely reflects genuine functional requirements. We note that this peculiar observation had been pointed out before for the case of VSG genes in *T. brucei*, the highest expressed gene, yet the different genes encoding VSG proteins have very weak codon bias [38,40]. The puzzling aspect of this observation is why and how is it possible that these organisms do not optimize the codon preferences in genes that represent such a substantial proportion of the protein mass. Two different explanations (yet not mutually exclusive) can be put forward. One of these is that these genes belong to multigene

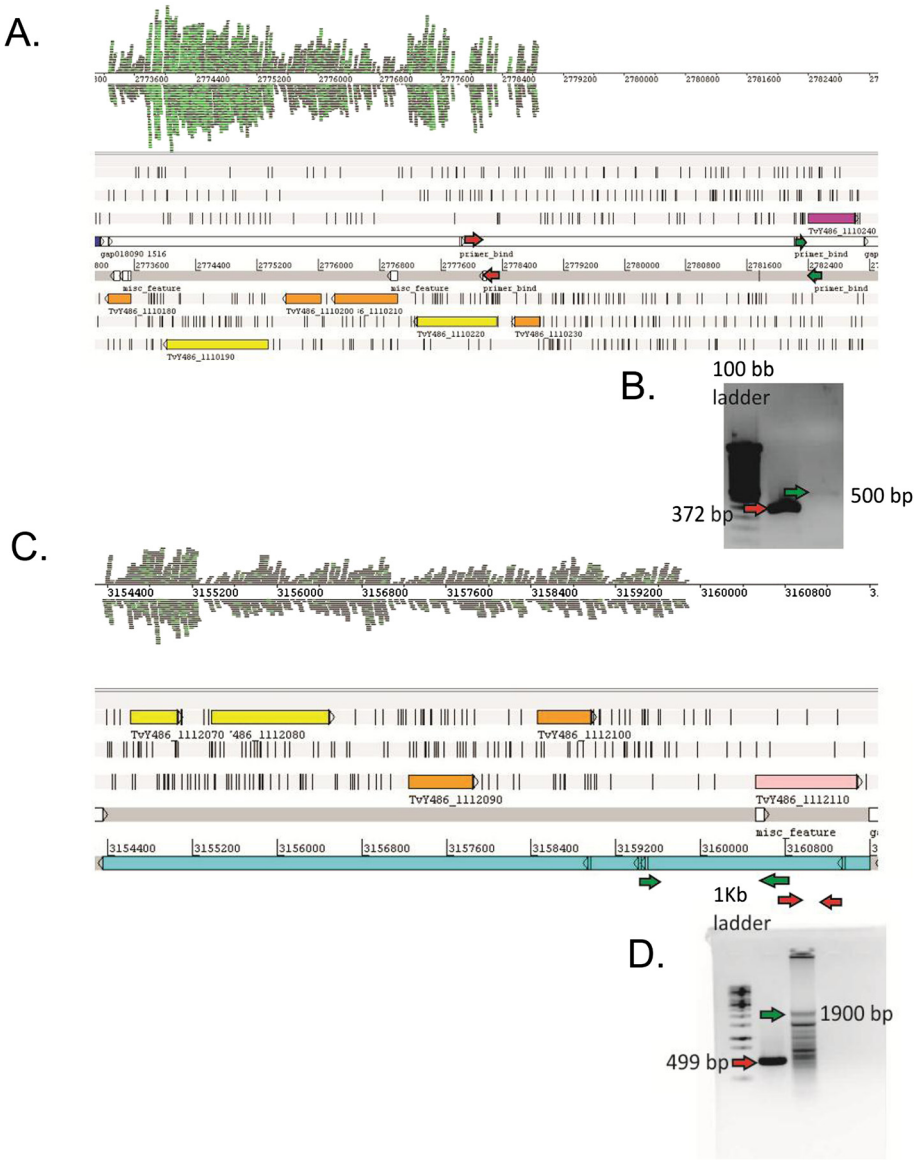


Figure 3 The figure shows two representative genomic regions that appear to be not transcribed in bloodstream stage of *T. vivax* (A,C). B and **D** PCR amplification of the genomic region represented in panels **A** and **C** respectively. The amplification confirms their presence in the genome of LIEM-176 strain. Arrows (red and green) represent the two primer sets used for each genomic region.

families that have emerged, or became highly expressed, only recently (on an evolutionary scale). Hence selection did not have enough time to optimize their codon biases. This can be the case of leucine-rich repeat protein in *T. vivax*, procyclins in *T. brucei* (and also Mucin Associated Surface Proteins (MASP) in *T. cruzi*), that are very highly expressed yet have AT or weak GC biases (see Additional file 10: Table S6). The second explanation is that translational selection is not effective enough for these genes because they are seldom expressed, namely they behave most of the time as silent ORFs (like pseudogenes), during which time natural selection does not have any effect on them. This second explanation applies to VSG coding genes. Some additional analyses give support to these proposals. Indeed, when the analysis of the relationship between codon biases and gene expression levels is restricted to those coding sequences that have bona fide (and conserved) orthologs in other trypanosome species (what could be called the trypanosome gene core), most genes are "well-behaved", that is the differences in GC codon biases between highly and lowly expressed genes become sharper in both species (Additional file 11: Figure S5).

Finally, we would like to mention that in spite of the fact that these explanations may account in part for this

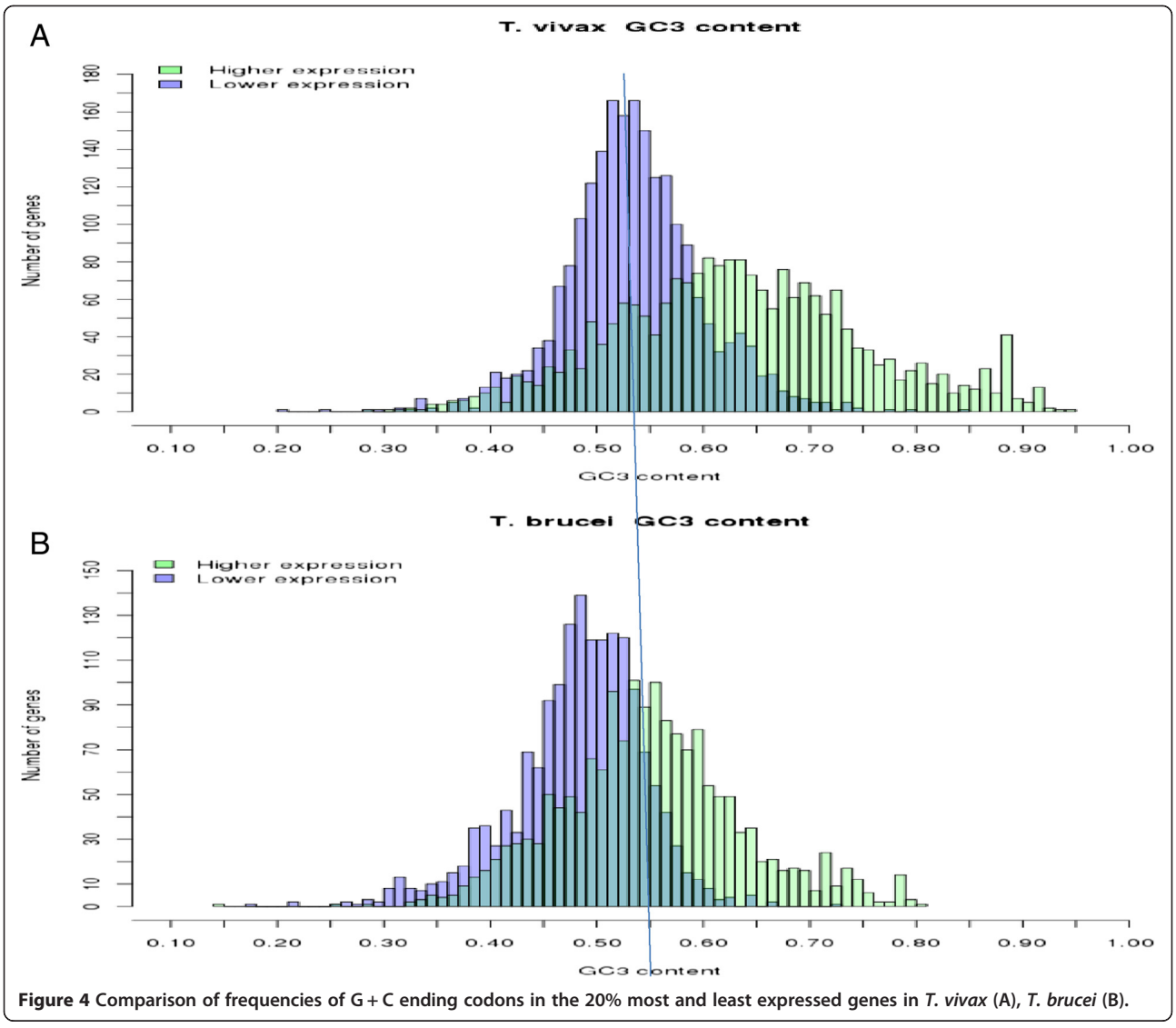


Figure 4 Comparison of frequencies of G + C ending codons in the 20% most and least expressed genes in *T. vivax* (A), *T. brucei* (B).

atypical behavior displayed by trypanosomes, it is also evident that they do not apply to the case of ribosomal and other conserved proteins that exhibit low o none GC_3 biases and very high expression. Mapping trans-splice sites.

To identify trans-splice sites, we mapped 159395 miniexon containing Illumina reads onto the *T. vivax* draft genome that has been recently made available in Genbank. This allowed us to identify the trans-splice sites in 5959 genes. Among these genes, 3350 had only one bloodstream splice site and 2609 genes had two or more. The distribution of splicing sites per gene is presented in Figure 5A. The maximum number of sites per gene was 9 and the average 1.48. This figure is considerably lower than that described for *T. brucei* (mean 2.7-2.9 sites/gene [32]). Using the splice site location, the distribution of 5′ UTR lengths was also determined (Figure 5B). The mean sizes for the first and second splice sites were 132 and 164 nts, respectively. These are in the same range as it has been described for *T. brucei* [21,32,33].

Next we analyzed the consensus sequences around the splice site. Figure 6A, shows the logo representation of the major site, which is virtually identical to that described for *T. brucei*, basically consisting in a long (>50 nt) poly-pyrimidine rich track. The consensus for the second and the remaining minor sites are also very similar yet the signal is not as strong as for the major site (Additional file 12: Figure S6) Furthermore, the canonical AG dinucleotide was found at 98% of the major splice sites (Figure 6C), whereas minor sites had an AG dinucleotide in progressively decreasing proportions, 94% for the second, 90% for the third and 80% for the fourth site. Therefore, the frequency of AG at secondary splice sites is considerable higher than that observed in *T. brucei* (that on average is around 80%, see reference [33]). As it has been also

observed in *T. brucei*, the second most frequent dinucleotide at splice site is GG (Figure 6). A similar analysis was carried out also in *T. cruzi*; the logo illustration presented in Figure 6B shows that also in the American parasite the overall pattern is very similar to that of salivarians.

These results indicate that both the trans-splicing machinery, and the signals that this machinery recognizes, have been conserved not only in African trypanosomes, but also in *T. cruzi*, and therefore in all likelihood in all trypanosomes.

Along the same line, we also compared orthologous genes between *T. brucei* and *T. vivax* to investigate whether the spatial pattern of trans-splicing sites, namely their number and distances to the initiation codon, was similar between these two African parasites. Interestingly enough, the pattern exhibited considerably agreement in spite of the fact that the DNA sequences in the 5' UTR located between the sites of splicing and the initiation codon were poorly conserved (Figure 7).

As it has been already reported for *T. brucei*, a large number of *T. vivax* genes contain one or more (up to five) trans-splicing sites inside the coding region [21]. Moreover, we also found that a significant number of genes contain their main, or unique, splice site very close (sometimes immediately before and sometimes after) the start codon (AUG). We decided to investigate this peculiar aspect further by determining if this feature is characteristic of some groups of genes or functions. A Gene Ontology enrichment analysis was carried out to explore this aspect, namely if the genes exhibiting this feature encode proteins belonging to some particular categories. Interestingly enough, this group contains a much higher than expected frequency of ribosomal proteins, elongation factors and other proteins related with the translation machinery. Other type of proteins over-represented in this group are heat shock proteins and proteins that interact with RNA (Figures 8A and B).

Because the annotation of *T. vivax* genes available in GenBank is not precise in relation to the correct identification of start codons, and considering that this trouble can introduce serious biases in this analysis, the same ontology analyses were also conducted in *T. brucei*, whose annotation is expected to be much more depurated. As it can be observed in Figures 8C and D, the same pattern is also present in *T. brucei*, and hence allows us to conclude that it cannot be attributed to an artifact due to low quality annotation.

For these genes with splice site very close to the start codon, we identified the orthologs between *T. vivax* and *T. brucei*, and in many cases the splice sites were located upstream of the annotated start codon in one of the species but downstream in the other. We suspected that in all likelihood this was caused by the above mentioned trouble of misidentified start codons. Therefore their sequences were aligned to determine, on the basis of DNA and amino

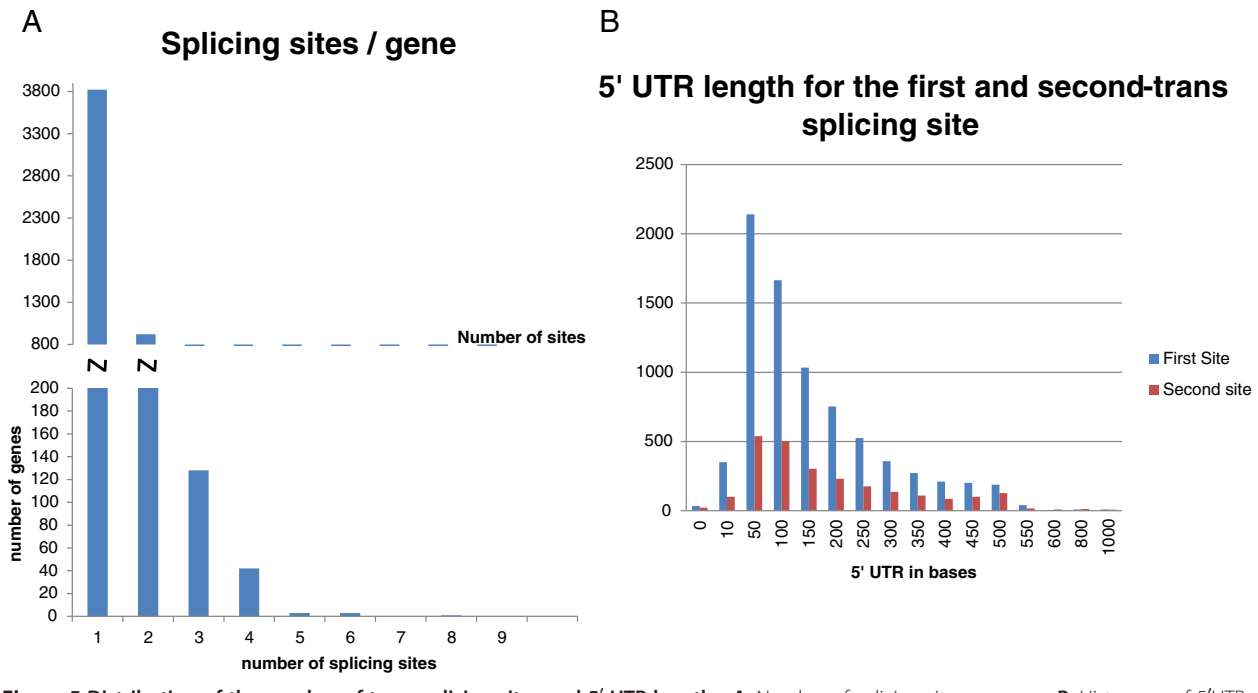


Figure 5 Distribution of the number of trans-splicing sites and 5′ UTR lengths. A. Number of splicing sites per gene. **B**. Histogram of 5′UTR lengths to the first (FTSS) and second trans-splicing site (STSS). 5′UTR lengths average values 132 nts and 164 nts for FTSS and STSS respectively.

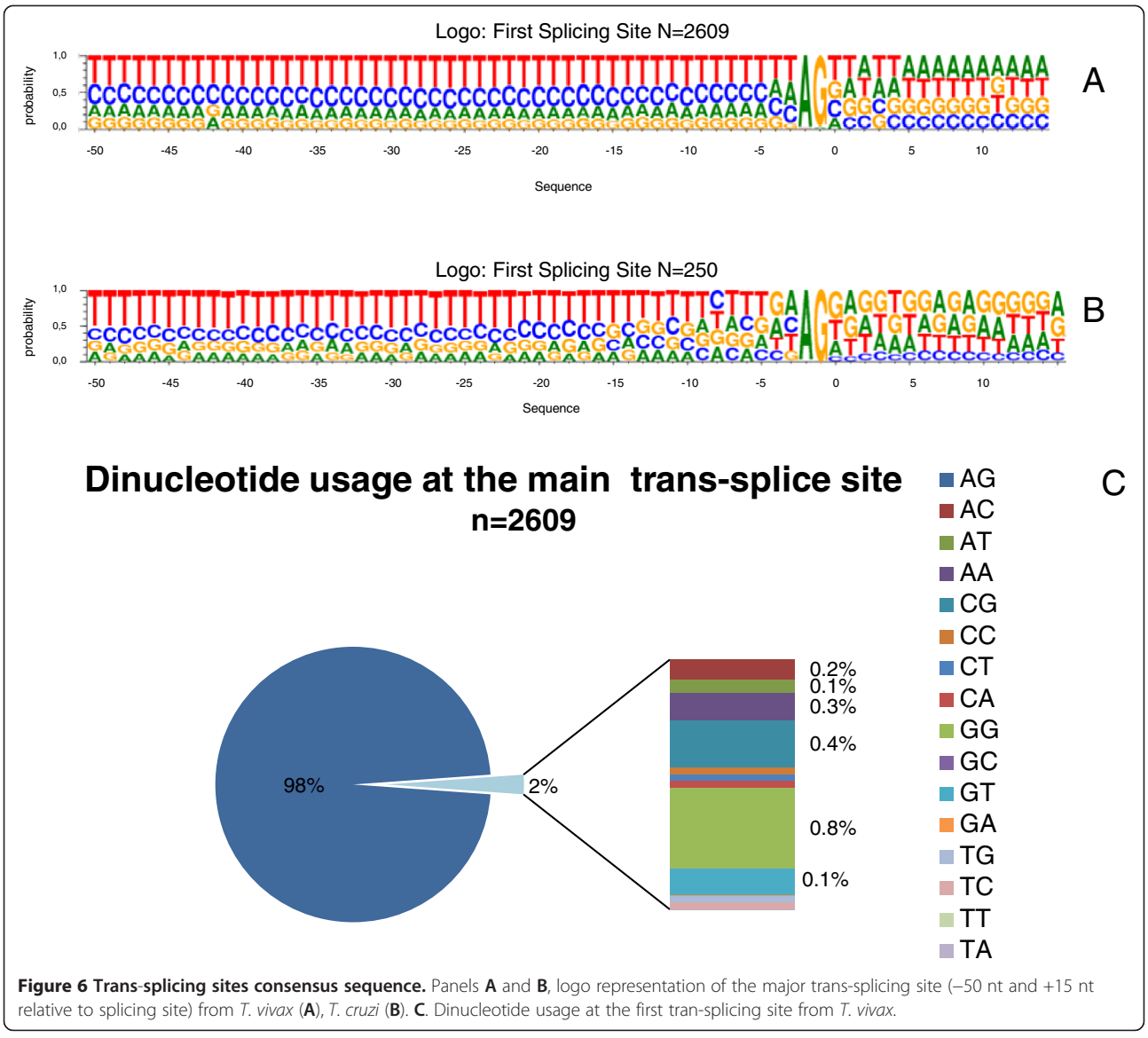


Figure 6 Trans-splicing sites consensus sequence. Panels **A** and **B**, logo representation of the major trans-splicing site (–50 nt and +15 nt relative to splicing site) from *T. vivax* (**A**), *T. cruzi* (**B**). **C**. Dinucleotide usage at the first tran-splicing site from *T. vivax*.

acid conservation, the most probable start codon. For these comparisons sequences from *T. congolense* (whenever available) were also included. The rationale for this approach for detecting more accurately AUG start codons is simple, and it is based on the fact that inside the coding part of the genes there are higher functional constraints and hence higher conservation. The approach allowed us to detect that many AUG codons were incorrectly annotated as the starting ones not only in *T. vivax* but also (and unexpectedly) in *T. brucei*. After correcting the annotation using conservation information, it was possible to determine that almost all downstream splice sites have in fact an upstream location (see Additional file 13: Figure S7 for representative examples). In addition when the orthologs between these two species are compared in relation to this feature, it is possible to observe that there is a very good agreement, namely the number of splice sites and their distances to the initiation codon is roughly the same (Additional file 14: Table S7). Noteworthy, while these distances remain, there is very little sequence conservation between the two species in the 5' UTR, which strongly suggests that what it is important is indeed the distance and not the sequence. Regretfully this analysis could not be extended to *T. cruzi* due to the limited number of reads that spanned the trans splicing junctions and retained a big enough sequence (>15 nt) after the Spliced leader was removed.

Although the biological significance of these observations is not fully clear, some hypotheses could be advanced on why this particular group of genes contain so

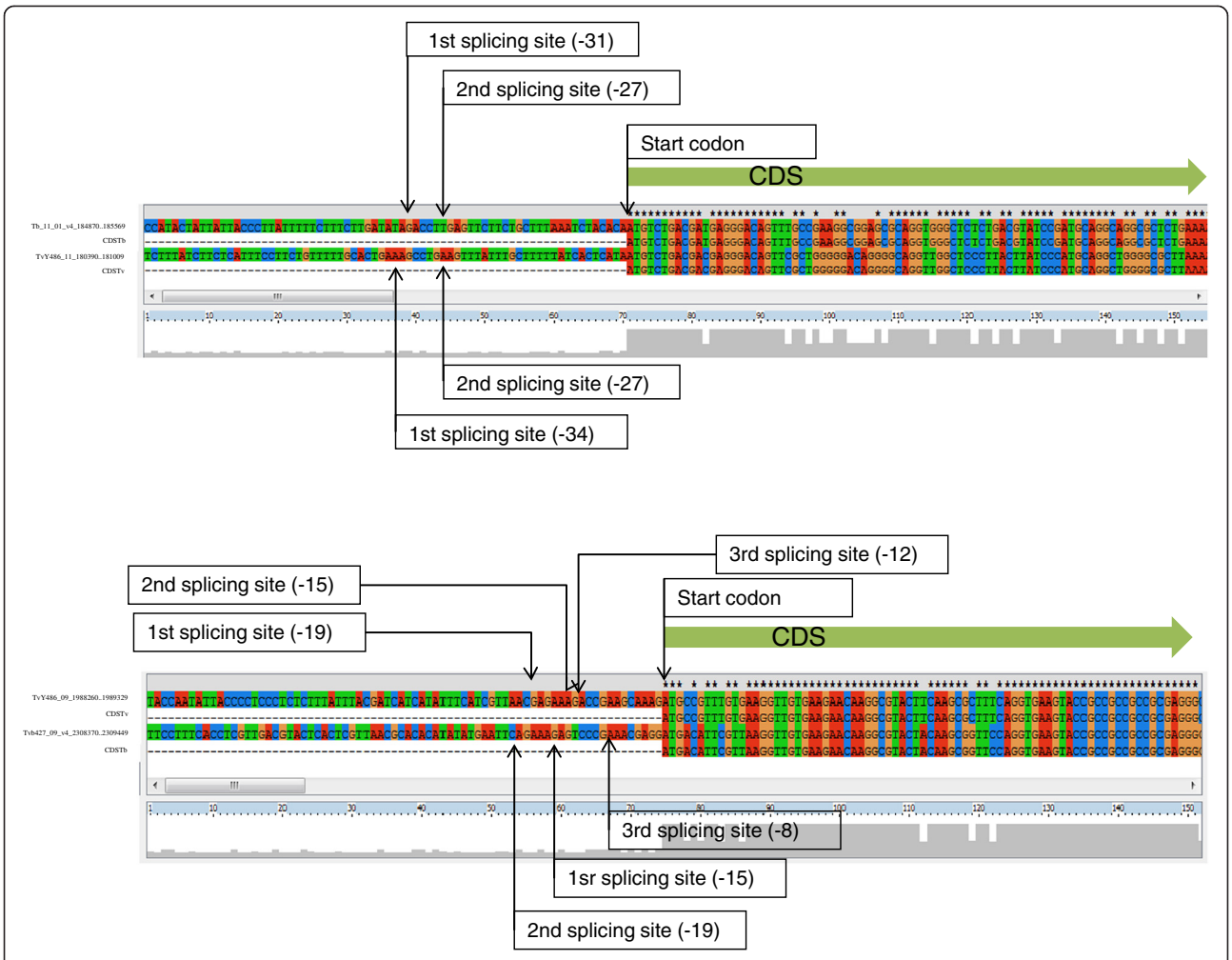


Figure 7 Representative examples of 5′UTR distance conservation. In the two examples (**A** and **B**) the distance to the different splice sites is fairly the same in both species, while sequence conservation in the 5'UTR is low.

short 5' UTR. It has been proposed that highly expressed genes tend to be more compact, shortening their 5' and 3' UTRs and introns to reduce energetic cost of protein synthesis [41]. At a first glance this explanation appears to fit the results presented herein provided that ribosomal proteins are normally highly expressed. However, the average expression level (as indicated by their transcript abundance) of short 5'UTR containing genes is not significantly different to the average expression level of the genome (*T*-test, $p < 0.05$).

Alternatively this feature could be related to genes that are constitutively expressed. This hypothesis becomes clearer if two aspects are taken into consideration; first translation initiation plays a key role in trypanosomatid expression regulation, and second it has been demonstrated in *Leishmania* that the sole presence of a Spliced Leader ensures the recruitment of the 40S ribosome complex to the mRNA 5′ (through the eIF4F initiation complex binding to the 5′ m7G-mRNA cap and/or to the SL itself) [42]. Therefore the lack of a segment between the Spliced Leader and the start codon (to which negative regulators could eventually bind), would imply that once the ribosomal initiation complex is assembled, there is almost no chance of blocking translation initiation. In this regard it is worth mentioning that it has been recently proposed that trypanosomes may contain posttranscriptional cis-regulatory elements located in the 5′ UTRs, which would be part of a mechanism to sense environmental changes (temperature) in a way reminiscent to bacterial RNA thermometers [35]. At any rate, the results presented here give initial hints that would require additional experiments (e.g. constructs containing specific modifications in the 5′ UTR) to test this or alternative hypotheses.

Conclusions

In this work we conducted a RNA-seq analysis in *T. vivax*, a species of great importance for comparative purposes owing to its evolutionary location as the

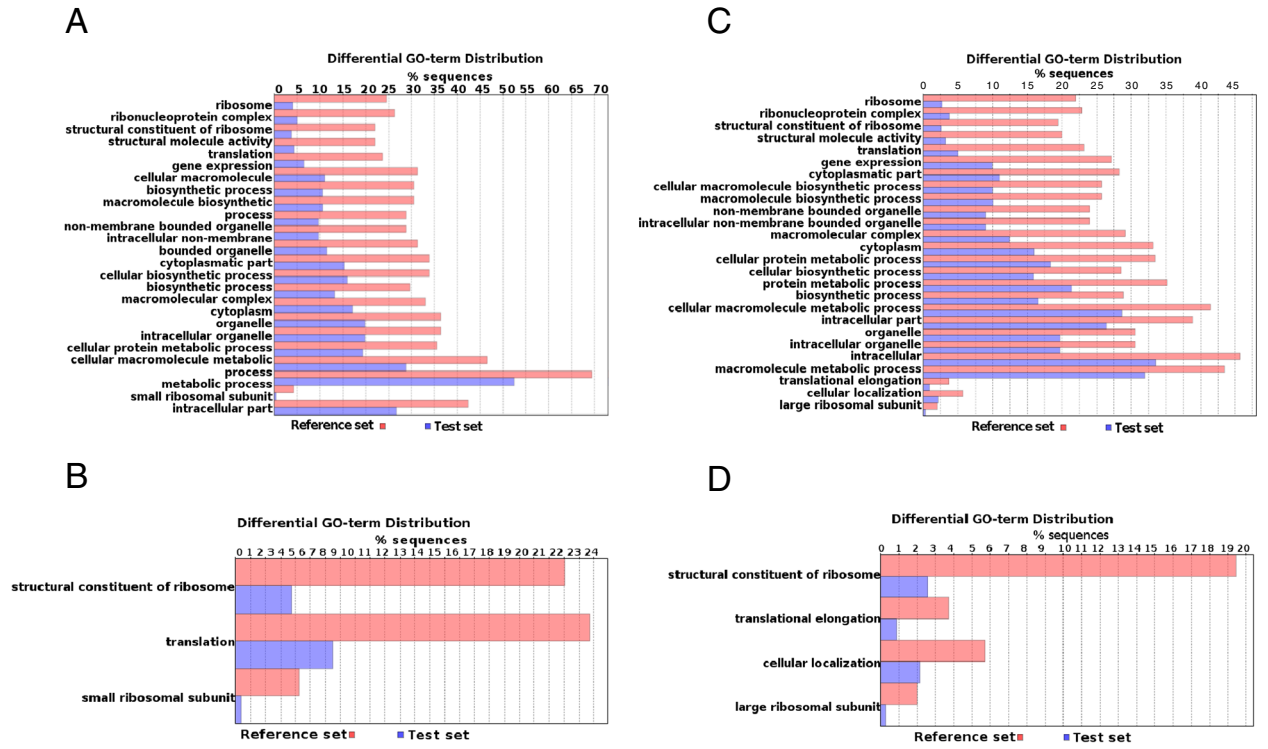


Figure 8 Gene Ontology Enrichment Analysis. These figures show the distribution of GO terms exhibiting statistical significant differences (Fisher Exact Test, filtering p-values for multiple testing using False Discovery Rate). The test set consisted in genes containing splicing site close to start codon (distance < = 10 nucleotides). Panels **A** and **B**: *T. vivax* In this case the testing set of short 5'UTR containing genes is composed by 196 sequences. The two panels correspond to different ontology levels (abstraction levels). Panels **C** and **D** show the equivalent GO analysis in *T. brucei* (n = 654).

earliest branching African trypanosome. To this aim we sequenced the bloodstream stage of its life cycle using two complementary sequencing technologies. The first of these technologies allowed us to obtain a high quality assembly without the restriction of a reference set. The annotation of the contigs thus obtained (using a battery of bioinformatic tools) permitted the identification of about 6500 protein coding genes and other non-coding RNAs. Noteworthy, more than 1000 genes were found to be species specific and about 50 exclusive of *T.vivax* LIEM-176. This information and the partial reconstruction of metabolic pathways, is publicly available through a searchable online database.

The use of Illumina technology in combination with the above mentioned assembly and genomic information was used to analyze several aspects in this species which in turn allowed us to draw relevant conclusions by means of comparative analysis with *T. brucei*.

One first aspect to be emphasized concerns the Variable Surface Glycoproteins, that exhibit levels of expression considerably lower than those observed in *T. brucei*; an observation that is consistent with previous indications obtained from microscopy. This denotes not only that the proteins composition of cellular surfaces is notably different between the two species; but also implies that in all likelihood the way VSG proteins accomplish their shielding role did not remain exactly the same since their emergence. In this regards it is worth reminding that in *T. brucei*, the VSG coat is a dense physical barrier around the parasite, which does largely modulate the ability of immunoglobulins to recognize other surface (invariant) proteins. This point, which is of chief importance to understand the primordial function of VSGs, requires further investigation on diverse aspects such as assessing the level of exposure to the immune system of *T. vivax* invariant surface proteins or determining their efficiency in antibody clearing and the VSG switching rate.

As long as the expression patterns is concerned, we would like to stress that we present in this work evidence that some regions of *T. vivax* genome (that contain coding genes) have no transcriptional activity. In fact, a detailed study shows that vast genomic regions encompassing about one third of the repertoire of variant genes and other regions containing other protein coding sequences are transcriptionally inactive (Lamolle et al., in preparation). This strongly suggests, in contrast to the generally accepted view, that in trypanosomatids the regulation of transcription initiation might also play

an important role in gene expression regulation. This works perhaps by switching off and on entire genome segments, something that might be accomplished by different mechanisms like condensation or loosening of chromatin in specific regions.

Finally, we would like to address the topic of trans-splicing patterns exhibited by *T. vivax*. A first conclusion that can be drawn in relation to this topic, is that the signals recognized by the trans-splicing enzymatic machinery (and thus the machinery itself) are substantially conserved not only in African trypanosomes but also in most distant species like *T. cruzi*. Another significant aspect is that the distance distribution of trans-splice sites, but not the sequence, is conserved for an important proportion of genes. The last important point regarding trans-splicing, is that a group of genes related to translation and interaction with RNAs, contain very short 5'UTR (i.e. the splice site is located just before the start codon). This observation cannot be attributed to any technical (bias in library preparation, sequencing) or bioinformatic (determination of AUG codon) artifact provided that the same pattern is found in both *T. brucei* and *T.vivax*. Although here we suggest some possible explanations and hypotheses that are in line with the regulatory role already proposed for the 5′UTRs in trypanosomatid RNAs, additional data from other trypanosomatid species will allow to determine the phylogenetic extent of this feature; and experiments (such as the use of manipulated DNA segments) would help shed light on its possible functional role.

Additional files

Additional file 1: Table S1. Details of sequence data obtained from 454 FLX and Illumina.

Additional file 2: Figure S1. Coverage Metrics for Top-Middle-Lowest 1000 Expressed Transcripts.

Additional file 3: Table S2. Quality of Assembly. Excel table containing Q values obtained by mira and Newbler assemblers.

Additional file 4: Table S3. Expression levels. Excel table containing expression levels (rpkm values) obtained with erange. Comparison between 454 and Illumina quantification.

Additional file 5: Table S4. Species specific proteins. Sheet 1. List and features of the contigs. Sheet 2. Summary table.

Additional file 6: Figure S2. Sequence alignment of VSG from American and African isolates.

Additional file 7: Figure S3. PCR of VSG. Genomic amplification with VSG specific primers in American and African isolates.

Additional file 8: Table S5. rpkm and percentage of total sequence reads corresponding to VSG and tubulin genes in *T. vivax* and *T. brucei*.

Additional file 9: Figure S4. GC_3 content discriminated by amino acid.

Additional file 10: Table S6. Group of genes having high expression levels (rpkm > average, 3 SD) and low GC3 frequency. MASP and ribosomal genes in *T. cruzi*.

Additional file 11: Figure S5. Comparison of frequencies of G + C ending codons in the most and least expressed genes in *T. vivax* and *T. brucei*. The comparison was done between conserved and non conserved orthologous genes (up and low panels).

Additional file 12: Figure S6. Sequence logo representation of 2nd to 4th trans-splicing sites.

Additional file 13: Figure S7. Examples of Trans-splicing sites in *T. brucei* and *T. vivax* and annotation correction.

Additional file 14: Table S7. *T. vivax* and *T. brucei* orthologs genes and trans splicing sites.

Competing interests
Authors declare that they have no competing interests.

Authors' contributions
GG and GL performed library preparation and sequencing. MPL and GL conducted the assembling of 454 contigs. MPL worked in the de annotation contigs. MPL and MR developed the online database and several Perl and Python scripts. MR took care of codon usage analysis and Perl and Python scripting. GG, MPL, MR, FAV were in charge of bioinformatic analysis (SL location, determination of expression level, splice leader analysis). GG, DP, LTM and ARB were in charge of experimental infection, parasite purification and RNA isolation. ARB was the veterinary that took care of sheep health condition. CR and FAV conceived the work. GG and FAV wrote the manuscript. All authors read and approved the final version of the manuscript.

Acknowledgements
T. vivax bovine Venezuelan isolate (LIEM-176) was a kindly provided by Laura Morón and Glenda Moreno, from Universidad de los Andes, Núcleo Táchira, Venezuela. Cryostabilites of *Trypanosoma vivax* (Y486 strain) were kindly provided by Philippe Büscher (Parasite Diagnostics Unit, Department of Parasitology, Institute of Tropical Medicine Antwerp, Belgium).
We thank Paula Tucci for critical reading of the manuscript and Daniel Ramón and Laia Pedrola (Life Sequencing S.L., Valencia, Spain) for technical assistance and helpful experimental suggestions on 454 sequencing. This work was supported by grants from Fondo Clemente Estable (ANII) and CSIC (Universidad de la República, Uruguay). DP, CR. and FAV are researchers from the Sistema Nacional de Investigadores (ANII), Montevideo Uruguay.

Author details
[1]Unidad de Biología Molecular, Institut Pasteur de Montevideo, Montevideo, CP 11400, Uruguay. [2]Sección Biomatemática, Facultad de Ciencias, Universidad de la Republica Uruguay, Montevideo, Uruguay. [3]Departamento de Bioquímica, Facultad de Medicina, Universidad de la República Uruguay, Montevideo, Uruguay. [4]Centro de Estudios Biomédicos y Veterinarios, Universidad Nacional Experimental Simón Rodríguez-IDECYT, Caracas, Venezuela.

Received: 14 July 2012 Accepted: 15 February 2013
Published: 5 March 2013

References
1. Ferenc SA, Stopinski V, Courtney CH: **The development of an enzyme-linked immunosorbent assay for Trypanosoma vivax and its use in a seroepidemiological survey of the Eastern Caribbean Basin.** *Int J Parasitol* 1990, **20**(1):51–56.
2. Desquesnes M, Dia ML: **Mechanical transmission of Trypanosoma vivax in cattle by the African tabanid Atylotus fuscipes.** *Vet Parasitol* 2004, **119**(1):9–19.
3. Desquesnes M: *Livestock trypanosomoses and their vectors in Latin America.* Paris: OIE & CIRAD; 2004.
4. Hoare CA: *The trypanosomes of mammals: a Zoological Monograph.* London: Blackwell Scientific Publications; 1972.
5. Berriman M, Ghedin E, Hertz-Fowler C, Blandin G, Renauld H, Bartholomeu DC, Lennard NJ, Caler E, Hamlin NE, Haas B, *et al*: **The genome of the African trypanosome Trypanosoma brucei.** *Science* 2005, **309**(5733):416–422.
6. Stevens JR, Teixeira MM, Bingle LE, Gibson WC: **The taxonomic position and evolutionary relationships of Trypanosoma rangeli.** *Int J Parasitol* 1999, **29**(5):749–757.

7. Adams ER, Hamilton PB, Rodrigues AC, Malele II, Delespaux V, Teixeira MM, Gibson W: **New Trypanosoma (Duttonella) vivax genotypes from tsetse flies in East Africa.** *Parasitology* 2010, **137**(4):641–650.
8. Rodrigues AC, Neves L, Garcia HA, Viola LB, Marcili A, Da Silva FM, Sigauque I, Batista JS, Paiva F, Teixeira MM: **Phylogenetic analysis of Trypanosoma vivax supports the separation of South American/West African from East African isolates and a new T. vivax-like genotype infecting a nyala antelope from Mozambique.** *Parasitology* 2008, **135**(11):1317–1328.
9. Jackson AP, Berry A, Aslett M, Allison HC, Burton P, Vavrova-Anderson J, Brown R, Browne H, Corton N, Hauser H, *et al*: **Antigenic diversity is generated by distinct evolutionary mechanisms in African trypanosome species.** *Proc Natl Acad Sci USA* 2012, **109**(9):3416–3421.
10. van Luenen HG, Farris C, Jan S, Genest PA, Tripathi P, Velds A, Kerkhoven RM, Nieuwland M, Haydock A, Ramasamy G, *et al*: **Glucosylated hydroxymethyluracil, DNA base J, prevents transcriptional readthrough in Leishmania.** *Cell* 2012, **150**(5):909–921.
11. Gonzalez LE, Garcia JA, Nunez C, Perrone TM, Gonzalez-Baradat B, Gonzatti MI, Reyna-Bello A: **Trypanosoma vivax: a novel method for purification from experimentally infected sheep blood.** *Exp Parasitol* 2005, **111**(2):126–129.
12. Chamond N, Cosson A, Blom-Potar MC, Jouvion G, D'Archivio S, Medina M, Droin-Bergere S, Huerre M, Goyard S, Minoprio P: **Trypanosoma vivax infections: pushing ahead with mouse models for the study of Nagana. I. Parasitological, hematological and pathological parameters.** *PLoS Negl Trop Dis* 2010, **4**(8):e792.
13. Langmead B, Trapnell C, Pop M, Salzberg SL: **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.** *Genome Biol* 2009, **10**(3):R25.
14. DeLuca DS, Levin JZ, Sivachenko A, Fennell T, Nazaire MD, Williams C, Reich M, Winckler W, Getz G: **RNA-SeQC: RNA-seq metrics for quality control and process optimization.** *Bioinformatics* 2012, **28**(11):1530–1532.
15. Chevreux B, Pfisterer T, Drescher B, Driesel AJ, Muller WE, Wetter T, Suhai S: **Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs.** *Genome Res* 2004, **14**(6):1147–1159.
16. Iseli C, Jongeneel CV, Bucher P: **ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences.** *Proc Int Conf Intell Syst Mol Biol* 1999:138–148.
17. Conesa A, Gotz S: **Blast2GO: A comprehensive suite for functional analysis in plant genomics.** *Int J Plant Genomics* 2008, **2008**:619832.
18. Kelly RJ, Vincent DE, Friedberg I: **IPRStats: visualization of the functional potential of an InterProScan run.** *BMC Bioinformatics* 2010, **11**(12):S13.
19. Otto TD, Guimaraes AC, Degrave WM, de Miranda AB: **AnEnPi: identification and annotation of analogous enzymes.** *BMC Bioinformatics* 2008, **9**:544.
20. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B: **Mapping and quantifying mammalian transcriptomes by RNA-Seq.** *Nat Methods* 2008, **5**(7):621–628.
21. Siegel TN, Hekstra DR, Wang X, Dewell S, Cross GA: **Genome-wide analysis of mRNA abundance in two life-cycle stages of Trypanosoma brucei and identification of splicing and polyadenylation sites.** *Nucleic Acids Res* 2010, **38**(15):4946–4957.
22. Langenberger D, Bermudez-Santana C, Hertel J, Hoffmann S, Khaitovich P, Stadler PF: **Evidence for human microRNA-offset RNAs in small RNA sequencing data.** *Bioinformatics* 2009, **25**(18):2298–2301.
23. Carbon S, Ireland A, Mungall CJ, Shu S, Marshall B, Lewis S: **AmiGO: online access to ontology and annotation data.** *Bioinformatics* 2009, **25**(2):288–289.
24. Gardiner PR, Nene V, Barry MM, Thatthi R, Burleigh B, Clarke MW: **Characterization of a small variable surface glycoprotein from Trypanosoma vivax.** *Mol Biochem Parasitol* 1996, **82**(1):1–11.
25. Cortez AP, Ventura RM, Rodrigues AC, Batista JS, Paiva F, Anez N, Machado RZ, Gibson WC, Teixeira MM: **The taxonomic and phylogenetic relationships of Trypanosoma vivax from South America and Africa.** *Parasitology* 2006, **133**(Pt 2):159–169.
26. Fabre H, Bernard M: **Sur un nouveau foyer de Trypanosomiase Bovine observé a la Guadaloupe.** *Bull Soc Path Exot* 1926, **19**:435–437.
27. Carougeau M: **Trypanosomiase bovine à la Guadeloupe.** *Bull Soc Path Exot* 1929, **22**:246–247.
28. Leger M, Vienne M: **Epizootic á Trypanosomes chez les bovides de la Guyane Francaise.** *Bull Soc Path Exot* 1919, **12**:258–266.
29. Curasson G: *Trypanosoma vivax et variétés.* Paris: Vigot Frères; 1943. vol. Tome 1.
30. Levine N: *The hemoflagellates.* 2nd edition. Minneapolis: Burgess Publishing; 1973.
31. Hutchinson OC, Picozzi K, Jones NG, Mott H, Sharma R, Welburn SC, Carrington M: **Variant Surface Glycoprotein gene repertoires in Trypanosoma brucei have diverged to become strain-specific.** *BMC Genomics* 2007, **8**:234.
32. Kolev NG, Franklin JB, Carmi S, Shi H, Michaeli S, Tschudi C: **The transcriptome of the human pathogen Trypanosoma brucei at single-nucleotide resolution.** *PLoS Pathog* 2010, **6**(9):e1001090.
33. Nilsson D, Gunasekera K, Mani J, Osteras M, Farinelli L, Baerlocher L, Roditi I, Ochsenreiter T: **Spliced leader trapping reveals widespread alternative splicing patterns in the highly dynamic transcriptome of Trypanosoma brucei.** *PLoS Pathog* 2010, **6**(8):e1001037.
34. Vickerman KP TM: *Biology of the kinteplastida.* London: Academic; 1976.
35. Kramer S: **Developmental regulation of gene expression in the absence of transcriptional control: the case of kinetoplastids.** *Mol Biochem Parasitol* 2012, **181**(2):61–72.
36. Grantham R, Gautier C, Gouy M, Mercier R, Pave A: **Codon catalog usage and the genome hypothesis.** *Nucleic Acids Res* 1980, **8**(1):r49–r62.
37. Sharp PM, Emery LR, Zeng K: **Forces that influence the evolution of codon bias.** *Philos Trans R Soc Lond B Biol Sci* 2010, **365**(1544):1203–1212.
38. Alvarez F, Robello C, Vignali M: **Evolution of codon usage and base contents in kinetoplastid protozoans.** *Mol Biol Evol* 1994, **11**(5):790–802.
39. Horn D: **Codon usage suggests that translational selection has a major impact on protein expression in trypanosomatids.** *BMC Genomics* 2008, **9**:2.
40. Michels PA: **Evolutionary aspects of trypanosomes: analysis of genes.** *J Mol Evol* 1986, **24**(1–2):45–52.
41. Li SW, Feng L, Niu DK: **Selection for the miniaturization of highly expressed genes.** *Biochem Biophys Res Commun* 2007, **360**(3):586–592.
42. Zeiner GM, Sturm NR, Campbell DA: **The Leishmania tarentolae spliced leader contains determinants for association with polysomes.** *J Biol Chem* 2003, **278**(40):38269–38275.

doi:10.1186/1471-2164-14-149
Cite this article as: Greif *et al.*: Transcriptome analysis of the bloodstream stage from the parasite *Trypanosoma vivax*. *BMC Genomics* 2013 **14**:149.

Parte B. Estudio evolutivo del genoma mitocondrial de *T. vivax*

En esta sección, se presentan los resultados relacionados con la comparación del genoma mitocondrial de diferentes cepas de *T. vivax* con el fin de realizar un análisis comparativo. En particular nos interesa investigar los cambios producidos durante el proceso de adaptación a la transmisión mecánica en las cepas americanas (Objetivo General Nº2).

Particularmente de los resultados obtenidos en esta sección queremos resaltar:

1. Obtención de la secuencia completa del maxicírculo de *T. vivax* (tanto de la cepa de referencia Y486, como de dos cepas americanas –Liem-176 y MT1-).

2. Análisis sobre el *editing* de genes mitocondriales y comparación entre las tres cepas.

3. Obtención del minicirculoma (set completo de secuencias de minicírculo) para la cepa americana MT1 e implicancias en la evolución de estos parásitos.

Secuenciación del genoma mitocondrial de diferentes cepas de *T. vivax*

Se secuenciaron los genomas (ADN nuclear y mitocondrial) de las cepas americanas MT1 y Liem-176. A partir de estos datos se identificaron por homología con el genoma mitocondrial de *T. brucei* las secuencias correspondientes al maxicírculo. Se obtuvo la secuencia del maxicírculo completa para la cepa MT1, con secuenciación profunda (Illumina, MiSeq y GaIIx) y secuenciación Sanger. En concreto, se realizó un refinamiento de la secuencia con la amplificación y secuenciado por Sanger de regiones específicas. Para el caso del maxicírculo de Liem-176, debido a la menor calidad en el secuenciado profundo realizado, se completó el ensamblaje utilizando las secuencias de RNAseq generadas para el primer trabajo en el equipo 454 (estas secuencias, de mayor tamaño que las obtenidas en la secuenciación por Illumina, permitieron el ensamblado final de este genoma). Por último, para el ensamblaje del maxicírculo de Y486, se utilizaron las secuencias disponibles en el GeneDB. Los detalles del ensamblaje de estos tres genomas mitocondriales se encuentran en el material suplementario (Anexo 2, Archivo suplementario 1 y Figura Suplementaria 1).

La comparación entre los tres genomas mitocondriales (maxicírculos) se muestra en la la Figura 27. Por un lado se observaron cambios nucleotídicos y deleciones en las cepas americanas que implicarían efectos deletéreos (Tabla 5). Estos no están presentes en la cepa africana, indicando que la ocurrencia de estos cambios fue posterior a la introducción de estos parásitos en América. Cuando comparamos las secuencias de las cepas americanas (MT1 vs Liem-176) sólo se observaron 6 cambios de bases, que en principio no afectan las proteínas codificadas en estos genomas.

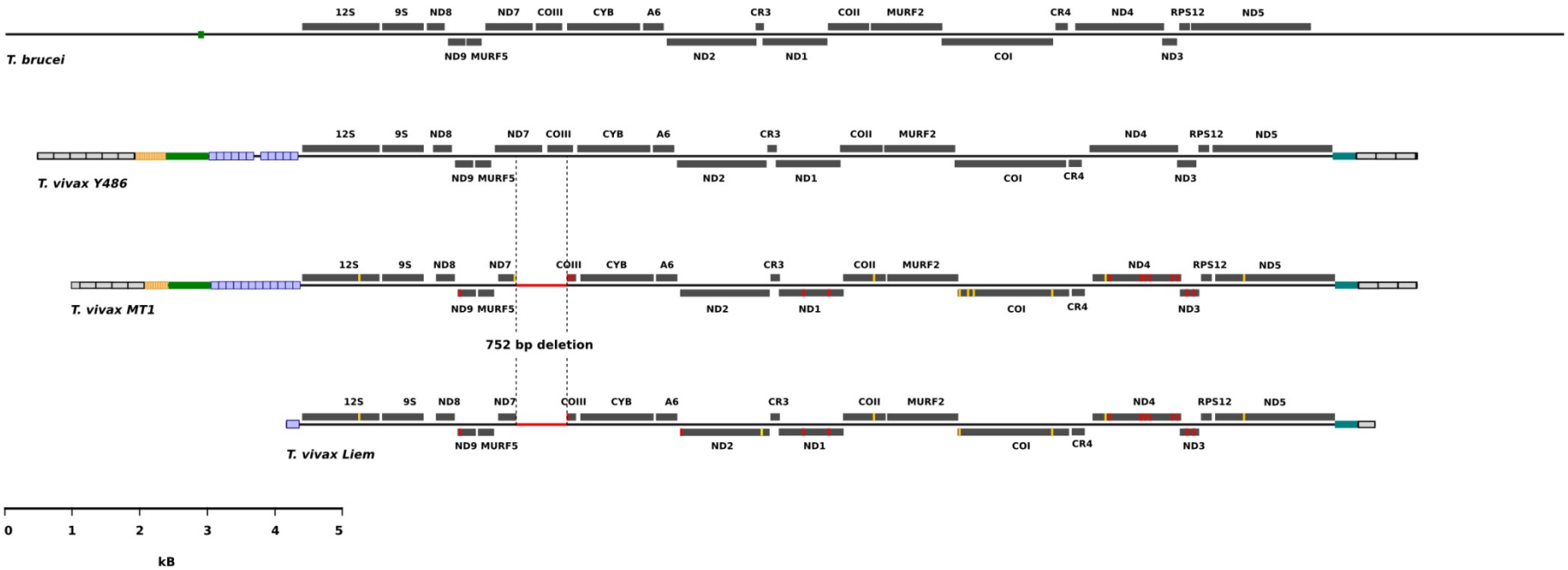


Figura 27. Representación y alineamiento de las secuencias completas del maxícirculo de tres cepas de *T. vivax.* En la figura se incluye, como referencia, la representación del maxicírculo de *T. brucei.* Los cuadros grises indican la posición de los genes y se indican sus nombres. Las mutaciones en los maxicírculos de MT1 y Liem-176 se indican con líneas verticales. Los cuadrados celestes representan secuencias repetidas en *cluster* de 105 pb. Los cuadros naranjas representan zonas repetidas de 24 pb y los cuadrados grises representan repetidos de 175 pb. La zona verde representa una región especie-específica, no repetitiva.

Tabla 5. Lista de genes del maxicírculo. Se muestra el tipo de edición, las mutaciones observadas y su efecto.* En inserciones, se indican las posiciones anterior y posterior a la inserción. En deleciones se indica la posición deletada. ** Cepa dónde se observa la mutación.

Gen	*Editing*	Posición: cambio *	Efecto esperado de la mutación (aa)	Cepa **
1. ND8	Pan-editado	---	---	---
2. ND9	Pan-editado	21_22: ins T	Posible *frameshift*	MT1, Liem
3 MURF5	No editado	---	---	---
4. ND7	Pan-editado	248: G>T 275_699: del 424 bp	ND Deleción mayor	MT1 MT1, Liem
5. COIII	Pan-editado	1_248: del 248 bp 253: G>C 261_262: del C 303_304: ins AG	Deleción mayor ND Posible *frameshift* Posible *frameshift*	MT1, Liem MT1, Liem MT1 MT1, Liem
6. Cyb	5' editado	---	---	---
7. A6-ATPase	Pan-editado	----	---	---
8. ND2 (MURF1)	No editado	16_19: del ATAC 1212:A>G	*Frameshift* Mutación puntual	Liem Liem
9. CR3	Pan-editado	----	----	---
10. ND1	No editado	491_492:ins T 863: del A	*Frameshift* Recupera ORF	MT1, Liem MT1, Liem
11. COII	*Editing* parcial	456_457: ins TGC	Inserción de 1 aa (ins C)	MT1, Liem
12. MURF2	5' editado	---	---	---
13. COI	No editado	24: T>A 154: G>T 237: G>A 1399: G>A	*Missense* (C>W) *Missense* (G>C) *Missense* (G>S) *Missense* (V>I)	MT1, Liem MT1 MT1 MT1, Liem
14. CR4	Pan-editado	---	---	---
15. ND4	Not editado	194: C>T 254_259: del ATATAC 712_713: ins TT 778: del T 846: G>A 1175_1176: ins AT 1257_1258: del AT	Missense(A>V) Deleción de 2 aa (del MY) *Frameshift* *Frameshift* *Sinónima* Recupera ORF *Frameshift*	MT1, Liem MT1, Liem MT1, Liem MT1, Liem MT1, Liem MT1, Liem MT1, Liem
16. ND3	Pan-editado	105_106: del TT 205_206: del TT	*Posible frameshift* *Posible frameshift*	MT1, Liem MT1, Liem
17. RPS12	Pan-editado	---	---	--
18. ND5	No editado	430: G>T	*Missens*e (G>C)	MT1, Liem

El mayor cambio entre las cepas americanas y la cepa Y486 es una deleción de 752 bases que afecta la región codificante de dos proteínas (COIII y ND7) (Tabla 5, Figura 27) de las cepas americanas. Esta deleción implica que, en las cepas americanas, estas proteínas (COIII y ND7) no son funcionales. Asimismo, se observaron *indels* en otros 6 genes que producen cambios en el marco abierto de lectura, que, si no fueran corregidos post-transcripcionalmente, provocarían traducciones no funcionales.

Para determinar si estos cambios eran revertidos post-transcripcionalmente (por ejemplo, mediante el *editing* que sufren los genes mitocondriales), evaluamos datos de RNAseq. En la Figura 28 se muestra un ejemplo de como los cambios a nivel genómico no son corregidos a nivel post-transcripcional, ya que la secuencia transcripta es igual a la secuencia genómica. Asimismo, estos datos indican que la población de maxicírculos no presenta heteroplasmia, es decir es "homocigota" para estas mutaciones. Más ejemplos pueden observarse en Figura Suplementaria 2, Anexo 2.

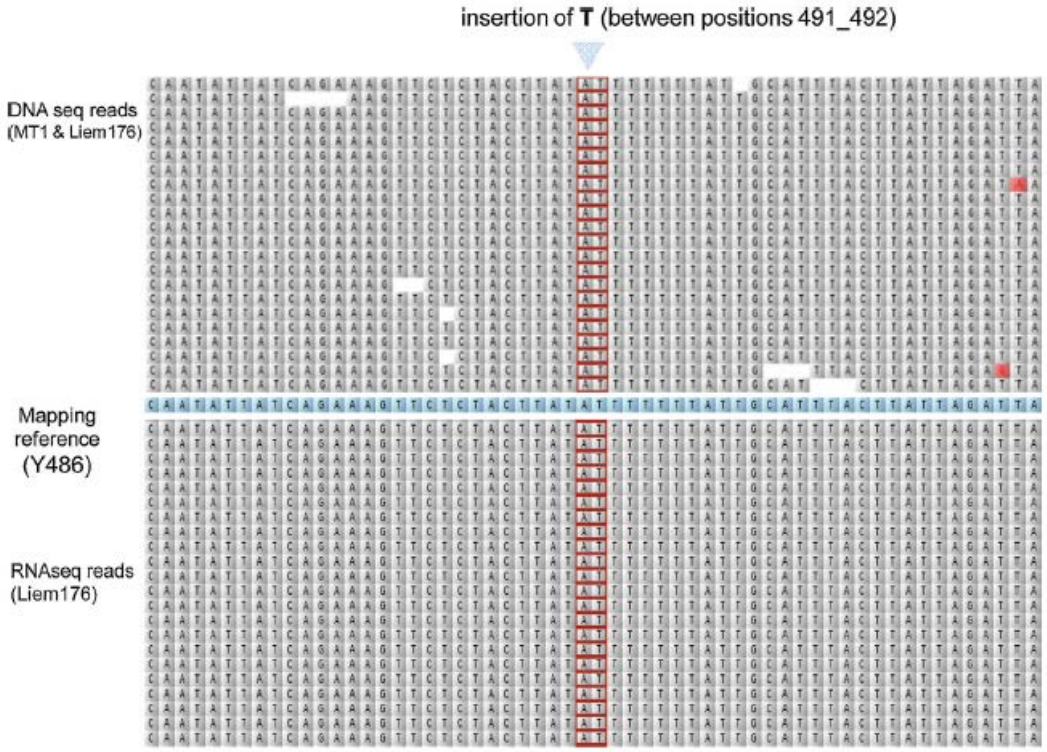


Figura 28. Mapeo de *reads* de DNAseq y RNAseq sobre secuencia del gen ND1. Se muestra la inserción T entre las posiciones 491 y 492 del gen ND1 en las cepas americanas (datos DNAseq, parte superior), y cómo la misma no es corregida post-transcripcionalmente (datos RNAseq, parte inferior). Asimismo se observa que los maxicírculos no presentan heteroplasmia.

Los cambios observados, entonces, muestran una pérdida importante de información genómica mitocondrial de las cepas americanas de *T. vivax*. Esto resultaría en la pérdida completa de funcionalidad de la misma pues al menos 10 genes mitocondriales presentan mutaciones inhabilitantes (Tabla 5).

Editing de genes mitocondriales.

Respecto al segundo punto, estudiamos si los genes mitocondriales eran editados en las cepas americanas. Cómo se mencionó en la introducción, varios genes mitocondriales sufren procesos de *editing* post-transcripcional mediante la inserción/deleción de residuos de uridina que restauran marcos abiertos de lectura y permite la traducción de proteínas funcionales. Estos cambios pueden ser muy pequeños (por ejemplo, en el gen COII se adicionan 4 residuos de uridina), hasta grandes cambios (adición de cientos de residuos de uridina, por ejemplo el caso de los genes ND8 y ND9). En la Tabla 5 se indican los diferentes grados de edición de los 18 genes mitocondriales en *T. vivax* (pan-editados, parcialmente editados o no editados).

Para poder determinar si el *editing* se lleva adelante correctamente fue necesario identificar las secuencias de ARNm maduras, es decir post-editadas. Esto se llevó a cabo usando datos de RNAseq de la cepa Y486 (detalles en la sección 3.2 en [102]). Luego de identificar las secuencias editadas (ARNm maduro) para cada gen se mapearon las secuencias de RNAseq de la cepa Liem-176. Esto permitió observar que si bien se transcriben todos los genes mitocondriales, únicamente encontramos el ARNm editado en 3 de los 12 genes que requieren edición. Estos genes son: la subunidad A6 de la ATP sintasa mitocondrial (A6-ATPase), la proteína ribosomal S12 (RPS12) y MURF2. Para los restantes 9 genes, no se obtuvieron *reads* que mapearan cuando utilizamos como referencia la secuencia de los genes editados. En la Figura 29 se muestra el mapeo de secuencias sobre el gen COII que evidencia la ausencia de *reads* que correspondan al transcripto editado (maduro). En el Anexo 3 de esta tesis, se muestran más ejemplos de genes que requieren *editing* parcial (Cyb) o deben ser pan-editados (ND3), donde no se observa mapeo de *reads* de RNAseq cuando se utiliza el transcripto maduro (editado) como referencia en las cepas americanas (Liem-176 y MT1) y sí hay mapeo de *reads* sobre el transcripto maduro en la cepa africana (Y486). Asimismo, en el Anexo 3, se muestra el caso del gen RPS12, que sí es editado completamente tanto en las cepas americanas como en la cepa africana.

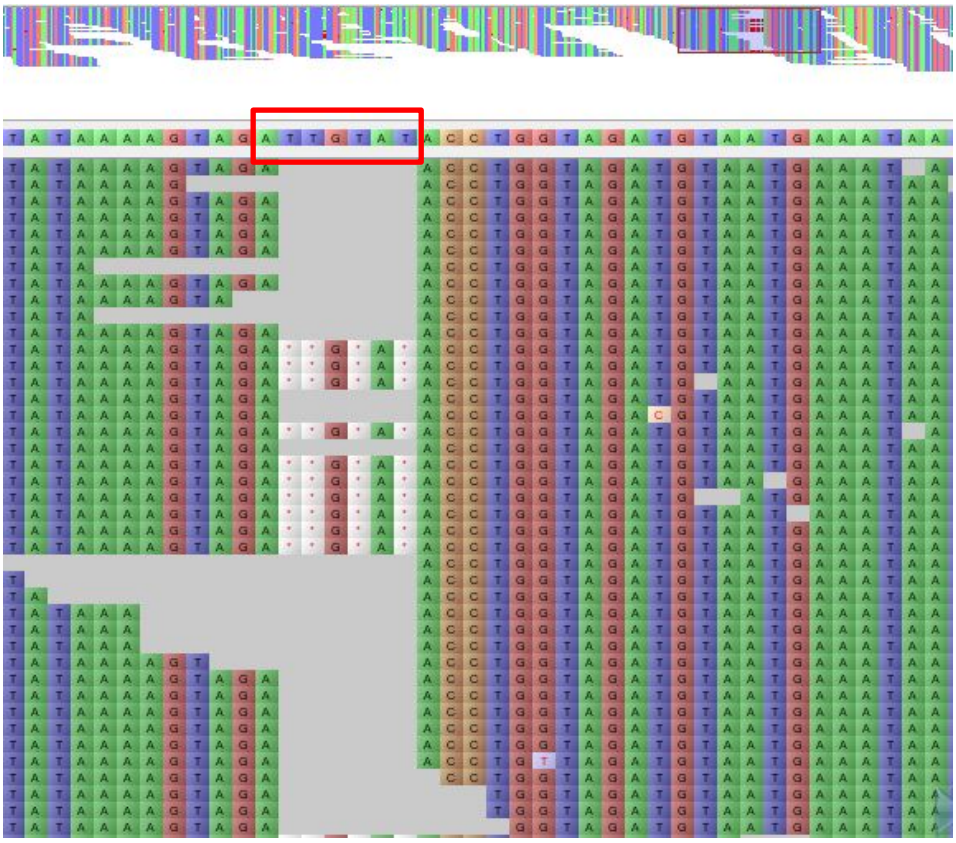


Figura 29. Mapeo de *reads* de RNAseq sobre secuencia del ARNm maduro (editada) del gen COII. En la figura se muestra la ausencia de *reads* que mapeen en la región editada (AuuGuAu, recuadro rojo, en minúscula se muestran las residuos que se adicionan en el ARNm editado). El mapeo se realizó permitiendo este tipo de alineamiento.

Se analizó además si la cepa africana Y486 (utilizando datos públicos de RNAseq, Número de acceso SRA: PRJEB3234) es capaz de realizar *editing* correctamente. Los mapeos de *reads* muestran que todos los genes de esta cepa son transcriptos y correctamente editados (Figura 30) tanto en las formas sanguíneas como en las procíclicas.

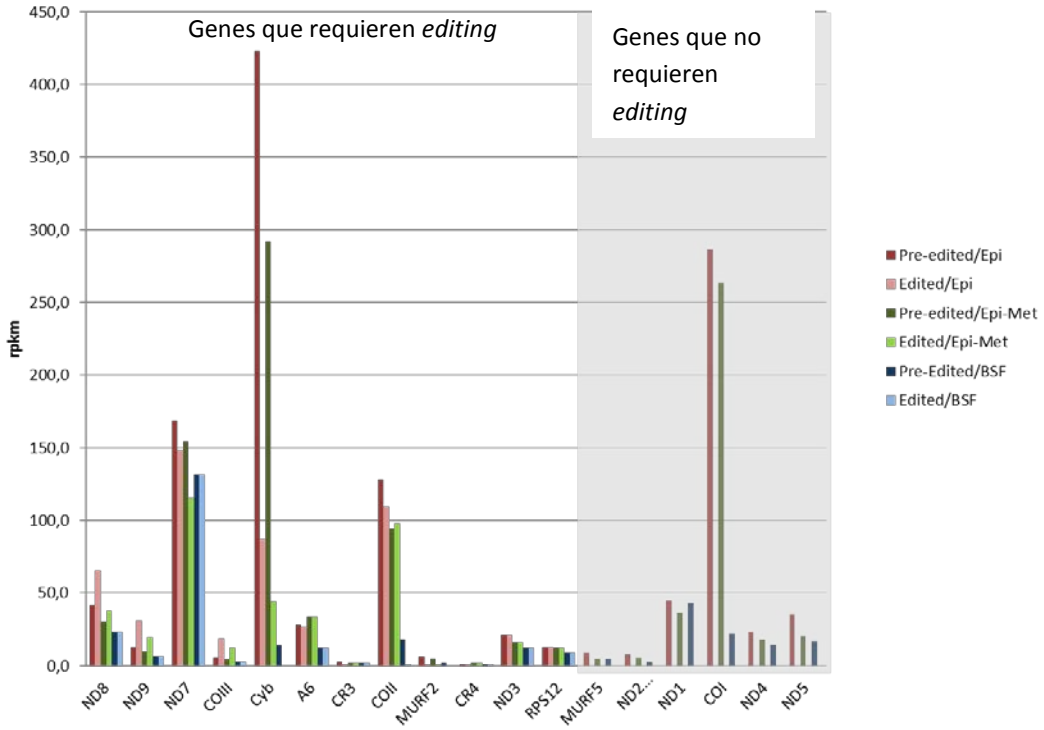


Figura 30. Niveles de expresión de genes mitocondriales en la cepa Y486. Se muestran los valores de expresión (rpkm) para todos los genes mitocondriales pre-editados y editados completamente en diferentes formas del parásito. Epi: epimastigota. Epi-Met: epimastigotas con 20% de metacíclicos-*like*. BSF: forma sanguínea (*bloodstream form*).

Determinación de minicirculoma de *T. vivax*.

El *editing* de genes requiere de los ARNs guía que mayormente se sintetizan en los minicírculos. Para evaluar si la ausencia de *editing* estaba vinculada a alteraciones a nivel del minicirculoma se realizó el ensamblaje del set de minicírculos de las cepas americanas.

En el caso de la identificación de los minicírculos se utilizaron 2 estrategias. Por un lado la búsqueda por homología, para lo cual utilizamos una secuencia de 120 bases que contiene tres bloques que se conservan en todos los tripanosomátidos, denominados CSB-1 a 3 (*Conserved Sequences Block*) [72]. El bloque CSB-3, también conocido como UMS (*Universal Minicircle Sequence*), presenta una secuencia de 12 nucleótidos, completamente conservada en todos los tripanosomátidos. El bloque CSB-1 es más pequeño y menos conservado (Figura 32). La búsqueda por homología utilizando las CSB de *T. brucei* permitió, en una primera etapa, la identificación de 18 secuencias de minicírculos (Blast HSP *e-value* < $1e^{-15}$). La secuencia de 120 bases de estos 18 *contigs* identificados fue utilizada para realizar una segunda ronda de búsqueda por homología. Además, se utilizó una segunda estrategia basada en las propiedades

estadísticas de las secuencias (análisis de componentes principales utilizando la frecuencia de dinucleótidos de los *contigs* ensamblados). En la siguiente figura (Figura 31) se observa como las secuencias correspondientes a los minicírculos agrupan en un *cluster*, separado de los *contigs* correspondientes al genoma y al maxicírculo.

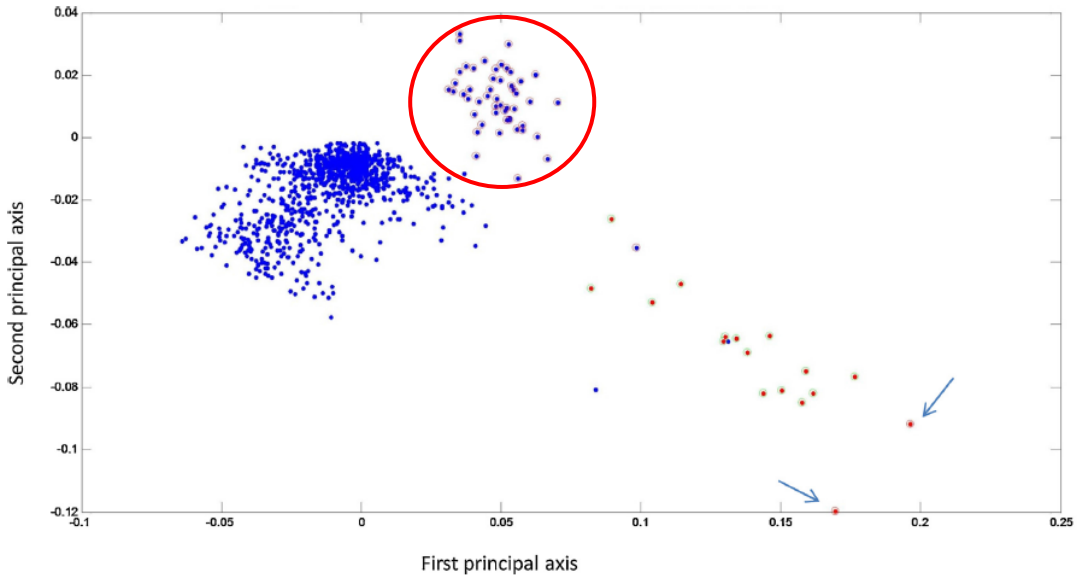


Figura 31. Análisis de componentes principales de la frecuencia de nucleótidos en *contigs* genómicos. Los puntos azules corresponden a secuencias (*contigs*) de origen nuclear, los puntos azules con borde rojo (dentro del círculo rojo en la figura) corresponden a *contigs* con secuencias de minicírculo. Los puntos rojos con borde verde corresponden a secuencias del maxícirculo, las flechas indican dos *contigs* con secuencias repetidas del maxicírculo.

La combinación de ambas fuentes de información permitió la identificación de 54 clases de secuencias de minicírculo en MT1 y 46 en Liem-176 (Figura 32).

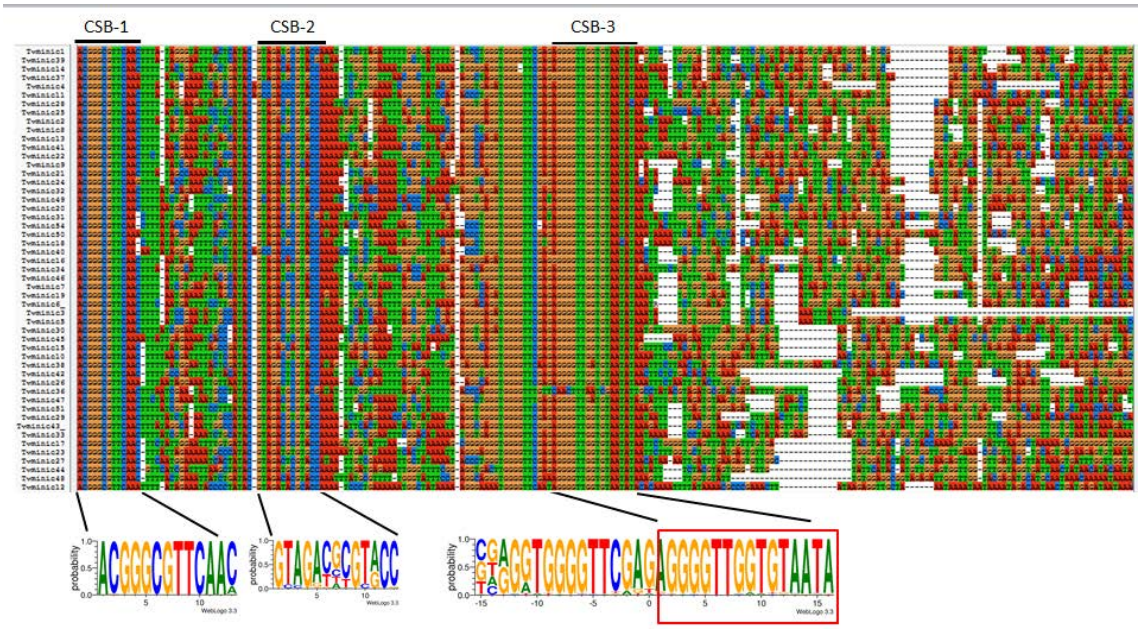


Figura 32. Alineamiento múltiple de secuencias de minicírculos. En la figura se observa el alineamiento de las diferentes clases de minicírculos de la cepa MT1. Arriba se indica la ubicación de los tres bloques de conservación (CSB 1 a 3). En la parte inferior se muestra la conservación de cada bloque.

Notablemente, encontramos un set de minicírculos reducido (54 clases en MT1 y 46 en Liem-176); recordemos que en *T. brucei* el set de clases de minicírculos que se requieren para la edición de todos los genes se estima entre 300 y 400.

Luego de la identificación de las secuencias correspondientes a minicírculos, realizamos la búsqueda de ARNs guía de los genes editados. Como esperábamos, prácticamente la mayoría de los ARNg encontrados (más del 66%, 42 ARNg) corresponden a aquellos requeridos para el *editing* de los genes que efectivamente son editados (A6-ATPase y RPS12). Estos representan practicamente la totalidad de ARNg necesarios para estos genes, los cuales se muestran en la Tabla 6 y en el material suplementario (ver Figura Suplementaria 7, Anexo 2). Se encontraron 21 ARNs guía que participan en la edición de los restantes genes mitocondriales (Tabla 6), aunque la gran mayoría de ellos (15) se encuentra en minicírculos que contienen además ARNs guía de A6-ATPase o de RPS12, lo cual indicaría que los ARNg correspondientes a los restantes genes se han mantenido como “polizones” en minicírculos que portan guías para los genes efectivamente editados. Únicamente 6 minicírculos presentan ARN guías no involucrados en la edición de A6-ATPase y RPS12, Esto se puede apreciar en la Figura 33.

Tabla 6. Detalle de ARNg encontrados para cada gen que requiere *editing*.

ARNg para ND8			
Gen inicio-fin	**Minicírculo**	**Mismatches**	**largo**
108-71	TvMinic15	6	38
102-79	TvMinic45	1	24
181-153	TvMinic3 *	4	30
181-153	TvMinic5	4	30
302-252	TvMinic7	6	51
ARNg para ND9			
Gen inicio-fin	**Minicírculo**	**Mismatches**	**largo**
78-40	TvMinic29	12	39
101-66	TvMinic6	7	36
128-97	TvMinic33	9	33
247-223	TvMinic6	3	26
383-341	TvMinic14	10	43
415-385	TvMinic24	5	31
ARNg para ND7			
Gen inicio-fin	**Minicírculo**	**Mismatches**	**largo**
110-76	TvMinic19	5	35
315-277	TvMinic41	7	39
434-405	TvMinic2	5	30
721-687	TvMinic44	4	36
721-695	TvMinic44	3	28
ARNg para COIII			
Gen inicio-fin	**Minicírculo**	**Mismatches**	**largo**
90-64	TvMinic47	5	27
262-226	TvMinic38	13	38
ARNg para Cyb			
Gen inicio-fin	**Minicírculo**	**Mismatches**	**largo**
ARNg no detectado			
ARNg para CR3			
Gen inicio-fin	**Minicírculo**	**Mismatches**	**largo**
159-140	TvMinic45	3	20
227-191	TvMinic54	7	37
ARNg para COII			
Gen inicio-fin	**Minicírculo**	**Mismatches**	**largo**
No gRNA detected			
ARNg para MURF2			
Gen inicio-fin	**Minicírculo**	**Mismatches**	**largo**
No gRNA found			
ARNg para CR4			
Gen inicio-fin	**Minicírculo**	**Mismatches**	**largo**
27-1	TvMinic28	4	27
210-182	TvMinic46	2	29
334-289	TvMinic26	8	46
412-364	TvMinic42	11	49
ARNg para ND3			
Gen inicio-fin	**Minicírculo**	**Mismatches**	**largo**
ARNg no detectado			

ARNg para A6			
Gen inicio-fin	**Minicírculo**	**Mismatches**	**largo**
39-1	TvMinic34	6	39
60-27	TvMinic14	5	34
81-43	TvMinic54	10	39
97-63	TvMinic8	7	35
111-75	TvMinic31	7	37
141-98	TvMinic35	12	44
166-116	TvMinic40	13	51
190-154	TvMinic51	8	37
166-192	TvMinic16 *	16	40
237-204	TvMinic52 *	5	34
266-210	TvMinic1	18	60
267-217	TvMinic29	13	51
318-273	TvMinic49	11	46
345-307	TvMinic25	12	39
366-330	TvMinic32	8	37
389-362	TvMinic11	3	28
425-378	TvMinic19	10	48
433-399	TvMinic37 *	5	35
462-414	TvMinic38	5	49
509-473	TvMinic48	7	37
548-513	TvMinic36 *	6	36
559-525	TvMinic13	7	35
573-542	TvMinic16 *	3	32
604-558	TvMinic20	12	47
614-585	TvMinic45	4	30
646-627	TvMinic30	2	20
673-635	TvMinic27	7	39
686-656	TvMinic44 *	4	31
711-673	TvMinic33	9	39
724-689	TvMinic46	7	35
754-715	TvMinic10	8	40
ARNg para RSP12			
Gen inicio-fin	**Minicírculo**	**Mismatches**	**largo**
34-1	TvMinic9	7	34
65-21	TvMinic6	12	45
51-32	TvMinic35	2	20
104-76	TvMinic21	2	29
124-93	TvMinic28	6	32
150-108	TvMinic24	11	43
175-140	TvMinic23	10	36
192-160	TvMinic18	6	34
218-180	TvMinic42	9	39
247-204	TvMinic26	7	44
242-207	TvMinic41	9	36
255-238	TvMinic50	1	18

* Minicírculo no detectado en Liem-176

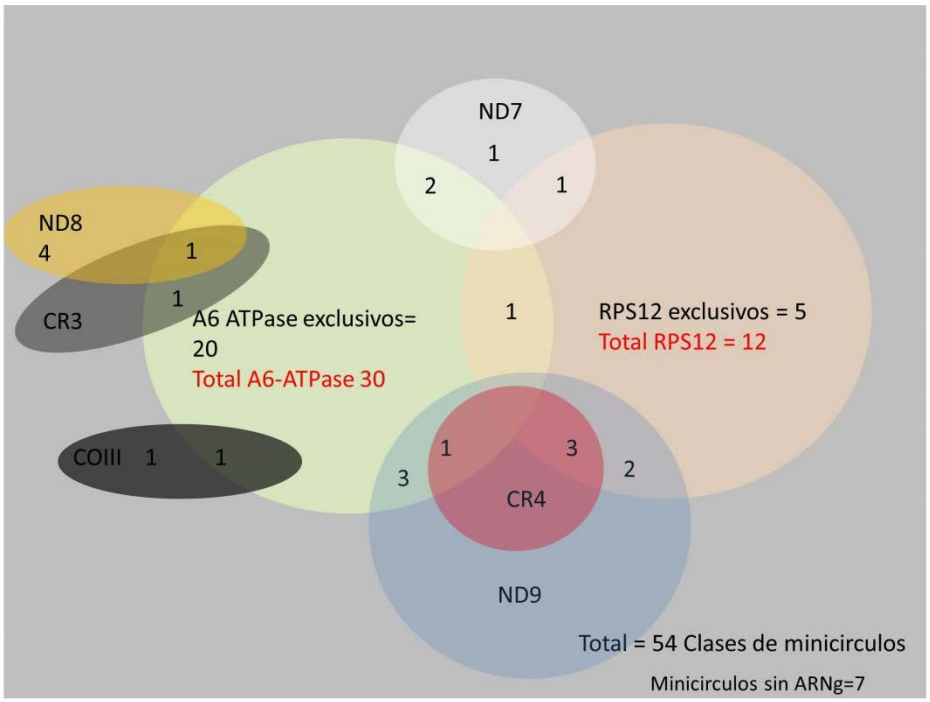


Figura 33. Diagrama de Venn mostrando la cantidad de ARNg encontrados para cada gen. Únicamente 6 minicírculos presentan ARN guías no involucrados en la edición de A6-ATPase y RPS12. En 7 minicírculos no se encontraron ARNg para ningún gen.

Los datos hasta aquí presentados indican que las cepas americanas, al no completar su ciclo (en América no se encuentra el insecto vector y los parásitos únicamente realizan la fase sanguínea en el mamífero) han prescindido en gran parte de la función mitocondrial. Recordemos que en las formas sanguíneas la mitocondria no posee una cadena respiratoria funcional, por lo que no realiza fosforilación oxidativa, y mantiene el potencial de membrana mitocondrial interna por la acción de la ATP sintasa mitocondrial [126]. En cambio en el insecto, donde los nutrientes son más limitados, la mitocondria es completamente funcional.

La fuente de energía más abundante en el insecto vector es principalmente Prolina, y posiblemente sea la principal fuente de generación de ATP (mediante fosforilación oxidativa mitocondrial) en las formas procíclicas de los parásitos. En contraste, la glucosa es abundante en las sangre de los hospederos mamíferos y por ello la glucólisis es el principal mecanismo de generación de ATP en las parásitos sanguíneos [126], no requiriéndose entonces esta función mitocondrial en la producción energética. La mitocondria sin embargo es aún requerida para llevar adelante otras funciones tales como el mantenimiento de su potencial de membrana, el transporte de metabolitos y tRNAs desde el citosol [127, 128], o la síntesis de ácidos grasos [129].

Resulta interesante analizar lo que sucede en los parásitos *T. brucei evansi* y *T. brucei equiperdum*, dos descendientes de *T. brucei* que permanecen exclusivamente en el

hospedero mamífero y han perdido la capacidad de sobrevivir y reproducirse en el insecto (mosca *tse-tsé*). Estos parásitos han perdido en forma completa o casi completa sus genomas mitocondriales. Un paso crítico en la adaptación a la vida independiente de los genes mitocondriales, manteniendo las funcionalidades requeridas en la etapa sanguínea (descritas en el párrafo anterior) son las modificaciones en la subunidad γ de la ATP sintasa nuclear.

Algunos trabajos han mostrado cómo mutaciones en la ATP sintasa nuclear compensan la pérdida de función de la enzima sintetizada en la mitocondria [130], como se esquematiza en la Figura 34.

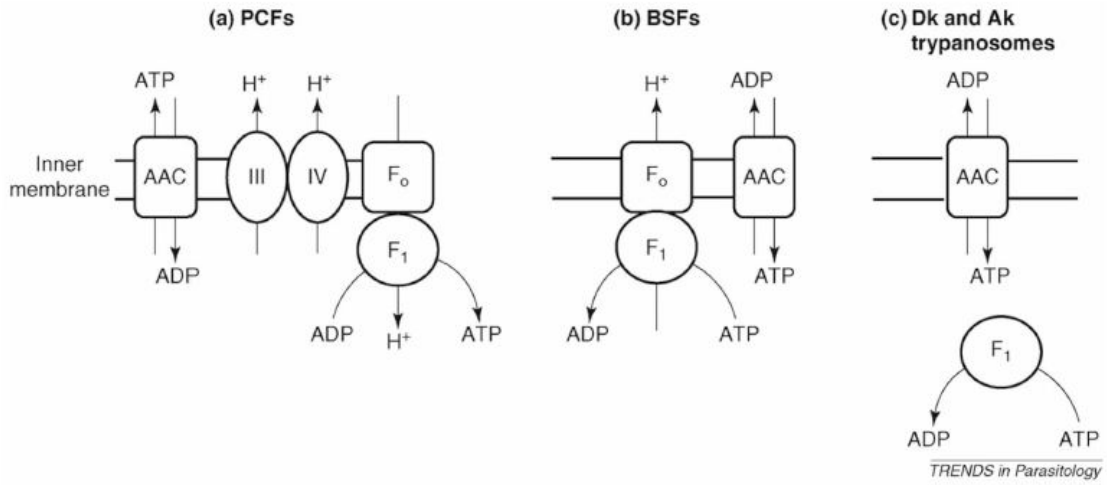


Figura 34. Esquema de función de la ATPasa mitocondrial en parásitos procíclicos, sanguíneos y en tripanosomas Di o Akinetoplastidos. El esquema tomado de [131] muestra en el panel (a) la ATP sintasa compuesta por dos subunidades, Fo (en el espacio intermembrana) trasloca protones, mientras que la subunidad F1, puede sintetizar o hidrolizar ATP. El transportador ATP-ADP (AAC) media el intercambio de ATP y ADP a través de la membrana. En las formas procíclicas, el potencial de membrana es generado a través de la cadena de transporte de electrones, y es utilizado para generar ATP (a). En el panel (b) se muestra lo que sucede en las formas sanguíneas (BSF). En este caso la enzima trabaja en reversa y utiliza la hidrólisis de ATP para bombear protones a través de la membrana y mantener el potencial de membrana (Δψ). Por último, en el panel (c), se muestra que sucede en los parásitos sin kinteoplasto (akinetoplástidos) o con kinetoplasto parcial (diskinetoplástidos): la porción Fo de la ATP sintasa mitocondrial está ausente, pero la porción F1 (codificada en el núcleo) cumple la función de hidrolizar ATP para mantener el Δψ.

Los pasos propuestos por Lai y colaboradores [130] para la adaptación a la pérdida del ciclo completo (es decir la permanencia únicamente en el hospedero mamífero) son los siguientes:

1. La presencia de mutaciones en genes mitocondriales esenciales o la deleción de algún gen esencial, aunque la subunidad A6 de la ATP sintasa (Fo) debe permanecer funcional. Estos parásitos pierden la capacidad de vivir en el

insecto vector y deben ser trasmitidos solamente entre hospederos mamíferos, perdiendo la capacidad de completar el ciclo.

2. La pérdida de clases de la diversidad de ARNg y los minicírculos que los portan (excepto para aquellos que se requieran para sintetizar la subunidad Fo). Según los autores este paso puede llevar cientos de años.

3. Mutaciones compensatorias en la ATP sintasa nuclear, que eliminan la necesidad de la síntesis de la enzima mitocondrial.

4. El ADNk podría desaparecer completamente.

Lun y colaboradores, han planteado dos pasos para dar lugar a la adaptación de *T. brucei evansi* y *T. brucei equiperdum* a la pérdida del ciclo completo [132]. Como se observa en la Figura 35, los pasos propuestos son:

1. La pérdida de heterogeneidad de minicírculos en las formas sanguíneas de T. b. brucei.

2. Una consecuencia posible es la eventual pérdida completa de la capacidad de diferenciarse en el insecto, y que su ciclo se limite a la transmisión mecánica en el hospedero mamífero, lo cual a lo largo de las generaciones llevaría a la pérdida total del maxicírculo.

Estos pasos son consistentes con la observación de pérdida de kinetoplasto en cepas de *T. b. brucei* con muchos pasajes en animales de laboratorio.

Cómo se esquematiza en la Figura 35, en el insecto la presión selectiva favorece la recombinación entre minicírculos y la diversidad de ARNs guía para permitir la sobrevida de los parásitos en esta etapa. En cambio, los parásitos que persisten en la forma sanguínea, sin esta presión selectiva, comienzan a perder su genoma mitocondrial y eventualmente puede suceder una pérdida completa de maxicírculos y minicírculos. Los autores de este trabajo, plantean que las mutaciones en la ATP sintasa nuclear podrían ocurrir antes de la pérdida de los ARNs guía que se necesitan únicamente en la fase procíclica. En el mismo sentido Jensen y colaboradores [131] plantean que las mutaciones compensatorias en la subunidad γ de la ATP sintasa nuclear ocurren tempranamente durante la progresión de la homogeneización de la población de los minicírculos. Únicamente en una cepa de *T. b. equiperdum* (V01395), las mutaciones pudieron haber ocurrido luego de la pérdida de homogeneidad de los minicírculos, manteniéndose la edición del gen A6-ATP sintasa mitocondrial hasta ese momento.

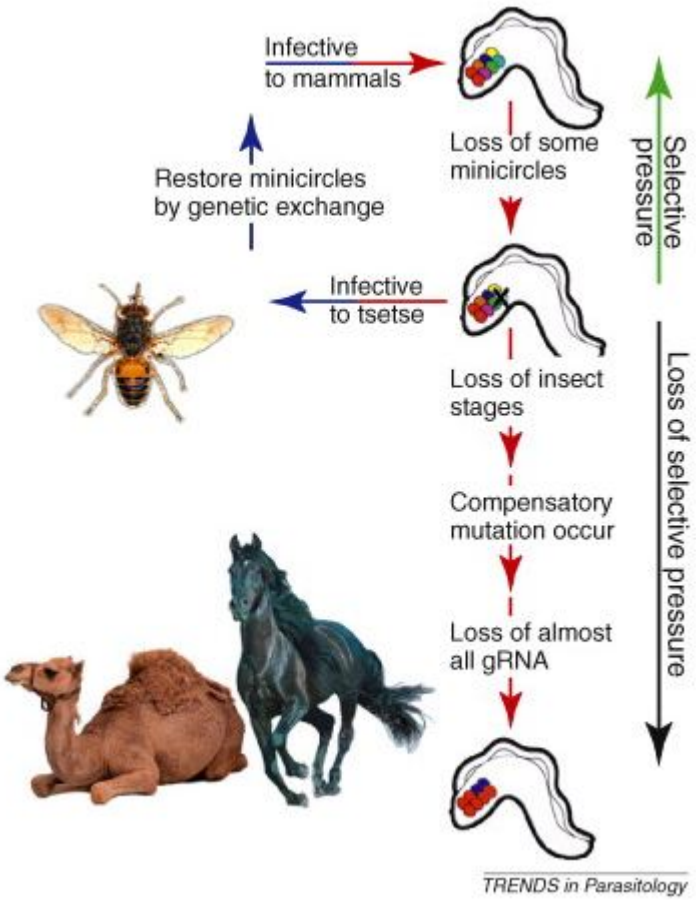


Figura 35. Pasos evolutivos propuestos para *T. b. evansi* y *T. b. equiperdum* como subespecies de *T. brucei* (tomado de [132]). Cuando se pierde el ciclo completo (sin pasar por el insecto vector), se favorece la reducción de la diversidad de ARNg (*T.b. equiperdum*) y eventualmente se pierde complemente el ADN mitocondrial (*T. b. evansi*), esta pérdida sucede luego de la aparición de mutaciones compensatorias en la ATP sintasa nuclear

En este trabajo se analizaron las mutaciones en la ATPasa nuclear de *T. vivax*. Los resultados muestran que al menos en las posiciones reportadas previamente, las tres cepas analizadas en este trabajo no presentan cambios (Figura 36). Asimismo, encontramos 2 cambios en las cepas americanas respecto a la cepa africana en las posiciones 60 y 61 (T60A y T61(AG)), aunque estas mutaciones se localizan en regiones de poca conservación, lo que sugiere que no son funcionalmente relevantes.

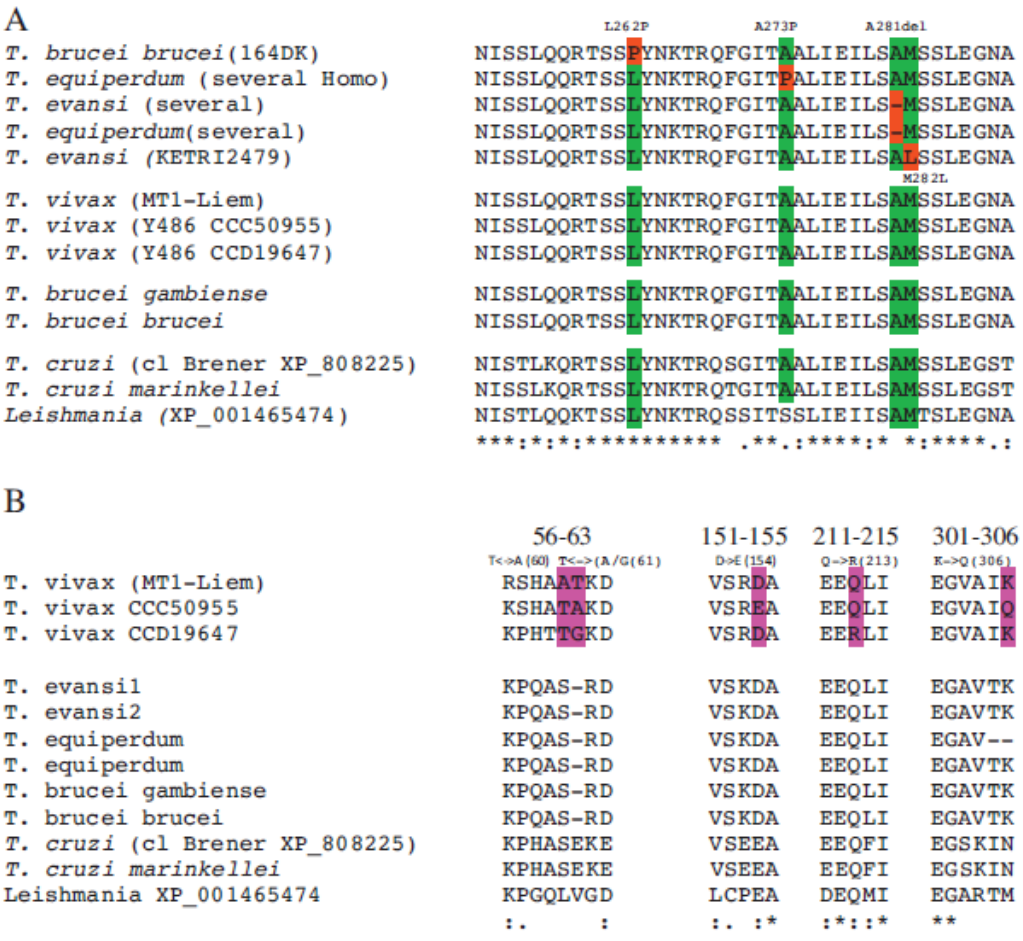


Figura 36. **Alineamiento de regiones relevantes de la ATP sintasa nuclear**. El alineamiento incluye especies representativas de tripansomátidos, las cepas americanas y africanas de *T. vivax*, así como cepas de *T. evansi* y *T. equiperdum* que contienen cambios de aminoácidos capaces de compensar la pérdida del genoma mitocondrial. **A.** Región entre los aminoácidos 250 y 289, que contienen cambios de aminoácidos asociados a nuevas funcionalidades de la ATP sintasa nuclear (indicados en rojo). En verde se indica la secuencia *wild type* para esa posición. **B.** Alineamientos en las regiones donde se observa variabilidad en *T. vivax* (indicado en violeta).

En suma, en los parásitos americanos se observa el proceso de degradación del genoma mitocondrial, y este proceso muy probablemente comenzó desde la introducción de *T. vivax* en América, ya que la cepa africana analizada Y486, pariente muy cercano del ancestro de las cepas introducidas en América, es capaz de editar todos los genes mitocondriales. El proceso observado, sería similar al postulado para las subespecies de *T. brucei*, *T.b equiperdum* y *T. b evansi* (como mencionamos más arriba); sin embargo en estas especies la pérdida de la ATP sintasa mitocondrial es compensada por mutaciones en la ATP sintasa nuclear que le permiten adquirir la función requerida para el mantenimiento del potencial de membrana mitocondrial. Este no es el caso de las cepas de *T. vivax* americanas analizadas en este trabajo.

Cabe destacar, que si bien en los estudios mencionados más arriba se postula la adquisición de mutaciones en la enzima nuclear como un paso intermedio posible, hasta el momento no se ha observado este estado en la naturaleza (o en el laboratorio). Lun y colaboradores, por ejemplo, plantean que las mutaciones en la enzima nuclear son pasos previos a la degradación del maxicírculo [132].

Por lo tanto cabe plantearse dos alternativas para explicar las observaciones presentadas en esta tesis. Una posibilidad es que, de acuerdo a lo planteado por Jensen y colaboradores [131], estemos observando una etapa en la transición hacia la pérdida completa del genoma mitocondrial, proceso que requiere de mutaciones compensatorias en la ATP sintasa nuclear que aún no han sucedido en el genoma de las cepas americanas de *T. vivax*. Aunque esta hipótesis ha sido formulada para explicar lo observado en *T. equiperdum* y *T. evansi*, nunca ha sido observado en la naturaleza. Por otra parte Lun y colaboradores [132], plantean que las mutaciones en la ATP sintasa nuclear son una precondición para la eliminación del kinetoplasto, y en este sentido estaríamos observando un camino evolutivo diferentes en las cepas americanas de *T. vivax* que han comenzado a perder su genoma mitocondrial pero que no presentan mutaciones en la ATP sintasa nuclear que le permitan vivir sin la función mitocondrial.

Kinetoplast adaptations in American strains from *Trypanosoma vivax*.

Greif G, Rodriguez M, Reyna-Bello A, Robello C, Alvarez-Valin F.
Mutation Research. 773 (2015);69-82. doi: 10.1016/j.mrfmmm.2015.01.008

Mutation Research 773 (2015) 69–82

Contents lists available at ScienceDirect

Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis

journal homepage: www.elsevier.com/locate/molmut
Community address: www.elsevier.com/locate/mutres

Kinetoplast adaptations in American strains from *Trypanosoma vivax*

Gonzalo Greif[a,1], Matías Rodriguez[b,1], Armando Reyna-Bello[c,d], Carlos Robello[a,e], Fernando Alvarez-Valin[b,*]

[a] *Unidad de Biología Molecular, Institut Pasteur de Montevideo, Uruguay*
[b] *Sección Biomatemática, Facultad de Ciencias, Universidad de la Republica, Uruguay*
[c] *Departamento de Ciencias de la Vida, Carrera en Ingeniería en Biotecnología, Universidad de las Fuerzas Armadas, Ecuador*
[d] *Centro de Estudios Biomédicos y Veterinarios, Universidad Nacional Experimental Simón Rodríguez-IDECYT, Caracas, Venezuela*
[e] *Departamento de Bioquímica, Facultad de Medicina, Universidad de la República Uruguay*

ARTICLE INFO

Article history:
Received 1 October 2014
Received in revised form 6 January 2015
Accepted 17 January 2015
Available online 25 January 2015

Keywords:
Fast evolution
Mechanical transmission
Genome degradation
Editing

ABSTRACT

The mitochondrion role changes during the digenetic life cycle of African trypanosomes. Owing to the low abundance of glucose in the insect vector (tsetse flies) the parasites are dependent upon a fully functional mitochondrion, capable of performing oxidative phosphorylation. Nevertheless, inside the mammalian host (bloodstream forms), which is rich in nutrients, parasite proliferation relies on glycolysis, and the mitochondrion is partially redundant. In this work we perform a comparative study of the mitochondrial genome (kinetoplast) in different strains of *Trypanosoma vivax*. The comparison was conducted between a West African strain that goes through a complete life cycle and two American strains that are mechanically transmitted (by different vectors) and remain as bloodstream forms only. It was found that while the African strain has a complete and apparently fully functional kinetoplast, the American *T. vivax* strains have undergone a drastic process of mitochondrial genome degradation, in spite of the recent introduction of these parasites in America. Many of their genes exhibit different types of mutations that are disruptive of function such as major deletions, frameshift causing indels and missense mutations. Moreover, all but three genes (A6-ATPase, RPS12 and MURF2) are not edited in the American strains, whereas editing takes place normally in all (editable) genes from the African strain. Two of these genes, A6-ATPase and RPS12, are known to play an essential function during bloodstream stage. Analysis of the minicircle population shows that its diversity has been greatly reduced, remaining mostly those minicircles that carry guide RNAs necessary for the editing of A6-ATPase and RPS12. The fact that these two genes remain functioning normally, as opposed to that reported in *Trypanosoma brucei*-like trypanosomes that restrict their life cycle to the bloodstream forms, along with other differences, is indicative that the American *T. vivax* strains are following a novel evolutionary pathway.

1. Introduction

Kinetoplastid protozoans owe their name to the peculiar mitochondrial DNA (kinetoplast or kDNA) that they contain. This consists of a very large network of interlocked circular DNA molecules. There are two types of such molecules: maxicircles and minicircles. The former, whose size varies among species between 20 and 35 kb, encompasses many of the typical mitochondrial protein coding genes (principally sub-units from respiratory chain complexes), as well as 2 ribosomal RNAs. They exhibit very limited (if any) intra-individual sequence variability. Minicircles instead are vastly more variable, both in size and sequence. In the case of trypanosomatids (the most important family of parasitic kinetoplastids), minicircle length varies from as short as 400 bp (*Trypanosoma vivax*) to about 10 kb (*Trypanosoma avium*). A normal set of minicircles is composed by 30,000–50,000 copies belonging to between 80 and 200 different classes [1]. The role of minicircles is related to the editing of maxicircle transcripts. This is a process in which these transcripts become translatable only after uridine insertion/deletion. Although editing is not exclusive of kinetoplastids, the extent it has attained in this group is unparalleled. In fact in some genes (termed cryptogenes) the degree of posttranscriptional modifications their transcripts experience is so extensive, that it is not even possible to recognize their corresponding genomic segments as coding some particular protein. In other genes editing is restricted to a particular gene segment (usually 5′) while in other cases it affects just a few nucleotides. Minicircles encode

* Corresponding author. Tel.: +598 25258618x7138; fax: +598 25258617.
E-mail address: falvarez@fcien.edu.uy (F. Alvarez-Valin).
[1] These authors contributed equally to this work.

http://dx.doi.org/10.1016/j.mrfmmm.2015.01.008
0027-5107/

short RNA molecules called guide RNAs (gRNA), necessary for directing the addition/deletion of uridines, and thus allow decoding the encrypted message of maxicircle transcripts. So far, the complete minicirculome has not been determined for any trypanosomatid species, but its complexity has been estimated to be between 80 (in *Leishmania*) and more than 200 (in *Trypanosoma brucei*) classes based on the total number of editing events (and hence gRNAs) required to produce all mature mRNAs [2]. A very recent characterization of gRNA populations in *T. brucei* shows that there are about 640 different classes that participate in the processing of all maxicircle mRNAs [3].

The mitochondrial role may change during the life cycle in some trypanosomatid groups. A clear example are the model species *T. brucei* and other African trypanosomes which have a digenetic life cycle that includes a bloodstream stage (BS) in a mammalian host, and a procyclic stage (PS) in the midgut of the insect vectors (tsetse flies). Owing to the low supply of energy sources in the insect vector's midgut [4,5], the parasite is dependent upon a fully functional mitochondrion, capable of performing oxidative phosphorylation. Nevertheless, inside the mammalian host, which is rich in nutrients, parasite proliferation relies on glycolysis, and most mitochondrial proteins are not essential. It should be noted, however, that even though oxidative phosphorylation is not carried out in BS forms, the mitochondrion is still necessary during this phase to accomplish other important functions.

Radical changes have taken place in the kDNA of *Trypanosoma equiperdum* and *Trypanosoma evansi*, two sub-species of *T. brucei*, adapted to tsetse-independent transmission. *T. equiperdum* is a sexually transmitted horse parasite, while *T. evansi* is mechanically transmitted (i.e. without completing the cycle) by other species of hematophagous flies (like tabanids and stomomix). These two trypanosomes have abandoned the procyclic stage associated with the tsetse vector, remaining permanently as BS forms. In all likelihood, the above mentioned changes in the kDNA are the cause of locking these parasites in the BS stage [6]. This is because these changes consist in the partial deletions (in *T. equiperdum*) or complete absence (in *T. evansi*) of maxicircles, as well as a severe (and sometimes complete) reduction of minicircle diversity. As a consequence these *T. brucei* relatives lack some, or all, of the mitochondrial genes encoding the respiratory proteins (oxidative phosphorylation) essential to proliferate in the tsetse fly [7]. The sequence of evolutionary events that leads from a "normal" *T. brucei* (i.e. capable of completing the whole cycle) to a BS restricted trypanosome still remains not fully clarified [6,8]. A particularly interesting facet is the fact that the absence of maxicircle encoded proteins is not necessarily circumvented by confining the parasite to the mammalian host, since some of these proteins are necessary in that stage too. This was first put in evidence by the observation that the editing machinery was also required in the BS forms in normal *T. brucei* strains [9]. From previous studies on petite mutants of yeast, it was immediately realized that the protein needed in BS trypanosomes was the A6 subunit of the ATP synthase (A6-ATPase), a subunit of the F0–F1 complex [10]. This is a proton pump that during the BS stage runs in reverse (hydrolyses ATP), enabling the parasite to generate electric potential. However, as shown by Lai et al. [7], none of the *T. equiperdum* or *T. evansi* strains analyzed up to date are able to produce mitochondrial A6-ATPase (because the gene is either absent or not edited), implying that in principle these parasites would not be able to survive, not only in the insect vector but also in the mammal host. Nevertheless, as it is evident these trypanosomes do reproduce in the mammalian host and also exhibit almost normal electric potential. The answer to this apparent contradiction was again provided by prior work in yeast, where it was observed that some mutations in the nuclearly encoded ATPase subunits are able to compensate the absence of A6-ATPase [11]. Lai et al. [7] searched for equivalents mutations in *T. evansi* and *T. equiperdum* and found that the γ-subunit of ATP synthase (encoded by the nuclear genome) bears mutations in evolutionary conserved amino acid positions (thus inferred to be functionally important). More recently Dean et al. [12] tested the functionality of these mutations and found that many of them are definitively capable to compensate the lack of A6-ATPase.

In this work we conduct a comparative study of the mitochondrial genome in *T. vivax*, a neglected African trypanosome. For this purpose we determine the complete genome sequence of their kDNA (maxicircle and minicircles) in three strains of this species, one originally from West Africa and two American strains. We also combine these data with transcriptomic information previously obtained by us [13], new RNAseq data produced for this work and data downloaded from public databases, to infer functional aspects of their mitochondria (in particular those related to editing).

There are a number of reasons that make the analysis of this species an interesting one. In the first place because it occupies a crucial phylogenetic position, since it is the earliest branching African trypanosome [14]. This phylogenetic location is of great relevance since it makes *T. vivax* a good model to address questions concerning the evolutionary genomics of African trypanosomes. In the second place, because *T. vivax* was able to leave Africa, and now affects regions, like America, devoid of tsetse flies. In these regions *T. vivax* is transmitted mechanically (like *T. evansi*) remaining permanently as BS forms [15,16]. It is of interest to investigate if in this species the process of adaptation to mechanical transmission parallels that already described for *T. evansi* or if it has followed a different evolutionary pathway. As it will be explained later, we point out that due to their evolutionary history and mode of transmission, the three strains used in this work are particularly suitable to tackle this topic.

2. Materials and methods

2.1. Parasites

2.1.1. Experimental infection and parasite purification

In this work we analyze three *T. vivax* strains, one originary from West Africa called Y486, which is the same used to determine the first genome in this species [17,18] and two strains from South American (MT1 and Liem176). The *T. vivax* samples were grown and purified as described before [13].

All animal work was conducted in accordance with relevant national and international guidelines. Mice were housed in the animal care facilities at Institut Pasteur of Montevideo (Uruguay). Animal housing conditions and protocols used in the present work were approved by the CEUA (Ethical Committee for Laboratory Animal Use) under the number 013-11 according to the Ethics Chart of animal experimentation which includes appropriate procedures to minimize pain and animal suffering. Infections in sheep were conducted under veterinary supervision with daily control of temperature and hematocrit which was never below 30%.

2.1.2. RNA and DNA purification and quality control

Total DNA was isolated from 10^9 parasites using QIAamp DNA Mini Kit (Qiagen, Germany) according to manufacturer's protocol. Obtained DNA was quantified in a Nanodrop (Thermo Scientific, USA) and its integrity was checked by agarose gel electrophoresis. Total RNA was purified from 10^9 parasites using Trizol (Sigma, USA) and Direct-Zol™ RNA MiniPrep (Zymoresearch, USA) according to the instructions of the kit. Obtained RNA was quantified in a Qubit

(Invitrogen, USA) and its integrity was checked in a Bioanalyzer (Agilent, USA).

2.1.3. Library construction and sequencing

Genomic DNA libraries were generated from 50 ng of total DNA using Nextera Kit (Illumina, USA) according to the manufacturer's instructions. The libraries were quality checked using Agilent High Sensitivity DNA Bioanalyzer Kit (Agilent, USA), and quantified using Qubit® dsDNA BR Assay Kit (Life Technologies, USA).

MT1 libraries sequencing was performed on an Illumina Genome Analyzer IIX platform and generated 26.494.848 paired-end reads (2 × 100 cycles). For the Liem176 strain, libraries were sequenced using a MiSeq (Illumina, USA), paired-end 2 × 150 cycles run, yielding 6.260.150 reads.

For RNAseq libraries (Y486), double-stranded cDNA was generated from 1 μg of total RNA using a SuperScript III Double-Stranded cDNA Synthesis Kit (Invitrogen). Libraries were made, indexed and normalized with NexteraXT kit (Illumina, USA), using manufacturer's protocol. Finally, libraries were sequenced on a MiSeq (Illumina, USA) paired-end 2 × 150 cycles run and 13.3 million reads were obtained.

2.1.4. PCR and Sanger sequencing confirmation

The primers and condition used in PCR and Sanger sequencing experiments for final maxicircle assembly and minicircle confirmation are summarized in supplementary file 1.

2.1.5. Data handling and analysis

Illumina reads from all datasets were trimmed from adapters and other contaminants using Scythe (v0.981) (http://github.com/vsbuffalo/scythe). Genomic reads from MT1 strain were quality filtered by trimming bases with a Phred score lower than 20 using Sickle (v1.2) (https://github.com/najoshi/sickle) and keeping only those reads whose length (after trimming) was at least 65 bp. After quality filtering and trimming, 24.422.908 usable reads were left for this strain. In the case of Liem176 strain reads were also quality filtered and trimmed with a Phred score threshold of 20, yet using a minimum length of 75 yielding 5.088.843 usable reads. The quality control in both datasets was done using FastQC (v0.11.1). De novo assembly and scaffolding in MT1 was conducted using ABySS (v1.3.5) [19] with k-mer options ranging from 40 to 64, retaining the assembly with a k-mer size of 50 which produced the longest contigs and best N50 value. For Liem176 the assembly was performed using SPAdes (v2.5.0) [20] with k-mer ranging from 45 to 85 with an increase step of 10. The assembly of the Y486 maxicircle genome was performed using raw Sanger reads downloaded from public databases. After mapping these reads against the MT1 maxicircle genome, 2.816 maxicircular reads were identified which were assembled using Mira (v3.9.17) [21] with the options – job = genome, de novo, accurate. This assembly produced a single contig of 18.730 bp comprising the whole coding region and part of the species specific (repetitive) regions from both ends. This contig was extended up to 20.400 bp by iteratively overlapping an extending the contig edges with more Sanger reads. Additional details of the sequencing and assembly are explained in supplementary file 1.

Mapping of reads (DNAseq and RNAseq) against the assemblies was done using Bowtie2 [22] alignment software with end-to-end default options. Rpkm values were calculated according to Mortazavi et al. [23] and Garber et al. [24]. All mappings were obtained in SAM format, and then converted to binary files, sorted and indexed using SAMtools. Tablet (1.13.07.31) [25] was used for visualization and data-navigation purposes.

3. Results

3.1. Maxicircle genome: gene degradation in American strains

The maxicircle genomes were determined for the three *T. vivax* strains analyzed here. In the case of MT1 it was done combining Illumina paired-end reads and the finishing was carried out amplifying and sequencing (Sanger) specific segments that were not resolved in the initial assembly. For this strain sequencing was not limited to the 15 kb well conserved genome segment that contains genes but also the region located upstream to the 12S ribosomal RNA gene and downstream to ND5 gene (the last protein coding gene). This latter region is poorly conserved among trypanosomes and basically contains repetitive sequences. The initial MT1 assembly contained only one part of this region, including a single copy region of approximately 1 kb in length (represented in green in Fig. 1) and two clusters containing different types of tandem repeats. One of them located near to the 12S rRNA gene, which appeared as a region with a notorious increase of sequencing depth (hence indicating its repetitive nature), is composed by a 105 bp repeating unit whereas the second cluster (orange box in Fig. 1) contains a 24 bp repeat. In MT1 strain an additional region of 5.2 kb in length was obtained using specific primers. This region is composed by another type of repetitive sequence, which is 170 bp in length and also has a tandem array disposition (see Fig. 1). Due to the length and repetitive nature of this segment it was only possible to solve it partially by Sanger sequencing. Nevertheless we could confirm that the remainder part of this maxicircle segment is composed by copies (complete and partial) of the same repeat (further details in supplementary file 1 and figure FS1A and FS1B).

For the other two strains the strategies to determine the maxicircle sequence were different. In the case of Liem176 we used an Illumina paired-end (2 × 150) library plus long RNAseq derived contigs. In Y486 in turn, raw Sanger reads were downloaded from public databases and mitochondrial reads were identified by mapping against the previously assembled mitochondrial genomes. The reads thus identified were assembled using Mira assembler producing a full genome sequence contained in a single contig. Additional details of the sequencing and assembling methodology are available in accompanying supplementary material (supplementary file 1 and supplementary figure FS1A).

In summary, the 15 kb maxicircle segment that contains genes was completely determined for the three strains, while the repetitive (species specific) region was not resolved in Liem176 and Y486 with the same level of detail as in MT1. Nevertheless, it was possible to determine that in these strains the species specific segment is composed by the same type of repeats and present in similar amounts (as estimated by sequencing depth and the size of amplicons).

Next we concentrated in the comparison among the three *T. vivax* strains, of the 15 kb maxicircle region that contains genes. As it can be observed in Fig. 1 and Table 1, this comparison reveals some interesting aspects. As expected, the two American strains are more similar to one another than to the African *T. vivax* (Y486). In effect, the two American strains do exhibit many nucleotide differences and several deletions relative to the African strain. The majority, but not all, of these changes are shared between the American strains indicating that they occurred in America before the separation of the two strain analyzed here. There are only 6 changes between MT1 and Liem176, and most of these changes do not affect any functional significant position (many are synonymous, i.e. they do not change the encoded amino acid). The most noticeable change between American strains and strain Y486 is a 752 bp deletion that affects two protein coding genes, ND7 (which results in a deletion of 427 bp in the 3′ end) and COIII that has a 248 bp deletion in the 5′ part of the gene. This relatively big deletion was confirmed by

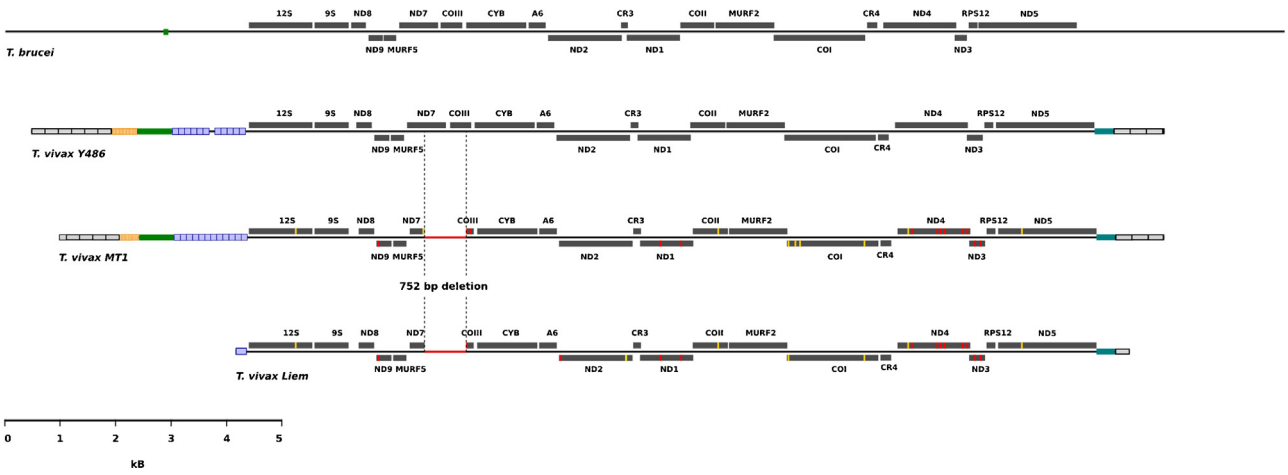


Fig. 1. Alignment of complete maxicircle sequences in the three *T. vivax* strains analyzed in this work. *T. brucei* maxicircle was included as a reference. Genes are indicated by gray boxes and the respective names indicated. Those genes placed above the line are located in the forward strand whereas those below in the complementary strand. For MT1 and Liem176 mutation are indicated by vertical lines: yellow lines represent transitions and transversions, red lines indels. Components of the species specific region of the genome are also schematized. Light blue boxes represent the cluster of 105 bp repeats, orange boxes, 24 bp repeats and gray boxes 175 bp repeats. The green box in turn, represents a species specific non-repetitive genomic sequence, a segment of which is faintly conserved in *T. brucei*.

PCR (supplementary figure FS1A, box a) since two sets of primers amplify in this region a much smaller than expected amplicon (252 nt, whereas a fragment of approximately 1000 nt is expected). In fact, the same primers amplify both in *T. vivax* Y486 and *T. brucei*, a segment of the expected size (1042 and 915 bp respectively) which contains the portions of ND7 and COIII that are missing in the American strains (supplementary figure FS1A, box a). It is important to stress that in the two American strains, only one amplification species was obtained meaning that all maxicircle copies lack this 752 bp fragment (i.e. they are "homogenous" for this loss). This

Table 1
List of maxicircle genes, editing status, mutations observed and their effect.

Gene	Editing[a]	Position: base change[b]	Mutation effect (aa)[c]	Strain[d]
1. ND8	Pan-edited	–	–	–
2. ND9	Pan-edited	21_22: ins T	Possible frameshift	MT1, Liem
3 MURF5	Not edited	–	–	–
4. ND7	Pan-edited	248: G>T	ND	MT1
		275_699: del 424 bp	Major deletion	MT1, Liem
5. COIII	Pan-edited	1_248: del 248 bp	Major deletion	MT1, Liem
		253: G>C	ND	MT1, Liem
		261_262: del C	Possible frameshift	MT1
		303_304: ins AG	Possible frameshift	MT1, Liem
6. Cyb	5′ edited	–	–	–
7. A6-ATPase	Pan-edited	–	–	–
8. ND2 (MURF1)	Not edited	16_19: del ATAC	Frameshift	Liem
		1212: A>G	Point mutation	Liem
9. CR3	Pan-edited	–	–	–
10. ND1	Not edited	491_492: ins T	Frameshift	MT1, Liem
		863: del A	Restores frame	MT1, Liem
11. COII	Partial editing	456_457: ins TGC	Insertion of 1 aa (ins C)	MT1, Liem
12. MURF2	5′ edited	–	–	–
13. COI	Not edited	24: T>A	Missense (C>W)	MT1, Liem
		154: G>T	Missense (G>C)	MT1
		237: G>A	Missense (G>S)	MT1
		1399: G>A	Missense (V>I)	MT1, Liem
14. CR4	Pan-edited	–	–	–
15. ND4	Not edited	194: C>T	Missense (A>V)	MT1, Liem
		254_259: del ATATAC	Deletion of 2 aa (del	MT1, Liem
		712_713: ins TT	MY)	MT1, Liem
		778: del T	Frameshift	MT1, Liem
		846: G>A	Frameshift	MT1, Liem
		1175_1176: ins AT	Synonymous	MT1, Liem
		1257_1258: del AT	Restores frame	MT1, Liem
			frameshift	
16. ND3	Pan-edited	105_106: del TT	Possible frameshift	MT1, Liem
		205_206: del TT	Possible frameshift	MT1, Liem
17. RPS12	Pan-edited	–	–	–
18. ND5	Not edited	430: G>T	Missense (G>C)	MT1, Liem

[a] Extent of editing during RNA maturation.
[b] Nucleotide position and type of change. For insertions, numbers correspond to positions that surround them, for insertions numbers correspond to the nucleotides deleted. Nucleotide changes are indicated.
[c] Expected modification at amino acid level. Amino acid change is provided inside the brackets. ND: effect not determined.
[d] Strains where the mutation is observed.

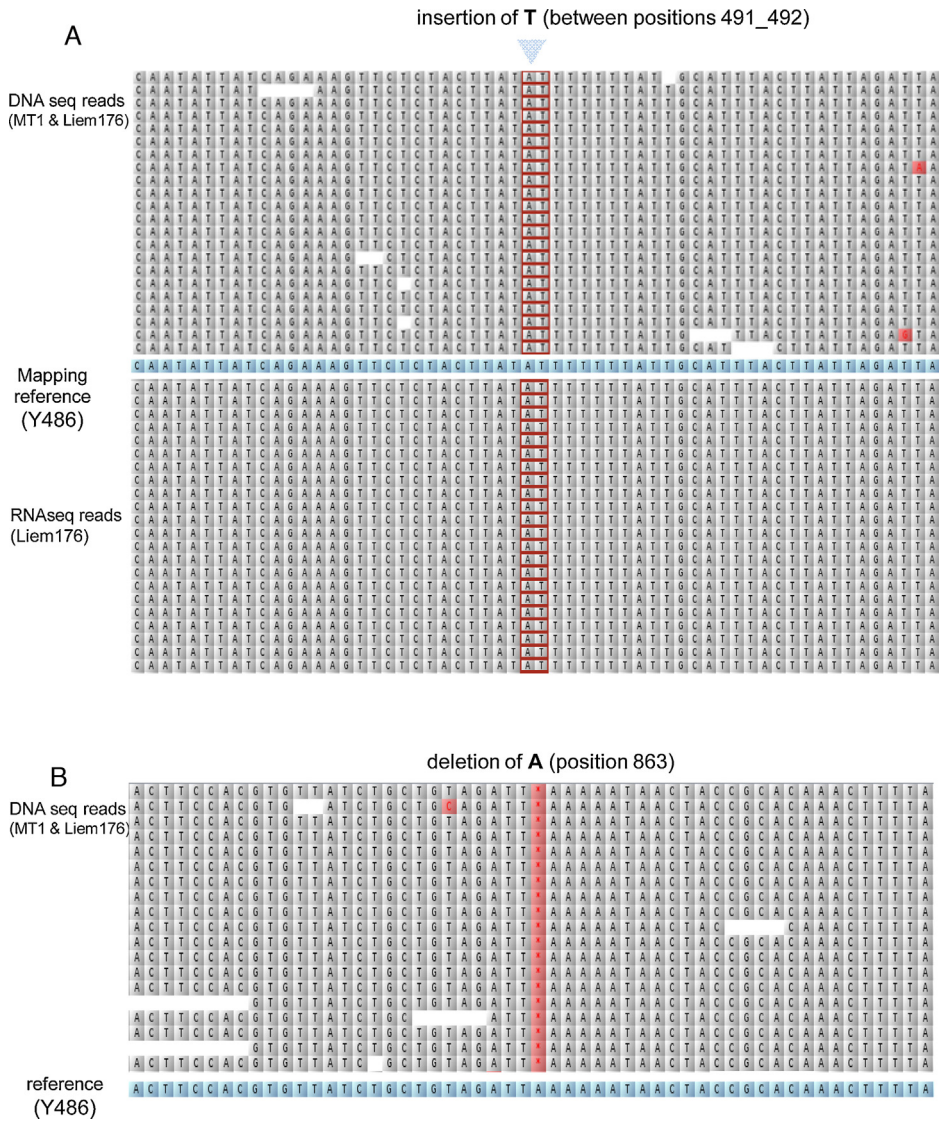


Fig. 2. Testing heteroplasmsy and novel editing sites in the ND1 gene from American strains. (A) Insertion of a **T** between positions 491 and 492. In the upper half of figure (above the mapping reference) DNA sequencing derived reads (from MT1 and Liem176) were mapped against the ND1 coding sequence. The gene from Y486 strain was used as mapping reference to evidence the difference. In the lower part of the image RNAseq reads (from Liem176) were mapped against the same reference to test whether this insertion is rectified post-transcriptionally by editing (i.e. whether it may represent a novel editing event). Note that the inserted **T** is not actually shown in the reads but represented by a red box outlining the two bases that surround the insertion (AT, 491:492). (B) Deletion of **A** at position 863. For this position only heteroplasmy was tested since this deletion cannot be corrected post-transcriptionally by editing. The red asterisks represent gaps (i.e. the base deleted in the American strains).

implies that these two genes (ND7 and COIII) are completely non-functional neither in MT1 nor in Liem176. Other small deletions and insertions are observed in 6 additional genes (listed in Table 1). Although these indels are small, they produce frameshifts that if not corrected (post-transcriptionally) the respective genes cannot be translated into a functional protein. In consequence we explored this aspect in more detail to confirm the possible functional effect of these mutations. In the first place we tested heteroplasmy for these mutations, namely if the population of maxicircles contains, apart from the mutated version of genes presented in Table 1, normal copies not affected by these indels that might eventually complement the loss of function. Back mapping of the maxicircle reads shows that this is not the case. Note that if there was heteroplasmy one would expect some reads bearing the indel, while others would be not mutated. This analysis shows that the maxicircle population is completely homogenous for these mutations (see Figs. 2 and FS2).

Another possibility tested was if these small indels might represent novel editing sites that in the genome, or in pre-edited mRNA, may appear as frameshifts, but in fact could be corrected during the editing process. In four of these genes it is difficult to discern this question because the genes are pan-edited. However, in three genes (MURF1, ND1and ND4), which do not require editing since in *T. brucei* and *T. vivax* Y486 their genomic sequences are the same as the mature mRNA, the observed indels would eventually render nonfunctional proteins if they did not undergo these "novel" corrective editing events and were translated exactly as they are transcribed. To distinguish between the two alternatives, Illumina RNAseq reads were mapped onto these genes. Attention was paid to those mutations that consist in deletions or insertions of thymidine, namely those ones that eventually could be corrected post-transcriptionally by editing. In the three genes, all RNA derived reads match perfectly with the genomic template, indicating that the indels are not rectified post-transcriptionally (results for ND1 are presented in Fig. 2, whereas MURF1 and ND4 are reported in supplementary figure FS2).

Two other genes, COI and ND5, have nucleotide changes that imply amino acid substitutions. By comparing these genes with the

corresponding homologs from *T. brucei* and more distantly related trypanosomatids like *Trypanosoma cruzi* and *Leishmania donovani*, it is possible to infer that these changes took place in the lineage leading to American strains (after their separation with Y486), and affect evolutionary conserved position, something suggestive of deleterious effect (supplementary figure FS3A).

In summary, the results presented in this section evidence that the American strains from *T. vivax* have accumulated numerous mutations in ten mitochondrial genes. In all likelihood these mutations are disruptive of function since they imply big deletions, frameshifts or point changes at amino acid positions that are probably functionally relevant.

3.2. Editing of maxicircle genes

Another relevant aspect to investigate is whether the mechanism of editing is working correctly and if all genes are productively edited in the American and African *T. viv*ax strains. For this purpose it is necessary first to identify mature (edited) mRNAs encoded by maxicircle genes. This is something relatively simple for those genes that have minimal editing such as COII, Cyb and MURF2, since the differences between the mRNA and genomic sequences are restricted to few positions. However for pan-edited genes this is far more complicated because the edited sequence is unknown, and cannot be inferred from the genomic sequences alone, given the great number of editing events that their (pre-edited) RNAs suffer to become mature mRNAs.

To identify the mature (edited) mRNAs, the assembled transcriptome (obtained as explained in detail in supplementary file 2) from Y486 strain was virtually translated and the amino acid sequences thus obtained were compared with those encoded by mitochondrial genes of *T. brucei* and other trypanosomatids. This allowed us to identify the mature mRNA from all maxicircle coding genes on the basis of amino acid conservation. Interestingly the 18 maxicircle protein coding genes are transcribed in three life cycle stages from Y486 strain. Twelve of these genes require editing to become translatable, something that appears to occur normally in all life cycle stages from the African strain (in epimastimotes transcription and editing were checked using RNAseq data from NCBI SRA repository). Supplementary file 2 presents the alignments between the genomic gene sequences and mature mRNA; the individual editing events are depicted for all genes. It is worth mentioning that the level of abundance of mature and pre-edited RNAs (as estimated by read mapping depth) varies enormously from one gene to another as well as among the three life cycle stages analyzed here for the African Y486 strain (Fig. 3).

In American strains transcription and editing activity of maxicircle genes was assessed by mapping RNAseq reads from Liem176 against the pre-edited and edited (mature) RNAs. The results of this analysis, presented in Table 2, show that in the America strain all genes appear to be successfully transcribed. However among the 12 genes that require editing for their correct translation only three undergo editing normally: A6 subunit of ATP synthase, ribosomal proteins 12 (RPS12) and MURF2. For the remaining 9 genes (ND8, ND9, ND7, COIII, Cyb, COII, CR4, CR3 and ND3), no reads indicative of editing activity could be detected in Liem176 (read counts were zero when the mature mRNA was used as mapping reference), with the only exception of CR3 (a pan-edited gene) in which only traces of editing were found toward the 3′ end of the gene (Table 2). In other words these 9 genes are transcribed but their RNAs remain immature (pre-edited), which implies that they cannot be translated into functional proteins. It is important to stress that the failure to detect editing cannot be attributed to insufficient sequencing depth, namely that the number of reads was not large enough to detect some low abundance RNA species, since the set of Illumina reads from Liem176 has a sequencing depth adequate to detect even minor RNA species [13]. In fact, the sequencing depth in Y486 was approximately one half of that of Liem176, and all mature maxicircle mRNAs were found.

3.3. Characterizing the population of minicircles: sequencing, assembly and identification

We decided to identify the population of guide RNAs encoded in the genome of American strains to investigate if this can explain the loss of editing in the 9 maxicircle genes mentioned before. To this end the minicirculome (i.e. whole population of minicircles) was determined, something that has an intrinsic interest because the information obtained can also be used to analyze other relevant aspects concerning their organization, divergence dynamics among strains and also (when combined with RNAseq data) to analyze their expression activity. Genome wide studies and comparisons of minicircle populations are almost inexistent being restricted to only one example in *T. cruzi* strains [26].

As in the case of maxicircle, sequencing and assembling of minicircles was carried out together with the whole nuclear genome. However, the identification of contigs corresponding to minicircles is not as straightforward as with maxicircles due to the lack of sequence conservation among trypanosomatids. Specifically there is only one segment of approximately 120 bp in length that contains three blocks with different levels of sequence conservation, called CSB-1 to -3 [27]. CSB-3 (or Universal Minicircle Sequence, UMS) has a length of only 12 nt and is completely conserved in all trypanosomatids. Such conservation has been associated with its function because it works as a replication origin [27]. CSB-1 is less conserved and even shorter (10 nt), while the conservation of CSB-2 is almost marginal (supplementary figure FS4A). It is thus obvious that even if these conserved blocks may help identifying minicircles, their short length and variable conservation might render minicircle identification based on their sole presence not fully reliable. Therefore we adopted a two steps strategy that combines two sources of information: conservation information and statistical signatures. A first group of 18 putative minicircles sequences was identified using Blast against *T. brucei* CSBs. This first group contains minicircles with highly significant Blast HSPs (*E*-value < 1e−15). The 120 nt segment containing the conserved blocks from these 18 putative minicircles sequences were used as new queries to search against the whole population of contigs. A complementary source of information was also used. Since previous studies on minicircles from other species of trypanosomes indicated that they are peculiar in their base composition [26] we conducted a principal component analysis (using dinucleotides as variables) to check if this feature also holds for *T. vivax*. Fig. 4 shows that this is case since minicircles cluster close to each other and are clearly separated from contigs corresponding to other genomic compartments (nuclear and maxicircles). This indicates that dinucleotide composition can be a very useful predictor, and hence appropriate to complement initial assignments done on the basis of blastn results.

By combining these two data sources it was possible to identify 54 minicircles classes in MT1 and 46 in Liem176 (Table 3 and supplementary figure FS4A). All the 46 Liem176 minicircles have their homologs in the MT1 set, with DNA identity values ranging from 95% to 100%. However some MT1 minicircles are exclusive from this strain. Table 3 presents information on the minicircles sequences from both strains. This result shows that the two American *T. vivax* strains reported here have a reduced set of minicircles, considering that in *T. brucei* the total number of different minicircles classes has been estimated to be between 200 and 300 [1,2]. This could be due to a real absence of certain minircircles classes in the American strains or to the fact that the samples are not fully representative. In our opinion the latter alternative is not very likely given that

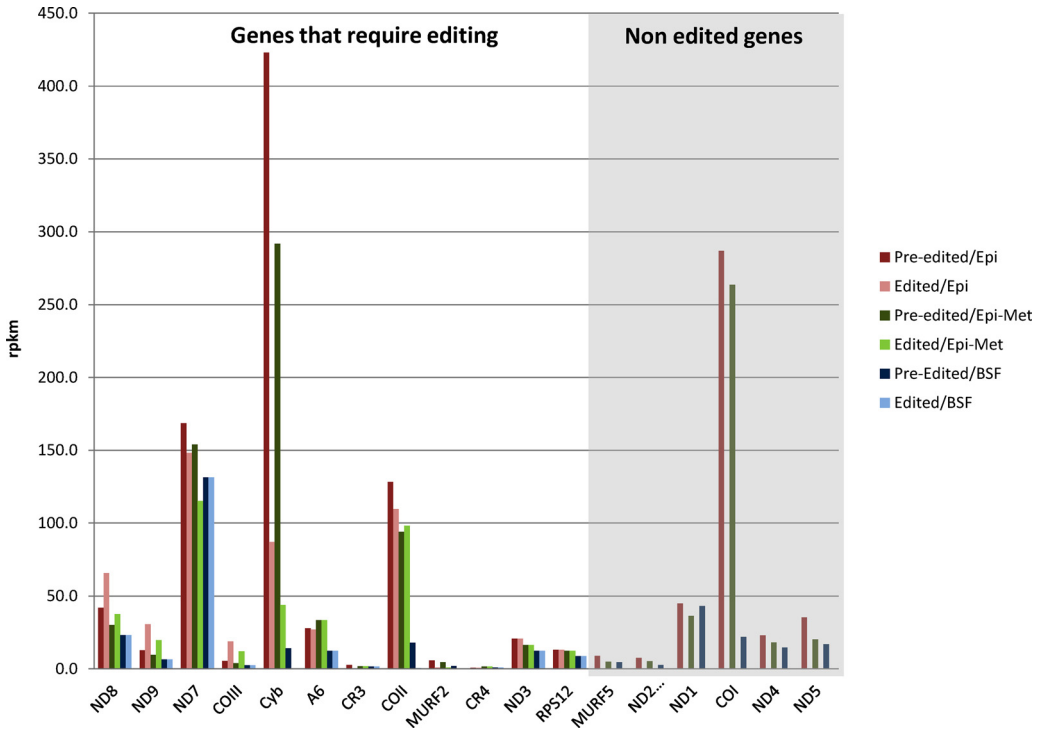


Fig. 3. Transcript levels of edited (mature mRNA) and pre-edited maxicircle genes in three different parasitic life stages (epimastigotes, bloodstream trypomastigotes and metacylic epimastigotes) in the strain Y486.

sequencing depth was large enough to guarantee that all (or most) minicircle classes were represented.

Another worth mentioning aspect is that there are two clearly defined size groups of minicircles: short ones (sizes ranging from 320 to about 600 nt), and a second group containing minicircles that have approximately the double length. Intra-sequence comparisons show that all but one of the long minicircles are multimers of a shorter repeating unit. Short contigs in turn, represent momomeric version of these units. Heterodimeric minicircles were not observed. There is only one case of a long minicircle that has no internal repetition (Tvminic53).

To investigate whether long (multimeric) minicircles are real molecules and not assembling artifacts, some of them were selected for further in silico and experimental analyses. The results, presented in supplementary figure FS5, suggest that these contigs are not assembling artifacts. It is interesting to note that our

Table 2
Transcript levels of maxicircle genes in the American strain Liem176, before and after editing. Both raw read count and rpkm values are shown. RNA lengths are also specified to illustrate size increase due to editing. In the case of genes with partial editing (Cyb, Murf2, COII), the read count on the right side of the table is restricted to the gene segment that undergoes editing because the gene regions which are not edited have non-zero count irrespective of editing taking place.

Gene	Pre-edited			Post-edited		
	Length (genomic sequences)	Reads mapped	rpkm	Length (mature mRNA)[e]	Reads mapping on edited segments	rpkm
ND8[a]	280	10,013	2059.7	440	0	0
ND9[a]	295	15,240	2975.5	585	0	0
Murf5[b]	240	37	8.9	240	Not edited	
ND7 (frag)[a]	275	12,456	2608.8	1164	0	0
COIII (frag)[a]	134	5873	2524.3	309	0	0
Cyb[c]	1081	1000	53.3	1113	0	0
A6[a]	320	29,340	5280.8	754	11,921	910.6
ND2[b]	1322	5782	251.9	1322	Not edited	
CR3[a]	119	1324	640.8	228	31[f]	7.8
ND1[b]	964	7602	454.2	964	Not edited	
COII[d]	632	250	22.8	636	0	0
Murf2[c]	1054	1188	64.9	1074	14	31.0
COI[b]	1650	1543	53.9	1650	Not edited	
CR4[a]	215	3118	835.3	495	0	0.0
ND4[b]	1309	4136	182.0	1309	Not edited	
ND3[a]	219	5043	1326.3	349	0	0
RSP12[a]	169	6175	2104.5	215	2068	554.0
ND5[b]	1773	4067	132.1	1773	Not edited	

[a] Pan-edited gene.
[b] Not edited gene.
[c] 5′ editing.
[d] Internal editing, frag: fragment, part of the gene is deleted in Americans strains.
[e] The sequences used as mapping reference for mature mRNA are those from Y486.
[f] Reads map on 3′ end only.

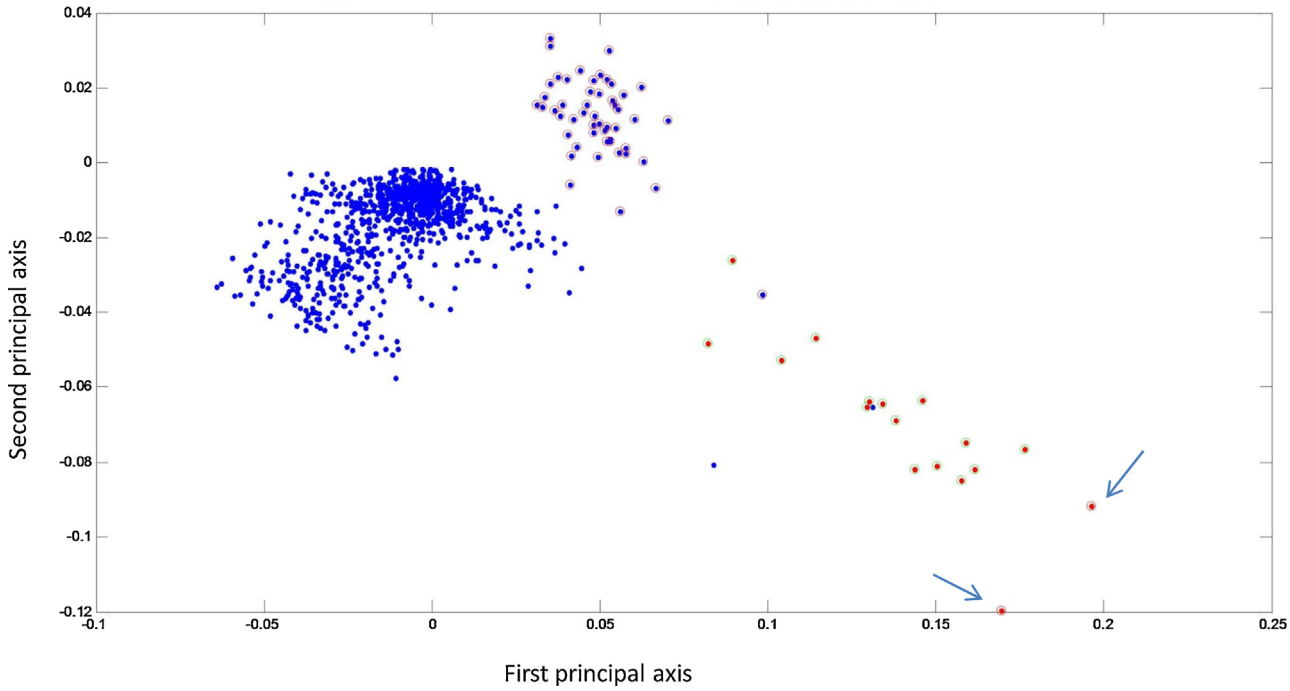


Fig. 4. Principal component analysis of dinucleotide frequencies. First and second axis plotted. Blue dots represent nuclear genome sequences (contigs), blue dots with red circles: putative minicircle sequences. Red dots inside green circles: maxicircle sequences. Maxicircle was splitted in segments of 1 kb each to allow intra genomic variability become apparent. Red dots inside red cicles: maxicircle repeated sequences (arrowed).

observation that *T. vivax* contains homodimeric and monomeric minicircles is in line with previous reports based on electron microcopy and restriction enzyme digestion [28]. It was found that in *T. vivax*, minicircles fall in two size categories: one category includes minicircles of approximately 460 bp and the second group of about 934 bp. The restriction pattern obtained by these authors was consistent with homodimeric minicircles. Finally we would like to make a cautionary note, because the approach used here to infer the sequences of minicircles can be somewhat unspecific for differentiating dimers and monomers when the two copies of the dimer are identical or almost identical, which may produce collapsed contigs during assembly.

3.4. Minicircle abundance and expression

Another aspect that was investigated is the relative abundance of each kind of minicircle. This was accessed my back-mapping the DNAseq reads onto the assembled contigs. The sequencing depth is a measure of both the success of the experiment and the relative abundance of a given sequence fragment. As it is evident from supplementary figure FS6A the relative abundances are not nearly homogenous among minicircles whereas the abundance of homologous minicircles is visibly similar between the two American strains.

Transcription activity of minicircles was analyzed using two sources of data. In the first place analysis based on Roche 454 derived RNA contigs shows that the initial transcripts are polycistronic molecules, in agreement with previous reports [29]. Interestingly, transcripts encompass the whole minicircle, spanning over the three CSB regions (supplementary figure FS4B shows 454 reads spanning the CSB blocks), in agreement with very early results by Thertulien et al. [30], but contrasting to what it had been suggested by other authors [29]. This observation was confirmed using RNAseq data from different and independent sources (Illumina reads from Liem176 and Y486, figure FS4C). Expression levels were also analyzed. Specifically they were inferred using customary RNAseq approaches, namely taking the number of Illumina reads mapping onto a given minicircle (normalized by length) as a measure of transcript abundance. Supplementary figure FS6B shows the scatterplots of RNAseq-RPKM vs DNAseq-RPKM. From what it can be observed in this figure, it is evident that the expression level of a given minicircle is highly correlated with its abundance. As a consequence, when RNAseq-RPKM figures are corrected by taking in consideration the abundance of the corresponding templates, it becomes apparent that the differences in transcript abundance are caused by the differential representation of each kind of template molecule rather than by differential transcription intensity. This implies that in minicircles, like in maxicircles and nuclear genes of trypanosomatids, regulation of transcription initiation plays a secondary role (or none at all) in determining transcript levels.

Finally the population of gRNAs was inferred from minicircle sequences using wu-blast with a modified scoring matrix as described in [31]. Table 4 presents the gRNA sequences identified grouped according to the gene where they exert their function. As expected, for the two genes that go through complete editing in the American strain (Liem176), namely A6-ATPase and RPS12, it was possible to identify almost all of the gRNAs necessary to guide their editing (details of gRNA sequences and their alignments to the mature mRNA are presented in Supplementary figure FS7). Two other genes undergo editing in the American strains, CR3 and MURF2. Regarding the former, although it is a pan-edited gene (in Y486), in the American strain Liem176 editing is largely incomplete, with few changes being added in a very restricted part of the gene (Table 2). MURF2 in turn, is 5′ edited (only the first 35 positions affected, see supplementary file 2), and just two gRNAs are required for its editing, one of which is encoded in the maxicircle [32]. Overall these results imply that only the genes that are productively edited have their corresponding gRNAs, while the vast majority of gRNAs responsible for the editing of the remaining genes is missing.

Table 3
Minicircle classes in the *T. vivax* strains MT1 and Liem176. Length and the % of identity between the different strains are indicated. It is also indicated which minicircle classes are observed as multimers.

Name	Length	As multimer (length)	Present in Liem	% Id MT1 consensus vs Liem consensus
TvMinic1	329	No	Yes	100% (142 bp del)
TvMinic2	360	Yes (1009)	Yes	100% (105 bp ins)
TvMinic3	396	No	No	–
TvMinic4	406	No	Yes	100% (63 bp del)
TvMinic5	422	Yes (1056)	Yes	100%
TvMinic6	427	No	Yes	100% (105 bp del)
TvMinic7	456	Yes (1262)	Yes	100% (105 bp ins)
TvMinic8	481	No	Yes	100% (31 bp ins)
TvMinic9	507	Yes (872)	Yes	100%
TvMinic10	521	Yes (1015)	Yes	100% (70 bp ins)
TvMinic11	538	Yes (975)	Yes	100%
TvMinic12	567	Yes (1211)	Yes	99.8%
TvMinic13	570	Yes (1102)	Yes	100%
TvMinic14	497	No	Yes	98.6% (26, 12 and 14 ins)
TvMinic15	535	No	Yes	100% (79 bp ins)
TvMinic16	482	No	No	–
TvMinic17	554	No	Yes	100%
TvMinic18	468	No	Yes	100%
TvMinic19	551	No	Yes	100%
TvMinic20	601	No	Yes	99.78% (143 bp ins)
TvMinic21	549	No	Yes	100% (82 bp ins)
TvMinic22	554	No	Yes	100%
TvMinic23	553	No	Yes	99.6% (85 bp ins/105 bp del)
TvMinic24	549	No	Yes	100% (85 bp ins)
TvMinic25	550	No	Yes	100% (109 and 79 bp ins)
TvMinic26	535	No	Yes	94.7% (85, 51, 22 and 3 bp ins)
TvMinic27	555	No	Yes	99.5% (59 bp ins)
TvMinic28	551	No	Yes	100% (85 bp ins, 66 bp del)
TvMinic29	554	No	Yes	97.6% (257 and 7 bp ins)
TvMinic30	533	No	Yes	100% (85 bp ins)
TvMinic31	553	No	Yes	100% (85 bp del)
TvMinic32	550	No	Yes	99.2% (164 bp ins)
TvMinic33	565	No	Yes	100% (85 bp ins)
TvMinic34	547	No	Yes	100% (75 bp ins)
TvMinic35	536	No	Yes	100% (56 bp del, 85 and 56 bp ins)
TvMinic36	566	No	No	–
TvMinic37	544	No	No	–
TvMinic38	595	No	Yes	95.4% (8, 20, 37, 12 and 12 ins, 177 del)
TvMinic39	509	No	No	–
TvMinic40	553	No	Yes	100% (85 bp ins)
TvMinic41	555	No	Yes	100% (85 bp ins)
TvMinic42	545	No	Yes	100% (85 bp ins)
TvMinic43	567	No	No	–
TvMinic44	549	No	No	–
TvMinic45	564	No	Yes	100% (85 bp ins)
TvMinic46	556	No	Yes	100% (40 bp ins)
TvMinic47	539	No	Yes	100% (80 bp ins)
TvMinic48	543	No	Yes	100% (85 bp ins)
TvMinic49	564	No	Yes	100% (86 bp ins, 115 bp del)
TvMinic50	576	No	Yes	100% (85 bp ins)
TvMinic51	542	No	Yes	100% (85 bp ins)
TvMinic52	489	No	No	–
TvMinic53	1168	No	Yes	–
TvMinic54	No	Yes (1306)	Yes	99.3%

In fact, few additional gRNAs are present, and in almost all cases as "hitchhikers", i.e. located in minicircles that also contain gRNAs for A6-ATPase or RPS12. As shown in Fig. 5, out of 54 minicircle classes, only 6 contain gRNAs that do not participate in the editing of A6-ATPase or RPS12 mRNAs (those that contain gRNAs for ND8, ND7 and COIII).

3.5. Analysis of γ subunit of ATP synthase

The situation described in the previous sections is highly suggestive that the American *T. vivax* strains are undergoing an "evolutionary journey" toward the derived mitochondrial genome somewhat similar to that observed in the *T. brucei* relatives which remain exclusively in the mammal host and have lost their ability to survive (reproduce) in tsetse flies. A critical step in the process of adaptation to become independent of mitochondrial genes is the series of modifications that affect the nuclear gene encoding γ subunit from ATP synthase. Lai et al. [7] identified in *T. equiperdum* and *T. evansi* some amino acid changes located in amino acid positions that are evolutionarily conserved across trypanosomatids. These authors suggested that these mutations might confer to this gene the ability to compensate the loss of F_0 portion from ATP synthase (encoded by the mitochondrial A6-ATPase gene, which is missing in these trypanosomes). Very recently Dean et al. [12] conducted an extensive study to test which ones of these (and other) changes in the γ subunit from ATP synthase are able to compensate the loss of kDNA in *T. brucei* strains and sub-species. It was observed that many of these variants (see Fig. 6) confer mutant *T. brucei* the ability to survive kDNA absence and even produce electric potential. It has been suggested that the acquisition of these

Table 4
Guide RNAs (gRNA) identified in MT1 and Liem176 strains. The gRNA are grouped according to the maxicircle gene where they exert their function. For each gRNA gene: length (of pairing region), the position where it matches in the cognate mRNA, number of mismatches and the minicircle that contains the gRNA are indicated.

Gene start–end[a]	Minicircle[b]	Mismatches[c]	Length[d]
gRNA for ND8			
108–71	TvMinic15	6	38
102–79	TvMinic45	1	24
181–153	TvMinic3[e]	4	30
181–153	TvMinic5	4	30
302–252	TvMinic7	6	51
gRNA for ND9			
78–40	TvMinic29	12	39
101–66	TvMinic6	7	36
128–97	TvMinic33	9	33
247–223	TvMinic6	3	26
383–341	TvMinic14	10	43
415–385	TvMinic24	5	31
gRNA for ND7			
110–76	TvMinic19	5	35
315–277	TvMinic41	7	39
434–405	TvMinic2	5	30
721–687	TvMinic44	4	36
721–695	TvMinic44	3	28
RNAg for COIII			
90–64	TvMinic47	5	27
262–226	TvMinic38	13	38
gRNA for Cyb			
No gRNA detected			
gRNA for CR3			
159–140	TvMinic45	3	20
227–191	TvMinic54	7	37
gRNA for COII			
No gRNA detected			
RNAg for MURF2 gRNA for MURF2			
No gRNA found			
RNAg for CR4			
27–1	TvMinic28	4	27
210–182	TvMinic46	2	29
334–289	TvMinic26	8	46
412–364	TvMinic42	11	49
gRNA for ND3			
No gRNA detected			
gRNA for A6			
39–1	TvMinic34	6	39
60–27	TvMinic14	5	34
81–43	TvMinic54	10	39
97–63	TvMinic8	7	35
111–75	TvMinic31	7	37
141–98	TvMinic35	12	44
166–116	TvMinic40	13	51
190–154	TvMinic51	8	37
166–192	TvMinic16[e]	16	40
237–204	TvMinic52[e]	5	34
266–210	TvMinic1	18	60
267–217	TvMinic29	13	51
318–273	TvMinic49	11	46
345–307	TvMinic25	12	39
366–330	TvMinic32	8	37
389–362	TvMinic11	3	28
425–378	TvMinic19	10	48
433–399	TvMinic37[e]	5	35
462–414	TvMinic38	5	49
509–473	TvMinic48	7	37
548–513	TvMinic36[e]	6	36
559–525	TvMinic13	7	35
573–542	TvMinic16[e]	3	32
604–558	TvMinic20	12	47
614–585	TvMinic45	4	30
646–627	TvMinic30	2	20
673–635	TvMinic27	7	39
686–656	TvMinic44[e]	4	31
711–673	TvMinic33	9	39
724–689	TvMinic46	7	35
754–715	TvMinic10	8	40

Table 4 (*Continued*)

Gene start–end[a]	Minicircle[b]	Mismatches[c]	Length[d]
gRNA for RSP12			
34–1	TvMinic9	7	34
65–21	TvMinic6	12	45
51–32	TvMinic35	2	20
104–76	TvMinic21	2	29
124–93	TvMinic28	6	32
150–108	TvMinic24	11	43
175–140	TvMinic23	10	36
192–160	TvMinic18	6	34
218–180	TvMinic42	9	39
247–204	TvMinic26	7	44
242–207	TvMinic41	9	36
255–238	TvMinic50	1	18

[a] Coordinates where the gRNA matches on its cognate mRNA (mature mRNA).
[b] Minicircle containing the gRNA.
[c] Number of differences between gRNA and its cognate mRNA in the region of pairing.
[d] Length of gRNA segment in the region of pairing (between gRNA and cognate mature mRNA).
[e] Minicircle not detected in Liem.

changes was a requirement to survive as a BS exclusively form [12].

We searched for these, or equivalent, amino acid substitutions in *T. vivax*. Fig. 6 shows the sub-alignments in relevant regions of γ ATPase in representative species from trypanosomatids, several strains of *T. evansi* and *T. equiperdum* and the three *T. vivax* strains studied in this work. As it can be observed all the amino acid positions previously indicated to be involved in conferring this new functionality to the γ subunit of ATPase exhibit, in the three *T. vivax* strains, the canonic amino acid (i.e. the same observed in *T. cruzi*, *Leishmania* and wild type *T. brucei*).

It is worth noting that there is some variability for this gene in *T. vivax*. On the one hand, Y486 has two alleles that are 98% identical at the amino acid level, while the two American *T. vivax* strains are almost identical to each other. In fact, SNP calling by read back-mapping shows that MT1 is heterozygous for this locus having only a synonymous change in Ala 35 (GCT ↔ GCC).

On the other hand, both Americans strains do exhibit two differences with Y486 at position 60 and 61 (T60A, T61(AG)); however, these differences (as well as the differences between the two Y486 alleles) are located in regions of poor interspecific amino acid conservation, suggesting that these variations are not functionally relevant.

These observations strongly suggest that this gene did not suffer the compensatory mutations observed in the other African trypanosomes that are unable to survive in the insect. Something that is compatible with the fact that the American *T. vivax* strains preserve the mitochondrial A6-ATPase gene fully functional.

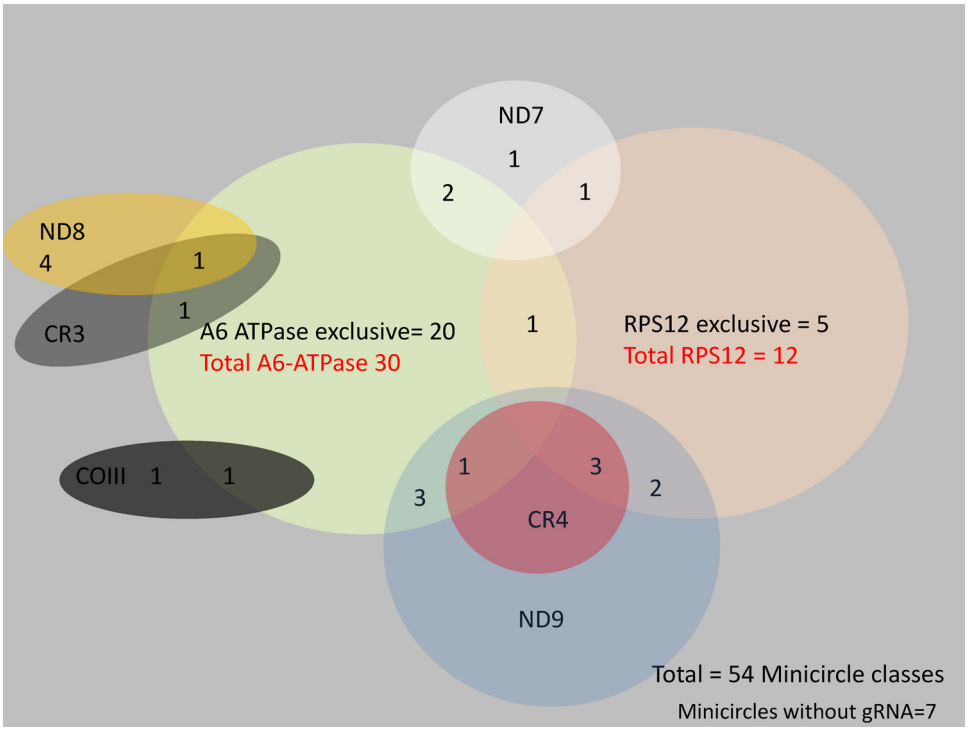


Fig. 5. Venn diagram showing the number of minicircle classes containing gRNA for each gene. Only six minicircle classes have gRNAs that do not participate in the editing of A6-ATPase and/or RPS12, whereas in seven minicircle classes no gRNAs were detected.

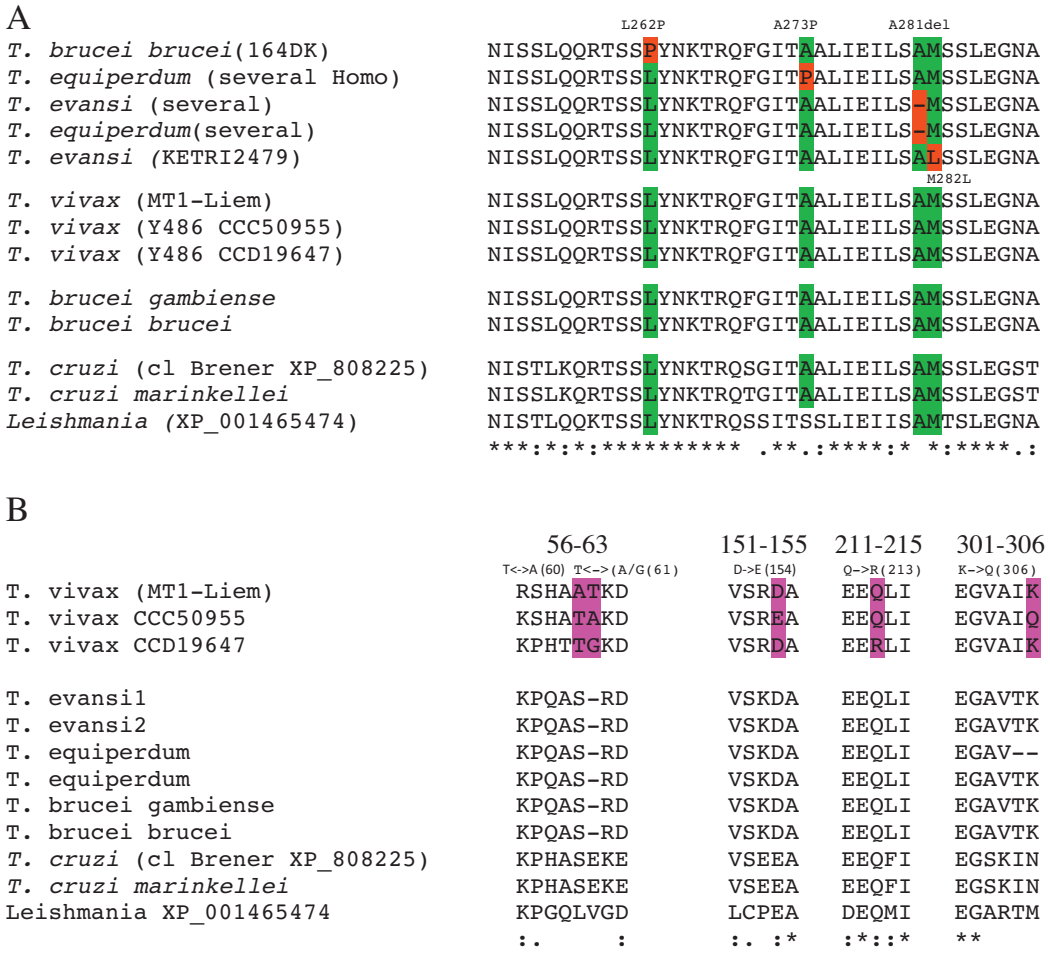


Fig. 6. Multiple alignment in relevant γ ATPase regions. The alignment includes representative trypanosomatid species, the African and American strains from *T. vivax* and several *T. evansi* and *T. equiperdum* strains that contain amino acid changes capable of compensating the loss of mitochondrial genome. (A) The region spans from amino acid positions 250–289, and contains amino acids that have been associated to conferring new functionalities to γ ATPase. These changes are shadowed in red, while the canonical variant (i.e. the wild type in the remaining species) in green. (B) Sub-alignments containing the regions that surround the amino acid positions (shadowed in violet) which present variability in *T. vivax*.

4. Discussion

In this work we analyze several aspects in relation to the evolutionary process that took place in the mitochondrial genomes of American strains from *T. vivax* with special stress on the changes that resulted from the new lifestyle that this parasite acquired in the Americas concerning its mechanical transmission.

The maxicircle genome sequences were determined for three *T. vivax* strains, two from America and the African strain Y486. In turn the "minicirculome" (minicircle population) was determined in detail for the two American strains.

This group of three strains is particularly suitable to tackle the analysis of the genome changes associated with the adoption of mechanical transmission because of their genetic proximity but different life cycle. In effect, two of the strains analyzed in this work are representative of American *T. vivax* strains, which are closely related to West African strains [33]. The third strain analyzed in this work, Y486, is of West African origin and a close relative to the African strain that migrated to America. It is worth mentioning that Y486 is derived from naturally infected cattle from Nigeria [17], and it has been shown that it can be cyclically transmitted by several species of tsetse flies [17,34,35].

It has been postulated that the introduction of *T. vivax* in South America took place around 1870 by infected Zebu cattle introduced in French Guiana [36–38]. A recent and comprehensive assessment of genetic variation in *T. vivax* indicates that American strains are genetically homogeneous clustering monophyletically when compared with West African strains [33]. Note that monophyly means that these strains coalesce to a unique ancestor and therefore supports the notion that at least for the most common circulating strains, the incursion in America from Africa occurred only once.

The comparison of maxircircle protein coding genes shows that in American strains two genes exhibit large deletions (ND7 and COIII), and three genes (ND1, ND2 and ND4) frameshift causing indels, all of which implying a complete loss of function, yet other genes have missense mutations. Analyses of expression and editing show that all coding genes are transcribed in the American strains, but only three (A6-ATPase, RPS12 and MURF2) are correctly edited. The fact that some genes are edited but others are not is a clear indication that whereas the enzymatic machinery of editing is fully functional, the guide RNAs necessary for the editing of the latter genes are absent. Guide RNAs were inferred by the analysis of minicircle sequences confirming this conjecture, since few gRNA genes were detected apart from those ones involved in the editing of A6-ATPase and RPS12. In addition, most of the gRNA genes not involved in the editing of these two genes are located in minicircles that also contain gRNA genes for A6-ATPase and RPS12, suggesting that they are passively carried. This allows one to conclude that the mitochondrial genome from American strains of *T. vivax* is suffering a drastic genome wide degradation process of

its coding capacity (severe mutations and loss of editing capacity).

It is probable that this process started after the introduction of *T. vivax* in America, since in its evolutionary close relative, the West African strain Y486, all maxicircle genes are transcribed and correctly edited not only in insect stage of life cycle but also in the bloodstream form. It is surprising that the degradation process has progressed to this extent, considering that it probably started so recently (less than two centuries ago). It must be taken into account that if the phylogenetic inferences and times of divergence among *T. vivax* strains mentioned before were confirmed with additional data, they would imply that the evolutionary speed described here for these parasites would be unprecedented for eukaryotes, even for mitochondrial genomes.

As long as the biological causes are concerned, it is reasonable to postulate that, like in *T. brucei*-like species *T. equiperdum* and *T. evansi*, the driving force responsible of this massive genome decay is the fact that oxidative phosphorylation is not essential in the BS stage of the parasite, and hence there is no selective pressure to favor the persistence of these genes. In turn A6-ATPase and RPS12 genes are kept intact in American strains because their functions are still necessary in the BS stage of the parasite. It is worth mentioning that this latter aspect differs radically from the strategy followed by the *T. brucei* relatives that remain as BS only parasites, where also A6-ATPase (and RPS12) is either nonfunctional or it was simply deleted like the remaining maxicircle genes [7,39]. In these species the loss of A6-ATPase protein (whose function is required to survive in mammals) is complemented by the nuclearly encoded γ ATPase which acquires new functionalities [7,12]. This is a phenomenon that has occurred several independent times in evolution as evidenced by the multiple (non-monophyletic) origin of *T. equiperdum* and *T. evansi* strains as well as by the existence of different, non-related, mutations that confer complementing capabilities to γ ATPase subunit (list appears in Fig. 6).

This alternative strategy observed in *T. vivax* to cope with mitochondrial genome disintegration has been theoretically postulated to exist as a transitional stage in the progression until complete elimination of maxicircles in non-cycling *T. brucei* relatives [8]. According to this viewpoint the process could only proceed after a compensatory mutation in the γ ATPase would eliminate the need of A6-ATPase [8]. However to the best our knowledge such transitional stage has never been observed before neither in nature nor in the laboratory. Other authors have proposed that the acquisition of mutations in the γ ATPase gene is a precondition rather than a final step in the evolution of kinetoplast elimination [6] and hence what we observe in today's circulating American strains from *T. vivax* would represent an alternative evolutionary direction instead of an intermediary step.

Funding

This work was supported by grant FCE_2_2011_1_6850 (Fondo Clemente Estable, Agencia Nacional de Investigación e Innovación (ANII) from Uruguay) and partially funded by FOCEM (MERCOSUR Structural Convergence Fund), grant COF 03/11.

G.G., C.R. and F.A.V. are researchers from the Sistema Nacional de Investigadores (ANII), Uruguay A.R. is a researcher from Prometeo program (Ecuador).

Conflict of interest statement

There are no conflicts of interest to declare.

Acknowledgement

We thank Paula Tucci for critical reading of the manuscript. Cryostabilites of *Trypanosoma vivax* (Y486 strain) were kindly provided by Philippe Büscher (Parasite Diagnostics Unit, Department of Parasitology, Institute of Tropical Medicine Antwerp, Belgium).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.mrfmmm.2015.01.008.

References

[1] R.A. Corell, J.E. Feagin, G.R. Riley, T. Strickland, J.A. Guderian, P.J. Myler, K. Stuart, *Trypanosoma brucei* minicircles encode multiple guide RNAs which can direct editing of extensively overlapping sequences, Nucleic Acids Res. 21 (1993) 4313–4320.
[2] L. Simpson, O.H. Thiemann, N.J. Savill, J.D. Alfonzo, D.A. Maslov, Evolution of RNA editing in trypanosome mitochondria, Proc. Natl. Acad. Sci. U.S.A. 97 (2000) 6986–6993.
[3] D. Koslowsky, Y. Sun, J. Hindenach, T. Theisen, J. Lucas, The insect-phase gRNA transcriptome in *Trypanosoma brucei*, Nucleic Acids Res. 42 (2014) 1873–1886.
[4] F. Bringaud, L. Riviere, V. Coustou, Energy metabolism of trypanosomatids: adaptation to available carbon sources, Mol. Biochem. Parasitol. 149 (2006) 1–9.
[5] A. Zikova, A. Schnaufer, R.A. Dalley, A.K. Panigrahi, K.D. Stuart, The F(0)F(1)-ATP synthase complex contains novel subunits and is essential for procyclic *Trypanosoma brucei*, PLoS Pathog. 5 (2009) e1000436.
[6] Z.R. Lun, D.H. Lai, F.J. Li, J. Lukes, F.J. Ayala, *Trypanosoma brucei*: two steps to spread out from Africa, Trends Parasitol. 26 (2010) 424–427.
[7] D.H. Lai, H. Hashimi, Z.R. Lun, F.J. Ayala, J. Lukes, Adaptations of *Trypanosoma brucei* to gradual loss of kinetoplast DNA: *Trypanosoma equiperdum* and *Trypanosoma evansi* are petite mutants of *T. brucei*, Proc. Natl. Acad. Sci. U.S.A. 105 (2008) 1999–2004.
[8] R.E. Jensen, L. Simpson, P.T. Englund, What happens when *Trypanosoma brucei* leaves Africa, Trends Parasitol. 24 (2008) 428–431.
[9] A. Schnaufer, A.K. Panigrahi, B. Panicucci, R.P. Igo Jr., E. Wirtz, R. Salavati, K. Stuart, An RNA ligase essential for RNA editing and survival of the bloodstream form of *Trypanosoma brucei*, Science 291 (2001) 2159–2162.
[10] S.V. Brown, P. Hosking, J. Li, N. Williams, ATP synthase is responsible for maintaining mitochondrial membrane potential in bloodstream form *Trypanosoma brucei*, Eukaryot Cell 5 (2006) 45–53.
[11] X.J. Chen, G.D. Clark-Walker, Specific mutations in alpha- and gamma-subunits of F1-ATPase affect mitochondrial genome integrity in the petite-negative yeast *Kluyveromyces lactis*, EMBO J. 14 (1995) 3277–3286.
[12] S. Dean, M.K. Gould, C.E. Dewar, A.C. Schnaufer, Single point mutations in ATP synthase compensate for mitochondrial genome loss in trypanosomes, Proc. Natl. Acad. Sci. U.S.A. 110 (2013) 14741–14746.
[13] G. Greif, M. Ponce de Leon, G. Lamolle, M. Rodriguez, D. Pineyro, L.M. Tavares-Marques, A. Reyna-Bello, C. Robello, F. Alvarez-Valin, Transcriptome analysis of the bloodstream stage from the parasite *Trypanosoma vivax*, BMC Genomics 14 (2013) 149.
[14] A.P. Cortez, R.M. Ventura, A.C. Rodrigues, J.S. Batista, F. Paiva, N. Anez, R.Z. Machado, W.C. Gibson, M.M. Teixeira, The taxonomic and phylogenetic relationships of *Trypanosoma vivax* from South America and Africa, Parasitology 133 (2006) 159–169.
[15] M. Desquesnes, M.L. Dia, Mechanical transmission of *Trypanosoma vivax* in cattle by the African tabanid *Atylotus fuscipes*, Vet. Parasitol. 119 (2004) 9–19.
[16] M. Desquesnes, Livestock trypanosomoses and their vectors in Latin America, OIE & CIRAD, Paris, 2004.
[17] W. Gibson, The origins of the trypanosome genome strains *Trypanosoma brucei* TREU 927, *T. b. gambiense* DAL 972, *T. vivax* Y486 and *T. congolense* IL3000, Parasit. Vectors 5 (2012) 71.
[18] A.P. Jackson, A. Berry, M. Aslett, H.C. Allison, P. Burton, J. Vavrova-Anderson, et al., Antigenic diversity is generated by distinct evolutionary mechanisms in African trypanosome species, Proc. Natl. Acad. Sci. U.S.A. 109 (2012) 3416–3421.
[19] J.T. Simpson, K. Wong, S.D. Jackman, J.E. Schein, S.J. Jones, I. Birol, ABySS: a parallel assembler for short read sequence data, Genome Res. 19 (2009) 1117–1123.
[20] A. Bankevich, S. Nurk, D. Antipov, A.A. Gurevich, M. Dvorkin, A.S. Kulikov, et al., SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing, J. Comput. Biol. 19 (2012) 455–477.
[21] B. Chevreux, T. Pfisterer, B. Drescher, A.J. Driesel, W.E. Muller, T. Wetter, S. Suhai, Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs, Genome Res. 14 (2004) 1147–1159.

[22] B. Langmead, S.L. Salzberg, Fast gapped-read alignment with Bowtie 2, Nat. Methods 9 (2012) 357–359.
[23] A. Mortazavi, B.A. Williams, K. McCue, L. Schaeffer, B. Wold, Mapping and quantifying mammalian transcriptomes by RNA-Seq, Nat. Methods 5 (2008) 621–628.
[24] M. Garber, M.G. Grabherr, M. Guttman, C. Trapnell, Computational methods for transcriptome annotation and quantification using RNA-seq, Nat. Methods 8 (2011) 469–477.
[25] I. Milne, G. Stephen, M. Bayer, P.J. Cock, L. Pritchard, L. Cardle, P.D. Shaw, D. Marshall, Using Tablet for visual exploration of second-generation sequencing data, Brief. Bioinform. 14 (2013) 193–202.
[26] S. Thomas, L.L. Martinez, S.J. Westenberger, N.R. Sturm, A population study of the minicircles in *Trypanosoma cruzi*: predicting guide RNAs in the absence of empirical RNA editing, BMC Genomics 8 (2007) 133.
[27] D.S. Ray, Conserved sequence blocks in kinetoplast minicircles from diverse species of trypanosomes, Mol. Cell. Biol. 9 (1989) 1365–1367.
[28] P. Borst, F. Fase-Fowler, P.J. Weijers, J.D. Barry, L. Tetley, K. Vickerman, Kinetoplast DNA from *Trypanosoma vivax* and *T. congolense*, Mol. Biochem. Parasitol. 15 (1985) 129–142.
[29] J. Grams, M.T. McManus, S.L. Hajduk, Processing of polycistronic guide RNAs is associated with RNA editing complexes in *Trypanosoma brucei*, EMBO J. 19 (2000) 5525–5532.
[30] R. Thertulien, G. Harth, C.G. Haidaris, Evidence that the entire length of a kinetoplast DNA minicircle is transcribed in *Trypanosoma cruzi*, J. Mol. Microbiol. 5 (1) (1991) 207–215.
[31] T. Ochsenreiter, M. Cipriano, S.L. Hajduk, KISS: the kinetoplastid RNA editing sequence search tool, RNA 13 (2007) 1–4.
[32] S.L. Clement, M.K. Mingler, D.J. Koslowsky, An intragenic guide RNA location suggests a complex mechanism for mitochondrial gene expression in *Trypanosoma brucei*, Eukaryot Cell 3 (2004) 862–869.
[33] H.A. Garcia, A.C. Rodrigues, C.M. Rodrigues, Z. Bengaly, A.H. Minervino, F. Riet-Correa, et al., Microsatellite analysis supports clonal propagation and reduced divergence of *Trypanosoma vivax* from asymptomatic to fatally infected livestock in South America compared to West Africa, Parasit. Vectors 7 (2014) 210.
[34] A.L. De Gee, K. Ige, P. Leeflang, Studies on *Trypanosoma vivax*: transmission of mouse infective *T. vivax* by tsetse flies, Int. J. Parasitol. 6 (1976) 419–421.
[35] S. D'Archivio, M. Medina, A. Cosson, N. Chamond, B. Rotureau, P. Minoprio, S. Goyard, Genetic engineering of *Trypanosoma* (*Dutonella*) *vivax* and in vitro differentiation under axenic conditions, PLoS Negl. Trop. Dis. 5 (2011) e1461.
[36] H.B.M. Fabre, Sur un nouveau foyer de Trypanosomiase bovine observé a la Guadaloupe, Bull. Soc. Pathol. Exot. 19 (1926) 435–437.
[37] M. Carougeau, Trypanosomiase bovine à la Guadeloupe, Bull. Soc. Pathol. Exot. 22 (1929) 246–247.
[38] A.V.M. Leger, Epizootic á Trypanosomes chez les bovides de la Guyane Francaise, Bull. Soc. Pathol. Exot. 12 (1919) 258–266.
[39] F.J. Li, D.H. Lai, J. Lukes, X.G. Chen, Z.R. Lun, Doubts about *Trypanosoma equiperdum* strains classed as *Trypanosoma brucei* or *Trypanosoma evansi*, Trends Parasitol. 22 (2006) 55–56, author reply 58–59.

Conclusiones y Perspectivas

En esta tesis se presentan los resultados obtenidos en los estudios realizados en el parásito africano *Trypanosoma vivax*, el cual debido a su crucial ubicación filogenética, estamos desarrollando como modelo para entender algunas de las notables particularidades de los tripanosomas africanos. Este trabajo se centra en el análisis de los genomas de las cepas americanas Liem-176 y MT1 así como su ancestro africano más cercano, la cepa Y486. Se combinan análisis transcriptómicos, genómicos y evolutivos. Los procesos adaptativos que ocurrieron en el genoma mitocondrial entre las cepas americanas (Liem-176 y MT1) y la africana (Y486) se analizan detalladamente.

Respecto a la primera parte de esta tesis, nuestro trabajo constituye el primer análisis transcriptómico en un tripanosomátido, exceptuando los realizados en el organismo modelo *Trypanosoma brucei* [60, 123]. Se generó una base de datos pública con las secuencias que desde su publicación presenta un promedio de ingreso cercano a 100 visitas mensuales provenientes de 53 países, por lo cual consideramos que la herramienta generada constituye un aporte a la comunidad científica nacional e internacional.

Los datos obtenidos fueron los primeros de este tipo en ser producidos, procesados, analizados e interpretados por nuestro grupo. A partir de esos datos se logró obtener información relevante sobre la cubierta de proteínas de superficie de estos parásitos aunque en *Trypanosoma vivax* no está claro aún el rol que juega la variación antigénica de superficie en la evasión de la respuesta inmune. No se han identificado los sitios de expresión ni el repertorio de genes VSG que posee esta especie. En relación a este punto, nuestro grupo de trabajo se encuentra abocado a resolver algunos de estas interrogantes.

Cabe preguntarse ¿porqué consideramos tan importante disecar el sistema de variación antigénica, determinar el repertorio de genes VSGs, los mecanismos de regulación de la variación antigénica y los sitios de expresión en *T. vivax*? ¿Su rol es el mismo que se ha caracterizado en *T. brucei*? La respuesta a estas preguntas es que todos los análisis filogenéticos indican que *T. vivax* es un descendiente directo del ancestro de los tripanosomas africanos, por lo que es muy probable que esta especie presente la organización de la superfamilia de genes VSG en el estado ancestral. Es lógico pensar, entonces, que el sistema de variación antigénica haya comenzado en estos parásitos. Esta es la razón por la que identificar sus genes VSGs, así como su organización genómica, es clave para comprender como surgió esta superfamilia de genes, y por ende intentar aportar al conocimiento de uno de los mecanismos de evasión del sistema inmunitario más sorprendentes de la naturaleza.

Asimismo seguir la dinámica de expresión en los diferentes picos de parasitemia, así como determinar posibles cambios a nivel genómico en los sitios de expresión, son preguntas que intentamos resolver con experimentos en curso. Desde el año 2012 contamos con el modelo murino de la cepa Y486. En este sentido, estamos realizando experimentos de infección en este modelo y aislando parásitos en diferentes momentos de la infección para intentar comprender los fenómenos y la dinámica de la variación antigénica en estos parásitos.

Queremos resaltar la importancia de clarificar los mecanismos de expresión pues estudios previos en esta especie no han arrojado resultados que validen la existencia de Sitios de Expresión similares a los observados en *T. brucei* [119]. Sin embargo resultados preliminares de secuenciación genómica que hemos obtenido recientemente (aún no publicados) aportan algunas evidencias en el sentido de que estos sitios efectivamente existen en *T. vivax*.

Dadas estas incertidumbres, las siguientes preguntas enmarcan los aspectos de mayor relevancia en los cuales estamos trabajando en la actualidad: ¿Cómo es la expresión del gen VSG activo en *T. vivax*? ¿Cómo se controla? ¿Existe un Sitio de Expresión como sucede en *T. brucei*, en el que junto al gen VSG se co-transcriben otros genes (ESAG) mediante la ARN Polimerasa I?

Para intentar responder estas interrogantes, hemos realizado en nuestro laboratorio, la secuenciación del genoma completo de la cepas MT1 y Liem-176 de *Trypanosoma vivax,* obteniendo un ensamblaje de alta calidad, el cual está siendo analizado al momento de escritura de esta tesis. Los análisis primarios sobre estos datos muestran resultados muy interesantes relacionados al repertorio de genes VSG, la organización del genoma nuclear y la presencia de regiones que se encuentran específicamente en el genoma de cepas americanas.

Con relación a la determinación de los sitios de *trans-splicing* para todos los genes expresados y su comparación con los datos disponibles públicamente del parásito relacionado *T. brucei* es de resaltar que cerca de 600 genes no presentaban secuencias 5’ UTR y que el *Spliced-leader* se encontraba inmediatamente antes del codón de inicio de la traducción. Análisis de Ontología de genes sobre este grupo mostraron un enriquecimiento de proteínas relacionadas con la traducción, indicando un posible mecanismo de regulación común en este grupo de genes.

Respecto a los patrones de expresión, en esta tesis presentamos evidencia de la existencia de regiones del genoma de *T. vivax* que contienen CDS pero que no muestran actividad transcripcional. La completa inactividad transcripcional de dichas regiones la hemos corroborado reiteradamente con datos de RNAseq provenientes de

diversas fuentes (manuscrito en preparación). Esto sugeriría la existencia de regulación al inicio de la transcripción, lo que representa un punto controversial en los tripanosomátidos, donde el punto de vista más aceptado es que se transcribe la totalidad de sus genes y la regulación se da nivel post-transcripcional.

Respecto a la segunda parte de esta tesis (Objetivo Nº2) resaltamos que se logró la obtención de los genomas del maxicírculo de 2 cepas americanas (Liem-176 y MT1) y de la cepa africana (Y486, derivada de ganado infectado naturalmente en Nigeria, proveniente de África occidental y filogenéticamente relacionada con las cepas americanas) de *Trypanosoma vivax*. Asimismo se logró la secuenciación e identificación del set completo de clases de minicírculos de las cepas americanas. El análisis comparativo de estos datos, así como su combinación con los datos de transcripción, generados por nuestro grupo y de bases de datos públicas (SRA), permitió obtener información relevante respecto a la evolución de estos parásitos.

En primer lugar, demostramos la degradación en su capacidad codificante del genoma mitocondrial, que han sufrido las cepas americanas de *T. vivax* desde su introducción en el continente. Esta degradación incluye una gran deleción (752 bases) y deleciones puntuales en el genoma mitocondrial, con consecuencias en la expresión de los genes maxicirculares.

En segundo lugar, mostramos una reducción en las clases de minicírculos necesarios para la edición de los genes mitocondriales, exceptuando los requeridos para el *editing* de tres genes (A6-ATP sintasa, la proteína ribosomal S12 –RPS12- y MURF2). Además, mostramos cómo únicamente se expresaban y editaban correctamente estos tres genes, mientras que en la cepa africana (Y486) se expresan y editan correctamente todos los genes mitocondriales.

Las subespecies de *T. brucei*, *T.b. evansi* y *T.b. equiperdum,* que al igual que las cepas americanas de *T. vivax* permanecen únicamente en su forma sanguínea, sin completar su ciclo en el insecto, han perdido parte o todo su genoma mitocondrial. En nuestro caso de estudio, estamos observando una fase intermedia de este proceso evolutivo (no observado en la naturaleza en los casos antes mencionados). Se ha postulado, que este proceso evolutivo se asocia con el cambio de ciclo de vida de estos parásitos y el rol cambiante que juega la mitocondria en este proceso. Dado que contamos con las secuencias genómica completas de ambas cepas americanas de *T. vivax* (MT1 y Liem-176) planeamos determinar si un proceso similar al observado en el genoma mitocondrial también ha afectado a los genes nucleares que codifican proteínas mitocondriales.

Bibliografía

1. Gluenz E, Povelones ML, Englund PT, Gull K: The kinetoplast duplication cycle in Trypanosoma brucei is orchestrated by cytoskeleton-mediated cell morphogenesis. *Mol Cell Biol* 2011, 31(5):1012-1021.
2. Chagas C: A new human trypanosomiasis. Morphology and life cycle of Schyzotrypanum cruzi, the cause of a new human disease. . *Mem Inst Oswaldo Cruz* 1909, 1:159-218.
3. Coutinho M, Freire O, Jr., Dias JC: The noble enigma: Chagas' nominations for the Nobel prize. *Mem Inst Oswaldo Cruz* 1999, 94 Suppl 1:123-129.
4. Datos obtenidos de World Health Organization, Marzo 2015 [www.who.org]
5. Bartholomeu DC, de Paiva RM, Mendes TA, DaRocha WD, Teixeira SM: Unveiling the intracellular survival gene kit of trypanosomatid parasites. *PLoS Pathog* 2014, 10(12):e1004399.
6. Martinez-Calvillo S, Vizuet-de-Rueda JC, Florencio-Martinez LE, Manning-Cela RG, Figueroa-Angulo EE: Gene expression in trypanosomatid parasites. *J Biomed Biotechnol* 2010, 2010:525241.
7. Steverding D: The history of African trypanosomiasis. *Parasit Vectors* 2008, 1(1):3.
8. FAO: http://www.fao.org/. 2015.
9. Leger M, Vienne, M.: Epizootic á Trypanosomes chez les bovides de la Guyane Francaise. *Bull Soc Path Exot* 1919, 12:8.
10. Curasson G (ed.): Traité de protozoologie vétérinaire et comparée. I. trypanosomes. . París; 1943.
11. Maillard J-C, Maillard N: Historique du peuplement bovin et de l'introduction de la tique #Amblyomma variegatum# dans les îles françaises des Antilles : synthèse bibliographique. *Ethnozootechnie* 1998:19-35.
12. Garcia HA, Rodrigues AC, Rodrigues CM, Bengaly Z, Minervino AH, Riet-Correa F, Machado RZ, Paiva F, Batista JS, Neves L *et al*: Microsatellite analysis supports clonal propagation and reduced divergence of Trypanosoma vivax from asymptomatic to fatally infected livestock in South America compared to West Africa. *Parasit Vectors* 2014, 7:210.
13. Osorio AL, Madruga CR, Desquesnes M, Soares CO, Ribeiro LR, Costa SC: Trypanosoma (Duttonella) vivax: its biology, epidemiology, pathogenesis, and introduction in the New World--a review. *Mem Inst Oswaldo Cruz* 2008, 103(1):1-13.
14. Cadioli FA, Barnabe Pde A, Machado RZ, Teixeira MC, Andre MR, Sampaio PH, Fidelis Junior OL, Teixeira MM, Marques LC: First report of Trypanosoma vivax outbreak in dairy cattle in Sao Paulo state, Brazil. *Rev Bras Parasitol Vet* 2012, 21(2):118-124.
15. Ocorrência de Trypanosoma evansi em eqüinos no município de Cruz Alta, RS, Brasil [http://biblioteca.universia.net/html_bura/ficha/params/id/44759865.html]
16. Campbell DA, Thomas S, Sturm NR: Transcription in kinetoplastid protozoa: why be normal? *Microbes Infect* 2003, 5(13):1231-1240.
17. Wheeler RJ, Gluenz E, Gull K: The limits on trypanosomatid morphological diversity. *PLoS One* 2013, 8(11):e79581.
18. Walshe DP, Pheng Ooi, C., Lehane, M.J., Haines, L.R.: Chapter 3 The Enemy Within: Interactions Between Tsetse, Trypanosomes and Symbionts. *Advances in Insect Physiology* 2010, 37:119-175.
19. Mony BM, MacGregor P, Ivens A, Rojas F, Cowton A, Young J, Horn D, Matthews K: Genome-wide dissection of the quorum sensing signalling pathway in Trypanosoma brucei. *Nature* 2014, 505(7485):681-685.

20. Matthews KR, Ellis JR, Paterou A: Molecular regulation of the life cycle of African trypanosomes. *Trends Parasitol* 2004, 20(1):40-47.

21. Desquesnes M, Dia ML: Mechanical transmission of Trypanosoma vivax in cattle by the African tabanid Atylotus fuscipes. *Vet Parasitol* 2004, 119(1):9-19.

22. Ferenc SA, Stopinski V, Courtney CH: The development of an enzyme-linked immunosorbent assay for Trypanosoma vivax and its use in a seroepidemiological survey of the Eastern Caribbean Basin. *Int J Parasitol* 1990, 20(1):51-56.

23. Desquesnes M: Livestock trypanosomoses and their vectors in Latin America. *Paris: OIE & CIRAD* 2004.

24. Van den Bossche P, Shumba W, Makhambera P: The distribution and epidemiology of bovine trypanosomosis in Malawi. *Vet Parasitol* 2000, 88(3-4):163-176.

25. Calvo-Alvarez E, Alvarez-Velilla R, Jimenez M, Molina R, Perez-Pertejo Y, Balana-Fouce R, Reguera RM: First evidence of intraclonal genetic exchange in trypanosomatids using two Leishmania infantum fluorescent transgenic clones. *PLoS Negl Trop Dis* 2014, 8(9):e3075.

26. Gaunt MW, Yeo M, Frame IA, Stothard JR, Carrasco HJ, Taylor MC, Mena SS, Veazey P, Miles GA, Acosta N *et al*: Mechanism of genetic exchange in American trypanosomes. *Nature* 2003, 421(6926):936-939.

27. Peacock L, Ferris V, Bailey M, Gibson W: Mating compatibility in the parasitic protist Trypanosoma brucei. *Parasit Vectors* 2014, 7:78.

28. El-Sayed NM, Myler PJ, Bartholomeu DC, Nilsson D, Aggarwal G, Tran AN, Ghedin E, Worthey EA, Delcher AL, Blandin G *et al*: The genome sequence of Trypanosoma cruzi, etiologic agent of Chagas disease. *Science* 2005, 309(5733):409-415.

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

30. Martinez-Calvillo S, Yan S, Nguyen D, Fox M, Stuart K, Myler PJ: Transcription of Leishmania major Friedlin chromosome 1 initiates in both directions within a single region. *Mol Cell* 2003, 11(5):1291-1299.

31. Siegel TN, Hekstra DR, Kemp LE, Figueiredo LM, Lowell JE, Fenyo D, Wang X, Dewell S, Cross GA: Four histone variants mark the boundaries of polycistronic transcription units in Trypanosoma brucei. *Genes Dev* 2009, 23(9):1063-1076.

32. Berriman M, Ghedin E, Hertz-Fowler C, Blandin G, Renauld H, Bartholomeu DC, Lennard NJ, Caler E, Hamlin NE, Haas B *et al*: The genome of the African trypanosome Trypanosoma brucei. *Science* 2005, 309(5733):416-422.

33. Ivens AC, Peacock CS, Worthey EA, Murphy L, Aggarwal G, Berriman M, Sisk E, Rajandream MA, Adlem E, Aert R *et al*: The genome of the kinetoplastid parasite, Leishmania major. *Science* 2005, 309(5733):436-442.

34. El-Sayed NM, Myler PJ, Blandin G, Berriman M, Crabtree J, Aggarwal G, Caler E, Renauld H, Worthey EA, Hertz-Fowler C *et al*: Comparative genomics of trypanosomatid parasitic protozoa. *Science* 2005, 309(5733):404-409.

35. Douzery EJ, Snell EA, Bapteste E, Delsuc F, Philippe H: The timing of eukaryotic evolution: does a relaxed molecular clock reconcile proteins and fossils? *Proc Natl Acad Sci U S A* 2004, 101(43):15386-15391.

36. Choi J, El-Sayed NM: Functional genomics of trypanosomatids. *Parasite Immunol* 2012, 34(2-3):72-79.

37. Peacock CS, Seeger K, Harris D, Murphy L, Ruiz JC, Quail MA, Peters N, Adlem E, Tivey A, Aslett M *et al*: Comparative genomic analysis of three Leishmania species that cause diverse human disease. *Nat Genet* 2007, 39(7):839-847.

38. **Downing T, Stark O, Vanaerschot M, Imamura H, Sanders M, Decuypere S, de Doncker S, Maes I, Rijal S, Sundar S *et al*: Genome-wide SNP and microsatellite variation illuminate population-level epidemiology in the Leishmania donovani species complex. *Infect Genet Evol* 2012, 12(1):149-159.**
39. **Downing T, Imamura H, Decuypere S, Clark TG, Coombs GH, Cotton JA, Hilley JD, de Doncker S, Maes I, Mottram JC *et al*: Whole genome sequencing of multiple Leishmania donovani clinical isolates provides insights into population structure and mechanisms of drug resistance. *Genome Res* 2011, 21(12):2143-2156.**
40. **Rogers MB, Hilley JD, Dickens NJ, Wilkes J, Bates PA, Depledge DP, Harris D, Her Y, Herzyk P, Imamura H *et al*: Chromosome and gene copy number variation allow major structural change between species and strains of Leishmania. *Genome Res* 2011, 21(12):2129-2142.**
41. **Raymond F, Boisvert S, Roy G, Ritt JF, Legare D, Isnard A, Stanke M, Olivier M, Tremblay MJ, Papadopoulou B *et al*: Genome sequencing of the lizard parasite Leishmania tarentolae reveals loss of genes associated to the intracellular stage of human pathogenic species. *Nucleic Acids Res* 2012, 40(3):1131-1147.**
42. **Jackson AP, Sanders M, Berry A, McQuillan J, Aslett MA, Quail MA, Chukualim B, Capewell P, MacLeod A, Melville SE *et al*: The genome sequence of Trypanosoma brucei gambiense, causative agent of chronic human african trypanosomiasis. *PLoS Negl Trop Dis* 2010, 4(4):e658.**
43. **Jackson AP, Berry A, Aslett M, Allison HC, Burton P, Vavrova-Anderson J, Brown R, Browne H, Corton N, Hauser H *et al*: Antigenic diversity is generated by distinct evolutionary mechanisms in African trypanosome species. *Proc Natl Acad Sci U S A* 2012, 109(9):3416-3421.**
44. **Real F, Vidal RO, Carazzolle MF, Mondego JM, Costa GG, Herai RH, Wurtele M, de Carvalho LM, Carmona e Ferreira R, Mortara RA *et al*: The genome sequence of Leishmania (Leishmania) amazonensis: functional annotation and extended analysis of gene models. *DNA Res* 2013, 20(6):567-581.**
45. **Stoco PH, Wagner G, Talavera-Lopez C, Gerber A, Zaha A, Thompson CE, Bartholomeu DC, Luckemeyer DD, Bahia D, Loreto E *et al*: Genome of the avirulent human-infective trypanosome--Trypanosoma rangeli. *PLoS Negl Trop Dis* 2014, 8(9):e3176.**
46. **Carnes J, Anupama A, Balmer O, Jackson A, Lewis M, Brown R, Cestari I, Desquesnes M, Gendrin C, Hertz-Fowler C *et al*: Genome and phylogenetic analyses of trypanosoma evansi reveal extensive similarity to T. brucei and multiple independent origins for dyskinetoplasty. *PLoS Negl Trop Dis* 2015, 9(1):e3404.**
47. **Johnson PJ, Kooter JM, Borst P: Inactivation of transcription by UV irradiation of T. brucei provides evidence for a multicistronic transcription unit including a VSG gene. *Cell* 1987, 51(2):273-281.**
48. **Sather S, Agabian N: A 5' spliced leader is added in trans to both alpha- and beta-tubulin transcripts in Trypanosoma brucei. *Proc Natl Acad Sci U S A* 1985, 82(17):5695-5699.**
49. **Shea C, Van der Ploeg LH: Stable variant-specific transcripts of the variant cell surface glycoprotein gene 1.8 expression site in Trypanosoma brucei. *Mol Cell Biol* 1988, 8(2):854-859.**
50. **Pays E, Coquelet H, Pays A, Tebabi P, Steinert M: Trypanosoma brucei: posttranscriptional control of the variable surface glycoprotein gene expression site. *Mol Cell Biol* 1989, 9(9):4018-4021.**

51. Vijayasarathy S, Ernest I, Itzhaki JE, Sherman D, Mowatt MR, Michels PA, Clayton CE: The genes encoding fructose bisphosphate aldolase in Trypanosoma brucei are interspersed with unrelated genes. *Nucleic Acids Res* 1990, 18(10):2967-2975.

52. Robello C: El locus H de Trypanosoma cruzi: caracterización de los genes ptr1 y tcp17. *Tesis Doctoral Universidad de Granada* 1998.

53. Jensen BC, Ramasamy G, Vasconcelos EJ, Ingolia NT, Myler PJ, Parsons M: Extensive stage-regulation of translation revealed by ribosome profiling of Trypanosoma brucei. *BMC Genomics* 2014, 15:911.

54. Vasquez JJ, Hon CC, Vanselow JT, Schlosser A, Siegel TN: Comparative ribosome profiling reveals extensive translational complexity in different Trypanosoma brucei life cycle stages. *Nucleic Acids Res* 2014, 42(6):3623-3637.

55. Smircich P, Eastman G, Bispo S, Duhagon MA, Guerra-Slompo EP, Garat B, Goldenberg S, Munroe DJ, Dallagiovanna B, Holetz F *et al*: Ribosome profiling reveals translation control as a key mechanism generating differential gene expression in Trypanosoma cruzi. *BMC Genomics* 2015, 16:443.

56. Maniatis T, Goodbourn S, Fischer JA: Regulation of inducible and tissue-specific gene expression. *Science* 1987, 236(4806):1237-1245.

57. Veitch NJ, Johnson PC, Trivedi U, Terry S, Wildridge D, MacLeod A: Digital gene expression analysis of two life cycle stages of the human-infective parasite, Trypanosoma brucei gambiense reveals differentially expressed clusters of co-regulated genes. *BMC Genomics* 2010, 11:124.

58. Jensen BC, Sivam D, Kifer CT, Myler PJ, Parsons M: Widespread variation in transcript abundance within and across developmental stages of Trypanosoma brucei. *BMC Genomics* 2009, 10:482.

59. Minning TA, Weatherly DB, Atwood J, 3rd, Orlando R, Tarleton RL: The steady-state transcriptome of the four major life-cycle stages of Trypanosoma cruzi. *BMC Genomics* 2009, 10:370.

60. Siegel TN, Hekstra DR, Wang X, Dewell S, Cross GA: Genome-wide analysis of mRNA abundance in two life-cycle stages of Trypanosoma brucei and identification of splicing and polyadenylation sites. *Nucleic Acids Res* 2010, 38(15):4946-4957.

61. Nilsson D, Gunasekera K, Mani J, Osteras M, Farinelli L, Baerlocher L, Roditi I, Ochsenreiter T: Spliced leader trapping reveals widespread alternative splicing patterns in the highly dynamic transcriptome of Trypanosoma brucei. *PLoS Pathog* 2010, 6(8):e1001037.

62. Ingolia NT, Ghaemmaghami S, Newman JR, Weissman JS: Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* 2009, 324(5924):218-223.

63. Miller PF, Hinnebusch AG: cis-acting sequences involved in the translational control of GCN4 expression. *Biochim Biophys Acta* 1990, 1050(1-3):151-154.

64. Englund PT: A passion for parasites. *J Biol Chem* 2014, 289(49):33712-33729.

65. Englund PT: The replication of kinetoplast DNA networks in Crithidia fasciculata. *Cell* 1978, 14(1):157-168.

66. Verner Z, Basu S, Benz C, Dixit S, Dobakova E, Faktorova D, Hashimi H, Horakova E, Huang Z, Paris Z *et al*: Malleable mitochondrion of Trypanosoma brucei. *Int Rev Cell Mol Biol* 2015, 315:73-151.

67. Lukes J, Guilbride DL, Votypka J, Zikova A, Benne R, Englund PT: Kinetoplast DNA network: evolution of an improbable structure. *Eukaryot Cell* 2002, 1(4):495-502.

68. Westenberger SJ, Cerqueira GC, El-Sayed NM, Zingales B, Campbell DA, Sturm NR: Trypanosoma cruzi mitochondrial maxicircles display species- and strain-specific

variation and a conserved element in the non-coding region. *BMC Genomics* 2006, 7:60.

69. Rauch CA, Perez-Morga D, Cozzarelli NR, Englund PT: The absence of supercoiling in kinetoplast DNA minicircles. *EMBO J* 1993, 12(2):403-411.

70. Corell RA, Feagin JE, Riley GR, Strickland T, Guderian JA, Myler PJ, Stuart K: Trypanosoma brucei minicircles encode multiple guide RNAs which can direct editing of extensively overlapping sequences. *Nucleic Acids Res* 1993, 21(18):4313-4320.

71. Ntambi JM, Shapiro TA, Ryan KA, Englund PT: Ribonucleotides associated with a gap in newly replicated kinetoplast DNA minicircles from Trypanosoma equiperdum. *J Biol Chem* 1986, 261(25):11890-11895.

72. Ray DS: Conserved sequence blocks in kinetoplast minicircles from diverse species of trypanosomes. *Mol Cell Biol* 1989, 9(3):1365-1367.

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

74. Aphasizhev R, Aphasizheva I: Mitochondrial RNA editing in trypanosomes: small RNAs in control. *Biochimie* 2014, 100:125-131.

75. Lukes J, Archibald JM, Keeling PJ, Doolittle WF, Gray MW: How a neutral evolutionary ratchet can build cellular complexity. *IUBMB Life* 2011, 63(7):528-537.

76. Gott JM, Emeson RB: Functions and mechanisms of RNA editing. *Annu Rev Genet* 2000, 34:499-531.

77. Hajduk S, Ochsenreiter T: RNA editing in kinetoplastids. *RNA Biol* 2010, 7(2):229-236.

78. Blum B, Simpson L: Guide RNAs in kinetoplastid mitochondria have a nonencoded 3' oligo(U) tail involved in recognition of the preedited region. *Cell* 1990, 62(2):391-397.

79. Blum B, Bakalara N, Simpson L: A model for RNA editing in kinetoplastid mitochondria: "guide" RNA molecules transcribed from maxicircle DNA provide the edited information. *Cell* 1990, 60(2):189-198.

80. Cross GA: Identification, purification and properties of clone-specific glycoprotein antigens constituting the surface coat of Trypanosoma brucei. *Parasitology* 1975, 71(3):393-417.

81. Vickerman K: On the surface coat and flagellar adhesion in trypanosomes. *J Cell Sci* 1969, 5(1):163-193.

82. Reinitz DM, Aizenstein BD, Mansfield JM: Variable and conserved structural elements of trypanosome variant surface glycoproteins. *Mol Biochem Parasitol* 1992, 51(1):119-132.

83. Manna PT, Boehm C, Leung KF, Natesan SK, Field MC: Life and times: synthesis, trafficking, and evolution of VSG. *Trends Parasitol* 2014, 30(5):251-258.

84. Vanhamme L, Pays E, McCulloch R, Barry JD: An update on antigenic variation in African trypanosomes. *Trends Parasitol* 2001, 17(7):338-343.

85. Vassella E, Reuner B, Yutzy B, Boshart M: Differentiation of African trypanosomes is controlled by a density sensing mechanism which signals cell cycle arrest via the cAMP pathway. *J Cell Sci* 1997, 110 (Pt 21):2661-2671.

86. Metcalf P, Blum M, Freymann D, Turner M, Wiley DC: Two variant surface glycoproteins of Trypanosoma brucei of different sequence classes have similar 6 A resolution X-ray structures. *Nature* 1987, 325(6099):84-86.

87. Leeflang P, Buys J, Blotkamp C: Studies on Trypanosoma vivax: infectivity and serial maintenance of natural bovine isolates in mice. *Int J Parasitol* 1976, 6(5):413-417.

88. D'Archivio S, Medina M, Cosson A, Chamond N, Rotureau B, Minoprio P, Goyard S: Genetic engineering of Trypanosoma (Dutonella) vivax and in vitro differentiation under axenic conditions. *PLoS Negl Trop Dis* 2011, 5(12):e1461.

89. **Alvarez F, Cortinas MN, Musto H: The analysis of protein coding genes suggests monophyly of Trypanosoma. *Mol Phylogenet Evol* 1996, 5(2):333-343.**
90. **Simpson AG, Stevens JR, Lukes J: The evolution and diversity of kinetoplastid flagellates. *Trends Parasitol* 2006, 22(4):168-174.**
91. **Hoare C: The trypanosomes of mammals: a zoological monograph. : Blackwell Scientific Publications. Oxford and Edinburgh.; 1972.**
92. **Haag J, O'HUigin C, Overath P: The molecular phylogeny of trypanosomes: evidence for an early divergence of the Salivaria. *Mol Biochem Parasitol* 1998, 91(1):37-49.**
93. **Lukes J, Skalicky T, Tyc J, Votypka J, Yurchenko V: Evolution of parasitism in kinetoplastid flagellates. *Mol Biochem Parasitol* 2014, 195(2):115-122.**
94. **Hoare CA: The Trypanosomes of Mammals. *A Zoological Monograph* 1972, Blackwell Scientific Publications, Oxford, UK, p. 749.**
95. **Jones TW, Davila AM: Trypanosoma vivax--out of Africa. *Trends Parasitol* 2001, 17(2):99-101.**
96. **Gardiner PR, Mahmoud, M.M.: Salivarian trypanosomes causing disease in livestock outside sub-saharan Africa. *Baker, JR (Eds), Parasitic Protozoa Academic Press, New York, USA, pp 1–68* 1990, 3.**
97. **Leger M, Vienne, M.: Epizootie a trypanosomee ches les bovides de la Guyane franÇaise. *Bull Soc Path Exot,* 1919, 12.**
98. **Roubaud E, Provost, A.: Sensibilite de lapin au trypanosome des ruminants de Antilles Tr. viennei souche americaine de Tr. cazalboui (vivax). *Bull Soc Path Exot* 1939, 32(5):6.**
99. **Silva ASMC, M.;Flores PolenzII, M.; Polenz, C.; Geraldes Teixeira, M.M.; Dos Anjos Lopes, S.; Gonzalez Monteiro, S.: Primeiro registro de Trypanosoma vivax em bovinos no Estado do Rio Grande do Sul, Brasil. *Brasil Cienc Rural* 2009, 39(8):4.**
100. **Gonzalez LE, Garcia JA, Nunez C, Perrone TM, Gonzalez-Baradat B, Gonzatti MI, Reyna-Bello A: Trypanosoma vivax: a novel method for purification from experimentally infected sheep blood. *Exp Parasitol* 2005, 111(2):126-129.**
101. **Chamond N, Cosson A, Blom-Potar MC, Jouvion G, D'Archivio S, Medina M, Droin-Bergere S, Huerre M, Goyard S, Minoprio P: Trypanosoma vivax infections: pushing ahead with mouse models for the study of Nagana. I. Parasitological, hematological and pathological parameters. *PLoS Negl Trop Dis* 2010, 4(8):e792.**
102. **Greif G, Rodriguez M, Reyna-Bello A, Robello C, Alvarez-Valin F: Kinetoplast adaptations in American strains from Trypanosoma vivax. *Mutat Res* 2015, 773:69-82.**
103. **Greif G, Ponce de Leon M, Lamolle G, Rodriguez M, Pineyro D, Tavares-Marques LM, Reyna-Bello A, Robello C, Alvarez-Valin F: Transcriptome analysis of the bloodstream stage from the parasite Trypanosoma vivax. *BMC Genomics* 2013, 14:149.**
104. **DeLuca DS, Levin JZ, Sivachenko A, Fennell T, Nazaire MD, Williams C, Reich M, Winckler W, Getz G: RNA-SeQC: RNA-seq metrics for quality control and process optimization. *Bioinformatics* 2012, 28(11):1530-1532.**
105. **Chevreux B, Pfisterer T, Drescher B, Driesel AJ, Muller WE, Wetter T, Suhai S: Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. *Genome Res* 2004, 14(6):1147-1159.**
106. **Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z *et al*: Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 2005, 437(7057):376-380.**
107. **Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: Basic local alignment search tool. *J Mol Biol* 1990, 215(3):403-410.**

108. **Langmead B, Trapnell C, Pop M, Salzberg SL: Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 2009, 10(3):R25.**

109. **Langmead B, Salzberg SL: Fast gapped-read alignment with Bowtie 2. *Nat Methods* 2012, 9(4):357-359.**

110. **Iseli C, Jongeneel CV, Bucher P: ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. *Proc Int Conf Intell Syst Mol Biol* 1999:138-148.**

111. **Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, Lopez R: InterProScan: protein domains identifier. *Nucleic Acids Res* 2005, 33(Web Server issue):W116-120.**

112. **Otto TD, Guimaraes AC, Degrave WM, de Miranda AB: AnEnPi: identification and annotation of analogous enzymes. *BMC Bioinformatics* 2008, 9:544.**

113. **Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M: Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 2005, 21(18):3674-3676.**

114. **Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B: Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* 2008, 5(7):621-628.**

115. **Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I: ABySS: a parallel assembler for short read sequence data. *Genome Res* 2009, 19(6):1117-1123.**

116. **Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD *et al*: SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 2012, 19(5):455-477.**

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

118. **Gilles A, Meglecz E, Pech N, Ferreira S, Malausa T, Martin JF: Accuracy and quality assessment of 454 GS-FLX Titanium pyrosequencing. *BMC Genomics* 2011, 12:245.**

119. **Jackson AP, Allison HC, Barry JD, Field MC, Hertz-Fowler C, Berriman M: A cell-surface phylome for African trypanosomes. *PLoS Negl Trop Dis* 2013, 7(3):e2121.**

120. **Gardiner PR, Nene V, Barry MM, Thatthi R, Burleigh B, Clarke MW: Characterization of a small variable surface glycoprotein from Trypanosoma vivax. *Mol Biochem Parasitol* 1996, 82(1):1-11.**

121. **Hutchinson OC, Picozzi K, Jones NG, Mott H, Sharma R, Welburn SC, Carrington M: Variant Surface Glycoprotein gene repertoires in Trypanosoma brucei have diverged to become strain-specific. *BMC Genomics* 2007, 8:234.**

122. **Cortez AP, Ventura RM, Rodrigues AC, Batista JS, Paiva F, Anez N, Machado RZ, Gibson WC, Teixeira MM: The taxonomic and phylogenetic relationships of Trypanosoma vivax from South America and Africa. *Parasitology* 2006, 133(Pt 2):159-169.**

123. **Kolev NG, Franklin JB, Carmi S, Shi H, Michaeli S, Tschudi C: The transcriptome of the human pathogen Trypanosoma brucei at single-nucleotide resolution. *PLoS Pathog* 2010, 6(9):e1001090.**

124. **Vickerman K: Biology of the kinetoplastida. *London, Academic* 1976.**

125. **Li SW, Feng L, Niu DK: Selection for the miniaturization of highly expressed genes. *Biochem Biophys Res Commun* 2007, 360(3):586-592.**

126. **Huang G, Vercesi AE, Docampo R: Essential regulation of cell bioenergetics in Trypanosoma brucei by the mitochondrial calcium uniporter. *Nat Commun* 2013, 4:2865.**

127. Dean S, Gould MK, Dewar CE, Schnaufer AC: Single point mutations in ATP synthase compensate for mitochondrial genome loss in trypanosomes. *Proc Natl Acad Sci U S A* 2013, 110(36):14741-14746.

128. Cristodero M, Seebeck T, Schneider A: Mitochondrial translation is essential in bloodstream forms of Trypanosoma brucei. *Mol Microbiol* 2010, 78(3):757-769.

129. Stephens JL, Lee SH, Paul KS, Englund PT: Mitochondrial fatty acid synthesis in Trypanosoma brucei. *J Biol Chem* 2007, 282(7):4427-4436.

130. Lai DH, Hashimi H, Lun ZR, Ayala FJ, Lukes J: Adaptations of Trypanosoma brucei to gradual loss of kinetoplast DNA: Trypanosoma equiperdum and Trypanosoma evansi are petite mutants of T. brucei. *Proc Natl Acad Sci U S A* 2008, 105(6):1999-2004.

131. Jensen RE, Simpson L, Englund PT: What happens when Trypanosoma brucei leaves Africa. *Trends Parasitol* 2008, 24(10):428-431.

132. Lun ZR, Lai DH, Li FJ, Lukes J, Ayala FJ: Trypanosoma brucei: two steps to spread out from Africa. *Trends Parasitol* 2010, 26(9):424-427.

Anexo 1. Material suplementario: "Transcriptome analysis of the bloodstream stage from the parasite *Trypanosoma vivax*"

A continuación se presenta el material suplementario del artículo "Transcriptome analysis of the bloodstream stage from the parasite *Trypanosoma vivax*". El mismo se encuentra disponible para descargar online en la siguiente web:
http://www.biomedcentral.com/1471-2164/14/149/additional

Tabla Suplementaria 1. Detalles de secuenciaciones

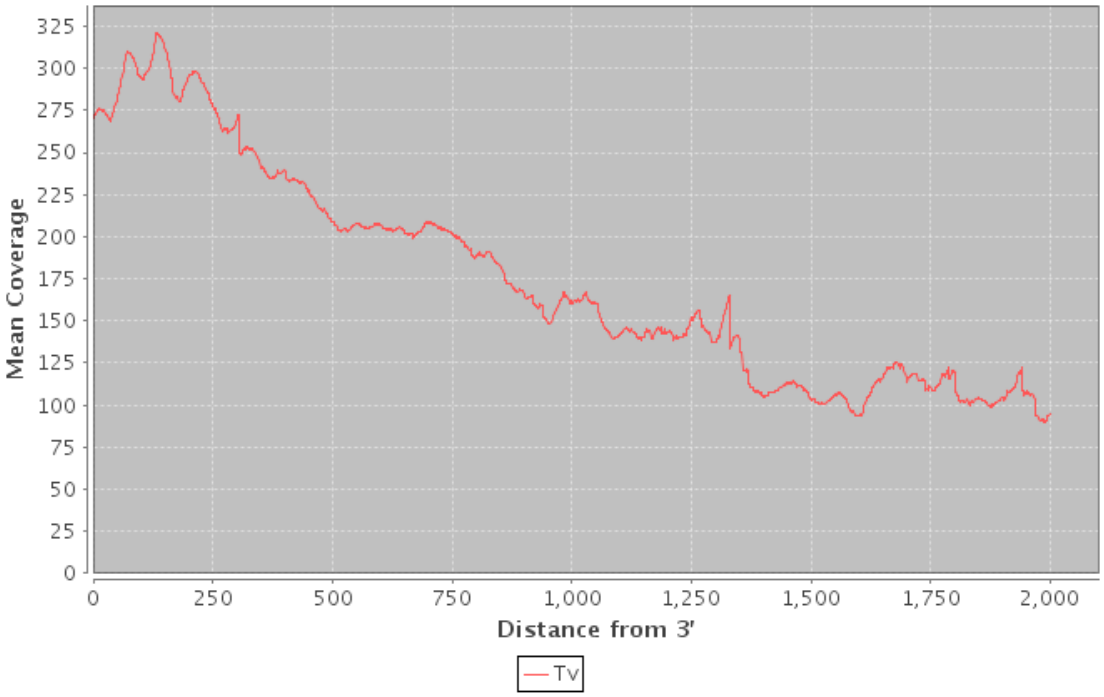
Details of sequence data obtained from 454 FLX and Illumina.

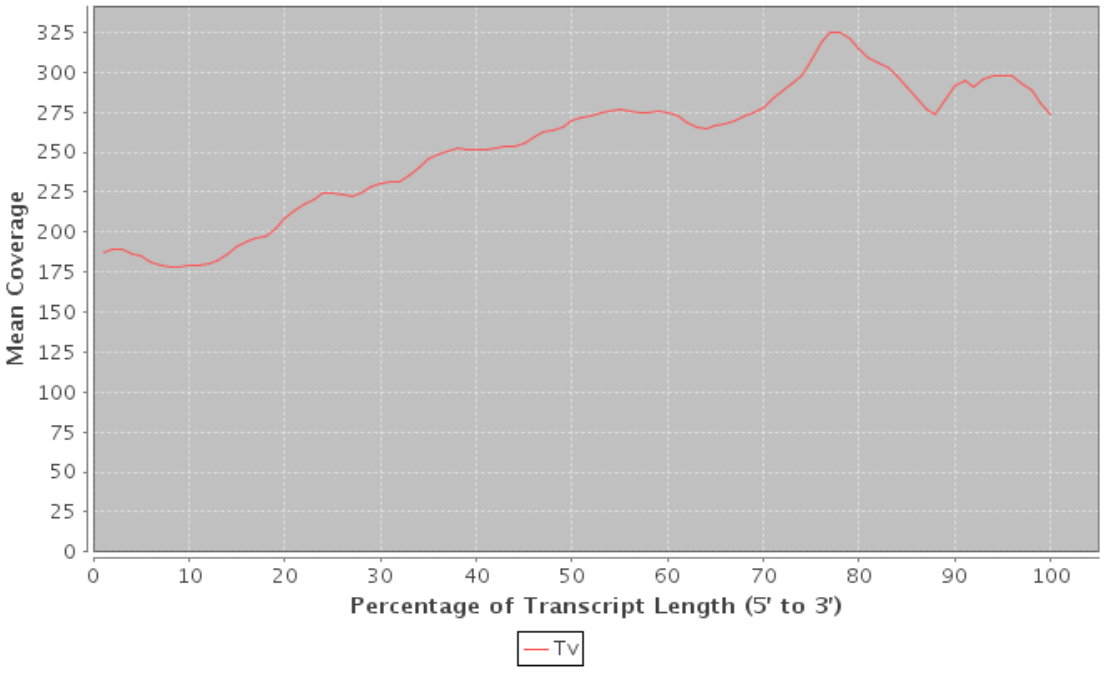
	454 FLX	**Illumina**
Number of Reads	187128	37406418
Average Read Length	289	36
Number of Reads after eliminating low quality	-----	34128677
Artificially Repeated Reads	15000	-----
Host contamination	445 (0.20%)	166000 (0.48%)
Reads corresponding to Ribosomal RNA	8385 (4.48%)	2035269 (5.95%)
Reads correspoding Maxicircles	8587 (4.58%)	778146 (2.28%)
Reads with SL	3022 (1.61%)	171239 (0.5%)

Figura Suplementaria 1. Control de Calidad de datos de secuenciación

Illumina Coverage Metrics for Top 1000 Expressed Transcripts

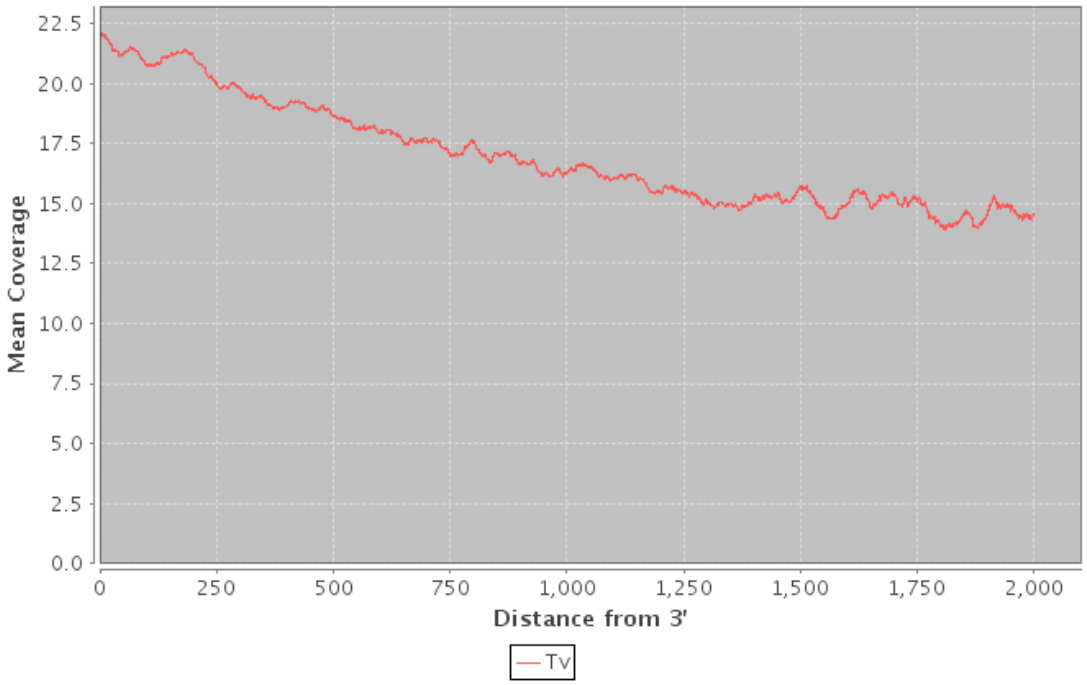
Sample	Note	Mean Per Base Cov.	Mean CV	No. Covered 5'	5'200Base Norm	No. Covered 3'	3' 200Base Norm	Num. Gaps	Cumul. Gap Length	Gap %
Tv	RNAseq	216.25	0.52	982	0.75	986	787	46707	4.9	

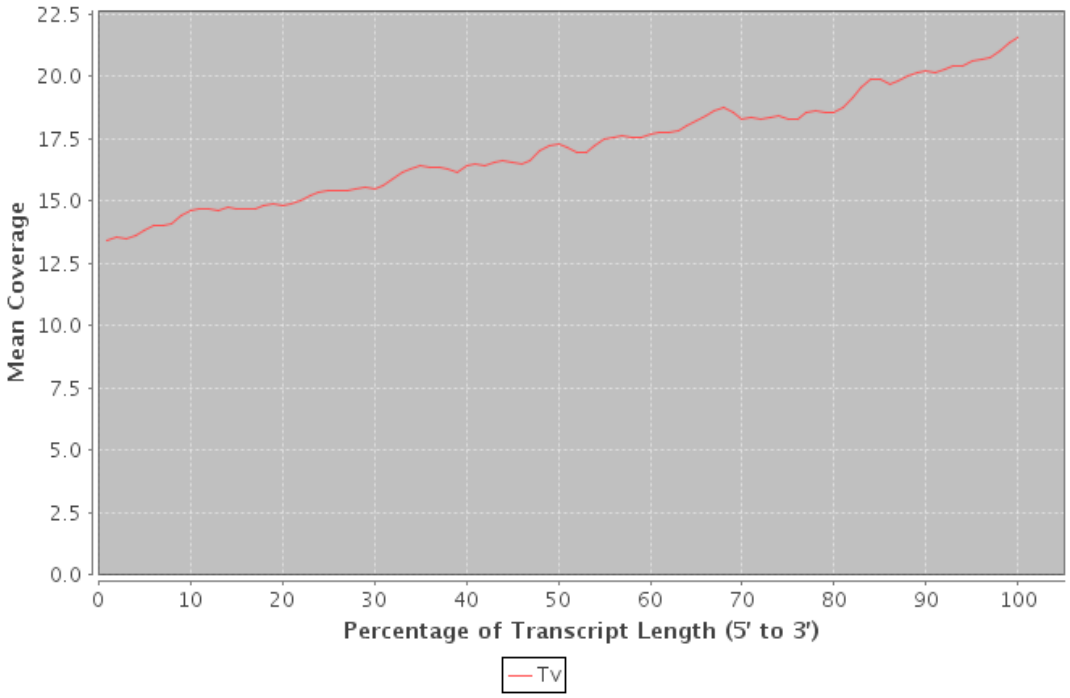




Illumina Coverage Metrics for Middle 1000 Expressed Transcripts

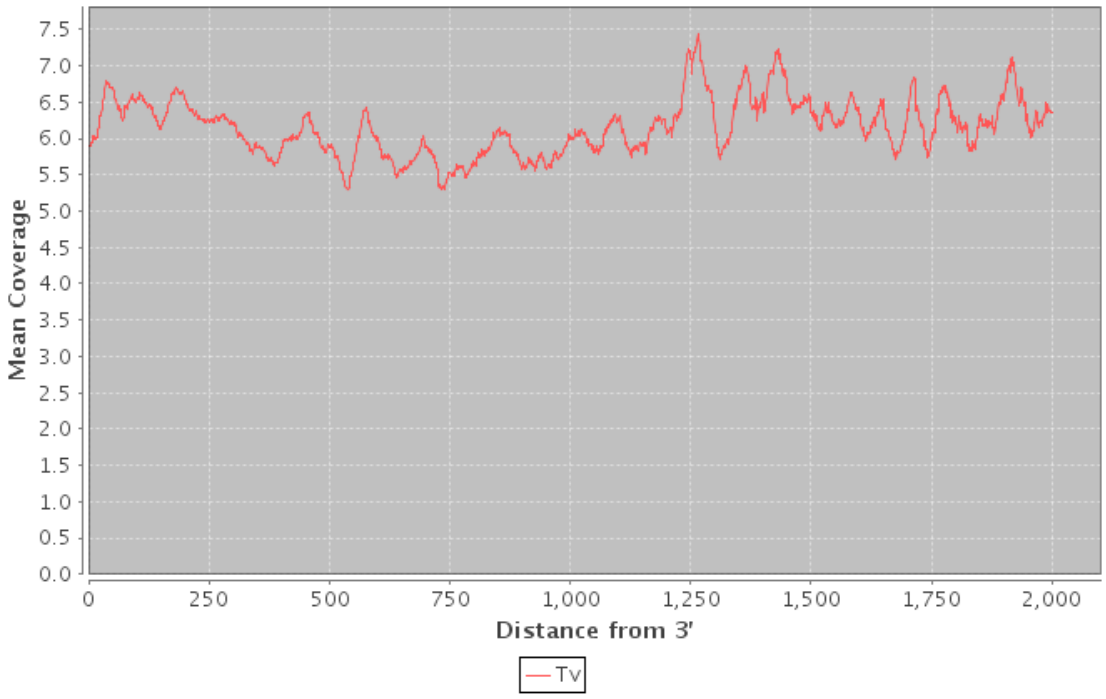
Sample	Note	Mean Per Base Cov.	Mean CV	No. Covered 5'	5'200Base Norm	No. Covered 3'	3' 200Base Norm	Num. Gaps	Cumul. Gap Length	Gap %
Tv		17.28	0.42	997	0.76	999	387	11599	0.8	

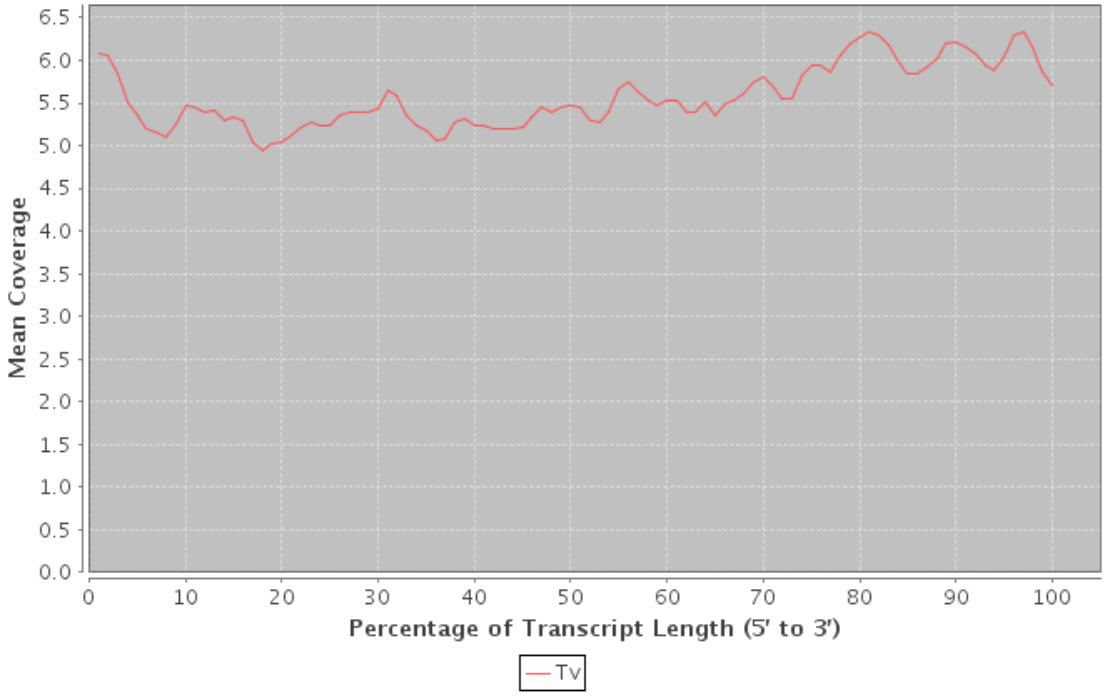




Illumina coverage Metrics for Bottom 1000 Expressed Transcripts

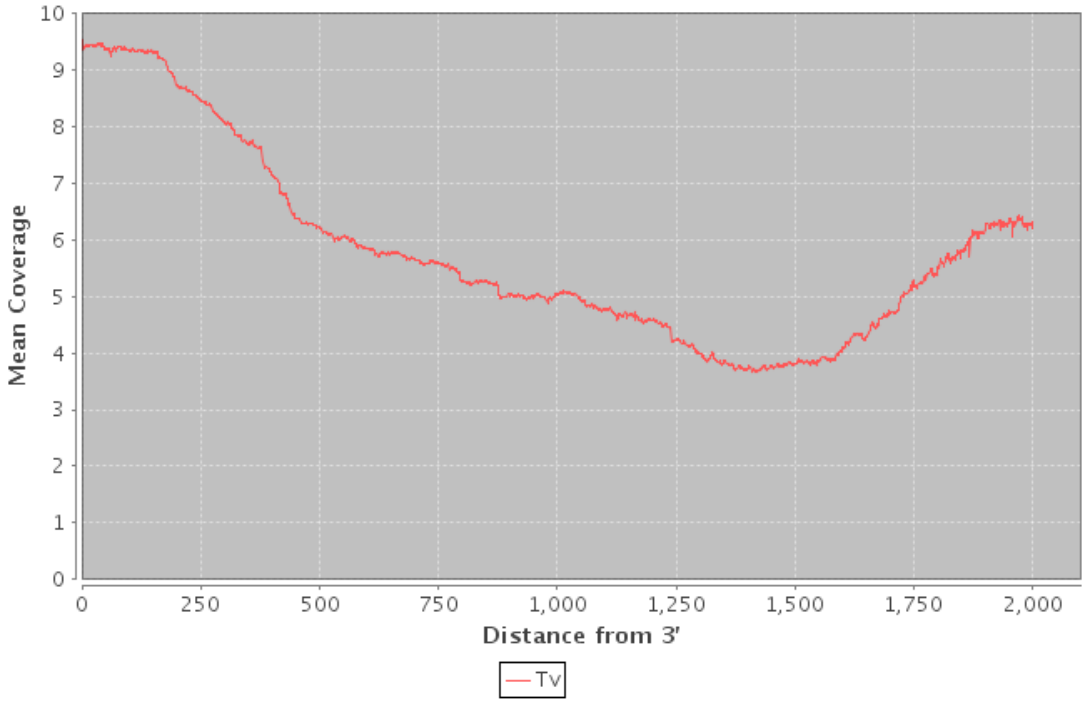
Sample	Note	Mean Per Base Cov.	Mean CV	No. Covered 5'	5'200Base Norm	No. Covered 3'	3' 200Base Norm	Num. Gaps	Cumul. Gap Length	Gap %
Tv	RNAseq	216.25	0.52	982	0.75	986	787	46707	4.9	

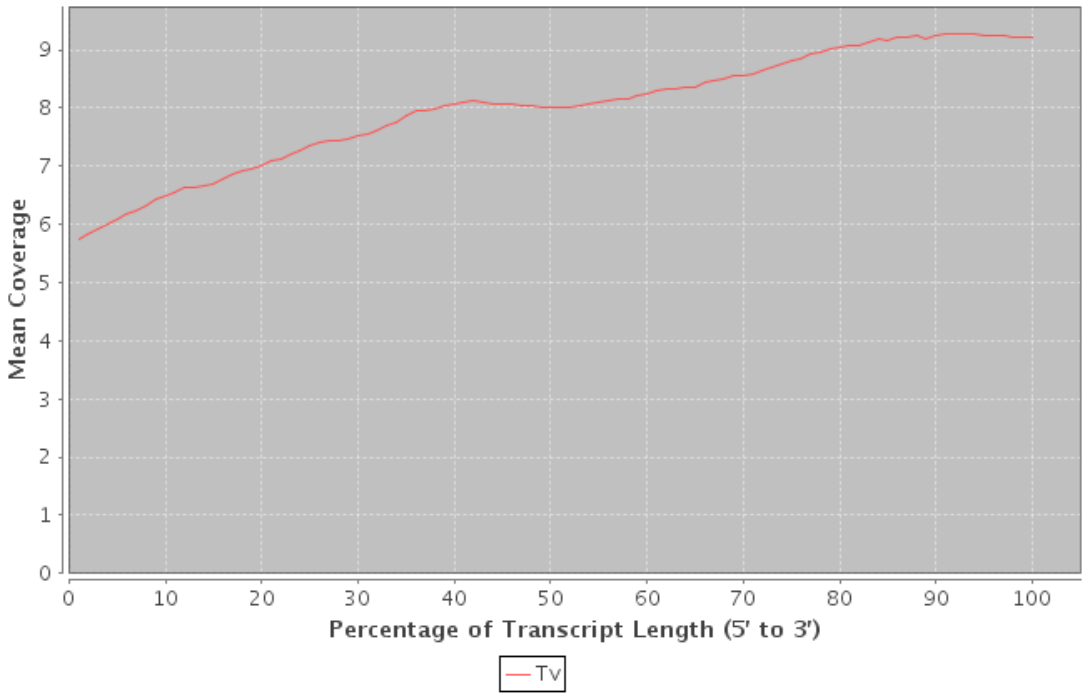




454 FLX Coverage Metrics for Top 1000 Expressed Transcripts

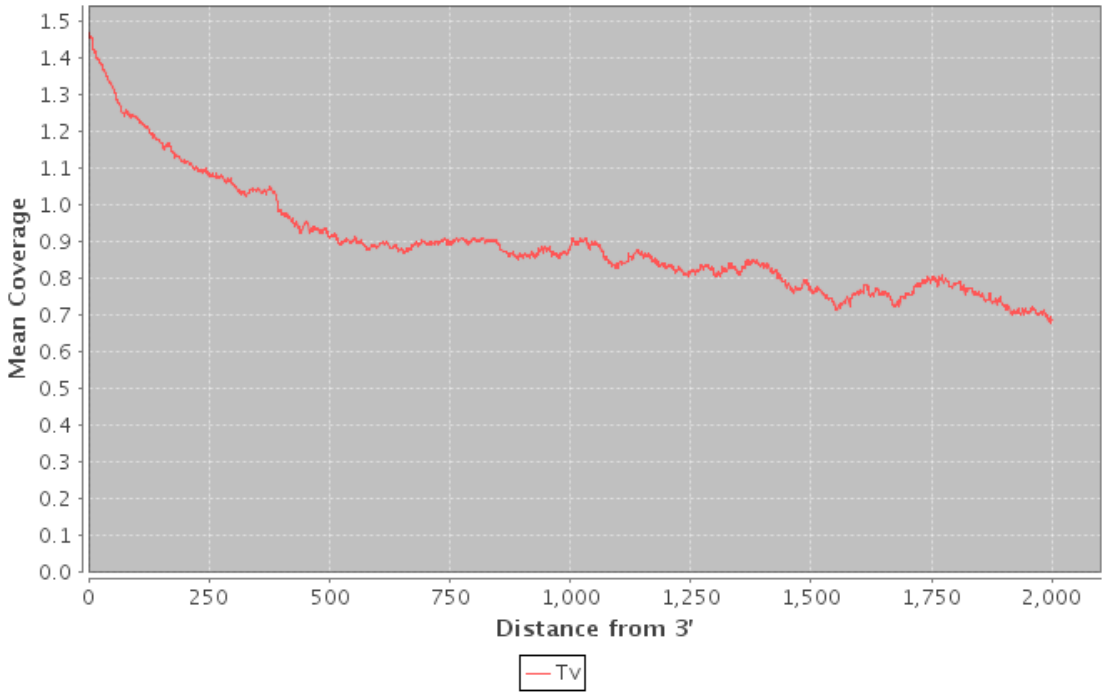
Sample	Note	Mean Per Base Cov.	Mean CV	No. Covered 5'	5'200Base Norm	No. Covered 3'	3' 200Base Norm	Num. Gaps	Cumul. Gap Length	Gap %
Tv	RNAseq	6.83	0.62	806	0.72	856	661	120692	13.2	

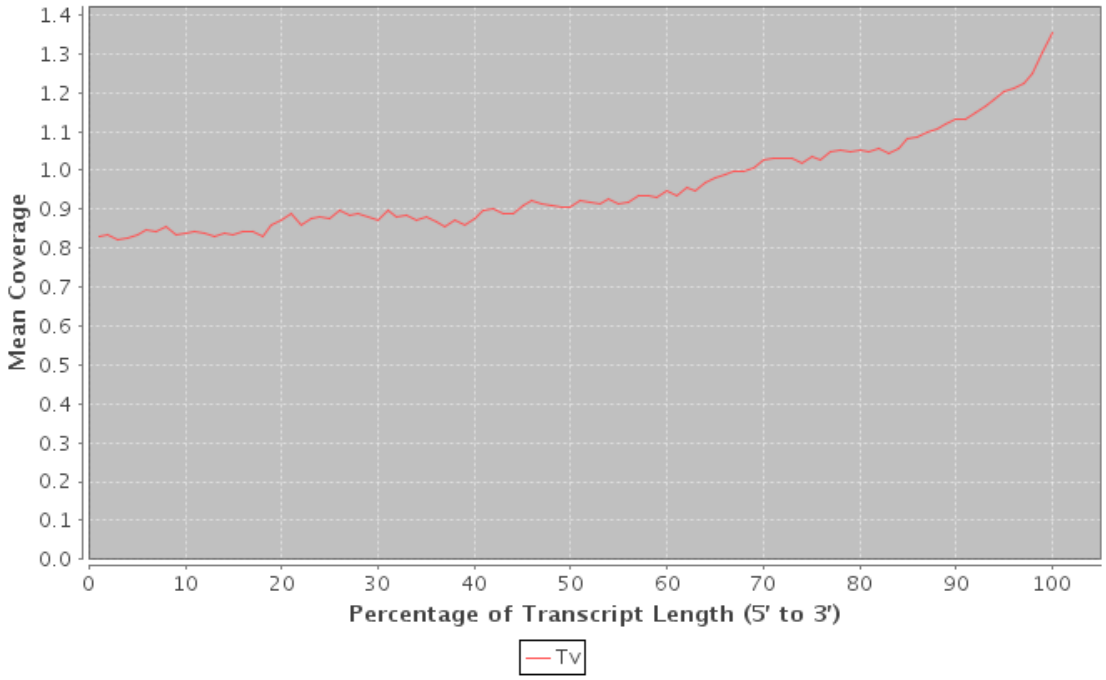




454 FLX Coverage Metrics for Middle 1000 Expressed Transcripts

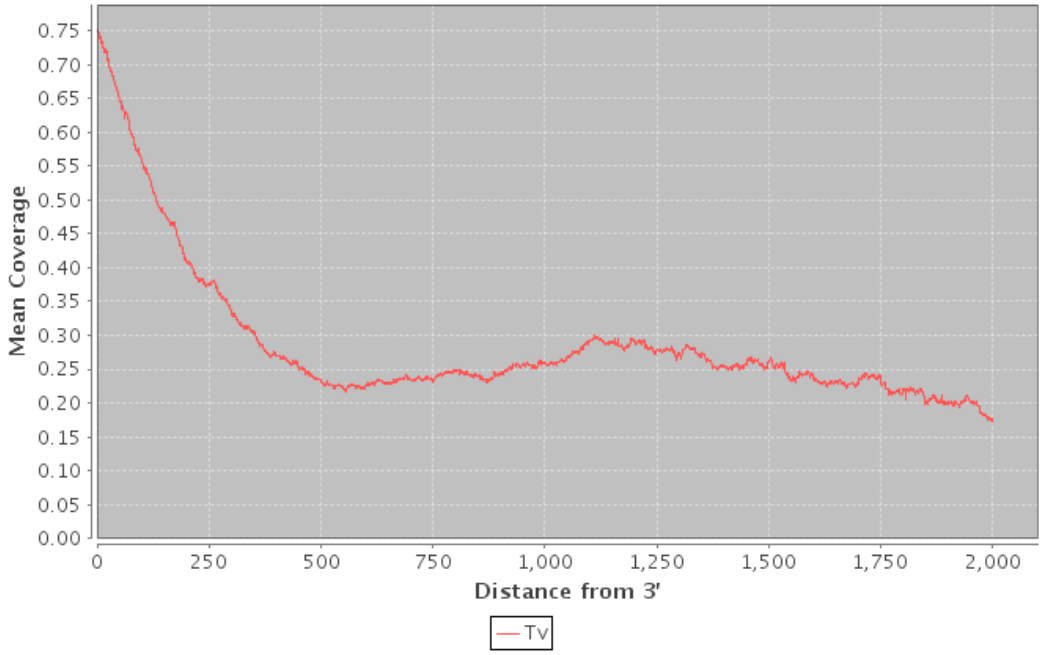
Sample	Note	Mean Per Base Cov.	Mean CV	No. Covered 5'	5'200Base Norm	No. Covered 3'	3' 200Base Norm	Num. Gaps	Cumul. Gap Length	Gap %
Tv	RNAseq	0.90	1.05	418	0.84	528	2016	691519	45.0	





454 FLX Coverage Metrics for Bottom 1000 Expressed Transcripts

Sample	Note	Mean Per Base Cov.	Mean CV	No. Covered 5'	5'200Base Norm	No. Covered 3'	3' 200Base Norm	Num. Gaps	Cumul. Gap Length	Gap %
Tv	RNAseq	0.28	2.15	200	1.47	311	1977	1649691	76.6	



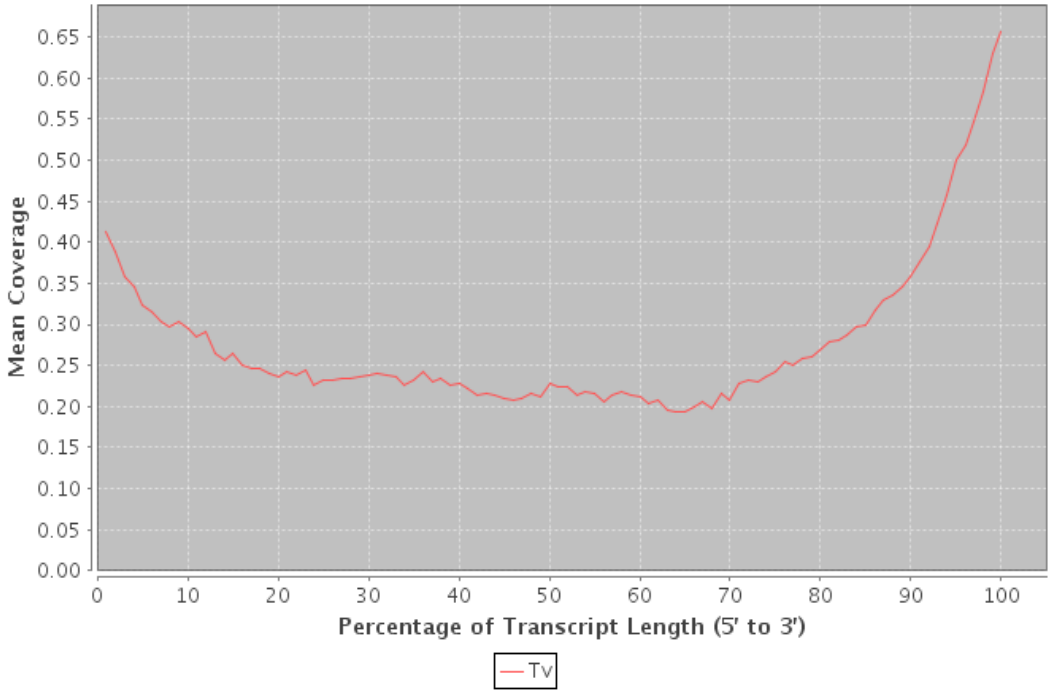


Tabla Suplementaria 2. Calidad de Ensamblajes

Calidad de ensamblaje. Las tablas muestran los valores Q (ver texto de manuscrito) obtenidos con los ensambladores Mira y Newbler.

bestHitsStats	**mira**			**454 Newbler**		
Total of genes	684	% total	**Q mira**	684	% total	**Q 454**
Not Found	51	**7,5**		64	9,4	
Coverage						
10%	54	**7,9**	5,4	56	8,2	5,6
20%	135	**19,7**	27	156	22,8	31,2
30%	97	**14,2**	29,1	116	17,0	34,8
40%	92	**13,5**	36,8	89	13,0	35,6
50%	60	**8,8**	30	50	7,3	25
60%	44	**6,4**	26,4	34	5,0	20,4
70%	15	**2,2**	10,5	22	3,2	15,4
80%	15	**2,2**	12	11	1,6	8,8
90%	22	**3,2**	19,8	17	2,5	15,3
100%	99	**14,5**	99	69	10,1	69
Total found	633	**92,5**	**296**	620	90,6	**261,1**

Tabla suplementaria 3. Cuantificación de transcriptos (Erange)

Niveles de expresión obtenidos con Erange (se muestran los valores para los primeros 35 genes). La tabla completa se encuentra disponible online.

gene ID	length_kb	454 (Blast) Read Count	Illumina (Blast) Read count	rpkm Erange (Bowtie)
TvY486_0000010	1,666	0	0	0
TvY486_0000020	1,327	188	40312	66,02
TvY486_0000030	0,804	24	11592	47,23
TvY486_0000040	0,773	0	0	0
TvY486_0000050	1,326	3	478	16,02
TvY486_0000060	1,27	137	24454	7,99
TvY486_0000070	1,293	3	1761	1,67
TvY486_0000080	1,239	4	1328	7,53
TvY486_0000090	1,926	106	6912	20,78
TvY486_0000120	1,476	0	0	0
TvY486_0000130	0,976	9	701	1,85
TvY486_0000140	1,846	5	612	16,4
TvY486_0000150	4,015	4	1635	19,8
TvY486_0000160	0,905	2	226	11,14
TvY486_0000170	1,017	13	3519	52,48
TvY486_0000180	1,397	0	0	0
TvY486_0000190	1,903	96	12527	51,18
TvY486_0000200	1,348	68	11848	17,53
TvY486_0000210	1,83	39	6985	50,65
TvY486_0000220	0,979	0	0	0
TvY486_0000230	0,979	28	3500	8,4
TvY486_0000240	1,684	70	12688	269,34
TvY486_0000250	2,386	119	19363	3,52
TvY486_0000260	1,531	29	8128	22,23
TvY486_0000270	2,226	6	726	10,76
TvY486_0000280	0,823	11	413	13,09
TvY486_0000300	1,251	0	0	0
TvY486_0000310	1,369	0	0	0
TvY486_0000320	1,263	0	62	0
TvY486_0000330	1,518	0	0	0
TvY486_0000350	1,518	0	0	0
TvY486_0000360	1,434	0	0	0
TvY486_0000370	1,44	20	2179	33,88
TvY486_0000380	1,632	35	1610	3
TvY486_0000390	1,074	39	11069	39,04

Tabla suplementaria 4. Análisis de proteínas especie-específicas

Listado de proteínas especie específicas (n=570, se listan las primeras 35 filas de la tabla).

Contig Name	Predicted TMH	GPI	Signal P	Presence in Y486 genome (no/yes)
TvMiraNov_c1477	0	No	Yes	yes
TvMiraNov_c320	0	No	No	yes
TvMiraNov_c668	0	No	Yes	yes
TvMiraNov_c878	0	No	No	yes
TvMiraNov_c1066	0	No	No	yes
TvMiraNov_c711	0	No	No	yes
TvMiraNov_c555	0	No	No	yes
TvMiraNov_c3860	0	No	No	yes
TvMiraNov_c438	0	No	No	yes
TvMiraNov_c5987	0	No	No	yes
TvMiraNov_c2447	0	No	No	yes
TvMiraNov_c1210	2	No	Yes	yes
TvMiraNov_c2889	0	No	No	yes
TvMiraNov_c5442	0	No	No	yes
TvMiraNov_c357	2	No	Yes	yes
TvMiraNov_c6880	0	No	No	yes
TvMiraNov_c2411	0	No	No	yes
TvMiraNov_c1353	0	No	No	yes
TvMiraNov_c2089	0	No	No	yes
TvMiraNov_c5591	0	No	No	yes
TvMiraNov_c2031	0	No	Yes	yes
TvMiraNov_c331	1	No	No	yes
TvMiraNov_c2058	0	No	No	yes
TvMiraNov_c1986	0	No	No	yes
TvMiraNov_c2161	0	No	No	yes
TvMiraNov_c6618	0	No	No	yes
TvMiraNov_c2337	0	No	No	yes
TvMiraNov_c1015	0	No	No	yes
TvMiraNov_c6670	0	No	No	yes
TvMiraNov_c1603	0	No	No	yes
TvMiraNov_c7913	0	No	No	yes
TvMiraNov_c1057_3	8	Yes	No	no
TvMiraNov_c1257	1	No	No	yes
TvMiraNov_c1231_4	6	No	No	yes

Figura Suplementaria 2. Comparación secuencia VSG Liem-176 e Y486

Alineamiento de secuencias de gen VSG de cepa americana (Liem-176) y africana (Y486).

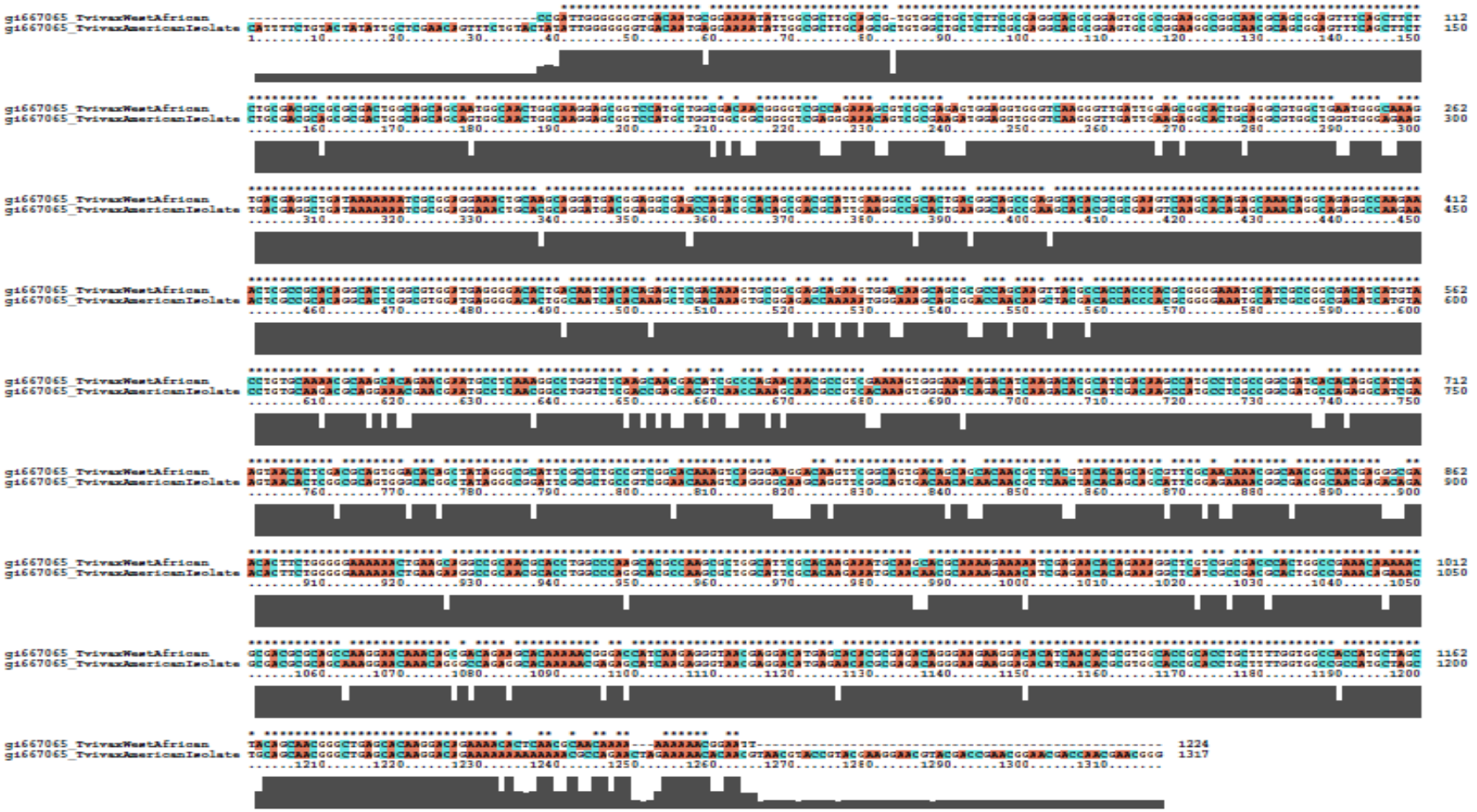


Figura Suplementaria 3. Comprobación expresión gen VSG

Amplificación por PCR de regiones genómicas con cebadores específicos para el gen VSG expresado en la cepa americana Liem-176.

Figure S3

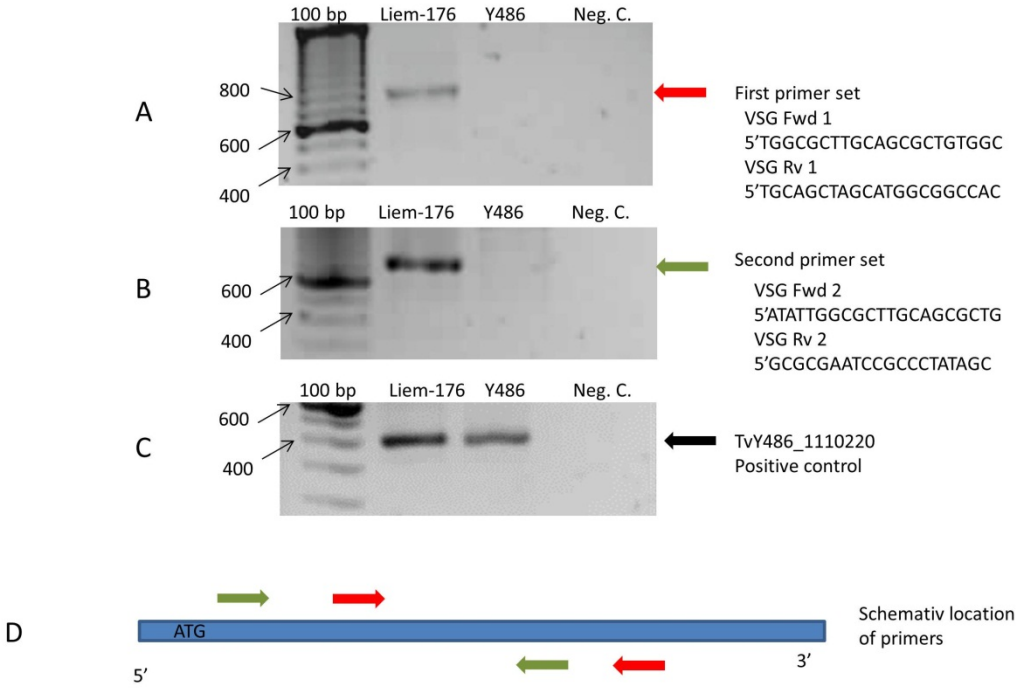


Figure S3. VSG genomic comparison. We used VSG specific primers (two sets) that confirm the presence of this gene in Liem-176 and its absence in Y486 genome (A and B). C. This gel contains a positive control corresponding to the amplification of a common region for both genomes. D. The box shows a schematical representation of the VSG gene and the localization of primers is depicted by green and red arrows.

Tabla suplementaria 5. Análisis de composición de membrana

rpkm y porcentaje total de secuencias correspondientes al gen VSG y los genes de tubulina de *T. vivax* y *T. brucei*.

Illumina reads	*T. vivax*	*%*	*T. brucei **	*%*
VSG	224379	0.70	510424	5
alpha tubulin	63766	0.18	26791	0.27
beta tubulin	112740	0.33	19837	0.20

The figures correspond to the number (and percentage) of reads that map in the corresponding CDS (indicated on the left most column).

* Data from RNAseq of *T. brucei* (Siegel et al, 2010).

Figura Suplementaria 4. Frecuencias G+C en genes de alta o baja expresión

Contenido GC3 discriminados para cada aminoácidos en genes alta y mínimamente expresados.

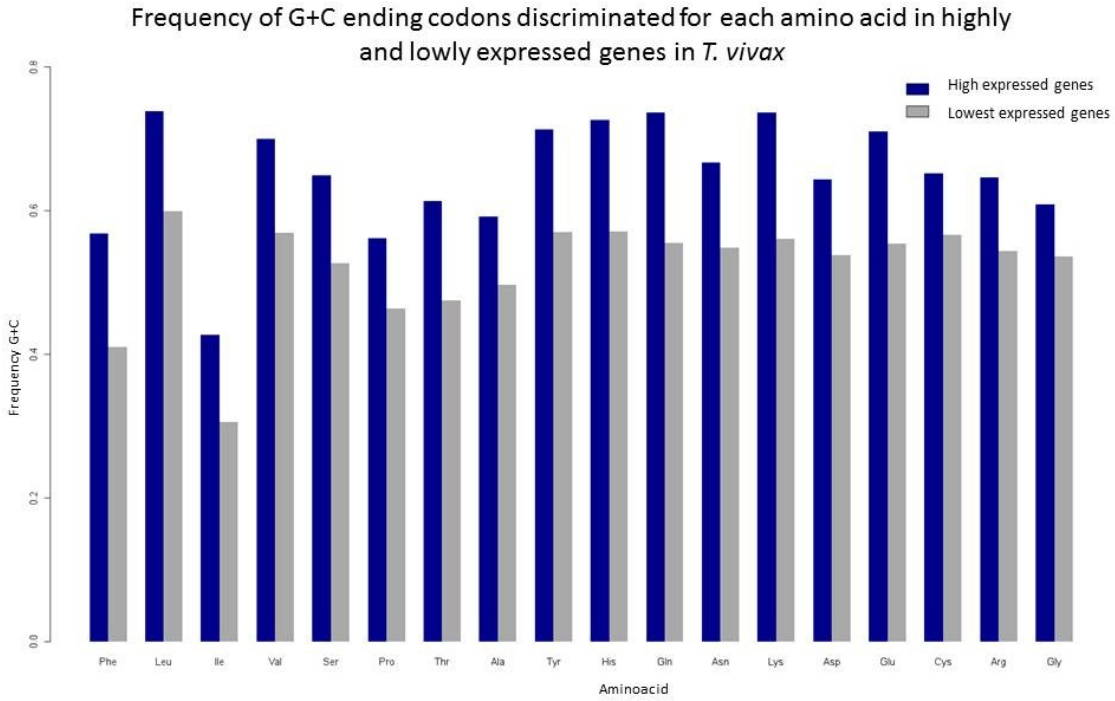


Figure S4.

Tabla suplementaria 6. Genes de alta expresión y baja frecuencia GC3

Grupos de genes con alta expresión (rpkm > promedio, 3SD) y baja frecuencia GC3 en *T. brucei* (pimera tabla) y *T.vivax* (segunda tabla).

Table S6

Gene ID	GC3	rpkm	Gene Product
Tb427.02.5910	0,56291391	282,456	40S ribosomal protein S13, putative
Tb427.04.1790	0,57380457	341,212	ribosomal protein L3, putative
Tb427.04.4620	0,56329114	101,751	cytochrome oxidase subunit VIII
Tb427.05.2160	0,57777778	311,111	hypothetical protein, conserved
Tb427.05.2170	0,55555556	318,691	hypothetical protein, conserved
Tb427.05.2200	0,57037037	319,037	hypothetical protein, conserved
Tb427.05.2230	0,57037037	319,037	hypothetical protein, conserved
Tb427.05.2260	0,57462687	336,469	hypothetical protein, conserved
Tb427.06.450	0,4047619	1024,505	
Tb427.06.480	0,40336134	1072,392	
Tb427.06.510	0,46086957	1657,188	GPEET2 procyclin precursor
Tb427.06.5120	0,46296296	150,864	60S acidic ribosomal protein P2, putative
Tb427.06.5130	0,46296296	150,864	60S acidic ribosomal protein P2, putative
Tb427.06.520	0,40769231	1256,256	EP3-2 procyclin
Tb427.07.2980	0,55497382	171,727	hypothetical protein, conserved
Tb427.07.3020	0,55497382	171,727	hypothetical protein, conserved
Tb427.08.5440	0,54794521	104,961	flagellar calcium-binding protein
Tb427.08.5465	0,55045872	108,828	flagellar calcium-binding protein
Tb427.08.5470	0,54506438	101,139	flagellar calcium-binding protein
Tb427.10.10250	0,37209302	1153,948	EP2 procyclin
Tb427.10.10260	0,35251799	1122,981	EP1 procyclin
Tb427.10.10280	0,41574074	625,309	microtubule-associated protein, putative
Tb427.10.10360	0,53009961	151,285	microtubule-associated protein, putative
Tb427.10.10450	0,45355191	301,056	hypothetical protein
Tb427.10.15120	0,56953642	279,21	40S ribosomal protein S13, putative
Tb427.10.3370	0,40869565	393,826	60S acidic ribosomal protein P2, putative
Tb427.10.3380	0,40869565	393,826	60S acidic ribosomal protein P2, putative
Tb427.10.7180	0,45510026	501,087	cysteine-rich, acidic integral membrane protein precursor
Tb427.10.8490	0,53396226	157,371	glucose transporter, putative
Tb427.10.8500	0,53497164	155,754	glucose transporter, putative
Tb427.10.8510	0,53396226	157,371	glucose transporter, putative
Tb427.10.8520	0,53396226	157,371	glucose transporter, putative
Tb427.10.8530	0,53396226	157,371	glucose transporter
Tb427.10.9080	0,51482059	120,535	pteridine transporter, putative
Tb427tmp.01.1680	0,57918552	171,603	polyubiquitin, putative
Tb427tmp.02.3760	0,4587156	201,131	hypothetical protein, conserved
Tb427tmp.02.3770	0,46099291	190,372	hypothetical protein, conserved
Tb427tmp.03.0410	0,54491018	132,634	eukaryotic translation initiation factor 5a, putative
Tb427tmp.160.4200	0,57017544	150,994	60S acidic ribosomal protein, putative
Tb427tmp.160.4250	0,53	102,216	tryparedoxin peroxidase
Tb427tmp.160.4280	0,53	102,216	tryparedoxin peroxidase
Tb427tmp.160.4560	0,53580247	107,901	arginine kinase
Tb427tmp.160.4570	0,57412399	110,601	arginine kinase
Tb427tmp.160.4590	0,56862745	113,454	arginine kinase
Tb427tmp.211.0120	0,53968254	172,38	nascent polypeptide associated complex subunit, putative
Tb427tmp.211.0130	0,53968254	172,38	nascent polypeptide associated complex subunit, putative
Tb427tmp.211.2730	0,53688525	162,527	Gim5A protein
Tb427tmp.211.2740	0,54545455	153,085	Gim5B protein
Tb427tmp.244.0810	0,55357143	1169,91	hypothetical protein
Tb427tmp.42.0005	0,56976744	50,055	histone H1, putative
Tb427tmp.42.0007	0,57142857	129,523	histone H1, putative

Table S6				
TvY486_0401510	0,589783282	1100,15		ribosomal protein L3, putative, (fragment)
TvY486_0401870	0,594835263	3818,83		antigenic protein, putative, (fragment)
TvY486_0401980	0,566666667	495,37		ribosomal protein L35A, putative
TvY486_0404360	0,592771084	255,68		calreticulin, putative
TvY486_0500710	0,575609756	139,42		60S ribosomal protein L2, putative, (fragment) 60S ribosomal protein L8, putative, (fragment)
TvY486_0501280	0,504587156	292,13		60S acidic ribosomal protein, putative
TvY486_0601720	0,515151515	389,72		4-methyl-5(beta-hydroxyethyl)-thiazole monophosphate synthesis protein, putative
TvY486_0604300	0,579310345	185,07		40S ribosomal protein S14
TvY486_0604430	0,592592593	337,12		60S acidic ribosomal protein P2, putative
TvY486_0604440	0,583333333	50,2		60S acidic ribosomal protein P2, putative
TvY486_0700070	0,5625	326,08		40S ribosomal protein S33, putative
TvY486_0700080	0,527607362	265,35		40S ribosomal protein S33, putative
TvY486_0903590	0,568253968	282,99		RNA-binding protein, putative
TvY486_0907470	0,564593301	245,82		ribosomal protein L15, putative
TvY486_0907480	0,559633028	221,27		ribosomal protein L36, putative
TvY486_1001600	0,559633028	650,31		ribosomal proteins L36, putative
TvY486_1007165	0,566037736	117,28		40S ribosomal protein S24E, putative, (fragment)
TvY486_1007240	0,586387435	287,08		succinyl-CoA ligase [GDP-forming] beta-chain, putative
TvY486_1008320	0,47260274	651,7		40S ribosomal protein S12, putative
TvY486_1012330	0,593023256	329,71		60S ribosomal protein L34, putative
TvY486_1013580	0,569514238	51,83		transporter, putative, (fragment)
TvY486_1013590	0,578787879	299,83		pyruvate kinase 1, putative, (fragment)
TvY486_1014490	0,564102564	234,29		40S ribosomal protein S13, putative
TvY486_1100080	0,571428571	411,25		DNA-directed RNA polymerase subunit, putative, (fragment)
TvY486_1100590	0,473053892	502,54		eukaryotic translation initiation factor 5a, putative
TvY486_1103320	0,588785047	484,42		60S ribosomal protein L44
TvY486_1104380	0,503448276	266,08		40S ribosomal protein S12, putative
TvY486_1108430	0,572413793	438,34		40S ribosomal protein S14, putative
TvY486_1110340	0,578431373	490,38		nascent polypeptide associated complex alpha subunit, putative
TvY486_1116280	0,445255474	260,25		Tb-291 membrane associated protein, putative, (fragment) membrane associated protein, putative, (fragment)
TvY486_1116970	0,532467532	462,11		nucleoside diphosphate kinase, putative
TvY486_1117190	0,552188552	306,67		60S ribosomal protein, putative

Figura Suplementaria 5. Comparación frecuencias G+C en tripanosomas africanos

Comparación de frecuencias G + C en la tercer posición del codón en genes más y menos expresados en *T. vivax* y *T. brucei*. La comparación fue realizada entre genes ortólogos conservados (arriba) y no conservados (debajo).

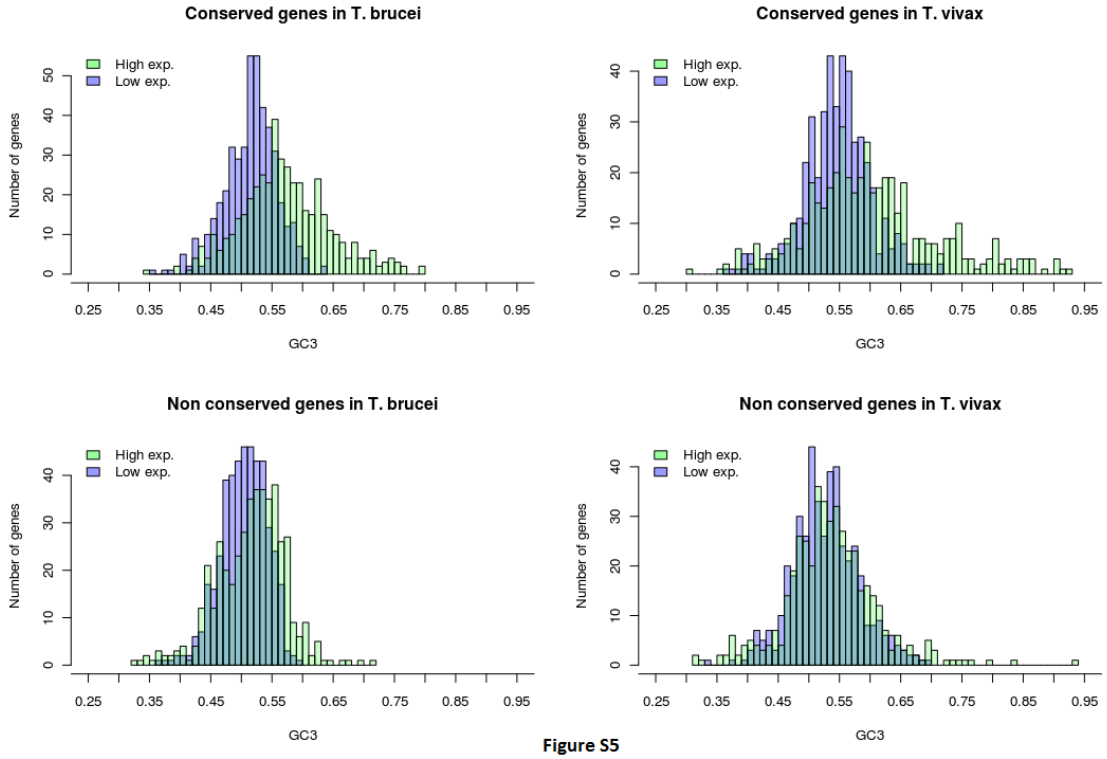


Figure S5

Figura Suplementaria 6. Sitios de *trans-splicing* en *T. vivax*

Representación de secuencia de los sitios 1 a 4 de *trans-splicing* en *T. vivax* (TSS: *trans-splicing site*).

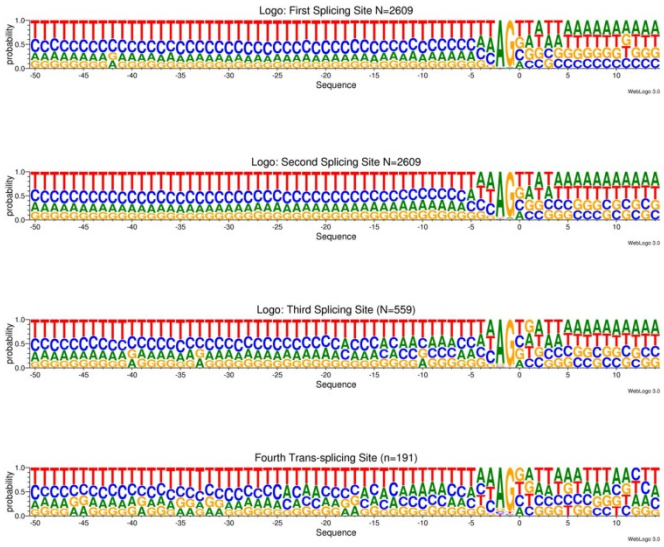


Figure S6. Logo representation of trans-splicing sites (-50 nt and +15 nt relative to splicing site). **A.** Logo representation of the major site. **B.** Logo of the second TSS. **C.** Logo of the third TSS. **D.** Logo of the fourth TSS.

Figura Suplementaria 7. Corrección de anotación genómica basados en *trans-splicing site*

Ejemplos de corrección de anotación utilizando información de sitios de *trans-splicing* en *T. vivax* y *T. brucei*.

Example 1:

Gene ID	Description	Reads	1st Splicing Site	Reads	2nd Splicing Site
Tb427.08.5810	mitochondrial carrier protein putative	67	31	10	34
TvY486_0805310	mitochondrial carrier protein putative	12	-23		

First alignment:

```
Tb427.08.5810          MHALVHFGNEGWLVSTCTWMTEEYVHTVVAGTISGAAGVLLEYPLDTIKVRLQMGGGRYT
TcIL3000.0.57240       -------------------MAEEYALTAVAGTVSGAAGVLLEYPLDTVKVRLQTLGTRYS
TvY486_0805310         -------------------MVPDVCHTLIAGIVSGVAGTVIEYPLDTLKVRLQTCGGRYS
                                          *. :   * :** :**.**.::******:*****  * **:
Green: Annotated start methionine
Yellow: proposed start methionine
```

```
Tb427.08.5810          AATGTTGCGCAGTGTGCCCACCCGATGCGTCACTTAGCTCCTCTTTTTCTATGCACGCCC
TcIL3000.0.57240       -----------------------------------------------------TTTTTTC
TvY486_0805310         ----------------------------------------------------GACCTTCTTT

Tb427.08.5810          TTGTACATTTTGGAAATGAAGGATGGTTAGTTTCCACCTGCACTTGGATGACTGAGGAGT
TcIL3000.0.57240       TTAAAAGCCCGGCAACTG----GCCCCTAGCTCACTGCACCATAGAAATGGCGGAGGAGT
TvY486_0805310         TTACCCACACACCAAC------GTAGACGCCCTGCA-CTGTGTTGCAATGGTGCCCGACG
                       **           **                  *  *         ***      **


Tb427.08.5810          ATGTCCACACGGTTGTAGCGGGTACAATATCTGGCGCGGCTGGTGTATTGCTAGAGTACC
TcIL3000.0.57240       ACGCGCTGACGGCTGTAGCTGGAACGGTATCAGGGGCAGCGGGCGTGTTGCTGGAATATC
TvY486_0805310         TTTGCCATACACTCATTGCGGGCATCGTTTCGGGCGTTGCAGGCACGGTGATTGAGTACC
                            *  **     * ** ** *   * ** ** *  ** **     ** * ** ** *
Green: annotated start codón
Yellow: Proposed start codon
Red: First trans splicing site
Orange: Second trans splicing site
```

Results after ATG correction:

Gene ID	Description	Reads	1st Splicing Site	Reads	2nd Splicing Site
Tb427.08.5810	mitochondrial carrier protein putative	67	-26	10	-23
TvY486_0805310	mitochondrial carrier protein putative	12	-23		

Alignment after correction:

```
Tb427.08.5810       mitochondrial carrier protein putative
Tb              --GAAATGAAGGATGGTTAGTTTCCACCTGCACTTGGATG 38
Tv              CACACACCAACG-TAGACGCCCTGCAC-TGTGTTGCAATG 38
                   * *  ** * * *      * *** **   *   ***

Green: annotated start codón
Red: First trans splicing site
Orange: Second trans splicing site
```

Example 2:

Gene ID	Description	Reads	1st Splicing Site	Reads	2nd Splicing Site	Reads/3rd Splicing Site	Reads/4th Splicing Site
Tb427.08.5800	Hypothetical protein conserved	47	-32	32	-2	5/1	6/-29
TvY486_0805300	Hypothetical protein conserved	5	9	--	--	--	--

First Alignment:

```
Tb427.08.5800         -----MKHKDARGGS-TPYFAITNNKTGEVLLEVAGPLPPTAASPP---PVEEECCGLFN
TcIL3000.0.57230      -----MKHKDGRGQTPPPLFSITNQKTGEVLLEIT-SVPPNNAPLP---PVEEECCGQFN
TvY486_0805300        MPSREMSKNEQNAEQLTPYFAITNQQTGAVLLEITSFGKEFAASTLDEMVVEEECRGLFK
                           *.::: ..   .* *:***::** ****::       *.      ***** * *:

Green: Annotated start methionine
Yellow: T. vivax proposed start methionine
```

```
Tb427.08.5800         --------------ATGAAACACAAAGATGCTCGCGGTGGTTCGACAC-CGTAC---TTT
TcIL3000.0.57230      --------------ATGAAGCACAAAGATGGACGGGGCCAGACGCCAC-CGCCGCTGTTT
TvY486_0805300        ATGCCTTCTAGAGAAATGAGTAAAAATGAGCAAAATGCGGAACAGCTCACCCCGTACTTT
                                    *   *  * ***   *      *     *  * * *       ***
Green: Annotated start codon
Yellow: T. vivax proposed start codon
```

Results after ATG correction:

Gene ID	Description	Reads	1st Splicing Site	Reads	2nd Splicing Site	Reads/3rd Splicing Site	Reads/4th Splicing Site
Tb427.08.5800	Hypothetical protein conserved	47	-32	32	-2	5/1	6/-29
TvY486_0805300	Hypothetical protein conserved	5	-6	--	--	--	--

Alignment after correction:

```
Tb427.08.5800          ---------------TCTCCCCTAAACCACCAGTCGCTGTTTTTTTTTGTTTTTGCGGTG
TcIL3000.0.57230       ---------------TAAAATTTTGTCGTCCCACCACGATATCTTTTTCTTTGAGTTGCG
TvY486_0805300         CCCCTCCCACCCCCATTTTCAATTATCTTTTTCTTCTTCTCTGTTTTTTATGCCTTCTA
                                      *      *   *            *    **** * *

Tb427.08.5800          TAGTAATGAAACACAAAGATGCTCGCGGTGGTTCGACAC-CGTAC---TTTGCCATAACG
TcIL3000.0.57230       TAGCAATGAAGCACAAAGATGGACGGGGCCAGACGCCAC-CGCCGCTGTTTTCTATAACC
TvY486_0805300         GAGAAATGA-GTAAAAATGAGCAAAATGCGGAACAGCTCACCCCGTACTTTGCAATAACG
                        ** *****   * ***   *      *     *  * * *       *** * *****

Green: annotated start codón
Yellow: T. vivax proposed start codon
Red: First trans-splicing site
Orange: Second trans-splicing site
Gray: Third trans-splicing site
Blue: Fourth trans-splicing site
```

Example 3

Gene ID	Description	Reads	1st Splicing Site	Reads	2nd Splicing Site	Reads/3rd Splicing Site	Reads/4th Splicing Site
Tb427.08.6180	60S ribosomal protein L26 putative	3445	-20	414	-16	6/19	5/14
TvY486_0805782	60S ribosomal protein L26 putative	296	-20	24	-11	--	--

```
Tb427.08.6180          MVGIKCRNRRKARRAHFQAPSHVRRILMSAPLSKELRAKYNVRSMPVRKDDEVRVKRGKF
TcIL3000.8.6030        MVGIKCRNRRKARRAHFQAPSHIRRILMSAPLSKELRAKYNVRSMPVRKDDEVRVKRGNF
TvY486_0805780         MPSIKCRNRRKARRAHFQAPSHVRRILMSAPLSKELRAKYNVRSLPVRKDDEVRVKRGAF
                       * .******************:*********************:************ *

Green: Annotated methionine start
Yellow: proposed methionine start
```

```
Tb427.08.6180          --CTGGCAATCGGTATTATCATCT-TCCTTTAGCGCCGCACAATCAACCATATATGGTCG
TcIL3000.8.6030        -AATATACACATACACACATATATGTTCAAAAGCTTAACACAATCATCAAC--ATGGTTG
TvY486_0805780         TTTTGGTTGGTTGTTTGCTTGTTTGCTCCAT  AGCGCATAAGTGTCCACA--ATGCCAA
                          *                  * *   *   ***    ** **    * *   ***

Tb427.08.6180          GCATTAAGTGTAGGAACCGCCGAAAGGCCCGTCGCGCACACTTCCAAGCGCCCAGTCATG
TcIL3000.8.6030        GCATCAAGTGCAGGAACCGCCGAAAGGCCCGCCGCGCCCACTTCCAGGCGCCCAGCCACA
TvY486_0805780         GCATAAAGTGTAGGAACCGTCGCAAGGCGCGTCGTGCACACTTTCAGGCCCCGAGTCATG
                       **** ***** ******** ** ***** ** ** ** ***** ** ** ** ** **

Green: Annotated start Codon
Red: First trans-splicing site
Orange: Second trans-splicing site
Blue: Third and fourth trans-splicing site.
```

Example 4:

Gene ID	Description	Reads	Splicing Site
Tb427.03.4810	hypothetical protein conserved	41	35
TvY486_0304140	hypothetical protein conserved	4	-8

First alignment:

```
Tb427.03.4810          MSFSHFVPPISRRRMFFEDQLDEALSREGSPRLSTSNTVGGADLVSAGAANDETFPFPSH
TcIL3000.0.42620       --------------MFFEEHLDEALSREGSPVIGSGHAANSVN-ISVGGVVDNAFLLASR
TvY486_0304140         --------------MFFEEHLDEPLSREDSPNLVAG---NGADNVSNGAP----FLFAAR
                                      ****::***.****.** : :.   ...: :* *.     * :.::
Green: Annotated methionine start
Yellow: T. brucei proposed methionine start
```

```
Tb427.03.4810         CAAGAATGTCATTTTCCCATTTTGTACCTCCAATTTCTA----GGAGACGTATGTTTTTTG
TcIL3000.0.42620      TTATTCCTTAATGCTGCCCTTTCATAAGTGTAGCCTCGA----AAAGGTGCATGTTTTTTG
TvY486_0304140        ACAACCGTCAATTACATCCCTTTAAATCCCTTGCTGTGACATTAGGGAGGAATGTTTTTTG
                        *       **     *  **   *            *       *  * **********
Green: Annotated start Codon
Red: Trans-splicing site
Yellow: T. brucei proposed start codon
```

Results after ATG correction:

Gene ID	Description	Reads	Splicing Site
Tb427.03.4810	hypothetical protein conserved	41	-7
TvY486_0304140	hypothetical protein conserved	4	-8

Alignment after correction:

```
Tb              --TCCAATTTCTAGGAGACGTATG
Tv              GCTGTGACATT AGG-GAGGAATG
                  *   *  *   ****  * ***
Red: Trans-splicing site
Green: Start Codon
```

Tabla Suplementaria 7. Comparación TSS en genes ortólogos

Sitios de *trans-splicing* en genes ortólogos entre *T. vivax* y *T.brucei* (primeros 35 genes).

Table S7

n SL	Gene ID	Product	reads	pos	reads	pos	reads	pos
3	TvY486_0905260	60S ribosomal protein L23 putative	4	6	9	-13	109	-18
3	TvY486_0907460	60S ribosomal protein L5 putative	2	-8	106	-15	36	-19
3	TvY486_1005370	40S ribosomal protein S18 putative	3	-10	21	-13	175	-19
3	TvY486_0601590	40S ribosomal protein S30 putative	12	-11	107	-19	189	-27
3	TvY486_1107730	hypothetical protein conserved	3	-13	11	-22	61	-26
3	TvY486_0604330	hypothetical protein conserved	2	-16	5	-22	60	-32
2	TvY486_0604440	60S acidic ribosomal protein P2 putative	2	3	30	-20		
2	TvY486_1103320	60S ribosomal protein L44	15	-3	300	-8		
2	TvY486_0300750	ribosomal protein S25 putative	19	-5	209	-34		
2	TvY486_0301950	hypothetical protein conserved	2	-5	2	-10		
2	TvY486_0805540	60S ribosomal protein L12 putative	10	-5	226	-17		
2	TvY486_1004150	60S ribosomal protein L30 putative	6	-6	11	-23		
2	TvY486_1005810	hypothetical protein conserved	4	-6	2	10		
2	TvY486_1109400	DNA-dependent RNA polymerases putative	3	-6	2	-12		
2	TvY486_0304410	60S ribosomal protein L4 putative	315	7	15	34		
2	TvY486_1001040	40S ribosomal protein S23 putative	2	7	85	-15		
2	TvY486_1003350	60S acidic ribosomal protein P2 putative	2	-7	49	-24		
2	TvY486_1008280	legume-like lectin putative	2	-7	19	-17		
2	TvY486_0602900	hypothetical protein conserved	3	-8	3	-13		
2	TvY486_1114810	LSM4psmall nucleolar ribonucleoprotein-lik	6	-8	2	-32		
2	TvY486_1116710	ribosomal protein L27 putative	2	-8	146	-14		
2	TvY486_0805470	protein kinase putative fragment	6	-9	2	9		
2	TvY486_1005090	ubiquitinribosomal protein S27a putative	7	-9	45	-12		
2	TvY486_0805750	40S ribosomal protein S8 putative	5	-10	230	-13		
2	TvY486_0907580	40S ribosomal protein S6 putative	3	-10	154	-24		
2	TvY486_1008420	hypothetical protein conserved	5	-10	2	-14		
2	TvY486_0402830	chaperone protein DNAj putative	3	-11	2	-14		
2	TvY486_0503740	NUDIX hydrolase putative	10	-11	2	-22		
2	TvY486_0805780	60S ribosomal protein L26 putative	24	-11	296	-20		
2	TvY486_0807660	amino acid transporter putative	3	12	4	22		
2	TvY486_1000500	ribosomal protein S17 putative	34	-12	19	-15		
2	TvY486_1006440	hypothetical protein conserved	3	-13	7	-21		
2	TvY486_0806380	IgE-dependent histamine-releasing factor p	2	-15	12	-24		
2	TvY486_1008320	40S ribosomal protein S12 putative	37	-15	306	-26		
2	TvY486_0102140	serine peptidase putativeserine peptidase (	2	-16	2	-23		
2	TvY486_0904200	hypothetical protein conserved	2	-16	4	-25		
2	TvY486_1004060	hypothetical protein conserved	2	16	3	-21		
2	TvY486_1006330	60S ribosomal proteins L37 putative	2	-16	23	-22		
2	TvY486_1100870	hypothetical protein conserved	3	-17	8	-26		
2	TvY486_1004640	peptidyl-prolyl cis-trans isomerase putative	2	-18	5	-31		
2	TvY486_1103730	40S ribosomal protein S4 putative	5	-18	48	-25		
2	TvY486_1112230	40S ribosomal protein L14 putative	2	-18	141	-27		
2	TvY486_1008040	hypothetical protein conserved	2	-20	10	-23		
2	TvY486_1008140	eukaryotic translation initiation factor 3 sub	3	-20	10	-24		
2	TvY486_1111790	hypothetical protein conserved	2	-20	3	-33		
2	TvY486_1114710	hypothetical protein conserved	3	-21	4	-18		
2	TvY486_0904730	hypothetical protein conserved	2	-22	10	-27		
2	TvY486_1006610	COP-coated vesicle membrane protein erv2	3	-23	19	-31		
2	TvY486_1009970	hypothetical protein conserved	3	-23	5	-26		
2	TvY486_1112850	40S ribosomal protein S17 putative	9	-23	110	-28		
2	TvY486_0601740	hypothetical protein conserved fragment	5	25	2	31		

n SL	Gene ID	Product	reads	pos	reads	pos	reads	pos	reads	pos	reads	pos
5	Tb427tmp.211.2630	60S ribosomal protein L23 putative	17	-1	118	-5	9	9	514	-11	4985	-15
3	Tb427tmp.244.2730	60S ribosomal protein L5 putative	24	-12	557	-15	3109	-19				
2	Tb427.10.5340	40S ribosomal protein S18 putative	621	-8	4745	-12						
2	Tb427.06.2110	40S ribosomal protein S30 putative	5	-9	1749	-18						
3	Tb427tmp.02.5150	hypothetical protein conserved	8	1	55	-4	276	-8				
2	Tb427.06.5010	hypothetical protein conserved	15	-28	175	-32						
3	Tb427.06.5130	60S acidic ribosomal protein P2 putative	10	-10	357	-14	2843	-18				
3	Tb427.02.6090	60S ribosomal protein L44	16	-14	131	-19	1040	-23				
2	Tb427.03.1370	40S ribosomal protein S25 putative	28	-15	261	-34						
3	Tb427.03.2660	hypothetical protein conserved	16	-3	17	-15	20	-19				
3	Tb427.08.6030	60S ribosomal protein L12 putative	10	-12	266	-15	1568	-19				
4	Tb427.10.4120	60S ribosomal protein L30	7	9	14	-6	5609	-20	25	-23		
1	Tb427.10.5870	hypothetical protein conserved	121	-5								
1	Tb427tmp.01.0625	DNA-dependent RNA polymerases putati	6	-34								
4	Tb427.03.5050	60S ribosomal protein L4	12	-4	106	-7	632	-25	5471	-29		
2	Tb427.10.1090	40S ribosomal protein S23 putative	15	-6	3117	-31						
4	Tb427.10.3370	60S acidic ribosomal protein P2 putative	34	-1	157	-5	633	-15	5972	-19		
3	Tb427.10.8400	legume-like lectin putative	21	2	151	-3	307	-8				
1	Tb427.06.3420	hypothetical protein conserved	51	-7								
3	Tb427tmp.01.5535	U6 snRNA-associated Sm-like protein LS	154	0	14	3	8	11				
4	Tb427tmp.01.7535	60S ribosomal protein L27 putative	160	-2	9	6	6852	-13	10	-18		
2	Tb427.08.5950	protein kinase putative	15	-31	120	-35						
3	Tb427.10.5030	ubiquitinribosomal protein S27a putative	291	-1	2564	-5	6	-11				
5	Tb427.08.6150	40S ribosomal protein S8 putative	353	-1	3179	-5	67	-11	574	-15	2509	-20
3	Tb427.10.190	40S ribosomal protein S6 putative	13	-6	221	-10	1435	-14				
2	Tb427.10.8640	hypothetical protein conserved	11	-17	80	-21						
2	Tb427.04.2970	chaperone protein DNAj putative	16	-5	167	-8						
2	Tb427.05.4350	NUDIX hydrolase putative	8	-9	73	-23						
4	Tb427.08.6180	60S ribosomal protein L26 putative	5	14	414	-16	6	19	3445	-20		
2	Tb427.08.8260	amino acid transporter putative	36	-27	183	-31						
3	Tb427.01.3180	40S ribosomal protein S11 putative	10	-21	365	-25	3492	-29				
2	Tb427.10.6480	hypothetical protein conserved	9	-20	48	-24						
1	Tb427.08.6760	IgE-dependent histamine-releasing factor	25	-22								
5	Tb427.10.8430	40S ribosomal protein S12 putative	15	-8	116	-12	523	-19	4855	-23	24	-29
1	Tb427.01.4780	serine peptidase putative	6	-35								
1	Tb427tmp.211.1280	hypothetical protein conserved	91	-17								
1	Tb427.10.4030	hypothetical protein	59	-26								
2	Tb427.10.6370	60S ribosomal proteins L37 putative	6	-12	3427	-27						
2	Tb427tmp.03.0115	hypothetical protein conserved	10	-18	114	-21						
1	Tb427.10.4620	peptidyl-prolyl cis-trans isomerase putati	379	-14								
4	Tb427tmp.02.1085	40S ribosomal protein S4 putative	5	8	8	-15	13	-22	5	-35		
3	Tb427tmp.01.3020	40S ribosomal protein L14 putative	10	-16	188	-20	1574	-24				
1	Tb427.10.8160	hypothetical protein conserved	31	-22								
2	Tb427.10.8290	eukaryotic translation initiation factor 3 s	13	-22	647	-31						
1	Tb427tmp.01.2630	hypothetical protein conserved	32	-34								
2	Tb427tmp.01.5440	hypothetical protein conserved	51	-25	386	-29						
1	Tb427tmp.211.1890	hypothetical protein conserved	19	-18								
3	Tb427.10.6640	COP-coated vesicle membrane protein er	25	-16	230	-19	137	-23				
2	Tb427.10.10080	hypothetical protein conserved	23	-11	254	-15						
7	Tb427tmp.01.3676	40S ribosomal protein S17 putative	6	-2	79	-5	9	9	7	-11	316	-14
5	Tb427.06.2220	hypothetical protein conserved	5	-11	38	-14	41	-20	6	21	391	-23

Anexo 2. Material suplementario de "Kinetoplast adaptations in American strains from *Trypanosoma vivax*".

A continuación se presenta el material suplementario del artículo "Kinetoplast adaptations in American strains from *Trypanosoma vivax*".

Archivo suplementario 1. Detalle de ensamblado de genomas mitocondriales

Detalles sobre la secuenciación y el ensamblado de los genomas mitocondriales de *T. viv*ax.

Maxicircle sequencing and assembly

The maxicircle from MT1 strain was sequenced (together with the nuclear genome) using Illumina GAIIx sequencer. 26 million of 2x100 bp reads were obtained which were quality filtered. Assembling was conducted together with the whole genome using Abyss assembler (K-mer size equal to 50). Contigs belonging to the maxicircle were identified by comparing with *T. brucei* using blastn. In this way two large contigs were found having long HSPs with *T. brucei* maxicircle. The alignment of these two initial contigs to *T. brucei* maxicircle is schematized in figure FS1 (a). Using *T. brucei* as a reference to infer gene ordering is fully justified by the fact synteny is conserved among *T. cruzi*, *T. brucei* and *L. tarentolae* (Westnberger et al, BMC Genomics, 7:60, 2006)

One of the two contigs (contig A) has a length of 3.3 kb and includes the two mitochondrial rRNA 12s and 9S and a 1.5 Kb region (located 5' to the rRNA genes) containing *T. vivax* species specific sequences. Three different types of such sequences are located in this region: one of them of about 1 Kb (represented by green box in FS1) and two clusters of tandem repeats composed of different types of repetitive sequences (represented by orange and blue rectangles in FS1). One of them, located close to the rRNA genes (blue rectangle), is composed by a repeating unit of 105 bp in length, the other one contains 24 nt repeat units (orange rectangle). These tandem arrays were collapsed in the assembly, and their repetitive nature was evidenced by back-mapping the reads on this contig (repetitive regions appear as segments of increased depth).

The second contig (B) had a length of 13.1 kb and contains 18 protein coding genes. This long contig has a 470 bp deletion that is expected to involve the 3' of ND7 and 5' from COIII. The existence of this deletion was confirmed by PCR (explained in box a, green box).

This initial assembly implied that two maxicircle segments were missing, a gap between the two contigs on one side (3' from contig A, and 5' from contig B) and a relatively large region

that joins these two contigs on the other side. Namely, that spanning between the leftmost extreme from contig A (orange and black segments of contig A) and the region next to the last protein coding gene (ND5) on the other side of contig B. In other trypanosomatid species this region has between 5 and 10 kb and contains species specific repetitive sequences. In order to sequence the two missing segments the following strategy was adopted. For the first (smaller) segment, two sets of primers were designed (only one set of such primers is represented by green arrowheads in figure F1S). The resulting amplicons, which exhibited the expected size according to the gap presented in figure F1S, were cloned and sequenced by Sanger technology (explained in **box a**, green box).

The second missing segment was amplified using primers schematized in figure FS1 by black arrowheads that yielded a 5.2 KB amplicon (represented by dashed and details of amplification in box d, black box). This was sequenced from each end using Sanger technology sequencing. It was possible to obtain a 2kb and a 1 kb sequence from each side. This DNA segment is composed by a different type of repeat unit that has 175 bp in length. Apart from the 3 kb sequenced by Sanger method, we could confirm that the whole 5.2 segment is composed only by this type of repeating unit. This was done using two different methods apart from Sanger just mentioned. For the first method we performed PCR using one primer that matches inside the repeat zone and a second primer that hybridize outside the repeat cluster. As it can be observed in FS1_B, the amplification products produce a ladder whose band sizes differ in 175 nt. The second approach was to construct an Illumina library using only this 5.2 Kb segment as starting material. This library was sequenced, and 20 thousand 2x150 bp reads were obtained. These reads are composed exclusively by the 175 nt repetitive sequence.

The sequencing of the maxircle from strain Liem176 was performed by combining DNA sequencing data and long RNAseq contigs obtained using 454 technology.

In this case sequencing of DNA was carried out with an Illumina Miseq sequencer. 5 millions of 2x150 bp reads were obtained that corresponds to both nuclear and mitochondrial genomes. By mapping the resulting reads on the MT1 maxircle, it was possible to observe that coverage was not homogenous, and some parts did not present any mapping reads. This anticipated that the assembly of this genome was going to be fragmented. This was effectively the case, and the maxicircle genome was obtained in 12 different contigs. The scaffolding of these contigs was done using the aid of the previously assembled genome (that of MT1) and long RNAseq contigs (retrieved from http://bioinformatica.fcien.edu.uy/Tvivax) obtained using 454 sequencing technology. Many of these contigs (listed in table TS1) are quite long and very

likely correspond to polycistronic transcription units before they are processed to become mature mRNA (i.e pre-edited and lacking any other modifications). This latter assumption is further supported by the fact that they map precisely on to the MT1 maxicircle genome.

By combining these two sources of data it was possible to assemble a 15.3 kb containing the 12S and 9S rRNA genes and the 18 protein coding genes.

For the case of the Y486 strain the strategy was completely different since its nuclear genome was obtained before using Sanger sequencing technology and is publicly available. Although the mitochondrial genome was not determined, we suspected that the reads corresponding to maxicircle were present in the raw sequencing data. We therefore mapped raw Sanger reads (downloaded from GeneDB) on to the MT1, Liem176 as well as *T. brucei* maxircles. 2816 reads were identified as of maxicircle origin. These reads were assembled using Mira yielding a single 20 kb contig that encompasses the whole maxicirle genome including not only the coding and conserved segment but also the other part of the maxicircle containing species specific and repeats. A schematic alignment of the tree genomes is presented in figure 1.

The final versions of assembled maxicircle genomes were deposited in Genbank under the accession numbers: KM386508 (*Tvivax*_MT1_maxicircle) and KM386509 (*Tvivax*_Liem_maxicircle).

Table TS1. RNAseq derived contigs (454 sequencing) download from http://bioinformatica.fcien.edu.uy. These contigs represent pre-edited ,mulicistronic ARNs

NAME	LENGTH
TvMiraNov_c3402	904
TvMiraNov_rep_c5265	718
TvMiraNov_c1118	1635
TvMiraNov_c37	1394
TvMiraNov_c2353	774
TvMiraNov_rep_c5287	1040
TvMiraNov_c457	2929
TvMiraNov_rep_c5279	1448
TvMiraNov_c553	2007
TvMiraNov_rep_c5667	634
TvMiraNov_rep_c5373	887

All minicircles classes were deposited in Genbank and their corresponding accession numbers:

Minicircle	Acc. Num
Tvminic1	KM386454
Tvminic2	KM386455
Tvminic3	KM386456
Tvminic4	KM386457
Tvminic5	KM386458
Tvminic6	KM386459
Tvminic7	KM386460
Tvminic8	KM386461
Tvminic9	KM386462
Tvminic10	KM386463
Tvminic11	KM386464
Tvminic12	KM386465
Tvminic13	KM386466
Tvminic14	KM386467
Tvminic15	KM386468
Tvminic16	KM386469
Tvminic17	KM386470
Tvminic18	KM386471
Tvminic19	KM386472
Tvminic20	KM386473
Tvminic21	KM386474
Tvminic22	KM386475
Tvminic23	KM386476
Tvminic24	KM386477

Minicircle	Acc. Num
Tvminic25	KM386478
Tvminic26	KM386479
Tvminic27	KM386480
Tvminic28	KM386481
Tvminic29	KM386482
Tvminic30	KM386483
Tvminic31	KM386484
Tvminic32	KM386485
Tvminic33	KM386486
Tvminic34	KM386487
Tvminic35	KM386488
Tvminic36	KM386489
Tvminic37	KM386490
Tvminic38	KM386491
Tvminic39	KM386492
Tvminic40	KM386493
Tvminic41	KM386494
Tvminic42	KM386495
Tvminic43	KM386496
Tvminic44	KM386497
Tvminic45	KM386498
Tvminic46	KM386499
Tvminic47	KM386500
Tvminic48	KM386501

Minicircle	Acc. Num
Tvminic49	KM386502
Tvminic50	KM386503
Tvminic51	KM386504
Tvminic52	KM386505
Tvminic53	KM386506
Tvminic54	KM386507

List of primers and PCR conditions were used to final maxicircle assembly and minicircle confirmation are summarized in the next tables:

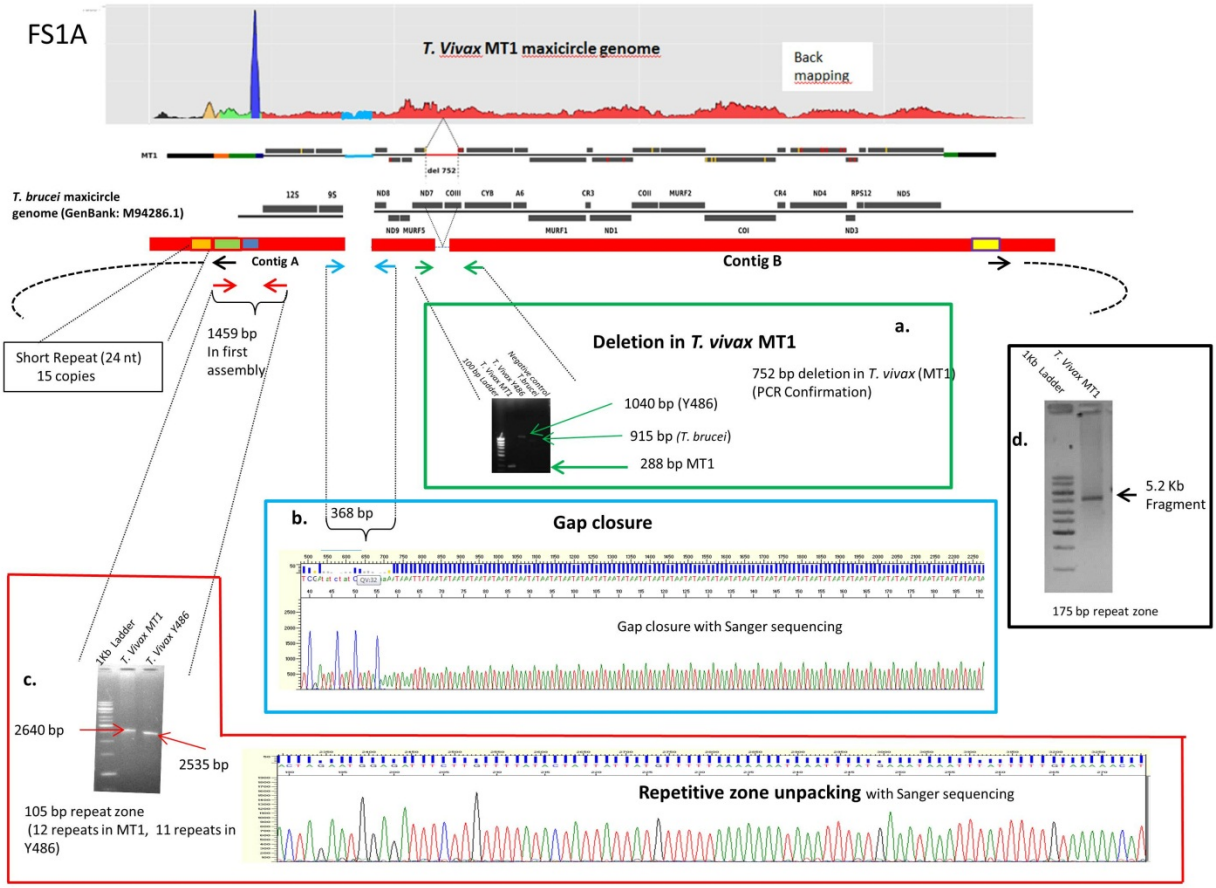
Primers Table,PCR conditions and Sanger sequencing

Maxicircle Final Assembly:

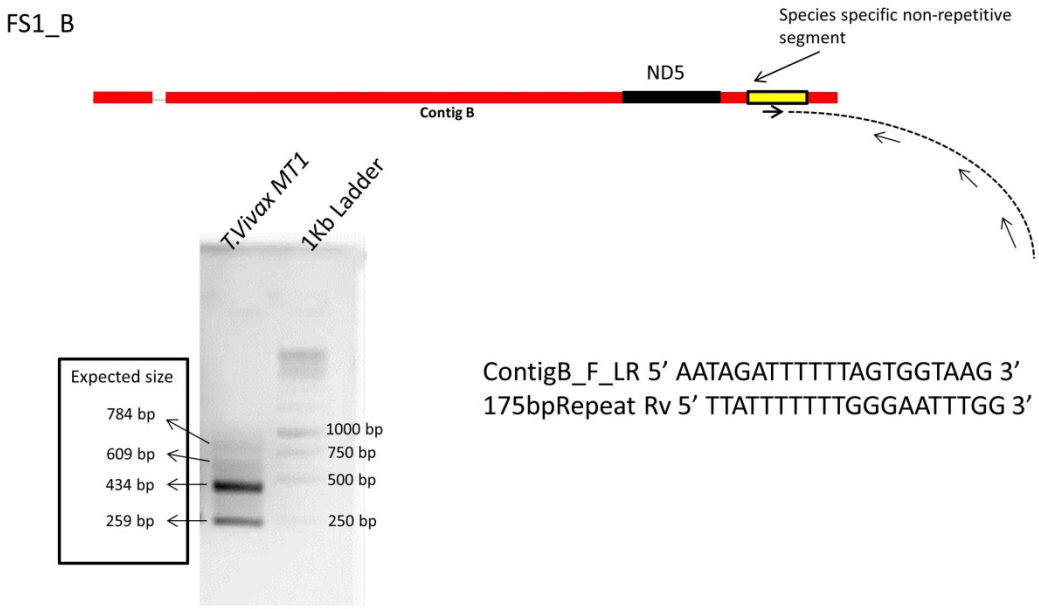
Primer name	Sequence (5´-3´)	Coordinates (respect to final assembly)	Comment	PCR conditions	Enzyme used
ContigB_F	AAGGTATTGTTGCCCACC	5190-5207	Close gap between ContigA-ContigB	94°C, 5'; **37 cycles: 94°C,30''-50°C,30''-72°C,30'**; 72°C 20'	Taq Thermo Scientific (USA)
ContigA_R	CTAAACCCCCCCGCTCTTGCTC	5565-5544			
ContigB_F_deletion	GAGTGATTGAGTGGGAAAG	6403-6421	Confirm deletion respect *T. brucei* maxicircle	94°C, 5'; **37 cycles: 94°C,30''-50°C,30''-72°C,30'**; 72°C 20'	Taq Thermo Scientific (USA)
ContigB_R_deletion	AACTCCTCCCTCTCTGC	6690-6674			
ContigB_F_LR	AATAGATTTTTTAGTGGTAAG	18260-18280	Amplification and sequencing of long repeated region (5.2 Kb product)	94°C, 5'; **37 cycles: 94°C,30''-52°C,30''-65°C,10'**; 65°C 60'	LongAmp®. NEB (USA)
ContigA_R_LR	GAATTTTTTTTGGCGTTTC	853-835			
ContigA_F_DeepZone	TGCTCCACCCACCATTAATTTTATC	1467-1484	Amplification of 105 bp repeated zone	95°C, 2'; **30 cycles: 95°C,30''-50°C,30''-72°C,2'**; 72°C 20'	FailSafe DNA polymerase. Epicentre (USA)
ContigA_R	GACAAACGCATTTAAACGC	4090-4076			
Rev_175RepeatZone	TTATTT TTTTGGGAA TTT	in 175 bp repeated zone	Amplification to confirme 175 bp repeat zone (using ContigB_F_LR as forward primer)	94°C, 5'; 38 cycles: 94°C,30''-51°C,30''-72°C,60''; 72°C 10'	Taq Thermo Scientific (USA)
ContigA_R_2_LR	AAAATTAATGGTGGGAGCA	1481-1463	Sequencing Primers to Unknown zone	Not correspond	Not correspond

Figura Suplementaria 1. Ensamblaje final de maxícirculo de MT1

A. Detalle del ensamblaje final del maxicírculo de la cepa MT1 de *T. vivax*.



B. Detalle del ensamblaje final del maxicírculo de la cepa MT1 de *T. vivax*.



Archivo suplementario 2. Identificación de ARNm editado.

Identificación de ARNm editado. Búsqueda de la presencia de cada ARNm en las cepas americanas y africanas de *T. viv*ax.

The identification of edited mRNA sequences was conducted for all mRNA maxicircle genes. To this end the complete transcriptome was assembled. The resulting contigs were virtually translated and compared with the amino acid sequences of maxicircle encoded proteins from *Trypanosoma brucei* (using blastp). The rationale of this approach is that while maxicircle DNA sequences are not necessarily well conserved among trypanosomatids (even inside the coding part of mitochondrial genes), there is pronounced amino acid conservation for the proteins encoded by these genes as revealed by previous studies.

Assembling of RNA sequences from Y486 was conducted with spades using bloodstream RNAseq reads (2x150 paired end reads) produced in our laboratory as explained in Material and Methods. Epimastigote and metacyclic epimastigote derived reads (2x100 paired end reads) were used to check the expression of maxicircle genes in these life cycle stages. These latter reads were downloaded from the NCBI Sequence Read Archive (SRA) public database (http://www.ncbi.nlm.nih.gov/sra) (SRA accession numbers ERX211384-ERX211392).

Assembled maxicircle RNA contigs (edited mRNAs) were deposited in GenBank under the following accession numbers:

Gene Name	GenBank Acc. Num.
ND8_*T.vivax*	KM386442
ND9_*T.vivax*	KM386443
ND7_*T.vivax*	KM386444
COIII_*T.vivax*	KM386445
Cyb_*T.vivax*	KM386446
A6_*T.vivax*	KM386447
CR3_*T.vivax*	KM386448
COII_*T.vivax*	KM386449
MURF2_*T.vivax*	KM386450
CR4_*T.vivax*	KM386451
ND3_*T.vivax*	KM386452
RPS12_*T.vivax*	KM386453

Edited mRNAs. Alignments between genomic and edited RNAs for all maxicircle genes that undergo editing.

ND8

```
>T.vivax (Y486) ND8 mRNA (edited)

ATGTTTTTTTTTGATTTTTTGTTTTTTTTTTTTGTTTGTTTTTATATGTGTTTTGTTTGTTGTGTTACCATTTGTTTACCCATTGAATTAACATTT
TGTAGTTTGTTGACCCGTGGTAACCATTTTTTGCGTTTTTATTGATGTGGTTTAGAACGTTGTATTGCTTGTCGTTTATGTGATTTGATTTGTCCC
AGTTTAGCATTAGATGTTCGTGTTGGGTGGAGTTTTGGTGGTCATCGTTTTGCAGATTGATTTACATTGAGTTACCGACGTTGTATTTATTGTGGT
TTTTGTATGCATGTTTGTCCAACAGATGCCATTACACATTCATTGTTTGTTATGTGTTTTTGTTGTTTAGCCATGTATTTATTGGCGCCCAAGTTT
TTGTTATTTGGTTGTTGTTTTATGTTGTTTGATTTTTATTTGTGTTTTGTGTAGTT

>ND8_edited Translated
MFFFDFLFFFFVCFYMCFVCCVTICLPIELTFCSLLTRGNHFLRFYWCGLERCIACRLCDLICPSLALDVRVGWSFGGHRFADWFTLSYRRCIYCG
FCMHVCPTDAITHSLFVMCFCCLAMYLLAPKFLLFGCCFMLFDFYLCFV*
```

Confirmation by aligning with *T. brucei*

```
ND8_brucei        MFFFDFLFFFFVCFYMCFVCCVTICLPIELTIVSLLVRGNHFLRFYWCGLERCIACRLCD 60
ND8_edited(Trans) MFFFDFLFFFFVCFYMCFVCCVTICLPIELTFCSLLTRGNHFLRFYWCGLERCIACRLCD 60
                  ******************************: ***.**********************

ND8_brucei        LICPSLALDVRVGWSFGGHRFADWFTLSYRRCIYCGFCMHVCPTDAITHSLFVMCFCCLA 120
ND8_edited(Trans) LICPSLALDVRVGWSFGGHRFADWFTLSYRRCIYCGFCMHVCPTDAITHSLFVMCFCCLA 120
                  ************************************************************

ND8_brucei        MYLLAPKFLLFGCCFMLFDFYLCFV 145
ND8_edited(Trans) MYLLAPKFLLFGCCFMLFDFYLCFV 145
                  *************************
```

Alignment: Genomic vs Edited (mRNA) sequences

```
ND8_genomic   A-G---------GA------G------------G---G-----A-A-G-G----G---G-
ND8_edited    AuGuuuuuuuuuGAuuuuuuGuuuuuuuuuuuuGuuuGuuuuuuAuAuGuGuuuuGuuuGu

ND8_genomic   -G-G--ACCATT-G---ATTTTCCCA--GAA--AACA----G-AG---G--GATCCCG-G
ND8_edited    uGuGuuACCATTuGuuuA----CCCAuuGAAuuAACAuuuuGuAGuuuGuuGA-CCCGuG

ND8 genomic   G-AATTTCTCTTTA------GCG-----ATTTGA-G-GG---AGAACGT-G-A--GC--G
ND8_edited    GuAA---C-C---AuuuuuuGCGuuuuuATT-GAuGuGGuuuAGAACGTuGuAuuGCuuG

ND8 genomic   -CG---A-G-GA---GA---G-CTTTTTCCAG---AGCA--AGA-G--CG-G--GGG-GG
ND8_edited    uCGuuuAuGuGAuuuGAuuuGuC-----CCAGuuuAGCAuuAGAuGuuCGuGuuGGGuGG

ND8 genomic  TTTAG----GG-GG-CA-CG----GCATTGA--GTTA---ACA--GAG--ATTTCCGTTT
ND8_edited   ---AGuuuuGGuGGuCAuCGuuuuGCA--GauuG--AuuuACAuuGAGuuA---CCG---

ND8 genomic  ACTTTTG--G-A---A--G-GG-----G-A-GCA-G---G-CCAACAGA-GTTTCCTTTT
ND8_edited   AC----GuuGuAuuuAuuGuGGuuuuuGuAuGCAuGuuuGuCCAACAGAuG---CC----

ND8 genomic  A--ATTTTTCACATTTTTCA--G---G--A-G-G-----G--G---ATTTTTTTGCCTTA
ND8_edited   AuuA-----CACATT---CAuuGuuuGuuAuGuGuuuuuGuuGuuuA-------GCC--A

ND8 genomic  TG-A---A--GGCGCCCTTTTAAG-----G--A---GGTTTTTG--G----A-G--G---
ND8_edited   TGuAuuuAuuGGCGCCC----AAGuuuuuGuuAuuuGGTT---GuuGuuuuAuGuuGuuu

ND8 genomic  GA-----A---G-G----G-G-AG
ND8_edited   GAuuuuuAuuuGuGuuuuGuGuAG
```

ND9

```
>T.vivax (Y486) ND9 mRNA (edited)
ATGGGTTTGTTGTTGTGTTTATTTCGTTTGTGTTTTTGTTTGATTTTGTTTTATTGTTTGTGTTGTAGGTGGTGTTTTGTTTGTTTTGTTGATTGT
AGTTTTTTGTTTTTTTATTGTTTTGTTAGTTTTTTTTTGTTTTATTGTATGTTTTTATTTTTTAATTTATGGTTTTTGTTTTTGTATTGTTGTGAT
TTATTGTTGATTGATTTTTGTGGTTTTTGTTTTTGTCGTTTTATGTTATTATATATTTTGTTTTGTTTGTTTTTGTGTTTTCGTTTATGTTTTGTT
TTGTGTTGTTTGTTTTTGTTTTTTGGATTGTGTTTTAGTTTTAGTTGTTTTTGTTATGCGTTTTTATTGTTAGAACGTGAGTGTTTTGATTTGTTT
GGTTTTTATTTTGTTGGTAATGATGTTTTACACCGTTTATTTGTTGATTGATTTTTTGTTGGTTTTTTTTTGTTGAAGTGTTATCCATTATTTGGT
TTGTTTGTATTATTGTTTTGTGTTTTAGTTAATGAGATGATATGTACGTTTACAATGGTTTTTTTGTTGTTGCATACCAATTTTTATTTGCATTTT
TTTGTTTAG

>ND9_edited Translated
MGLLLCLFRLCFCLILFYCLCCRWCFVCFVDCSFLFFYCFVSFFLFYCMFLFFNLWFLFLYCCDLLLIDFCGFCFCRFMLLYILFCLFLCFRLCFV
LCCLFLFFGLCFSFSCFCYAFLLLERECFDLFGFYFVGNDVLHRLFVDWFFVGFFLLKCYPLFGLFVLLFCVLVNEMICTFTMVFLLLHTNFYLHF
FV*
```

Confirmation by aligning with *T. brucei*

```
ND9_brucei          MCIFLCLFRLCFCLILFYCLCCRWCFVCFVDCSFLFFYCFVSFFLFYCMFLFFNLWFLFL 60
ND9_edited(Trans)   MGLLLCLFRLCFCLILFYCLCCRWCFVCFVDCSFLFFYCFVSFFLFYCMFLFFNLWFLFL 60
                    * ::*********************************************************

ND9_brucei          YCCDLLLIDFCGFCFCRFMLLYILFCLFLCVRLCFVLCCLFVFFGLCFSFSCFCYAFLLL 120
ND9_edited(Trans)   YCCDLLLIDFCGFCFCRFMLLYILFCLFLCFRLCFVLCCLFLFFGLCFSFSCFCYAFLLL 120
                    ******************************.**********:******************

ND9_brucei          ERECFDLFGFYFVGNDILHRLFVDWFFVGFFLLKCYPLFGLFVLLFCVLVEEIVCTFTML 180
ND9_edited(Trans)   ERECFDLFGFYFVGNDVLHRLFVDWFFVGFFLLKCYPLFGLFVLLFCVLVNEMICTFTMV 180
                    ***************:**********************************:*::*****:

ND9_brucei          FLLLHTNFYLHYFI 194
ND9_edited(Trans)   FLLLHTNFYLHFFV 194
                    ***********:*:
```

Alignment: Genomic vs Edited (mRNA) sequences

```
ND9_genomic     A-GGG---G--G--G-G---A---CGTT-G-GTTTTTG---GAT---G----AT-G--
ND9_edited      AuGGGuuuGuuGuuGuGuuuAuuuCGTTuGuGTTTTTGuuuGATuuuGuuuuATuGuu

ND9_genomic     -GTG--G-ATTTTTTTTTTGGTGG-G----G---G----G--GATTTG-AG------G
ND9_edited      uGTGuuGuA----------GGTGGuGuuuuGuuuGuuuuGuuGATT-GuAGuuuuuuG

ND9_genomic     -------ATTGTTT-G--ATG---------G----A--G-A-G-----ATT----AA-
ND9_edited      uuuuuuuuATTGTTTuGuuA-Guuuuuuuuuu Guuuu AuuGuAuGuuuuuATTuuuuAAu

ND9_genomic     --A-GTG-----G-----G-A--G--G-GTA---A--G--GA--GA-----G-GGT--
ND9_edited      uuAuG-GuuuuuGuuuuuGuAuuGuuGuG-AuuuAuuGuuGAuuGAuuuuuGuGGTuu

ND9_genomic     --G-----G-CG----A-G--A--A-A-A----G----G---GT----G-G----CG-
ND9_edited      uuGuuuuuGuCGuuuuAuGuuAuuAuAuAuuuuGuuuuGuuuGTuuuuGuGuuuuCGu

ND9 _brucei     --A-G----G----G-GTTGTT-G-----GTTTT--GGA--GTTTTTG----ATTG--
ND9_edited      uuAuGuuuuGuuuuGuGTTGTTuGuuuuuGTTTTuuGGAuuGT----GuuuuA--Guu

ND9_genomic     --AG--G-----G--A-GCTTTG-----ATTG--AGATTTTACG-GAG-G----GA--
ND9_edited      uuAGTTGuuuuuGuuAuGC---GuuuuuATTGuuAGA----ACGuGAGuGuuuuGAuu

ND9_genomic     -G---GG-----A----GTTTTGG-AA-GA-G----ACACCG---AT--G--GTTA--
ND9_edited      uGuuuGGuuuuuAuuuuGTT--GGuAAuGAuGuuuuACACCGuuuATuuGuuG--Auu
```

```
ND9_genomic     GATT----G--GTG---------G--GATAG-G--A-CCATTTTA---GGT--G---G
ND9_edited      GATTuuuuGuuG-Guuuuuuuuu GuuGA-AGuGuuAuCCATT--AuuuGGTuuGuuuG

ND9_genomic     -A--A--G----G-G----AG--AA-GATTGATGA-A-G-ACG---ACAA-GG-----
ND9_edited      uAuuAuuGuuuuGuGuuuuAGuuAAuGA--GATGAuAuGuACGuuuACAAuGGuuuuu

ND9_genomic     --G--G--GCA-ACCTTAA-----A---GCATT-----G---ATG
ND9_edited      uuGuuGuuGCAuACC--AauuuuuAuuuGCATTuuuuuGuuuA-G
```

ND7

```
>ND7_edited (Y486) ND7 mRNA (edited)
ATGTTTATTTTTGTTGTAGTTTTTTTGCATTTGTATCGTTTTACATTTGGCCCACAGCATCCCGCAGCACATGGTGTTTTATGTTGTTTATTGTAT
TTTTGTGGTGAATTTATTGTTTATATTGATTGTATTATAGGTTATTTGCATCGAGGTACAGAAAAGTTATGTGAGTATAAGAGCGTAGAGCAGTGT
CTTCCGTATTTTGATAGATTAGATTATGTTAGTGTTTGTTGTAATGAACATTTATTGTCTTTATGTTTTGAGTATATGTTACGGTGTTGTTTATCA
TTGCGTGTTGCATTTATGCGTTTATTGATTGTAGAGTTTACTCGTAGTTTTAATGGTTTATTGTGTGTGTCGTGTATGATTTTAGATTTAGGTTGT
TTATCCCCGTTATTATGGTCATTTGAGGAGCGTGATAAGTTAATGACGTTTTTTGATTTGTGTTGTGGTTGTCGTATGCATTTGGCTTTCATGGTT
TTATTGGGTATTCTTGATGATTTTGTTTTTGGTTTTGTTGATTTTTTGTTGTTATTGATAATATCGTGTTTGTTTGTTATGGATTGTTATGATTTA
TTGTTTGTGGGTAATCGTTTGTTTTATTTGCGTTTGCGTGGTTTGTCATTTTTTGATTTGTATGATTTGGTTTTTAATAGTTTGAGTGGTGTTTTG
TCACGTTCATTGGGTATGGTATGAGATTGCCGTTTATATTGTTGTTATGAGTTGTATTTTATGTTTTGTTATGATTATTGTTTTTGTTTTATAGGT
GATGCATTTGACCGTTTGTTTTTGCGTTTGTTTGATATGCGTATGAGTTTGTTGATTTGTAAGCAATGTTTTTTTGTTGGTTTTTTTGTTTTTGGA
TTTGTTTGTTTGTTTGATTATTTGTATTGTGATGTTACCATTGAGACTATTATTATGTTGTTTTATAGTTTATGGTGTTGTTGTTTACCAGGTATA
TCATTTGCTTGTGTTGAGCATCCCAAGGGTGAGTATTGTTTGTTATTATGTTTTTGTGTTGGTTTGTGTTCCCGTTTGCGTTTGCGTTGTGCGGAT
TTTTTACATATTTGTTTGTTGGATGTTTGTTTACGTGGTTTTTTATTGCATGATTTAGTTGCCGTTATTGGTAATATTGATGTTGTTTTTGGATCT
GTGGATCGTTAG
>ND7_edited Translated
MFIFVVVFLHLYRFTFGPQHPAAHGVLCCLLYFCGEFIVYIDCIIGYLHRGTEKLCEYKSVEQCLPYFDRLDYVSVCCNEHLLSLCFEYMLRCCLS
LRVAFMRLLIVEFTRSFNGLLCVSCMILDLGCLSPLLWSFEERDKLMTFFDLCCGCRMHLAFMVLLGILDDFVFGFVDFLLLLIISCLFVMDCYDL
LFVGNRLFYLRLRGLSFFDLYDLVFNSLSGVLSRSLGMVWDCRLYCCYELYFMFCYDYCFCFIGDAFDRLFLRLFDMRMSLLICKQCFFVGFFVFG
FVCLFDYLYCDVTIETIIMLFYSLWCCCLPGISFACVEHPKGEYCLLLCFCVGLCSRLRLRCADFLHICLLDVCLRGFLLHDLVAVIGNIDVVFGS
VDR*
```

Confirmation by aligning with *T. brucei*

```
ND7_brucei          M-LFLVVFLHLYRFTFGPQHPAAHGVLCCLLYFCGEFIVYIDCIIGYLHRGTEKLCEYKS 59
ND7_edited(Trans)   MFIFVVVFLHLYRFTFGPQHPAAHGVLCCLLYFCGEFIVYIDCIIGYLHRGTEKLCEYKS 60
                    *::*:**************************************************

ND7_brucei          VEQCLPYFDRLDYVSVCCNEHLLSLCFEYMLRCCLSLRCAFMRLLIVEFTRSFNGLLCIS 119
ND7_edited(Trans)   VEQCLPYFDRLDYVSVCCNEHLLSLCFEYMLRCCLSLRVAFMRLLIVEFTRSFNGLLCVS 120
                    ********************************** ******************:*

ND7_brucei          CMVLDLGCLSPLLWSFEERDKLMTFFDLCCGCRMHLAFMVLLGILDDFVFGFVDFLLLLI 179
ND7_edited(Trans)   CMILDLGCLSPLLWSFEERDKLMTFFDLCCGCRMHLAFMVLLGILDDFVFGFVDFLLLLI 180
                    **:********************************************************

ND7_brucei          ISCLFVMDCYDLLFVGNRLFYLRLRGLSFFDLYDLVFNSLSGVLSRSLGMVWDCRLFSCY 239
ND7_edited(Trans)   ISCLFVMDCYDLLFVGNRLFYLRLRGLSFFDLYDLVFNSLSGVLSRSLGMVWDCRLYCCY 240
                    ************************************************:.**

ND7_brucei          ELYFMFCYDYCFCFIGDAFDRLFLRLFDMRMSLLICKQCFFVGFFVFGFVCLFDYLYCDI 299
ND7_edited(Trans)   ELYFMFCYDYCFCFIGDAFDRLFLRLFDMRMSLLICKQCFFVGFFVFGFVCLFDYLYCDV 300
                    ************************************************************:

ND7_brucei          TIETIIMLFYSLWCCCLPGISFACVEHPKGEYCLLLCFCVGLCSRLRLRCADFLHICLLD 359
ND7_edited(Trans)   TIETIIMLFYSLWCCCLPGISFACVEHPKGEYCLLLCFCVGLCSRLRLRCADFLHICLLD 360
                    ************************************************************

ND7_brucei          VCLRGFLLHDLVAVLGNIDVVFGSVDR 386
ND7_edited(Trans)   VCLRGFLLHDLVAVIGNIDVVFGSVDR 387
                    **************:************
```

Alignment: Genomic vs Edited (mRNA) sequences

```
ND7_genomic     A-G---A-----G--G-AG-------GCA---G-ATCGT---ACAT--GGCCCACAGCA-
ND7_edited      AuGuuuAuuuuuGuuGuAGuuuuuuuGCAuuuGuATCGTuuuACATuuGGCCCACAGCAu

ND7_genomic     CCCGCAGCACA-GTGTTG----A-GT-G---ATTG-A-----GTGG-GAA---A--G---
ND7_edited      CCCGCAGCACAuG-GT-GuuuuAuGTuGuuuATTGuAuuuuuGTGGuGAAuuuAuuGuuu

ND7_genomic     A-ATTTTTGA--G-A--A-AGG--ATTTTTGCATCGAGGTACAGAAAAGTTATGTGAGTA
ND7_edited      AuA---TTGAuuGuAuuAuAGGuuATTT--GCATCGAGGTACAGAAAAGTTATGTGAGTA

ND7_genomic     TAAGAGCGTAGAGCAGTGTCTTCCG-ATTT-GATTTTAGA--AGA--A-G--ATTTTG-G
ND7_edited      TAAGAGCGTAGAGCAGTGTCTTCCGuATTTuGAT---AGAuuAGAuuAuGuuA----GuG

ND7_genomic     ---G--G-AA-GAACA---A--G-C---A-G----GAG-ATTTA-G--ACGG-G--G---
ND7_edited      uuuGuuGuAAuGAACAuuuAuuGuCuuuAuGuuuuGAGuAT--AuGuuACGGuGuuGuuu

ND7_genomic     A-CA--GCG-G--GCA---A-GCG---A--GA--G-AGAG---ACTTTCTG-AG----AA
ND7_edited      AuCAuuGCGuGuuGCAuuuAuGCGuuuAuuGAuuGuAGAGuuuACT--C-GTAGuuuuAA

ND7_genomic     -GG---A--G-G-G-G-CG-G-A-GA----AGA---AGG--G---A-CCCCG--AT-A-G
ND7_edited      uGGuuuAuuGuGuGuGuCGuGuAuGAuuuuAGAuuuAGGuuGuuuAuCCCCGuuATuAuG

ND7_genomic     G-CA---GAGGAGTTTTCG-GATTAAG--AA-GACG------GA---G-G--G-GG--G-
ND7_edited      GuCAuuuGAGGAG----CGuGAT-AAGuuAAuGACGuuuuuuGAuuuGuGuuGuGGuuGu

ND7_genomic     CG-A-GCA---GGCTTTCA-GG----A--GTTGG-A--CTTGA-GA----G-----GG--
ND7_edited      CGuAuGCAuuuGGCTTTCAuGGuuuuAuuG--GGuAuuCTTGAuGAuuuuGuuuuuGGuu

ND7_genomic     --G--GATTT---G--G--A--GA-AA-A-CG-G---G---G--A-GGA--G--A-GA--
ND7_edited      uuGuuGATTTuuuGuuGuuAuuGAuAAuAuCGuGuuuGuuuGuuAuGGAuuGuuAuGAuu

ND7_genomic     -A--G---GTTTGGG-AATTTTCG---G----A---GCG---GCG-GG---G-CA-----
ND7_edited      uAuuGuuuGT--GGGuAAT---CGuuuGuuuuAuuuGCGuuuGCGuGGuuuGuCAuuuuu

ND7_genomic     -GA---GTTTTA-GA---GG-----AA-AG---GAG-GG-G----G-CACG--CA--GGG
ND7_edited      uGAuuuGT---AuGAuuuGGuuuuuAAuAGuuuGAGuGGuGuuuuGuCACGuuCAuuGGG

ND7_genomic     -A-GG-A-GAGA--GCCG---A-A--G--G--ATTTTGAG--GTTA----ATTTTG----
ND7_edited      uAuGGuAuGAGAuuGCCGuuuAuAuuGuuGuuAT---GAGuuGT-ATTTTAT---Guuuu

ND7_genomic     G--A-GA--A--G-----G----A-ATGG-GA-GCA---GACTCG---G-----GCG---
ND7_edited      GuuAuGAuuAuuGuuuuuGuuuuAuA-GGuGAuGCAuuuGAC-CGuuuGuuuuuGCGuuu

ND7_genomic     G---GA-A-GCG-A-GAG---G--GA---G-AAGCAATTG-------G--GG-------G
ND7_edited      GuuuGAuAuGCGuAuGAGuuuGuuGAuuuGuAAGCAAT-GuuuuuuuGuuGGuuuuuuuG

ND7_genomic     -----GGA---G---G---G---GA--A---G-A--G-GA-G--ACCA--GAGAC-A--A
ND7_edited      uuuuuGGAuuuGuuuGuuuGuuuGAuuAuuuGuAuuGuGAuGuuACCAuuGAGACuAuuA

ND7_genomic     --A-G--G----A-AG---ATTTGG-G--G--G---ACCAGG-A-ATTTTTTTCA---GC
ND7_edited      uuAuGuuGuuuuAuAGuuuAT--GGuGuuGuuGuuuACCAGG-AuAT------CATTTGC

ND7_genomic     TTGTTG--GAGCA-CCCAAGGTTTTTG-GAG-A--G---G--A--ATTTTTG-----G-G
ND7_edited      TTGT-GuuGAGCAuCCCAAGG-----GuGAguAuuGuuuGuuAuuAT----Guuuuu GuG

ND7_genomic     --GG---G-GTTCCCG---GCG---GCG--G-GCGGA------ACATTTTATTTTTG---
ND7_edited      uuGGuuuGuGTTCCCGuuuGCGuuuGCGuuGuGCGGAuuuuuuACAT---ATTT--Guuu

ND7_genomic     G--GGA-G---G---ACG-GG------A--GCA-GA---AG--GCTTTTTCTG--A--GG
ND7_edited      GuuGGAuGuuuGuuuACGuGGuuuuuuAuuGCAuGAuuuAGuuGC-----C-GuuAuuGG

ND7_genomic     -AA-A--GA-G--G-----GGA-CTTTG-GGATCG--ATG
ND7_edited      uAAuAuuGAuGuuGuuuuuGGAuCT--GuGGATCGuuA-G
```

COIII

(frag)

```
>T.vivax (Y486) COIII mRNA (edited)
TTGTTGTTGTTTTTTTTGTTTTTGTTGAATCAGTTGTGGGAATTTACGTTGTTATTTATTACGTTGAGTTGTGGTGTTTTTGGTTCTATTTTATTT
TTATTGGATTTGTTGCATTTTATGCATGTTTTTTTAGGTGTTTTATTGTTGTTTTATGTTTTTATGCGTTTGTTTAGTTTTTTATGTATGGATACA
CGTTTTGTTTTTTTGTATGTTGTTTGTTTGTATTGACATTTTGTTGATTTGGTTTGGTTTTTTTTGTTGCGATTTGTTTATTTTGATGTTTTATGT
GTTATGTATTTGTGTGTGTAG

>COIII_edited Translated
LLLFFLFLLNQLWEFTLLFITLSCGVFGSILFLLDLLHFMHVFLGVLLLFYVFMRLFSFLCMDTRFVFLYVVCLYWHFVDLVWFFLLRFVYFDVL
CVMYLCV*
```

Confirmation by aligning with *T. brucei*

```
COIII_brucei         MFLFRCIFVGVSGVFVFLSLPAIVIVYWLFCLLGFICLLFGSFLFVDCGFIFFFVGFCIC 60
COIII_edited (Trans) ------------------------------------------------------------

COIII_brucei         LLLLLLDLFCDFLRGLFDFCVLLRCIQYCFLWFLCSEFVLFMAFFVVLFGLCLFLCCEFA 120
COIII_edited (Trans) ------------------------------------------------------------

COIII_brucei         FVFCLPYMFCCLLCDYGFVFYWYFLDLFNLLINTFYLFVSGLFVNFFVLCFWFRFFCCCC 180
COIII_edited (Trans) ------------------------------------------------------------

COIII_brucei         FVLWLSLLFGFLFLWNQLWEFALLFITLSCGVFGSILFLLDLLHFMHVFLGVLLLFICFM 240
COIII_edited (Trans) ------LLLFFLFLLNQLWEFTLLFITLSCGVFGSILFLLDLLHFMHVFLGVLLLFYVFM 54
                           **: **** ******:**********************************  **

COIII_brucei         RLFNFLCMDTRFVFLYCVCLYWHFVDLVWFFLLRFVYFDVLCVMYLCV 288
COIII_edited (Trans) RLFSFLCMDTRFVFLYVVCLYWHFVDLVWFFLLRFVYFDVLCVMYLCV 102
                     ***.************ ******************************
```

Alignment: Genomic vs Edited (mRNA) sequences

```
COIII_genomic        --G--G--G--------G-----G--GAA-CATTG--GTTTTGGGATTA---ACG--G--
COIII_edited         uuGuuGuuGuuuuuuuuGuuuuuGuuGAAuCA--GuuGT---GGGA--AuuuACGuuGuu

COIII_genomic        A---A--ACG--GAG--G-GG-GTTTTTTGG--C-A----A-----A--GTTTGA---G-
COIII_edited         AuuuAuuACGuuGAGuuGuGGuGTTTTT-GguuCuAuuuuAuuuuuAuuG---GAuuuGu

COIII_genomic        -GCA----A-GCA-G-------AGG-G----A--G--G----A-GTT---A-GTTCG---
COIII_edited         uGCAuuuuAuGCAuGuuuuuuuAGGuGuuuuAuuGuuGuuuuAuGTTuuuAuG--CGuuu

COIII_genomic        G---AG------A-G-A-GGA-ACACG----G-------G-A-G--G---G---G-A--G
COIII_edited         GuuuAGuuuuuuAuGuAuGGAuACACGuuuuGuuuuuuuGuAuGuuGuuuGuuuGuAuuG

COIII_genomic        ACA----G--GATTTTGG---GG--------G--GCGA---G---A----GA-G----A-
COIII_edited         ACAuuuuGuuGATTT-GguuuGGuuuuuuuuGuuGCGAuuuGuuuAuuuuGAuGuuuuAu

COIII_genomic        GTTTTTG--A-G-A---G-G-G-G-AG
COIII_edited         GT----GuuAuGuAuuuGuGuGuGuAG
```

Cyb

```
>T. vivax (Y486) Cyb mRNA (edited)
ATGTTTCGTGTTAGATTTTTGTTATTTTTTTTATTGTTTAGGAATTTGTGTTGTCTTTTAATGTCTGGTTGTTTATATAGGCTTTATGGTGTAGGA
TTTAGTTTAGGTTTTTTTATTTTATTGCAAGTGATATGTGGTTTATTTTTATCTTGATTATTTTTTAGTTGTTTTATATGCGTCAATTGATATTTT
ATATTATTTTTGTGAGATTTTGATTTAGGTTTTATTATTAGAAGTGTACATATATGTTTTACTTCTTTACTATATTTTTTATTATATGTACATATA
TTTAAGGCAATAATATTAATATTGTTATTTGATACGCATATTATAGTTTGGTTTATAGGGTTTTTAATTTTAATATTAATAATAATAATTGCATTT
ATAGGATATGTATTACCATGTACTATGATGTCATATTGAGGGTTAACAGTGTTTAGTAATATAATAGCTACAGTCCCTGTGTTTGGCAAATGGTTA
TGTTATTGAATATGAGGTAGTGAATTCATTAACGATTATACATTATTAAAATTACAAGTATTGCATATTGTTTTACCTTTTATGTTAATATTTATA
TTAATATTACATTTATTTTGTTTACATTATTTTATGAGTTCTGATGCTTTTTGTGATAGATTTGTATTTTATTGTGAAAGATTATGTTTTTGTTTG
TGATTTTATTTGAGAGATTTAGTATTATGTTTTTTAATTTTAATTTGTGTATTGTATTGTATTTTAATTAATTGGTATTTTGTTTTTCACGAAGAA
TCCTGATTAATTGTAGATGTTCTAAAAACATCTGACAAAATATTACCAGAATGATTTTTTTTATATTTATTTGGTTTTTTAAAAGCGGTGTCAGAT
AAGTTTTTTGGATTATTTTTGATGGTATTATTATTAATATCGTTATTCTTATTTATATTGAATTGTATTTTATGATTTGTATATTGTAGAAGTTCA
TTATTGTGATGCACATATGCATTTATTTTATTTTATTGTTTATGTATGAGTGGCTATTTAGCTATGTATGTTATTTTAATATATCCAATATGAATG
GAATTACAATTTTGAGTATTATTATTGTTTTTATTAATTATTTGTAGATTAGATTAA

>Cyb_edited (Translated)
MFRVRFLLFFLLFRNLCCLLMSGCLYRLYGVGFSLGFFILLQVICGLFLSWLFFSCFICVNWYFILFLWDFDLGFIIRSVHICFTSLLYFLLYVH
IFKAIILILLFDTHIIVWFIGFLILILIIIIAFIGYVLPCTMMSYWGLTVFSNIIATVPVFGKWLCYWIWGSEFINDYTLLKLQVLHIVLPFMLI
FILILHLFCLHYFMSSDAFCDRFVFYCERLCFCLWFYLRDLVLCFLILICVLYCILINWYFVFHEESWLIVDVLKTSDKILPEWFFLYLFGFLKA
VSDKFFGLFLMVLLLISLFLFILNCILWFVYCRSSLLWCTYAFILFYCLCMSGYLAMYVILIYPIWMELQFWVLLLFLLIICRLD*
```

Confirmation by aligning with *T. brucei*

```
Cyb_brucei         MFRCRFLLFFLLFRNLCCLLMSGCLYRIYGVGFSLGFFIALQIICGVCLAWLFFSCFICS 60
Cyb_edited(Trans)  MFRVRFLLFFLLFRNLCCLLMSGCLYRLYGVGFSLGFFILLQVICGLFLSWLFFSCFICV 60
                   *** **********************:********** **:***: *:*********

Cyb_brucei         NWYFVLFLWDFDLGFVIRSVHICFTSLLYLLLYIHIFKSITLIILFDTHILVWFIGFILF 120
Cyb_edited(Trans)  NWYFILFLWDFDLGFIIRSVHICFTSLLYFLLYVHIFKAIILILLFDTHIIVWFIGFLIL 120
                   ****:**********:*************:***:****:* **:******:******:::

Cyb_brucei         VFIIIIAFIGYVLPCTMMSYWGLTVFSNIIATVPILGIWLCYWIWGSEFINDFTLLKLHV 180
Cyb_edited(Trans)  ILIIIIAFIGYVLPCTMMSYWGLTVFSNIIATVPVFGKWLCYWIWGSEFINDYTLLKLQV 180
                   ::*********************************::* **************:*****:*

Cyb_brucei         LHVLLPFILLIILILHLFCLHYFMSSDAFCDRFAFYCERLSFCMWFYLRDMFLAFSILLC 240
Cyb_edited(Trans)  LHIVLPFMLIFILILHLFCLHYFMSSDAFCDRFVFYCERLCFCLWFYLRDLVLCFLILIC 240
                   **::***:*::**********************.******.**:******:.*.* **:*

Cyb_brucei         MMYVIFINWYFVFHEESWVIVDTLKTSDKILPEWFFLYLFGFLKAIPDKFMGLFLMVILL 300
Cyb_edited(Trans)  VLYCILINWYFVFHEESWLIVDVLKTSDKILPEWFFLYLFGFLKAVSDKFFGLFLMVLLL 300
                   ::* *:*************:***.**********************:.***:******:**

Cyb_brucei         FSLFLFILNCILWFVYCRSSLLWLTYSLILFYSIWMSGFLALYVVLAYPIWMELQYWVLL 360
Cyb_edited(Trans)  ISLFLFILNCILWFVYCRSSLLWCTYAFILFYCLCMSGYLAMYVILIYPIWMELQFWVLL 360
                   :********************** **::****.: ***:**:**:* ********:****

Cyb_brucei         LFLLIVCRLD 370
Cyb_edited(Trans)  LFLLIICRLD 370
                   *****:****
```

```
Alignment: Genomic vs Edited (mRNA) sequences

Cyb_genomic     A-G---CG-G--AGA-----G--A--------A--G---AGAAA---G-G--G-CTTT-A
Cyb_edited      AuGuuuCGuGuuAGAuuuuuGuuAuuuuuuuuAuuGuuuAGGAAuuuGuGuuGuCTTTuA

Cyb_genomic     ATTTGTCTTGGTTGTTTATATAGGCTTTATGGTGTAGGATTTAGTTTAGGTTTTTTTATT
Cyb_edited      A--TGTCT-GGTTGTTTATATAGGCTTTATGGTGTAGGATTTAGTTTAGGTTTTTTTATT

Cyb_genomic     TTATTGCAAGTGATATGTGGTTTATTTTTATCTTGATTATTTTTTAGTTGTTTTATATGC
Cyb_edited      TTATTGCAAGTGATATGTGGTTTATTTTTATCTTGATTATTTTTTAGTTGTTTTATATGC

Cyb_genomic     GTCAATTGATATTTTATATTATTTTTGTGAGATTTTGATTTAGGTTTTATTATTAGAAGT
Cyb_edited      GTCAATTGATATTTTATATTATTTTTGTGAGATTTTGATTTAGGTTTTATTATTAGAAGT

Cyb_genomic     GTACATATATGTTTTACTTCTTTACTATATTTTTTATTATATGTACATATATTTAAGGCA
Cyb_edited      GTACATATATGTTTTACTTCTTTACTATATTTTTTATTATATGTACATATATTTAAGGCA

Cyb_genomic     ATAATATTAATATTGTTATTTGATACGCATATTATAGTTTGGTTTATAGGGTTTTTAATT
Cyb_edited      ATAATATTAATATTGTTATTTGATACGCATATTATAGTTTGGTTTATAGGGTTTTTAATT

Cyb_genomic     TTAATATTAATAATAATAATTGCATTTATAGGATATGTATTACCATGTACTATGATGTCA
Cyb_edited      TTAATATTAATAATAATAATTGCATTTATAGGATATGTATTACCATGTACTATGATGTCA

Cyb_genomic     TATTGAGGGTTAACAGTGTTTAGTAATATAATAGCTACAGTCCCTGTGTTTGGCAAATGG
Cyb_edited      TATTGAGGGTTAACAGTGTTTAGTAATATAATAGCTACAGTCCCTGTGTTTGGCAAATGG

Cyb_genomic    TTATGTTATTGAATATGAGGTAGTGAATTCATTAACGATTATACATTATTAAAATTACAA
Cyb_edited     TTATGTTATTGAATATGAGGTAGTGAATTCATTAACGATTATACATTATTAAAATTACAA

Cyb_genomic    GTATTGCATATTGTTTTACCTTTTATGTTAATATTTATATTAATATTACATTTATTTTGT
Cyb_edited     GTATTGCATATTGTTTTACCTTTTATGTTAATATTTATATTAATATTACATTTATTTTGT

Cyb_genomic    TTACATTATTTTATGAGTTCTGATGCTTTTTGTGATAGATTTGTATTTTATTGTGAAAGA
Cyb_edited     TTACATTATTTTATGAGTTCTGATGCTTTTTGTGATAGATTTGTATTTTATTGTGAAAGA

Cyb_genomic    TTATGTTTTTGTTTGTGATTTTATTTGAGAGATTTAGTATTATGTTTTTTAATTTTAATT
Cyb_edited     TTATGTTTTTGTTTGTGATTTTATTTGAGAGATTTAGTATTATGTTTTTTAATTTTAATT

Cyb_genomic    TGTGTATTGTATTGTATTTTAATTAATTGGTATTTTGTTTTTCACGAAGAATCCTGATTA
Cyb_edited     TGTGTATTGTATTGTATTTTAATTAATTGGTATTTTGTTTTTCACGAAGAATCCTGATTA

Cyb_genomic    ATTGTAGATGTTCTAAAAACATCTGACAAAATATTACCAGAATGATTTTTTTTATATTTA
Cyb_edited     ATTGTAGATGTTCTAAAAACATCTGACAAAATATTACCAGAATGATTTTTTTTATATTTA

Cyb_genomic    TTTGGTTTTTTAAAAGCGGTGTCAGATAAGTTTTTTGGATTATTTTTGATGGTATTATTA
Cyb_edited     TTTGGTTTTTTAAAAGCGGTGTCAGATAAGTTTTTTGGATTATTTTTGATGGTATTATTA

Cyb_genomic    TTAATATCGTTATTCTTATTTATATTGAATTGTATTTTATGATTTGTATATTGTAGAAGT
Cyb_edited     TTAATATCGTTATTCTTATTTATATTGAATTGTATTTTATGATTTGTATATTGTAGAAGT

Cyb_genomic    TCATTATTGTGATGCACATATGCATTTATTTTATTTTATTGTTTATGTATGAGTGGCTAT
Cyb_edited     TCATTATTGTGATGCACATATGCATTTATTTTATTTTATTGTTTATGTATGAGTGGCTAT

Cyb_genomic    TTAGCTATGTATGTTATTTTAATATATCCAATATGAATGGAATTACAATTTTGAGTATTA
Cyb_edited     TTAGCTATGTATGTTATTTTAATATATCCAATATGAATGGAATTACAATTTTGAGTATTA

Cyb_genomic    TTATTGTTTTTATTAATTATTTGTAGATTAGATTAA
Cyb_edited     TTATTGTTTTTATTAATTATTTGTAGATTAGATTAA
```

A6

```
>T. vivax (Y486) A6 mRNA (edited)
ATGTTTTTGTTTTTTTTTTGTGATTTGTTTTGGTTGCGTTTGTTATTATGTGTGTATTATTGTGTGTGATCTAGGTTATGTTTTGTTGTGTATTTT
AATTGTTTGATGTTAATTTTTGATTTTTTGTTGTTTTGTTTGTTTGATTTGTATTTGTTTATTGGTTTATGTTTATTTTTGTTATTGTGGTTTATG
TTGTTTAATTTGTATAGTTTGATTTTGTATTATTGTATTACCTATTTGAATTTGTATTTGTTGTTTTGTATTGTTTTTTTATTGTATATGTCATTT
TTGTTTTTGTTTTGTTTTTTGTTGGATTTTTTTTTTGTTTAATAGTTTGTTGTGTGGTGATAGTTTTATGGATGTTTTTTTTTTGCGTTTTTTGTTG
TGTTTTTTAGAATGTTTTTCTTTGTTATGTCGTTGTTTGTCAACATTTTTACGTTTGTTTTGTAATTTATTGTCATCCCATTTTTTATTGTTAATG
TTTTTTGTATTTTTTTTTATTTTATTTTTTGTTTTTTTTTTTGTTTTTTTTTGTTTTGTGTATTTTATTTGGTTTATTTTGTTTGTTTGTGTTTTT
TGTTTTTGTTTGTTGTTTTATTTGTTGTTGTATTTATTGGATTTATTTGCCGCGTTATTACAGTTGTTTATTTTTTGTAATATGATTTTGCAATTG
GTAATGGATTTTTTATTGTTTTTGTTGTTTTGTTAG

>A6_edited Translated
MFLFFFCDLFWLRLLLCVYYCVWSRLCFVVYFNCLMLIFDFLLFCLFDLYLFIGLCLFLLLWFMLFNLYSLILYYCITYLNLYLLFCIVFLLYMSF
LFLFCFLLDFFLFNSLLCGDSFMDVFFLRFLLCFLECFSLLCRCLSTFLRLFCNLLSSHFLLLMFFVFFFILFFVFFFVFFCFVYFIWFILFVCVF
CFCLLFYLLLYLLDLFAALLQLFIFCNMILQLVMDFLLFLLFC*
```

Confirmation by aligning with *T. brucei*

```
A6_brucei          MFLFFFCDLFWLRLLLCMYYCVWSRLCFIVYFNCLMLIFDFLLFCLFDLYLFVGLCLFLL 60
A6_edited(Trans)   MFLFFFCDLFWLRLLLCVYYCVWSRLCFVVYFNCLMLIFDFLLFCLFDLYLFIGLCLFLL 60
                   ****************:**********:**********************:*******

A6_brucei          LWFMLFNLYSLILYYCITYLNLYLLFCIVFLLYIAFLFLFCFLCDFFLFNNLLVGDSFMD 120
A6_edited(Trans)   LWFMLFNLYSLILYYCITYLNLYLLFCIVFLLYMSFLFLFCFLLDFFLFNSLLCGDSFMD 120
                   ********************************::******** ******.** ******

A6_brucei          VFFIRFLLCFLECFSLLCRCLSTFLRLFCNLLSSHFLLLMFFDFFYFIFVFFFYGVF--- 177
A6_edited(Trans)   VFFLRFLLCFLECFSLLCRCLSTFLRLFCNLLSSHFLLLMFFVFFFILFFVFFFVFFCFV 180
                   ***:************************************ **:::*..**: .*

A6_brucei          -CYWFILFIFVFCFCLLFYVFLYLLDLFAAILQLFIFCNMILQLIMDFLLFLLFV 231
A6_edited(Trans)   YFIWFILFVCVFCFCLLFYLLLYLLDLFAALLQLFIFCNMILQLVMDFLLFLLFC 235
                      *****: *********::*********:*************:*********
```

Alignment: Genomic vs Edited (mRNA) sequences

```
A6_genomic         A-G-----G----------G-GA---GTTTTGTG--GCG---G--ATTTTTTA-GTG-G-
A6_edited          AuGuuuuuGuuuuuuuuuuGuGAuuuGTTTTG-GuuGCGuuuGuuATT----AuGTGuGu

A6_genomic         A--A--G-G-G-GA-C-AGG--A-G----G--G-G-A----AATTTTTG---GA-G--AA
A6_edited          AuuAuuGuGuGuGAuCuAGGuuAuGuuuuGuuGuGuAuuuuAATT---GuuuGAuGuuAA

A6_genomic         -----GTA------G--G----G---G---GA---G-A---G---A--GG---A-G---A
A6_edited          uuuuuG-AuuuuuuGuuGuuuuGuuuGuuuGAuuuGuAuuuGuuuAuuGGuuuAuGuuuA

A6_genomic         -----G--A--G-GG---A-G--G---ATTTTA---G-A-AG---GATTT-G-A--ATTT
A6_edited          uuuuuGuuAuuGuGGuuuAuGuuGuuuA----AuuuGuAuAGuuuGATTTuGuAuuATT-

A6_genomic         G-A--ACCTTTTTA---GTAA---G-ATTTG--GTTT-G-A--G-------A--G-ATTT
A6_edited          GuAuuACCT----AuuuG-AAuuuGuATTTGuuGTTTuGuAuuGuuuuuuuAuuGuAT--

A6_genomic         TTTA-G-CA-----G-----G----G------G--GGA---------GTTTAA-AG---G
A6_edited          ---AuGuCAuuuuuGuuuuuGuuuuGuuuuuuGuuGGAuuuuuuuuuGTTTAAuAGuuuG
```

```
A6_genomic        --GTG-GG-GA-AG----A-GGA-G----------GTCG------G--G-G------AGA
A6_edited         uuGTGuGGuGAuAGuuuuAuGGAuGuuuuuuuuuuG-CGuuuuuuGuuGuGuuuuuuAGA

A6_genomic        A-G-----C---G--A-GTCTTTTG--G---G-CAACA-----ACG---GTTTTTG-AA-
A6_edited         AuGuuuuuCuuuGuuAuGTC----GuuGuuuGuCAACAuuuuuACGuuuGTTTT-GuAAu

A6_genomic        --A--G-CA-CCCATTTT--A--G--AA-G------G-AT--------A----A------
A6_edited         uuAuuGuCAuCCCATTTTuuAuuGuuAAuGuuuuuuGuATuuuuuuuuuAuuuuAuuuuuu

A6_genomic        G-----------G---------G----G-G-A----A---GG---A----G---G---GT
A6_ed             GuuuuuuuuuuuGuuuuuuuuuuGuuuuGuGuAuuuuAuuuGGuuuAuuuuGuuuGuuuGT

A6_genomic        G------G-----G---G--G----AT--G--G--G-A---A--GGA---A---GCCTTT
A6_edited         GuuuuuuGuuuuuGuuuGuuGuuuuATuuGuuGuuGuAuuuAuuGGAuuuAuuuGCC---

A6_genomic        TTGCTTG--A--ACAG--G---A------G-AA-A-GA----GCAA--GG-AA-GTTGTT
A6_edited         --GC--GuuAuuACAGuuGuuuAuuuuuuGuAAuAuGAuuuuGCAAuuGGuAAuG--G--

A6_genomic        A------A--G-----G--G----G--AG
A6_edited         AuuuuuuAuuGuuuuuGuuGuuuuGuuAG
```

CR3

```
>T.vivax (Y486) CR3 mRNA (edited)
ATGTTTGATTGTTTAGTTTTGTTGTTTTTTTTATTGTTGTTTGTACATTTTTTTTGTTTTTTGTTTATTTGTGATTTGTTTTTATGTTTGTTGTTT
GTTTTTTGTTTATTTGTGGATTTTTGTTTTTTGTTTAATATGGGTTTGTTGTTGTGTTTATTTTTTTTTTTTGTTTTATCATTTGATATGTTATTA
TCGTTTTTGTTATTATATATAAGTTTTCGTTATTAA

>CR3_edited Translated
MFDCLVLLFFLLLFVHFFCFLFICDLFLCLLFVFCLFVDFCFLFNMGLLLCLFFFFVLSFDMLLSFLLLYISFRY*
```

Confirmation by aligning with *T. brucei*

```
CR3_brucei          MFDCLVLLFFYCLFVHFFCFLFVCDLFLCLLFSFCFLLDFCFLFNMGLLLCLFFFFILSF 60
CR3_edited(Trans)   MFDCLVLLFFLLLFVHFFCFLFICDLFLCLLFVFCLFVDFCFLFNMGLLLCLFFFFVLSF 60
                    **********  **********:********* **:::******************:***
CR3_brucei          DMCIKLLFII----- 70
CR3_edited(Trans)   DMLLSFLLLYISFRY 75
                    ** :.:*::
```

Alignment: Genomic vs Edited (mRNA) sequences

```
CR3_genomic          A-G---GATTTG---AG----GTTTTTG--------A--G--G---G-ACA--------G
CR3_edited           AuGuuuGATT-GuuuAGuuuuGTT---Guuuuuuuu AuuGuuGuuuGuACAuuuuuuuuG

CR3_genomic          -TTTTTG---A---G-GTTTTTTTTA---G-----A-G---G--G---G------G--TA
CR3_edited           uTTTTTGuuuAuuuGuG--------AuuuGuuuuuAuGuuuGuuGuuuGuuuuuuGuuTA

CR3_genomic          ---G-GGA-----G------G---AATTTTTATTTGGG---GTTTTTG--GTG---A---
CR3_edited           uuuGuGGAuuuuuGuuuuuuGuuuAAT----AT--GGGuuuGTT---GuuGTGuuuAuuu

CR3_genomic          ---------G----A-CA---GA-A-G--A--A-CG-TTTTG--A--A-A-A-AAG-TTT
CR3_edited           uuuuuuuuuGuuuuAuCAuuuGAuAuGuuAuuAuCGuTTTTGuuAuuAuAuAuAAGuTTT

CR3_genomic          CGTTATTAA
CR3_edited           CGTTATTAA
```

COII

```
>T.vivax (Y486) COII mRNA (edited)
ATGAGTTTTATATTATCCTTTTGATTATTATTTTTAATAGATTCTGTAATAGTTTTATTATCTTTATCTTTTTTTCTCGTTTTATGAATATGTATT
TTATTAATTTCTTCTATATTATGTATTATAAAGATAAATATAGTTTATTGTTCATGAGATTTTATTGCCTCAAAATTTTTAGATGTTTATTGATTT
GTTGTTGGAACTATGTTTATGCTATGTTTATTAATGCGTTTATGCATATTATTATATTTTGGATGTCTCAATTTTGTCAGTTTTGATATATGTAAA
GTTATAGGATTCCAATGGTATTGAGTTTATTTTTTATTTGGTGACACAACTATTTTTAGTAATTTAATATTAGAAAGTGATTACATTTTAGGAGAT
TTAAGAATATTGCAATGTAATCATGTGCTAACTTTATTAAGTTTAGTAATTTATAAATTATGATTATCTGCAGTAGATGTAATACATTCATTCTCA
TTATCAAGTTTAGGTATAAAAGTAGATTGTATACCTGGTAGATGTAATGAAATAATTTTATATGCATCAAATGCAGCAACTATTTATGGTCAGTGT
AGTGAATTATGTGGTGTATTACATGGATTTATGCCTATTGTAATAAATTTTATATAA

>COII_edited Translated
MSFILSFWLLFLIDSVIVLLSLSFFLVLWICILLISSILCIIKINIVYCSWDFIASKFLDVYWFVVGTMFMLCLLMRLCILLYFGCLNFVSFDICK
VIGFQWYWVYFLFGDTTIFSNLILESDYILGDLRILQCNHVLTLLSLVIYKLWLSAVDVIHSFSLSSLGIKVDCIPGRCNEIILYASNAATIYGQC
SELCGVLHGFMPIVINFI*
```

Confirmation by aligning with *T. brucei*

```
COII_brucei         MSFILTFWMIFLMDSIIVLISFSIFLSVWICALIIATVLTVTKINNIYCTWDFISSKFID 60
COII_edited(Trans)  MSFILSFWLLFLIDSVIVLLSLSFFLVLWICILLISSILCIIKINIVYCSWDFIASKFLD 60
                    *****:**::**:**:***:*:*:** :*** *:*:::* : *** :**:****:***:*

COII_brucei         TYWFVLGMMFILCLLLRLCLLLYFSCINFVSFDLCKVIGFQWYWVYFLFGETTIFSNLIL 120
COII_edited(Trans)  VYWFVVGTMFMLCLLMRLCILLYFGCLNFVSFDICKVIGFQWYWVYFLFGDTTIFSNLIL 120
                    .****:* **:****:***:****.*:******:****************:*********

COII_brucei         ESDYLIGDLRILQCNHVLTLLSLVIYKLWVSAVDVIHSFTISSLGIKVDCIPGRCNEIIL 180
COII_edited(Trans)  ESDYILGDLRILQCNHVLTLLSLVIYKLWLSAVDVIHSFSLSSLGIKVDCIPGRCNEIIL 180
                    ****::**********************:*********::******************

COII_brucei         FATNNATLYGQCSELCGVLHGFMPIVINFI 210
COII_edited(Trans)  YASNAATIYGQCSELCGVLHGFMPIVINFI 210
                    :*:* **:**********************
```

Alignment: Genomic vs Edited (mRNA) sequences

```
COII_genomic        ATGAGTTTTATATTATCCTTTTGATTATTATTTTTAATAGATTCTGTAATAGTTTTATTA
COII_edited         ATGAGTTTTATATTATCCTTTTGATTATTATTTTTAATAGATTCTGTAATAGTTTTATTA

COII_genomic        TCTTTATCTTTTTTTCTCGTTTTATGAATATGTATTTTATTAATTTCTTCTATATTATGT
COII_edited         TCTTTATCTTTTTTTCTCGTTTTATGAATATGTATTTTATTAATTTCTTCTATATTATGT

COII_genomic        ATTATAAAGATAAATATAGTTTATTGTTCATGAGATTTTATTGCCTCAAAATTTTTAGAT
COII_edited         ATTATAAAGATAAATATAGTTTATTGTTCATGAGATTTTATTGCCTCAAAATTTTTAGAT

COII_genomic        GTTTATTGATTTGTTGTTGGAACTATGTTTATGCTATGTTTATTAATGCGTTTATGCATA
COII_edited         GTTTATTGATTTGTTGTTGGAACTATGTTTATGCTATGTTTATTAATGCGTTTATGCATA

COII_genomic        TTATTATATTTTGGATGTCTCAATTTTGTCAGTTTTGATATATGTAAAGTTATAGGATTC
COII_edited         TTATTATATTTTGGATGTCTCAATTTTGTCAGTTTTGATATATGTAAAGTTATAGGATTC

COII_genomic        CAATGGTATTGAGTTTATTTTTTATTTGGTGACACAACTATTTTTAGTAATTTAATATTA
COII_edited         CAATGGTATTGAGTTTATTTTTTATTTGGTGACACAACTATTTTTAGTAATTTAATATTA

COII_genomic        GAAAGTGATTACATTTTAGGAGATTTAAGAATATTGCAATGTAATCATGTGCTAACTTTA
COII_edited         GAAAGTGATTACATTTTAGGAGATTTAAGAATATTGCAATGTAATCATGTGCTAACTTTA


COII_genomic        TTAAGTTTAGTAATTTATAAATTATGATTATCTGCAGTAGATGTAATACATTCATTCTCA
```

```
COII_edited          TTAAGTTTAGTAATTTATAAATTATGATTATCTGCAGTAGATGTAATACATTCATTCTCA

COII_genomic         TTATCAAGTTTAGGTATAAAAGTAGA--G-A-ACCTGGTAGATGTAATGAAATAATTTTA
COII_edited          TTATCAAGTTTAGGTATAAAAGTAGAuuGuAuACCTGGTAGATGTAATGAAATAATTTTA

COII_genomic         TATGCATCAAATGCAGCAACTATTTATGGTCAGTGTAGTGAATTATGTGGTGTATTACAT
COII_edited          TATGCATCAAATGCAGCAACTATTTATGGTCAGTGTAGTGAATTATGTGGTGTATTACAT

COII_genomic         GGATTTATGCCTATTGTAATAAATTTTATATAA
COII_edited          GGATTTATGCCTATTGTAATAAATTTTATATAA
```

MURF2

```
>T.vivax (Y486) MURF2 mRNA (edited)
ATGTTTGGTTGTTTTAATTTAGTTTTATTTTTGTGCTTTGATTGTAGCCGTATTTTTGATTTATTATGTATTAGAACATATGATTTTATATTATGA
TGATTTGATTTAGATTTCATTTTATACGATTTCGTTTTTGATTTTGTTGTTTGCATAACTTTTATCTTTATTTTTGTGTTGGGTTTTTTTTTGAGG
ATTTTTTTCAGTTTTGTATTTGTCTTATTATTTATAACATTTTTTGGTTTTTTTGCAACAGCTTTATTATATTCTGGATATTATTTATTTTATATA
TATATATTATATAATTTTATCTGTTTTTTTTTTGTATTTGGTATTAATTTTATTTTATTTTACCTCGAATTTTTTACATATATAAATCTTTTTATA
TTTTTTGATTTCGTTAGTTTTTCAAGTTATTTATATAATTTTTTTGGATTAGTTATATTATTTAATATTATATTTTGTCATTATTTGTTTTGTTTA
TTTTATTTTGTAATTCATTTTTTATTTTGTTTTATATTTTTTGTTATTCGTTGTTTATTTATAATAATTACAGATTTTTTATTTTTTAATTTTGAT
ATATTTGTTTCTATTTTATTGTGTGATATGTCATTTTTAGATTTTATTAGTTTATTATTATTATATTTTAATTTTATATTTAATTATATGTATGGT
TTTATTAGTTTTATTATTATTTTAGGTTTTTTATTTTTATTATTATTTTTTGTTATTAATTTATTTTTTGGTTTTATATTTTTGGTTTATGGATTT
CATTTGCATTTATTAAATTATATTTTTTGATTATATATTATTTACAGTAAAAGTTGTTTTGTATTATTACCAGCTGTATTAATATTTTTTAAATTT
ATGTATTTCGATGTTTTTTTTGTATTTATTTTTGTATTAATATTATTTATAATATGTTTTTTTAGTTTTTTCTTAAAAGATTTTTTATTTTTATCT
TTATTTTTTGATATATTTACATCATTATATAGTTATGATATATATTCATATATTGCATTTTATGCAAATATTCAATATTTTAGTATAACGCAGTTA
TTATTTATTTATATGTAA

>MURF2_edited Translated
MFGCFNLVLFLCFDCSRIFDLLCIRTYDFILWWFDLDFILYDFVFDFVVCITFIFIFVLGFFLRIFFSFVFVLLFITFFGFFATALLYSGYYLFYI
YILYNFICFFFVFGINFILFYLEFFTYINLFIFFDFVSFSSYLYNFFGLVILFNIIFCHYLFCLFYFVIHFLFCFIFFVIRCLFIIITDFLFFNFD
IFVSILLCDMSFLDFISLLLLYFNFIFNYMYGFISFIIILGFLFLLLFFVINLFFGFIFLVYGFHLHLLNYIFWLYIIYSKSCFVLLPAVLIFFKF
MYFDVFFVFIFVLILFIICFFSFFLKDFLFLSLFFDIFTSLYSYDIYSYIAFYANIQYFSITQLLFIYM*
```

Confirmation by aligning with *T. brucei*

```
MURF2_brucei         MFGCFNLVLFLCFDCSRVFDLLCIRTYDFILWWFDLDFILYDFVFDFVVCITFIFIFVLG 60
MURF2_edited(Trans)  MFGCFNLVLFLCFDCSRIFDLLCIRTYDFILWWFDLDFILYDFVFDFVVCITFIFIFVLG 60
                     *****************:****************************************

MURF2_brucei         FFIRIFFSFVFVLLFITFFGICSLTMLFTGYYIYYIYILYNFICFFFAFGINFLIYYIEF 120
MURF2_edited(Trans)  FFLRIFFSFVFVLLFITFFGFFATALLYSGYYLFYIYILYNFICFFFVFGINFILFYLEF 120
                     **:*****************: : ::*::***::*************.*****:::*:**

MURF2_brucei         FIFITFHIFFDFISFSNYIYNYFGILYMFNVMFCAYLFCLFYFVIYFLFCFIFFVIRCLF 180
MURF2_edited(Trans)  FTYINLFIFFDFVSFSSYLYNFFGLVILFNIIFCHYLFCLFYFVIHFLFCFIFFVIRCLF 180
                     * :*.:.*****:***.*:**:**:: :**::** **********:**************

MURF2_brucei         IVIMDFLFFNFDIFVSILLCDIVYLDFISLLLLYFNFIFNFIYGFFSFVIILGLLFLLLF 240
MURF2_edited(Trans)  IIITDFLFFNFDIFVSILLCDMSFLDFISLLLLYFNFIFNYMYGFISFIIILGFLFLLLF 240
                     *:* *****************: :****************::***:**:****:******

MURF2_brucei         LVINLFFGFTFLVYGIQIILLYYVYWLYMIYSRSCYILMPAILIFFKFIYFDVFFVFVFI 300
MURF2_edited(Trans)  FVINLFFGFIFLVYGFHLHLLNYIFWLYIIYSKSCFVLLPAVLIFFKFMYFDVFFVFIFV 300
                     :******** *****::: ** *::***:***:**::*:**:******:********:*:

MURF2_brucei         LILFIISFFSFFLKDFLFLSLYFDIFGSLYNYDILSYSIFYYQNNQ-FCLTQLLSIYI 357
MURF2_edited(Trans)  LILFIICFFSFFLKDFLFLSLFFDIFTSLYSYDIYSY-IAFYANIQYFSITQLLFIYM 357
                     ******.**************:**** ***.*** ** * :* * * *.:**** **:
```

Alignment: Genomic vs Edited (mRNA) sequences

```
MURF2_genomic       A-G---GG--G----ATTTTA---AG----A-----G-GCTTTGATTG-AGCCGTATTTT
MURF2_edited        AuGuuuGGuuGuuuuA----AuuuAGuuuuAuuuuuGuGCTTTGATTGuAGCCGTATTTT

MURF2_genomic       TGATTTATTATGTATTAGAACATATGATTTTATATTATGATGATTTGATTTAGATTTCAT
MURF2_edited        TGATTTATTATGTATTAGAACATATGATTTTATATTATGATGATTTGATTTAGATTTCAT

MURF2_genomic       TTTATACGATTTCGTTTTTGATTTTGTTGTTTGCATAACTTTTATCTTTATTTTTGTGTT
MURF2_edited        TTTATACGATTTCGTTTTTGATTTTGTTGTTTGCATAACTTTTATCTTTATTTTTGTGTT

MURF2_genomic       GGGTTTTTTTTTGAGGATTTTTTTCAGTTTTGTATTTGTCTTATTATTTATAACATTTTT
MURF2_edited        GGGTTTTTTTTTGAGGATTTTTTTCAGTTTTGTATTTGTCTTATTATTTATAACATTTTT

MURF2_genomic       TGGTTTTTTTGCAACAGCTTTATTATATTCTGGATATTATTTATTTTATATATATATATT
MURF2_edited        TGGTTTTTTTGCAACAGCTTTATTATATTCTGGATATTATTTATTTTATATATATATATT

MURF2_genomic       ATATAATTTTATCTGTTTTTTTTTTTGTATTTGGTATTAATTTTATTTTATTTTACCTCGA
MURF2_edited        ATATAATTTTATCTGTTTTTTTTTTTGTATTTGGTATTAATTTTATTTTATTTTACCTCGA

MURF2_genomic       ATTTTTTACATATATAAATCTTTTTATATTTTTTGATTTCGTTAGTTTTTCAAGTTATTT
MURF2_edited        ATTTTTTACATATATAAATCTTTTTATATTTTTTGATTTCGTTAGTTTTTCAAGTTATTT

MURF2_genomic       ATATAATTTTTTTTGGATTAGTTATATTATTTAATATTATATTTTGTCATTATTTGTTTTG
MURF2_edited        ATATAATTTTTTTTGGATTAGTTATATTATTTAATATTATATTTTGTCATTATTTGTTTTG

MURF2_genomic       TTTATTTTATTTTGTAATTCATTTTTTATTTTGTTTTATATTTTTTGTTATTCGTTGTTT
MURF2_edited        TTTATTTTATTTTGTAATTCATTTTTTATTTTGTTTTATATTTTTTGTTATTCGTTGTTT

MURF2_genomic       ATTTATAATAATTACAGATTTTTTATTTTTTAATTTTGATATATTTGTTTCTATTTTATT
MURF2_edited        ATTTATAATAATTACAGATTTTTTATTTTTTAATTTTGATATATTTGTTTCTATTTTATT

MURF2_genomic       GTGTGATATGTCATTTTTAGATTTTATTAGTTTATTATTATTATATTTTAATTTTATATT
MURF2_edited        GTGTGATATGTCATTTTTAGATTTTATTAGTTTATTATTATTATATTTTAATTTTATATT

MURF2_genomic       TAATTATATGTATGGTTTTATTAGTTTTATTATTATTTTAGGTTTTTTATTTTTATTATT
MURF2_edited        TAATTATATGTATGGTTTTATTAGTTTTATTATTATTTTAGGTTTTTTATTTTTATTATT

MURF2_genomic       ATTTTTTGTTATTAATTTATTTTTTGGTTTTATATTTTTGGTTTATGGATTTCATTTGCA
MURF2_edited        ATTTTTTGTTATTAATTTATTTTTTGGTTTTATATTTTTGGTTTATGGATTTCATTTGCA

MURF2_genomic       TTTATTAAATTATATTTTTTGATTATATATTATTTACAGTAAAAGTTGTTTTGTATTATT
MURF2_edited        TTTATTAAATTATATTTTTTGATTATATATTATTTACAGTAAAAGTTGTTTTGTATTATT

MURF2_genomic       ACCAGCTGTATTAATATTTTTTAAATTTATGTATTTCGATGTTTTTTTTGTATTTATTTT
MURF2_edited        ACCAGCTGTATTAATATTTTTTAAATTTATGTATTTCGATGTTTTTTTTGTATTTATTTT

MURF2_genomic       TGTATTAATATTATTTATAATATGTTTTTTTAGTTTTTTCTTAAAAGATTTTTTATTTTT
MURF2_edited        TGTATTAATATTATTTATAATATGTTTTTTTAGTTTTTTCTTAAAAGATTTTTTATTTTT

MURF2_genomic       ATCTTTATTTTTTGATATATTTACATCATTATATAGTTATGATATATATTCATATATTGC
MURF2_edited        ATCTTTATTTTTTGATATATTTACATCATTATATAGTTATGATATATATTCATATATTGC

MURF2_genomi        ATTTTATGCAAATATTCAATATTTTAGTATAACGCAGTTATTATTTATTTATATGTAA
MURF2_edited        ATTTTATGCAAATATTCAATATTTTAGTATAACGCAGTTATTATTTATTTATATGTAA
```

CR4

```
>T.vivax (Y486) CR4 mRNA (edited)
ATTTTTTTGTTTGTTGTGCATTTTTTTTTTTTTTATTTGGTGTGTTTGTGTTTTATGTATAGTTTGTGGTATATTTTGTTGTTGTTTTTTTTTTTG
TTTTTTTTGTTTGTGTGTTTGGGTATGTTTTATTTGTTTTGTTATAGTTGTTTGTTTTTTTTTTTTGTTTTGGGTTGTGATTTTTTGTTAGTTTAT
TGATTGTATAGTTTGTTTTTTTTGTGACGTTATAATTTTGTTTATTTTTTTTTTTTGTTTTGTTTTGTGTTTTTTGTTTTGTTGTTTTTTTTTGGT
TTGTTTGGTTTTTTTTTGTATTTTTTGTTATGTTTTGTGTTATTTTTTGATTTGTTTTTTATGTTGTTTTTTGTATTGGGTGGTTTTTTTGTTTTT
GTTTTTTTTTTTTGTTTATGTTTATTGTTTTTTGTAGTTATTGTTGTTTTATTAGTTTGGTTGTTGTTGTTATTTGTGTATAGGTTTATTTATATG
CGTTTTGTGTTTTAA

>CR4_edited Translated
IFLFVVHFFFFYLVCLCFMYSLWYILLLFFFLFFLFVCLGMFYLFCYSCLFFFFVLGCDFLLVYWLYSLFFLWRYNFVYFFFLFCFVFFVLLFFFG
LFGFFLYFLLCFVLFFDLFFMLFFVLGGFFVFVFFFCLCLLFFVVIVVLLVWLLLLFVYRFIYMRFVF*
```

Confirmation by aligning with *T. brucei*

```
CR4_brucei          NLLSLCIYYYFILILVVHFFFFYLVCLCFMYSLWYILLLFCFLFLLFVCVGMFYLFCYSC 60
CR4_edited(Trans)   -----------IFLFVVHFFFFYLVCLCFMYSLWYILLLFFFLFFLFVCLGMFYLFCYSC 49
                              *:::************************ ***:****:**********

CR4_brucei          LFFFVVLGCDFLLVFWLYSLFFLWRYNFVYFFFLFCFVFFVLLFLFGLFGFFLYFLLCFV 120
CR4_edited(Trans)   LFFFFVLGCDFLLVYWLYSLFFLWRYNFVYFFFLFCFVFFVLLFFFGLFGFFLYFLLCFV 109
                    ****.*********:*****************************:***************

CR4_brucei          LFFDLFFMLFFVLGGFFVFVFFFCLCLFLFVVVVILLVWLLLLFVYRFIYMRFLF 175
CR4_edited(Trans)   LFFDLFFMLFFVLGGFFVFVFFFCLCLLFFVVIVVLLVWLLLLFVYRFIYMRFVF 164
                    **************************::***:*:*****************:*
```

Alignment: Genomic vs Edited (mRNA) sequences

```
CR4_genomic          ATT-----G---G--G-GCTTTTTA--------------A---GG-G-G---G-G----A
CR4_edited           ATTuuuuuGuuuGuuGuGC-----AuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuA

CR4_genomic          -G-A-ATTTTTG---G-GG-A-A----GT-G--G-----------G--------G---GT
CR4_edited           uGuAuA-----GuuuGuGGuAuAuuuuGTuGuuGuuuuuuuuuuuuGuuuuuuuuGuuuGT

CR4_genomic          TTTG-G---GGG-ATTTG----A---G----G--A-ATTTTG--G---G-----------
CR4_edited           ---GuGuuuGGGuAT--GuuuuAuuuGuuuuGuuAuA----GuuGuuuGuuuuuuuuuuu

CR4_genomic          -G----GGG--G-GATTTT--G--AG---AT-GA--G-A-ATTTGT--G--------G-G
CR4_edited           uGuuuuGGGuuGuGATTTTuuGuuAGuuuATuGAuuGuAuA---GTuuGuuuuuuuuGuG

CR4_genomic          ACTG--A-AAT---G---AT-----------GT---GTT--G-G------G----G--G-
CR4_edited           AC-GuuAuAATuuuGuuuATTTTuuuuuuuuGTuuuGTTuuGuGuuuuuuGuuuuGuuGu

CR4_genomic          --------GGTTTTG---GG---------G-A------G--A-GT---G-G--A------
CR4_edited           uuuuuuuuGGTTT-GuuuGGuuuuuuuuuGuAuuuuuuGuuAuGTuuuGuGuuAuuuuuu

CR4_genomic          GA---G------A-G--G------G-A--GGGTGG-------G-----G-----------
CR4_edited           GAuuuGuuuuuuAuGuuGuuuuuuGuAuuGGGTGGuuuuuuuGuuuuuGuuuuuuuuuuu

CR_genomic           -G---ATG---A--G------G-AG--ATTGTTTTG----A--AG---GG--G--G--GT
CR4_edited           uGuuuATGuuuAuuGuuuuuuGuAGuuATTGTT--GuuuuAuuAGuuuGGuuGuuGuuGT

CR4_genomic          TTA---G-GTTTTATTTATTGG---A---A-A-GCG----G-GTTTTTTAA
CR4_edited           T-AuuuGuGT---ATuuATTGGuuuAuuuAuAuGCGuuuuGuGTTTT--AA
```

ND3

```
>T.vivax (Y486) ND3 mRNA (edited)
ATGTTATCAATTTTTTTGTATTTGGTTTTGTGTTTGTGGTTTATTATTTTATTGTTAGGTTTTTTTCTATGTTTTTTATGTTTTTTGTTACATTTTT
TTGATTGTTTTCGTTGTTGTTTATGATTTTCATGTGGTTTGTATGATATGAATTCACGTTTGGTGTTTTATACATTAGATTTATGTTTTGTTAGTTG
TTTGTTTTTTGTGTTATTGAATTCTGTTATTTGTGTTTTGTTATTTGTATTTGTGTTGGTATTGTTTTATTTTTGTTATGGTTTTTTATTTTTGTGG
TTTTTGTTTTTTGTGTTGTTATTAGGGTTTGTGTGGTATTTTTGAGATCATGTATATT

>ND3_edited Translated
MLSIFLYLVLCLWFIILLLGFFLCFLCFLLHFFDCFRCCLWFSCGLYDMNSRLVFYTLDLCFVSCLFFVLLNSVICVLLFVFVLVLFYFCYGFLFLW
FLFFVLLLGFVWYFWDHVY
```

Confirmation by aligning with *T. brucei*

```
ND3_brucei          LLSLFLYLFLILWFIILFIGFFLCFLCFLLHFFDCFRCCLWFSCGLYDMNSRLVFYTLDL 60
ND3_edited(Trans)   MLSIFLYLVLCLWFIILLLGFFLCFLCFLLHFFDCFRCCLWFSCGLYDMNSRLVFYTLDL 60
                    :**:****.* ******::****************************************

ND3_brucei          CFVSCLFFVLLNSIICVLLFVFVIVLFYFCYGFLFLWFLFFVVCIGFVWYFWDHVY 116
ND3_edited(Trans)   CFVSCLFFVLLNSVICVLLFVFVLVLFYFCYGFLFLWFLFFVLLLGFVWYFWDHVY 116
                    ************:*********:******************: :***********
```

Alignment: Genomic vs Edited (mRNA) sequences

```
ND3_genomic          ATGTTA-CAAT------G-A---GGTTT-G-G---G-GG---AT-A----A--G--AGG-
ND3_edited           ATGTTAuCAATuuuuuuGuAuuuGGTTTuGuGuuuGuGGuuuATuAuuuuAuuGuuAGGu

ND3_genomic          ------CTTTA-G------A-G------G--ACA-------GTTTTA--G----CG--G-
ND3_edited           uuuuuuCT--AuGuuuuuuAuGuuuuuuGuuACAuuuuuuuG----AuuGuuuuCGuuGu

ND3_genomic          -G---A-GA----CA-G-GTTTG---G-ATTTTGA-A-GTTATTTTTA--CACG---GG-
ND3_edited           uGuuuAuGAuuuuCAuGuG---GuuuGuAT---GAuAuG--A-----AuuCACGuuuGGu

ND3_genomic          G----ATTTACA--AGA---A-G----G--ATTTTTTTTTTTTTG--G---GT-----G-
ND3_edited           GuuuuAT--ACAuuAGAuuuAuGuuuuGuuA-------------GuuGuuuGTuuuuuGu

ND3_genomic          G--ATTTTTGAAT-C-G--A---GTTTTTG----G--ATTTG-A---GTTTTTTTTTTG-
ND3_edited           GuuATT---GAATuCuGuuAuuuGT----GuuuuGuuATTTGuAuuuGT---------Gu

ND3_genomic          -GG-A--G----A-----GTTATTTTTGG------ATTTTTTG-GG-----G------G-
ND3_edited           uGGuAuuGuuuuAuuuuuGTTAT----GGuuuuuuATTTTT-GuGGuuuuuGuuuuuuGu

ND3_genomic          G--G--ATTTTAGGG---G-GTGG-A-----GAGATTCA-G-ATATT
ND3_edited           GuuGuuATT--AGGGuuuGuGTGGuAuuuuuGAGAT-CAuGuATATT
```

RPS12

```
>T.vivax (Y486) RPS12 mRNA (edited)
ATGTGATTTTTGTATGGTTGTTGTTTGCGTTTTGTTTTGTTTGTTTTATGTTATTATATGAGTCCCCGATTGCCCAGTTCCGGTAATCGA
CGTGTGTTGTATGCCGTGTTTTATTTGTATAATTTTGTGTGGTTGTTGCGTTGTTTTTTTTGTTGTGTGTTTTTTGGTTTGCATTTGTCG
TTATTTATTATAGAGGGTGGTGGTTTTGTTGATTTACCCGGTATAAAGTATTATACACGTATGTTTATTAATTAA

>RPS12_edited Translated
MWFLYGCCLRFVLFVLCYYMSPRLPSSGNRRVLYAVFYLYNFVWLLRCFFCCVFFGLHLSLFIIEGGGFVDLPGIKYYTRMFIN*
```

Confirmation by aligning with *T. brucei*

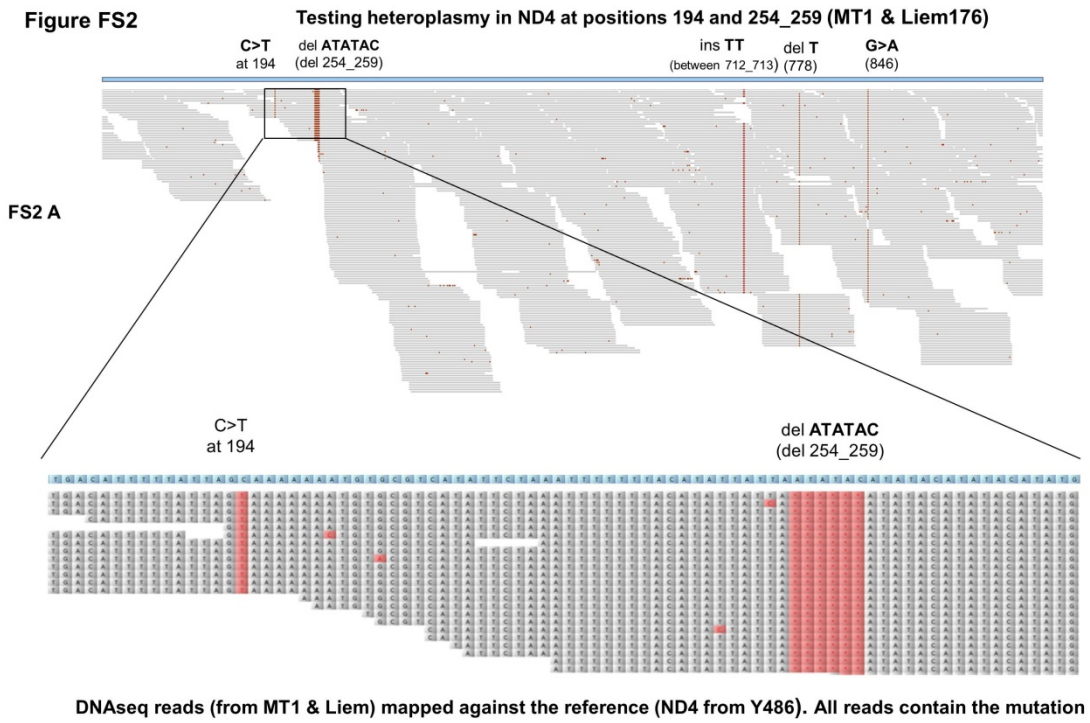
```
RPS12_brucei        MWFLYGCCLRFVLFVLCYYMSPRLPSSGNRRVLYAVFYLYNFVWMLRCFFCC-FIGLVMS 59
RPS12_edited(Trans) MWFLYGCCLRFVLFVLCYYMSPRLPSSGNRRVLYAVFYLYNFVWLLRCFFCCVFFGLHLS 60
                    *******************************************:******* *:** :*

RPS12_brucei        LFIIEGGGFVDLPGVKYYTRIVS- 82
RPS12_edited(Trans) LFIIEGGGFVDLPGIKYYTRMFIN 84
                    **************:*****:.
```

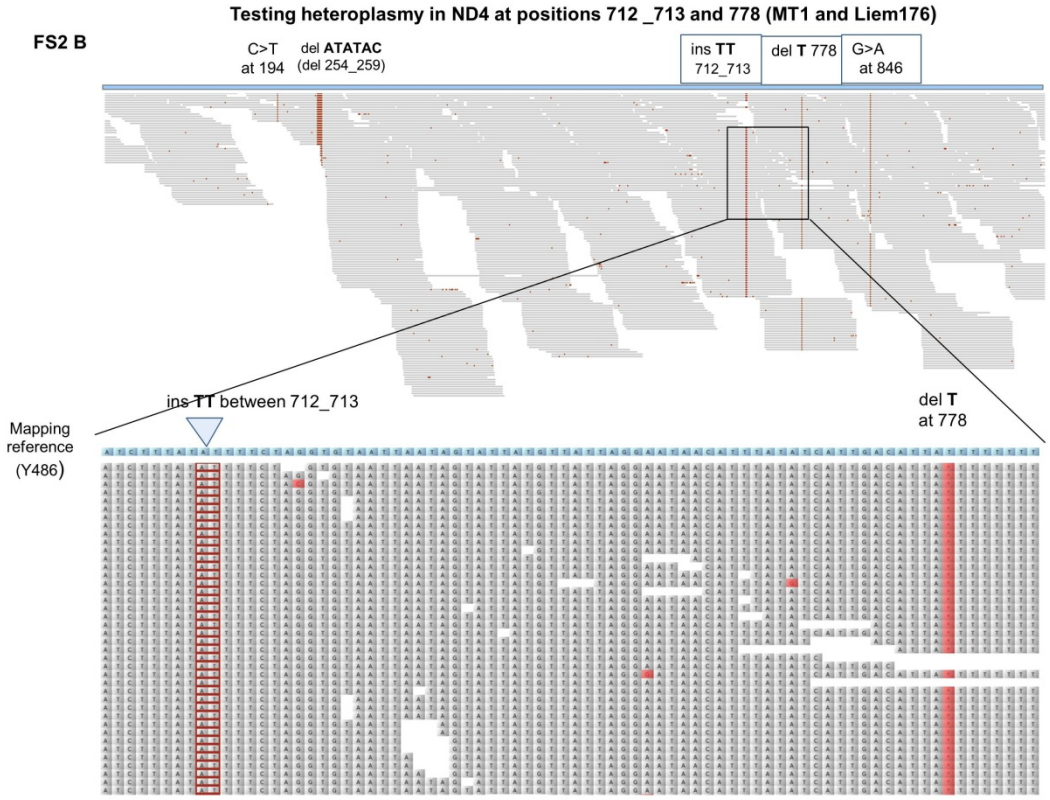
Alignment: Genomic vs Edited (mRNA) sequences

```
RPS12_genomic       A-G-GATTTTTTG-A-GG--G--GTTTGCTTTTTG----G----G---G----A-G--A-
RPS12_edited        AuGuGATTTTT-GuAuGGuuGuuGTTTGC-----GuuuuGuuuuGuuuGuuuuAuGuuAu

RPS12_genomic       -A-A-GAGTCCTTTTCCGA--GCCCAG--CCGG-AA-CGACG-G-G--G-A-GCCTTG-G
RPS12_edited        uAuAuGAGTCC----CCGAuuGCCCAGuuCCGGuAAuCGACGuGuGuuGuAuGCC--GuG

RPS12_genomic       ----A---G-A-AA----G-G-GG--G--GCG--G--------G--G-G-GTTTT--GG-
RPS12_edited        uuuuAuuuGuAuAAuuuuGuGuGGuuGuuGCGuuGuuuuuuuuGuuGuGuGTTTTuuGGu

RPS12_genomic       --GCATTTGTCG--A---A--A-AGATTTTTGTTTTGG-GG-GG----G--GA---ACCC
RPS12_edited        uuGCATTTGTCGuuAuuuAuuAuAGA-----G----GguGGuGGuuuuGuuGAuuuACCC

RPS12_genomic       TTTTTGTTG-ATAAAG-A--A-ACACG-A-G---A--AA--AA
RPS12_edited        -----G--GuATAAAGuAuuAuACACGuAuGuuuAuuAAuuAA
```

Figura Suplementaria 2. Análisis de heteroplasmia (MT1 y Liem-176)

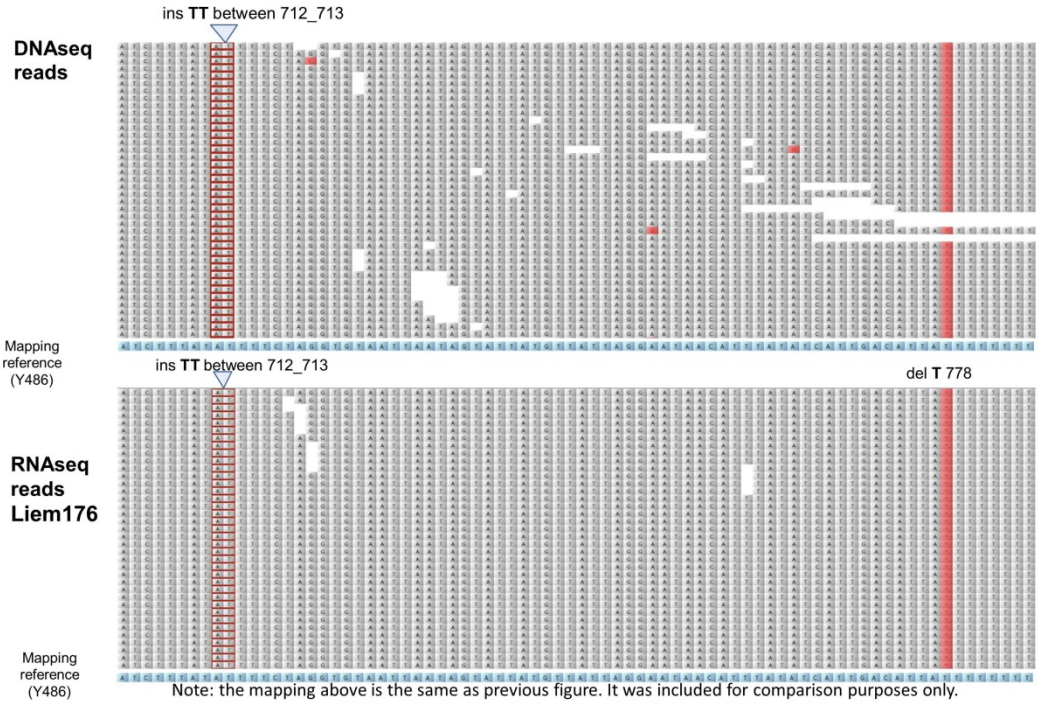
Búsqueda de heteroplasmia en diferentes regiones del maxicírculo de MT1 o Liem-176 (A-D).





DNAseq reads (from MT1 and Liem) mapped against the reference (ND4 from Y486). All reads contain the mutation

FS2 C Testing whether **TT** insertion between positions 712_713 and deletion of **T** 778 of ND4 represent novel editing sites. Analysis conducted using RNAseq reads mapped against the genomic sequence fromY486



Note: the mapping above is the same as previous figure. It was included for comparison purposes only. As the figure shows, all RNAseq reads have the insertion of **TT** (between 712 and 713) and the deletion of **T** at 778, indicating the these mutations do not represent novel editing sites.

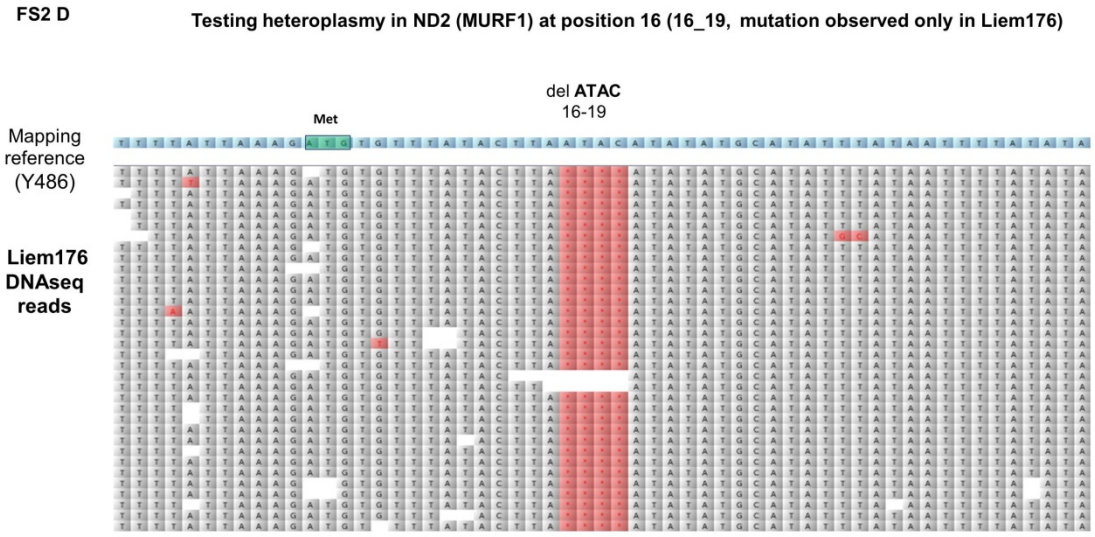

FS2 D
Testing heteroplasmy in ND2 (MURF1) at position 16 (16_19, mutation observed only in Liem176)
del ATAC
16-19
Met
Mapping reference (Y486)
Liem176 DNAseq reads

Figura Suplementaria 3. Análisis de mutaciones en COI y ND5

Mutaciones observadas en los genes COI y ND5 presentes exclusivamente en las cepas americanas de *T. vivax*. Alineamientos múltiples de los genes ortólogos de especies relacionadas y confirmación de homoplasmia para estas mutaciones.

A. Alignments show that the point mutations observed in COI and ND5 affect evolutionary conserved amino acid positions (they are conserved between the African strain Y486 and T. brucei, Trypanosoma cruzi and Leishmania). Positions of interest are emphasized with colors: *the mutant variant in red, while the canonical variant (i.e. the wild type in the remaining species) in green.* The corresponding change at the DNA level and nucleotide position is indicated inside the brackets.

A.

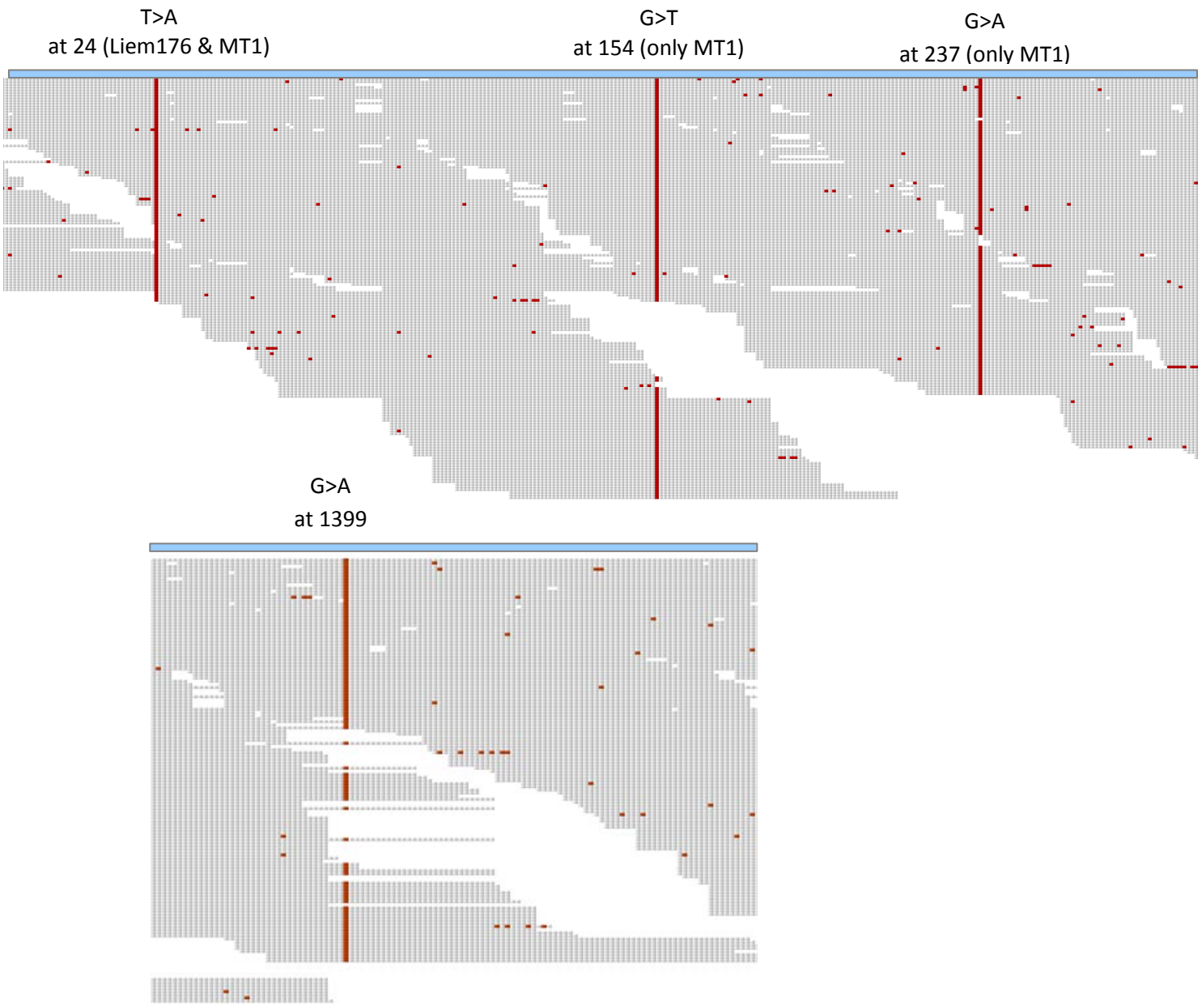
Cytochrome oxidase I

```
                          C>W (nt 24, T->A )                                  G>C (nt 154, G->T)
T.vivax_Liem      MFFICLTWLSVSHKMIGICYLLTAILCGFIGYVYSLFIRLELSLVGCGVLFGDYQFYNVL 60
T.vivax _MT1      MFFICLTWLSVSHKMIGICYLLTAILCGFIGYVYSLFIRLELSLVGCGVLFCDYQFYNVL 60
T.vivax _Y486     MFFICLTCLSVSHKMIGICYLLTAILCGFIGYVYSLFIRLELSLVGCGVLFGDYQFYNVL 60
T.brucei          MFFLCLVCLSVSHKMIGICYLLVAILCGFIGYIYSLFIRLELSLIGCGVLFGDYQFYNVL 60
T.cruzi           MFFICLVCLSVSHKMIGICYLLVAILCGFVGYVYSLFIRLELSLVGCGVLFGDYQFYNVL 60
Leishmania dono   MFWLCLVCLSVSHKMIGLCYLLVAILSGFVGYVYSLFIRLELSIIGCGILFGDYQFYNVL 60
                  **::**. *********:****.***.**:**:**********::***:** ********

                                    G>S (nt 237,  G->A)
T.vivax _Liem     ITAHGLIMVFGFIMPVAMGGFMNYFAPIIVGFPDMVFPRLNNMSFWMFLGGVGCLISGFL 120
T.vivax _MT1      ITAHGLIMVFGFIMPVAMGSFMNYFAPIIVGFPDMVFPRLNNMSFWMFLGGVGCLISGFL 120
T.vivax _Y486     ITAHGLIMVFGFIMPVAMGGFMNYFAPIIVGFPDMVFPRLNNMSFWMFLGGVGCLISGFL 120
T.brucei          ITSHGLIMVFAFIMPITMGGFTNYFAPVMVGFPDMVFPRINNMSFWMFIGGFGCLVSGFL 120
T.cruzi           ITAHGLIMVFAFIMPVIVGGFVNYFAPVMVGFPDMVFPRLNNMSFWMFVGGFGCLVSGFL 120
Leishmania dono   ITSHGLIMVFAFIMPVMMGGLVNYFIPVMAGFPDMVFPRLNNMSFWMYLAGFGCVVNGFL 120
                  **:*******.****: :*.: *** *::.*********:*******::.*.**::.***

                                                                  V>I  (nt 1399, G->A)
T.vivax_Liem      FGSNMVFFPLHSIGMFAFPRRISDYPISFLFWSSFTLFGMLLLSFLILFCCCLFNTFLFW 480
T.vivax_MT1       FGSNMVFFPLHSIGMFAFPRRISDYPISFLFWSSFTLFGMLLLSFLILFCCCLFNTFLFW 480
T.vivax_Y486      FGSNMVFFPLHSIGMFAFPRRISDYPISFLFWSSFTLFGMLLLSFLVLFCCCLFNTFLFW 480
T.brucei          FGSNMVFFPLHSLGMFAFPRRISDYPISFLFWSAFTLYGMLLLTFLVIFCCCLFNVILFW 480
T.cruzi           VGVNMLFFPLHSLGMFAFPRRISDYPISFLYWSSFSLYGMLLITSLVLFCCCLFSILFFW 480
Leishmania dono   IGSNMVFFPMHSLGMYAFPRRISDYPISFLFWSSFMLYGMLLLASLILFLCALFCVFLFW 480
                  .* **:***:**:**:***************:**:* *:****:: *::* *.**  ::**
```

ND5

```
                                    G>C (nt 430 G->T)
T.vivax_Liem       VLSMNIFILSFDFLTAYCGWELLGLFSFQLISYFWFRFFALKYGFKAFLIGKIGDVLLII 180
T.vivax_MT1        VLSMNIFILSFDFLTAYCGWELLGLFSFQLISYFWFRFFALKYGFKAFLIGKIGDVLLII 180
T.vivax _Y486      VLSMNIFILSFDFLTAYCGWELLGLFSFQLISYFWFRFFALKYGFKAFLIGKIGDVLLII 180
T.brucei           VVCMNLFILSYDFLTAYCGWELLGLFSFFLISYFWYRFFALKFGFKAFFIGKIGDVLLIF 180
T.rangeli          IMCMNFFILSYDFLTAYCGWELLGLFSFFLISYFWYRFFSLKFGFKAFFIGKIGDILLMS 180
Leishmania         VLCMNFFILSYDFLTAYCGWELLGLFSFFLISYFWYRFYALKFGFKAFFISKVGDIMLLL 180
                   ::.**:****:************ **** ******:**::**:*****:*.*:**::*:
```

B. Testing heteroplasmy in Cytochrome Oxidase I at position 24, 154, 237 and 1399



Checking heteroplasmy in Cytochrome Oxidase I. The figure shows two segments of alignment containing the 4 mutations listed in table 2. They show that all reads exhibit the point mutations (variation between mapping reference and reads appear marked in red) indicating absence of heteroplasmy for these mutations. Randomly scattered red squares are due to sequencing errors.

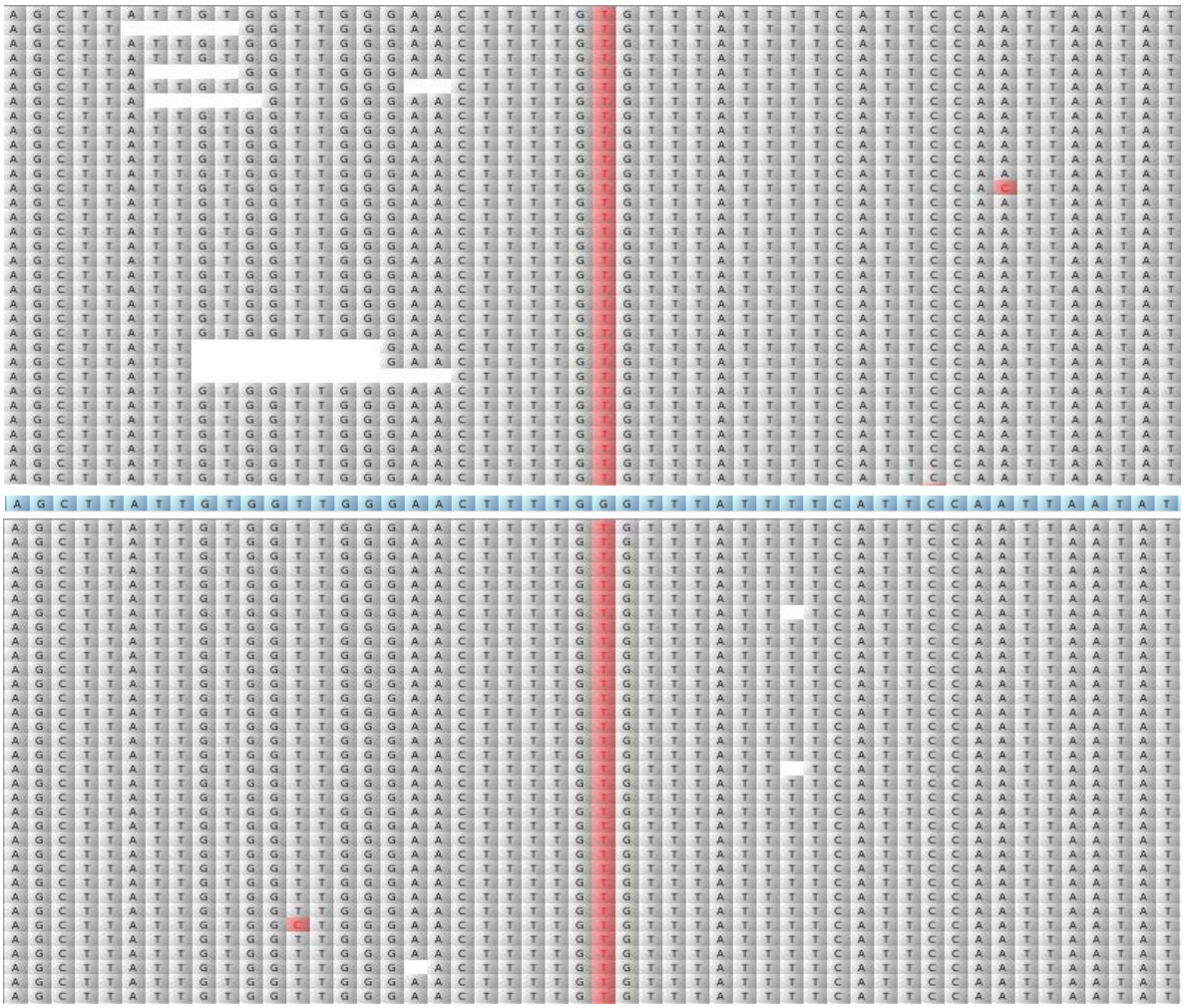
C. Testing heteroplasmy in ND5 at position 430

ND5 (MT1 and Liem)

G>T (430)

DNAseq MT1

DNAseq Liem-176

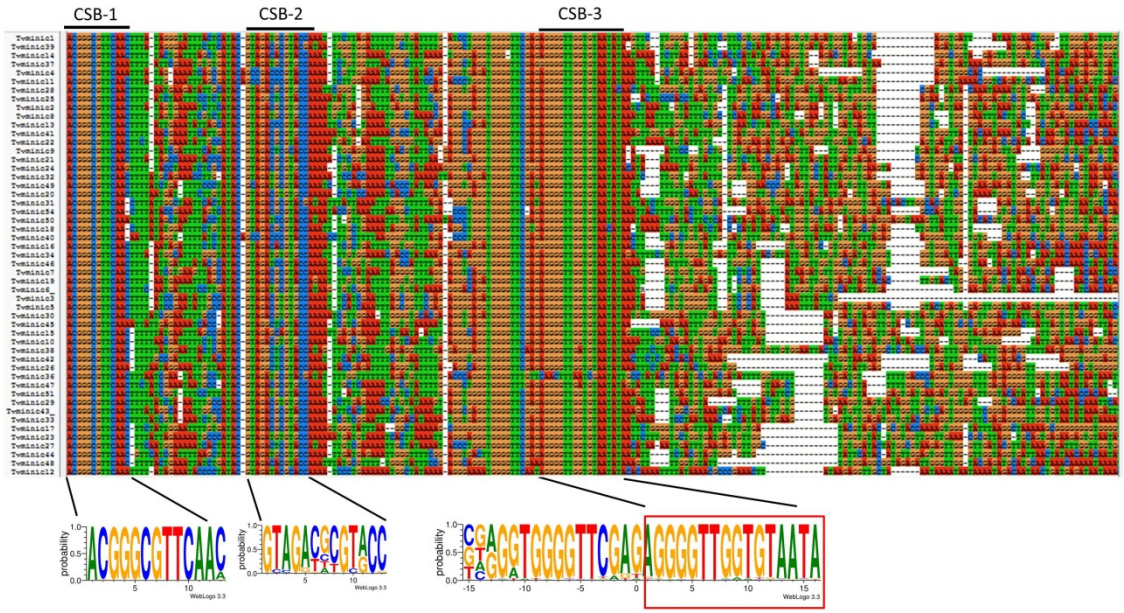


Checking heteroplasmy in ND5. The figure shows the segment of alignment containing the mutation (G->T, at 430). In this case reads from Liem-176 and MT1 were mapped separately. Figure confirms absence of heteroplasmy for this mutation.

Figura Suplementaria 4. Análisis sobre minicirculoma

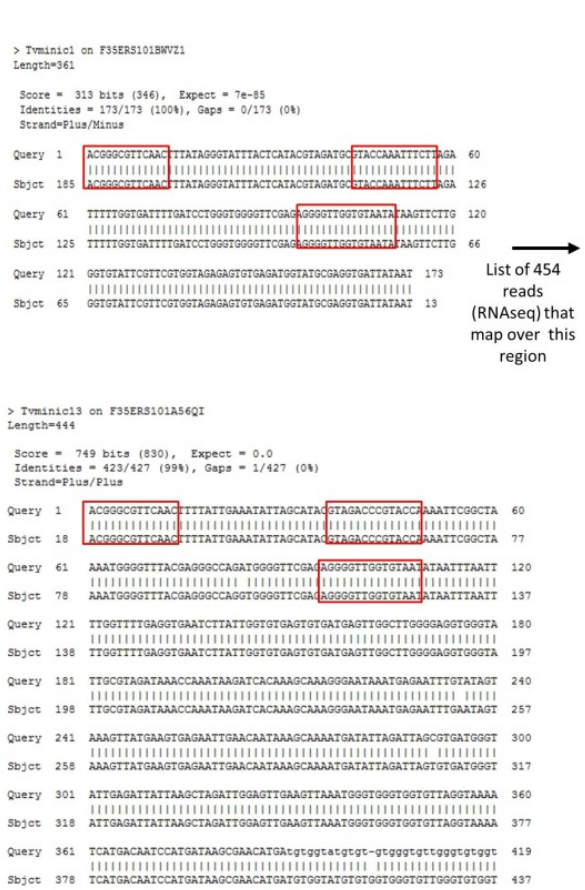
A. Alineamiento de los 54 minicírculos identificados en la cepa MT1 (la region contiene las zonas de conservación CSB 1 a 3).

Figure FS4



Supplementary Figure FS4. Alignment of conserved region that contains the CSB blocks (data set includes the 54 MT1 minicircles). The three conserved blocks are highlighted and their degree of conservation is illustrated by a logo representation of each one of the three blocks

B. Mapeo de *reads* de RNAseq mostrando transcripción de bloques conservados (CSBs).



List of 454 reads (RNAseq) that map over this region

Sequences producing significant alignments:	Score	e-value
F35ERS101BWVZ1	313	7,00E-85
F491XYR02GH0D	266	1,00E-70
F35ERS101A0NB2	187	7,00E-47
GCI4HUN02JLFCL	131	5,00E-30
GCI4HUN02JAON3	131	5,00E-30
GCI4HUN02G3OAZ	131	5,00E-30
GCI4HUN02IBGKH	131	5,00E-30
F491XYR02JE6EO	131	5,00E-30
F35ERS101A1A2Q	131	5,00E-30
F35ERS101AMFOV	131	5,00E-30
F35ERS101BINMT	131	5,00E-30
F491XYR02FMEUG	129	2,00E-29
F35ERS101AJOZL	127	6,00E-29
F35ERS101AID8T	127	6,00E-29
GCI4HUN02JOQQ5	125	2,00E-28
GCI4HUN02GQUX1	123	7,00E-28
GCI4HUN02H4WOK	123	7,00E-28
GCI4HUN02IJVDY	123	7,00E-28
GCI4HUN02FUMT2	123	7,00E-28
GCI4HUN02INM65	123	7,00E-28
GCI4HUN02JATFW	123	7,00E-28
GCI4HUN02GZZ0D	123	7,00E-28
F35ERS101B8Y8A	123	7,00E-28
F35ERS101APRD2	123	7,00E-28
F35ERS101B6L2H	123	7,00E-28
F35ERS101B9N6Y	123	7,00E-28
F35ERS101B6GRZ	123	7,00E-28
F35ERS101BFIV7	123	7,00E-28
F35ERS101BWDT6	123	7,00E-28
F35ERS101CBJ6A	123	7,00E-28
F35ERS101BKXFU	123	7,00E-28
F35ERS101B1TY3	123	7,00E-28
F35ERS101CCC6Z	123	7,00E-28
F35ERS101B6HDY	123	7,00E-28
F35ERS101BG9WF	123	7,00E-28
F35ERS101BG2G8	123	7,00E-28

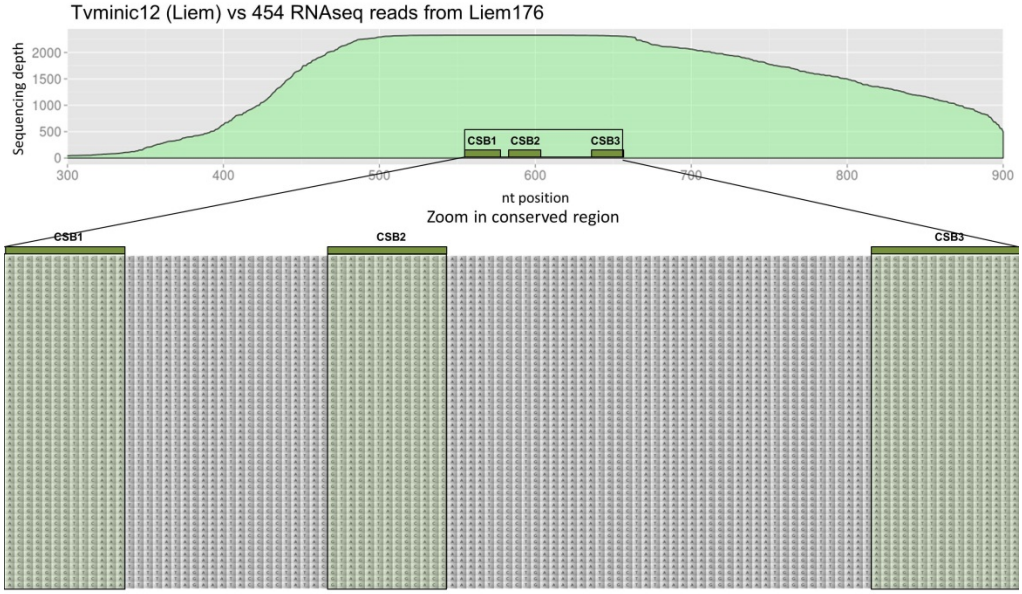
List of 454 RNAseq mapping reads

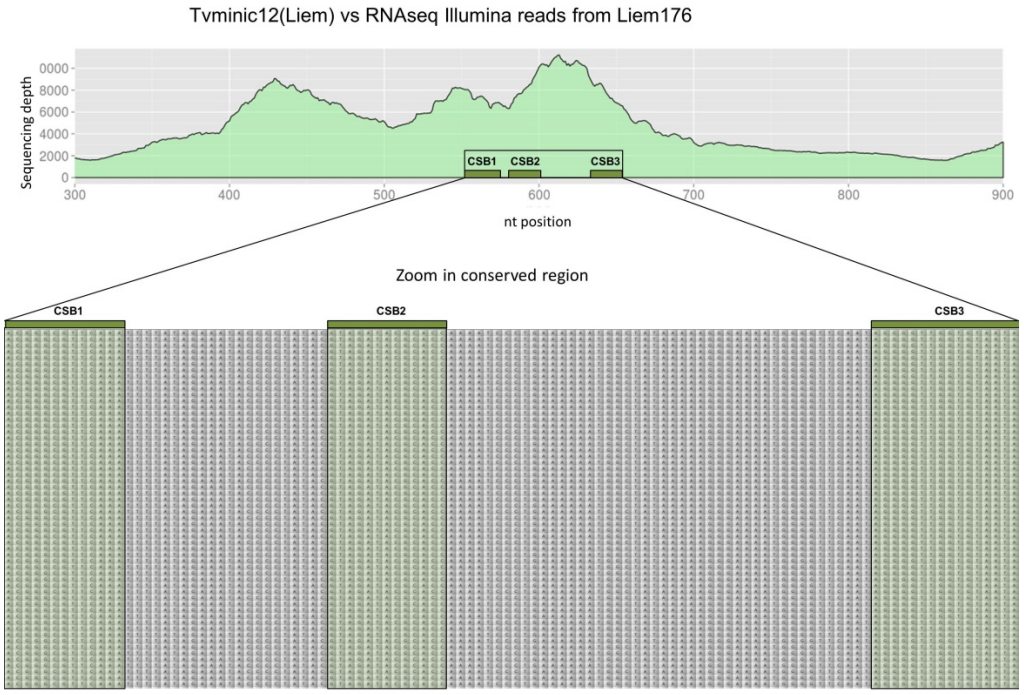
Sequences producing significant alignments:	Score	e-value
F35ERS101A56QI	749	0.0
F35ERS101A6QM3	713	0.0
F491XYR02G5VXL	708	0.0
F35ERS101BALG9	661	0.0
F35ERS101BVM7Z	632	0.0
GCI4HUN02GWUK2	572	7,00E-163
F35ERS101BCYH1	533	6,00E-151
GCI4HUN02FTH3D	471	2,00E-132
GCI4HUN02HVPO6	407	6,00E-113
F35ERS101APVRO	378	3,00E-104
GCI4HUN02IC3D8	376	1,00E-103
GCI4HUN02IZNRS	376	1,00E-103
F35ERS101BTJCT	374	4,00E-103
F35ERS101CGGF5	374	4,00E-103
F491XYR02GEHNC	360	8,00E-99
F35ERS101BRWR6	358	3,00E-98
GCI4HUN02GTHTP	338	3,00E-92
F491XYR02FY7TY	333	1,00E-90
F35ERS101BNRGA	275	3,00E-73
F35ERS101BZOCZ	257	7,00E-68
GCI4HUN02GC5K6	250	1,00E-65
F35ERS101BQGEZ	199	2,00E-50
GCI4HUN02H661T	181	4,00E-45
GCI4HUN02HOBSI	181	4,00E-45
F35ERS101AUYHL	178	5,00E-44
F35ERS101B1VM2	172	2,00E-42
GCI4HUN02GS5QU	165	3,00E-40
GCI4HUN02HXX22	143	1,00E-33
F491XYR02G2MGP	143	1,00E-33

Supplementary Figure FS4 B. **Transcription spans the conserved block in minicircles**. Minicircle alignment against 454 RNAseq database (bioinformatica.fcien.edu.uy/Tvivax). The alignments show two examples of the expression of the CSB-1 to 3 region in different minicircles (CSB-1 to 3 are marked by red boxes). The list of 454 reads spanning the region is presented in the tables located to the rigth of each aligment.

C. Mapeo de *reads* de experimentos diferentes de RNAseq (Liem-176 e Y486) sobre una de las secuencias de minicírculo, mostrando la transcripción de las regiones conservadas.

Figure FS4 C. Three different RNAseq datasets were mapped on the Tvmic12 minicircle. This minicircle was chosen in this illustrative example because it contains an **A** as the last nt in CSB1, i.e the less frequent, non canonical base at this position(see logo in FS4 A). This aspect and the fact that the alignment of reads to this reference was 100% strict (no mismatches were allowed) guarantees that mapping results do not represent cross-mapping on a conserved region.




Tvminic12(Liem) vs RNAseq Illumina reads from Liem176
Sequencing depth
0000
8000
6000
4000
2000
0
300
400
500
600
700
800
900
CSB1
CSB2
CSB3
nt position
Zoom in conserved region
CSB1
CSB2
CSB3

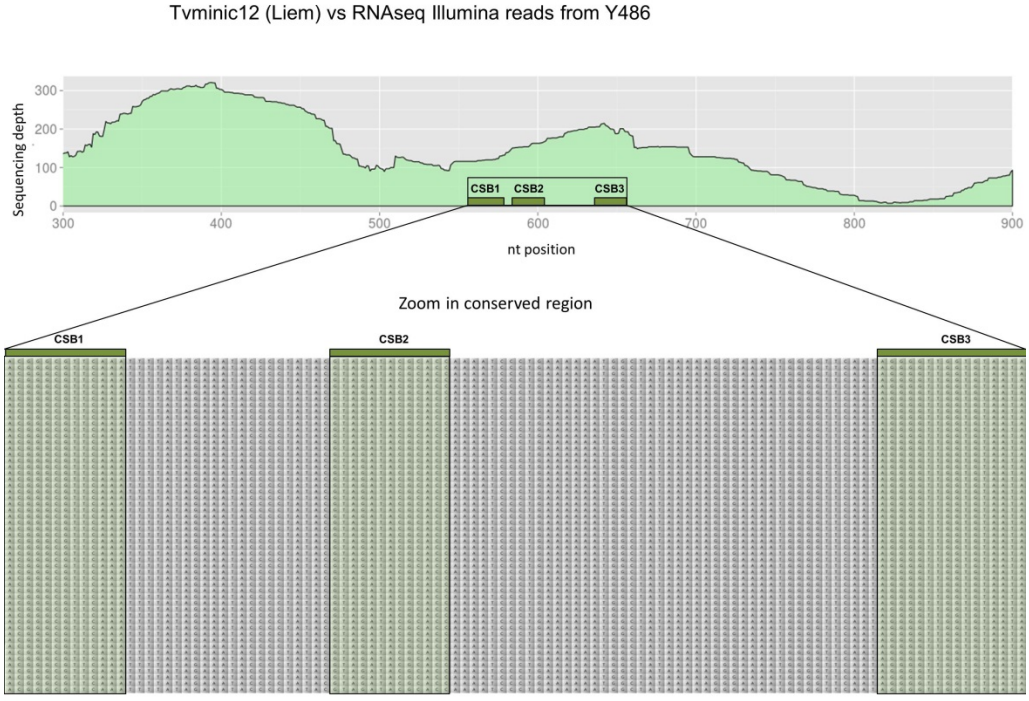
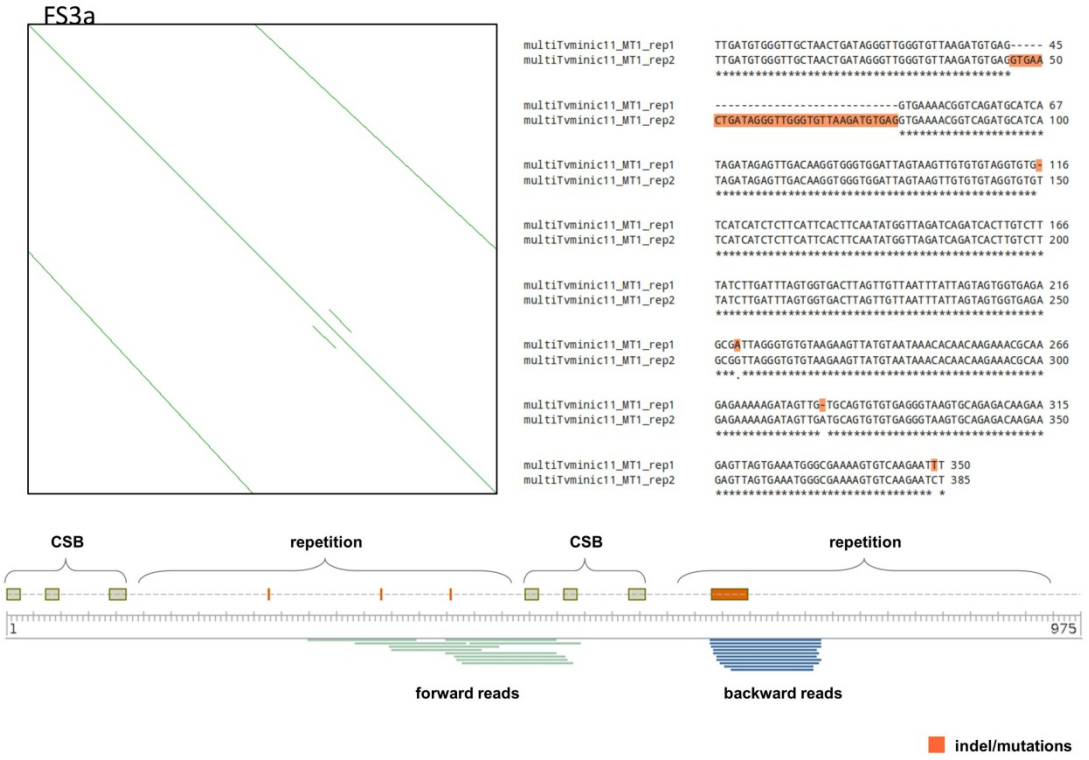

Tvminic12 (Liem) vs RNAseq Illumina reads from Y486
Sequencing depth
300
200
100
0
300
400
500
600
700
800
900
CSB1
CSB2
CSB3
nt position
Zoom in conserved region
CSB1
CSB2
CSB3

Figura Suplementaria 5. Análisis de minicírculos diméricos

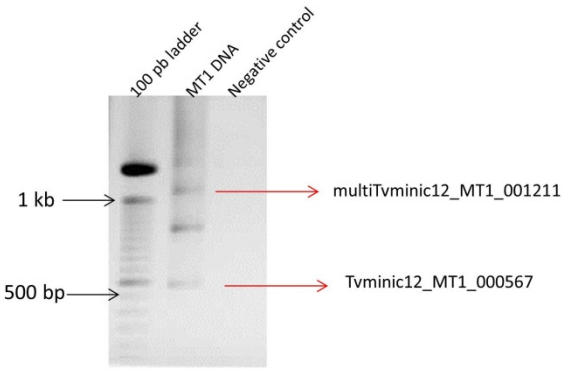
Análisis de minicírculos diméricos.

To corroborate if long (multimeric) minicircles are real molecules and not assembling artifacts, some of them were selected for further analyses. One type of analysis was merely in silico, and consisted on mapping paired-end reads. Note that when the two halves of a dimeric minicircle are divergent enough (say contains indels and point differences), and the two reads of a pair (paired-end) match perfectly, each one of them on a different half of the dimeric minicircle (see bellow FS3 a), is indicative that the molecule does exist as a dimer. Moreover specific primers were designed for their amplification. The sizes of resulting amplicons allowed us to verify the existence of singletons and dimers. Furthermore, these were sequenced using Sanger (capillary) methodology and in all cases the obtained sequences were almost identical to the respective assembled contigs (FS3 b).



FS3b

A. Minicircle confirmation by PCR-sequencing



B. Primers and sequence

miniC_12_F 5´ ATCTTATTCTTATCGATCAAC 3´
miniC_12_R 5´ CGATAAGAATAAGATAATAATAC 3

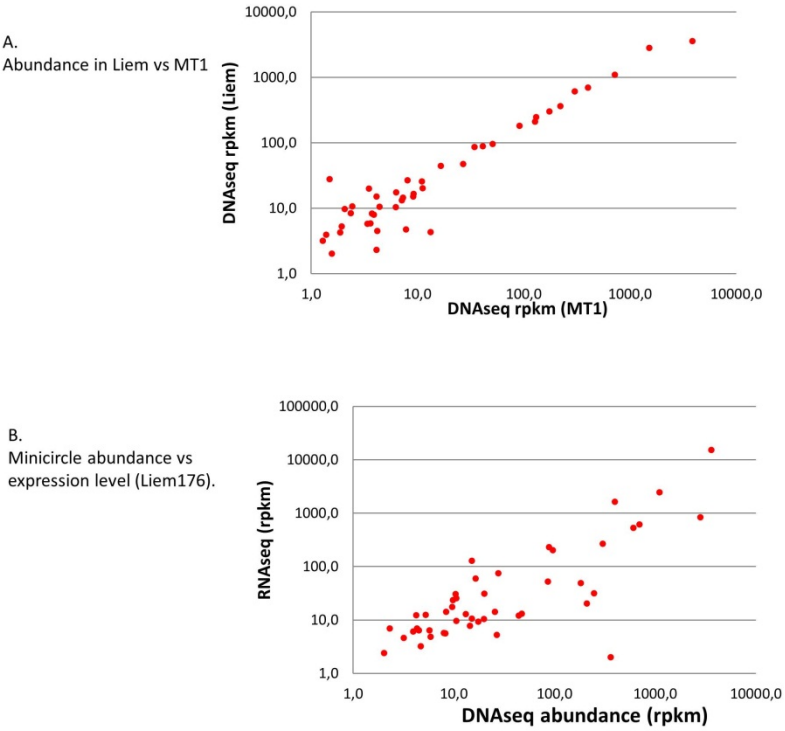
atcttattcttatcgatcaacttattgttattgttttatctttattgactatcttatacttatctttattgtttattatctttcatcatatt
accttcttatcaccgtttattatctcatcagttattgtctattttattatttttatttagcaacctctcataagtttcgggcgttttataa
aattttgtgtatattacaccaaccccttattgaaccccatttccccttttttatagccatttttttcagggattttggtgcgtatctacgt
ataggggtatttctataaaatttgaacgcccgtaaatttttgaggttttcagcgattttcttgactattttattttattctcattaac
tatctttcacgctgtttctcaattttattgattgttgttcaagttattgctcggtcttattgcacggttgttaggtttgtgtttcgcgac
ttgtctcttaacacttaaccttcaaatcagtgacatctctccacttctctaatctaacacaactaaagaggattatatattatatt
atattagtattattatcttattcttatcgatta

C. Illumina Assembly vs Sanger confirmation sequence

```
Sequence ID: lcl|16259Length: 567Number of Matches: 1
Related Information
Range 1: 1 to 564Graphics Next Match Previous Match
Alignment statistics for match #1 Score Expect  Identities      Gaps    Strand
998 bits(540)   0.0     557/565(99%)    1/565(0%)       Plus/Plus

Query  2    ATCTTATTCTTATCGATCAACttattgttattgttttatctttattgactatcttatact  61
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1    ATCTTATTCTTATCGATCAACTTATTGTTATTGTTTTATCTTTATTGACTATCTTATACT  60

Query  62   tatctttattgtttattatctttcatcatattaccttcttattatcgtttattatctcat  121
            ||||||||||||||||||||||||||||||||||||||||||| | ||||||||||||||
Sbjct  61   TATCTTTATTGTTTATTATCTTTCATCATATTACCTTCTTATCACCGTTTATTATCTCAT  120

Query  122  cagttattgtctatttattatttattttagGCAACCTCTTATAAGTTTCGGGCGTTTTAT  181
            ||||||||||||||||||||||||||||| |||||||||||||||||||||||||||||
Sbjct  121  CAGTTATTGTCTATTTTATTATTTTATTTAGCAACCTCTCATAAGTTTCGGGCGTTTTAT  180

Query  182  AAAATTTTGTGTATATTACACCAACCCCTATTGAACCCCATTTCCCTTTTTATAGCCAT  241
            ||||||||||||||||||||||||||||||||||||||||||| |||||||||||||||
Sbjct  181  AAAATTTTGTGTATATTACACCAACCCCTATTGAACCCCATTTCCCCTTTTTATAGCCAT  240

Query  242  TTTTTCAGGGATTTTGGTGCGTATCTACGTATAGGGGTATTTCTATAAAATTTGAACGCC  301
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  241  TTTTTCAGGGATTTTGGTGCGTATCTACGTATAGGGGTATTTCTATAAAATTTGAACGCC  300

Query  302  CGTAAATTTTTGAGGTTTTTCAGCGATTTTCTTGACTATTTTATTTTATTCTCATTAACT  361
            ||||||||||||| ||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  301  CGTAAATTTTTGAGG-TTTTCAGCGATTTTCTTGACTATTTTATTTTATTCTCATTAACT  359

Query  362  ATCTTTCATGCTGTTTCTCAATTTTATTGATTGTTGTTCAAGTTATTGCTCGGTCTTATT  421
            ||||||| ||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  360  ATCTTTCACGCTGTTTCTCAATTTTATTGATTGTTGTTCAAGTTATTGCTCGGTCTTATT  419

Query  422  GCACGGTTGTTAGGTTTGTGTTTCGTGACTTGTCTCTTAACACTTAACCTACAAATCAGT  481
            |||||||||||||||||||||| |||||||||||||||||||||| ||||||| ||||||
Sbjct  420  GCACGGTTGTTAGGTTTGTGTTTCGCGACTTGTCTCTTAACACTTAACCTTCAAATCAGT  479

Query  482  GACATCTCTCCACTTCTCTAATCTAACACAACTAAAGAGGATTATATATTATATTATATT  541
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  480  GACATCTCTCCACTTCTCTAATCTAACACAACTAAAGAGGATTATATATTATATTATATT  539

Query  542  agtattattatcttattcttatCGA  566
            |||||||||||||||||||||||||
Sbjct  540  AGTATTATTATCTTATTCTTATCGA  564
```

Figura Suplementaria 6. Análisis de abundancia y expresión de minicírculos

A. Comparación entre abundancia de clases de minícirculos en las cepas MT1 y Liem-176.
B. Comparación entre la abundancia de minicírculos y su expresión (en Liem-176)

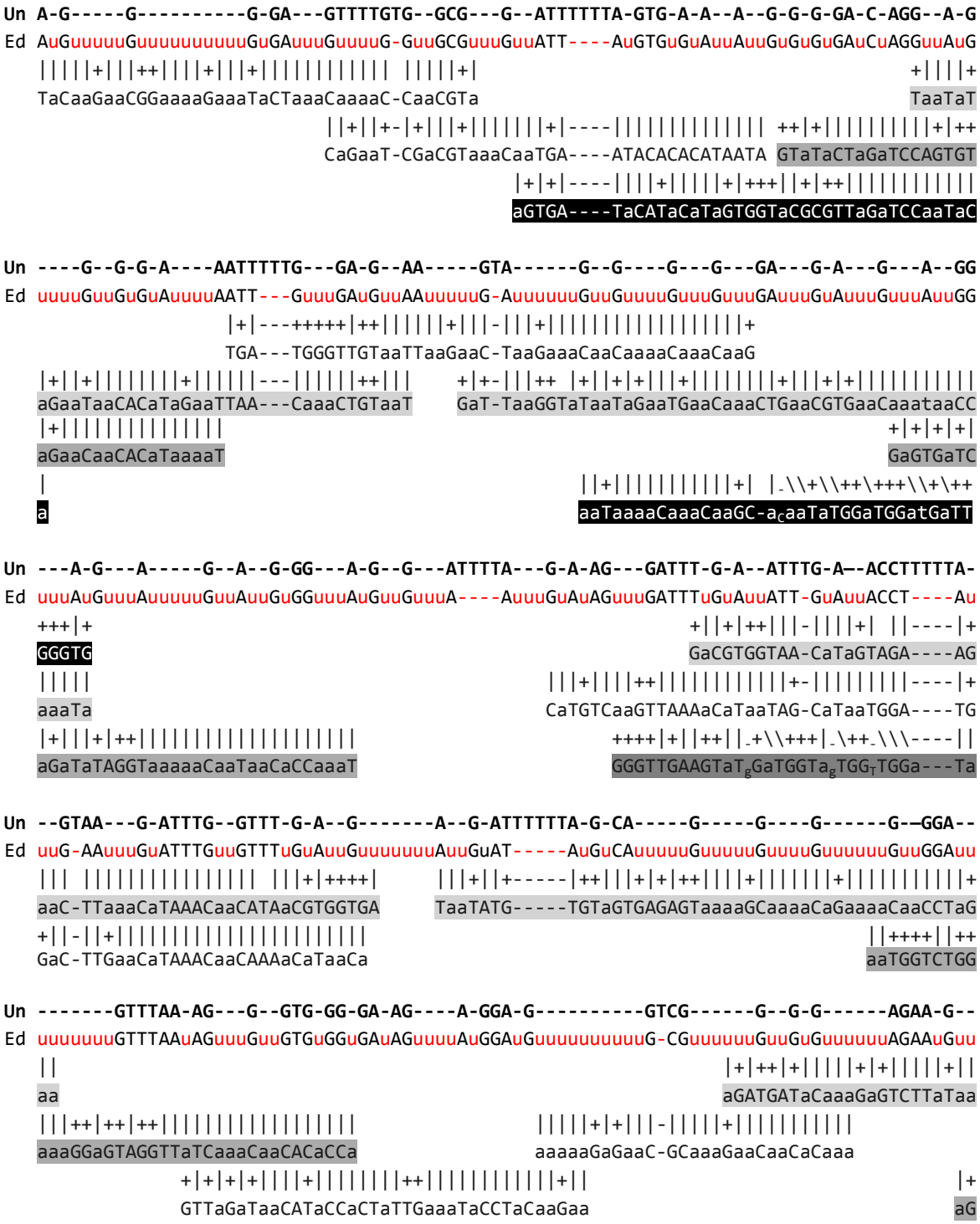


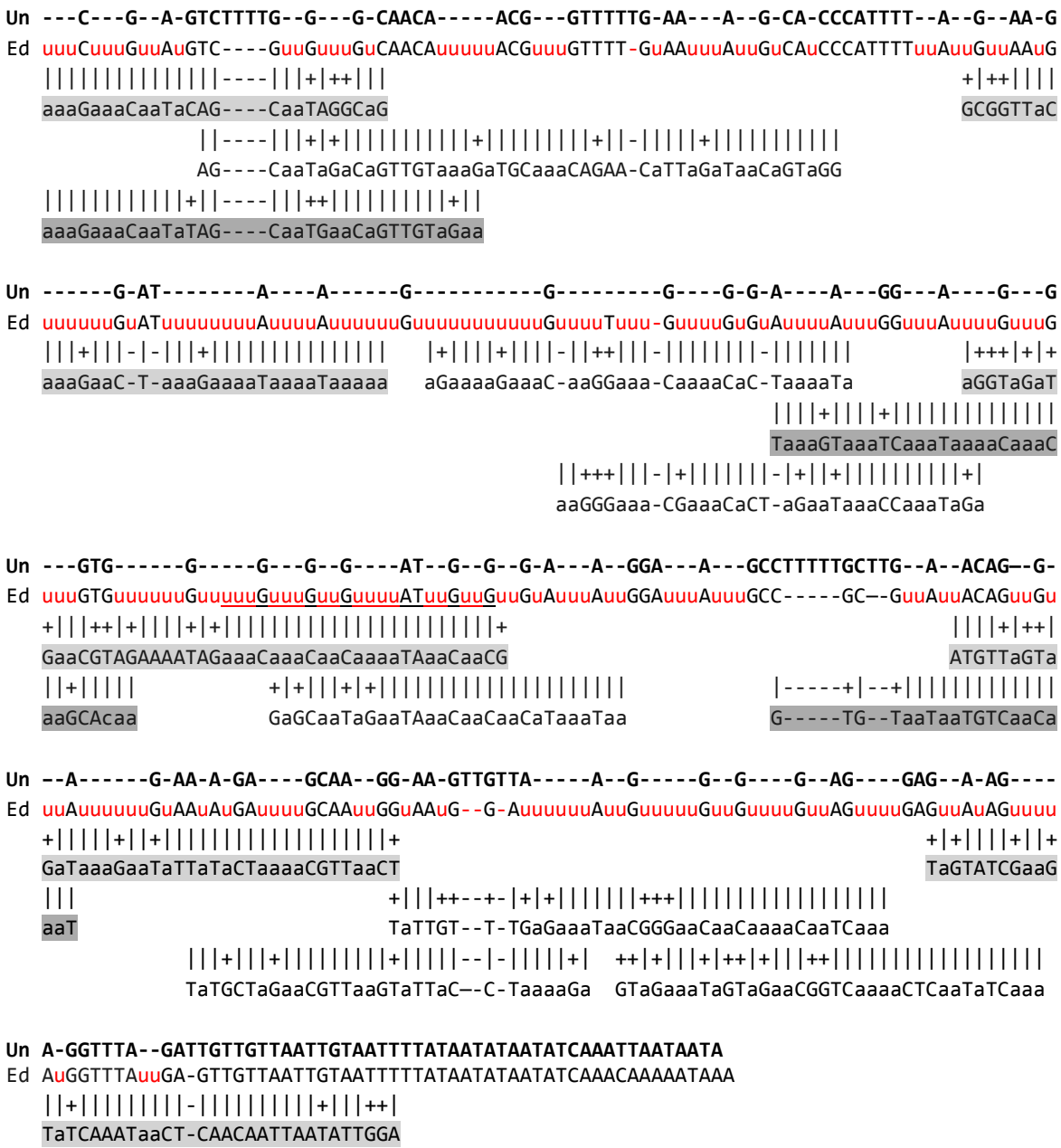
Supplementary Figure FS6. Scatter plot showing the relative abundance of homologous minicircles in Liem176 and MT1 American strains measured as DNA rpkm (A). Expression level of Liem176 minicircles (RNAseq rpkm) related to the abundance of each one (DNAseq rpkm).

Figura Suplementaria 7. Mapeo de ARNg identificados en A6 y RPS12

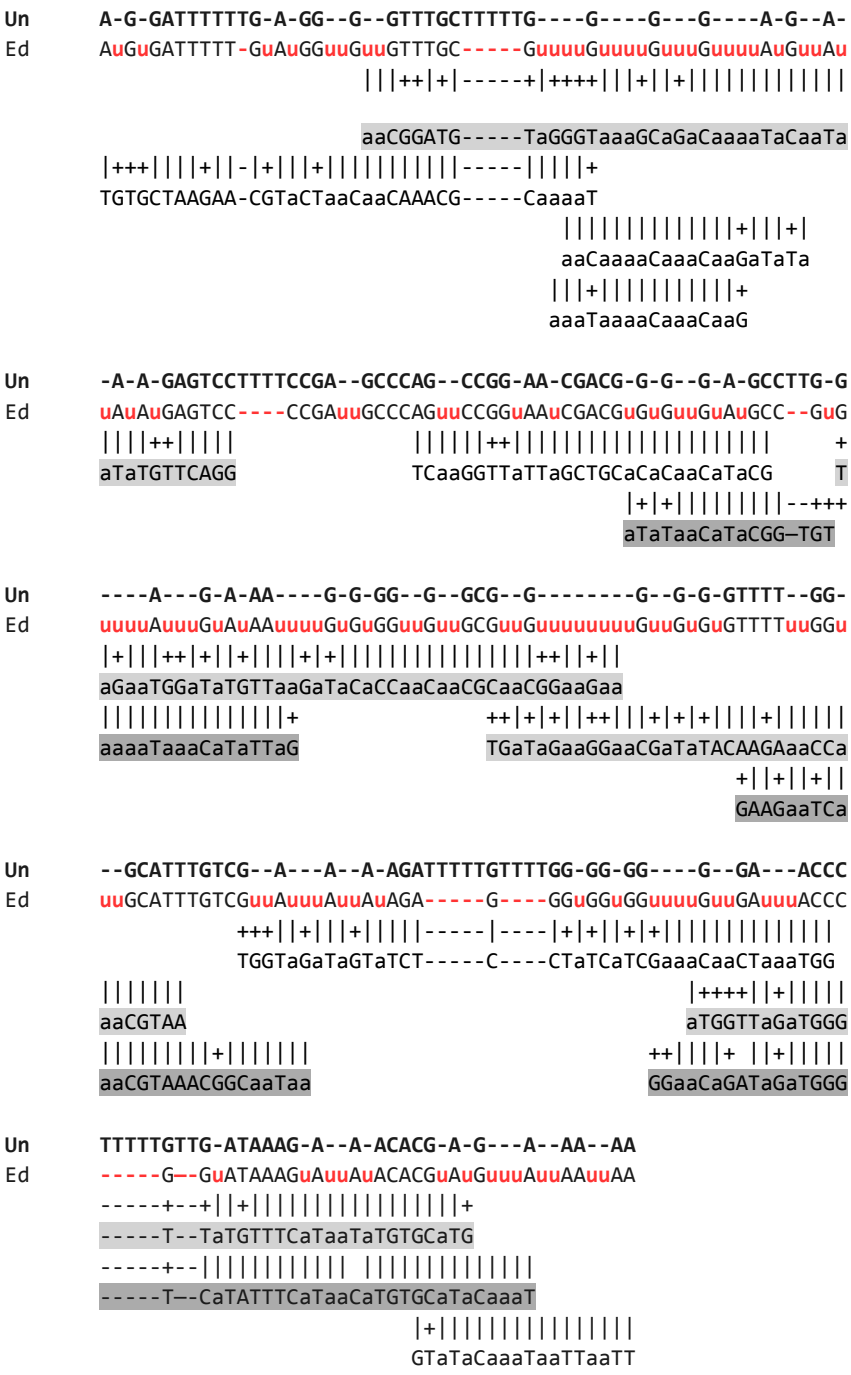
Mapeo de ARNg sobre las secuencias de los genes A6-ATPasa y RPS12. Se muestra la secuencia genómica (sin editar, Un), la secuencia editada completamente (Ed) y los ARNs guía identificados (con + se indican apareamientos G:U).

A6 Genomic, Edited (mRNA) sequences and gRNAs





RSP12 Genomic, Edited (mRNA) sequences and gRNAs



Anexo 3. Expresión y *editing* de genes mitocondriales en cepas americanas y africanas.

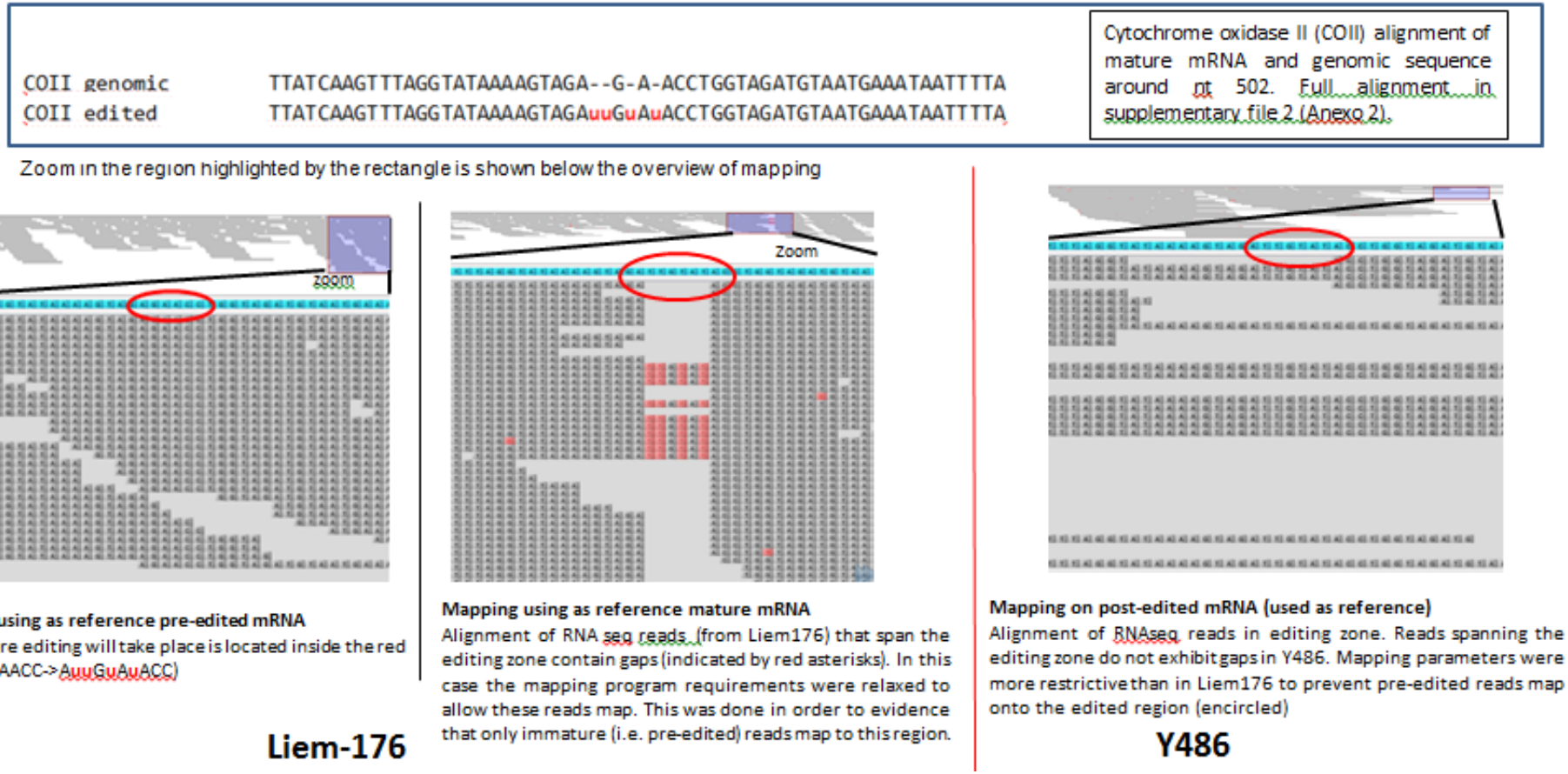
Ejemplos de genes que requieren *editing* parcial (Cyb) o deben ser pan-editados (ND3), donde no se observa mapeo de *reads* de RNAseq cuando se utiliza el transcripto maduro (editado) como referencia en las cepas americanas (Liem-176 y MT1) y sí hay mapeo de *reads* sobre el transcripto maduro en la cepa africana (Y486) (A-C).

El último ejemplo muestra el mapeo de *reads* de RNAseq sobre el transcripto maduro (editado) del gen RPS12 tanto en las cepas Liem-176 y MT1 (americanas) como en la cepa Y486 (africana) (D).

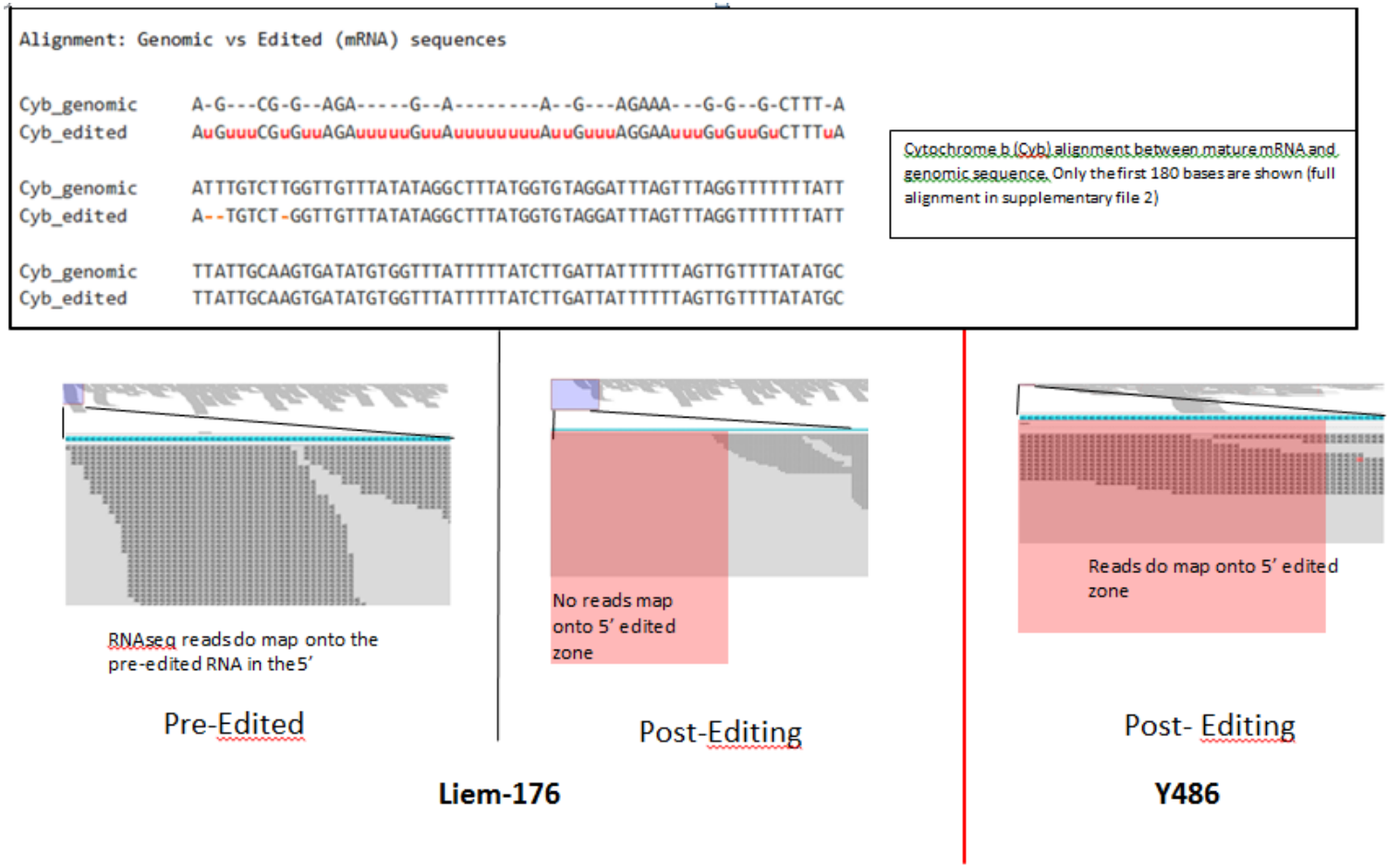
These figures complement the results presented in table 2*. We have chosen as illustrative examples four genes that undergo different degrees of editing. COII, a gene where editing implies the insertion of four Us around position 502. Cyb, a gene in which editing consists in the insertion in the 5' region of about 40 uridines (and three deletions) and two pan-edited genes ND3 and RPS12. Although ND3 is transcribed, it does not undergo editing in Liem176 (i.e. remains as immature RNA) while RPS12 is completely and correctly edited. For the case of Liem176 the mapping on pre- and post-edited RNAs is presented while for Y486 only the post-edited mapping is shown.

* Hace referencia a la Tabla 2 del manuscrito "Kinetoplast adaptations in American strains from *Trypanosoma vivax".*

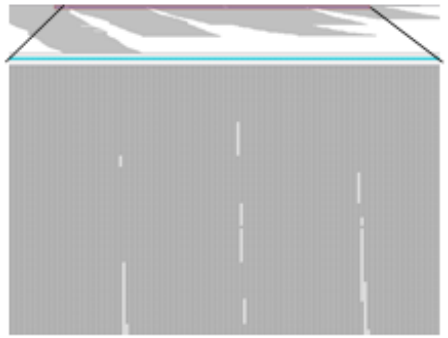
A. Example 1. COII, editing implies the insertion of only four Us around nt position 502. As it can be observed, RNAseq reads corresponding to edited form or mRNA are not detected in Liem-176, thus indicating absence of editing.



B. Example 2. Cytochrome b, edited only in the 5' region. RNAseq data mapped against pre-edited and post-edited RNA forms in Liem-176 and Y486 strains. No reads mapping onto the edited region could detected in Liem-176. Yet in strain Y486, the same region (post-edited) contains abundant RNAseq reads that map on it.

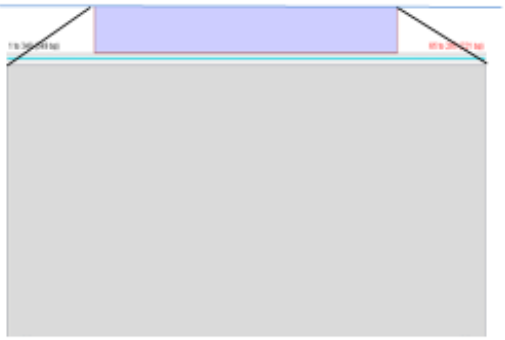


C. Example 3. ND3, a pan-edited gene. This gene that has lost its ability to undergo editing in Liem-176. RNAseq reads map only to pre-edited RNA in Liem-176. But in Y486 strain reads map to both pre-edited and post-edited RNA (only the latter is shown).



RNA seq reads do map onto the pre-edited RNA (5043 reads in total, see table 2)

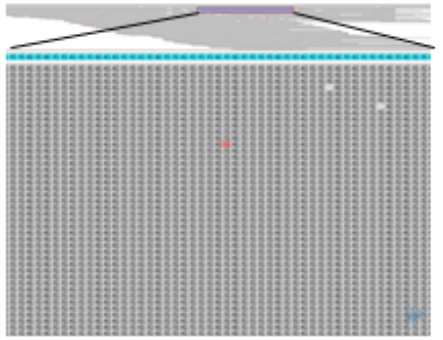
Pre- edited



Post-edited template (mature mRNA) used as mapping reference.
Not a single read maps onto this template in Liem176.

Post-Edited

Liem-176



Post-edited template (mature mRNA) used as reference. Reads map along the whole sequence (523 reads map onto this template in Y486)

Post-Edited

Y486

D. Example 4. RSP12, a pan-edited gene. RNAseq reads mapped against pre-edited and post-edited RNAs in Liem-176 and Y486 (fort the latter strain only post-edited mRNA is shown). RNAseq reads map to both pre-edited and post-edited forms of mRNA in Liem-176 and Y486 strains.