Tesis de doctorado

Organización genómica en los tripanosomas africanos

Estudio de un modelo: *Trypanosoma vivax*

Guillermo Lamolle
Orientador: Fernando Álvarez

Resumen 3
Introducción 4
Trypanosoma vivax 18
Materiales y métodos 21
Genoma de *T. vivax* 21
Obtención de ARN y secuenciación del transcriptoma de *T. vivax* 22
Anotación funcional 26
Nivel de transcripción 28
Análisis estadístico 29
Análisis de componentes principales 29
Análisis de clusters 32
Generación de secuencias por simulación 32
Bioinformática 35
Blast 35
Artemis 36
FigTree 37
Emboss 37
Bowtie 37
Scripts personalizados 39
Planteo del problema y objetivos 40
Resultados 43
Reanotación primaria de contigs genómicos 43
Búsqueda del repertorio silencioso de genes VSG 44
ORF huérfanos 45
Análisis multivariados 47
Orfogenicidad 53
Compartimientos genómicos 61
Expresión 65
Búsqueda de la VSG activa 71
Conclusiones 73
Anexo 78
markov2.pl 78
hebra.pl 82
Abreviaturas 84
Referencias bibliográficas 86
Paper adjunto 90

Resumen

De interés sanitario y económico, los tripanosomas africanos presentan un particular modo de evasión del sistema inmune del hospedador vertebrado, denominado variación antigénica, basado en la alternancia cíclica (a nivel poblacional) de diferentes variantes de la proteína que cubre la superficie celular, denominada VSG (*Variant Surface Glycoprotein*). En esta tesis se analiza la organización y evolución genómica de *Trypanosoma vivax*, dando especial énfasis a las regiones genómicas que contienen los genes codificantes de proteínas VSG. La importancia de estudiar esta especie radica en su posición evolutiva, pues es la separación más ancestral del árbol evolutivo de los tripanosomas africanos.

La búsqueda de marcos abiertos de lectura (ORF) revela que esta especie presenta un número excepcionalmente alto de los mismos, y sorprendentemente más de la mitad estos ORF son huérfanos, es decir no presentan homología con genes previamente descriptos. Se realizaron análisis multivariados (en base a frecuencias de codones, es decir, trinucleótidos en marco), los cuales muestran que la mayoría de estos ORF huérfanos tienen propiedades estadística sustancialmente diferentes de los restantes genes del genoma. Análisis adicionales de estos ORFs huérfanos revelan la existencia de un fenómeno que denominamos *orfogenicidad*, que consiste en la capacidad intrínseca de una molecula de ADN de presentar largos y numerosos ORF superpuestos en distintos marcos de lectura y hebras. Usando simulaciones computacionales basadas cadenas de Markov (de diverso orden), pudimos determinar que dicho fenómeno no se explica meramente por el contenido GC = 0.6 de estos ORF (en secuencias ricas en GC la probabilidad de que se formen codones STOP por azar es menor), sino que responde fundamentalmente a la sobre o subrepresntación de ciertas palabras (oligonucleótidos). En particular destacamos la casi completa ausencia del dinucleótido TpA (que forma parte de 2 de los tres codones stop) como elemento clave en la generación de la orfogenicidad.

Análisis adicionales, usando contigs genómicos, muestran que existen dos espacios genómicos bien definidos: uno que alberga los genes *housekeeping*, presenta un nivel de GC de 0.5 y baja orfogenicidad, y otro, orfogénico, en el que se ubican los genes de VSG y varios miles de los ORF huérfanos arriba descritos, y que posee un GC de 0.56. Basándonos en el estudio de transcriptoma realizado en *T. vivax*, concluimos que muchos contigs de este último espacio no se transcriben (si bien contienen genes de VSG, otras proteínas de superficie y proteínas hipotéticas), poniendo en duda el concepto ampliamente aceptado de que en tripanosomatidos el nivel transcripcional la regulación de la expresión génica no juega ningún rol.

Un conjunto de importantes diferencias con *T. brucei*, el tripanosoma mejor estudiado hasta hoy, deja preguntas abiertas acerca del surgimiento y la evolución de las VSG.

Introducción

Los tripanosomátidos son protozoarios flagelados de amplia distribución y conforman una de las dos familias del orden *Kinetoplastida* (*Bodonidae* y *Trypanosomatidae*). *Bodonidae* comprende organismos de vida libre con dos flagelos. Los miembros de *Trypanosomatidae* presentan un único flagelo y son parásitos (en principio de insectos, aunque varios grupos se han adaptado a un segundo hospedador, que puede ser un vertebrado o una planta).

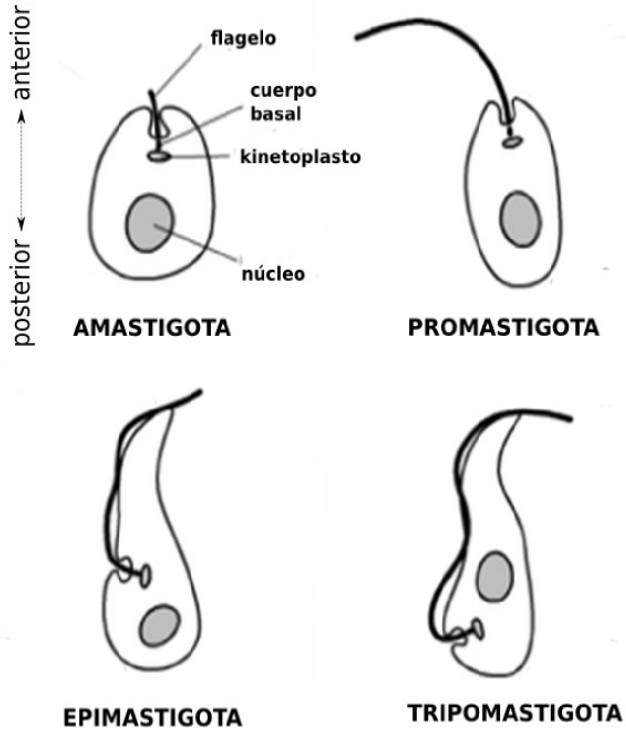


Figura 1. **Esquema simplificado de las fases del ciclo vital de los tripanosomátidos.**

Los kinetoplástidos deben su nombre a un cuerpo denominado, justamente, kinetoplasto (o cinetoplasto), ubicado en la zona basal del flagelo, que consiste en una zona modificada de la única mitocondria, donde se encuentra el ADN mitocondrial anular, organizado en minicírculos (varios miles) y maxicírculos (unas decenas).

Estos presentan diferencias morfológicas en distintas etapas del ciclo. En la figura 1 se presenta un esquema simplificado de las principales formas, que 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 estas son totalmente equivalentes, biológicamente, entre distintas especies. *T. brucei*, por ejemplo, presenta una fase tripomastigota en el vertebrado, en la cual los individuos se reproducen y, en el insecto, dos fases: tripomastigota (que se reproduce en el intestino dando lugar a epimastigotas) y epimastigota (cuyos individuos migran a las glándulas salivares, donde a su vez se reproducen dando lugar a tripomastigotas, llamados tripomastigotas metacíclicos, que tienen capacidad infectiva). En *T. cruzi*, en cambio, se da, además, una fase amastigota, que se reproduce en el interior de las células infectadas del vertebrado. El epimastigota se reproduce, al igual que en *T. brucei*, en el intestino del insecto, pero el tripomastigota (forma infectiva en la sangre del vertebrado, que se produce tras la rotura de la célula hospedadora y favorce la invasión de nuevas células), a diferencia del de *T. brucei*, no se reproduce.

Dentro de los tripanosomátidos, el género *Trypanosoma* tiene especial importancia debido a que posee especies y subespecies que infectan a las personas. Estas son, por un lado, *T. brucei gambiense y T. brucei rhodesiense* (enfermedad del sueño o tripanosomiasis africana) y por otro *T. cruzi* (enfermedad de Chagas o tripanosomiasis americana). También hay especies que afectan especialmente al ganado: *T. brucei brucei, T. vivax*, *T. evansi* y *T. congolense.* La enfermedad que provocan se denomina genéricamente *nagana.* Se estima que más de cinco millones de personas del África subsahariana padecen la enfermedad del sueño, con decenas de

miles[1] de nuevos casos registrados al año y muchos millones (diez millones sólo en la República Democrática del Congo) de personas en situación de riesgo. Además, la misma es la causa de aproximadamente cincuenta mil muertes anuales (*WHO-report*, 2004). La nagana, por su parte, puede producir un fuerte impacto en la economía de los países que la padecen, debido a las pérdidas de importancia en la cría de animales domésticos. De hecho, existen regiones de África —con poblaciones de moscas tse-tse especialmente densas— que están vedadas a las actividades ganaderas debido a la incidencia de esta enfermedad, de lo que resultan no sólo pérdidas económicas sino un daño a la salud pública, debido a la dieta deficiente en proteínas de las poblaciones locales.

Si bien estos tripanosomas son de origen africano, se han dispersado por casi toda la franja tropical y subtropical, y más allá de ella. En Sudamérica y el Caribe, dos especies de tripanosomas africanos (*T. vivax* y *T. evansi*) fueron introducidas a mediados del siglo XIX, probablemente por la Guayana francesa (a través de ganado infectado importado de Senegal) y se expandieron mediante varias especies de moscas hematófagas (tabánidos y *Stomoxys*), que actúan como factores mecánicos de transmisión. Esto último se refiere a que estos insectos actúan como una «aguja infectada», a diferencia de lo que ocurre en su vector natural, la mosca tse-tsé, donde el parásito puede completar el ciclo antes descrito. Esto no es exclusivo de los *T. vivax* presentes en América: en África se da este tipo de transmisión mecánica por tabánidos, además de la que utiliza como vector a la mosca tse-tsé. Este hecho le ha permitido a *T. vivax* expandirse por áreas libres de esa mosca (Gardiner y Mahmoud, 1992; Jones y Davila, 2001). Actualmente, afectan áreas dedicadas a la ganadería, desde México hasta Paraguay (Osorio *et*

[1] En 1995 (en que se reportaron 30.000 casos) la Organización Mundial de la Salud (OMS) estimaba que, por cada caso que era reportado, había diez que no lo eran. Sin embargo, en 2006, con mejores registros, dicha estimación se corrigió, cambiándose ese factor de 10 por uno de cada 3. Además, los casos registrados disminuyeron: en 2009 no llegaron a 10.000 (Simarro *et al.*, 2011).

al., 2008). La nagana ha sido recientemente reportada incluso en el sur de Río Grande do Sul, a pocos kilómetros de la frontera con Uruguay (Schafer *et al.*, 2010), lo que indica un alto riesgo de que nuestro país se vea afectado en el futuro próximo por esta zoonosis. Esto ubica a estas tripanosomiasis africanas en la categoría de parasitosis emergentes en la región.

El genoma de los tripanosomátidos es diploide, y sus poblaciones son monoclonales. Existe evidencia de intercambio sexual. Los primeros datos corresponden a *T. brucei* y fueron publicados en la década de los ochenta (Jenni *et al.*, 1986). Con posterioridad se advirtió el mismo fenómeno en *T. cruzi* (Gaunt *et al.*, 2003*)*, e incluso en *Leishmania major* (Akopiants *et al.*, 2009). Aparentemente, en todos los casos, el intercambio genético ocurre en escasa proporción.

Los genes están organizados en unidades de transcripción con asimetría o polaridad de hebra. Con frecuencia existe un cambio de hebra en la región intermedia (región de *switch* transcripcional), de modo que estas unidades de transcripción se alternan en cuanto a la hebra que ocupan. Como producto de la transcripción se generan transcriptos policistrónicos; esto es, grandes moléculas de ARN mensajero, que contienen la información codificante para distintas proteínas. Es importante destacar que esto implica que los genes estén organizados en grupos, y que todos los genes de uno de estos bloques transcripcionales tengan que estar ubicados en la misma hebra. Esta situación también implica que la regulación de su expresión sea de tipo postranscripcional pues, como ya fue mencionado, debemos considerar el grupo como una unidad policistrónica. Las distintas unidades policistrónicas alternan a lo largo de los cromosomas. La transición entre dos bloques implica un cambio de sentido en la transcripción (*switch* transcripcional) y la existencia de una región intermedia localizada entre dos bloques consecutivos.

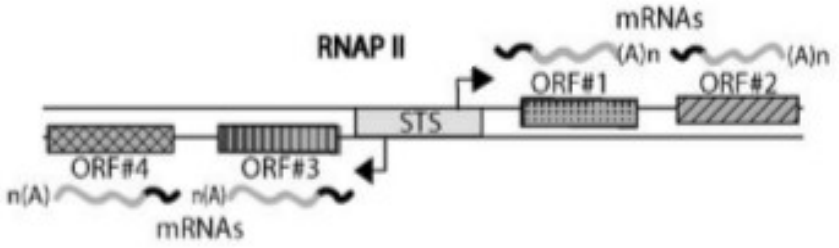


Figura 2. **Ejemplo de una de las regiones *strand switch* (STS) desde cuyos extremos radian dos grupos de genes (Liang *et al.* 2003).**

Las moléculas de ARN mensajero maduro comienzan con una secuencia de 39 nucleótidos denominada *spliced leader* o miniexón, que se agrega durante el proceso de maduración. A continuación siguen la región 5' (no codificante), el marco abierto de lectura (ORF), la región 3' (no codificante), y la cola de poli-A (que también se agrega durante la maduración). El proceso durante el cual se fragmenta el ARN policistrónico y se agrega el *spliced leader* se denomina *trans splicing*. En la mayoría de los genes de eucariotas la maduración se produce por un proceso diferente denominado *cis splicing*, consistente en la eliminación de intrones y empalme de los exones. El agregado de la cola poli-A en los tripanosomas es similar al del resto de los eucariotas (Liang *et al.*, 2003; Mandelboim *et al.*, 2003).

La transcripción policistrónica utiliza una sola corrida de la enzima ARN polimerasa, a diferencia del resto de los eucariotas, en que la transcripción implica la acción de una molécula de ARN polimerasa para cada gen, previa intervención de diversidad de elementos cis y trans (promotores y factores de transcripción), cuya acción conjunta conduce a la iniciación de la transcripción.

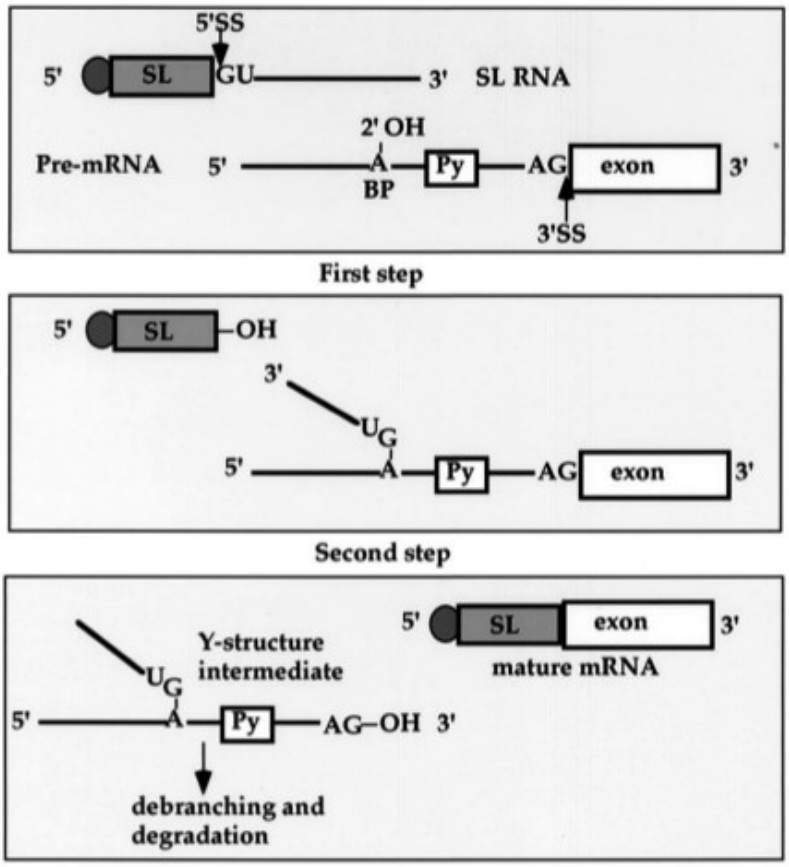


Figura 3. **Serie de pasos que involucra el evento de *trans splicing* de SL. Están señalados el sitio de *trans splicing* 5' GU en el SL RNA y el sitio 3' AG en el pre-mRNA, BP (*branching point*) el sitio de ramificación y Py, el tracto de polipirimidinas (Liang *et al.*, 2003).**

El gasto energético que implica transcribir prácticamente por igual todos los genes, es viable, probablemente, debido al tipo de vida parasitario de estos organismos, para los cuales el abastecimiento de nutrientes es un problema relativamente menor en comparación con lo que ocurre en el caso de organismos de vida libre. Se ha sugerido, además, que este tipo de regulación favorece una respuesta rápida a los cambios bruscos del ambiente, que resulta de vital importancia para estos organismos al pasar de una etapa a otra de su ciclo de vida (Palenchar y Bellofatto, 2006).

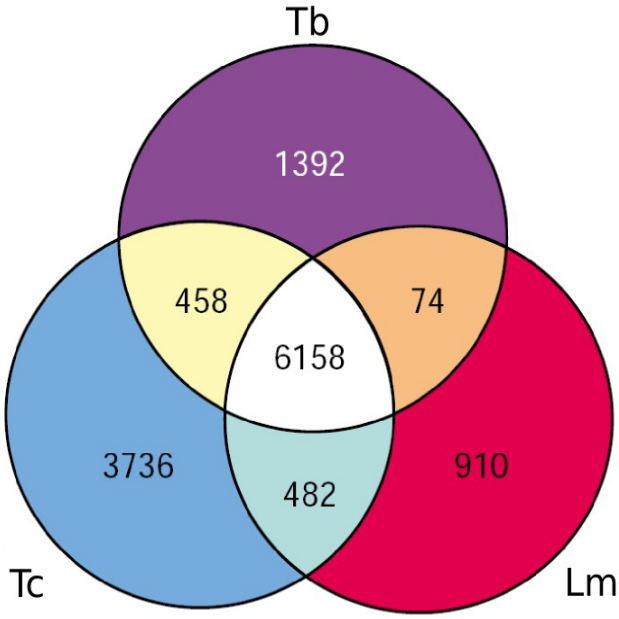


Figura 4. **Diagrama de Venn donde se muestran las cantidades de genes exclusivos o compartidos por dos o las tres especies *T. cruzi, T. brucei y L. major*.**

Existen otras peculiaridades, como la utilización de ARN polimerasa I (ARNPI) para la transcripción de genes codificantes de ciertas proteínas (usualmente esta enzima participa en la transcripción de genes de ARN ribosomal). Es el caso, por ejemplo, de las glicoproteínas variables de superficie, VSG, que forman una cubierta protectora en la fase sanguínea de los tripanosomas africanos, los genes asociados al sitio de expresión (ESAG), que se transcriben junto a las VSG, o las prociclinas, otro tipo de proteínas de superficie que sustituyen a las VSG durante la fase procíclica en el insecto (Gunzl *et al.*, 2003). Aparte de esto, la ARNPI también interviene en la transcripción de pre-ARN ribosomales, como en el resto de los eucariotas.

Entre los distintos tripanosomátidos existe una alta conservación, tanto a nivel de ortología como de sintenia. En la figura 4 se ve el resultado de la búsqueda de COG (*Clusters of Orthologous Groups*) en tres especies de tripanosomátidos, representativas de la diversidad filogenética del grupo: *Leishmania major*, *T. brucei* y *T. cruzi*. En primer lugar, vemos que hay

un gran número de genes ortólogos en común entre las tres especies, lo que habla de un alto grado de conservación. En segundo lugar, *L. major* comparte muchos más genes con *T. cruzi* que con *T. brucei*; esto es atribuible al modo de vida intracelular de ambos organismos (El-Sayed *et al.*, 2005). Por último, los genes exclusivos de cada especie son en su mayoría proteínas de superficie; por ejemplo, las VSG en *T. brucei*, vinculadas con la inmunoevasión, o las mucinas en *T. cruzi*, posiblemente relacionadas con la capacidad del organismo de tolerar el medio ácido del estómago del vertebrado en caso de infección por ingestión (Yoshida, 2009); lo que reafirma la idea de que se trata de proteínas que actúan en la relación con el hospedador.

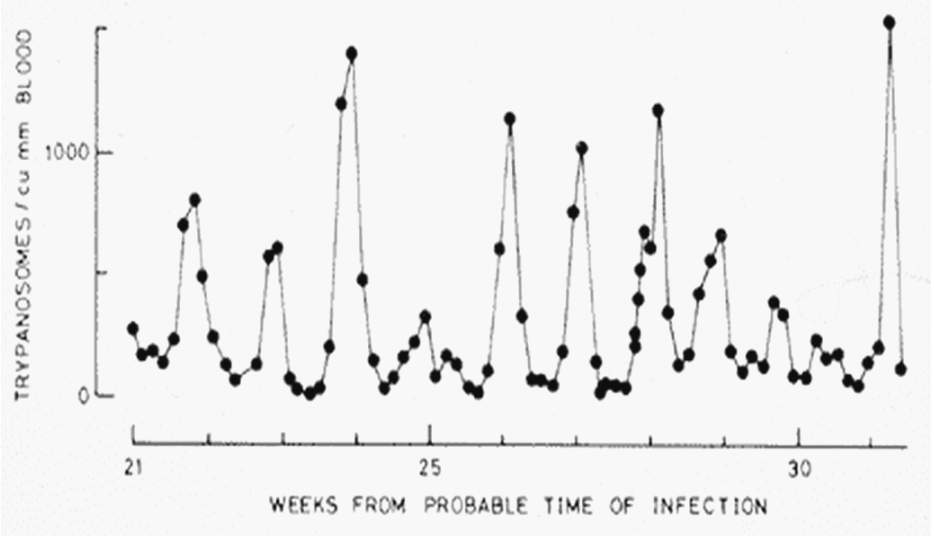


Figura 5. **Ciclos de variación periódica de la concentración de *T. b. gambiense* en sangre (modificado de Ross y Thomson, 1910).**

En cuanto a la sintenia, básicamente se puede decir que se encuentra bien conservada en el *core* cromosómico, y que se pierde hacia los extremos y en las regiones de *switch* transcripcional (ubicadas entre un policistrón y el siguiente), regiones en las que hay retroelementos, genes de ARN estructurales así como genes y pseudogenes especie-específicos, especialmente proteínas de superficie (El-Sayed *et al.*, 2005).

Un hecho remarcable de los tripanosomas africanos es que no presentan estadio intracelular en el hospedador vertebrado. Muy probablemente por este motivo han desarrollado una característica que los distingue: la variación antigénica. En efecto, mientras normalmente el sistema inmune tiene la capacidad de eliminar eventuales microorganismos invasores que ingresan al organismo, los tripanosomas africanos son capaces de producir infecciones de larga duración. Durante las mismas, el número de parásitos aumenta y disminuye drásticamente, siguiendo un patrón periódico que coincide con los picos de fiebre del mamífero infectado (figura 5). Si se estudian parásitos pertenecientes a distintos «picos» poblacionales, se ve que son antigénicamente distintos. La explicación de este fenómeno se encuentra en la densa cubierta de proteínas que envuelve totalmente su membrana celular. En otras palabras, cada tripanosoma individual está literalmente oculto bajo una capa de glicoproteínas variables de superficie (VSG). En cada célula hay unas 10^7 copias de la misma VSG (Pays *et al.*, 2004). El organismo tiene un reservorio de genes de VSG y muchos más pseudogenes y genes incompletos. En *T. brucei* se estiman más de un millar, del que sólo un 7 % codifica VSG totalmente funcionales (Berriman et al, 2005). En un momento dado, solo uno de esos genes se encuentra activo. Eventualmente, con una tasa de 10^{-3} cambios por división celular (Turner y Barry, 1989; Vanhamee *et al.*, 2001), el gen activo es sustituido por otro y la cubierta exterior pasa a estar constituida por una VSG diferente. La dinámica que rige la aparición de anticuerpos para una variante de VSG y el surgimiento de nuevas variantes, es lo que provoca ese ciclo de aumentos y disminuciones en el número de parásitos arriba mencionado.

El gen activo se encuentra en un sitio de ubicación telomérica denominado ES (*Expression Site*), y la VSG se transcribe junto a un grupo de genes asociados denominado ESAG (*Expression Site Associated Genes*). Existen, en *T. brucei*, unos 20 ES. El cambio en la

VSG que se expresa ocurre básicamente de tres maneras: 1) represión del ES activo y activación de alguno de los restantes, o *in situ switch*; 2) intercambio de telómeros entre el cromosoma que contiene el ES activo y otro; y 3) conversión génica, consistente en el reemplazo del gen VSG localizado en el ES activo por una copia de un gen de VSG del repertorio silencioso (Vanhamee *et al.*, 2001). Los tres mecanismos se ilustran en la figura 6.

Un tipo muy particular de conversión génica, que se presenta en etapas avanzadas de la infección, merece especial atención. Este mecanismo consiste en la creación de un gen a partir de fragmentos de diferentes copias de VSG, cada una de las cuales es un pseudogén, que por sí mismo es incapaz de codificar una proteína funcional. Sin embargo, la utilización de varios de estos pseudogenes permite combinar partes potencialmente funcionales de los mismos para formar una nueva copia completa de tipo mosaico. Esto implica que aquellas copias básicas de tipo pseudogén, o incluso copias que son solo fragmentos de secuencias codificantes de VSG, también contribuyan al repertorio de variabilidad antigénica de *T. brucei*, pudiendo generar así nuevas VSG no existentes previamente en el genoma (figura 7), siendo esta, además, una modalidad especialmente importante de variación antigénica en las etapas tardías de la infección (Marcello *et al.*, 2007).

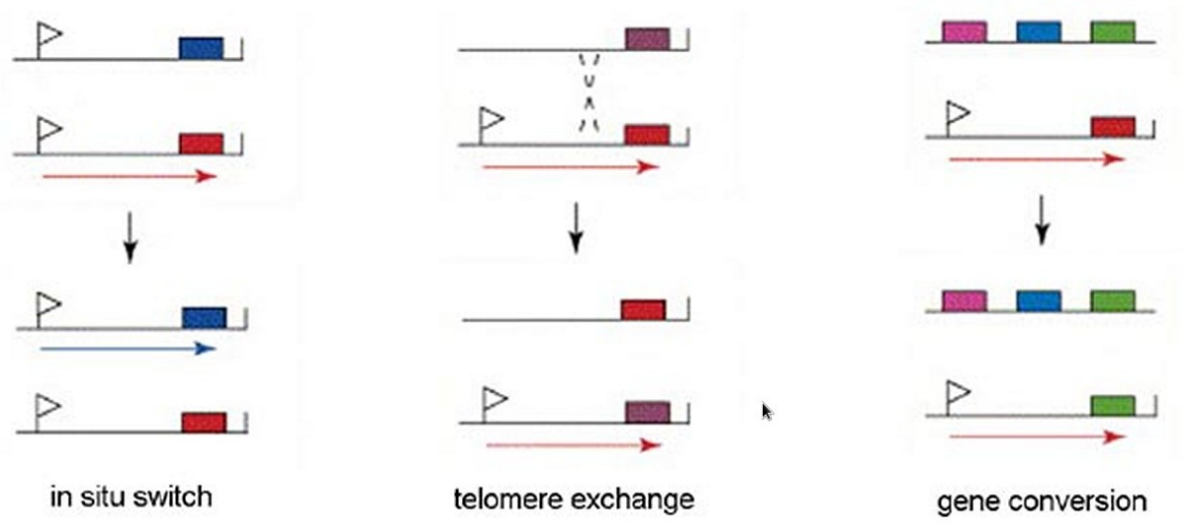


Figura 6. **Mecanismos de sustitución de la VSG activa (<http://www.tulane.edu/~wiser/protozoology/notes/kinet.html>).**

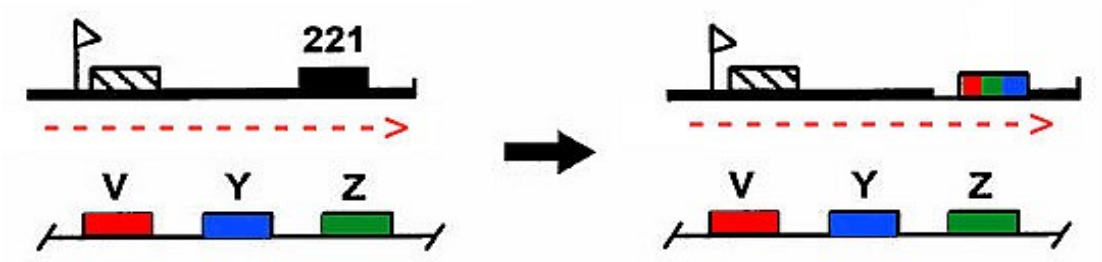


Figura 7. **Creación de gen «mosaico».**

Las VSG presentan muy baja similitud a nivel de secuencia. En muchos casos, dicha similitud es inferior al 20 % (tenerlo en cuenta resulta de vital importancia a la hora de anotar nuevas VSG; punto que se tratará más adelante en este trabajo). Sin embargo, esto no ocurre en su estructura terciaria (figuras 8 y 9), en la que se evidencia un alto grado de similitud. En general, las proteínas VSG poseen una región conservada (~ 100-120 aminoácidos), cerca del extremo C-terminal. Es en este extremo en que la proteína se une a la membrana celular a

través de una molécula de glicosilfosfatidilinositol (*GPI anchor*). El resto de la proteína (unos 360-380 aminoácidos) es una región variable (figura 10). De acuerdo a cómo se pliegan las proteínas y cómo se disponen sobre la superficie celular, la región conservada C-terminal resulta inaccesible, siendo la región variable N-terminal la única expuesta al sistema inmune del hospedador.

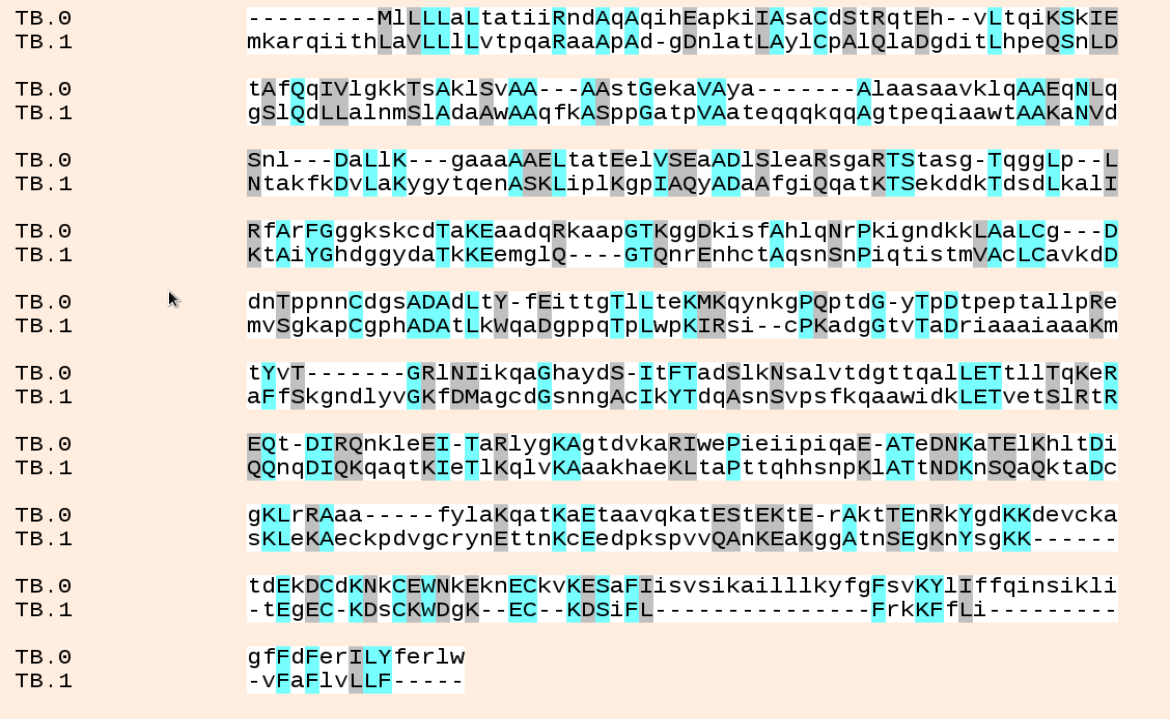


Figura 8. **Alineamiento entre dos proteínas de VSG, donde se evidencia visualmente la escasa similitud de secuencia.**

Es importante remarcar que el hecho de que las VSG presenten un bajo grado de similitud a nivel de secuencia aminoacídica, hace difícil e insegura su identificación por los métodos tradicionales basados, justamente, en dicha similitud. En su tesina de grado, Luisa Berná utilizó diversas aproximaciones a este problema, basadas en complementar la búsqueda por similitud de secuencia con otras técnicas que tienen en cuenta similitudes en los patrones de frecuencias aminoacídicas, así como de tri y tetranucleótidos. Fundamentalmente, lo que hizo

fue constatar, en primer lugar, que las frecuencias aminoacídicas de las VSG en *T. brucei*, difieren de las del resto de las secuencias codificantes de proteínas y, en segundo lugar, que dichas frecuencias presentan una clara homogeneidad a nivel interno de las propias VSG. Estas dos características permitían, dada una secuencia, afirmar con cierto grado de certeza si se trataba o no de una VSG lo cual, complementado con otros aspectos (similitud, aunque fuera baja, de secuencia con otras VSG conocidas, presencia de señales –signalPep, GPI anchor– que la identificaran como una proteína de superficie), se transformaba en una herramienta eficaz para identificar y anotar nuevas secuencias de VSG.

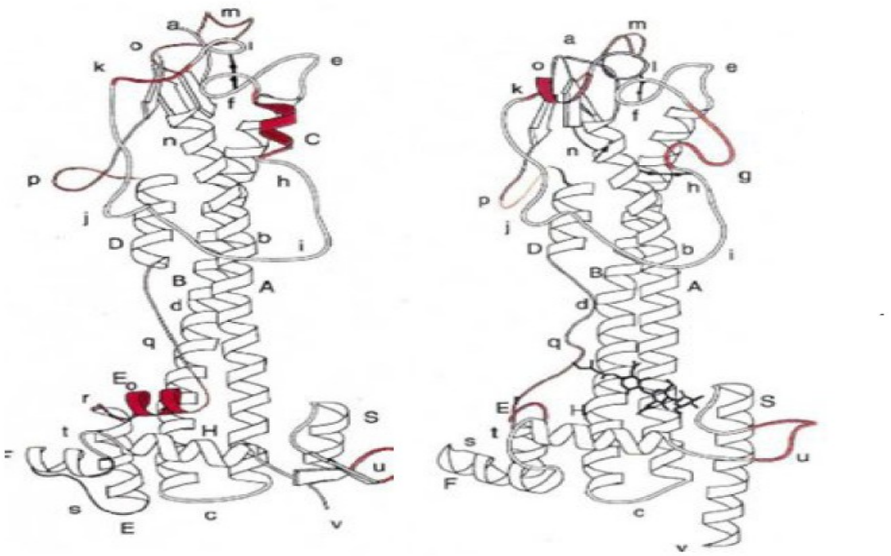


Figura 9. **Imagen de la estructura tridimensional de dos VSG. A este nivel el parecido es mucho mayor que a nivel de secuencias.**

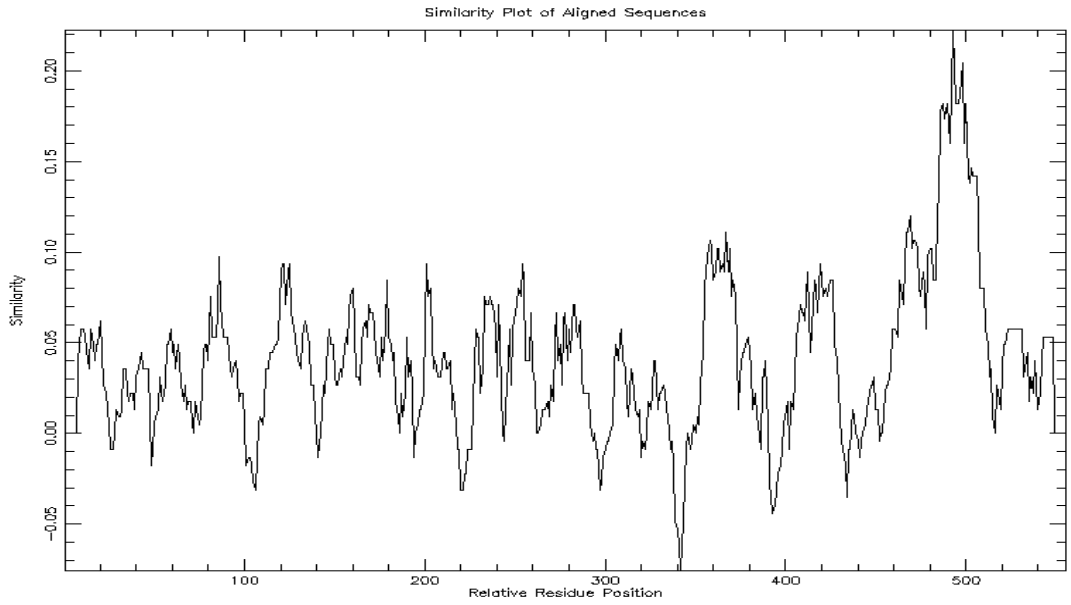


Figura 10. **Plot de similitud entre dos VSG alineadas. A la derecha, el extremo C-terminal, más conservado.**

En las siguientes dos figuras, vemos, resumidas, algunas de las pruebas que se realizaron en dicho trabajo: en la figura 11 se representan las frecuencias medias de aminoácidos de las proteínas de *T. brucei* (azul) y de las VSG del mismo organismo (rojo). Vemos que ambos perfiles no son coincidentes; y sin embargo no podemos saber si la variabilidad interna de cada grupo no es tan grande que, dado un caso concreto, no podríamos afirmar si pertenece a uno o a otro. Para ver esto se hizo un análisis de componentes principales (ACP) con dichas frecuencias. Los resultados de dicho análisis se muestran más adelante (ver *Resultados/Reanotación primaria de contigs genómicos*), y ponen en claro la efectividad de este tipo de análisis para identificar proteínas (o sus genes codificantes) que, como las VSG, presentan dificultades a la hora de realizar búsquedas mediante métodos convencionales basados en la similitud de secuencia.

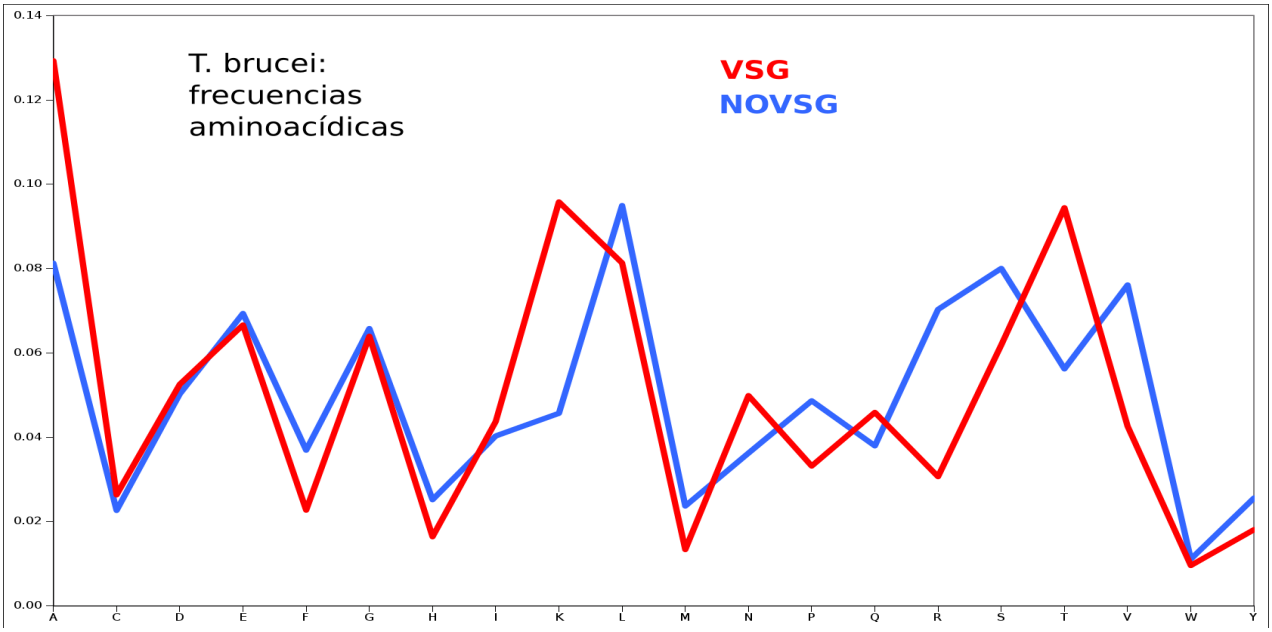


Figura 11. **Frecuencias aminoacídicas de las VSG (rojo) y de las proteínas no VSG (azul).**

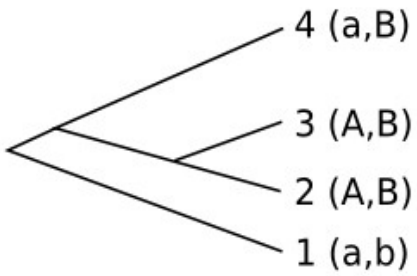


Figura 12. **Deducción de ancestralidad por el principio de parsimonia.**

Trypanosoma vivax

Para este trabajo decidimos tomar como modelo, en el estudio del origen y evolución de la variación antigénica, a *Trypanosoma vivax*, por ser un tripanosoma de origen africano (que, como se ha dicho, se ha extendido por América); y por tanto poseer también una cubierta protectora de VSG. La ubicación, filogenética de *T. vivax*, 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 cómo el mecanismo de variación antigénica se encontraba en su estado ancestral, lo que permitirá arrojar pistas importantes sobre la emergencia y diferenciación de las proteínas VSG. Esto es: de la comparación de caracteres presentes en un organismo tempranamente escindido de un grupo, con los mismos caracteres tal y como se presentan en el resto de los organismos de ese grupo, se pueden extraer inferencias filogenéticas acerca de las relaciones de ancestralidad de dichos caracteres. Veamos un ejemplo simple (figura 12). La especie **1** fue la primera en escindirse del grupo. Las letras representan dos caracteres (a y b) y sus variantes (mayúscula/minúscula). Podemos suponer, por parsimonia, que la variante **a** es ancestral frente a **A**, ya que se halla presente en la especie **1** y en una de las del resto del grupo. Lo contrario implicaría aceptar que **a** apareció dos veces, en forma independiente. En cambio, de la otra variante no tenemos información, ya que ambas posibilidades (**b** o **B** ancestrales) implican la misma cantidad de cambios. En todo caso, si la forma ancestral es **B**, la variante **b** tuvo mucho más tiempo para aparecer (lo representa el largo de la rama que conduce a **1**), ya que **B** debería haber surgido en el segmento que va desde el ancestro común del grupo hasta el ancestro de **2**, **3** y **4**, por lo cual podríamos suponer que es más probable que la forma ancestral sea **B**, aunque el argumento es menos contundente que en el caso del par (**a**, **A**).

Para el caso de *T. vivax* se puede tomar como punto de partida lo que se ha estudiado previamente en *T. brucei*, e intentar encontrar semejanzas y, especialmente, diferencias, analizando estas a la luz de que, siendo *T. vivax* un descendiente directo de la forma ancestral del género *Trypanosoma*, es de suponer que presenta algunas características que podemos considerar «arcaicas» de los tripanosomas africanos.

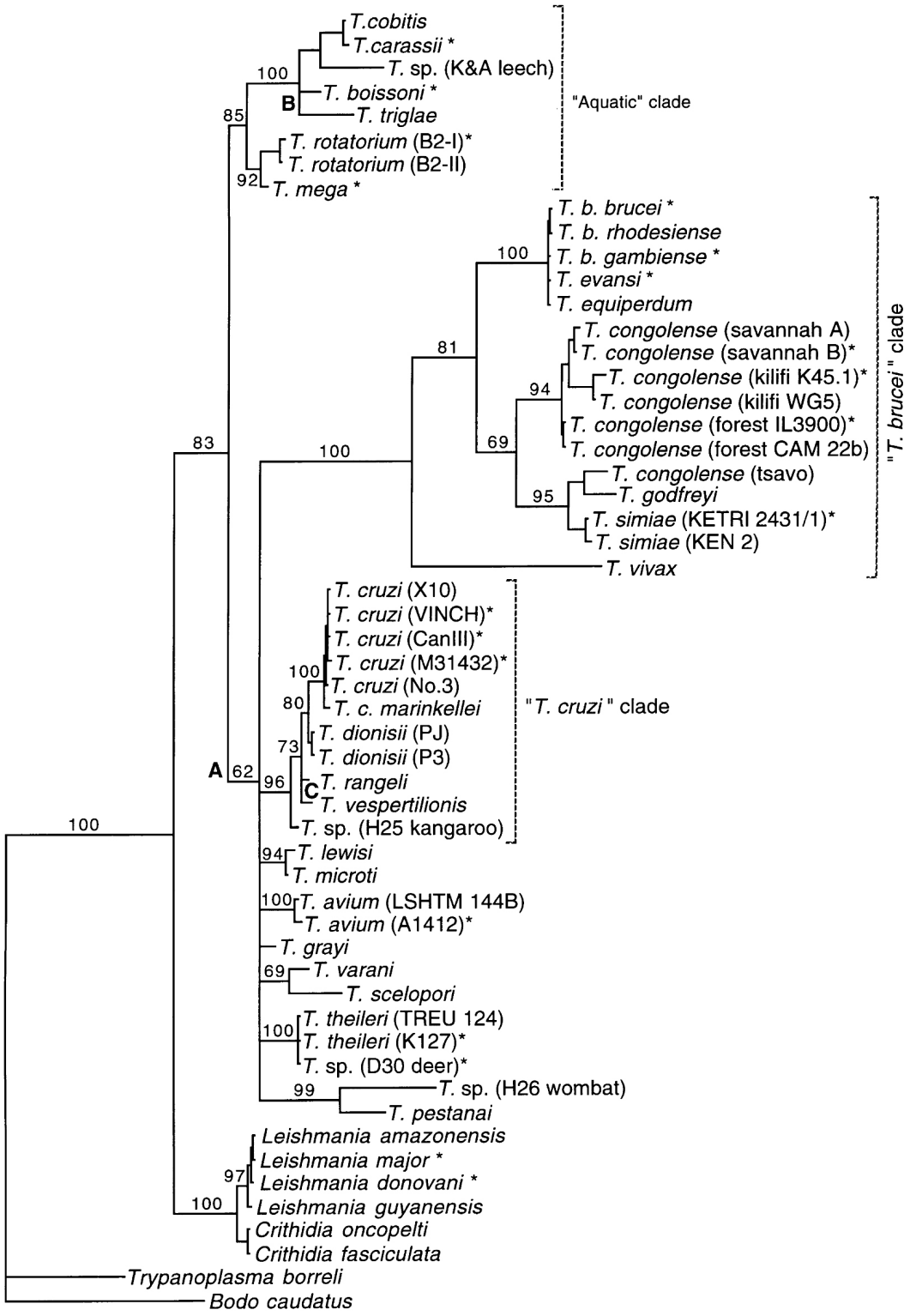


Figura 13. **Los distintos clados de tripanosomátidos y sus relaciones filogenéticas, según Stevens *et al.* (1998).**

Ambos organismos tienen una etapa definitiva en que viven en el torrente sanguíneo del hospedador vertebrado. El término «definitiva» se refiere a que no es, como ocurre en *T. cruzi*, un período de breve transición, en el cual los individuos pasan a la sangre en la que permanecen hasta que encuentran nuevas células para invadir, sino a que la población se establece allí en forma permanente. Si bien esto es una verdad a medias, ya que en las etapas superiores de la enfermedad puede haber invasión de distintos tejidos (por ejemplo, nerviosos), en ningún momento deja de haber tripanosomas en sangre, y en ningún caso estos adoptan una forma de parasitismo intracelular.

Materiales y métodos

Genoma de T. vivax

Se tomaron como base los datos de secuenciación genómica de la cepa Y486 (publicados por el Instituto Sanger, disponibles en el GenBank, <ftp://ftp.ncbi.nlm.nih.gov//genbank/wgs/>). Dichas secuencias están ensambladas en 2053 *scaffolds* (largo medio ~ 8000 pb), contenidos en 11 cromosomas (entre 65 y 475 scaffolds por cromosoma, en función del largo de estos) ensamblados por homología (conservación de sintenia) con los once cromosomas de *T. brucei*, y 8277 contigs que no fueron incluidos en ese ensamblaje (largo medio ~ 3000 bases de largo). Estos últimos contienen datos de anotación de 3998 genes (83 % «*hypothetical protein*», 2 % VSG, 3 % RHS y 12 % genes varios, de los cuales más de 1/3 son pseudogenes o secuencias fragmentarias). La anotación no posee datos de homología. El número de contigs con al menos una secuencia anotada es de 2850, y se

reparten según la tabla de la figura 14. A este *set* se agregaron nueve secuencias teloméricas de unas 30.000 bases de largo medio. Si bien hay cierta redundancia (uno de los 8277 contigs aparece íntegro en uno de los cromosómicos, y algún fragmento de 2000 o 3000 bases —perteneciente a contigs mucho más largos— aparece en los teloméricos), la misma es lo suficientemente baja como para que se puedan considerar grupos independientes. Como el conjunto contenía muchos contigs demasiado cortos como para realizar estudios estadísticos confiables, para muchos análisis se eliminaron todos aquellos que tuvieran menos de 1500 bases de largo, con lo que quedaron 6821.

secs. anotadas	contigs
0	5427
1	2151
2	460
3	134
4	56
5	25
6	10
7	6
8	6
9	2

Figura 14. **Número de contigs clasificados según la cantidad de secuencias anotadas que poseen.**

Obtención de ARN y secuenciación del transcriptoma de T. vivax

El ARN utilizado fue obtenido de la cepa LIEM-176, a partir de ovinos de Venezuela. Siguiendo los protocolos habituales, se realizó la purificación y se crearon las bibliotecas de 454 (Roche) e Illumina para su posterior secuenciado en Life Sequencing, Biópolis (Valencia, España) y en la Universidad de Washington (EE.UU.), respectivamente. Gonzalo Greif (IPMONT) participó en diversas etapas del proceso. La extracción de muestras fue realizada

por Dolores Piñeyro y Lucinda Tavárez en Venezuela (Centro de Estudios Biomédicos y Veterinarios, Universidad Nacional Experimental Simón Rodríguez-IDECYT, Caracas).

Desde que se inició la secuenciación automática (fines de los ochenta) y, especialmente con el reciente advenimiento de los métodos de secuenciación masiva caracterizados por proporcionar —a mucho menor costo y a mucha más velocidad que los métodos tradicionales — grandes cantidades de datos en forma de secuencias relativamente cortas (*reads*), se han desarrollado métodos cada vez más eficientes para ensamblar esas secuencias a partir de sus regiones solapantes. El constante avance del desarrollo de nuevos algoritmos y, muy especialmente, la mejora exponencial en el rendimiento de las computadoras, ha posibilitado manejar cantidades cada vez mayores de datos y trabajar con genomas a su vez más grandes y complejos.

En el caso del ARN se ven reducidos los problemas que genera la gran presencia de secuencias repetidas en las regiones intergénicas, que habitualmente dificultan el ensamblaje de *reads* genómicos. Sin embargo, debido a las particularidades del proceso de transcripción de los tripanosomátidos, encontramos una no despreciable cantidad de *reads* (o sea, en una etapa previa al ensamblaje) compuestos por repetidos cortos. Esto es: si bien el paradigma de que el ARN mensajero representa exactamente a las regiones codificantes no es exacto en ningún organismo, en el caso de los tripanosomátidos, con su mecanismo de transcripción policistrónica, ocurre que determinadas regiones intergénicas se transcriben a la par que las regiones codificantes.

En nuestro caso, utilizamos, para ensamblar los *reads* obtenidos del secuenciado del transcriptoma del secuenciador Roche, dos programas diferentes: Newbler (Roche, Suiza) y Mira (Chevreux *el al.*, 2004).

El total de *reads* obtenidos con Roche es de ~187.5 x 10^3, y su largo medio es de 287 nt (frente a aproximadamente 37 millones de reads de 36 nt de Illumina). El largo de los *reads* de Roche permite la aplicación de métodos de ensamblaje *de novo* (esto es, sin utilizar una secuencia de referencia como molde) tanto usando Newbler como Mira. En el análisis con Mira se obtienen 22063 contigs de un largo medio de 564 nt, mientras que queda sin ser ensamblado (sin incluirse en ningún contig) un 24 % de los *reads*.

En el caso de Newbler, se obtienen 8120 contigs de un largo medio de 627 nt, y queda sin ser ensamblado un 32 % de los *reads*.

Resumiendo, con Mira se obtienen más contigs y de menor largo medio (aunque de mayor largo total), y participa del ensamblaje final un mayor porcentaje de *reads*. Los números precisos se ven en la tabla de la figura 15.

	reads	roche ctgs	mira ctgs
cantidad	187491	8120	22063
largo medio (bases)	287	627	564
largo total (X 10^6 bases)	53.8	5.1	12.4
reads usados (%)		68	76

Figura 15. **Comparación de *reads* de Roche e Illumina,**

La calidad de los ensamblajes fue determinada comparando los contigs ensamblados con un conjunto de secuencias de ARN mensajero bien definidas. Se tomaron en cuenta dos variables: el *coverage o* proporción del ARN de referencia que es bien reconstruida (P) y el número de contigs (N) que caen en cada clase (las clases se definen por el coverage % y corresponden a las filas de la figura 16). A partir de estas dos variables se construyó un índice ($Q = \sum_i N_i P_i$) consistente en la suma de todos los productos de estas dos variables. Se utilizó un

conjunto de referencia formado por 684 secuencias de *T. vivax*, ortólogas de secuencias codificantes de proteínas de *T. brucei* (disponibles en el GenBank), que aparecen expresadas en el estado sanguíneo.

	mira			454 Newbler		
Total genes	684	% total	Q mira	684	% total	Q 454
Not Found	51	7.5		64	9.4	
Coverage						
10,00%	54	7.9	5.4	56	8.2	5.6
20,00%	135	19.7	27	156	22.8	31.2
30,00%	97	14.2	29.1	116	17.0	34.8
40,00%	92	13.5	36.8	89	13.0	35.6
50,00%	60	8.8	30	50	7.3	25
60,00%	44	6.4	26.4	34	5.0	20.4
70,00%	15	2.2	10.5	22	3.2	15.4
80,00%	15	2.2	12	11	1.6	8.8
90,00%	22	3.2	19.8	17	2.5	15.3
100,00%	99	14.5	99	69	10.1	69
Total found	633	92.5		620	90.6	

Figura 16. **Comparación de resultados de dos distintos *software* de ensamblaje.**

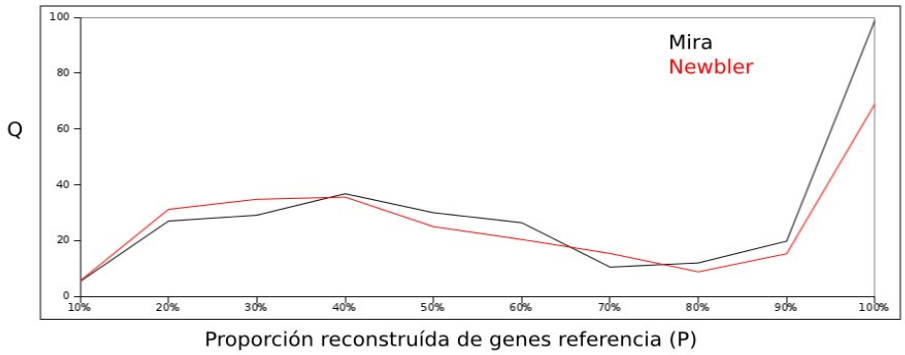


Figura 17. **Plot P vs Q para Mira y Newbler.**

En la figura 17 se ve gráficamente el comportamiento de la variable Q para Mira y Newbler. Queda claro que Mira se comporta, en general, mejor en los genes que fueron «bien reconstruidos» (a partir del 40 % de *coverage*).

Anotación funcional

Se denomina así al proceso de agregar información a una secuencia de ADN. Esto es, entre otras cosas, determinar qué regiones de dicha secuencia son codificantes, tienen una función reguladora, pertenecen a la categoría de elementos transponibles o presentan, simplemente, características dignas de mención (por ejemplo, «Región altamente repetida rica en AT»). Para el caso de secuencias codificantes de proteínas, además, se agrega información acerca de cuál es la función biológica de la proteína que codifica, el nombre de la proteína, la traducción, la función que cumple, qué dominios o motivos funcionales presentan, si es un gen funcional, una parte de un gen, o un pseudogén, e incluso en qué vía metabólica participa el producto de la transcripción y eventual traducción. Todo el proceso puede requerir diversos tipos de estudios biológicos, utilización del conocimiento previo sobre qué proteínas se encuentran en el organismo, y comparación con secuencias ya anotadas en organismos más o menos afines. Este último caso requiere especial cuidado, ya que la llamada «transferencia horizontal de información» va perdiendo precisión cuando nos alejamos del gen original cuya función se demostró empíricamente. Sin embargo, a medida que disponemos de mayores cantidades de datos genómicos de los organismos más diversos, y si se utilizan como fuente genomas de organismos muy cercanos y se trabaja con estrictos criterios de comparación, estas anotaciones previas son una fuente de información sumamente útil.

La anotación funcional de los contigs derivados del ensamblaje de ARN fue realizada a partir de un conjunto de herramientas complementarias: ESTscan (Iseli *et al.*, 1999), Blast2Go (Conesa *et al.*, 2005), InterProScan (Zdobnov *et al.*, 2001) y AnEnPi (Otto *et al.*, 2008).

En primer lugar, para identificar genes que codifiquen proteínas conocidas o no, es necesario contar con una traducción virtual de alta calidad. Esto no se consigue simplemente mediante la traducción mecánica de los marcos abiertos de lectura (ORF), debido a que errores de secuenciación pueden generar *indels* que provoquen corrimientos del marco de lectura, quitando sentido a la ulterior traducción. Para minimizar este factor, se recurrió al programa ESTscan que, utilizando *Hidden Markov Models* a partir de frecuencias de codones, detecta dichos corrimientos y los corrige introduciendo *indels* que restauran el marco de lectura correcto. El programa necesita ser entrenado sobre la base de un conjunto de genes conocidos, de modo que, con base en sus propiedades estadísticas, pueda detectar los errores. Para entrenarlo se utilizaron secuencias codificantes e intergénicas de *T. vivax* disponibles en bases de datos públicas. Posteriormente, se realizó la anotación funcional, en base a Blastp contra la base de datos nr (*non redundant)* del NCBI, y otras bases de datos bien curadas. También se realizó búsqueda de dominios basada en InterProScan, que se basa en la comparación de la secuencia *query* con una base de datos. Los resultados de ambas búsquedas se integraron usando el *software* Blast2Go, que permite introducir términos de GO (*Gene Ontology*) utilizando un vocabulario estructurado y estandarizado. Como Blast2Go es relativamente conservador a la hora de asignar términos GO, la anotación fue complementada con una simple búsqueda Blastp usando como base de datos de comparación la base nr (*non redundant*) traducida del NCBI. Paralelamente al pipeline AnEnPi, un instrumento que, entre otras cosas, permite identificar enzimas análogas no homólogas, vale decir, con igual función pero sin similaridad de secuencia, se utilizó KEGG (Kanehisa *et al.*, 2012), una colección de bases de datos genómicos asociados con vías metabólicas, con la intención de definir posibles rutas que se encuentren activas en el estado sanguíneo de *T. vivax*.

Nivel de transcripción

Una de las utilidades más directas de la secuenciación de ARN es la estimación de los niveles de transcripción de diferentes regiones del genoma. Esto se hace simplemente asumiendo que la cantidad de reads que caen dentro de un contig (o dentro de un CDS) es directamente proporcional al largo del contig y a la cantidad del ARN correspondiente que se encontraba en la célula en el momento de la extracción, es decir, su nivel de expresión. Por lo tanto, normalizando la cantidad de reads por el largo, se obtiene un estimativo confiable de cuánto se transcribe cada una de las partes del genoma.

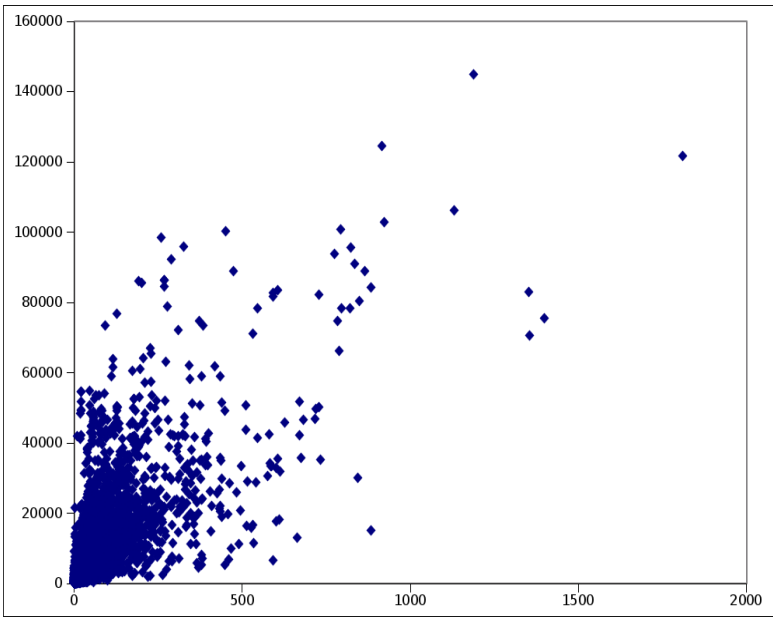


Figura 18. **Cantidad de reads Roche por contig (abcisas) versus la cantidad de reads Illumina por contig (ordenadas).**

Los resultados de las dos secuenciaciones, Roche e Illumina, permiten tener estimaciones independientes de estos valores, los cuales fueron comparados. En general existe una gran similitud entre ambos; si contamos los reads de uno y otro correspondientes a cada contig, obtenemos una correlación de ~0.6 (p~0), que aumenta hasta valores cercanos a 0.8

cuando se eliminan unos pocos *outliers,* siendo los contigs con extremo sesgo composicional y alta presencia de repetidos los que muestran mayores diferencias entre las cantidades de uno y otro conjunto de reads.

Sobre los datos de secuencia obtenidos con la tecnología Illumina se aplicó el *software* Erange (Mortazavi *et al.*, 2008) para estimar niveles de expresión, corrigiendo los asignamientos erróneos de reads a familias multigénicas en las que no necesariamente todos los parálogos se están expresando por igual.

Se eliminaron los reads de baja calidad y se mapeó el resto sobre el genoma de *T. vivax* disponible en Genbank.

En el caso de los datos de Roche dicho problema (el de la asignación incorrecta de reads a regiones similares) no existe o es mínimo, debido al mayor largo de los reads.

Se escribieron scripts de Bash y Perl para parsear las salidas de Blast y Bowtie (ver Bioinformática) a los efectos de estimar el nivel de transcripción a partir del número de reads.

Análisis estadístico

Los análisis fueron realizados sobre plataformas Matlab (ACP) y R (análisis de clusters) (R Core Team, 2012).

Análisis de componentes principales

Este análisis se utiliza para reducir la cantidad de variables que explican la varianza de un conjunto de datos, convirtiéndolo en una síntesis más inteligible del conjunto original. Consiste básicamente en una transformación lineal que, a partir de una matriz de correlaciones o de covarianza (en este caso utilizamos la matriz de correlaciones) genera un nuevo sistema de

coordenadas en el que las nuevas variables se ordenan en función de su aporte a la varianza general (componente uno, dos, etc.). Suele ocurrir que gran parte de la varianza global se explique en función de unas pocas variables, pudiendo descartarse las demás en la explicación. Posteriormente, se puede intentar dar «significado» a estas variables, correlacionándolas generalmente con variables conocidas (que pueden o no estar presentes en el modelo original), lo que permite explicar la varianza global. Pero también se puede utilizar el método, simplemente, para corroborar que las propiedades estadísticas de determinado subconjunto de casos son lo suficientemente homogéneas y distintas a las del resto como para que se puedan utilizar como factor discriminante, de clasificación o detección. Esto se realiza mediante la observación del gráfico 2D o 3D resultante (para los que se pueden seleccionar distintos componentes como ejes) que, en Matlab, además, se puede realizar variando el ángulo de observación y coloreando los grupos previamente definidos con el fin de facilitar la comprensión. Dicha definición previa de grupos se puede realizar a través de otros métodos estadísticos (por ejemplo, análisis de clusters) o por el conocimiento previo que tenemos de los casos utilizados (en otras palabras: a determinado grupo de secuencias —por ejemplo, las que codifican VSG— se les adjudica un color).

Como variables para realizar dicho análisis se utilizaron las frecuencias de trinucleótidos fuera de marco para el caso de los contigs, y frecuencias de codones (es decir, trinucleótidos en marco) para los ORF. Esto es, haciendo frecuencias de codones, en la secuencia AGGGGT tenemos 0.5 AGG y 0.5 GGT, pero si utilizamos trinucleótidos tendremos 0.25 AGG, 0.5 GGG y 0.25 GGT. La utilidad de utilizar frecuencias de codones radica en que se pueden diferenciar los perfiles de un marco con respecto al otro. Así, dos ORF prácticamente superpuestos (empezando y terminando muy cerca uno de otro, pero en distinto marco) son claramente

diferenciables por sus propiedades estadísticas, mientras que si utilizáramos frecuencias fuera de marco serían prácticamente indistinguibles, y muy probablemente aparecerían como el mismo tipo de ORF. En el caso de los contigs, no tiene sentido dar preferencia a un marco de lectura por sobre los otros, por lo que se prefiere utilizar frecuencias de trinucleótidos. Los CDS, en una región del genoma (en este caso hablamos de contigs genómicos), se pueden ubicar en cualquiera de los marcos de lectura, por lo que la información contenida en las frecuencias de codones «en marco» se anula si los consideramos a todos como en un mismo marco. Incluso, en este caso, se trabajó no directamente sobre los contigs, sino sobre una variante de estos (contigs «espejados»), consistente en crear un falso contig del doble de largo, que contiene la secuencia y su hebra reversa complementaria. De este modo, una secuencia AAACTG se convierte en AAACTGCAGTTT. Mediante este paso previo evitamos que contigs muy similares pero de los cuales fue secuenciada distinta hebra, aparezcan como totalmente distintos, formando por lo tanto parte de diferentes «nubes» o agrupamientos de puntos en el gŕafico del ACP, ya que las frecuencias de bases o palabras pueden diferir mucho entre ambas hebras de un contig (para el caso, tendríamos 0.25 AAA, 0.25 AAC, 0.25 ACT y 0.25 CTG, y 0 para todos los demás tripletes posibles, mientras que en la complementaria habría CAG, AGT, GTT y TTT en la mismas proporciones). La versión «espejada» de ambos contigs será, sin embargo, muy similar (en este caso, exactamente igual), y por lo tanto tendrá una misma «huella» de frecuencias de trinucleótidos.

Análisis de clusters

Este tipo de análisis permite dividir una muestra en un pequeño número de grupos, dentro de los cuales los casos difieren poco entre sí, en comparación con su distancia con casos pertenecientes a otros grupos. Existen métodos denominados jerárquicos, que separan primero grupos y dentro de estos, a su vez, crean subgrupos, de modo que el resultado puede representarse como un árbol (de aspecto similar, por ejemplo, a los árboles filogenéticos). Por otra parte, hay métodos no jerárquicos, que simplemente adjudican a cada caso la pertenencia a un grupo. Aquí usamos el método no jerárquico K-means. En este caso utilizamos frecuencias de codones porque estábamos trabajando con ORF. Este tipo de análisis parte de un conjunto de observaciones (siendo cada una un vector; para el caso de codones, en principio, de 64 dimensiones, si se incluyen codones STOP, o 61 si se excluyen), y las agrupa en k conjuntos (k<n) o *clusters*, que se crean siguiendo la regla de mínimos cuadrados: la diferencia entre cada valor y la media del grupo se eleva al cuadrado, y la suma de todos esos cuadrados, debe ser la mínima posible. El usuario debe definir previamente el valor de k (cuántos grupos desea), lo cual se realizó a partir de la inspección ocular del gráfico resultante del ACP. El algoritmo aquí utilizado es el de Hartigan-Wong (Hartigan y Wong, 1979), según la implementación provista en el paquete R.

Generación de secuencias por simulación

La simulación de secuencias se realizó utilizando scripts de Perl (ver detalles en Apéndice), usando probabilidades condicionales y cadenas de Markov.

Cuando generamos mediante simulación secuencias aleatorias podemos usar diversos modelos. Por ejemplo, podemos crear una secuencia que posea algunas de las propiedades

estadísticas de otra secuencia real, de forma tal que podemos poner a prueba algunas hipótesis sobre el comportamiento de dichas secuencias reales. Con el fin de generar secuencias que tengan la misma composición de nucleótidos que una secuencia real, utilizamos como probabilidades estimadas las frecuencias de bases de las secuencias reales. En estos casos lo que hacemos es ir colocando los nucleótidos uno a uno, y a cada nuevo nucleótido se le asigna como estado una de las cuatro bases de acuerdo a las probabilidades estimadas. A esto se le llama también cadenas de Markov de orden cero. Sin embargo, el modelo puede incorporar otros elementos más realistas: podemos, por ejemplo, modelar secuencias en las que la ocurrencia de cada una de las bases dependa de la base que se encuentre en el sitio previo, los dos previos, etcétera.

Se llama probabilidad condicionada a la probabilidad de que ocurra un evento A dado que también ocurrió un evento B, y se define como

$$p(A|B) = p(A \cap B) / p(B)$$

lo cual se lee como «probabilidad de *A*, dado *B*, es la probabilidad de *A* intersección *B*, sobre la probabilidad de *B*». $p(B)$ representa, por ejemplo, la probabilidad de sufrir un accidente de tránsito. $p(A \cap B)$ es la probabilidad de que cualquier persona muera algún día en un accidente de tránsito, lo cual se puede inferir de datos estadísticos: muertes en accidentes de tránsito sobre muertes totales. $p(A|B)$ es la probabilidad de que muera *una vez que sabemos que sufrió dicho accidente.*

Se denominan *cadenas de Markov* a determinados procesos estocásticos en los que se cumple la *condición de Markov*, que dice que el estado de un proceso depende del estado anterior, sin importar el resto de la historia, y se nota formalmente así:

$$P(X_n = j \mid X_{n-1} = i_{n-1}, \ldots, X_0 = i_0) = P(X_n = j \mid X_{n-1} = i_{n-1})$$

Para definir una cadena de Markov es necesario determinar un conjunto (E) de *estados* del sistema (los distintos valores que puede tomar las variable aleatoria X; para el caso de una secuencia de ADN, los valores son A, T, G y C). El cambio de un estado a otro se denomina *transición*. Hay que definir también el *estado* inicial, que consiste en un vector fila con las probabilidades de que X_n tenga cada uno de los valores de E. Por último, se define la probabilidad de cambio de un estado a otro. Cuando las probabilidades de transición son independientes del tiempo (o sea, invariables a lo largo del proceso) decimos que la cadena es *homogénea*. Volviendo al ADN, si queremos generar una secuencia aleatoria en la cual cada nuecleótido dependa del anterior, podemos modelarlo como una cadena homogénea de orden 1, y tendremos una matriz, válida para todas las posiciones, como la que sigue:

	A	**T**	**G**	**C**
A	p_{aa}	p_{at}	p_{ag}	p_{ac}
T	p_{ta}	p_{tt}	p_{tg}	p_{tc}
G	p_{ga}	p_{gt}	p_{gg}	p_{gc}
C	p_{ca}	p_{ct}	p_{cg}	p_{cc}

donde las filas suman 1, y, por ejemplo, p_{at} representa la probabilidad de la trasición $A->T$, que expresa, en este caso, la probabilidad de que en un sitio haya una T dado que en el sitio anterior hay una A. Este se puede estimar a partir de las frecuencias empíricas, específicamente la probabilidad condicional mencionada más arriba $(P(A->T))$, estará dada por la frecuencia del dinucleótido AT (el cual es un estimador de $p(T \cap A)$ sobre la frecuencia de la base A $(p(A))$.

Sintéticamente, agreguemos que se puede hablar de probabilidades de transición de distintos órdenes. El caso aquí definido correspondería al primer orden; en las transiciones de segundo orden la probabilidad de un evento depende de los dos casos (nucleótidos) anteriores. Asimismo, y tal cual fue mencionado anteriormente, se puede denominar probabilidad de transición de orden 0 a la que no depende en absoluto de los casos anteriores. Sobre la base de estos tres modelos construimos los scripts utilizamos para crear secuencias simuladas.

En otras palabras, lo que vamos a crear es una secuencia aleatoria que, sin embargo, posee las mismas frecuencias ya sea de bases, de dinucleótidos o trinucleótidos que una real, con el fin de averiguar si otras propiedades que observamos en dicha secuencia modelo están relacionadas con dichas frecuencias. En el caso de este trabajo, como se describirá más adelante, se utilizan para ver si la composición a nivel de nucleótidos, dinucleótidos y trinucleótidos puede explicar un fenómeno concreto al que denominamos *orfogenicidad*.

Bioinformática

Blast

El nombre viene de las siglas en inglés de *Basic Local Alignment Search Tool* (Altschul *et al.*, 1990), y es probablemente la herramienta bioinformática más ampliamente utilizada. Con versiones en línea y de instalación local, este programa es usado para encontrar, en una base de datos, fragmentos homólogos (que comparten ancestría) a uno determinado (denominado *query*), basándose en la similitud de secuencia. Existen diversas variantes, dependiendo del grado de similitud entre la secuencia *query* y las que espero encontrar por azar, y también permite realizar búsquedas adn-adn (blastn) o proteína-proteína (blastp), e

incluso usando —traducción mediante— uno de estos tipos de secuencia como query y el otro como base de datos (blastx, tblastn).

Se trata de un programa heurístico, que intenta una solución de compromiso entre la exactitud y la velocidad computacional. Se basa en dividir al query en una serie de minisecuencias («palabras» o «semillas»), que se cotejan contra la base de datos. En caso de existir en la base esas mismas secuencias, se irá ampliando el alineamiento hacia ambos lados hasta que cierto valor de similitud caiga por debajo de determinado umbral. De ahí la expresión «*local alignment*»: no se trata de alinear lo mejor posible la secuencia completa, sino los segmentos que encuentren similitud con segmentos de las secuencias de la base de datos.

El Blast, además, suele formar parte de numerosos *pipelines* (programas ejecutados en secuencia, donde la salida de uno es la entrada del siguiente) utilizados, por ejemplo, en la creación de COG (*Clusters of Orthologs Groups*) entre distintas especies, o para búsqueda de pares parálogos gen-pseudogén en estudios de sesgos mutacionales. Uno de los usos más frecuentes (especialmente en el claso del Blast *online*) es cuando tenemos una secuencia desconocida y queremos averiguar su probable función a partir de su semejanza con otras secuencias conocidas.

Artemis

Se trata de una herramienta de anotación y visualización genómica, que permite superponer a la secuencias de ADN detalles de anotación, diversos gráficos (GC, GCsqew, etc., así como gráficos de índices desarrollados por el usuario); además de encontrar ORF, corregir manualmente su extensión y guardarlos en diversos formatos, como GBK o EMBL (Rutherford *et al.*, 2000).

FigTree

Visualizador y editor de árboles filogenéticos, desarrollado inicialmente por Vikas Raykar y Changjiang Yang, y en las versiones más recientes por Vlad Morariu (Morariu *et al.*, 2008). La versión aquí utilizada está disponible en: <http://tree.bio.ed.ac.uk/download.html?name=figtree&version=v1.3.1&id=82&num=3>.

Emboss

La *European Molecular Biology Open Software Suite* (<http://emboss.sourceforge.net/>) es un conjunto de programas de diverso uso para el análisis y procesamiento de secuencias. En este trabajo hemos usado principalmente, con fines exploratorios o de análisis complementarios, los programas Needle (alineamiento de dos secuencias de ADN por el método Needle-Wunsch), Plotcon (grafica el grado de conservación de un alineamiento), Transeq (traduce de ADN a proteína), Getorf (a partir de una secuencia dada, genera un fasta con los ORF encontrados, ya sea de STOP a STOP o de Met a STOP, con diversos parámetros modificables por el usuario (largo mínimo, tipo de secuencia —ADN o proteína— del archivo de salida, etc.) y distmat (crea una matriz de distancias a partir de un alineamiento). En este marco, también se usó el programa de alineamientos múltiples para ADN y proteínas clustalw (Thompson *et al.*, 1994; Larkin *et al.*, 2007).

Bowtie

Debido a que las nuevas tecnologías de secuenciación de ultra-alto-rendimiento producen volúmenes de datos en cantidades enormes, no es posible realizar ni la tarea de

ensamblaje ni el mapeo de reads sobre un genoma utilizando los alineadores tradicionales (tales como Blast), que no están diseñados para lidiar con cantidades de datos tan elevadas (un solo *lane* del secuenciador Illumina Hiseq2000 puede producir doscientos millones de reads). Programas como Bowtie (Langmead *et al.*, 2009) o BWA (Li y Durbin, 2009) se usan para alinear estas grandes cantidades de reads cortos sobre un genoma. Estos programas se basan en el uso de la transformada de Burrows–Wheeler, que consiste en un reordenamiento de los elementos de una string de modo que todas las secuencias que comparten el mismo sufijo (en otras palabras, que terminan igual) tienden a aparecer juntas en la matriz resultante, facilitando así el proceso de búsqueda de strings. Además, la transformación de BW se utiliza en algoritmos de compresión de datos, debido a su reversibilidad y a que la secuencia transformada contiene más repeticiones consecutivas de letras, lo que la hace más factible de ser expresada de un modo más conciso. A modo de ejemplo, la secuencia AAAACCCAATTTGGGG (16 letras) se puede expresar en la forma abreviada a4c3a2t3g4 (10 letras), sin pérdida de información.

Por la forma en que están ordenadas las secuencias, tenemos que las substrings con el mismo sufijo tienden a quedar juntas dentro de una zona de la matriz (figura 19, centro). Obsérvense, a modo de ejemplo, las letras «c», a la derecha, o la secuencias «cgc» y «gca». Esto es aprovechado por el algoritmo para realizar una búsqueda más eficiente de substrings, lo que, en nuestro caso, implica alinear más rápidamente y con menor consumo de memoria, millones de reads a un genoma de referencia.

c t c g c a c g c a c $

$	c	t	c	g	c	a	c	g	c	a	c
c	$	c	t	c	g	c	a	c	g	c	a
a	c	$	c	t	c	g	c	a	c	g	c
c	a	c	$	c	t	c	g	c	a	c	g
g	c	a	c	$	c	t	c	g	c	a	c
c	g	c	a	c	$	c	t	c	g	c	a
a	c	g	c	a	c	$	c	t	c	g	c
c	a	c	g	c	a	c	$	c	t	c	g
g	c	a	c	g	c	a	c	$	c	t	c
c	g	c	a	c	g	c	a	c	$	c	t
t	c	g	c	a	c	g	c	a	c	$	c
c	t	c	g	c	a	c	g	c	a	c	$

ordeno por columna 1

$	c	t	c	g	c	a	c	g	c	a	c
c	$	c	t	c	g	c	a	c	g	c	a
c	g	c	a	c	g	c	a	c	$	c	t
c	g	c	a	c	$	c	t	c	g	c	a
c	a	c	g	c	a	c	$	c	t	c	g
c	a	c	$	c	t	c	g	c	a	c	g
c	**t**	**c**	**g**	**c**	**a**	**c**	**g**	**c**	**a**	**c**	**$**
g	c	a	c	g	c	a	c	$	c	t	c
g	c	a	c	$	c	t	c	g	c	a	c
a	c	g	c	a	c	$	c	t	c	g	c
a	c	$	c	t	c	g	c	a	c	g	c
t	c	g	c	a	c	g	c	a	c	$	c

c a t a g g $ c c c c c

proceso inverso

1	2		3		4			5		
$	**c**	$	$	c	**c**	$	c	$	c	t
c	**a**	c	c	$	**a**	c	$	c	$	c
c	**t**	c	c	g	**t**	c	g	c	g	c
c	**a**	c	c	g	**a**	c	g	c	g	c
c	**g**	c	c	a	**g**	c	a	c	a	c
c	**g**	c	c	a	**g**	c	a	c	a	c
c	**$**	c	c	t	**$**	c	t	c	t	c
g	**c**	g	g	c	**c**	g	c	g	c	a
g	**c**	g	g	c	**c**	g	c	g	c	a
a	**c**	a	a	c	**c**	a	c	a	c	g
a	**c**	a	a	c	**c**	a	c	a	c	$
t	**c**	t	t	c	**c**	t	c	t	c	g

Figura 19. ***Izquierda*****: todas las rotaciones posibles de la string.** ***Centro*****: ordeno alfabéticamente según la primera columna. La última columna contendrá muchas series de caracteres repetidos, debido a la estructura del lenguaje: por ejemplo, en español, para las «ele» de la primera columna —que están todas juntas porque acabo de ordenarlas—, habrá muchos grupos de «e» en la última, debido a que la palabra «el» es muy común. Esta forma de expresión de la secuencia original, por ser más repetitiva, es más fácilmente comprimible.** ***Derecha*****: proceso reverso. Ordeno la versión codificada, obteniendo (1) la primera columna de la matriz del centro; (2) le agrego a la izquierda la misma sin ordenar; (3) ordeno por la la de la izquierda (tengo ahora las dos primeras columnas) y (4, 5) repito el proceso hasta que, al final, vuelvo a obtener la matriz del centro. La fila que termina en $ (en negritas) es la string original.**

Scripts personalizados

Además de estas herramientas, se crearon especialmente un sinnúmero de scripts en lenguajes Perl y Bash, como traductores, simuladores, generadores de secuencias “espejo” (consistentes en una secuencia y su reversa complementaria, con el fin de realizar determinados análisis multivariados, como el ACP, cuando el material disponible consiste en fragmentos que pueden existir en las versiones de una u otra hebra), parseadores de salida Blast, contadores de frecuencias de bases, di o trinucleótidos, codones o aminoácidos, contadores de GC global y en las posiciones 1, 2 y 3 de secuencias «en marco», y calculadores del índice AT/TA de una secuencia, entre otros.

En el Anexo se detalla, a modo de ejemplo, el funcionamiento de dos de esos scripts.

Planteo del problema y objetivos

Sin duda, es de sumo interés entender a fondo el proceso de la variación antigénica que acabamos de describir, así como elaborar hipótesis acerca de su evolución; en concreto, cómo surgió, si funciona de la misma forma en distintas especies, o en qué consisten las diferencias, en caso de haberlas. También nos interesa estudiar la organización genómica de estos organismos con particular énfasis en el segmento que contiene el repertorio silencioso de genes VSG.

Tanto *T. vivax* como *T. brucei* utilizan las VSG como cubierta protectora, lo que produce el ciclo de picos de parásitos en sangre, descritos anteriormente, y que se correlacionan clínicamente con la recaídas y picos de fiebre del mamífero infectado. Sin embargo, observaciones de microscopía electrónica sugieren que la capa de VSG de *T. vivax* es menos densa que la de *T. brucei* (Vickerman, 1976).

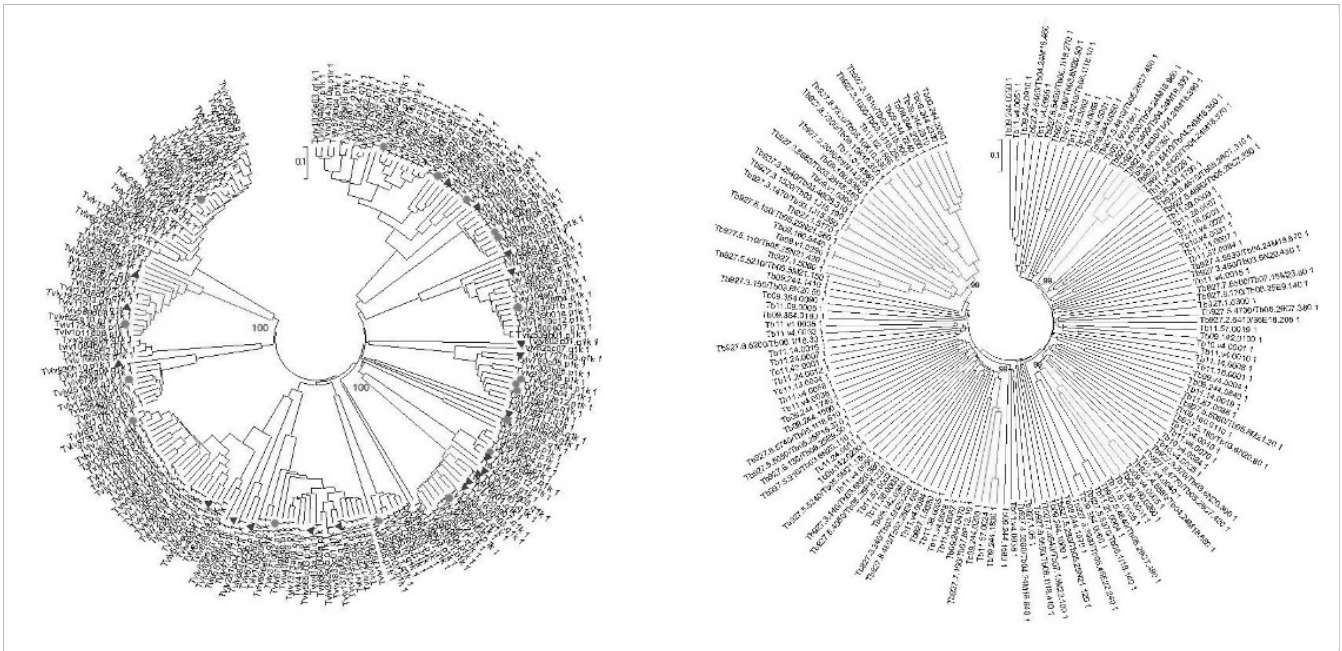


Figura 20. **Árboles construidos (Neighbor-Joinig) a partir de secuencias aminoacídicas de VSG de *T. vivax* (izq.) y *T. brucei* (der.)**

Otra diferencia es que las VSG de *T. vivax* (figura 20, izq.) se agrupan en clusters, dentro de las cuales la similitud de secuencias es mayor, mientras que las de *T. brucei* (figura 20, der.) se organizan en un árbol de tipo «en estrella», donde la similitud de dos VSG tomadas al azar tiende a mantenerse constante, y a un nivel cercano al 20 %. En el caso de *T. brucei*, a simple vista, puede apreciarse que no existen grandes agrupamientos: cualquier par de genes tomados al azar tenderá a presentar el mismo grado de divergencia que cualquier otro par. En cambio, cuando observamos las relaciones filogenéticas al interior de *T. vivax*, entre las distintas copias de genes VSG, podemos apreciar una diferencia notoria con *T. brucei*. El genoma de *T. vivax* contiene agrupamientos muy claros, dentro de las cuales las distintas secuencias exhiben niveles de divergencia aminoacídica que varían entre 20 y 30 %. Sin embargo, cuando comparamos secuencias pertenecientes a distintos agrupamientos, las distancias son similares a aquellas que presentan los genes de *T. brucei*. Esto habla de distintos mecanismos utilizados por estos organismos para crear y mantener su acervo de VSG.

En un trabajo reciente (Jackson *et al.*, 2012), se hipotetiza que las VSG de unos y otros tripanosomas, así como ciertas proteínas emparentadas (que se suponen VSG que perdieron funcionalidad o la cambiaron por otra), derivan de varios linajes que ya estaban presentes en el ancestro común, pero los distintos caminos que tomó cada uno de esos linajes en las distintas especies hacen que, por ejemplo, algunas VSG de *T. vivax* estén más emparentadas con proteínas que en *T. brucei* han perdido la función de VSG, que con otras VSG de *T. vivax*.

En *T. brucei* los genes de VSG ocupan regiones del genoma específicas, por lo general de ubicación subtelomérica y separadas de aquellas regiones en que están los genes *housekeeping*, y en general los genes de función conocida que tienen homólogos en otras

especies. Habíamos dicho también que en esta región existen muchos RHS y transposones en general. Una pregunta que intentaremos responder es si la organización del espacio genómico que ocupan las VSG es similar, en *T. vivax*, a lo que se ha observado en *T. brucei*.

Un aspecto crucial del mecanismo de evasión del sistema inmune es la forma en que cambia la VSG que se está expresando en un momento dado. Mediante datos de transcriptoma obtenidos por nuestro equipo a partir de una cepa —LIEM-126, de Sudamérica—, se podrá corroborar si en *T. vivax* es aplicable el modelo de *Expression Site* descrito en *T. brucei*. ¿Se expresa sólo una VSG por vez? ¿Cómo se regula su expresión?

Otro asunto a atender se refiere a la diversidad intraespecífica de *T. vivax*. Diversos estudios sugieren que, dentro de África, existen al menos dos formas principales de *T. vivax*, la oriental y la occidental (Malele *et al.*, 2003; Cortez *et al.*, 2006). La cepa Y486 pertenece al tipo occidental, y es muy cercana a la presentes en el continente americano. Dicha diversidad (además de su interés evolutivo), es de crucial importancia a la hora de desarrollar fármacos contra la tripanosomiasis. La bibliografía disponible en tal sentido no es muy amplia, y existen datos contradictorios, probablemente debidos a las distintas técnicas aplicadas en diferentes estudios (Hamilton, 2012).

En resumen, los objetivos de este trabajo son:

Estudiar la organización genómica de los tripanosomátidos, haciendo hincapié en *Trypanosoma vivax*.

Contribuir al esclarecimiento del origen de las VSG.

Comparar los repertorios de VSG entre *T. vivax* y *T. brucei*.

Identificar y estudiar la organización genómica de las VSG en *T. vivax*, así como sus mecanismos de regulación de transcripción y expresión.

Resultados

Para comprender la variación antigénica en *T. vivax* es necesario, en primer lugar, identificar genes codificantes de proteínas VSG, que son la materia prima sobre la cual se produce dicha variación. Posteriormente, estaremos en condiciones de analizar otros aspectos tales como su disposición en el genoma, junto a qué genes se encuentra, y cómo se regula su transcripción.

Con el fin de afinar esta búsqueda, procedimos como se detalla a continuación.

Reanotación primaria de contigs genómicos

Como se describe en la sección «Materiales y métodos», la anotación de los contigs disponibles en el GenBank es sumamente limitada, por lo cual se procedió a realizar una anotación por homología basada en los resultados de la anotación de los contigs de ARN descrita en la misma sección, obteniéndose 6821 archivos en formato GBK (visualizables en el programa Artemis). En total, se encontraron unos 12.000 CDS que son potenciamlente codificantes, ya sea con producto conocido u ORF conservados en especies afines. Asimismo, se identificó más de un millar de genes candidatos a VSG (incluyendo pseudogenes, con indels o codones STOP internos, y secuencias incompletas de VSG); cerca de 600 RHS y una enorme cantidad (más de 180.000) de ORF «huérfanos». Esto es, que no muestran homología con ninguna secuencia conocida pero que, sin embargo, son mucho más largos que lo esperado por azar. Sin embargo, una vez eliminadas las secuencias de menos de 500 bp, quedaron unas 15.000. Caben aquí un par de consideraciones: el hecho de que exista un ORF no implica funcionalidad alguna, en el sentido de secuencia que se transcriba, y sea luego traducida.

Búsqueda del repertorio silencioso de genes VSG

En la introducción se mencionó la utilización del análisis de componentes principales basado en las frecuencias de aminoácidos en la búsqueda de VSG. El mismo se muestra en la figura 21, donde los puntos rojos representan las VSG y los azules al resto de las secuencias. Vemos que las VSG se agrupan claramente (con pocas excepciones) aparte del resto de las secuencias. Este resultado es clave y permite afirmar que las frecuencias de aminoácidos pueden utilizarse para, al menos *grosso modo*, discriminar las VSG del resto de los CDS.

Estas secuencias candidatas a VSG se encontraron mediante tblastn (query = proteína, base de datos = ADN traducido en los seis marcos), utilizando como query las secuencias VSG que habían sido previamente identificadas por haberse determinado las secuencias de sus ARN correspondientes. La presencia de ARNm maduro en el citoplasma es contundente evidencia a favor del carácter codificante de una secuencia de ADN pero, en estos casos, al tratarse de las VSG que se estaban expresando, es la enorme abundancia de esos ARNm, lo que indicaría que se trataban de VSG pues, como ya se mencionó, estas proteínas son —por mucho— las más altamente expresadas. El criterio de búsqueda usado fue mucho más laxo de lo habitual en búsquedas por homología pero, a su vez, demasiado estricto para buscar proteínas donde la variabilidad de la secuencia es tan grande como en las VSG. El análisis de componentes principales puede ser usado como un indicador adicional a favor de que, efectivamente, dichas secuencias son VSG, o al menos, proteínas relacionadas con estas. Es decir, sus frecuencias aminoacídicas corresponden con el patrón-VSG, son homogéneas y difieren de otros patrones que puedan encontrarse en el genoma. Sin embargo, muchas VSG pueden no haber sido detectadas, por quedar debajo del umbral en la búsqueda realizada mediante Blast, que era 30 % de identidad en un fragmento de largo mínimo de 40 % del query. Con este método se

identificaron 228 VSG, y utilizando un criterio de búsqueda aún más laxo, se encontraron 820 más, muchas de las cuales poseían las propiedades estadísticas de las VSG. A partir de aquí veremos cómo utilizamos aproximaciones similares en el caso de *T. vivax.*

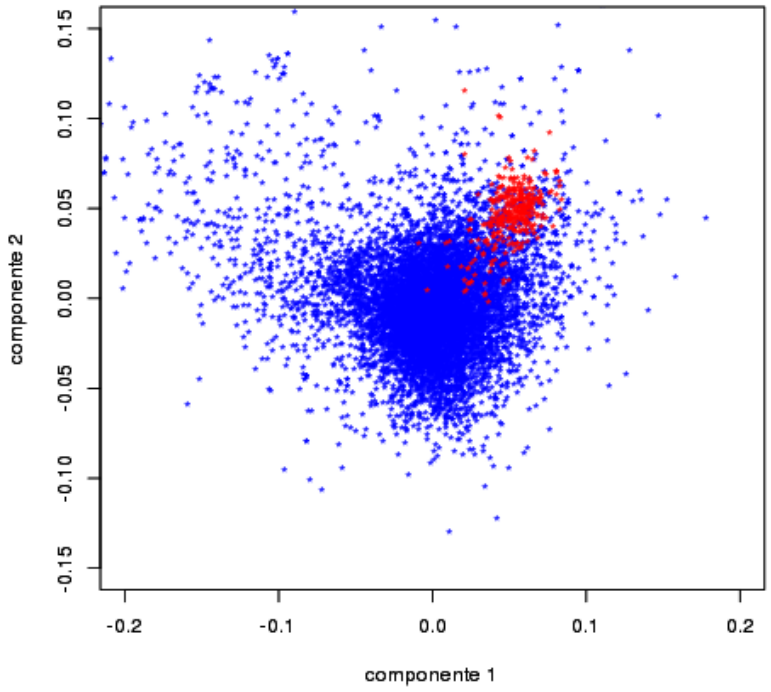


Figura 21. **Representación gráfica del ACP realizado a partir de las frecuencias aminoacídicas del proteoma de *T. vivax*. En rojo, las VSG.**

ORF huérfanos

La aproximación antes descrita para *T. brucei* es viable en esa especie gracias al importante número de secuencias codificantes de genes VSG que ya habían sido determinados en esta especie a partir de ARNm mediante biología molecular tradicional. Dicha información no existe para *T. vivax*, razón por la cual decidimos utilizar una estrategia diferente que, en primer lugar, consiste en reducir el espacio de búsqueda restrigiéndonos a analizar sólo

aquellos segmentos de ADN que son potencialmente codificantes de proteínas; es decir, ORF de una longitud significativamente superior a lo que esperamos por azar.

Téngase en cuenta que, en una secuencia al azar, en la que las cuatro bases aparezcan igualmente representadas, se producen como media tres codones STOP cada 64 codones (en un mismo marco); esto es, un STOP cada 64 bases. Lo que interesa, sin embargo, es ver con qué probabilidad se generan ORF de determinado largo. Para estudiar esto hicimos una simulación, que consistió en generar una secuencia de 10^6 bases con igual cantidad de A, T, G y C. Luego buscamos los ORF (de Met a STOP) en los seis marcos de lectura, con lo que obtuvimos 23660 ORF de un largo medio de 63.76 nucleótidos, de los cuales sólo siete (promediando numerosas simulaciones) superaban los 500 nt de largo.

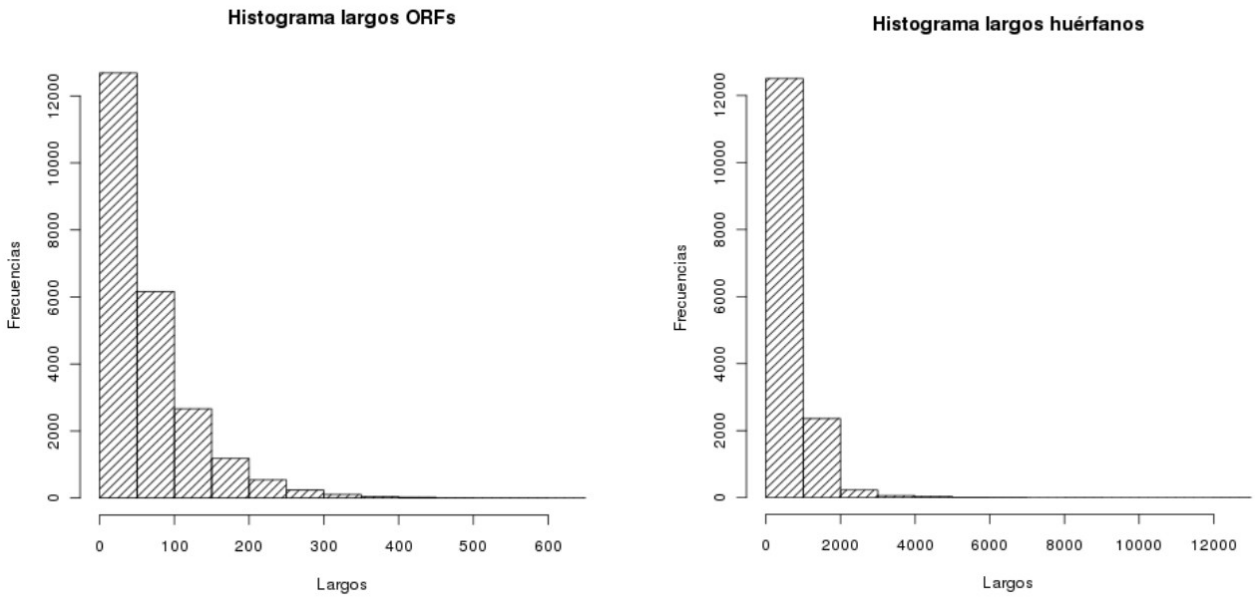


Figura 22. **Histogramas de los largos de los ORF simulados (izq.) y ORF «huérfanos» reales (der.)**

Por este motivo nos pareció razonable optar por esa cifra como umbral, descartando los ORF más cortos. Así, en el largo total de los contigs analizados, cabría esperar unos 260-270 ORF que superaran el umbral establecido, un 2,1 % de los aproximadamente 15.000 que

efectivamente encontramos en los contigs, que, teniendo más de 500 nt de largo, no presentaron homología con ninguna secuencia conocida. En otras palabras, los ORF que encontramos son casi cincuenta veces más que los que cabría esperar por azar; pero no sólo eso, sino que, como se ve en la figura 22, son mucho más largos (recuérdese que en este caso sólo se tuvieron en cuenta los huérfanos, o sea, los de largo >= 500 nt, por lo que la primera columna del histograma de la derecha —figura 21— contiene ORF cuyo largo oscila entre 500 y 1000 nt).

La presencia de tal cantidad de ORF huérfanos es llamativa. Podría tratarse de genes especie-específicos, lo cual explicaría que no se hallaran ortólogos en otras especies. Sin embargo, no tiene sentido pensar en tal cantidad de ellos. Recordemos que en las especies modelo *T. cruzi*, *T. brucei* y *L. major*, el que tenía más genes especie-específicos, *T. cruz*i, no llegaba a 4000 y, en el caso de *T. brucei*, la cifra era inferior a 1400; y esto comparando especies relativamente lejanas; aquí se realizó la búsqueda de ortólogos incluyendo en la base de datos, obviamente, especies sumamente cercanas, lo que llevaría a disminuir el número de genes exclusivos. Los genes exclusivos, en las tres especies mencionadas, corresponden a proteínas vinculadas con la interacción parásito-hospedador (por ejemplo, las VSG de *T. brucei* o las mucinas de *T. cruzi*).

Análisis multivariados

Los análisis multivariados se utilizan para estudiar el comportamiento de modelos que incluyen muchas variables. Una de sus ventajas es que determinan qué variables inciden menos en el comportamiento global, para eventualmente eliminarlas, simplificando así el análisis, y centrando el problema en las variables que efectivamente están vinculadas en forma sustancial con el problema estudiado. Aquí utilizamos dos tipos de análisis: el análisis de componentes

principales (ACP), con fines fundamentalmente exploratorios, y el análisis de clusters, con fines clasificatorios.

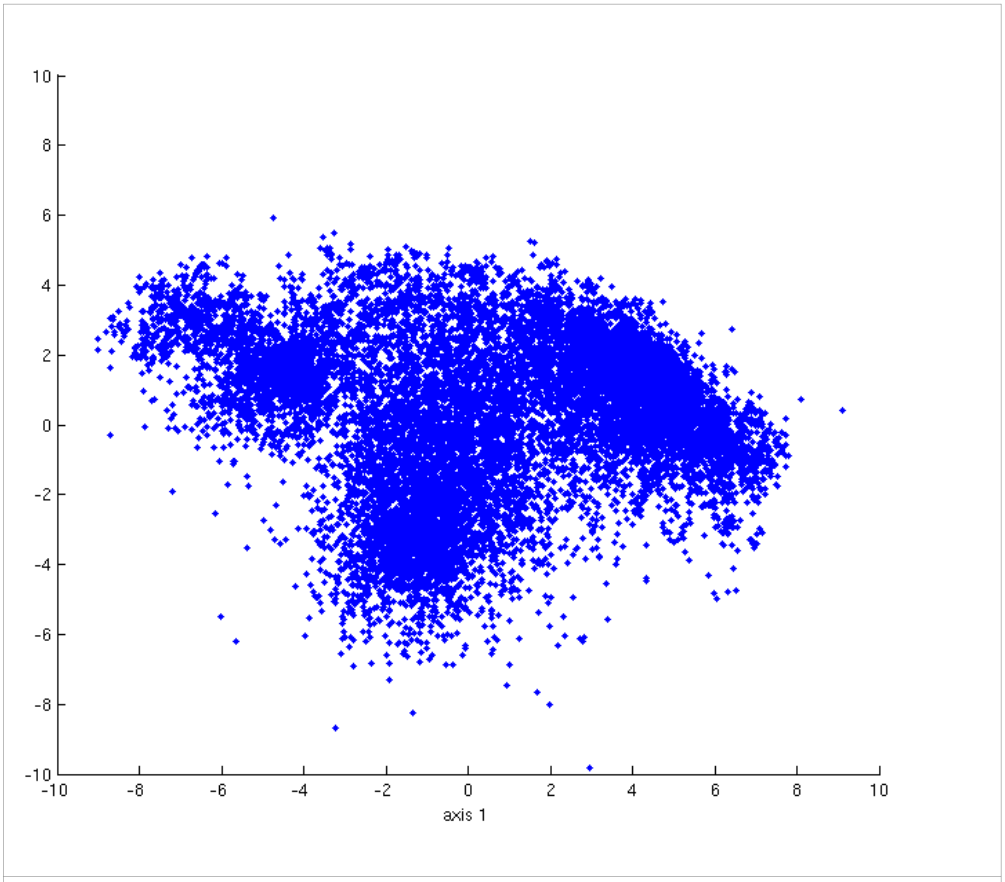


Figura 23. **Representación del ACP basado en las frecuencias de codones de los ORF huérfanos. Cada punto corresponde a un ORF.**

En primer lugar, para estimar el grado de homogeneidad de los ORF huérfanos se realizó un ACP usando las frecuencias de codones. En la figura 23 vemos la representación de dicho análisis donde se grafican los dos primeros componentes. Dos hechos se pueden observar ya en este estado de cosas: en primer lugar, que parece haber grupos bastante marcados y, en segundo, que la estructura aparenta tener cierta simetría (lo cual se confirmará más adelante). A continuación se hizo un análisis de clusters con los mismos datos de frecuencias, con el fin de identificar la pertenencia de cada secuencia a los agrupamientos que se insinúan en la imagen

del ACP. Para realizar el análisis de clusters hay que definir previamente cuántos grupos se desea obtener; esta decisión se tomó mediante la inspección ocular del gráfico original en 3D desde todos los ángulos posibles. Llegamos así a definir cinco grupos de ORF huérfanos a los que llamaremos cl1-5, habiendo obtenido 4488 en el primer cluster (cl1), 2046 en el segundo (cl2), 2646 en cl3, 2081 en cl4 y 3954 en cl5. Estos datos corresponden al total de contigs; al limitar el análisis, como ya se mencionó, a los contigs que tienen más de 1500 nt, esas cifras se redujeron.

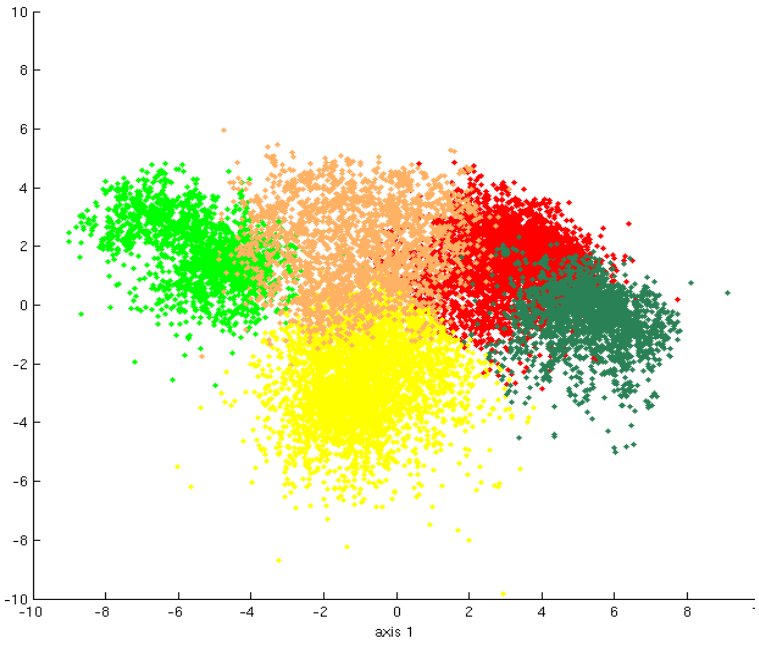


Figura 24. **PCA donde los colores representan la pertenencia de cada punto (cada ORF) a los distintos grupos definidos en el análisis de clusters.**

Al volver a graficar los resultados del análisis de componentes principales basado en las mismas frecuencias de codones, pero esta vez conociendo a qué grupo corresponde cada ORF, de modo de poder visualizarlo separadamente, obtenemos la imagen de la figura 24.

Basándonos en los antecedentes descritos en la introducción, podríamos suponer que alguno de esos grupos correspondería a los genes VSG. Considerando que las candidatas a VSG de *T. vivax* habían sido, en principio, resultado de una búsqueda por homología con criterios mucho más laxos que los habituales (por ejemplo, para definir a estos ORF como «huérfanos» se utilizó un umbral más exigente), era lógico que, al realizar las búsquedas mediante Blastn (se usó esta variante de Blast debido a que las secuencias están fuera de marco, existiendo la posibilidad de *frameshifts* internos, lo cual impedía utilizar secuencias traducidas), ninguno de estos ORF hubiera «matcheado» con secuencias de VSG de otras especies. Sin embargo, al tener VSG del mismo organismo, se eliminaba el problema de la gran variabilidad de secuencia de las VSG, ya que en este caso se trataría de encontrar las mismas secuencias, y no otras emparentadas. En efecto, el resultado del Blastn fue que las VSG de este trabajo correspondían con muchas de las secuencias del grupo cl2. Al agregarlas al análisis de componente principal, como era de esperar, resultaron comprendidas dentro de la nube de puntos correspondiente a ese grupo. Del total de 1048 VSG utilizadas como queries, solo 29 (cuatro de las encontradas con la búsqueda más estricta, y 25 con el criterio más laxo) no encontraron ORF correspondiente. La única explicación que encontramos para esa ausencia es que el ensamblaje utilizado en este trabajo es una versión posterior al que se usó para encontrar las VSG que sirvieron como query en esta búsqueda. Dichas secuencias dejaron de pertenecer al grupo cl2 para pasar a ser clasificadas directamente como VSG o, mejor dicho, «secuencias tipo VSG», que comparten propiedades estadísticas más específicas con ellas. Las cantidades definitivas de ORF en cada cluster son: 3908 cl1, 766 cl2, 2279 cl3, 1839 cl4 y 3672 cl5.

Sin embargo, los resultados de Blastn arrojaron otro dato sorprendente: las secuencias de VSG también matcheaban con muchas secuencias del grupo cl4, las cuales presentan

propiedades estadísticas casi opuestas —tanto a las secuencias codificantes de VSG como a los ORF de grupo cl2. Un primer aspecto esclarecedor de esta situación, aparentemente paradójica, es que dichos hits de Blastn se dan *antisentido*, es decir que corresponden a ORF localizados sobre el mismo segmento de ADN, pero en la hebra complementaria (también se observaron casos de hits entre ORF de distintos grupos, correspondientes al mismo segmento y a la misma hebra, pero ubicados en distinto marco). De hecho, las VSG también encontraban secuencias similares en los otros grupos, aunque en menor número y no de forma caprichosa: en la tabla de la figura 25 vemos cómo se distribuyen los «best hits» para las VSG entre los distintos grupos de ORF.

	VSG hits	misma hebra	hebra compl
cl1	106	0	106
cl2	566	566	0
cl3	15	7	8
cl4	323	0	323
cl5	9	8	1

Figura 25. ***Best hits*** **de VSG y su distribución entre los distintos grupos de ORF huérfanos.**

El análisis se completa al incorporar al ACP, además de las VSG, diversos tipos de secuencia, como ser «genes anotados» (ORF que encontraron secuencias homólogas en otras especies), RHS y las VSG o candidatas a VSG (figura 26).

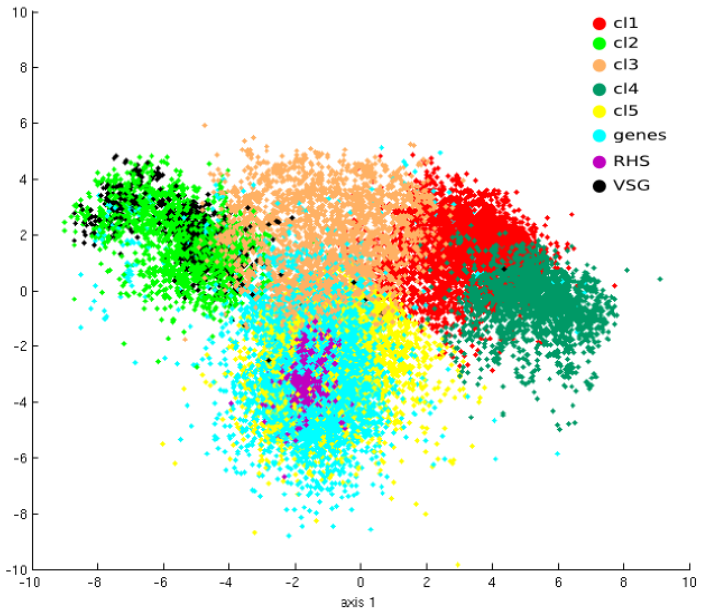


Figura 26. **El mismo ACP de tres y dos figuras antes, al que se agregaron datos de VSG, genes anotados y RHS.**

Nos detendremos aquí ante un hecho que llama poderosamente la atención: ¿por qué las hebras complementarias de las VSG presentan tan pocos codones STOP, como para generar ORF tan largos? Lo habitual en la hebra complementaria de una secuencia codificante (o más en general, en cualquiera de los otros cinco marcos de lectura) es que exista una proporción de codones STOP no muy diferente de la que resulta esperable en virtud de la composición de bases se la secuencia. En otras palabras, no tiene sentido que haya gran cantidad de marcos abierto de lectura de más de 500 nt, que sean complementarias ni de VSG ni de cualquier otra secuencia codificante.

Otra pregunta que surge de este análisis es la siguiente: ¿esta compartimentación de los genes en virtud de sus propiedades estadísticas se corresponde con una compartimentación

física? En otras palabras, ¿ocupa cada uno de estos grupos una región separada del genoma? ¿Hasta qué punto?

A continuación trataremos de responder a estas preguntas.

GC%	media stops/marco	dist media entre stops	con largo >500
40	21053	48	0
60	10626	94	245
80	2992	334	4003

Figura 27. **Conteos de STOP y largos de ORF en secuencias simuladas con distinto contenido GC.**

Orfogenicidad

Llamaremos «orfogenicidad» a la capacidad de ciertas secuencias de generar grandes ORF en cualquiera de los seis marcos de lectura. Del mismo modo, pasaremos a denominar OOO (por las siglas en inglés de *Orphan Orfogenic ORF*) a cada uno de los ORF de más de 500 nt de largo, que no presentan homología con secuencias conocidas y en cuya composición y disposición en el genoma se constata la propiedad de orfogenicidad.

Una primera explicación de esta propiedad radica en el nivel de GC. Siendo los codones STOP (TAG, TGA y TAA) ricos en AT, resulta evidente que una menor proporción de estas bases disminuye la probabilidad de que se forme por azar un codón STOP (de hecho, en *T. vivax*, los genes codificantes de VSG tienen un GC% más alto que la media del genoma). Para verificar o descartar esta explicación, realizamos simulaciones de secuencias aleatorias de distinto largo y composición, sobre las que realizamos determinados conteos y medidas. En la figura 27 vemos el resultado de algunas de estas simulaciones, en las que se generaron

secuencias de largo = 10^6 nt con contenidos GC de 40 %, 60 % (equivalente al de las VSG de *T. vivax*) y 80 %.

Vemos que la cantidad de secuencias que superan los 500 nt en la simulación realizada con un GC = 60 % es muy baja. En la figura 28 se ven los histogramas de frecuencias de largos derivados de estas simulaciones.

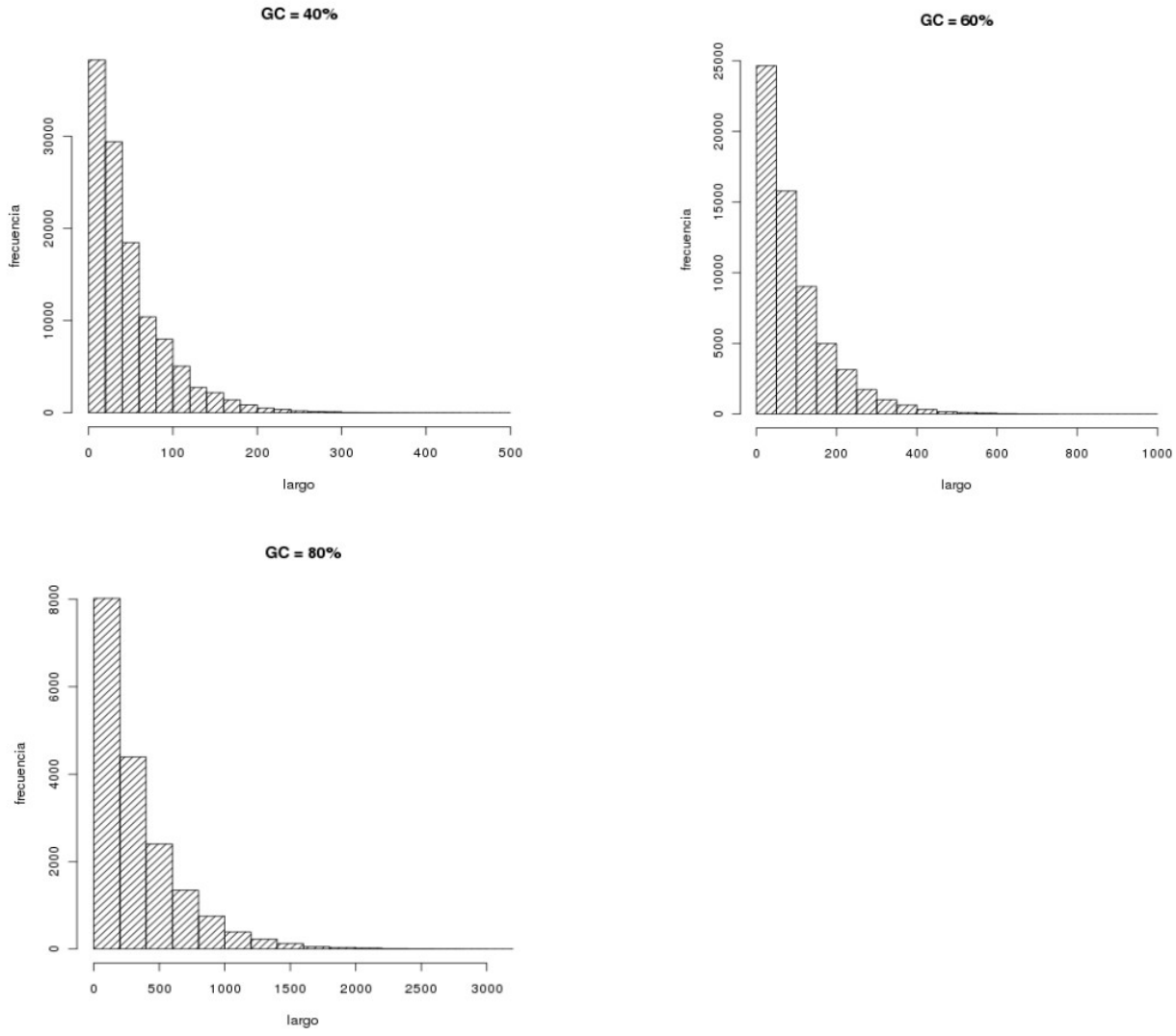


Figura 28. **Histogramas de largos de ORF simulados con distintos niveles de GC.**

En los siguientes análisis se realizaron simulaciones basadas en cadenas de Markov de distintos órdenes, según se describe en la sección «Materiales y métodos».

Ejemplificamos aquí este tipo de análisis con el contig CAEX01007145.1, de 5945 nucleótidos de largo, y que tiene varias secuencias tipo VSG y una alta orfogenicidad, que se hace evidente al mirar la representación en artemis de la figura 29. Se generaron tres secuencias a partir de sus características composicionales, siguiendo las reglas de las cadenas de Markov de orden 0, 1 y 2. En un principio, las secuencias generadas eran de 10^5 nt de largo, con el fin de obtener cierta precisión estadística en los conteos. El resultado fue de 63, 110 y 175 ORF (empezando por Met y terminando con STOP) de más de 500 nt de largo, usando cadenas de Markov de orden 0, 1 y 2, respectivamente. Si no se impone la restricción de que los ORF empiecen con Met, las cifras se multiplican aproximadamente por tres (por ejemplo, un ORF de 550 bases que tenga la primera Met más allá de la base 50, pasaría a tener menos de 500 bases si se lo hace empezar desde dicha Met, y no sería tenido en cuenta), pero se mantiene la relación.

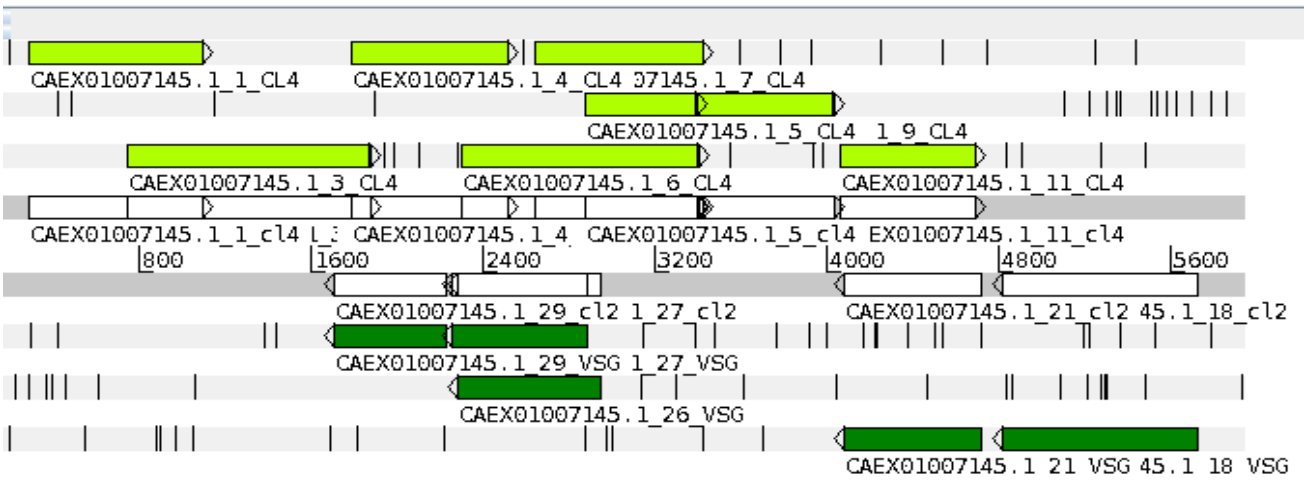


Figura 29. **Vista en Artemis de un contig orfogénico con secuencias de tipo cl2 «tipo VSG» (verde oscuro) y cl4 (verde claro). Las flechas que rematan cada rectángulo coloreado indican la dirección 5'->3'. Se presentan ambas hebras del ADN, y de cada hebra los tres marcos de lectura por separado. Las líneas verticales pequeñas en cada uno de los marcos de lectura indican la presencia de codones stop.**

Con el fin de visualizar estos resultados en una escala comparable a la del contig original, hicimos las mismas simulaciones para secuencias de 5045 nt (el largo del contig) y elegimos un caso que se aproximara a lo obtenido con la secuencia de 10^5 nucleótidos, que estadísticamente es más verosímil. En la figura 29, el contig real, vemos en verde claro los ORF de tipo cl4, y en verde oscuro los cl2 (que además, en este caso, pasaron el umbral de similitud de secuencia como para ser catalogados de «candidatos a VSG»).

En las figuras 30, 31 y 32 se ve el resultado de simulaciones con cadenas de Markov de orden 0, 1 y 2, las cuales, como vimos en «Materiales y métodos», se construyen, respectivamente, a partir de las frecuencias de bases, de dinucleótidos y de trinucleótidos observadas. Una vez obtenidos los ORF de más de 500 nt (celestes) se desplazó su extremo 5' hasta la primera metionina, dando como resultado los ORF representados en negro. De estos últimos, los que no llegaban a 500 nt de largo fueron eliminados.

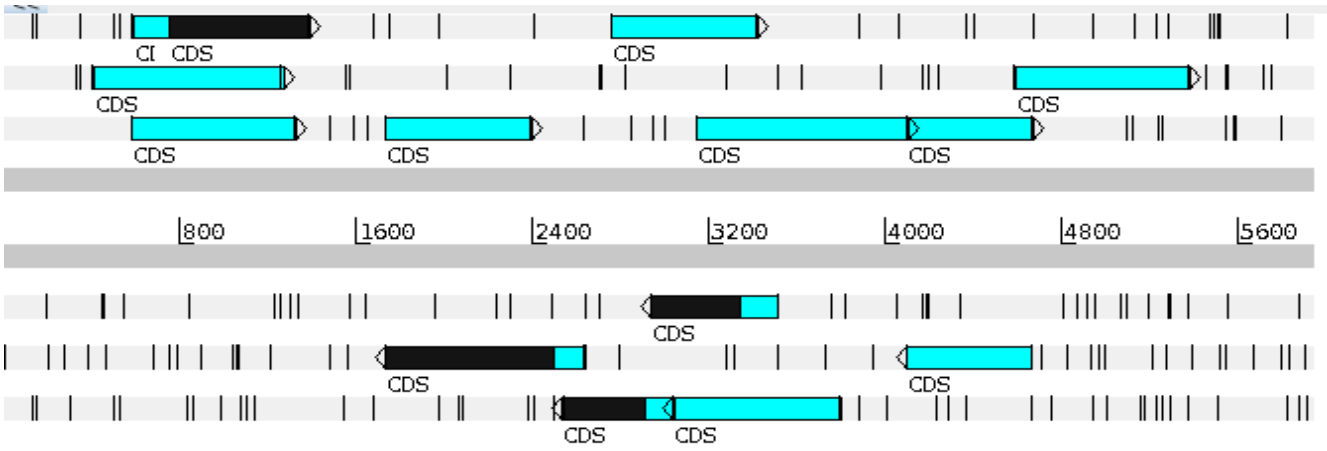


Figura 30. **ORF de más de 500 bases resultantes de una simulación basada en una cadena de Markov de orden 0. Los ORF negros son resultado de correr el inicio del ORF celeste correspondiente hasta la Met más próxima, En caso de no superar los 500 nt, fueron eliminados de la imagen.**

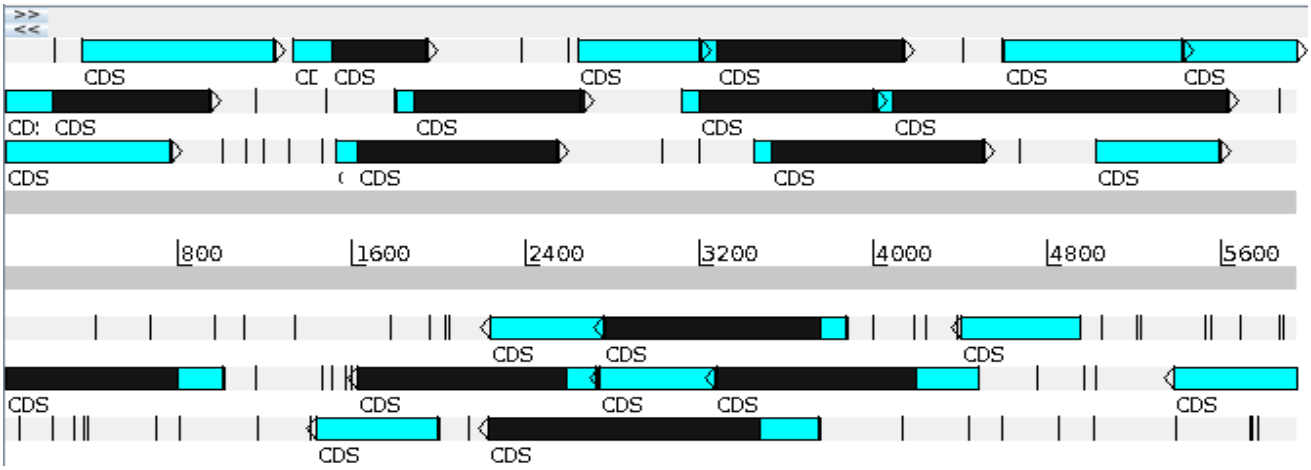


Figura 31: **Lo mismo, con cadenas de Markov de orden 2.**

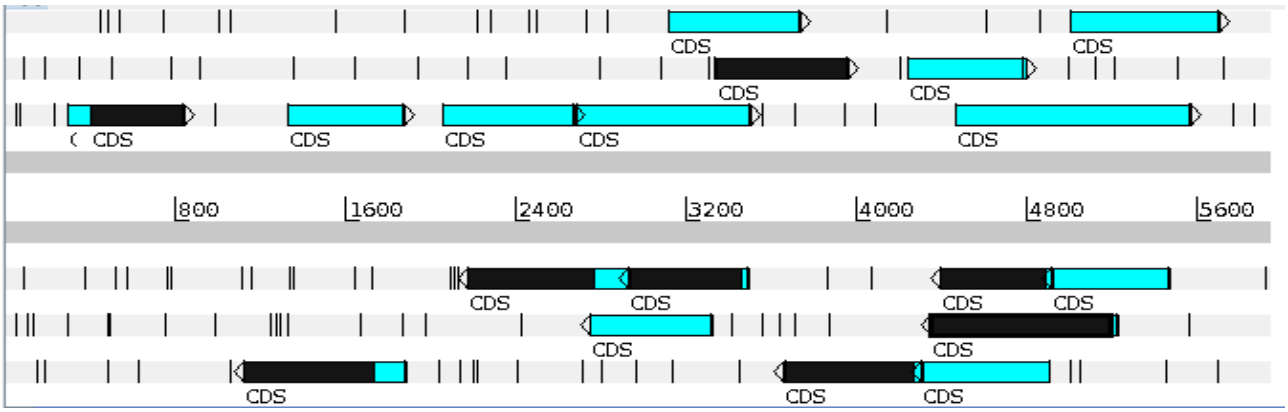


Figura 32. **Lo mismo que en el caso anterior, pero a partir de cadenas de Markov de orden 1.**

Vemos que las simulaciones en base a cadenas de orden 1 y —especialmente— de orden 2, son «más orfogénicas». Vale decir, estudiando las frecuencias de dinucleótidos y trinucleótidos de las secuencias orfogénicas, deberíamos poder comprender algo sobre la orfogenicidad. Esto no es sencillo, pero podemos hacer razonamientos previos. Por ejemplo: así como dijimos que una secuencia rica en GC tendría menos codones STOP en virtud de que estos son ricos en AT, podemos ver que, a nivel de dinucleótidos, el par TA pertenece a dos codones STOP: TAA y TAG, faltando en el tercero, TGA. Pero tanto este último como TAA

corresponden a la forma general TNA. Los codones de esta forma son, o bien STOP (TAA y TGA) o bien sus hebras complementarias (TTA y TCA, respectivamente).

	A	C	T	G
A	0.088	0.090	0.084	**0.025**
C	0.112	0.039	0.097	0.038
T	0.074	0.126	0.085	0.029
G	**0.012**	0.032	0.048	0.021

Figura 33. **Frecuencias de dinucleótidos de los genes de VSG.**

Al analizar la matriz de frecuencias absolutas de trinucleótidos, esto es, sumadas las 64 frecuencias de las VSG (figura 34), se ve que las de nuestros cuatro trinucleótidos TNA (en azul), son relativamente bajas. Esto es en cierto modo esperable por dos motivos:

1. la composición de bases. Como ya dijimos, las VSG tienen un contenido GC cercano al 60 %. Pero además, en la hebra codificante, existe una gran asimetría entre A y T: dicha composición es (hablamos de valores medios) de 30,90 % (A), 30,71 % (G), 28,50 % (C) y 9,89 % (T), una relación de aproximadamente 3:3:3:1, lo cual dificulta aún más la generación de tripletes del tipo TNA.
2. la restricción que implica el hecho de trabajar con ORF, ya que, en uno de los seis marcos, está «prohibida», por definición, la presencia de la mitad de esos tripletes (los que son efectivamente codones STOP; no así los que son su reversa complementaria).

	A	C	G	T
AA	0.0278	0.0259	0.0281	0.0065
AC	0.0371	0.0122	0.0284	0.0120
AG	0.0184	0.0327	0.0259	0.0065
AT	0.0039	0.0074	0.0099	0.0038
CA	0.0322	0.0343	0.0343	0.0110
CC	0.0144	0.0044	0.0140	0.0062
CG	0.0230	0.0413	0.0236	0.0095
CT	0.0038	0.0117	0.0157	0.0072
GA	0.0249	0.0248	0.0186	0.0054
GC	0.0501	0.0200	0.0399	0.0163
GG	0.0237	0.0336	0.0196	0.0079
GT	0.0027	0.0077	0.0143	0.0041
TA	**0.0035**	0.0047	**0.0026**	0.0013
TC	**0.0102**	0.0025	0.0152	0.0039
TG	0.0086	0.0189	0.0159	0.0049
TT	**0.0017**	0.0050	0.0084	0.0060

Figura 34. **Frecuencias de trinucleótidos de los genes de VSG.**

Para eliminar estos dos problemas comparamos las frecuencias de estos cuatro tripletes en las VSG con las que se obtienen de los ORF resultantes de una secuencia simulada (Markov de orden 0) a partir del contenido de bases de las VSG. Como media, se obtiene que la frecuencia de estos tripletes en las VSG reales es de ~ 70 % de la de los ORF simulados con la misma composición de bases. En otras palabras, en las VSG están subrepresentados los codones de la forma TNA (que son STOP en una u otra hebra), comparando con lo esperable en virtud de la composición de bases y del hecho de que se trata de marcos abiertos de lectura. En cuanto a los codones STOP en general y sus secuencias reversas complementarias (o sea, los tripletes que generan codones stop en la hebra no codificante), todos están subrepresentados —con la excepción de TGA, que aparece con un valor similar en las VSG y en las simulaciones—, con una media de 44 % de lo esperado.

En cuanto al par TA (presente en TAA y TAG, dos de los tres codones de terminación), ocurre algo similar. Siendo TA un *palíndrome*, que es igual a su reversa complementaria, al haber menos TA en una hebra también lo habrá en la otra, con lo cual el efecto «promotor de stop» de este par se cumple por igual en ambas hebras. Esto permite crear un indicador que puede estimar rápidamente y de un modo sencillo la orfogenicidad: el cociente AT/TA. Cuanto más alto ese valor, más orfogénica tenderá a ser una secuencia. La probabilidad de generar un par TA es la misma que un par AT (dado que tienen las mismas bases, es esperable que aparezcan en la misma proporción), con la salvedad, ya considerada, de que si estamos hablando de ORF, hay algunos codones que incluyen TA cuya presencia está vedada en uno de los marcos, por ser codones de terminación. Para verificar la intensidad de este sesgo también realizamos simulaciones, de las que resultó que el AT/TA de los ORF simulados es, efectivamente, superior a 1 (1.15), pero muy inferior al de las VSG (2.06).

Por último, era necesario verificar que la composición aminoacídica de las VSG no esté introduciendo, a su vez, un nuevo tipo de sesgo. Si los aminoácidos codificados por codones con presencia de TA fueran muy poco numerosos, podría considerarse que el alto valor de AT/TA se debe a esta causa. Esto no atentaría contra el concepto de orfogenicidad, simplemente daría una posible explicación alternativa de la misma. De todos modos, observando las frecuencias de aminoácidos en nuestro conjunto de VSG (del que se eliminaron previamente los casos que presentaban corrimiento del marco de lectura) vemos justamente que la suma de todos los aminoácidos codificados por CTA (Leu), GTA (Val), TAC (Tyr) y TTA (Leu), que da 44.796, es ligeramente mayor que en el caso de AAT(Asn), ATC (Ile), ATG (Met), ATT (Ile), CAT (His) y GAT (Asp), que suman 41.223. En la cuenta se incluyeron todas las apariciones de estos aminoácidos (estuvieran codificados por estos codones o no), ya que lo

que queríamos analizar era si la composición aminoacídica en sí era tal que podía explicar que hubiera más codones con AT que con TA. Fueron excluidos del análisis los codones ATA y TAT, por contener, simultáneamente, AT y TA.

A	C	D	E	F	G	H	I	K	L
0.163	0.025	0.044	0.057	0.014	0.082	0.028	0.025	0.058	0.068
M	N	P	Q	R	S	T	V	W	Y
0.014	0.038	0.029	0.056	0.077	0.064	0.096	0.037	0.014	0.007

Figura 35. **Frecuencias de aminoácidos de VSG.**

Compartimientos genómicos

A los efectos de acercarnos a la comprensión de la estructura del genoma de *T. vivax* en función de los distintos tipos de secuencia hasta ahora encontrados, hicimos un nuevo análisis de componente principal. Para ello utilizamos las frecuencias de trinucleótidos de los contigs espejados (ver «Materiales y métodos»). En el gráfico de la figura 36 se ve claramente que existen dos grupos de contigs, separables por sus propiedades estadísticas. Como poseemos datos de anotación de esos contigs, podemos ver si, además, se pueden definir características más tangibles (presencia o no de determinados tipos de genes en ellos) que el mero hecho de compartir similares patrones de frecuencias de trinucleótidos.

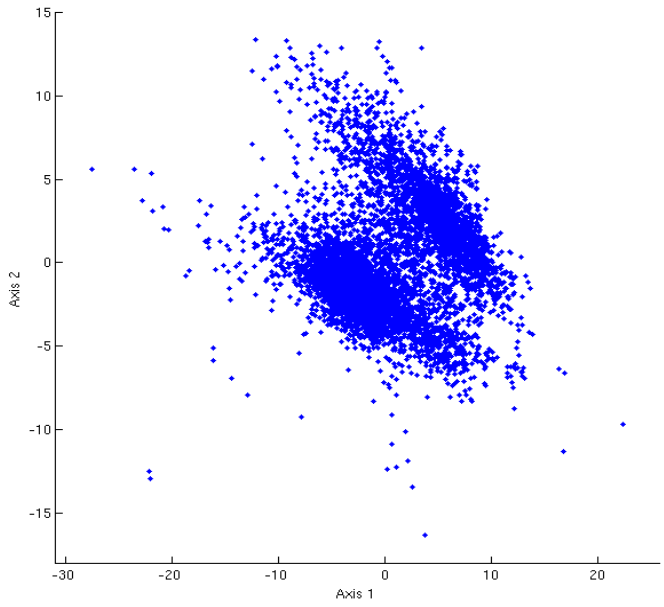


Figura 36. **ACP realizado sobre las frecuencias de trinucleótidos de los contigs «espejados».**

Con el objeto de disminuir el ruido, eliminamos arbitrariamente una franja de contigs cuya pertenencia a uno u otro grupo no era evidente. Los eliminados fueron 309, quedando dos grupos bien separados, de 4.200 (abajo, izquierda) y 2.312 (arriba, derecha) contigs, a los que denominaremos tipo A y B, respectivamente.

A continuación intentaremos encontrar características comunes a los contigs de cada grupo, basándonos entre otras cosas en la anotación, tanto de genes como de ORF huérfanos. En el siguiente cuadro (figura 37) se resumen algunos datos de anotación de los contigs. Como los largos sumados de los contigs de los grupos A y B son muy distintos (la relación es de 3 a 1: los 4200 contigs del grupo A suman 2.76 X 10^7 bases, y los 2312 del grupo B llegan a 9.2 X 10^6 bases), hemos normalizado los números para una secuencia de diez millones de bases.

grupo	len	GC	cl1	cl2	cl3	cl4	cl5	VSG	RHS	ANOT	AT/TA
A	1E+07	0.50	150	23	336	12	1282	1	208	3479	1.41
B	1E+07	0.56	3644	725	1255	1918	52	1095	6	1956	1.74

Figura 37. **Nivel de GC, valores normalizados por largo de conteos de distintos tipos de ORF, y valores de AT/TA en los contigs de los grupos A y B.**

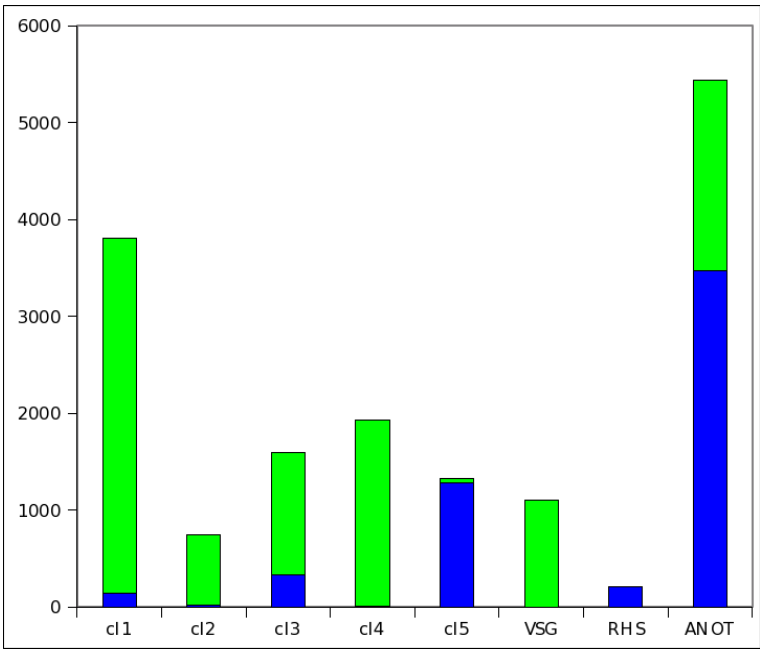


Figura 38. **Representación gráfica de los conteos del cuadro de la figura 36. Grupo A: azul, grupo B, verde.**

Con respecto a la composición de bases, en el grupo A es prácticamente de 25 % cada uno, mientras que en el grupo B la relación es de A ≃ T ≃ 22%, y G ≃ C ≃ 28%. Pero hay diferencias que llaman la atención: tanto las secuencias que son candidatas a codificar proteínas VSG como los ORF pertenecientes a los grupos cl1 a cl4, presentan notablemente más apariciones en los contigs de tipo B, mientras que los Retroposon Hot Spot (RHS) y los genes *housekeeping* en general (ANOT) tienden a aparecer en los contigs del tipo A. En algunos casos (cl2, cl4, VSG y los RHS) se puede hablar de casi una exclusividad en cuanto a la pertenencia a uno u otro grupo. El indicador de orfogenicidad AT/TA es más alto en los del

grupo B, lo que es consistente con la mayor presencia de ORF tipo OOO y secuencias tipo VSG en ellos (figura 38).

Cabe destacar que un conteo realizado sobre los ORF que llamamos «anotados» (esto es, ORF que presentaron homología con secuencias conocidas) que están presentes en contigs de tipo B (color verde en la parte superior de la columna «ANOT»), dio como resultado que, en alrededor de un 90 %, se trata de proteínas de superficie, *hypothetical proteins* y elementos transponibles, en ese orden. De modo que podemos caracterizar al espacio B como conteniendo genes de VSG y otras proteínas de superficie, transposones de diversos tipos y OOO de los tipos 1 a 4, mientras que el espacio A contiene el resto de los genes y OOO de tipo cl5. Estos últimos tienen un comportamiento diferente al del resto de los OOO. En el cuadro de la figura 38 vemos algunos indicadores que parecen mostrar que, efectivamente, corresponden a otra cosa; tal vez muchos de ellos sean en realidad secuencias codificantes para las que no se encontró homología.

	AT/TA	media (nt)	GC	GC1	GC2	GC3
cl1	2	764	0.61	0.60	0.60	0.62
cl2	2.2	775	0.58	0.59	0.55	0.61
cl3	2.44	718	0.65	0.65	0.62	0.67
cl4	2.07	778	0.55	0.55	0.55	0.56
cl5	1.48	967	0.52	0.56	0.46	0.54

Figura 39. **AT/TA, largo medio de ORF, GC y GC por posición en el codón para los OOO de los distintos grupos.**

Como vemos, los cl (1 a 4) presentan un alto AT/TA, comparable al de los genes codificantes para proteínas VSG, que era de 2.06. En cambio el valor de los cl5 es similar al de los contigs de tipo A (1.41).

El largo medio de los ORF tipo cl5, por otra parte, es superior al de los demás OOO, sin llegar al de los genes comunes, que es de alrededor de 1200 nt.

En cuanto a su nivel de GC, vemos que es el más bajo de los cinco grupos, algo también compatible con el GC general del grupo A, o el de los genes no VSG. Pero además, si observamos por separado los GC de las posiciones 1, 2 y 3 (de cada codón, de acuerdo con el marco de lectura del ORF) vemos que en el caso de cl5 se da la mayor disparidad entre estos tres valores. Esto es típico de una secuencia codificante (o que lo fue, en el caso de fragmentos génicos). Debido a la estructura del código genético, cuando existe presión selectiva tendiente a variar el contenido GC de un organismo, la posibilidad de realizar sustituciones sinónimas (que no alteran el aminoácido codificado) es distinta en cada una de las tres posiciones, lo cual produce esas variaciones en el contenido GC.

Todo lo anterior, unido al hecho de que los ORF de tipo cl5 aparecen en los contigs tipo A, mientras que el resto de los OOO predominan en los contigs tipo B, hace pensar que cl5 es un grupo diferente, probablemente formado por ORF correspondientes a genes y pseudogenes especie-específicos (lo que explicaría que no se les encuentre similitud con secuencias de otras especies), con propiedades estadísticas similares a los genes *housekeeping*, no estando vinculado al fenómeno de la orfogenicidad.

Expresión

Una de las principales utilidades del análisis de transcriptomas mediante tecnologías de RNA-seq es determinar con gran certeza el nivel de transcripción de diferentes regiones. Dado que el número de reads que matchean con determinada secuencia es proporcional a cuánto se transcribe esa secuencia, así como a su largo, normalizando por largo tenemos una forma

relativamente precisa de medir dicho nivel de trancripción. Ya vimos que los contigs del grupo A matchearon con una mayor cantidad de reads, tanto de Roche como de Illumina, siendo la relación (después de normalizar por largo total de los contigs) que va desde 2 a 1 (Illumina) a 5 a 1 (Roche) con respecto a la cantidad de reads que matchearon con los contigs B. Si bien estos últimos contienen las secuencias donde están codificadas las VSG, cabe recordar que se supone que sólo una de ellas se expresa en un momento dado. Estando la enorme mayoría de los genes en los contigs del grupo A, cabría esperar una relación bastante más alta. La explicación de esto radica, de nuevo, en la peculiaridad del sistema de transcripción y regulación de estos organismos, en los que, en principio, todos los genes se transcriben por igual, y la diferencia en la concentración de los distintos productos se debe a factores de acción posterior (diferente velocidad de degradación del ARN mensajero, control de la iniciación de la traducción, o incluso regulación postraduccional) (Kramer, 2012). La diferencia entre lo mostrado por los reads de Roche y de Illumina se debe a que los primeros, al ser más largos (una media de 287 nt, casi ocho veces el largo de los reads de Illumina) distinguen mejor entre genes parálogos, los cuales pueden presentar regiones homólogas idénticas de un largo que pueda contener varios reads de Illumina, y sin embargo ser bastante más cortas que un read de Roche, el cual por lo tanto no sería asignado a esas regiones similares sino a la correcta.

Sin embargo, un hecho llama poderosamente la atención: existen muchos contigs que presentaron una transcripción nula. Veamos los datos que muestra la figura 40.

	largo	GC	cl1	cl2	cl3	cl4	cl5	Vsg	rhs	ano	AT/TA	454	illumina
A	**163021**	**0.45**	**13**	**1**	**4**	**1**	**9**	**0**	**3**	**0**	**1.26**	**0**	**0**
B	**6415621**	**0.56**	**2363**	**484**	**702**	**1389**	**11**	**780**	**3**	**1091**	**1.69**	**0**	**0**
A	***1E+07***	***0.45***	***797***	***61***	***245***	***61***	***552***	***0***	***184***	***0***	***1.26***	***0***	***0***
B	***1E+07***	***0.56***	***3683***	***754***	***1094***	***2165***	***17***	***1216***	***5***	***1701***	***1.69***	***0***	***0***
A	27610377	0.5	413	64	927	33	3539	4	573	9605	1.40	284881	37703496
B	9256737	0.56	3373	671	1162	1775	48	1014	6	1811	1.73	19468	5768502
A	*1E+07*	*0.5*	*150*	*23*	*336*	*12*	*1282*	*1*	*208*	*3479*	*1.41*	*103179*	*13655553*
B	*1E+07*	*0.56*	*3644*	*725*	*1255*	*1918*	*52*	*1095*	*6*	*1956*	*1.74*	*21031*	*6231680*

Figura 40. **Comparación de los contigs «sin reads» con el total.**

Las cuatro primeras filas (en negrita) corresponden a valores sin normalizar (arriba) y normalizados por largo (abajo, en cursiva) realizados sobre contigs «sin reads», es decir, contigs con los cuales no matcheó ningún read, para los contigs de tipo A y B. Abajo, a efectos de comparación, se reproducen los mismos conteos pero sobre los contigs A y B totales; en este caso el cuadro «normalizado» es el mismo que vimos antes (página 51), cuando caracterizamos, justamente, los compartimientos genómicos A y B.

Algunos puntos a destacar son:

1. En largo total, los contigs sin reads son el 18 % de todos los contigs totales. De ese 18 %, apenas 2,5 % pertenecen al grupo A, mientras que 97,5 % al B.
2. Los contigs de tipo A sin reads corresponden al 0,6 % del largo total de contigs A. Los contigs B sin reads, en cambio, representan el 69 % del largo total de contigs B.
3. Mientras los genes anotados del grupo A están todos en los contigs con reads, los anotados del grupo B están en su mayoría en los contigs sin reads: 60 % en valores

reales, aunque su concentración en estos contigs es algo menor (1701 contra 1956 cada 10^7 bases).

Más allá de esto último, puede afirmarse que de aquellos genes de tipo «anotado» en el grupo B que veíamos en la figura 38 —que eran en realidad proteínas hipotéticas (anotados como proteínas hipotéticas en otras especies de tripanosomátidos), proteínas de superficie y elementos transponibles—, en su mayoría no se expresan en la etapa sanguínea.

4. Los genes candidatos a codificar proteínas VSG que se ubican en contigs del grupo A son sumamente escasas (apenas 4), y están todas en los contigs con reads.
5. Un 77 % de los genes codificantes de VSG del grupo B (en la práctica, un 77 % de todas las VSG) se ubica en los contigs sin reads. En este caso, incluso su concentración es mayor: 1216 en los contigs B sin reads contra 1095 en los contigs B con reads, cada 10^7 bases.
6. Un último elemento a intentar explicar es la drástica caída de cinco puntos en el GC% de los contigs A sin reads. Existe cierta correlación positiva entre el largo de los contigs y su contenido GC (0,38 para n=70 contigs A sin reads, $p<0,001$), siendo la media[2] de GC de los diez contigs más cortos inferior a 0,41, mientras que para los diez más largos supera los 0,50. Así mismo, los diez contigs con más AT tienen un largo medio de 1691, mientras que en los 10 con más GC ese largo es de 2346. Es de remarcar que este hecho se puede deber al azar o, en todo caso, a alguna peculiaridad del ensamblaje —por ejemplo, que los contigs con AT muy alto suelen tener secuencias repetidas y ser más difíciles de ensamblar—, y no a una causa biológica. Sin embargo, resulta difícil explicar por qué no ocurre lo mismo en el extremo rico en GC. Al observar el

[2] Al decir «media» me refiero, en todos estos casos, a la media ponderada por largos.

histograma de los valores GC de los contigs A, se ve que (salvo algunos contigs con valores de GC% < 30, que por otra parte no están incluidos entre los contigs «sin reads»), la distribución es altamente simétrica y va desde GC ~ 0.3 hasta ~ 0.7.

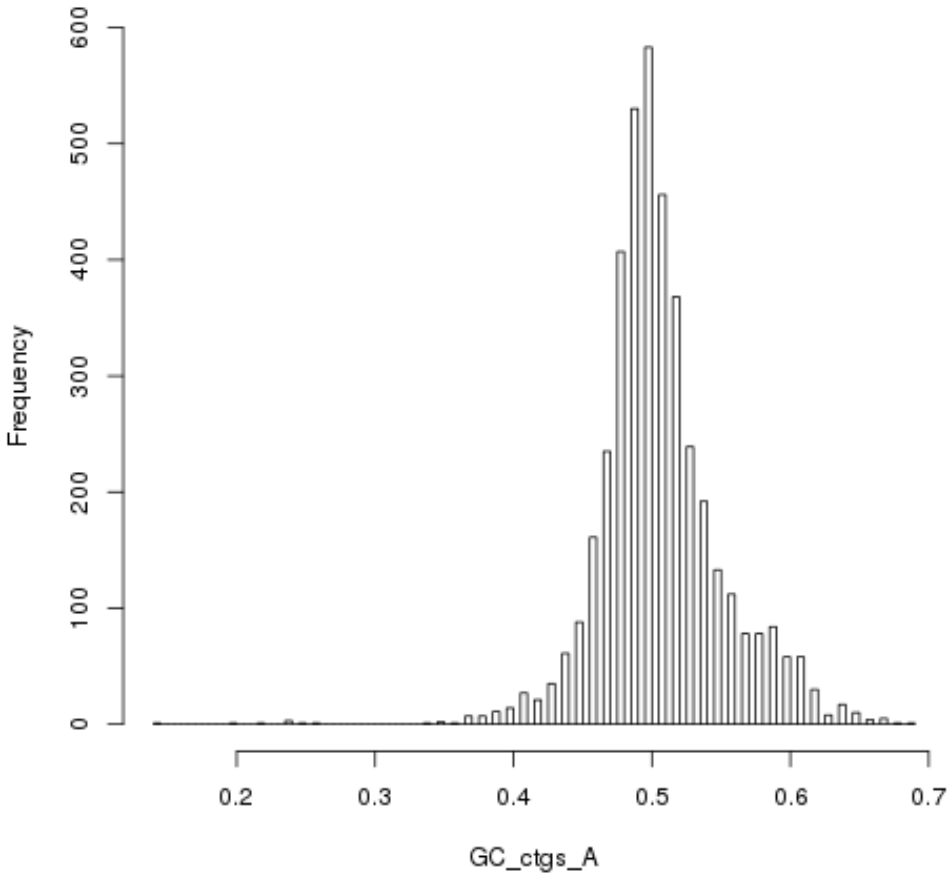


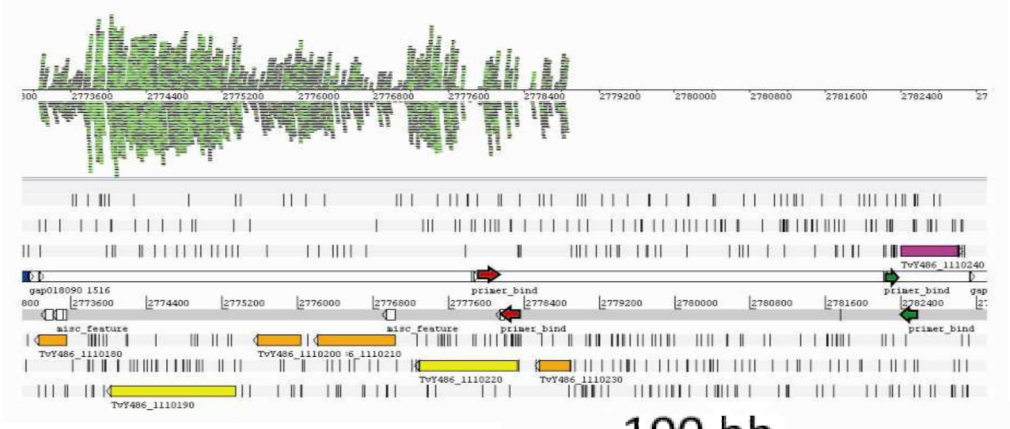
Figura 41. **Histograma de GC (contigs del grupo A).**

El hecho es que, entre los contigs sin reads tipo A, los que tienen un elevado valor de AT % tienen una ligera tendencia a ser más cortos. Si a esto sumamos la mayor cantidad de STOP en secuencias ricas en AT de que ya hablamos) el resultado es que hay muy pocos ORF medianamente largos como para ser codificantes, lo que podría ser una explicación de por qué, en promedio, estos contigs presentan un nivel de transcripción más bajo.

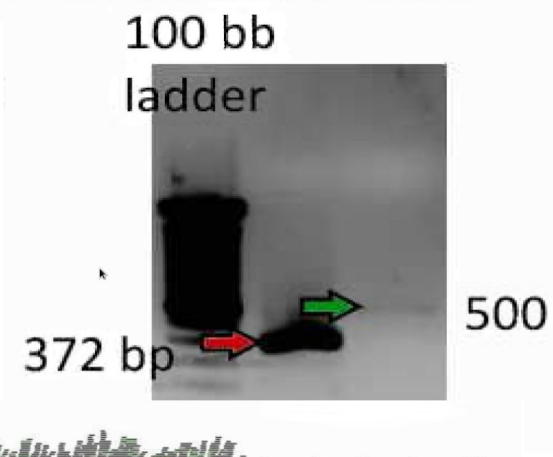
Además de haber contigs sin reads, hay otros en los que, al mapear los reads sobre ellos, presentan una zona que parece transcribirse activamente y otra en la que no aparece ningún read (figura 42 A).

Una explicación posible para todos esos contigs sin reads es que, en realidad, no estén en la cepa LIEM-176 de la que se obtuvo el ARN y se secuenciaron los reads, y sí estén en la Y486 cuyo genoma es el que estudiamos aquí y al cual estos contigs genómicos pertenecen. Para averiguarlo, fueron diseñados *primers* específicos para algunas de esas regiones (en rojo y verde en la figura 42). Los resultados del PCR mostrados en la figura 42 B y D muestran que claramente esas regiones en las que no matcheó ningún read están presentes en el genoma de la cepa LIEM-176, por lo que debemos concluir que lo más probable es que su ausencia en el transcriptoma de dicha cepa se deba a que, en efecto, no son transcriptos (Greif *et al.*, 2013). Esto tiene implicaciones importantes; el punto de vista generalmente aceptado es que «todo» el genoma es transcripto permanentemente en estos organismos, siendo la regulación de tipo postrancripcional. Sin embargo, estos resultados sugieren fuertemente que la regulación a nivel de la iniciación de la transcripción podría estar jugando un rol importante.

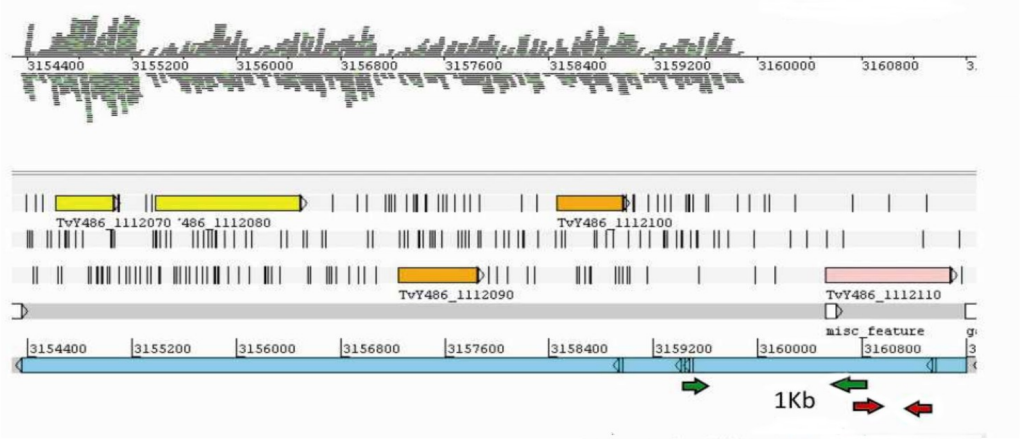
A



B



C



D

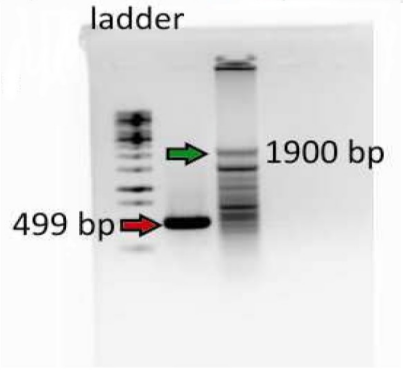


Figura 42: **Resultados de PCR y visualizaciones en Artemis del contig analizado (explicación en el texto) de Greif *et al* (2013).**

Búsqueda de la VSG activa

Con el fin de averiguar cuál era el nivel de transcripción del gen codificante de la VSG presente en la membrana del parásito (VSG activa), se recurrió a un método sencillo: se buscó entre los contigs que tenían más reads, y se encontró sólo una candidata a VSG. Sorpresivamente, tenía gran similitud de secuencia (algo más de 90 %) con la VSG reportada para *T. vivax* en una cepa de África occidental llamada Ildat 2.1 (Gardiner *et al.*, 1996). Esta VSG no aparece en el genoma de *T. vivax* disponible en el GenBank (cepa Y486) y, a la inversa, la VSG reportada como expresándose en la cepa Y486 no aparece en el transcriptoma presentado aquí de la cepa Liem-176. Siendo todas estas cepas originarias de África occidental (de donde proviene el *T. vivax* hoy presente en Sudamérica), parecería que la evolución por mutaciones puntuales y la originada a un nivel más macro por pérdida o ganancia de genes no marchan a la misma velocidad, al menos a este nivel temprano de diferenciación.

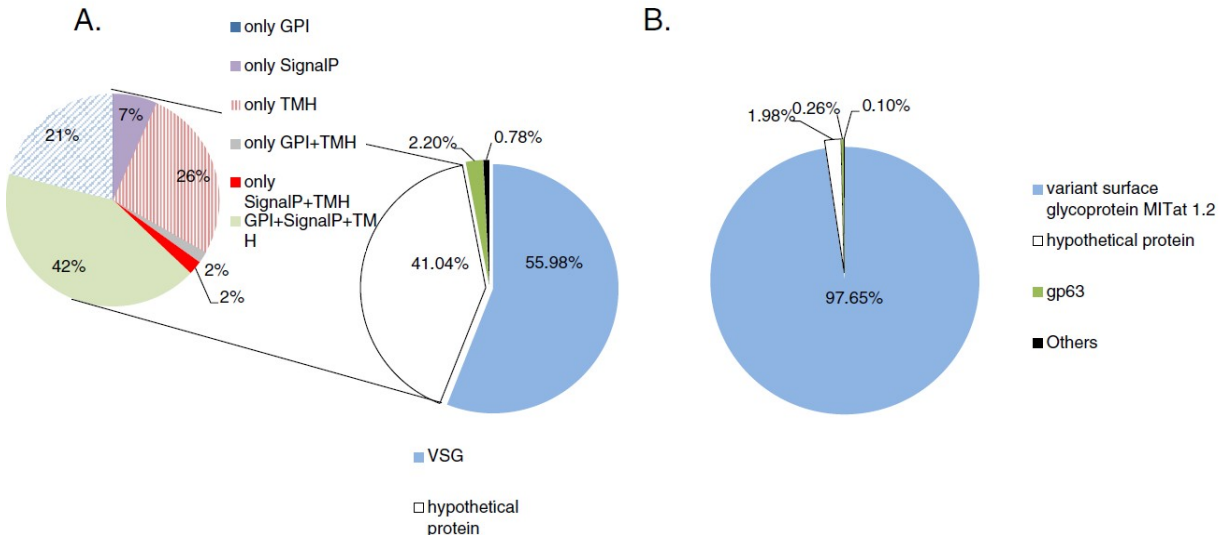


Figura 43. **Proporción de VSG en el total de proteínas de superficie en (A) *T. vivax* y (B) *T. brucei* (Greif *et al.*, 2013).**

Otro dato interesante es que a nivel de transcripción, la VSG activa de *T. vivax*, si bien muy abundante, lo es menos que la correspondiente a *T. brucei*: 56 % de los reads correspondientes a proteínas de membrana, contra 98 % en *T. brucei* (figura 43). Esto lleva a preguntarnos cuán eficiente pueden ser las VSG en cuanto a su rol inmunoevasor, en *T. vivax*, y posiblemente en el estado ancestral.

Conclusiones

Una de las principales conclusiones que podemos extraer de este trabajo es que el genoma de *T. vivax* se encuentra claramente compartimentado, distinguiéndose en él dos regiones: una (A) que ocupa el 75 % del genoma (o más exactamente, el 75 % del largo total de los contigs mayores de 1500 nt) y contiene la gran mayoría de los genes *housekeeping*, rica en genes codificantes de proteínas RHS y otros elementos transponibles, y otra (B), que representa la cuarta parte restante del genoma contiene a las secuencias codificantes de VSG y otras proteínas de superficie, así como la gran mayoría de las secuencias de tipo OOO (marcos abiertos de lectura de considerable longitud sin homólogos en otras especies). Esto marca dos diferencias con *T. brucei*, donde los RHS se ubican preferentemente en las regiones subteloméricas (junto a las VSG), y donde no han sido descritas regiones orfogénicas.

La existencia de secuencias de tipo OOO es en sí misma una novedad importante. Aparte de las genomas muy ricos en GC como las micobacterias u otros de similares características, no existen antecedentes de secuencias que posean gran cantidad de largos ORF superpuestos, en los seis marcos de lectura, y que ocupen una parte importante del genoma, que

además tiene otras características propias (entre otras, es en el espacio orfogénico donde residen los genes de VSG).

Hemos caracterizado y clasificado estos ORF en virtud de sus propiedades estadísticas, siendo divididos operativamente en cinco grupos. Cabe destacar que estos grupos no son absolutamente diferenciados, sino que, por ejemplo, a pesar de haber sido clasificadas aparte, dos secuencias pueden presentar regiones comunes. Si estas regiones están en distintas hebras (o en la misma hebra pero en distinto marco), pueden ocupar gran parte de al menos una de las secuencias; al menos en teoría y sin considerar los STOP, una secuencia perteneciente a un grupo puede ser idéntica a la complementaria de otra secuencia de otro grupo. En rigor, estaríamos hablando del mismo fragmento de ADN. Evidentemente, es necesario continuar los estudios en este tema para, entre otras cosas, darle sentido a esa clasificación (o, llegado el caso, modificarla). Sin embargo, en lo general, podemos decir con seguridad que el grupo 5 tiene propiedades similares a los genes previamente anotados y, en coherencia con esto, sus miembros se ubican en el compartimiento genómico A. Los otros cuatro grupos se ubican en el compartimiento B. En el grupo 2 están las secuencias tipo VSG, y en el 4, *grosso modo*, sus secuencias reversas complementarias. Los dos grupos restantes (1 y 3) también presentan complementariedad parcial entre sí, y en el caso del 1, un número no despreciable de ellos presenta homología (en distinta hebra) con algunas VSG, lo cual es consistente con la posición contigua que ambos grupos (1 y 4) ocupan en el gráfico tridimensional resultante del ACP basado en las frecuencias de codones.

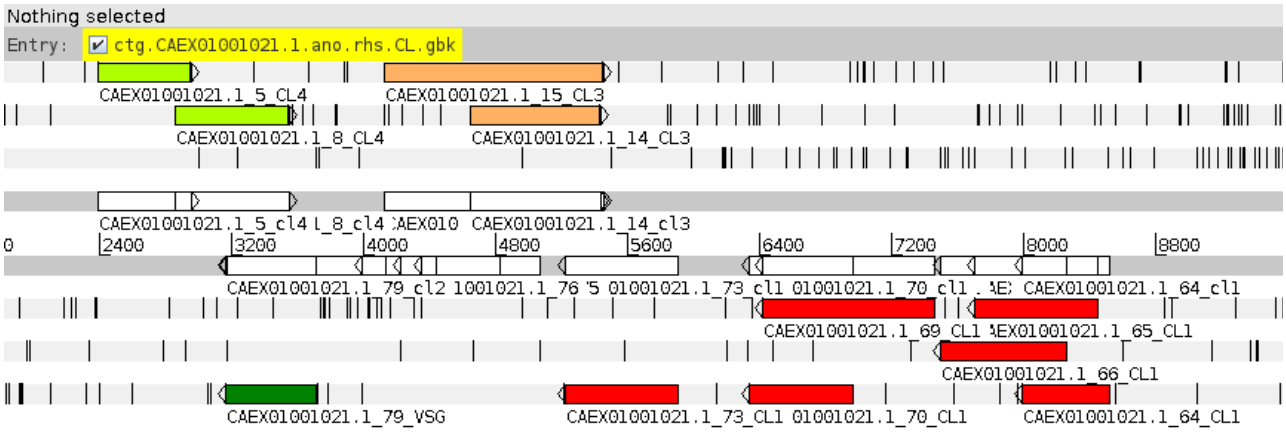


Figura 44. **Contig CAEX01001021.1, perteneciente al grupo B, en que se notan claramente las relaciones de complementariedad de los distintos tipos de ORF: cl1 (rojo), cl3 (anaranjado), cl2-VSG (verde oscuro), cl4 (verde claro).**

Volviendo a los resultados, por ahora sólo podemos especular acerca del significado biológico de los mismos. Para ello, hay que tener en cuenta que *T. vivax* se escindió tempranamente del resto de los tripanosomas africanos, por lo que las características peculiares que posee pueden: a) haber surgido con posterioridad a dicha separación, o b) ser ancestrales y haberse perdido en la rama restante. Tal vez para elucidar esta cuestión sea necesario esperar a que existan más genomas secuenciados: la presencia de alguna de estas características (por ejemplo, la existencia de OOO) en uno solo de los organismos pertenecientes a la otra rama, sería evidencia fuerte a favor del carácter ancestral de la misma.

Hemos presentado evidencia de cómo la composición de la secuencia, sobre todo a nivel de di y trinucleótidos, es esencial para favorecer la aparición de ORF largos en distintos marcos de lectura. Es tentador sugerir que, siendo estos ORF —o al menos parte de ellos— fragmentos que combinan partes de VSG con secuencias flanqueantes (previamente no codificantes), y existiendo la posibilidad (ya mencionada) de que se formen nuevas VSG a partir de, justamente, fragmentos de genes o pseudogenes silenciosos, la existencia de un gran número de

marcos abiertos de lectura con las características antedichas aumenta las posibilidades reales de creación de nuevas VSG, esto es, VSG cuya probabilidad de haber sido detectadas por el sistema inmune de un hospedador previamente infectado sea nula. Por otra parte, el hecho de que haya pocos codones stop permite la variación a través de la aparición de *frameshifts*, con un bajo riesgo de producir proteínas truncas (figura 45).

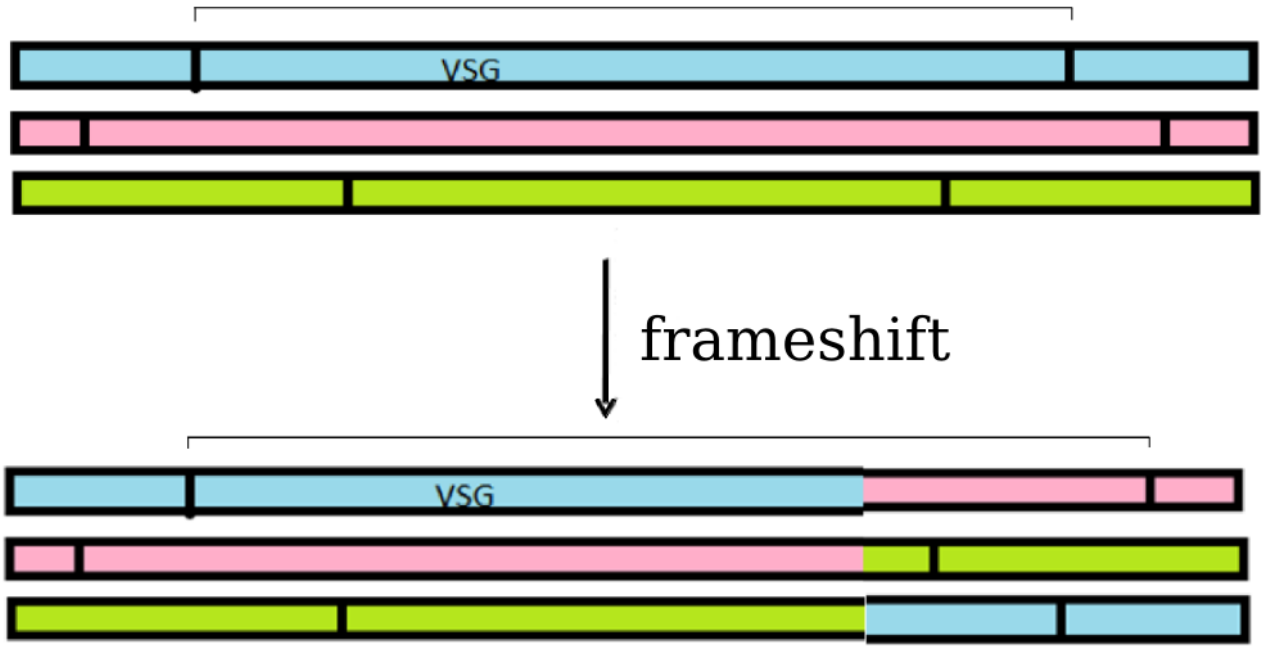


Figura 45. **Posible ventaja de la escasez de codones stop a la hora de generar variabilidad por medio de la aparición de indels que provoquen corrimiento del marco de lectura.**

Qué incidencia tiene esto en la dinámica de la relación entre parásitos y hospedadores, en su juego de evasión/detección, es difícil de cuantificar, pero no deja de ser una explicación atractiva. Es posible que muchos de nuestros genes hayan surgido a partir de genes con otra función, que tras alguna modificación aleatoria pasaron a codificar péptidos con una función diferente, incierta al principio, que fue haciéndose más clara tras un proceso de cambios regido por la selección natural. El sistema de generación rápida de nuevos genes de VSG a partir de

fragmentos de secuencias parcialmente no codificantes de tan amplia disponibilidad sería un banco de pruebas genético en tiempo acelerado (debido a las menores restricciones a nivel de secuencia que estas proteínas presentan para seguir siendo funcionales) y, por lo tanto, un modelo de estudio único.

Anexo I

markov2.pl

La idea central de las cadenas de Markov es que la probabilidad de que una base se encuentre en un sitio concreto no está determinada sólo por la frecuencia de dicha base (salvo en las llamadas cadenas de orden 0), sino por cuál es la base (orden 1) o el par de bases (orden 2) que la preceden. En la práctica, aplicado a la generación de secuencias de ADN, implica predefinir las frecuencias de cada dinucleótido (orden 1) o trinucleótido (orden 2), de modo que la suma de los cuatro trinucleótidos posibles que empiezan con determinado par, sea igual a 1, y lo mismo para los dinucleótidos que empiecen con determinada base. Pero dichas frecuencias se pueden calcular previamente a partir de una secuencia conocida. Para estos análisis escribimos tres scripts: markov0.pl, markov1.pl y markov2.pl. Cada uno de ellos tiene como output una secuencia con las mismas propiedades que una secuencia utilizada para «entrenar» el script. Lo que queríamos era averiguar si cierta propiedad, la generación de ORF largos, muchas veces superpuestos en distintos marcos de lectura, tenía que ver con la composición de la secuencia, y llevar el concepto de «composición» más allá de la frecuencia de cada una de las bases. Detallaré a modo de ejemplo cómo funciona el script markov2.pl.

El script consiste en algo más de trescientas líneas de código escrito en lenguaje perl. Utiliza como input una secuencia (puede estar en formato fasta, multifasta —en este caso se tratarán todas las secuencias como una sola—, o simplemente la secuencia sin más datos), preferentemente larga (por ejemplo, un contig o un conjunto de genes), cuyas propiedades queremos explorar. El output es otra secuencia, aleatoria, pero con prácticamente las mismas frecuencias de trinucleótidos que la secuencia de entrenamiento. El script utiliza la función rand(), de modo que la asignación de cada base se realiza en función de una probabilidad

preestablecida por lo que, cuanto más larga sea la secuencia output (el largo lo define el usuario), más seguridad tendremos con respecto a la fidelidad de la simulación, por la misma razón que si tiramos repetidas veces una moneda, cuanto más tiradas hagamos más seguros estaremos de acercarnos a un 50 % de caras. El pseudocódigo del script es el siguiente:

Se abre el archivo de entrada, y se lee su contenido, guardando la secuencia en una variable tipo string.

Se envía la secuencia a la subrutina makefrecs.

Dentro de la subrutina se cuentan los trinucleótidos utilizando un único hash. Por la forma en que está escrito el código, las keys de este hash son los trinucleótidos, y los values la cantidad de veces que dicho trinucleótido fue encontrado en la secuencia, de modo que si escribo, por ejemplo, print $hash{TTA}, me aparecerá en pantalla la cantidad de TTA, en cualquier marco, que hay en la secuencia de entrenamiento. Los trinucleótidos son extraídos de uno en uno, avanzando una base cada vez. A continuación se crean dieciséis variables denominadas $sumaAA, $sumaAC, etc., una por cada dinucleótido con el que puede empezar un trinucleótido. Tienen un valor numérico; por ejemplo, $sumaAA resulta de sumar $hash{AAA} + $hash{AAC} + $hash{AAG} + $hash{AAT}.

Posteriormente, utilizando un loop for, se imprimen en un archivo los cocientes resultantes de dividir cada uno de los valores del hash, por la variable suma correspondiente. Para el ejemplo, cada uno de los valores del hash cuya key empieza con AA, se divide por $sumaAA. El archivo tendrá el siguiente aspecto:

```
0.275244299674267 0.221498371335505 0.234527687296417 0.268729641693811
0.225641025641026 0.261538461538462 0.254700854700855 0.258119658119658
0.246376811594203 0.233494363929147 0.256038647342995 0.264090177133655
0.230421686746988 0.262048192771084 0.256024096385542 0.251506024096386
0.226475279106858 0.245614035087719 0.221690590111643 0.30622009569378
0.275806451612903 0.243548387096774 0.238709677419355 0.241935483870968
0.273489932885906 0.256711409395973 0.218120805369128 0.251677852348993
0.248012718600954 0.24483306836248 0.238473767885533 0.268680445151033
0.211726384364821 0.247557003257329 0.271986970684039 0.268729641693811
0.245222929936306 0.238853503184713 0.248407643312102 0.267515923566879
0.245928338762215 0.236156351791531 0.257328990228013 0.260586319218241
0.248366013071895 0.241830065359477 0.243464052287582 0.266339869281046
0.275039745627981 0.227344992050874 0.271860095389507 0.225755166931638
0.2660406885759 0.259780907668232 0.223787167449139 0.250391236306729
0.231974921630094 0.289968652037618 0.261755485893417 0.216300940438871
0.251497005988024 0.244011976047904 0.252994011976048 0.251497005988024
```

donde las filas corresponden a los 16 dinucleótidos de inicio (ordenados alfabéticamente) y las columnas a la tercera letra (en el mismo orden) de cada trinucleótido; de modo que el número de arriba a la derecha de la tabla, por ejemplo, corresponde a la frecuencia del trinucleótido AAT, calculada con respecto a todos los que empiezan con AA. Las filas, por lo tanto, suman 1. Para este caso se utilizó una secuencia aleatoria con 25 % de cada base; de ahí a que todos los valores ronden esa cifra.

Nótese que la subrutina no devuelve ningún valor (no hay *return*); simplemente crea el archivo, que será leído a continuación; es un recurso para simplificar el código del script, pero también permite conservar la tabla guardada, de modo que podemos leerla directamente después de correr el script si así lo deseamos.

A continuación se abre este archivo, y se guarda cada línea en una variable *string*, que contendrá por lo tanto los cuatro valores, separados por espacios.

Se envían las dieciséis variables a la subrutina simul.

Dentro de la sub simul, se guarda cada uno de las dieciséis strings pasadas en una variable, cuyo nombre son las dos letras correspondientes; la variable $aa tendrá la primera fila de la tabla, y así sucesivamente.

A continuación se generan aleatoriamente las dos primeras letras de la secuencia simulada. Recordemos que la probabilidad de cada letra, en esta secuencia, depende de las dos letras anteriores, por lo que hay que generar, de algún modo arbitrario, las dos primeras. Estas dos letras se guardan en dos variables: $simulada y $par_ant.

Después, en un loop for que correrá tantas veces como el largo querido de la secuencia simulada menos dos, se van enviando las cuatro frecuencias correspondientes a las letras de par_ant a una subrutina denominada new_base. Por ejemplo, si par_ant es AT, se enviará la variable $at, que contiene los valores de la cuarta línea de la tabla.

Dentro de la sub new_base se crea un array que contiene, como elementos separados, los cuatro valores que fueron pasados dentro de una misma string. Se genera un número aleatorio entre 0 y 1, y en función del número generado y los valores de probabilidad de cada uno de los elementos del array, se retorna una letra.

Otra vez dentro de la sub simul, dicha letra es anexada al final de la variable $simulada. Por último, se redefine el valor de $par_ant en una nueva subrutina, que elimina la primera letra del valor viejo de $par_ant, y adiciona, al final, la base recientemente generada (la que se acaba de anexar a $simulada). De este modo, $simulada va teniendo cada vez más letras, mientras que $par_ant siempre tiene las dos últimas letras de $simulada.

Cuando termina de correr el loop de la sub simul, el largo de $simulada es el que el usuario eligió. Se retorna por lo tanto la variable $simulada al cuerpo principal del script, donde se la imprime en un nuevo archivo de salida.

Para dar una idea acerca de la velocidad de este tipo de scripts, generar una secuencia de cien mil bases a partir de las propiedades de una de diez mil, tomó aproximadamente medio segundo en una computadora común de escritorio.

hebra.pl

Este es un script mucho más simple, de los que se hacen típicamente por decenas para estudiar aspectos puntuales, durante el desarrollo de cualquier trabajo de investigación *in silico*. Cuando analizamos los grupos en que se dividen los ORF de más de 500 nt que no tienen homólogos en otras especies, los «Orfogenic Orfan ORFs» (OOO), encontramos casos en que ORF de uno de los grupos tendían a ubicarse en la hebra complementaria de ORF de otro grupo. Estos datos salen del output del Blast. Al inspeccionar dicha salida se puede deducir, de acuerdo con el sentido de las coordenadas de la zona de similitud, tanto en el query como en el subject, si ambas secuencias son «iguales» (estadísticamente hablando) o si una es igual a la reversa complementaria de la otra. Ahora bien, la salida del Blast tiene miles de líneas que encierran gran cantidad de datos; frecuentemente se necesita «parsearla» (término derivado del inglés *parse*, que significa algo así como analizar, aunque en estos casos el sentido que se le da es más bien «filtrar») para presentar en una forma más clara los datos que nos interesan particularmente en ese momento. Con este script, ante una salida de un Blastn realizado entre dos conjuntos de secuencias, cada uno de los cuales tiene miles de elementos, obtenemos simplemente dos números: uno que nos indica cuántos hits hubo «en la misma hebra» y cuántos

en la complementaria. Al decir hits nos referimos en realidad al *best hit*, es decir, en este caso, por cada secuencia tomamos aquel hit para el cual el e-value es menor.

También escrito en perl, tiene unas treinta líneas de código (diez veces menos que el anterior). A diferencia de aquel no utiliza subrutinas, aunque sí imprime archivos (en este caso temporales) que son releídos por el programa. Además, desde el propio script se corren otros programas (nativos del sistema Linux). Estos son: cat, cut, sort y grep.

En primer lugar, utilizando los tres primeros de estos scripts, se genera un archivo con los nombres de todos los queries que encontraron al menos un hit en la base de datos.

Después se abre ese archivo y se guarda la lista en un array.

Con un loop for, y utilizando el comando grep con la opción m 1 (que nos devuelve la primera aparición de la string que buscamos en un archivo, y en la salida de Blast la primera aparición corresponde al best hit), se recrea una salida de Blast igual a la original pero que incluye sólo los best hits.

Se abre ese nuevo archivo, y mediante un loop *while* se lo va leyendo línea a línea.

Cada línea es convertida, a su vez, en un array (con el comando split). Sabemos que los elementos [8] y [9] de ese array corresponden a las coordenadas de la región de similitud en el hit. Como las coordenadas en el query siempre se dan en orden creciente, sabemos que si las del hit están en orden inverso (por ejemplo, 345 23) esto se debe a que la similitud se dio no con la secuencia del hit sino con su reversa complementaria. Tal sería el caso de estas dos secuencias:

```
query       nnnnnnTCCCGTGCnnnnn
subject     nnGCACGGAnnnnnnnnn
```

Siendo los segmentos en mayúsculas reversos complementarios (con la excepción de un indel de una base), las coordenadas del query serían 7 14 y las del subject 9 3. En el caso de las salidas de Blast que incluyen el alineamiento, se sustituye la secuencia del subject por su reversa complementaria y se aclara que la similitud fue encontrada en la hebra (-). En este caso, el alineamiento se parecería a

```
query    7 TCACGTGC 14
             || |||||
subject  9 TC-CGTGC 3
```

A continuación, simplemente se contabiliza en cuántas líneas se da que array[8] < array[9] (misma hebra) o lo opuesto (hebra complementaria).

Por último, se imprimen esas sumas en pantalla.

Anexo 2 - Abreviaturas[3]

ACP Análisis de Componentes Principales.

ARNPI ARN Polimerasa I.

C-terminal Extremo carboxilo terminal de una proteína.

CDS Coding DNA Sequence

cl1-5 Cada uno de los 5 grupos (clusters) en que fueron clasificados los OOO.

COG Clusters of Orthologous Groups.

EMBL European Molecular Biology Laboratory.

EMBOSS European Molecular Biology Open Software Suite.

ES Expression Site.

ESAG Expression Site Asociate Genes.

GO Gene Onotology.

GPI Glycophosphatidylinositol.

Met Metionina, aquí se lo usa generalmente referido al codón ATG que la codifica, que también cumple la función de codón de iniciación de la traducción.

[3] Se siguió el criterio hispánico de no pluralizar las siglas, incluso cuando provengan del inglés. Así, se escribió *los ORF*, y no *los ORFs.*

N-terminal Extremo amino terminal de una proteína.

NCBI National Center for Biotechnology Information.

nt Nucleótidos.

OOO Orphan Orfogenic ORF.

ORF Open Reading Frame.

pb Pares de bases.

RHS Retrotransposon Hot Spot.

Signal-pep Secuencia corta en el extremo N-terminal, relacionada con el destino final de la proteína y su capacidad de atravesar membranas.

SL Spliced Leader.

VSG Variant Surface Glycoprotein.

Referencias bibliográficas

Akopyants N, N Kimblin, N Secundino, R Patrick, N Peters, P Lawyer, D E. Dobson, S M. Beverley, D L. Sacks, (2009). Demonstration of Genetic Exchange During Cyclical Development of *Leishmania* in the Sand Fly Vector. Science 10 (324) 5924: 265-268 DOI: 10.1126/science.1169464.

Altschul S, W Gish, W Miller, E Myers, D Lipman (1990). Basic local alignment search tool. Journal of Molecular Biology 215 (3): 403-410. doi:10.1016/S0022-2836(05)80360-2. PMID 2231712

Chevreux B, T Pfisterer, B Drescher, AJ Driesel, WE Muller, T Wetter, S Suhai (2004). Using the mira EST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. Genome Res, 14(6):1147-1159.

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

Cortez AP, RM Ventura, AC Rodrigues, JS Batista, F Paiva, N Añez, RZ Machado, WC Gibson, MM Teixeira (2006). The taxonomic and phylogenetic relationships of *Trypanosoma vivax* from South America and Africa. Parasitol 133: 159-169. doi:10.1017/ S0031182006000254.

El-Sayed NM, PJ Myler, G Blandin, M Berriman *et al.* (2005). Comparative Genomics of Trypanosomatid Parasitic Protozoa. Science 309, 404. DOI: 10.1126/science.1112181.

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

Gardiner, PR y MM Mahmoud (1992). Salivarian trypanosomes causing disease in livestock outside sub-saharan Africa. In Parasitic Protozoa (ed. Kreier, J. P. y Baker, J. R.): 277-313. Londres, Academic Press.

Gaunt, M W, M Yeo, I A Frame, J R Stothard, H J Carrasco, M C Taylor, S S Mena, P Veazey, G A Miles, N Acosta, A R de Arias, M A Miles (2003). Mechanism of genetic exchange in American trypanosomes. Nature, 2003; 421(6926):936-9

Greif, G, M Ponce de León, G Lamolle, M Rodriguez, D Piñeyro, LM Tavares-Marques, A Reyna-Bello, C Robello y F Alvarez-Valin (2013). Transcriptome analysis of the bloodstream stage from the parasite *Trypanosoma vivax*. BMC Genomics. 2013 Mar 5;14(1):149. [Epub ahead of print]

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

Hamilton P B (2012). Is *Trypanosoma vivax* genetically diverse? (Letter). Trends in Parasitology, 28 (5): 173.

Iseli C, C V Jongeneel, P Bucher (1999). ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. Proc Int Conf Intell Syst Mol Biol.138-48.

Jackson A P, A Berrya, M Asletta, HC Allisonb, P Burtonc, J Vavrova-Andersonc, R Brownd, H Brownea, N Corton, H Hauser, J Gamble, R Gilderthorp, L Marcello, J McQuillan, T D Otto, MA Quail, MJ Sanders, A van Tonder, ML Ginger, MC Field, JD Barry, C Hertz-Fowler, M Berriman (2012). Antigenic diversity is generated by distinct evolutionary mechanisms in African trypanosome species. PNAS, 109 (9): 3416-3421.

Jenni, L, S Marti, J Schweizer, B Betschart, R W F Le page, J M Wells, A Tait, P Paindavoine, E Pays, M Steinert (1986). Hybrid formation between African trypanosomes during cyclical transmission. Nature 322: 173-175; doi:10.1038/322173a0.

Jones, TW y AM Davila (2001). *Trypanosoma vivax*-out of Africa. Trends in Parasitology 17: 99-101. doi:10.1016/S1471-4922(00)01777-3.

Kanehisa M, S Goto, Y Sato, M Furumichi, M Tanabe (2012). KEGG for integration and interpretation of large-scale molecular datasets. Nucleic Acids Res. 40, D109-D114.

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

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

Larkin MA, G Blackshields, NP Brown, R Chenna, PA McGettigan, H McWilliam, F Valentin, IM Wallace, A Wilm, R Lopez, JD Thompson, TJ Gibson and DG Higgins (2007). Clustal W and Clustal X version 2.0, Bioinformatics (2007) 23 (21): 2947-2948. doi: 10.1093/bioinformatics/btm404

Li H y R Durbin (2009). Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics, 25:1754-60. [PMID: 19451168]

Liang, X H, A Haritan, S Uliel, y S Michaeli (2003) Trans and cis splicing in trypanosomatids: mechanism, factors, and regulation. Eukaryot Cell 2: 830-840.

Malele, I, L Craske, C Knight, V Ferris, Z Njiru, P Hamilton, S Lehane, M Lehane, WC Gibson (2003). The use of specific and generic primers to identify trypanosome infections of wild tsetse flies in Tanzania by PCR. Infection, Genetics and Evolution 3, 271-279. doi:10.1016/S1567-1348(03)00090-X.

Mandelboim M, S Barth, M Biton, XH Liang, S Michaeli (2003). Silencing of Sm proteins in *Trypanosoma brucei* by RNA interference captured a novel cytoplasmic intermediate in spliced leader RNA biogenesis. J Biol Chem 278: 51469-51478.

Marcello L y J David Barry (2007). Genome Res.; 17(9): 1344-1352. doi: 10.1101/gr.6421207

Morariu V I, B Vasan Srinivasan, VC Raykar, R Duraiswami, LS Davis (2008). Automatic online tuning for fast Gaussian summation. Advances in Neural Information Processing Systems (NIPS).

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

Osório AL, CR Madruga, M Desquesnes, CO Soares, LR Ribeiro, SC Costa (2008). *Trypanosoma (Duttonella) vivax*: its biology, epidemiology, pathogenesis, and introduction in the New World–a review. Mem Inst Oswaldo Cruz; 103: 1-13.

Otto TD, AC Guimarães, WM Degrave, AB de Miranda (2008). AnEnPi: identification and annotation of analogous enzymes. MC Bioinformatics; 9: 544.

Palenchar JB, V Bellofatto (2006). Gene transcription in trypanosomes. Mol Biochem Parasitol 146: 135-141.

Pays E, L Vanhamme, D Pérez-Morga D (2004). Antigenic variation in *Trypanosoma brucei*: facts, challenges and mysteries. Curr Opin Microbiol 7(4): 369-74.

R Core Team (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, <http://www.R-project.org/>.

Rutherford K, J Parkhill, J Crook, T Horsnell, P Rice, MA Rajandream, B Barrell (2000). Artemis: sequence visualization and annotation, Bioinformatics (Oxford, Inglaterra);16;10;944-5.

Schafer Da Silva A, M Machado Costa, M Flores Polenz; CH Polenz; MM Geraldes Teixeira, ST Dos Anjos Lopes, S Gonzalez Monteiro (2010). Ciência Rural ISSN 0103-8478.P.

Simarro PP, A Diarra, JA Ruiz Postigo, JR Franco, JG Jannin (2011). The Human African Trypanosomiasis Control and Surveillance Programme of the World Health Organization 2000-2009: The Way Forward. PLoS Negl Trop Dis 5(2): e1007. doi:10.1371/journal.pntd.0001007

Stevens, JR, HA Noyes, GA Dover, WC Gibson (1988). The ancient and divergent origins of the human pathogenic trypanosomes, *Trypanosoma brucei* and *T. cruzi*. Parasitology 118: 107-116.

Thompson JD *et al.* (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22:4673-4680.

Turner CMR y JD Barry (1989). High frequency of antigenic variation in *Trypanosoma brucei rhodesiense* infections. Parasitology, 99: 67-75 doi:10.1017/S0031182000061035.

Vanhamee L, E Pays, R Mcculloch, JD Barry, (2001). An update on antigenic variation in African trypanosomes. Trends in Parasitology (17) 7: 338-343.

Vickerman KP (1976). TM: Biology of the kinteplastida. Londres, Academic.

WHO-report (2004). <http://www.who.int/whr/2004/en/>.

Yoshida N (2009). Molecular mechanisms of *Trypanosoma cruzi* infection by oral route. Mem Inst Oswaldo Cruz, Rio de Janeiro, Vol. 104(Suppl. I): 101-107.

Zdobnov E M, Apweiler R (2001). InterProScan - an integration platform for the signature-recognition methods in InterPro. Bioinformatics, 17(9): 847-8.

BMC Genomics

BioMed Central
The Open Access Publisher

Transcriptome analysis of the bloodstream stage from the parasite Trypanosoma vivax

BMC Genomics 2013, **14**:149 doi:10.1186/1471-2164-14-149

Gonzalo Greif (ggreif@pasteur.edu.uy)
Miguel Ponce de Leon (miguelponcedeleon@gmail.com)
Guillermo Lamolle (lamoogle@gmail.com)
Matías Rodriguez (matidae@gmail.com)
Dolores Piñeyro (pineyro@pasteur.edu.uy)
Lucinda M Tavares-Marques (lmtm17@gmail.com)
Armando Reyna-Bello (areyna@inmunobiologia.net.ve)
Carlos Robello (robello@pasteur.edu.uy)
Fernando Alvarez Valin (falvarez@fcien.edu.uy)

ISSN 1471-2164

Article type Research article

Submission date 14 July 2012

Acceptance date 15 February 2013

Publication date 5 March 2013

Article URL http://www.biomedcentral.com/1471-2164/14/149

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

Gonzalo Greif[1,†]
Email: ggreif@pasteur.edu.uy

Miguel Ponce de Leon[2,†]
Email: miguelponcedeleon@gmail.com

Guillermo Lamolle[2]
Email: lamoogle@gmail.com

Matías Rodriguez[2]
Email: matidae@gmail.com

Dolores Piñeyro[1,3]
Email: pineyro@pasteur.edu.uy

Lucinda M Tavares-Marques[4]
Email: lmtm17@gmail.com

Armando Reyna-Bello[4]
Email: areyna@inmunobiologia.net.ve

Carlos Robello[1,3]
Email: robello@pasteur.edu.uy

Fernando Alvarez Valin[2*]
[*] Corresponding author
Email: falvarez@fcien.edu.uy

[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

[†] Equal contributors.

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.

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 $2x10^7$ 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: $= \sum_i N_i P_i$. 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

1. 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.

2. 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.

3. 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.

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

4. 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, 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).

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.

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 appreciated in Figure 3 panels A and B, 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.

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.

5. Gene expression levels and codon biases in trypanosomes.

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.

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)

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 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 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.

6. 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].

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.

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.

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*.

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).

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.

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 8 A and B).

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).

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 8 C 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 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 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 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.

Competing interests

Authors declare that they have no competing interests

Authors' contribution

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.

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 Ia 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.

Additional files

Additional_file_1 as DOC

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

Additional_file_2 as DOCX

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

Additional_file_3 as XLS

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

Additional_file_4 as XLS

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 as XLS

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

Additional_file_6 as PDF

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

Additional_file_7 as PDF

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

Additional_file_8 as DOC

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 as JPG

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

Additional_file_10 as XLS

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 as PNG

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 as PPT

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

Additional_file_13 as DOCX

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

Additional_file_14 as XLS

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

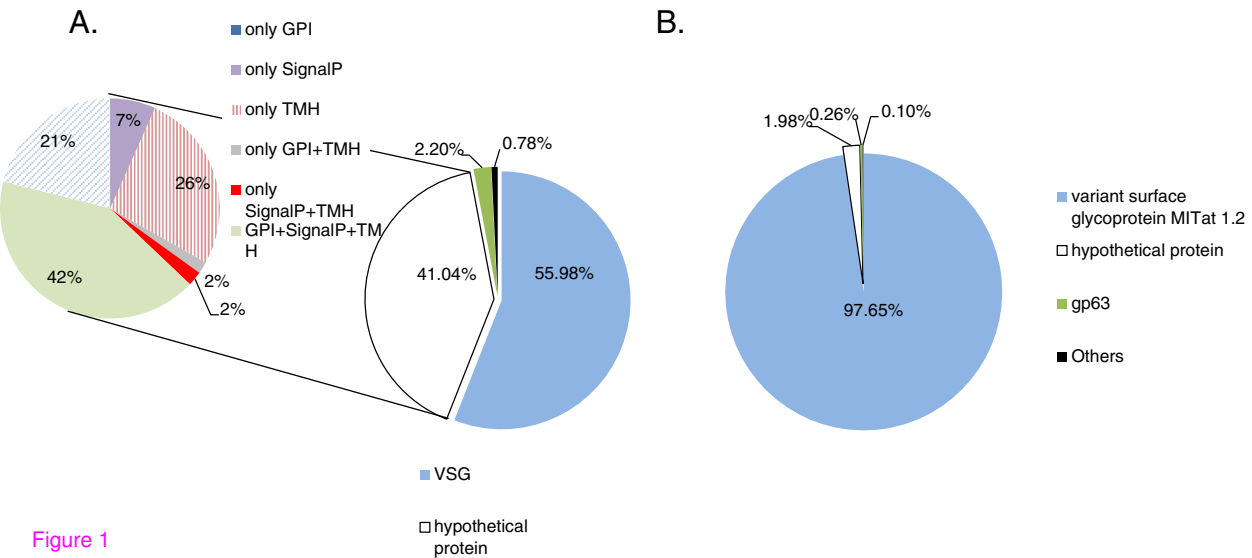

A.
only GPI
only SignalP
only TMH
only GPI+TMH
only SignalP+TMH
GPI+SignalP+TMH
7%
21%
26%
42%
2%
2%
2.20%
0.78%
41.04%
55.98%
VSG
hypothetical protein
B.
1.98%
0.26%
0.10%
97.65%
variant surface glycoprotein MITat 1.2
hypothetical protein
gp63
Others

Figure 1

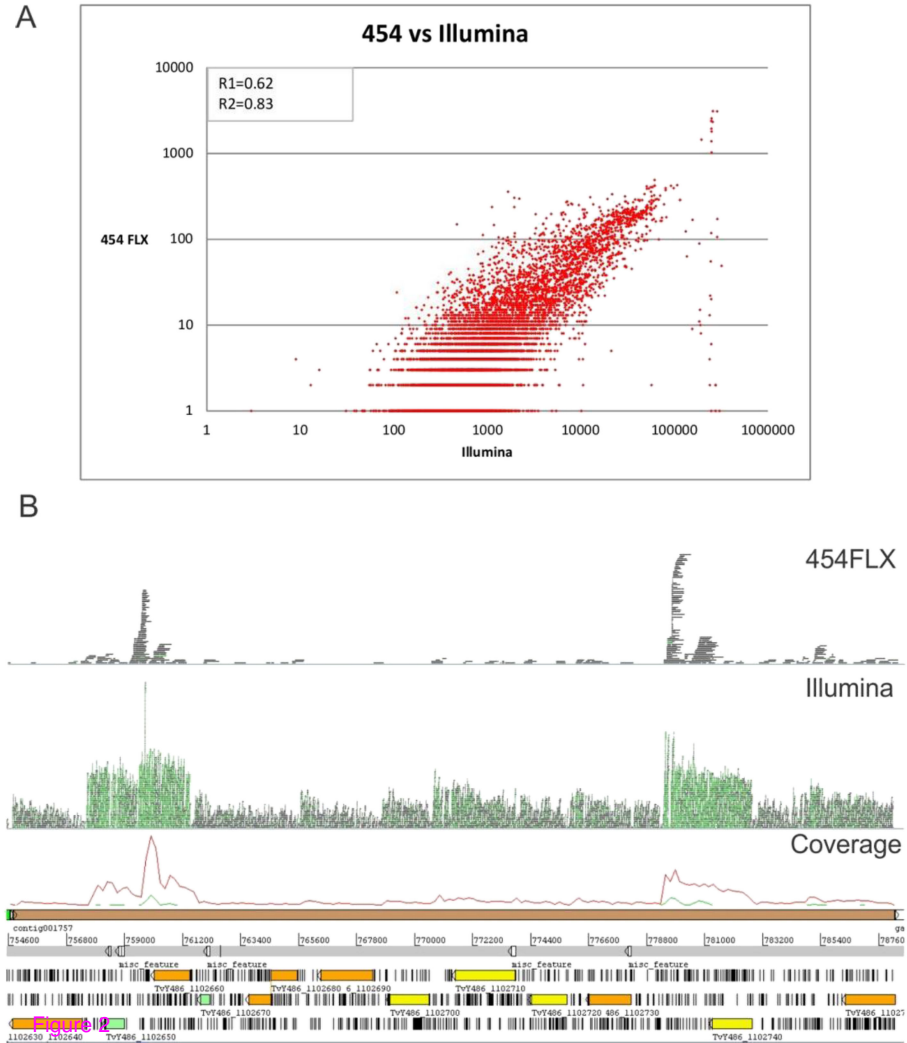

A
454 vs Illumina
R1=0.62
R2=0.83
454 FLX
Illumina
10000
1000
100
10
1
1
10
100
1000
10000
100000
1000000
B
454FLX
Illumina
Coverage
contig001757
misc_feature
TvY486_1102660
TvY486_1102680_6_1102690
TvY486_1102710
TvY486_1102670
TvY486_1102700
TvY486_1102720_486_1102730
TvY486_1102740

Figure 1B

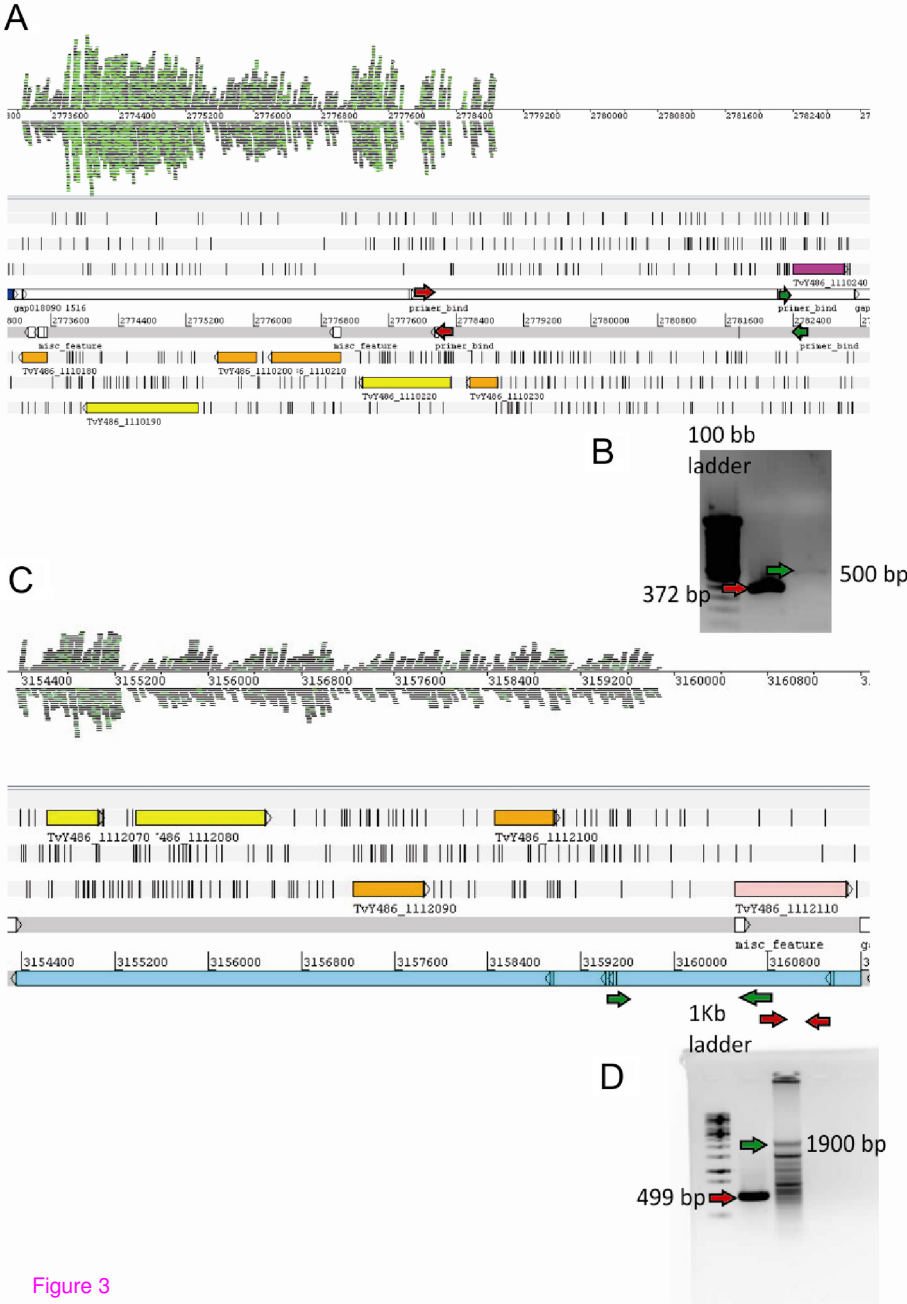


Figure 3

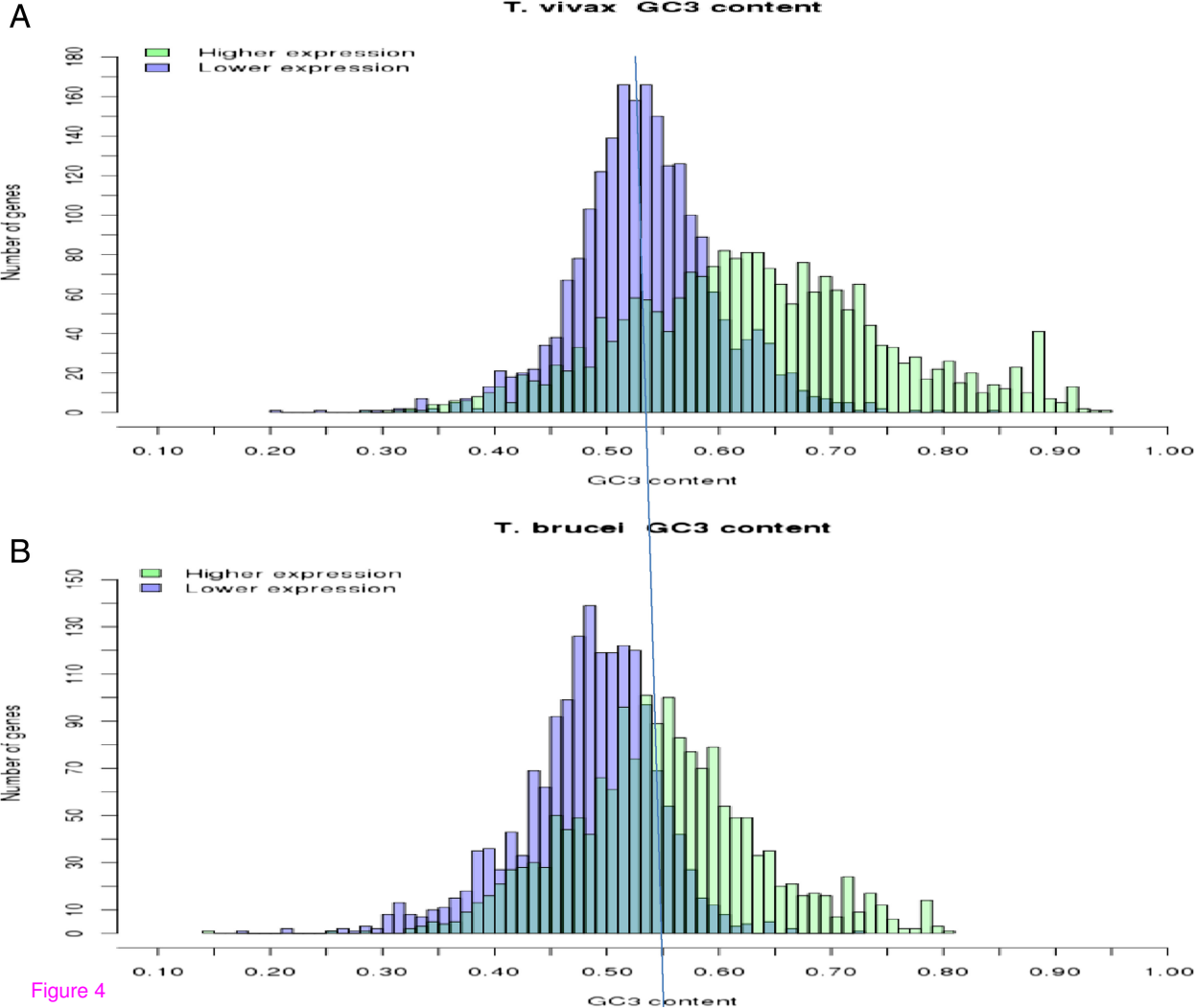
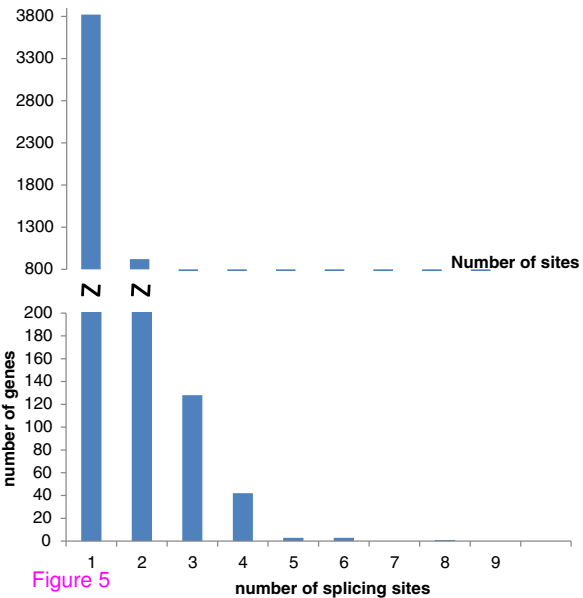

A
T. vivax GC3 content
Higher expression
Lower expression
Number of genes
0
20
40
60
80
100
120
140
160
180
0.10
0.20
0.30
0.40
0.50
0.60
0.70
0.80
0.90
1.00
GC3 content
B
T. brucei GC3 content
Higher expression
Lower expression
Number of genes
0
10
30
50
70
90
110
130
150
0.10
0.20
0.30
0.40
0.50
0.60
0.70
0.80
0.90
1.00
GC3 content

Figure 4

A


Splicing sites / gene
3800
3300
2800
2300
1800
1300
800
200
180
160
140
120
100
80
60
40
20
0
Number of sites
number of genes
1
2
3
4
5
6
7
8
9
number of splicing sites

B

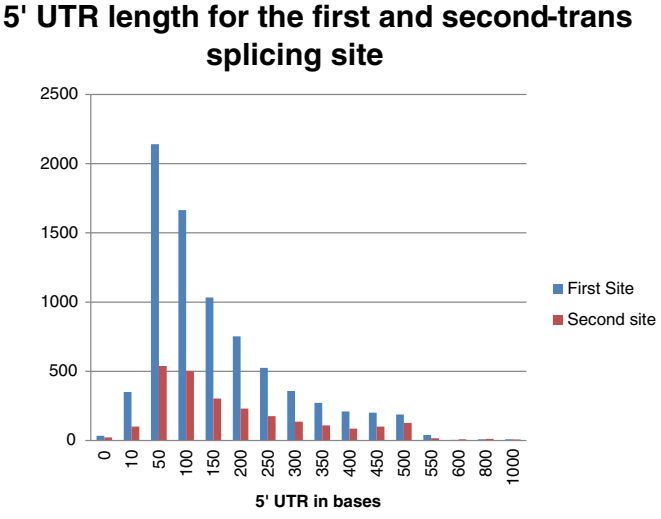
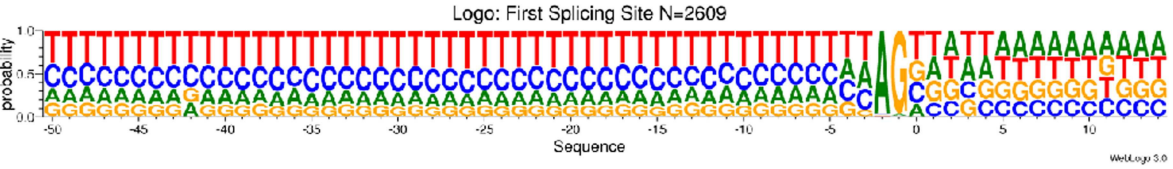
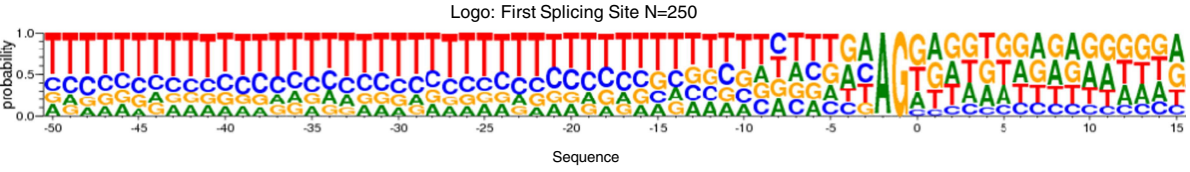

5' UTR length for the first and second-trans splicing site
2500
2000
1500
1000
500
0
0
10
50
100
150
200
250
300
350
400
450
500
550
600
800
1000
5' UTR in bases
First Site
Second site

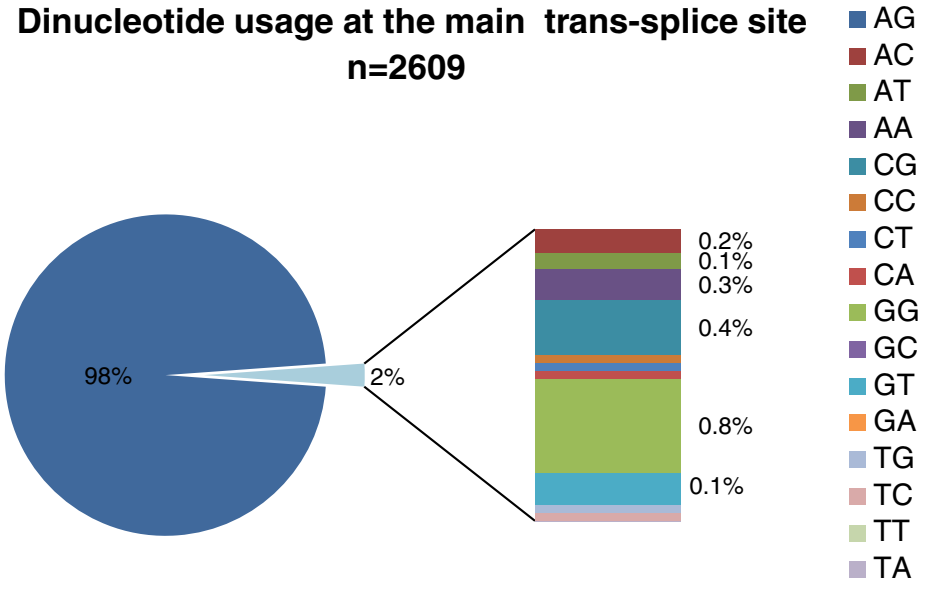
Figure 5


Logo: First Splicing Site N=2609
probability
1.0
0.5
0.0
-50
-45
-40
-35
-30
-25
-20
-15
-10
-5
0
5
10
Sequence

A


Logo: First Splicing Site N=250
probability
1.0
0.5
0.0
-50
-45
-40
-35
-30
-25
-20
-15
-10
-5
0
5
10
15
Sequence

B


Dinucleotide usage at the main trans-splice site
n=2609
98%
2%
0.2%
0.1%
0.3%
0.4%
0.8%
0.1%
AG
AC
AT
AA
CG
CC
CT
CA
GG
GC
GT
GA
TG
TC
TT
TA

C

Figure 6

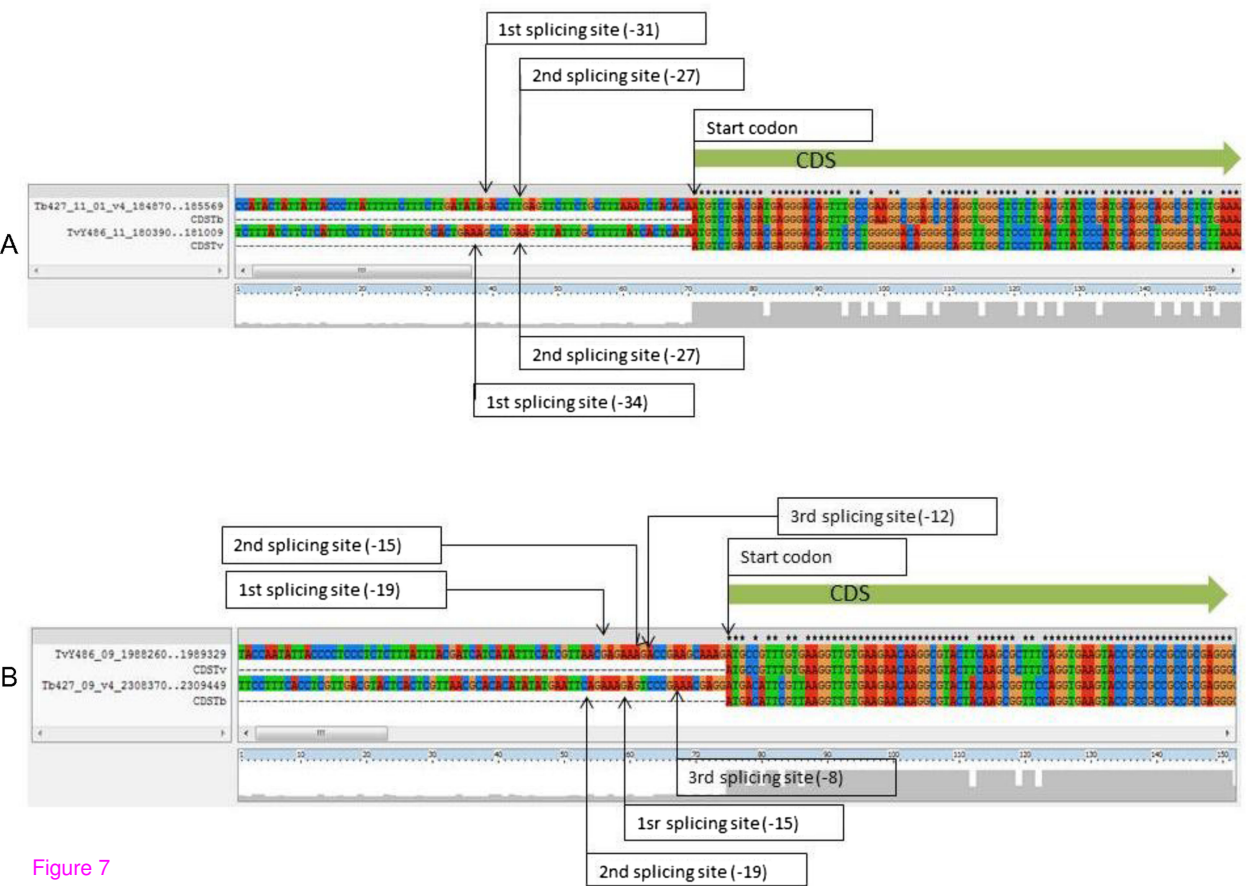

A
1st splicing site (-31)
2nd splicing site (-27)
Start codon
CDS
Tb427_11_01_v4_184870..185569
CDSTb
TvY486_11_180390..181009
CDSTv
2nd splicing site (-27)
1st splicing site (-34)
B
2nd splicing site (-15)
1st splicing site (-19)
3rd splicing site (-12)
Start codon
CDS
TvY486_09_1988260..1989329
CDSTv
Tb427_09_v4_2308370..2309449
CDSTb
3rd splicing site (-8)
1sr splicing site (-15)
2nd splicing site (-19)

Figure 7

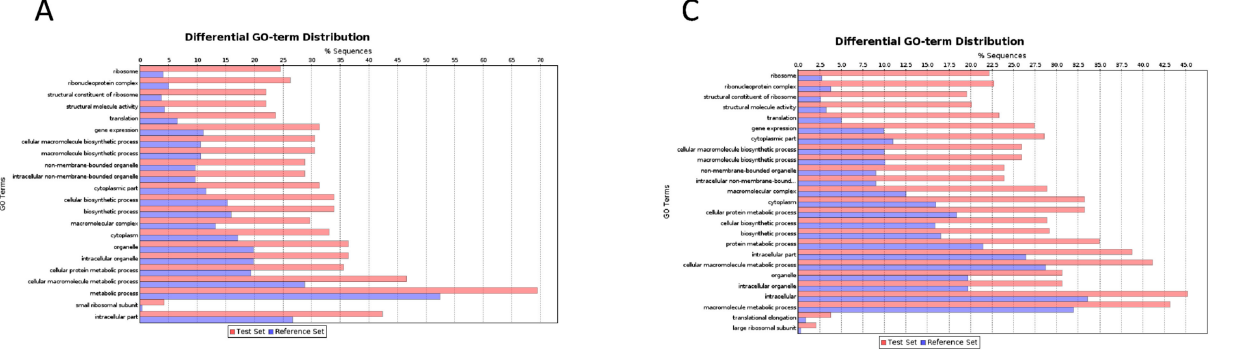

A
Differential GO-term Distribution
% Sequences
0 5 10 15 20 25 30 35 40 45 50 55 60 65 70
GO Terms
ribosome
ribonucleoprotein complex
structural constituent of ribosome
structural molecule activity
translation
gene expression
cellular macromolecule biosynthetic process
macromolecule biosynthetic process
non-membrane-bounded organelle
intracellular non-membrane-bounded organelle
cytoplasmic part
cellular biosynthetic process
biosynthetic process
macromolecular complex
cytoplasm
organelle
intracellular organelle
cellular protein metabolic process
cellular macromolecule metabolic process
metabolic process
small ribosomal subunit
intracellular part
Test Set
Reference Set
C
Differential GO-term Distribution
% Sequences
0.0 2.5 5.0 7.5 10.0 12.5 15.0 17.5 20.0 22.5 25.0 27.5 30.0 32.5 35.0 37.5 40.0 42.5 45.0
GO Terms
ribosome
ribonucleoprotein complex
structural constituent of ribosome
structural molecule activity
translation
gene expression
cytoplasmic part
cellular macromolecule biosynthetic process
macromolecule biosynthetic process
non-membrane-bounded organelle
intracellular non-membrane-bound...
macromolecular complex
cytoplasm
cellular protein metabolic process
cellular biosynthetic process
biosynthetic process
protein metabolic process
intracellular part
cellular macromolecule metabolic process
organelle
intracellular organelle
intracellular
macromolecule metabolic process
translational elongation
large ribosomal subunit
Test Set
Reference Set
B
Differential GO-term Distribution
% Sequences
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
GO Terms
structural constituent of ribosome
translation
small ribosomal subunit
Test Set
Reference Set
D
Differential GO-term Distribution
% Sequences
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
GO Terms
structural constituent of ribosome
translational elongation
large ribosomal subunit
Test Set
Reference Set

Figure 8

Additional files provided with this submission:

Additional file 1: 9044020997617635_add1.doc, 30K
http://www.biomedcentral.com/imedia/3584841739364512/supp1.doc
Additional file 2: 9044020997617635_add2.docx, 210K
http://www.biomedcentral.com/imedia/1947604922936451/supp2.docx
Additional file 3: 9044020997617635_add3.xls, 37K
http://www.biomedcentral.com/imedia/6600064993645136/supp3.xls
Additional file 4: 9044020997617635_add4.xls, 1228K
http://www.biomedcentral.com/imedia/6692358839364513/supp4.xls
Additional file 5: 9044020997617635_add5.xls, 90K
http://www.biomedcentral.com/imedia/1394775892936451/supp5.xls
Additional file 6: 9044020997617635_add6.pdf, 53K
http://www.biomedcentral.com/imedia/3339581119364514/supp6.pdf
Additional file 7: 9044020997617635_add7.ppt, 736K
http://www.biomedcentral.com/imedia/1573727839936451/supp7.ppt
Additional file 8: 9044020997617635_add8.doc, 29K
http://www.biomedcentral.com/imedia/8421817099364514/supp8.doc
Additional file 9: 9044020997617635_add9.jpg, 112K
http://www.biomedcentral.com/imedia/3173083069364514/supp9.jpeg
Additional file 10: 9044020997617635_add10.xls, 238K
http://www.biomedcentral.com/imedia/6485035229364514/supp10.xls
Additional file 11: 9044020997617635_add11.png, 66K
http://www.biomedcentral.com/imedia/1098086251936451/supp11.png
Additional file 12: 9044020997617635_add12.pptx, 383K
http://www.biomedcentral.com/imedia/1929535059936451/supp12.pptx
Additional file 13: 9044020997617635_add13.docx, 25K
http://www.biomedcentral.com/imedia/1712532197936451/supp13.docx
Additional file 14: 9044020997617635_add14.xls, 123K
http://www.biomedcentral.com/imedia/1429396012936451/supp14.xls