Estudio de la organización y dinámica evolutiva del genoma de *Trypanosoma vivax*

Matias Rodriguez

Maestría en Bioinformática

Sección Biomatemática, Unidad de Genómica Evolutiva

Orientadores: Fernando Alvarez

Alberto Ferrarini

FACULTAD DE CIENCIAS
UDELAR | fcien.edu.uy

PEDECIBA

ANII

Resumen

Los tripanosomas africanos son protozoarios parásitos flagelados que se desarrollan y multiplican en la sangre, el fluido tisular de sus hospederos. Su vector de transmisión son insectos hematófagos. En América *Trypanosoma vivax* fue introducido a mediados del siglo XIX al importar ganado infectado procedente de África, desde donde se ha expandido rápidamente. La principal diferencia entre las cepas americanas y africanas de *T.vivax* es la forma en la cual el parásito es transmitido. En África los parásitos realizan parte de su ciclo de vida en el aparato digestivo del insecto, mientras que en América los parásitos se transfieren de forma mecánica ya que no se han adaptado al vector. Una característica de los tripanosomas africanos es la capacidad para mantener infecciones crónicas gracias a un complejo mecanismo de evasión de la respuesta inmune denominado variación antigénica que le permite cambiar de forma periódica la capa de antígenos de superficie (VSG) que lo recubren y protegen de la respuesta inmune del hospedero. *T.vivax* es de particular interés para el estudio de la organización genómica y la variación antigénica debido a que estudios filogenéticos indican que es un descendiente directo del ancestro común de los tripanosomas africanos y es muy probable que esta especie presente una organización genómica y de la superfamilia de genes VSG en un estado ancestral. En el presente trabajo se realizó un estudio de genómica evolutiva y comparativa centrado en la comparación del genoma de las cepas americanas y africanas enfocándose en el estudio de la región genómica con genes asociados a los mecanismos de variabilidad antigénica. Además se estudiaron en detalle los cambios observados en el genoma mitocondrial como resultado de la adaptación a la transferencia mecánica en el continente americano. Se ensamblaron y anotaron los genomas de dos cepas americanas de *T.vivax* las cuales se analizaron junto a una cepa africana depositada en el Genbank. Se definió el concepto de "orfogenicidad", como la potencialidad de una secuencia de generar marcos abiertos de lectura, una característica singular de algunas regiones del genoma de *T. vivax*. Utilizando métodos de análisis multivariados se identificaron regiones o compartimentos del genoma que presentan características en cuanto a su frecuencia nucleotídica y se estudió como se relacionan con diferentes tipos de compartimentos genómicos. Con la finalidad de cuantificar la orfogenicidad del genoma se desarrollaron varios índices que estiman la contribución de diferentes factores, como las frecuencias nucleotídicas, el contenido GC y los codones stop. Estos análisis permiten distinguir en el genoma nuclear dos grandes compartimentos genómicos, uno asociado a genes del core y otro asociado a regiones orfogénicas, donde se encuentran los genes involucrados en la variación antigénica. Además se encontró que en las cepas americanas, donde no realizan su ciclo de vida en el insecto, el genoma mitocondrial se encuentra en proceso de degradación ya que no existe una presión selectiva para mantenerlo.

Introducción

La tripanosomiasis comprende un conjunto de patologías que afectan a vertebrados y se encuentran asociadas a una infección por parásitos pertenecientes al género *Trypanosoma*. Su distribución es mundial aunque el área de mayor incidencia son las regiones tropicales y subtropicales del planeta. Son de particular interés porque afectan a humanos, animales salvajes, domésticos y de consumo por lo cual revisten gran importancia tanto médica como veterinaria.

Los tripanosomas son protozoarios parásitos flagelados que se desarrollan y multiplican en la sangre, el fluido tisular de sus hospederos así como intracelularmente. Su vector de transmisión son comúnmente insectos hematófagos. Durante su ciclo de vida ocupan dos hospederos, el insecto vector y un mamífero, aunque también se pueden encontrar especies de tripanosomas con un solo hospedero.

Dos especies de tripanosomas producen tripanosomiasis humana, *Trypanosoma cruzi* en América que causa el llamado "mal de Chagas" y *Trypanosoma brucei* en África que provoca la "enfermedad del sueño".

El vector de *T. cruzi* son insectos hemípteros hematófagos de la subfamilia *Triatominae* conocidos bajo el nombre de vinchuca. Cuando un triatomino infectado se alimenta de la sangre de un mamífero, libera los parásitos en sus heces, cerca del sitio de la herida. Los parásitos entran a través de la herida o a través de una membrana mucosa como la conjuntiva (Giddings et al. 2006). Otra forma de infección probablemente más ancestral es a través de la mucosa gástrica al ingerir alimentos contaminados (Hoft et al. 1996). Según la OMS *T. cruzi* afecta entre 6 y 7 millones de personas, la mayoría en América del Sur.

En África subsahariana *T. brucei* es responsable de la tripanosomiasis africana humana, la cual presenta dos patologías diferentes; una infección crónica característica de las subespecies de África Occidental y otra aguda en África Oriental. El vector de transmisión son varias especies de moscas del género *Glossina* comúnmente conocidas como moscas tsé-tsé. La infección se produce cuando la mosca se alimenta de sangre e inocula los parásitos que migran a nódulos linfáticos, se reproducen en la sangre e invaden tejidos causando daños en varios órganos, pudiendo provocar incluso problemas neurológicos. La enfermedad es endémica en treinta y seis países, poniendo en riesgo a millones de personas. Los esfuerzos para controlar la enfermedad y disminuir la aparición de nuevos casos han logrado que la mortandad disminuya de 34.000 muertes anuales en 1990 a 9.000 en 2012 (Lozano et al. 2012).

Además de los graves efectos sobre la salud humana, varias especies de tripanosomas afectan animales salvajes y domésticos. Cuando se encuentra en animales la enfermedad es llamada "nagana", un vocablo derivado del idioma zulú que significa "debilitado" o "deprimido". Los

síntomas incluyen pérdida sustancial de peso, tasa reducida de crecimiento, disminución de la fertilidad en machos, abortos y daños en órganos, lo cual sin un tratamiento puede provocar una elevada mortalidad.

En general los tripanosomas suelen tener un amplio rango de hospederos y son varias las especies de importancia veterinaria; por ejemplo *T. brucei brucei, T. congolense* y *T. vivax* afectan principalmente bovinos y ovinos, *T. simiae* suinos, *T. equiperdum* y *T. evansi* infectan mayormente equinos provocando la enfermedad llamadas "durina" y "surra" respectivamente.

La tripanosomiasis africana animal ha tenido un profundo impacto en la capacidad ganadera en extensas zonas de África y ha afectado profundamente el asentamiento y desarrollo económico en una gran parte del continente. Históricamente el impacto de la tripanosomiasis animal africana ha sido tal que ha marcado el camino de las rutas migratorias de tribus con ganado y los movimientos de los primeros exploradores árabes y europeos que dependían de caballos y bueyes como medio transporte (McKelvey 1973).

En América *T. vivax* fue introducido a mediados del siglo XIX al importar ganado infectado procedente de África Occidental a la Guyana Francesa e islas del mar Caribe. Fue identificado por vez primera en América en la Guyana Francesa en 1919, desde donde se ha expandido rápidamente por América Central y del Sur encontrándose ganado infectado desde el sur de México hasta el estado de Rio Grande do Sul en Brasil (Schafer et al 2009).

Aunque *T. congolense* produce mayor mortalidad, *T. vivax* es el patógeno más extendido en el ganado en África y está distribuido en las áreas donde se encuentra la mosca tsé-tsé que es el vector del parásito y donde realiza parte de su ciclo de vida.

También se ha encontrado en zonas adyacentes donde se presume la ausencia del vector, lo cual sugiere que se ha adaptado a un modo de transmisión independiente de tsé-tsé. En América el principal vector son especies de *Tabanus spp.* y la transmisión es solamente mecánica, es decir que no realiza parte de su ciclo de vida en el insecto (Desquenes & Dia 2003).

A diferencia de *T. cruzi* y *Lehismania spp.* los tripanosomas africanos son parásitos extracelulares que se encuentran constantemente expuestos al sistema inmune del hospedero y poseen como defensa una densa cubierta de glicoproteínas variables de superficie (VSG).

Una característica de los tripanosomas africanos es la capacidad para mantener infecciones crónicas gracias a un complejo mecanismo de evasión de la respuesta inmune denominado variación antigénica que le permite cambiar de forma periódica la capa de antígenos de superficie que lo recubren y protegen de la respuesta inmune del hospedero.

T. vivax es de particular interés para el estudio de la organización genómica y la variación antigénica debido a que estudios filogenéticos indican que este tripanosoma es un descendiente directo del ancestro común de los tripanosomas africanos (Cortes et al. 2006), por lo que es muy

probable que esta especie presente una organización genómica y de la superfamilia de genes VSG en un estado ancestral.

Además la dispersión de *T. vivax* en regiones tropicales de América indica una rápida adaptación a las condiciones del continente y su expansión hacia los estados del sur de Brasil señala que la tripanosomiasis africana es una parasitosis emergente en la región.

Figura 1.- Frotis de una muestra de sangre de un bovino infectado con *Trypanosoma vivax* obtenida en Mato Grosso do Sul, Brasil. Tinción panóptico, vista de inmersión 1000x (Osorio et al. 2008).

Patología y distribución

Los síntomas que se desarrollan en todos los mamíferos infectados por tripanosomas africanos son similares. En humanos, la tripanosomiasis africana se denomina "enfermedad del sueño" y los síntomas son dolor de cabeza, fiebre, fatiga, hinchazón de nódulos linfáticos, dolor en músculos y articulaciones. Puede provocar problemas neurológicos luego de la invasión del sistema nervioso central y si no es tratada puede ser mortal (Stich 2002).

Existen dos tipos de tripanosomiasis africana en humanos dependiendo de la especie involucrada en la infección. La forma más habitual es la infección por el parásito *T. b. gambiense* que representa el 98% de los casos de la reportados. El área de distribución de este parásito es África occidental y central (los países más afectados son RD del Congo, Angola, Congo, Sudán del Sur, Uganda, República Centroafricana, Guinea, Costa de Marfil, Camerún y Nigeria), y provoca una infección crónica donde pueden pasar meses o años sin mayores síntomas, pero cuando ocurren la enfermedad suele encontrarse en un estado avanzado y afectar el sistema nervioso central. En África oriental (principalmente Uganda, Tanzania, Malawi y Zambia) se encuentra la especie *T. b. rhodesiense* que se caracteriza por provocar síntomas agudos que se desarrollan en semanas o meses luego de la infección invadiendo rápidamente el sistema nervioso central (Smith et al. 1998).

En la tripanosomiasis humana africana una vez que los parásitos atraviesan la piel crecen y causan hinchazón localizada. Luego migran a los nódulos linfáticos, y más tarde a la sangre donde se multiplican. En estadios avanzados invaden tejidos causando daños en varios órganos, además pueden llegar al sistema nervioso central causando problemas neurológicos.

En América Latina cuatro especies de tripanosomas son de importancia médica y veterinaria: *T. cruzi*, *T. evansi*, *T. equiperdum* y *T. vivax*. Solo *T. cruzi*, un tripanosoma del grupo Stercoraria es nativo de América, mientras que las otras tres especies del grupo Salivaria han sido importados junto con sus hospederos.

El tripanosoma de mayor dispersión y el más común en el ganado es *T. evansi*, encontrándose en áreas tropicales y subtropicales de África, Asia y América. Provoca la surra y es particularmente patogénico en camélidos y equinos. *T. equiperdum* también afecta a equinos, provocando la durina, y es el único tripanosoma conocido hasta el momento que tiene transmisión sexual; se lo encuentra en África, América, Asia y Europa Oriental.

T. vivax es un parásito mayormente de rumiantes (vacunos, ovinos, cabras y búfalos) causando anemia y pérdidas significativas de productividad

Ciclo de vida

Las etapas del ciclo de vida (figura 2), los estadios del parásito y el sitio anatómico en el insecto vector donde ocurren varían dependiendo de la especie de tripanosoma involucrada.

El ciclo de vida en el vector comienza cuando la sangre de un animal infectado es ingerida por un insecto hematófago; en el caso de *T. brucei* y *T. congolense* experimenta una serie de cambios importantes, adquiriendo la forma procíclica. La membrana celular pierde su cobertura de glicoproteínas variables de superficie (VSG) siendo sustituidas por prociclinas como GPEET, EP1 y EP3 (Günzl et al, 2003). Luego *T. brucei* y *T. congolense* migran hacia el intestino medio donde pueden multiplicarse por fisión binaria. Esta etapa no tiene similar en *T. vivax*. Se ha observado en *T. brucei* que la fisión binaria resulta en parásitos con dos formas asimétricas bien diferenciadas, una de epimastigotas largos que no continúan en el ciclo y mueren y otra con epimastigotas cortos que sobreviven (Ziegelbauer et al. 1990).

Posteriormente el parásito en su forma epimastigota migra en el insecto hacia las glándulas salivales en *T. brucei*, el proventrículo en *T. congolense* o la probóscide en *T. vivax*, se adhiere al epitelio y continúa dividiéndose y la superficie celular de *T. brucei* se reviste con las proteínas BARP (Brucei Alanine Rich Proteins). En esta etapa se ha observado intercambio genético vía meiosis en *T. brucei* (Peacock et al. 2011).

Finalmente se convierten en tripomastigotas metacíclicos, un estadio donde el parásito ha detenido su crecimiento y es la forma que será inoculada cuando el insecto vaya a alimentarse en un individuo no infectado. Las proteínas BARP de superficie son reemplazadas por VSG (Nolan et al. 2000).

En el mamífero hospedero los tripomastigotas metacíclicos se convierten en tripomastigotas sanguíneos de forma alargada y delgada ("slender") que son transportados a través del cuerpo del animal hasta alcanzar otros fluidos como el linfático y el espinal. La forma tripomastigota del parásito es capaz de multiplicarse por fisión binaria en la sangre y linfa, y el hospedero ahora puede transmitir el parásito cuando otra insecto vector se alimente.

En el caso de elevada parasitemia se ha observado en *T. brucei* una detención en el ciclo celular y una diferenciación en la forma recortada ("stumpy") que es la forma infectiva para el insecto. Bajo esta forma el organismo se encuentra detenido en la etapa G1 del ciclo celular (Shapiro et al. 1984).

A los tripanosomas africanos se los denomina Salivaria, y reciben este nombre porque el parásito es transmitido con la saliva del insecto durante la picadura que realiza para alimentarse. Parte del ciclo de vida de este grupo de parásitos es llevado a cabo en las partes bucales del insecto vector.

Estos agentes vectores son varias especies de “tábanos” del género *Glossina*, comúnmente conocidos como moscas tsé-tsé. Ejemplo de Salivaria son los tripanosomas africanos como *T. brucei*, *T. congolense* y *T. vivax*.

La mayor parte de las transmisiones a través de los vectores es cíclica, es decir el parásito realiza parte de su ciclo de vida en el vector, donde prolifera, aumenta su infectividad y experimenta cambios morfológicos y metabólicos notables. Algunos de estos cambios son el reemplazo de las glicoproteínas de superficie y cambios en la actividad mitocondrial.

Los tripanosomas africanos han logrado expandirse por fuera del área de distribución del vector *Glossina* utilizando otros vectores hematófagos donde no realizan su ciclo de vida y funcionan solamente transmitiendo el parásito en forma mecánica junto con la saliva al momento de alimentarse.

Debe tenerse en cuenta que la transmisión cíclica y mecánica coexisten en proporciones altamente variables dependiendo del área geográfica. No hay duda del papel de *Glossina* como principal vector de transmisión de la tripanosomiasis africana, sin embargo en zonas limítrofes al rango de expansión de su principal vector es importante considerar el impacto de la transmisión mecánica (Allsop & Newton 1985).

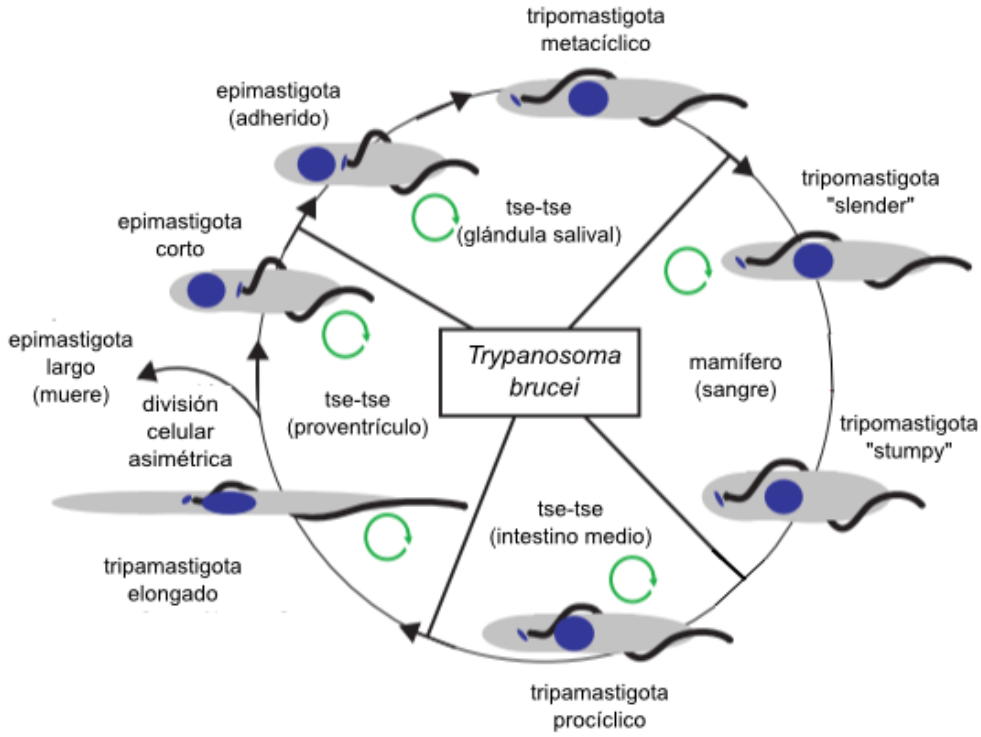

Figura 2.- Ciclo de vida de *T. brucei*, un Salivaria tipo. En *T. vivax* todo el ciclo en el insecto se desarrolla en las partes bucales. Con flecha circular verde se indican las etapas proliferativas (Kramer 2011).

Morfología

Los kinetoplástidos deben su nombre al genoma de su única mitocondria gigante ubicada en la zona basal del flagelo que tiene forma circular o elíptica y es de 1.1 µm de diámetro. Sus características son muy particulares y representa un carácter unificador. De hecho la presencia de DNA extranuclear fácilmente visible con tinciones de Giemsa es uno de los indicadores utilizados para identificar kinetoplástidos.

Este organelo se caracteriza por su compleja estructura de DNA y proceso de edición de los mRNAs. En los tripanosomas africanos su funcionamiento muy diferente en dos de las principales etapas del ciclo de vida, siendo metabólicamente muy activa durante su estadio en el insecto vector y luego con una actividad muy reducida en la etapa sanguínea en el hospedero vertebrado (Vickerman 1965).

El DNA del kinetoplasto (kDNA) se ubica en una posición periflagelar y su estructura tiene la forma de un disco con numerosos filamentos transmembrana que se unen al cuerpo basal del flagelo.

En *Trypanosoma brucei* se han observado tres formas diferentes circulando en la sangre de

animales inoculados, una forma alargada "slender", una más corta "stumpy" y una forma intermedia. Estas formas se corresponden con diferentes etapas de la infección; durante la fase ascendente de parasitemia la forma predominante es la "slender", mientras que en la fase descendente la forma "stumpy" es la más frecuente (Langousis & Hill 2014).

En su forma sanguínea *T. vivax* es un tripanosoma de tamaño medio, con un largo de 18-31 µm y 1.5-3 µm de ancho; en comparación con las cepas africanas, el tamaño de los parásitos observados en América es algo menor, de entre 16-26.5 µm. El tamaño del flagelo es de unos 7 µm y su membrana ondulante se encuentra generalmente atrofiada; el kinetoplasto se encuentra en posición terminal y la región posterior es redondeada (Desquenes 2004).

En África, *T. vivax* presenta un ciclo de desarrollo polimórfico y es posible encontrar las formas "slender" y "stumpy". Esto no se ha observado en América donde se transmite de forma mecánica y solo se ha observado la forma "slender" (Desquenes 2004).

Filogenia y evolución

Los miembros del clado *Trypanosoma* forman parte del orden *Trypanosomatida* que comparten con parásitos como *Crithidia*, *Leishmania* y *Phytomonas* entre otros. El orden *Trypanosomatida* forma junto a tres órdenes de bodónidos la subclase *Metakinetoplastina*, la cual junto a la subclase *Prokinetoplastina*, conforma la clase *Kinetoplastea* (figura 3) (Simpson et al. 2006).

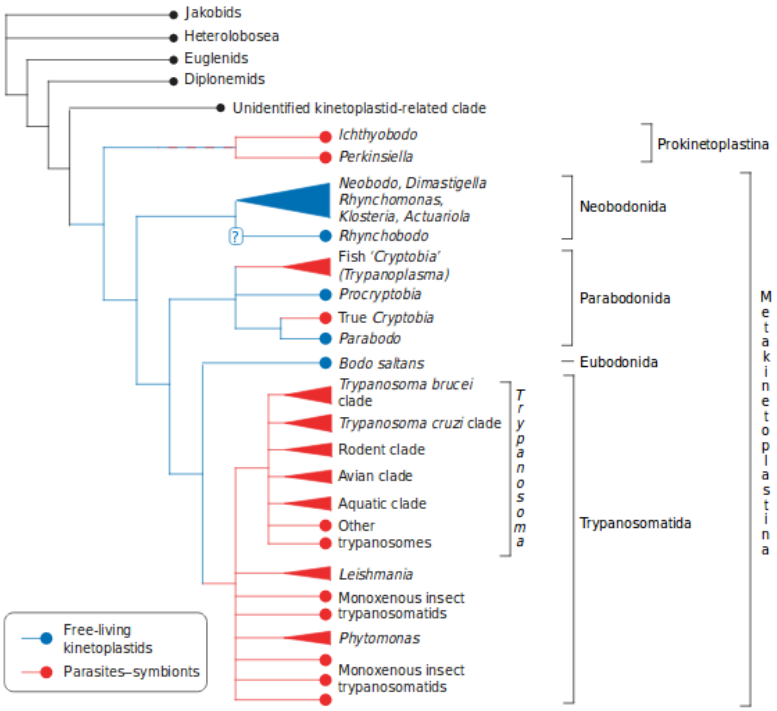

Figura 3.- Figura tomada de Simpson et al. 2006, mostrando las relaciones filogenéticas entre los tripanosomas y el resto de los kinetoplástidos. En rojo se indican las especies de parásitos o simbiontes y en azul los de vida libre.

Existen diferentes hipótesis sobre el origen del parasitismo en los kinetoplástidos. Un posible escenario es que los kinetoplástidos de vida libre primero fueran parásitos de invertebrados y que luego desarrollaron un ciclo de vida que incluyó un hospedero vertebrado luego que su hospedero invertebrado se volviera hematófago. Un ciclo de vida similar puede suponerse para aquellos parásitos de plantas como *Phytomonas*. Análisis filogenéticos sugieren que el paso hacia el parasitismo debió ocurrir varias veces en la historia evolutiva de estos organismos (Lukeš el al. 2014).

Los tripanosomas que infectan mamíferos y tienen importancia en la salud humana se dividen como Salivaria y Stercoraria dependiendo de su modo de transmisión. Todos los Salivaria conocidos evaden la respuesta humoral inmune del hospedero a través de una cubierta de proteínas VSG. Ésta cubierta es altamente inmunogénica pero capaz de activar genes de forma diferencial y adelantarse a la respuesta inmune produciendo infecciones crónicas.

En los Stercoraria la adaptación del parásito al vector triatomino es muy robusta, por lo que se estima que es muy antigua, cerca del 100% de los vectores son capaces de ser infectados (Desquenes 2004).

Por el contrario, se sospecha que el grupo Salivaria es el resultado de una evolución mucho más tardía a partir de tripanosomas que se desarrollaron cíclicamente en el intestino posterior del insecto, pero que también podían transmitirse mecánicamente, en particular por *Glossina*.

Se piensa que esta es la razón por la cual es más vulnerable, estando el nivel de vulnerabilidad relacionado con las complejidades que presenta su ciclo de vida.

Esto se ve reflejado en que las tasas de inefectividad en *Glossina* sean de un 20% para *T. vivax*, que tiene la primera etapa evolutiva con desarrollo en la proboscis, 10% para *T. congolense*, que representa una segunda etapa evolutiva con desarrollo en el intestino, seguido por migración a las partes bucales y solo un 1% para *T. brucei* que representa una instancia evolutiva mucho más compleja, con una forma metacíclica infectiva encontrada en las glándulas salivales luego de pasar a través del intestino. Esta hipótesis es apoyada por lo frágil que resulta la transmisión cíclica en *Glossina*, donde se ha observado que ciertas cepas de tripanosomas cultivadas en laboratorio tras cierto tiempo en ausencia de transmisión cíclica pierden su capacidad de infectar el vector nuevamente (Desquenes 2004).

Análisis filogenéticos basados en pequeñas subunidades de RNA ribosomales (SSU rRNA) muestran que los tripanosomas Salivaria forman un clado monofilético. Como se observa en la figura 4, dentro de los Salivaria el primer linaje en diverger fue el de *T. vivax*, del subgénero *Duttonella*, seguido por la división del subgénero *Trypanozoon* (*T. brucei* y *T. equiperdum*) y luego *Nannomonas* (*T. congolense* y *T. simiae*) (Haag et al. 1998).

La separación temprana de *T. vivax* dentro de los Salivaria representa una etapa inicial en la

adaptación a la transmisión por el vector *Glossina*. Estudios del 18S rRNA muestran que el grupo Salivaria evoluciona más de cuatro veces más rápido que otros tripanosomas y dentro del grupo Salivaria *T. vivax* es el de más rápida evolución a una tasa que duplica a otros Salivaria (Stevens & Rambaut 2001).

Aunque todos los tripanosomas Salivaria son transmitidos por la saliva de la mosca tsé-tsé el ciclo de vida de *T. vivax* no se desarrolla en el intestino como en otros subgéneros sino que completa su desarrollo en las partes bucales de la mosca.

En *T. vivax* la organización de los cromosomas muestra que tiene 1-2 minicromosomas a diferencia de los 50-100 de otros Salivaria. Además es un parásito predominantemente de bóvidos presentando un rango de hospederos mamíferos más reducido que otras especies del clado. Estas características biológicas y moleculares son compatibles con su ubicación en la periferia de los Salivaria (Cortez et al 2006).

Las cepas de *T. vivax* provenientes de Venezuela y centro de Brasil muestran una mayor relación con las de África Occidental, (Nigeria) que con las de las de África Oriental (Kenia) (Cortez et al 2006).

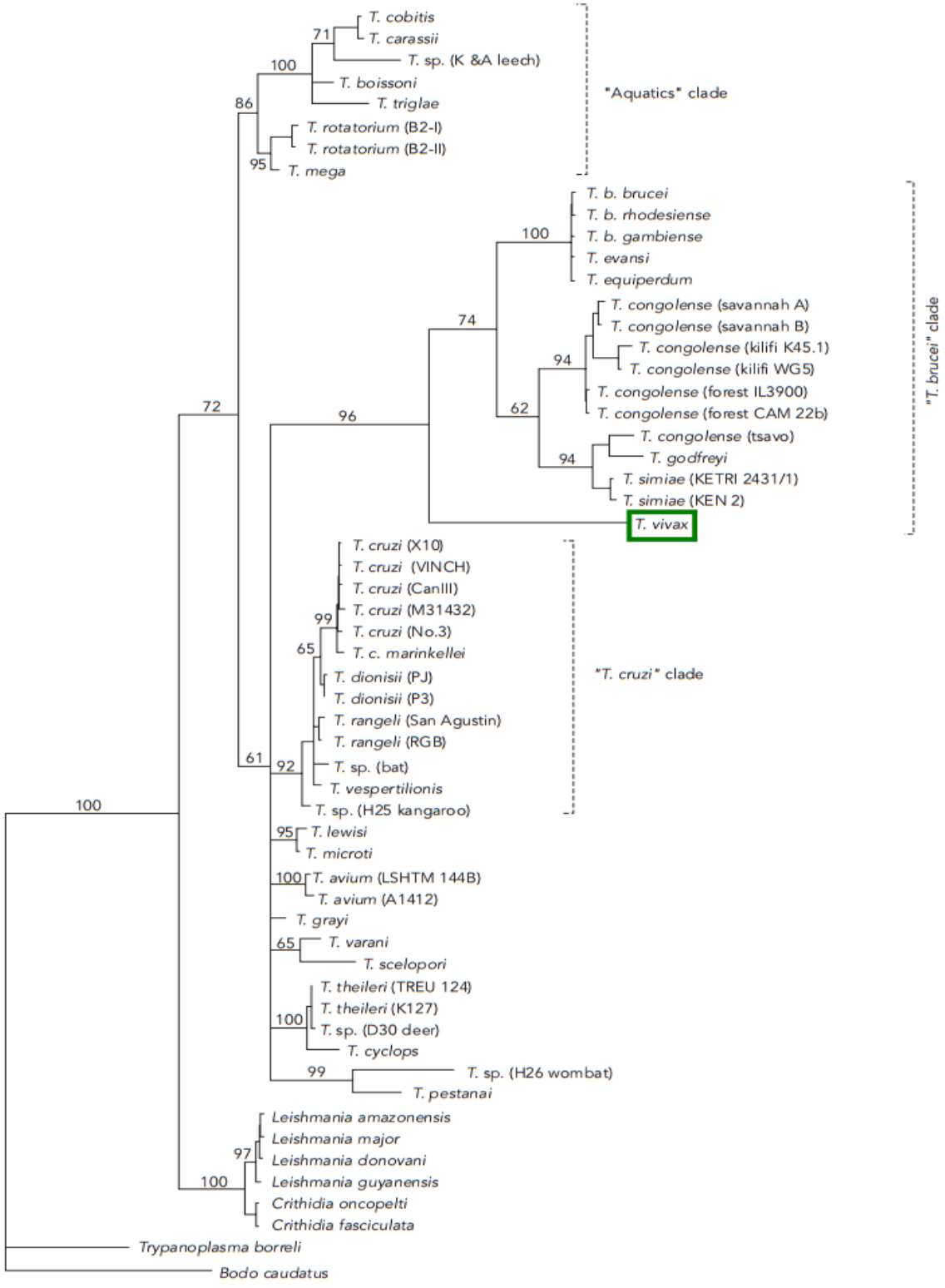

Figura 4.- Árbol filogenético de máxima parsimonia con bootstrap de secuencias SSU rRNA (tomado de Stevens et al. 2001).

Evasión de la respuesta inmune

Los tripanosomas africanos como *T. brucei* y *T. vivax* son exclusivamente extracelulares por lo cual están siempre expuestos al sistema inmune del hospedero mamífero.

Aun así son capaces de mantener una infección prolongada en el tiempo con patrones de parasitemia que muestran picos a intervalos regulares que se reflejan en los síntomas clínicos que presenta el hospedero, particularmente la fiebre. El éxito de estos patógenos depende de su capacidad para evadir la respuesta inmune de su hospedero el mayor tiempo posible y así poder establecer una infección duradera.

Probablemente debido a esta constante interacción de parásito con el sistema inmune es que han evolucionado como mecanismo de defensa la variabilidad antigénica.

Clearance

Un mecanismo utilizado por los tripanosomas para evadir la respuesta inmune del hospedero mamífero consiste en deshacerse de las moléculas del sistema inmune que se hayan unido a las proteínas de superficie y degradarlas, este proceso es conocido como “clearance”.

Mientras que la variación antigénica favorece la persistencia de infecciones a largo plazo, el clearance beneficia a cada organismo individualmente aumentando las posibilidades de evadir la respuesta de sistema inmune del hospedero.

La cubierta de VSG en los tripanosomas africanos consta de una plataforma densamente empaquetada, en donde las moléculas del sistema inmune que se adhieran, como los anticuerpos, aparecen como protuberancias.

Los VSG anclados a través de residuos GPI a la membrana no interaccionan con componentes intracelulares y son libres para desplazarse en el plano de la membrana.

En *T. brucei* se ha observado una estrategia de defensa ante el sistema inmune que combina la constante movilidad celular de estos organismos con las funciones de reciclado de las proteínas de membrana. Ambos procesos se combinan para eliminar de forma continua las moléculas de anticuerpos generadas por el hospedero que se hayan unido a las proteínas de cubierta VSG.

El proceso de eliminación de las moléculas de anticuerpo de las superficie del parásito puede explicar la razón por la cual los tripanosomas han sido seleccionados para mantenerse en movimiento continuo, explicando además el sentido de ese movimiento y el papel que desempeñan la estructura del flagelo y el bolsillo flagelar.

El flagelo en los tripanosomas corre a lo largo del cuerpo e impulsa la célula en dirección opuesta al bolsillo flagelar, ubicado en la base donde comienza el flagelo. Esto provoca que complejos moleculares de gran tamaño, como los del sistema inmune unidos a las proteínas VSG,

se vean arrastrados por fuerzas hidrodinámicas hacia la región donde se encuentra ubicado el bolsillo flagelar que actúa como un pozo donde acaban los complejos VSG-anticuerpo y son internalizados. En este proceso los anticuerpos son transportados a los lisosomas para ser degradados mientras que las proteínas VSG son recicladas y regresan a la superficie celular (Engstler et al. 2007).

Este proceso de clearance es direccional, es decir el desplazamiento de las moléculas del sistema inmune unidas a las VSG tiene el sentido de las fuerzas de arrastre del medio, siendo las moléculas de la región anterior del organismo las primeras en desplazarse hacia la región posterior. Esto ha podido ser observado experimentalmente utilizando anticuerpos IgG específicos contra una variante de VSG de *T. brucei* marcados con un fluoróforo, como puede verse en la figura 5.

Utilizando microscopía de inmunofluorescencia se observa al principio toda la célula del tripanosoma cubierta por el anticuerpo marcado, ya a los 20 segundos se observa que la región hacia donde se apunta el flagelo, o sea hacia donde se desplaza comienza a desplazar los anticuerpos marcados. A los 30-60 segundos la mayoría de los anticuerpos marcados se encuentran solamente en la región posterior, hacia la base del flagelo. Entre los 80-120 segundos se observan todos los anticuerpos marcados en el bolsillo flagelar donde son endocitados y a los 140 segundos solamente se detectan en los lisosomas (Engstler et al. 2007).

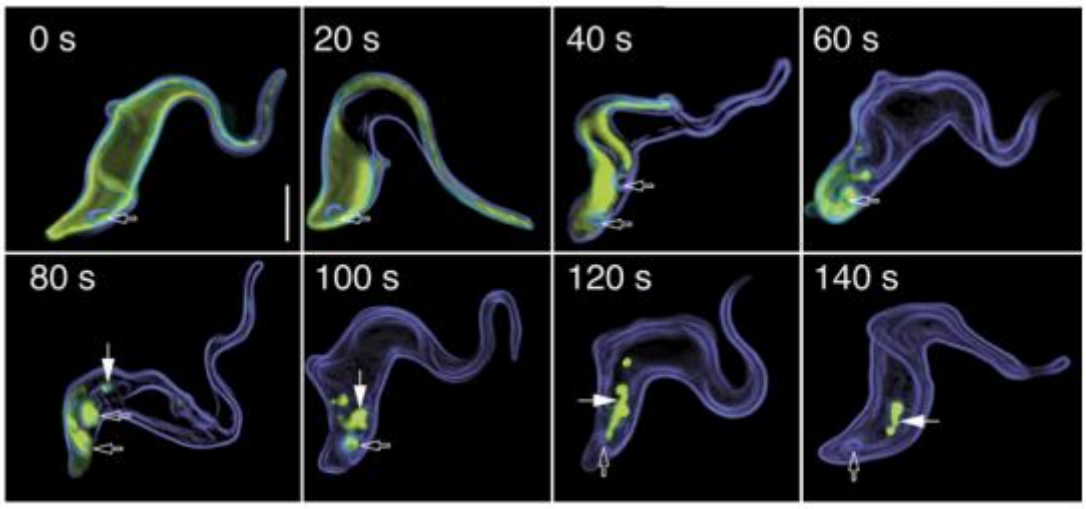

Figura 5.- Imágenes de microscopía de inmunofluorescencia que muestra el proceso en el tiempo de 'clearence' en *T.brucei* de anticuerpos IgG marcados con fluoróforos (Engstler et al. 2007).

Variación antigénica

Los tripanosomas africanos poseen una densa y altamente inmunogénica capa de proteínas VSG que los protege contra la lisis mediada por complemento. Los genes VSG forman una gran familia multigénica de características notables.

Solo uno de los genes de la familia VSG se encuentra activo en cada organismo y por tratarse de poblaciones clonales, en principio en toda la población se está expresando el mismo gen; sin embargo existe un nivel basal de intercambio del VSG activo. La variación antigénica consiste en cambios a nivel poblacional de la expresión de estos genes.

En el hospedero, una vez que han madurado anticuerpos específicos contra la VSG más frecuente en la población, las inmunoglobulinas generadas lisan los tripanosomas que tengan la misma cubierta de glicoproteínas. Sin embargo, debido a los mecanismos de variación antigénica una pequeña población en cada nueva generación de parásitos cambia a una glicoproteína antigénicamente diferente que resulta indetectable para el sistema inmune del hospedero, el cual debe reiniciar el proceso de maduración de la respuesta inmune.

Este cambio de antígenos de superficie se hace evidente al tomar muestras en los picos de parasitemia ya que cambia el gen VSG que está siendo expresado.

En *T. brucei* hay más de mil genes diferentes que codifican para VSG y la expresión secuencial de estos genes produce poblaciones de parásitos antigénicamente diferentes lo que permite la persistencia del parásito en el hospedero mamífero (Barret et al. 2003).

Se trata de un evento evolutivo único y la diversificación de un repertorio de estas dimensiones solo puede ocurrir bajo una fuerte presión selectiva como lo es la respuesta inmune humoral del hospedero vertebrado.

El elevado número de tipos de glicoproteínas antigénicas es lo que dificulta el desarrollo de una vacuna efectiva y permite las reinfecciones.

En *T. brucei* las VSGs están compuestas por un dominio N-terminal hipervariable de 350 a 400 residuos que puede tener una identidad de secuencia muy baja de tan solo 15% y un dominio C-terminal más conservado de 40-80 residuos que está anclado a la membrana celular a través de un enlace glicosil fosfatidil inositol (GPI) (Carrington et al. 1991). A pesar de la divergencia de secuencia, los dominios N-terminales pueden adoptar una estructura de hélice superenrollada formada por hélices alfa que contienen epítopes variables expuestos. Se han observado dominios hipervariables en donde coinciden espacialmente un 60% de los residuos y sin embargo su identidad a nivel de secuencia puede ser de solo un 16% (Blum et al. 1993).

Es de destacar que en el repertorio de genes VSG de *T. brucei*, alrededor del 80 a 90 % se encuentra en las regiones subteloméricas de los cromosomas (Berriman et al. 2005).

Solo una minoría, entre el 5 y 10% de estos genes son funcionales, mientras que los restantes son pseudogenes y forman una fuente importante de variación.

VSG switching

De los más de mil genes VSG que se encuentran en el genoma de *T.brucei*, solo uno se encuentra activo en un momento dado y es transcripto desde una región del genoma denominada sitio de expresión.

Los sitios de expresión (BES, "bloodstream expression sites") en *T. brucei* son alrededor de 20 y se encuentran en los extremos de los cromosomas, en la región telomérica.

Desde estas regiones se transcriben de forma policistrónica genes asociados al sitio de expresión (ESAG, "expression-site associated genes"). Muchos de estos genes codifican para moléculas de superficie o genes que están involucrados en la adaptación al hospedero (Becker et al. 2004).

Se ha sugerido que la presencia de varios sitios de expresión, pueda deberse a los diferentes genes ESAGs presentes en cada uno de ellos ya que algunos de estos genes codifican para diferentes receptores de transferrina, que provee al parásito de hierro. Cambiar entre distintos BES puede ser una adaptación a los diferentes hospederos que infecta (Bitter et al. 1998).

El cambio del VSG que está siendo expresado, puede ocurrir de varias formas diferentes (figura 6). Puede activarse otro BES donde se encuentra otro gen de VSG, proceso llamado *in-situ* switch.; pueden ocurrir rearreglos del DNA, a través de conversión génica de un nuevo VSG al sitio de expresión activo, o intercambio telomérico donde decenas de genes VSG pueden insertarse en un BES activo (Taylor & Rudenko 2006). La forma de activación de VSG por conversión génica duplicativa es la observada con más frecuencia (Robinson et al. 1999).

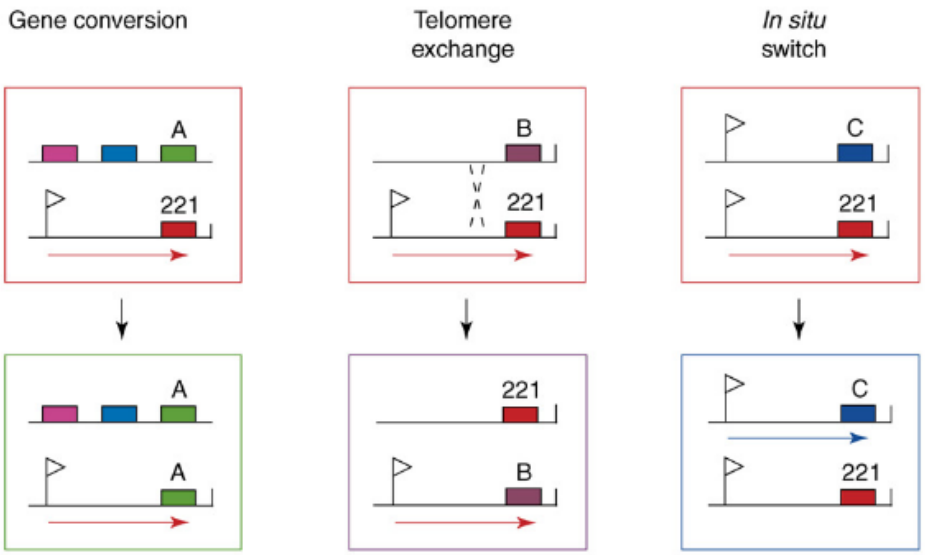

Figura 6.- Mecanismos por los cuales puede cambiarse el VSG activo. Los rectángulos de colores indican los genes VSG, la bandera blanca la región del promotor, y la flecha hacia la derecha indica la dirección de la transcripción. Se indican los pasos que deben darse para que ocurra conversión génica con copia del gen A (en verde) al sitio activo, intercambio telomérico donde el gen B (violeta) pasa al sitio activo y el cambio *in situ* donde se activa un nuevo promotor, silenciando el otro (Taylor & Rudenko 2006).

A pesar de la alta variabilidad de la secuencia de los genes VSG, los rearreglos de DNA (figura 7) utilizan recombinación homóloga ya que en la zona 5' de los genes se encuentran repetidos de 70 pb y en el extremo 3' secuencias conservadas. La conversión génica puede ser segmentada y ocurrir en varios lugares de la región 3' de los genes, lo cual puede dar lugar a genes quiméricos formados por varios pseudogenes de VSG (Morrison et al. 2009).

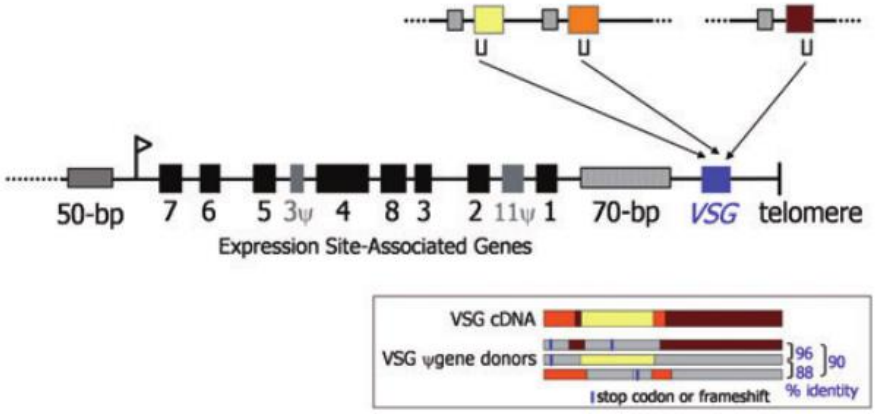

Figura 7.- Formación de un VSG mosaico por conversión génica en segmentos. El gen VSG en azul, está representado dentro de un sitio de expresión, acompañado de ESAGs. Los rectángulos grises de 70 y 50 pb representan secuencias repetidas de este largo, y el promotor por la bandera blanca. La conversión génica de tres pseudogenes de VSG no activos que forman parte del repertorio silencioso se combinan para reemplazar el VSG activo. El cuadro inferior muestra los resultados de dicha conversión mostrando el cDNA final del VSG y los tres pseudogenes que lo originan (Morrison et al. 2009).

Genoma nuclear de tripanosomas

El genoma de los tripanosomas es diploide y la forma de reproducción es por fisión binaria aunque existen evidencias de intercambio génico en *T.brucei*, *T.cruzi* y *Leishmania major* aparentemente la reproducción sexual parece ser muy limitada.

Existe una amplia diferencia en el tamaño de los genomas aunque existe conservación en los genes que comparten un mismo contexto genómico.

Organismo	Cepa	Tamaño genoma haploide (Mb)	Número de CDS
Trypanosoma brucei gambiense	DAL97	22.1	9822
Trypanosoma brucei brucei	927/4	26.1	8712
Trypanosoma vivax	Y486	24.8	11642
Trypanosoma congolense	IL3000	39.2	12927

Tabla 1.- Comparación del tamaño del genoma y número de genes codificantes de proteínas en tripanosomas africanos (datos tomados de NCBI Genome).

Uno de los genomas de tripanosomas africanos mejor estudiados es el de *T. brucei*; donde los cromosomas se clasifican en tres tipos, como se ilustra en la figura 8. El genoma lo componen once cromosomas diploides cuyo tamaño varía entre 0.9 y 6 Mb, los cuales por su tamaño son llamados "cromosomas de mega-bases" y pueden tener una considerable variación entre cepas.

En los cromosomas de mega-bases se encuentran los genes de housekeeping y es posible encontrar sitios de expresión de los VSG en aproximadamente la mitad de los subtelómeros de estos cromosomas (El-Sayed et al. 2000).

Luego en orden de tamaño, están los denominados "cromosomas intermedios" cuya longitud es de entre 200 a 700 Kb y son de ploidía incierta. Se pueden encontrar entre uno y siete de ellos dependiendo de la cepa. Los cromosomas intermedios poseen elementos repetitivos de 177 pb de largo que flanquean las regiones donde se encuentran los sitios de expresión para VSGs (El-Sayed et al. 2000).

Más pequeños aún son los denominados "minicromosomas" que son de entre 30 a 150 Kb de largo, de los cuales se han encontrado unas cien copias en *T. brucei*. Entre el 10 y el 20% del genoma nuclear de *T.bruce*i está compuesto por estos minicromosomas, en los cuales se encuentran repetidos de 177 pb en su región central que pueden llegar a abarcar hasta el 90% de un minicromosoma. En los extremos no cuenta con sitios de expresión de VSG pero si con genes VSG que son el único tipo de gen encontrado. Esto sugiere que su finalidad es la de mantener y expandir la población de estos genes (Wickstead et al. 2004).

Los genes VSG no expresados se encuentran ubicados en la mayoría, si no es que en todos los cromosomas y se estima que conforman un 5% del genoma (El-Sayed et al. 2000).

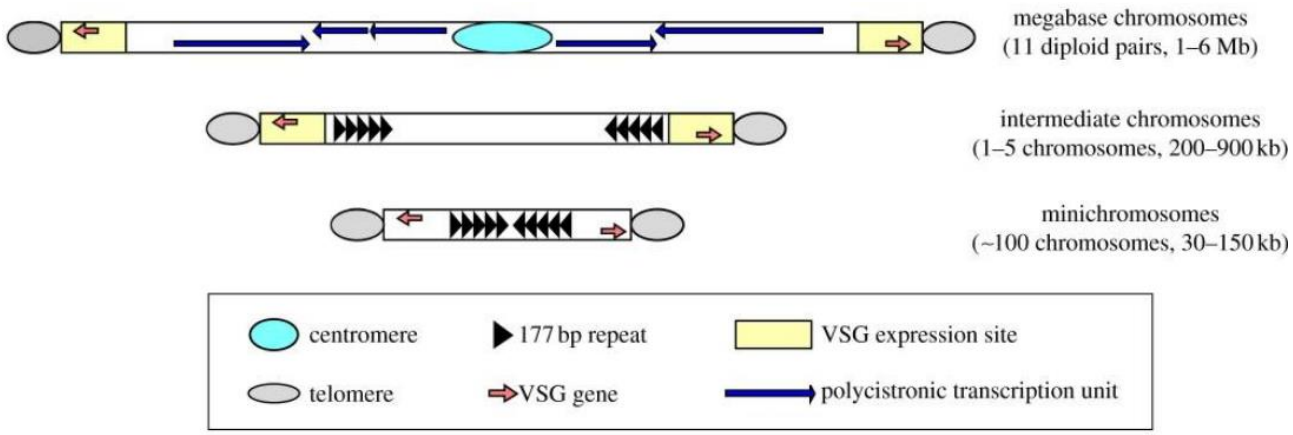

Figura 8.- Ilustración esquemática los tres tipos de cromosomas presentes en *T. brucei* (Akiyoshi & Bull 2013).

En diferentes especies de tripanosomas la mayoría de los genes conservan la posición y el orden a lo largo de los cromosomas indicando así proximidad filogenética que mantiene la sintenia. Sin embargo muchos genes especie-específico, en particular aquellos que forman parte las grandes familias de antígenos de superficie se encuentran en zonas donde no hay mantenimiento de la sintenia, en general en regiones internas de los cromosomas y regiones subteloméricas. Las regiones de quiebre de la sintenia se generan debido a la expansión de familias multigénicas, o son causados por retroelementos o RNA estructurales (El-Sayed et al 2005).

La localización de genes que codifican para proteínas de superficie junto a numerosos retrolementos en regiones subteloméricas podría servir al parásito para aumentar la frecuencia de recombinación y de esta manera generar variabilidad de secuencia rápidamente.

También hay pérdida de la conservación de la sintenia en cercanías a regiones de cambio de hebra lo cual puede que sea un reflejo de elevadas tasas de recombinación en estos sitios. Parece existir una fuerte presión selectiva para mantener el orden génico y los clústeres de genes intactos a pesar de la extensa divergencia de secuencia en los propios genes (Ghedin et al. 2004).

Las regiones subteloméricas de *T. brucei* representan un 20% de su genoma, la mayoría de los genes en esta ubicación son especie-específico y están relacionados con la variación antigénica y la evasión de la respuesta inmune (El-Sayed et al. 2005).

Recientemente se ha observado en *T.cruzi*, que estas grandes regiones con familias de genes especie-específicos no son exclusivamente subteloméricas sino que pueden encontrarse en cualquier región del cromosoma e incluso ocupando cromosomas enteros. De este modo puede considerarse el genoma de los tripanosomas como compartimentalizado, donde se pueden encontrar extensas regiones con genes pertenecientes a un core común a los tripanosomas y otras extensas regiones únicamente con genes especie-específicos (Berna et al. 2018).

Repertorio silencioso de VSG

De los 940 genes de VSG identificados en los cromosomas de mega-base de *T. brucei*, el 65% son pseudogenes y el 21% están incompletos. De los restantes un 9.5% son atípicos en cuanto a inconsistencias en la conservación de dominios de plegamiento o de la secuencia de anclaje a membrana GPI y solo un 4.5% pueden considerarse completamente funcionales (Marcello & Barry 2007).

La variación antigénica en *T. brucei* depende en un repositorio de genes VSG silenciosos compuesto en su mayor parte por pseudogenes que se encuentra en la regiones subteloméricas de la mayoría de los cromosomas.

La variación antigénica permite que se genere una colección de subpoblaciones de parásitos que expresan diferentes proteínas VSG. La expresión de estos genes no ocurre al azar sino que se da en con un orden determinado. Las primeras variantes de VSG expresadas están dictadas por el primer VSG expresado, sin embargo con el transcurso de la infección diferentes factores entran en juego, tales como la densidad de la población de parásitos y las respuestas del sistema inmune del hospedero (Lythgoe et al. 2007).

Se piensa que este orden es importante para mantener una infección prolongada en el tiempo asegurándose que surjan nuevas variantes lentamente pero a su vez limitando el número de estas a niveles subletales para el hospedero (Pays 1989).

Los genes VSG más próximos a los telómeros son los primeros en expresarse y las variantes más tardías expresadas que se han observado son VSGs que surgen como mosaicos de fragmentos de pseudogenes (Thon et al. 1990).

Transcripción

Una característica fundamental de los genomas de tripanosomas es la organización de sus genes en grupos que son transcriptos juntos en unidades denominadas policistrones.

Los pre-mRNA llevan información génica para la síntesis de varias cadenas polipeptídicas y son en cierta manera similar a los operones bacterianos; aunque con diferencias fundamentales, ya que son de mucho mayor tamaño y los genes no suelen estar funcionalmente relacionados entre sí, ni corresponden a una misma ruta o proceso metabólico como sucede en bacterias.

Las secuencias codificantes poseen una peculiar organización a lo largo de un cromosoma. Se ubican formando grupos o clústeres de genes que se encuentran todos en la misma hebra; luego le sigue una región de 1 a 13 Kb no transcripta denominada región de cambio de hebra y luego comienza otro clúster de genes en la hebra complementaria; de esta manera al ser transcriptos los mensajeros radian en ambas direcciones como policistrones desde la región de cambio de hebra (Palenchar et al. 2006).

El producto transcripto es un mRNAs policistrónico con información codificante para diferentes proteínas lo cual indica que la regulación de la expresión génica debe ocurrir a nivel post-transcripcional.

Al igual que en otros eucariotas los mRNAs de tripanosomas poseen modificaciones en los extremos 5' y 3' que resultan fundamentales para resolver mRNAs individuales a partir de mRNAs policistrónicos co-transcriptos.

Para la maduración de los pre-mRNAs policistrónicos en mRNAs traducibles estables los tripanosomas utilizan un cap 5' derivado de transcriptos del gen SL-RNA (Spliced Leader RNA),

presente en un alto número de copias. El gen SL-RNA es una secuencia de 39 nucleótidos que se añade al extremo 5' de los mRNAs maduros a modo de capping. En *T. brucei* los genes de SL-RNA forman clústeres en tándem y se encuentran unas 200 copias del gen.

El proceso por el cual se resuelve un pre-mRNA policistrónico en mensajeros maduros con la secuencia SL como cap 5' es un evento de trans-splicing.

En tripanosomas en el espacio intergénico de un transcripto policistrónico tiene lugar la poliadenilación en 3' del primer gen luego de que el segundo de ese pre-mRNA recibe un cap 5' con el SL-RNA. Debido a esta sucesión de eventos de procesado una única ronda de transcripción de la RNA polimerasa II (RNAP II) permite producir numerosos mRNAs funcionales.

Con ello los tripanosomas evitan el agregado de caps 5' de forma co-transcripcional como en otros eucariotas lo que lleva a una actividad pausada de la RNAP II y el reclutamiento de otras enzimas (Palenchar et al. 2006).

La actividad transcripcional y estructura de las RNA polimerasas en tripanosomas son similares a sus homólogas eucariotas, aunque aún no se han encontrado muchos de los factores basales de transcripción. En *T. brucei* la RNA polimerasa I transcribe los pre-rRNA del clúster de los genes ribosomales 18S, 5.8S y 28S, además de mRNAs de regiones llamadas "expression sites", donde se encuentran las VSG y los ESAGs (Günzl et al. 2003).

La RNAP II transcribe mRNAs así como el gen SL. La RNA polimerasa III transcribe tRNAs, 5S RNA y snRNAs (Douris et al 2010).

La transcripción policistrónica y la ausencia de promotores de RNAP II clásicos encontrados en eucariotas sugiere que el inicio de la transcripción no es un factor limitante en la producción de mRNAs y ocurre expresión constitutiva. Esto resulta particular por el enorme gasto energético que debe realizarse para mantener la totalidad de la población de transcriptos a un nivel de expresión constante; aunque en realidad puede resultar viable para un organismo parásito que no tiene demasiados problemas en obtener recursos y energía. Además es posible que mantener una transcripción basal sea útil en organismos que experimentan cambios bruscos de ambiente al cambiar de hospedero en distintas etapas de su ciclo de vida (Palenchar et al. 2006). Esto implica que un factor inherente a la supervivencia del parásito incluye la replicación exitosa en ambientes de características y disponibilidad de recursos muy dispares, la habilidad de adaptarse rápidamente al nuevo hospedero durante su ciclo de vida le otorga una cierta autonomía al parásito.

Un aspecto interesante del trans-splicing de SL es su distribución filogénetica, ya que ha sido descrito en grupos tan diversos como *Euglenozoa* (euglenoides y tripanosomas), dinoflagelados, cnidarios hidrozoarios, nemátodos, platelmintos, rotíferos bdelloideos, ctenóforos, quetognatos, urocordados y crustáceos (anfípodos y copépodos). Esto sugiere dos escenarios posibles, o el trans-splicing de SL es un mecanismo antiguo perdido de forma independiente en múltiples linajes o ha

evolucionado en múltiples ocasiones en eucariotas. La hipótesis de que el trans-splicing haya surgido como múltiples eventos independientes requiere que el mecanismo pueda evolucionar de una forma relativamente sencilla y esto es posible dado que utiliza la misma maquinaria que para cis-splicing solo que con el precursor SL en lugar de un snRNP (Douris et al 2010).

Regulación de la expresión génica

En la mayoría de los organismos la forma más frecuente de regulación de la expresión génica es al inicio de la transcripción. Sin embargo a pesar que casi la totalidad de los genes codificantes de proteínas son transcriptos por la RNAP II, no se han encontrado en kinetoplástidos elementos que regulen la actividad de la RNAP II lo cual sugiere que la regulación de la expresión génica debe estar determinado por mecanismos post-transcripcionales (Clayton & Shapira 2007).

En tripanosomas africanos solamente los genes que codifican para las proteínas de superficie como prociclinas o proteínas VSG, que son transcriptos por la RNAP I regulan su transcripción de manera estricta.

Se han planteado varios escenarios alternativos presentes también en otros eucariotas que podrían contribuir a la regulación de la expresión génica post-transcripcional.

En *T.brucei* más de la mitad de los genes tienen varios sitios aceptores alternativos para el trans-splicing de SL en 5' y de adenilación en 3'. Esto indica que los mRNAs de *T. brucei* existen en diferentes isoformas, lo cual puede indicar una muy poca presión evolutiva para la precisión de estos eventos de trans-splicing. Sin embargo también se han observado algunos genes en los cuales un sitio de trans-splicing puede truncar el extremo 5' del mRNA y generar proteínas con ubicaciones intracelulares diferentes (Kramer 2011).

Otro factor que influye en la expresión génica son las tasas de decaimiento de mRNA. En algunos casos estudiados en tripanosomas la vida media de los mRNA está asociada a la secuencia del 3' UTR que permite la unión de factores que controlan su estabilidad (Kramer & Carrington 2011). Al igual que en otros eucariotas, en tripanosomas comienza por la eliminación de nucleótidos de la cola polyA y además el proceso puede acelerarse por eliminación del cap 5' por degradación con exonucleasas encontradas en *T. brucei* (Fadda et al. 2014).

También puede ocurrir regulación a nivel traduccional y existen numerosos mecanismos presentes en eucariotas los cuales operan en su mayoría al inicio de la traducción.

Una forma de control indiscriminada de la población total de mRNAs es la fosforilación de los factores de iniciación de la traducción, en particular del factor eIF2-alpha que inhibe la activación de tRNAs lo cual ha sido observado en *Lehismania* y *T. cruzi*.

Si bien no se ha detectado en kinetoplástidos, las regiones 5'UTR de los mRNAs podrían contener formas de regular el inicio de la traducción de diferentes maneras. La estructura secundaria en el 5'UTR podría inhibir el ensamblaje, detener escaneo del ribosoma o secuencias AUGs en estas regiones pueden afectar el inicio de la traducción disminuyendo su eficiencia (Clayton 2014).

Además en tripanosomas, al igual que en el resto de los eucariotas, se pueden encontrar gránulos de RNA, como P-bodies y gránulos de estrés. Estos gránulos formados por RNA y proteínas están involucrados en el secuestro de mRNAs ya sea para su posterior degradación, o como almacenamiento temporal de mensajeros con el fin de regular la traducción ante situaciones de estrés celular. En tripanosomas son de gran importancia debido a los cambios ambientales repentinos en los cuales pueden encontrarse, como cuando pasan de un mamífero a un insecto, o viceversa. Por lo cual disponer de transcriptos los mRNA codificantes de proteínas listos para ser usados en una nueva etapa del ciclo de vida, les permite responder a nuevas demandas en forma casi inmediata (Fritz et al, 2015).

Genoma mitocondrial de tripanosomas

El DNA mitocondrial de los tripanosomátidos, llamado también kDNA (DNA kinetoplástido) presenta características únicas en su estructura y funcionamiento. El kDNA forma una enorme red compuesta de miles de secuencias de DNA circular topológicamente entrelazadas formando una estructura similar a una red o malla.

Las moléculas de kDNA pueden ser de dos tipos: los maxicírculos de 20 - 40 kbp, que se encuentran en algunas decenas y codifican algunos rRNAs y proteínas mitocondriales, similar al genoma mitocondrial de otros eucariotas; y los minicírculos de 500 bp a 3 kbp que pueden ser varias decenas de miles

En el maxicírculo es similar al genoma mitocondrial de otros eucariotas, en este se encuentran genes que codifican para las unidades 9S y 12S de rRNA, varias subunidades de la NADH dehidrogenasa (ND1-5, ND8 y ND9), la subunidad 6 de la ATPasa (A6), las subunidades de la citocromo oxidasa (COI-III), la subunidad b de la citocromo reductasa (CYB), y varios marcos abiertos de lectura aún no identificados.

Los transcriptos de algunos de estos genes mitocondriales pueden ser traducidos directamente, pero los transcriptos de la mayoría deben ser editados, en un proceso post-transcripcional de inserción y deleción de uridinas para poder terminar en un mensajero traducible (Shaw et al. 1988).

Los minicírculos conforman más del 90% de la masa del kDNA y en la mayoría de las especies conservan el mismo tamaño. En *T. brucei* que realiza una extensa edición de transcriptos se han identificado unas 250 clases de minicírculos (Shapiro & Englund 1995).

Los minicírculos codifican para RNA guías (gRNAs) que participan en el proceso de edición de los transcriptos del maxicírculo (Blum et al 1990). Estos son de 70 nts de largo y presentan una corta cola poly(U) no codificada. Los gRNAs se encargan de la especificidad del editado (guían) de los transcriptos del maxicírculo. Aparte de los gRNAs, los minicírculos contienen una región de 100 - 180 nts que forman tres bloques con diferentes grados de conservación donde se encuentra el origen de replicación (Blum et al 1990).

Edición de mRNAs mitocondriales

El proceso de edición (figura 9) comienza con el clivaje endonucleolítico del pre-mRNA en el sitio determinado por la interacción entre un gRNA y su mRNA cognado. La región 5' de un gRNA puede formar un dúplex con la secuencia que reconoce de un mRNA en una región 3' a la posición a ser editada. Los gRNAs tienen una cola oligo(U) agregada post-transcripcionalmente que facilita la interacción con el 5' del pre-mRNA. Cientos de diferentes secuencias de gRNAs cuando son pareadas con su mRNA presentan diversas secuencias nucleotídicas a las endonucleasas del editosoma lo que implica que el reconocimiento del sitio de clivaje por las endonucleasas es complicado y se sugiere que características estructurales de la interacción gRNA-mRNA es lo que reconocen las endonucleasas. El clivaje del pre-mRNA ocurre en un nucleótido no pareado corriente arriba de la posición del sitio de anclaje gRNA-mRNA y se da en dirección 3' a 5'. Todo el proceso también avanza 3’a 5’, y la edición de un segmento del pre-mRNA genera el sitio (secuencia) de reconocimiento para el siguiente gRNA. La edición de eliminación es realizada por exonucleasas U específicas o exoUasa que eliminan nucleótidos U no pareados con el gRNA luego del sitio de clivaje; mientras que en la edición de inserción las Us son añadidas por una uridil transferasa o TUTasa (Stuart et al. 2005).

Una sola RNA polimerasa mitocondrial es necesaria para la síntesis de los RNA de los maxi y minicírculos. Los minicírculos son transcriptos como policistrones y clivados por el complejo de proteínas 20S denominado "editosoma 20S" en uno o más gRNAs. Los maxicírculos también son transcriptos como policistrones y el proceso de edición mediado por los gRNAs ocurre independientemente del proceso de clivado en transcriptos monocistrónicos (Lukes 2010).

Se han sugerido varias hipótesis sobre las posibles ventajas evolutivas del editado de RNA, como puede ser un nivel extra de regulación de la expresión génica mitocondrial, la corrección de mutaciones acumuladas sobre una mitocondria no funcional, la evolución acelerada aumentando la variación genética o la posibilidad de obtener varios productos proteicos con un solo gen. La persistencia de la edición de algunos transcriptos en el estadio sanguíneo a pesar de no ser

requeridos en esta etapa contribuye a la idea de que el editado de RNA podría ser usado como una fuente para generar diversidad proteica (Lukes 2010).

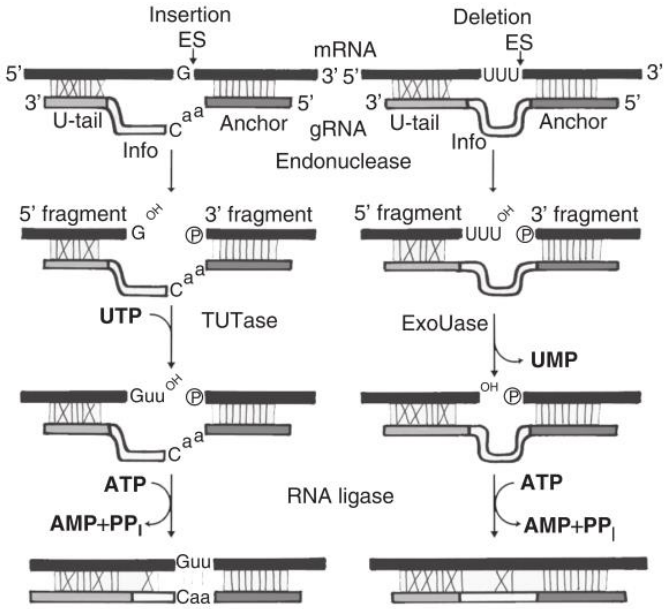

Figura 9.- Mecanismo de editado de RNA. Los mismatches entre el mRNA y el gRNA definen el sitio de edición, que es clivado por una endonucleasa. Luego una TUTasa añade Us o una ExoUasa las elimina. Finalmente una RNA ligasa une los dos fragmentos de mRNA (Lukes 2010).

Objetivos

El objetivo del presente trabajo es realizar un estudio de genómica evolutiva y comparativa de diferentes cepas del tripanosoma africano *T.vivax*. El trabajo se centrará en la comparación del genoma de las cepas que se han expandido al continente americano y las encontradas en África enfocándose en el estudio de la región genómica con genes asociados a los mecanismos de variabilidad antigénica.

Es de particular interés estudiar si han ocurrido cambios en el genoma de las cepas americanas de *T.vivax* como adaptación al continente americano donde la transmisión del parásito es solamente mecánica.

- Con datos de secuenciado genómico de dos cepas americanas de *T. vivax,* Liem y MT1 se ensamblará el genoma nuclear y mitocondrial de ambas.
- Se identificarán las regiones codificantes utilizando herramientas de software y se realizará la anotación consultando bases de datos de genes y motivos proteicos.
- Luego de ensamblar y anotar las cepas americanas de *T. vivax,* se harán análisis de genómica comparativa con la cepa africana de *T. vivax* Y486 disponible en el Genbank.
- Se explorará el espacio genómico de las cepas americanas y africanas de *T. vivax* caracterizando las regiones genómicas utilizando métodos estadísticos que tengan en cuenta las frecuencias nucleotídicas.
- Se realizarán estudios de genómica comparativa entre las cepas americanas y la cepa africana, identificando a nivel del genoma nuclear y mitocondrial las características propias de cada una y las eventuales adaptaciones de las cepas americanas al continente americano.
- Con los resultados obtenidos en los análisis anteriores se construirá una base de datos y se utilizará como interfaz de acceso una página web interactiva que facilite la exploración de los datos y los resultados obtenidos.
- Se estudiarán en detalle los cambios que puedan observarse en el genoma mitocondrial como resultado de la adaptación a la transferencia mecánica en las cepas de *T.vivax* americanas.

Materiales y métodos

Software

Blast

Blast (Basic Local Alignment Search Tool) (Altschul et al. 1990) es un software heurístico de alineamiento local de secuencias que utiliza el algoritmo de Smith-Watterman. La estrategia que utiliza consiste en dividir la secuencia ("query") que será utilizada en el mapeo en segmentos de menor tamaño denominados "seeds", que se buscan en la base de datos de referencia teniendo en cuenta como alinean estos segmentos y a que distancia se encuentran entre sí.

El siguiente paso consiste en extender estos segmentos utilizando el algoritmo Smith-Waterman de programación dinámica. El alineamiento se extiende a partir de los "seeds" y en cada paso de extensión se evalúa el alineamiento calculando un puntaje a partir de matrices de sustitución que han sido calculadas en base a la probabilidad de la frecuencia de observar sustituciones. Cuando el puntaje para un alineamiento decrece hasta cierto valor umbral que representa el mínimo aceptable, se detiene la etapa de extensión.

El último paso del programa consiste en evaluar cada uno de los alineamientos calculando valores de significancia estadística teniendo en cuenta la probabilidad de que el alineamiento haya sido obtenido por azar, esto es calculado con el parámetro e-value. Al final se reportan solamente los alineamientos cuyo valor de e-value sea menor a cierto valor umbral.

Existen varios programas de alineamiento dentro de la familia Blast, como lo son blastn que permite alinear secuencias de nucleótidos entre sí, blastp para alinear secuencias de proteínas, blastx traducir una secuencia de nucleótidos en todos sus marcos y alinearla contra una base de datos de proteínas, tblastn que compara una secuencia de proteínas contra una base de nucleótidos que son traducidos en sus seis marcos y otras combinaciones.

Una variante de Blast utilizada en la identificación de gRNAs y sus sitios de interacción en el mRNA, es el programa wu-blast desarrollado por la Universidad de Washington que permite estipular puntajes entre los distintos tipos de apareamientos entre nucleótidos. La ventaja de esta flexibilidad es que se puede tener en consideración la posibilidad de alinear secuencias de RNA en donde las interacciones de las base G-U también son aceptadas.

Bowtie2

Bowtie2 (Langmead & Salzberg, 2012) es un software de alineamiento de secuencias cortas, generalmente de unas pocas decenas o centenares de bases con secuencias del orden de mega o gigabases. Fue desarrollado con el objetivo de alinear rápidamente reads producto de secuenciados a genomas. La estrategia que utiliza consiste en generar con las secuencias de referencia un índice FM (Ferragina & Manzini, 2000) que comprime subcadenas de texto completo basándose en la transformada de Burrows-Wheeler y cuenta con la ventaja de que permite búsquedas rápidas de subcadenas informando el número de veces que se encuentra un patrón de texto y sus coordenadas.

Los pasos que se emplean para generar este índice se muestran a continuación (tomado de Langmead 2013). Se comienza aplicando la transformada de Burrows-Wheeler a un texto, para lo cual se debe generar una matriz de todas las rotaciones de caracteres del texto de entrada. Para hacer esto se forman distintas cadenas de texto en donde se toma el carácter del extremo final y se lo coloca al principio hasta completar todas las rotaciones posibles. Se utiliza el carácter $ para denotar el final de la cadena de texto.

Luego se ordenan estas rotaciones de forma lexicográfica y se toma el último carácter de cada una de ellas, como muestra la columna en rojo de la figura 10. Esto es similar a obtener el primer carácter de un árbol de sufijos. Esta transformación es comprimible e indexable porque agrupa ocurrencias del mismo carácter y es fácilmente reversible para la obtención del texto original.

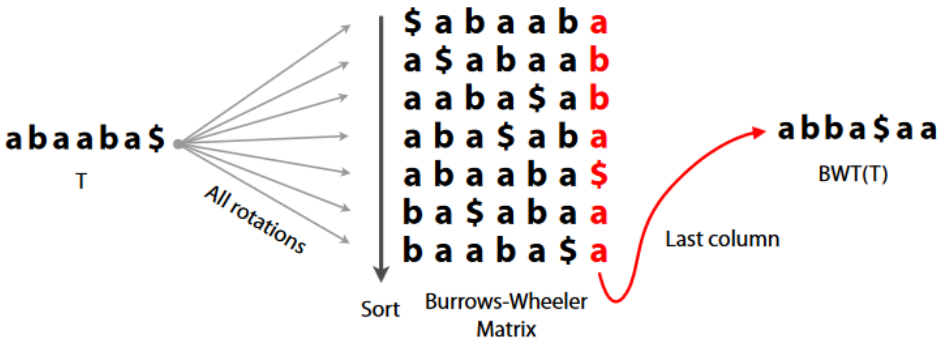

Figura 10.- Partiendo de la cadena de texto 'abaaba' se crea la matriz de Burrows-Wheeler generando todas las rotaciones posibles ordenadas. Tomando la última columna se obtiene el resultado de aplicar la transformada de Burrows-Wheeler (BWT).

Para recuperar el texto original es posible aplicar la propiedad "LF mapping". Previo a realizar las rotaciones se debe asignar un índice a cada carácter según su ocurrencia en la cadena de texto, (por ejemplo en la cadena anterior al añadir el índice quedaría como $a_0b_0a_1a_2b_1a_3$) de este modo solo es necesaria la información de la primera y la última columna de la matriz (columnas F y L). Con esta información ya es posible reconstruir la cadena original, sin embargo para facilitar el indexado

y la compresión de datos es conveniente asignar a cada carácter de la columna F un índice con valores ascendentes lo que se muestra en la figura 11.

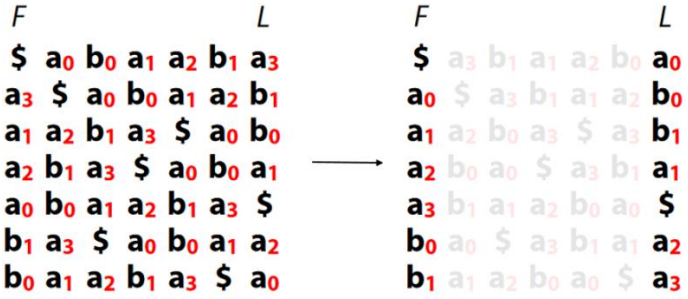

Figura 11.- A la izquierda, asignación de valor de posición en cadena de texto original (números en rojo) a la matriz. A la derecha reasignación de rangos con el fin de ordenarlos de manera ascendente. El objetivo de los rangos es poder recuperar el texto original utilizando "LF mapping".

Al aplicar "LF mapping" la cadena de texto original se reconstruye de derecha a izquierda. Se debe comenzar con la columna F de la primera fila que siempre contiene el carácter de fin de cadena $, luego se interroga la columna L de dicha fila, que va a contener el último carácter de la cadena (en este caso a_0). A continuación se debe buscar a_0 en la columna F e interrogar nuevamente la columna L, (en este caso b_0). Procediendo de esta manera de forma sucesiva, se reconstruye el texto hasta encontrarse con $ en la columna L lo que indica que se ha recuperado el texto completo. Los pasos a seguir están indicados con flechas en la figura 12.

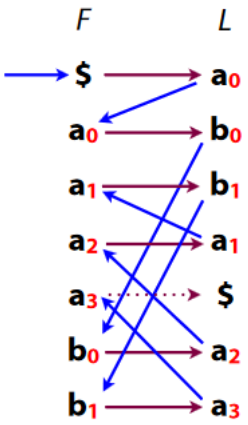

Figura 12.- Reconstrucción del texto de entrada resuelto por LF mapping.

El índice FM genera una estructura de datos con los elementos F y L de la matriz de Burrows-Wheeler, y datos auxiliares para facilitar la búsqueda de subcadenas. Con algunas transformaciones la columna F puede ser representada con un entero por carácter del alfabeto y la columna L es comprimible, por lo cual es potencialmente muy económico en términos del espacio en memoria que ocupa.

Para realizar una búsqueda en esta estructura de datos se deben recorrer las columnas utilizando “LF mapping”. Por ejemplo para encontrar la subcadena “aba”, comienzo buscando en la columna F donde se encuentran todas las “a”, luego en la columna L verifico cuales son precedidas por “b”;, luego cuales de estas “b” se encuentran precedidas por “a”. De esta manera reconstruyo la subcadena “aba” (figura 13).

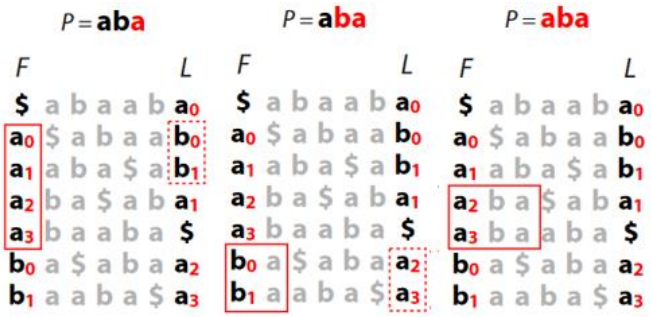

Figura 13.- Búsqueda de la subcadena “aba” utilizando “LF mapping”.

Este tipo de búsqueda no devuelve la posición en el texto original de la subcadena buscada y además es de una complejidad computacional de O(m), siendo m el número de enteros de los rangos de la matriz. Una forma de acelerar este proceso es pre-calcular una estructura de datos con los valores sobre cuantos caracteres preceden a cada carácter en la columna L, lo que lo transforma en un problema de complejidad O(1). Sin embargo esto requiere demasiada memoria, por lo cual se calculan solo “checkpoints”, y se resuelven los valores fuera de estos puntos.

Para encontrar la posición de la subcadena en el texto se requieren “suffix arrays”, que indican la posición en donde comienza cada subcadena posible. Debido a que almacenar esto requiere mucha memoria se almacenan solamente algunas entradas del “suffix array”, y las faltantes pueden ser obtenidas cruzando información realizando “LF mapping” y contando el número de pasos entre checkpoints.

Blast2GO

Blast2GO (Götz et al. 2008) es un software que facilita el proceso de anotación de secuencias. El programa toma como datos de entrada una secuencia de proteínas y las mapea contra las base de datos nr del GenBank o en su defecto un archivo de alineamientos de proteínas realizado con Blast en formato de salida XML. La descripción del mejor hit de cada secuencia es utilizada para realizar búsquedas contra las base de datos del proyecto Gene Ontology (GO). Además el programa permite mapear las secuencias contra bases de datos de motivos proteicos y sitios funcionales como ProDom, PRINTS, Pfam, SMART, TIGR, PROSITE, PANTHER, SUPERFAMILY, etc.

Otros programas integrados permiten hacer predicciones *de novo* de motivos como péptidos señal, utilizando el software SignalP y dominios transmembrana con TMHMM.

El modo en el cual Blast2GO realiza la anotación consta de tres pasos, descritos a continuación (Conesa & Götz, 2008). Primero se hace un mapeo de las secuencias de interés utilizando BLAST contra una base de datos de secuencias, que por defecto es la base de datos pública de proteínas, nr del NCBI, aunque puede configurarse para que utilice un base de datos local. También se pueden controlar parámetros del programa BLAST como el e-value, la longitud del HSP y el número de hits.

Blast2GO parsea la salida de los mapeos y muestra la información de modo tabulado, siendo de gran interés el texto que describe la secuencia, evitándose términos de contenido poco informativo.

En una segunda etapa se realiza un mapeo que consiste en obtener los términos GO asociados con los hits obtenidos del mapeo de Blast. Se hace tres mapeos diferentes, primero se utilizan los "accession numbers" para obtener información sobre el nombre del gen provisto por NCBI, el cual es usado para buscar el producto del gen en tablas de la base de datos GO. Segundo, los identificadores "GI" se usan para obtener IDs de UniProt haciendo uso de archivos de mapeos contra otras bases de datos de genes como Swiss-Prot, TrEMBL, RefSeq, GenPept, y PDB. Por último también se usan los "accession" para realizar búsquedas directas en la tabla de DBXRef de la base de datos GO.

El último paso consiste en la anotación, en donde se asignan términos funcionales a las secuencias de entrada de un pool de términos GO obtenidos en el paso de anterior. El algoritmo de anotación de Blast2GO tiene en consideración la similitud entre las secuencias de entrada y los hits, la calidad de la fuente de la asignación GO, y además se computa un score de anotación. Este score se compone de dos términos, el primero considera el más alto valor de similitud entre las secuencias que comparte el término GO, pesado por un factor que tiene en cuenta su código de evidencia, el cual está presente en cada anotación de la base de datos GO, y puede ser evidencia experimental, inferida por ensayos directos o asignación automática no supervisada. El segundo término introduce

la posibilidad de abstracción al algoritmo. Esta es definida como la anotación a un nodo padre cuando varios nodos hijos están presentes en el pool de términos GO candidatos.

Los parámetros por defecto de Blast2GO fueron elegidos para optimizar la relación entre la cobertura de la anotación y su precisión.

Gene Ontology organiza los productos génicos asociándolos a un glosario semántico de una serie de términos agrupados de forma jerárquica.

Las ontologías fueron desarrolladas considerando una célula eucariota genérica, estructuras especializadas no están representadas. Los términos GO están conectados en nodos de una red, por lo cual las conexiones entre los padres e hijos son conocidas, y forman grafos acíclicos dirigidos.

La información está organizada en tres bases de datos, "Cellular Component", "Molecular Function" y "Biological Process"; cada una con su propia jerarquía de término en forma de árbol.

Las bases de datos presentan las siguientes características: (Ashburner et al. 2000)

"Biological Process" hace referencia al objetivo biológico al cual un gen o producto génico contribuye. Los procesos son llevados a cabo a través de uno o más conjuntos ordenados de funciones moleculares e involucran transformaciones físicas o químicas. Ejemplos de procesos biológicos de alto nivel pueden ser "crecimiento celular y mantenimiento" o "transducción de señales". Ejemplos de términos de procesos más específicos, o sea de más bajo nivel, pueden ser "traducción" o "metabolismo de la pirimidina".

"Molecular Function" es definida como la actividad bioquímica de un producto génico. Esta definición también aplica a la capacidad potencial del producto de un gen. Describe solo lo que hace, sin especificar cuándo o dónde. Ejemplos de alto nivel de estos términos pueden ser "enzima", "transportador" o "ligando". Ejemplos más específicos podrían ser "adenilato ciclasa" o "ligando de receptor Toll".

"Cellular Component" hace referencia al lugar en la célula donde el producto génico es activo. Estos términos reflejan el conocimiento de la estructura celular eucariota. Incluye términos como "ribosoma" o "proteosoma" especificando donde el producto génico puede ser encontrado.

Emboss

Emboss (European Molecular Biology Open Source Software) (Rice et al. 2000) es una suite de herramientas para el análisis de secuencias.

De esta suite se utilizó el programa getORF que permite obtener una colección de los ORFs pertenecientes a una secuencia cumpliendo una serie de requisitos definidos como ser su largo mínimo, si debe comenzar con metionina y finalizar en stop, o entre dos stops, etc.

Otro programa utilizado de esta suite es etandem que permite definir en una secuencia las regiones repetidas utilizando como parámetros de entrada el tamaño mínimo del repetido y el número de veces que se encuentra presente para reportarlo como tal. La salida de este programa es una tabla con las coordenadas que indican donde se encuentran las secuencias repetidas que cumplen con los parámetros predefinidos. Este reporte fue utilizado como paso previo al análisis de frecuencias de trinucleótidos para descartar aquellas secuencias que tenían más del 75 % de repetidos en su secuencia.

Samtools

Samtools (Li et al 2009) se trata de una suite de herramientas de software para la manipulación de datos de alineamiento masivo de secuencias en formato SAM. Varios programas dedicados al análisis y visualización de alineamientos requiere que los datos se encuentren ordenados por coordenadas y en formato binario, aquí es donde los programas sort y view son utilizados para ordenar y convertir el formato de archivo de texto ASCII SAM al formato binario BAM. Otra herramienta de Samtools utilizada en este trabajo fue mpileup que toma como entrada un archivo BAM ordenada y devuelve un archivo TSV (tab separated value) donde cada línea representa un nucleótido de la secuencia de referencia e informa sobre la coordenada, la base de referencia, cuales son las bases de los reads que mapean en esa posición y cuál fue la calidad de secuenciado de estas bases; la información del mapeo se encuentra representada en un formato de texto comprimido denominado CIGAR. Esto es de utilidad para calcular la profundidad y cobertura de los mapeos, la presencia de polimorfismos y en general analizar en detalle las regiones donde los alineamientos resultaron conflictivos.

Otras herramientas utilizadas

Otro software utilizado con asiduidad para desarrollar scripts fue el lenguaje Python y en particular el módulo BioPython (Cock et al. 2009) que implementa métodos que facilitan la manipulación de datos biológicos, ya sean secuencias, filogenias, alineamientos, etc. Para los análisis estadísticos se utilizó el lenguaje R junto con la librería BioStrings (Pages et al. 2017) para el procesamiento de secuencias biológicas, la función pricomp para los análisis estadísticos multivariados y la librería ggplot2 para la generación de gráficas.

Las secuencias repetitivas se comprobaron con la aplicación online Yass (Noe & Kucherov, 2005) desarrollada por la Universidad de Lille. Yass toma como entrada dos secuencias, realiza un Blast de ambas y genera un dotplot con los resultados.

Ensambladores

Uno de los principales objetivos y uno de las etapas iniciales de los proyecto de genómica involucra obtener la secuencia completa del genoma de un organismo. Para ello luego de secuenciar el DNA se deben utilizar herramientas de software que implementan algoritmos de ensamblado.

Diferentes tecnologías de secuenciado resultan en lecturas de tamaños y profundidades muy dispares que requieren de algoritmos diferentes.

Actualmente se utilizan dos estrategias diferentes para realizar el ensamblado de secuencias, dependiendo del tipo y volumen de reads que generan los secuenciadores. Cuando se trata de algunas decenas de miles de reads con una longitud de varios centenares de bases, como los producidos en secuenciados con tecnologías Sanger, 454 o más recientemente PacBio, se suelen utilizar estrategias de ensamblado basadas en algoritmos de tipo Overlay-Layout-Consensus (OLC).

Con el surgimiento de la tecnología Solexa (Illumina) que produce millones de reads de decenas de bases de largo, se requirió de otro tipo de aproximación al problema de ensamblado y se comenzaron a utilizar algoritmos que implementan grafos de de-Bruijn los cuales son más adecuados para manejar grandes volúmenes de datos con poca longitud de secuenciado.

Algoritmos Overlay-Layout-Consensus

Los algoritmos OLC en general constan de tres etapas, el primer paso consiste en encontrar los solapamientos entre reads, esta es la parte más lenta del ensamblado, dado que debe crear índices de los reads y buscar las subcadenas compartidos por los reads para así construir el grafo de solapamientos. El siguiente paso es generar el layout, esto consiste en unir los segmentos resueltos por los grafos de solapamiento para formar contigs. Los grafos de solapamiento no suelen ser lineales sino que por el contrario existe redundancia, segmentos no resueltos como sucede en secuencias con regiones repetidas, o secuencias que aportan al grafo información no resuelta. Por eso la etapa de generar el layout del grafo consiste en eliminar las aristas que pueden inferirse transitivamente, dividir el grafo en los segmentos donde se generan ramificaciones, eliminar subgrafos producto de errores de secuenciado, eliminar las burbujas en los grafos causadas por SNPs heterocigotas o errores de secuenciado, etc.

En la última etapa se genera la secuencia consenso para un contig tomando todos los reads que componen un contig, alineándolos y aplicando un consenso de secuencias por mayoría y calidad de las secuencias de los reads.

En general los ensambladores que implementan algoritmos OLC utilizan una estrategia mixta de suffix trees para filtrar la mayoría de los pares no solapantes y luego programación dinámica para los reads que solapan y así permitir gaps y mismatches.

La principal desventaja de los algoritmos OLC radica en que construir el grafo de solapamiento es un paso muy lento que requiere de mucha memoria RAM ya que cada nodo es un read y el número de aristas tiene un crecimiento superlineal y los tiempos computacionales son O(N) para los suffix trees y O(N^2) para la etapa de programación dinámica.

Entre los ensambladores que implementan algoritmos OLC se encuentran Celera Assembler (Myers et al. 2000), Newbler (Miller 2010), Phrap (Green 1996) y Mira (Chevreux et al. 1999), entre otros.

Algoritmos de-Bruijn

La estrategia de ensamblado utilizando grafos de de-Bruijn fue originalmente propuesta con la finalidad de resolver de mejor manera el ensamblado de regiones repetitivas. Fue desarrollada por Pevzner et al. en 2001, con el desarrollo del ensamblador EULER.

La mayor dificultad que enfrentan los algoritmos OLC es encontrar el camino correcto en un grafo de solapamiento. Este problema es particularmente difícil utilizando una aproximación OLC la cual requiere visitar cada nodo una sola vez y encontrar el camino adecuado. Este problema es conocido como el camino Hamiltoniano, y es un problema de tipo NP-completo, por lo que no se conocen algoritmos eficientes para resolverlo.

En una aproximación de de-Bruijn los reads se dividen en fragmentos denominados k-mers, luego se construye un grafo dirigido conectando los pares de k-mers que se solapan por k-1 bases. Luego se debe recorrer este grafo para reconstruir la secuencia original.

Pevzner compara ambas estrategias de ensamblado, como se encuentra en la figura 7. La secuencia de DNA consta de cuatro segmentos únicos y un repetido triple. Cada read corresponde a un nodo y las aristas que los conectan a sus solapamientos. En la estrategia OLC el resolver el camino del grafo implica visitar cada nodo una sola vez, problema del camino Hamiltoniano. En la parte inferior de la figura se encuentra la aproximación de de-Bruijn, que consiste en cuatro nodos y cada repetido se corresponde con una arista en vez de con un grupo de nodo. El grafo resultante es una representación más simple del problema y encontrar el camino requiere visitar cada arista del grafo una sola vez, lo que se denomina el problema del camino Euleriano. Este problema cuenta con múltiples algoritmos que lo resuelven en tiempo lineal (Pevzner et al. 2001).

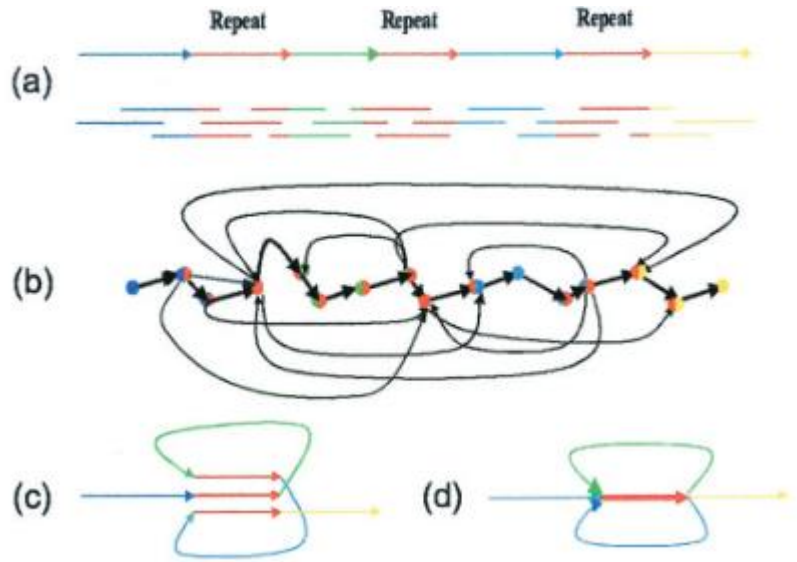

Figura 14.- Secuencia de DNA con tres repetidos (a). El grafo resultante en una estrategia OLC (b). Proceso de construcción de un grafo de-Bruijn (c). Resultado final del grafo de-Bruijn. (d) (Pevzner et al. 2001)

Secuenciado de las cepas MT1 y Liem de *T. vivax*

Para las secuenciaciones se utilizaron cepas de Liem-176 y MT1 obtenidas del laboratorio del Dr. Armando Reyna Centro de Estudios Biomédicos y Veterinarios, de la Universidad Simón Rodríguez, Caracas, Venezuela.

La secuenciación del genoma de *T.vivax* cepa MT1 fue realizada por Gonzalo Greif, en el laboratorio de Biología Molecular del Institut Pasteur de Montevideo, utilizando el equipo de secuenciamiento Genome Analyzer II de Illumina. De esta secuenciación se obtuvieron 23.5 millones de reads paired-end de 100 nucleótidos de longitud. El secuenciado de Liem-176 se realizó utilizando Illumina MiSeq paired-end de 150 bases de longitud obteniendo 6.2 millones de reads.

Ensamblado del genoma nuclear

Los reads obtenidos se analizaron en busca de restos de adaptadores de Illumina utilizando el software Scythe, versión 0.98. Este software utiliza una base de datos de secuencias de adaptadores y busca fragmentos de estos en los extremos de los reads, recortando los reads en caso que sea necesario.

Luego se recortaron los nucleótidos de baja calidad en los extremos de los reads utilizando el software Scythe, con la opción de recortar bases cuya calidad medida en la escala de Phred fuera menor a un valor de 30. Finalmente los reads cuya longitud final luego de ser recortados resultó ser menor a 65 nucleótidos se descartaron. En caso de que solo uno de los reads del par fuera descartado el otro pasó a un archivo de reads single-end.

Los resultados de los recortes de los reads por calidad se evaluaron utilizando el software FastQC (Andrews 2010), el cual permite observar rápidamente la calidad del set de datos con el que se va a trabajar computando la calidad por base y k-mers más frecuentes, lo cual es útil para encontrar sesgos o errores en el set de datos.

Este pipeline de corrección de errores fue utilizado para los datasets de MT1 y Liem-176.

Para realizar el ensamblado se testaron varios ensambladores: Velvet (Zerbino & Birney, 2008), SOAPdeNovo (Luo et al, 2012), Spades v. 3.1.0 (Bankevich et al. 2012) y Abyss (Simpson et al, 2009). La medida para evaluar la calidad del ensamblado fue el valor de N50, el tamaño del genoma ensamblado, y el alineamiento contra en genoma de la cepa Y486, utilizando el software alineador de genomas Mummer (Delcher et al. 1999).

El N50 es el tamaño del menor necesario contig en cubrir el 50% del genoma contando los contigs de mayor a menor longitud. Esta es una buena medida del tamaño de los contigs ensamblados, útil para evaluar que tan preciso fue el ensamblado y que grado de fragmentación se puede esperar.

Para MT1 en Abyss se probaron valores de k-mer de entre 40 y 64, y el mejor resultado fue obtenido con un k-mer de 50. En Liem-176 lo mejores resultados fueron con Spades, que utiliza varios k-mers combinados, y donde se utilizó una combinación de k-mers de 45 a 85 con paso 10.

Ensamblado del genoma mitocondrial

Para ensamblar el genoma mitocondrial de Y486 se utilizaron reads secuenciados con tecnología Sanger de la base de datos pública del NCBI (accession: CAEX00000000.1). Se utilizó el software Mira (Chevreux et al. 1999) cuya salida resultó en un único contig de 19.3 Kb de largo conteniendo el total de la región codificante del genoma.

Los contigs que conforman el genoma mitocondrial de MT1 y Liem fueron obtenidos de sus respectivos ensamblados de reads totales e identificados luego de ser mapeados con el maxicírculo de *T. brucei*, depositado en el Genbank, número de acceso M94286.1.

Orfogenicidad

En estudios previos llevados a cabo en nuestro laboratorio se ha observado que en el genoma de algunos tripanosomas se pueden encontrar extensas regiones con marcos abiertos de lectura (ORFs) solapantes que no son caracterizables utilizando métodos de identificación de secuencias por homología. Este fenómeno se ha denominado “orfogenicidad”.

Procesado de los datos

Los datos de partida utilizados fueron el genoma de la cepa Y486, descargado del GenBank (GCA_000227375.1) y ensamblados propios de las cepas MT1 y Liem. Con la finalidad de clasificar diferentes regiones del genoma nuclear se realizó una limpieza de regiones carentes de información y potencialmente conflictivas para posteriores análisis.

El primer paso consistió en eliminar las posiciones con Ns ya que no aportan información de secuencia y son generadas en el paso de scaffolding del ensamblado. Luego se dividieron las secuencias en fragmentos de 5 Kb con la finalidad de permitir que la variabilidad intra-contig pueda ser detectada. En caso de que el fragmento resultante fuera menor a 3 Kb, este fue descartado. El valor del largo fue elegido con el objetivo de generar la suficiente cantidad de datos para realizar análisis posteriores, si bien se pueden observar resultados similares con longitudes de fragmentos mayores y menores.

Luego se identificaron secuencias con altos niveles de repetición utilizando la herramienta etandem de la suite emboss. Este software identifica secuencia repetidas en tándem con altos niveles de identidad nucleotídica y reporta sus coordenadas en forma tabulada.

Utilizando la información de repetidos generada con etandem y utilizando scripts personalizados en Python se descartaron aquellos fragmentos en donde más de 75% de la secuencia la conforman repetidos. El procesado de los datos eliminado las secuencias altamente repetitivas permite descartar aquellas secuencias con los patrones nucleotídicos más extremos y así al realizar análisis estadísticos se evitan outliers y valores fuera de escala que distorsionen la representación e interpretación de los datos.

Cálculo de la frecuencia de trinucleótidos de cada fragmento secuenciado

Se utilizó el paquete Biostrings (Pages et al. 2017) en R que tiene herramientas para el procesado de archivos de secuencia en formato fasta.

Se carga el archivo de secuencias previamente fragmentado, se calcula la frecuencia de kmers de tamaño de palabra 3 para la secuencia original y para su reverso complementario, con el fin de evitar el sesgo de cada hebra. Luego se genera una tabla con las frecuencias de los trinucleótidos para cada una de las secuencias. Estos datos junto con los índices de orfogenicidad (ver más adelante) fueron utilizados como entrada para realizar análisis multivariados.

Para evitar el sesgo de hebra se calcularon las frecuencias tanto para la secuencia original como para su reversa complementaria, por lo cual hay valores repetidos para todos lo pares de trinucleótidos. Por ejemplo, la frecuencia calculada para AAA va a ser la misma que para TTT, en un DNA doble hebra, por lo cual ambas variables van a tener el mismo valor de frecuencia para todas las secuencias; y así con el resto de las variables. Por ello se eliminaron 32 variables que no aportan información sobre el sistema y representarían un problema de colinealidad en análisis futuros.

Análisis de componentes principales

El análisis de componentes principales (PCA) es una técnica estadística multivariada aplicada con el fin de reducir la dimensionalidad a un número menor de variables nuevas llamadas componentes principales. Estas son capaces de representar la variabilidad de los datos en un número sustancialmente menor de dimensiones intentando solo perder la variabilidad atribuible a ruido de fondo. Esto es posible porque es común que en un conjunto de datos solo existan unas pocas variables subyacentes reales (factores) que expliquen la mayor parte de la variabilidad observada. Encontrar los componentes principales consiste precisamente en identificar esas variables subyacentes basándonos en cómo se correlacionan las variables originales.

Los componentes principales por tanto tienen como finalidad transformar las variables originales a un nuevo conjunto de variables que tienen las siguientes características: son combinaciones lineales de las variables de los datos originales, no están correlacionadas y se encuentran ordenadas de acuerdo a la cantidad de variabilidad que pueden explicar (Everitt & Hothorn, 2011).

El resultado final será poder representar una nube de puntos en un número reducido de dimensiones que permita su representación gráfica evitando a su vez distorsionar las distancias originales entre los individuos (Husson, Le & Pages, 2010).

Un PCA realiza una proyección de los datos en nuevas dimensiones denominadas componentes principales; los componentes se eligen de tal modo que sean ortogonales busca minimizar la distancia entre las variables y su proyección en el componente. Al minimizar la distancia se maximiza la varianza de los puntos proyectados. El segundo y los siguientes componentes son seleccionados de igual manera y además no debe estar correlacionado con componentes anteriores. El requisito de no correlación entre los componentes hace que el máximo de componentes que se puedan calcular sea el número de elementos estudiados o el número de variables, el que sea menor.

PCA es una poderosa herramienta estadística que ayuda a reducir la variabilidad aleatoria (ruido de fondo), la redundancia (variables correlacionadas) y por tanto facilita la observación de patrones en los datos.

En el presente trabajo, y con el fin de obtener un perfil de las propiedades estadísticas de las secuencias del genoma de *T.vivax* se estudió la variabilidad de la frecuencia de trinucleótidos a través de un análisis de componentes principales.

Índices de orfogenicidad

Con la finalidad de caracterizar los fragmentos genómicos ensamblados en base a las propiedades orfogénicas de las secuencias, desarrollamos diferentes índices que reflejan la capacidad de una secuencia de generar marcos abiertos de lectura.

Una primera medida simple, consiste en determinar para cada nucleótido de una secuencia en cuantos marcos abiertos de lectura pertenece de longitud mayor a un umbral (por ejemplo166 codones o 500 nt). El valor de 500 nucleótidos claramente es arbitrario, pero es considerado como bastante mayor al largo esperado por azar en una secuencia con nucleótidos equiprobables. En concreto, 1 de cada 20 tripletes (3 de los 64) corresponde a un codón de terminación. Por tanto en una secuencia con composición igualitaria de bases, uno esperaría que el largo esperado de ORFs al azar sea 20 codones, es decir 60 nucleótidos, por lo que el valor de 500 es muy superior a la esperanza por azar. Lógicamente que el largo esperado por azar de las ORFs cambiará con la composición nucleotídica, especialmente si tenemos en cuenta que los codones stop son ricos en AT. Los valores que puede obtener este índice van desde 0, si el nucleótido no está contenido dentro de ningún ORF mayor a 166 tripletes hasta 6, si el nucleótido estuviera dentro de ORFs>166

en los 6 marcos de lectura posibles. Luego de estimado este valor para cada nucleótido, se obtiene el índice de orfogenicidad, el cual es el valor medio para el contig.

$$orfindex = \frac{\sum len_{\ ORFs>166}}{len_{contig}}$$

Frecuencias observadas y esperadas de codones de terminación

Para estimar la probabilidad de encontrar por azar codones stop en una secuencia hay que tener en cuenta la composición nucleotídica de dicha secuencia. La probabilidad de encontrar un codón de terminación se calcula a partir de la probabilidad de encontrar cada una de las bases que lo componen asumiendo independencia.

$$P(TAA) = P(T) * P(A)^2$$
$$P(TAG) = P(TGA) = P(T) * P(A) * P(G)$$

La probabilidad de encontrar un codón stop es la suma cualquiera de las probabilidades individuales y la denominaremos P1.

$$P_1 = P(stop_{exp}) = P(TAA) + P(TAG) + P(TGA)$$

La frecuencia observada de codones stop se contabiliza teniendo en cuenta ambas hebras. Esta frecuencia, se denominará P_2 y es un estimador empírico de la probabilidad de encontrar un codón stop en dicha secuencia.

$$P_2 = F(stop_{obs}) = \frac{(nTAA + nTAG + nTGA)}{2 * (n_{total})}$$

Podemos comparar ambos estimadores, es decir la estimación de probabilidad empírica P2, con aquella basada en la composición de nucleótidos: a esta relación la denominaremos índice stopOE.

$$stop_{OE} = \frac{F(stop_{obs})}{P(stop_{exp})} = \frac{P_2}{P_1}$$

Es evidente que si el índice stopOE es mayor a 1 implica que la frecuencia de codones de terminación en dicha secuencia es superior a lo que uno esperaría por azar basándose en la

frecuencia de aparición de la bases individuales, mientras que una deficiencia en la frecuencia de codones stops se reflejaría en valores menores a uno para este índice.

Distribución del largo de ORFs

La probabilidad de encontrar marcos abiertos de lectura de un largo dado se puede estimar basándose en la probabilidad de encontrar un codón stop utilizando la distribución de probabilidad geométrica. En una distribución geométrica se contabilizan el número de experimentos necesarios hasta obtener un éxito. Cada experimento es un ensayo de Bernoulli, es decir, ensayos independientes con sólo dos posibles resultados. En este caso la pregunta que hacemos es ¿cuál es la probabilidad de no encontrar un codón stop durante un número dado de codones?

Si definimos a ***p*** como la probabilidad de encontrar un codón de terminación, la probabilidad de encontrar un ORF de k codones de largo es (distribución geométrica de probabilidad):

$$P(L = k) = (1 - p)^k * p$$

Esto representa la probabilidad de tener un éxito (encontrar un codón de terminación) luego de k fracasos. En nuestro caso podría considerarse como la probabilidad de encontrar un ORF de largo k (por ejemplo 166 codones).

Sin embargo el universo de ORFs en el que estamos interesados también incluye a todos aquellos mayores a $\underline{k}$, por lo cual se debe recurrir a la sumatoria de la serie geométrica para obtener la probabilidad de encontrar ORFs mayores o iguales a k.

Esto se puede expresar de la siguiente manera:

$$P(L \geq k) = S_n = (1 - p)^k * p + (1 - p)^{k+1} * p + (1 - p)^{k+2} * p + \ldots + (1 - p)^{k+n} * p$$

Definiendo $q = 1 - p$

La sumatoria de esta serie geométrica puede escribirse como:

$$S_n = \sum_{l=k}^{n} pq^l = pq^k + pq^{k+1} + \cdots + pq^n$$

Multiplicando por $(1 - q)$ a ambos lados:

$$S_n = (1 - q) \sum_{l=k}^{n} pq^k = (1 - q)(pq^k + \ pq^{k+1} + \cdots + pq^n)$$

$$S_n = (1-q)\sum_{l=k}^{n} pq^l = (pq^k + pq^{k+1} + \cdots + pq^n) - (pq^{k+1} + pq^{k+2} + \cdots + pq^{n+1})$$

$$S_n = (1-q)\sum_{l=k\ (k=166)}^{n} pq^l = pq^k + pq^{n+1}$$

$$S_n = \sum_{l=k}^{n} pq^l = \frac{pq^k + pq^{n+1}}{1-q}$$

Cuando n tiende a infinito y $0 < q < 1$ la sumatoria se puede expresar como:

$$S_n = \lim_{n\to\infty} \sum_{l=k}^{n} pq^l = \frac{pq^k + pq^{n+1}}{1-q} = \frac{pq^k}{1-q} =$$

Dado como se definió anteriormente q, p=1-q por lo tanto:

$$S_n = \lim_{n\to\infty} \sum_{l=k}^{n} pq^l = (1-p)^k$$

Por lo cual cuando n tiende a infinito, la probabilidad de encontrar ORFs mayores o iguales a un largo k es:

$$P(L \geq k) = (1-p)^k$$

Si bien los contigs no tienen una longitud infinita, los términos para valores grandes son muy pequeños dado que el valor de probabilidad (<1) se está elevando a potencias muy altas, por lo que no cambian los resultados de forma significativa.

En nuestro caso donde el universo de ORFs en los que estamos interesados es el de los mayores o iguales a 166 codones de longitud, la probabilidad quedaría definida como:

$$P(L \geq 166) = (1 - P(stop_{obs}))^{166}$$

Número de ORFs esperados

Para calcular el número de ORFs iguales o mayores a un largo umbral, se debe tener en cuenta la probabilidad de encontrarlo así como el número de "intentos". La probabilidad de encontrar ORFs mayores o iguales a 166 codones $P(L \geq 166)$ fue calculada anteriormente. Para estimar el número de intentos se debe tener en cuenta que un ORF puede encontrarse en cualquiera de los seis marcos

de lectura, así como el largo medio esperado. Esto último está dado por la esperanza en la distribución geométrica:

$$E(L) = \frac{1-p}{p}$$

Se puede estimar la esperanza de L para la sub-población de ORFs mayores o iguales a 166 codones (k=166) en una secuencia, como el valor de la esperanza de L más la cota.

$$E(L_{ORF}) = \frac{1 - P(stop_{obs})}{P(stop_{obs})} + 166$$

Estos valores deben normalizarse por el largo de la secuencia, sobre el largo esperado calculado como $E(L_{ORF})$ anteriormente.

Con los datos anteriores es posible estimar el número de ORFs mayores a un valor mínimo (k=166) que se esperan en una secuencia de largo L:

$$nORFs_{exp} = \left(1 - P(stop_{obs})\right)^{166} * 6 * \frac{L_{seq}}{E(L_{ORF}) * 3}$$

La finalidad de estos índices es utilizarlos como una herramienta que permita entender los patrones (clusters de genes) que se observen en las gráficas PCA. Para esto, los índices se utilizaron como variables de exploración visual/funcional en los análisis PCA descritos anteriormente.

Resultados y discusión

Comparación del genoma mitocondrial de cepas americanas y africanas de *T.vivax*

Tras realizar los ensamblados se pudieron determinar y caracterizar los genomas mitocondriales de las tres cepas de *T. vivax* estudiadas, llevar a cabo la anotación de sus genes y determinar la naturaleza de las secuencias repetidas presentes en los extremos.

El genoma mitocondrial (maxicírculo) de MT1 fue determinado tras mapear los contigs ensamblados contra el maxicírculo de *T. brucei* utilizando el software de alineamiento Blast. Además la secuencia del maxicírculo de esta cepa fue completada con secuenciado Sanger, con el objetivo de generar un contig único conteniendo todas las secuencias codificantes de genes rodeadas por repetidos.

En la figura 15A se grafican los resultados de profundidad de un mapeo de los reads utilizados para ensamblar el maxicírculo de la cepa MT1, contra los dos contigs que conforman al maxicírculo ensamblado inicialmente. Con colores se señalan diferentes regiones del genoma mitocondrial. En color negro, naranja, verde y azul se representa las regiones del maxicírculo que están conformadas por diferentes tipos de secuencias repetidas; nótese la elevada profundidad de la región en color azul que corresponde a una región de repetidos colapsada que no pudo ser resuelta por el ensamblador pero cuya longitud puede estimarse en base a su profundidad. En rojo se señalan las zonas que contienen los genes mitocondriales.

En la figura 15B se muestra de forma esquemática el ensamblado inicial obtenida para el genoma mitocondrial. El primer contig en 5’ contiene regiones repetidas y los genes ribosomales mitocondriales 12S y 9S. El segundo contig contiene el resto de los genes mitocondriales y una secuencia repetida en 3’. Utilizando primers en el 3’ del primer contig y el 5’ del segundo se secuenció el gap entre ambos, que resultó en una secuencia repetida de 368 pb. En la figura 15C se muestra la posición de los genes en el maxicírculo de *T.brucei*.

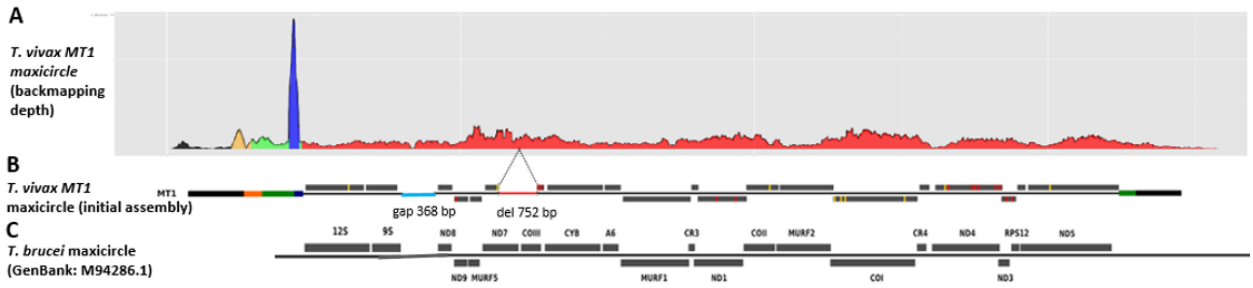

Figura 15.- (A) Gráfico de profundidad del backmapping de reads. En colores se indican diferentes regiones del genoma; en negro, naranja, verde y azul se representa las regiones repetidas, en rojo las zonas con genes mitocondriales. En la imagen inferior se muestra el ensamblado inicial de MT1 (B) y la secuencia del maxicírculo de *T.brucei* (C).

Para ensamblar el maxicírculo en la cepa Y486 se mapearon los reads de secuenciado Sanger disponibles en la base de datos del NCBI contra el ensamblado del maxicírculo de MT1. Con aquellos que mapearon se realizó un ensamblado utilizando el software Mira y se obtuvo en un solo contig de 18.730 bp de largo conteniendo la totalidad de la región codificante del genoma mitocondrial y parte de los repetidos de los extremos. Este contig fue extendido llevando a cabo un proceso manual de extensión de las regiones solapantes hasta que alcanzó los 20.400 bp de longitud.

En Liem176 se mapearon los reads utilizando Bowtie2 contra el maxicírculo de MT1, con los reads que mapearon se realizó el ensamblado utilizando el software SPAdes obteniendo varios contigs. El scaffolding de los contigs fue realizado con la ayuda de datos propios previamente obtenidos de secuenciado 454 del transcriptoma obteniendo así un único contig de 15.330 pb.

Para las tres cepas de *T. vivax* se identificó y anotó la región de 15 kb aproximadamente que contiene los genes mitocondriales, mientras que la región repetitiva en los extremos del maxicírculo fue caracterizada solo para la cepa MT1. Esta región fue secuenciada con Sanger debido a su longitud y naturaleza repetitiva, sin embargo se puedo determinar que estos repetidos especie-específicos son compartidos por las tres cepas y se encuentran el mismo número de veces (ver figura 17).

Luego se realizaron comparaciones de la región de 15 kb que contiene los genes mitocondriales en las tres cepas. Como es de esperar las cepas americanas son más similares entre sí que con la cepa africana; de hecho comparten la mayoría de las diferencias nucleotídicas y deleciones respecto a la cepa africana, siendo la mayoría de sus diferencias sustituciones (ver tabla 2).

El cambio más notable entre las cepas americanas y la africana es una deleción de 752 bp que afecta dos genes que codifican para ND7, que sufre una deleción de los últimos 427 bp en su región 3' y COII que tiene una deleción de 248 bp en la región 5' del gen. Esta deleción fue confirmada cuando se realizó una PCR para amplificar dicha región y no se obtuvieron fragmentos que cubrieran esa zona, lo cual también indica que la deleción está presente en todos los maxicírculos lo cual implica que la función de estos dos genes está perdida tanto en MT1 como en Liem176.

Otras pequeñas deleciones e inserciones se observaron en otros seis genes, y aunque son pequeñas producen cambios en el marco de lectura que si no son corregidos post-transcripcionalmente no pueden traducirse a una proteína funcional.

Gen	Tipo de edición	Pos:Base (AA)	Efecto de la mutación	Cepa
1 ND8	Pan-edición	-	ND	
2 ND9	Pan-edición	21: ins T	Possible frameshift	MT1, Liem
3 MURF5	No editado	-		
4 ND7	Pan-edición	248: G>T 275: del 424 pb	Missense Major deletion	MT1 MT1, Liem
5 COIII	Pan-edición	1: 248 del 5: G>C* 13: del C * 55: ins AG	Major deletion Posible frameshift	MT1, Liem MT1, Liem MT1 MT1, Liem
6 Cyb	Editado 5'	-		
7 A6-ATPase	Pan-edición	-		
8 ND2 (MURF1)	No editado	17: del TACA 1212:A>G	Frameshift Point mutation	Liem Liem
9 CR3	Pan-edición	-	ND	
10 ND1	No editado	491:ins T 863: del A	Frameshift Restores frame	MT1, Liem MT1, Liem
11 COII	Edición parcial	456: ins TGC (C)	Insertion of 1 aa	MT1, Liem
12 MURF2	Editado 5'	-		
13 COI	No editado	24: T>A (C>W) 154: G>T (G>C) 237: G>A (G>S) 1399: G>A (V>I)	missense missense missense missense	MT1, Liem MT1 MT1 MT1, Liem
14 CR4	Pan-edición	-	ND	
15 ND4	No editado	194: C>T 254: del ACATAT 712: ins TT 778: del T 846: G>A 1175: ins AT 1257: del AT	Missense Deletion of 2 aa Frameshift Frameshift Synonymous Restores frame Frameshift	MT1, Liem MT1, Liem MT1, Liem MT1, Liem MT1, Liem MT1, Liem MT1, Liem
16 ND3	Pan-edición	105: del TT 205: del TT	Frameshift Frameshift	MT1, Liem MT1, Liem
17 RPS12	Pan-edición			
18 ND5	No editado	430: G>T (G>C)	missense	MT1, Liem

Tabla 2.- Lista de genes mitocondriales anotados indicando el tipo de edición mutaciones con respecto a la cepa Y486, efecto de la mutación y cepa americana en la que ocurre.

Para confirmar si todas estas mutaciones están presentes en todos los maxicírculos o si existe heteroplasmia, es decir la mutación está presente solo en parte de la población se realizó el mapeo de los reads genómicos contra los ensamblados y se observó que las mutaciones son homogéneas en estas cepas (ver figura 16).

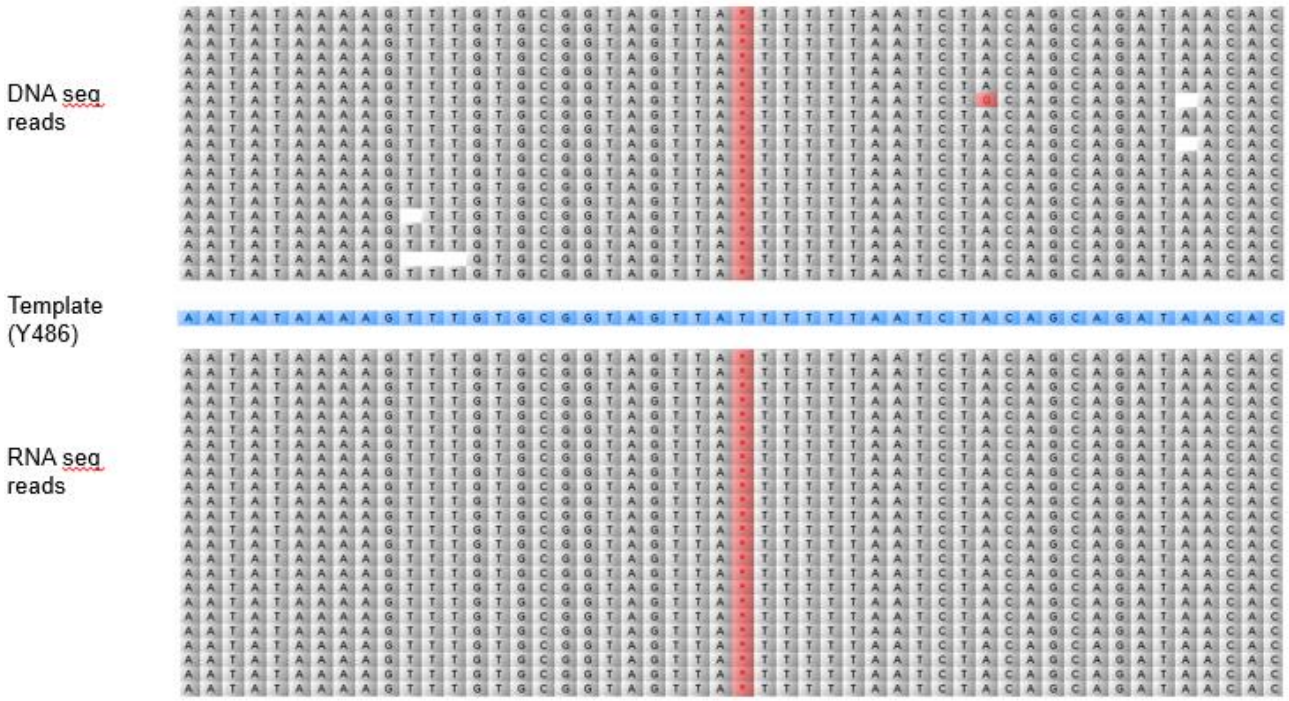

Figura 16.- Un fragmento del gen ND1 de Y486 fue utilizado como secuencia blanco del mapeo para evidenciar la deleción en MT1. En la mitad superior se muestra el mapeo de reads derivados de secuenciado de DNA y en la mitad inferior reads derivados del secuenciado de RNAseq de Liem-176.

También se evaluó la posibilidad de que estas mutaciones fueran corregidas post-transcripcionalmente durante el proceso de editado de mRNAs. En cuatro de los genes es difícil definir si las mutaciones son corregidas porque son pan-editados, sin embargo se observan en MURF1, ND1 y ND4 que no editan y sus secuencias genómicas son idénticas a las de mRNA. Otros dos genes COI y ND5 tienen cambios nucleotídicos que implican sustituciones aminoacídicas en las cepas americanas y se encuentra conservada en la cepa africana y en homólogos de *T. brucei* y especies filogenéticamente más lejanas como *T. cruzi* y *Leishmania donovani*, por lo cual cambios en estas posiciones conservadas sugieren mutaciones potencialmente deletéreas.

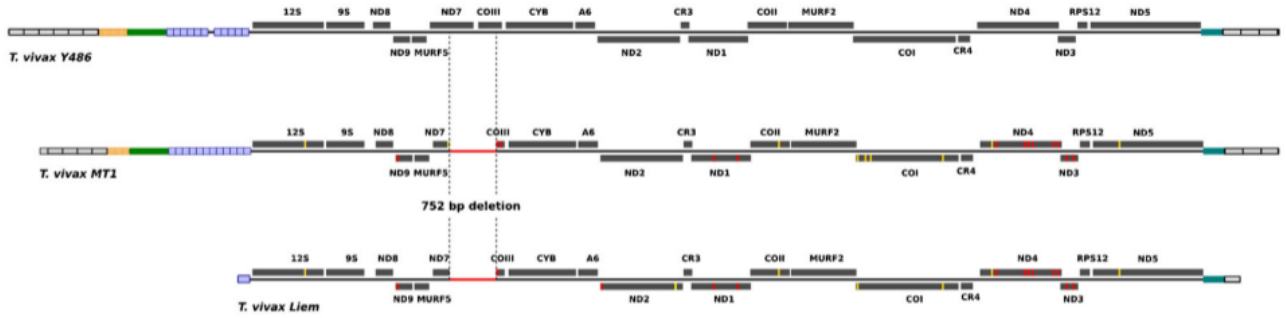

Figura 17.-Comparación del alineamiento de los tres genomas mitocondriales de *T. vivax* estudiados. Los bloques grises en 5' representan repetidos de 175 pb, los naranjas de 24 pb, los celestes de 105 pb; la región verde es especie-específica. Las marcas rojas representan indels y las amarillas sustituciones nucleotídicas.

Edición de mRNAs en cepas americanas y africanas

Se estudió además la edición de genes en las cepas americanas y la africana de *T. vivax*. Para identificar los mRNAs maduros o sea ya editados se ensambló el transcriptoma de Y486 y se tradujo a su secuencia de aminoácidos que luego fue comparada usando Blast con los genes mitocondriales traducidos de *T. brucei* debido a que la secuencia aminoacídica de las proteínas codificadas se espera que este conservada, no necesariamente así la secuencia nucleotídica. Esto permite identificar los mRNAs maduros. La comparación entre los mRNAs maduros y sus correspondientes genes codificantes permite determinar de que manera ocurre el proceso de edición y a su vez anotar con precisión las coordenadas de inicio y fin de los genes, lo cual es de particular complejidad en genes sujetos a un editado extremo.

De los dieciocho genes codificantes de proteínas que se encuentran en el maxicírculo, doce requieren de ediciones post-transcripcionales para poder ser traducidos. Los mRNAs de todos ellos se encontraron en la cepa africana de *T. vivax*, tanto en su forma editada como sin editar. En cambio en las cepas americanas de los doce genes mitocondriales que requieren edición solamente tres (A6 (ver figura 18), RPS12 y MURF2) llevan a cabo la edición de manera correcta, mientras que para los nueve restantes (ND3, ND7, ND8, ND9, COII, COIII, Cyb, CR3 y CR4) no hay indicios de que se lleve a cabo el proceso de edición. Tan solo en CR3 se encontraron trazas de edición en su extremo 3'. Esto implica que nueve de los genes mitocondriales no logran madurar sus mRNAs por lo que no pueden ser traducidos a una proteína funcional.

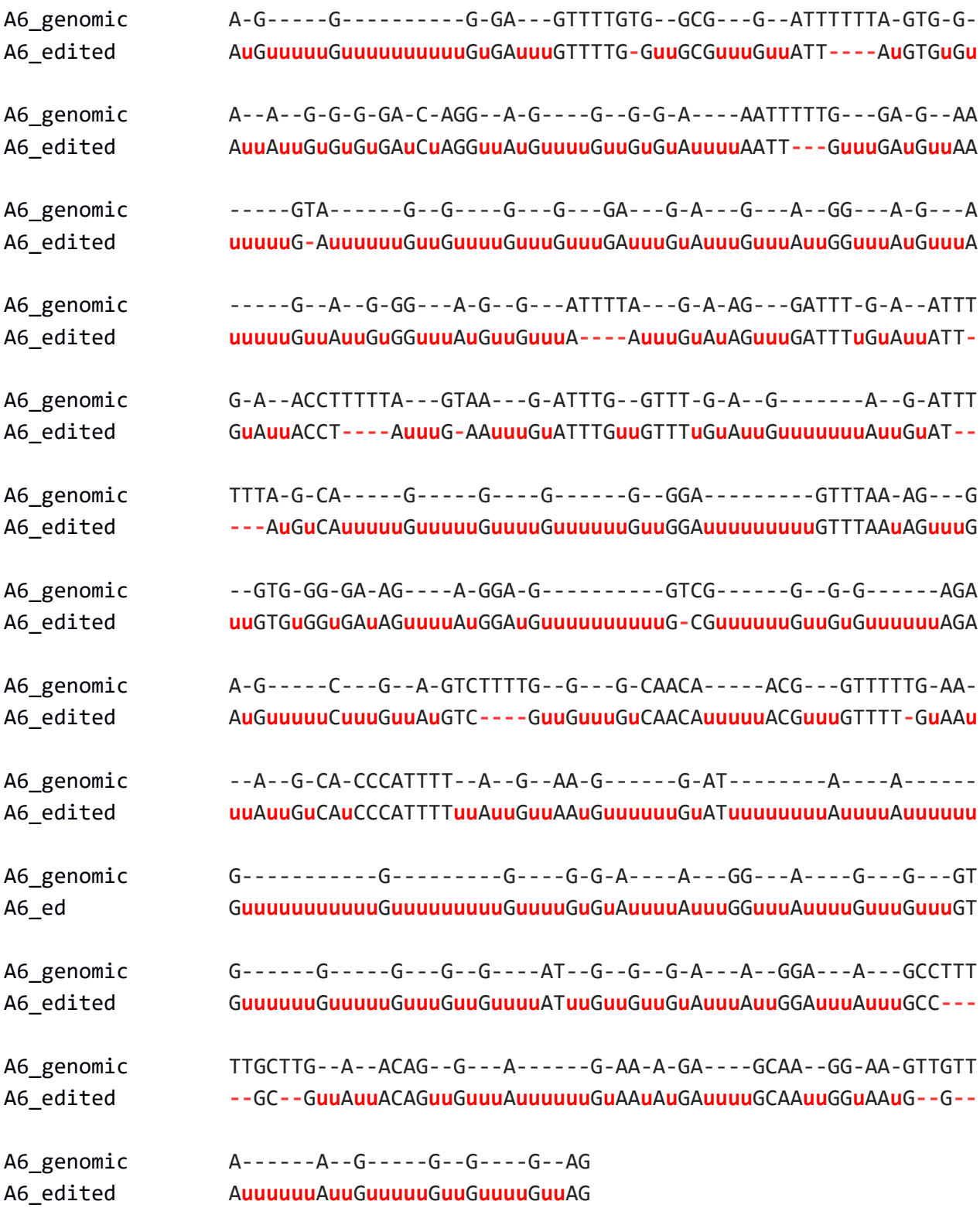

Figura 18.- Ejemplo de edición generalizada en el gen A6. En la línea superior se encuentra la secuencia genómica. En la segunda línea la secuencia del mRNA luego de ser editado. En color rojo se ilustran las inserciones y deleciones realizadas por el complejo de edición y los gRNAs.

Comparación de las poblaciones de minicírculos

Se estudió la población total de minicírculos con el fin de identificar los RNAs guía y determinar si la pérdida de la edición en nueve de los genes mitocondriales en las cepas americanas podía deberse a la pérdida de los RNAs guía que dirigen el proceso de edición.

Los minicírculos en tripanosomas no presentan conservación entre especies pero comparten una región de aproximadamente 120 bp de largo que contiene tres bloques llamados CSB-1, CSB-2 y CSB-3 (conserved sequence block, ver figura 19) de conservación variable; en particular el bloque más conservado en tripanosomas es CSB-3 o UMS (universal minicircle sequence) de 12 nt que es también el sitio de inicio de la transcripción, CSB-1 tiene 10 nt y presenta menos conservación y en CSB-2 la conservación de secuencia es marginal (Ray 1988).

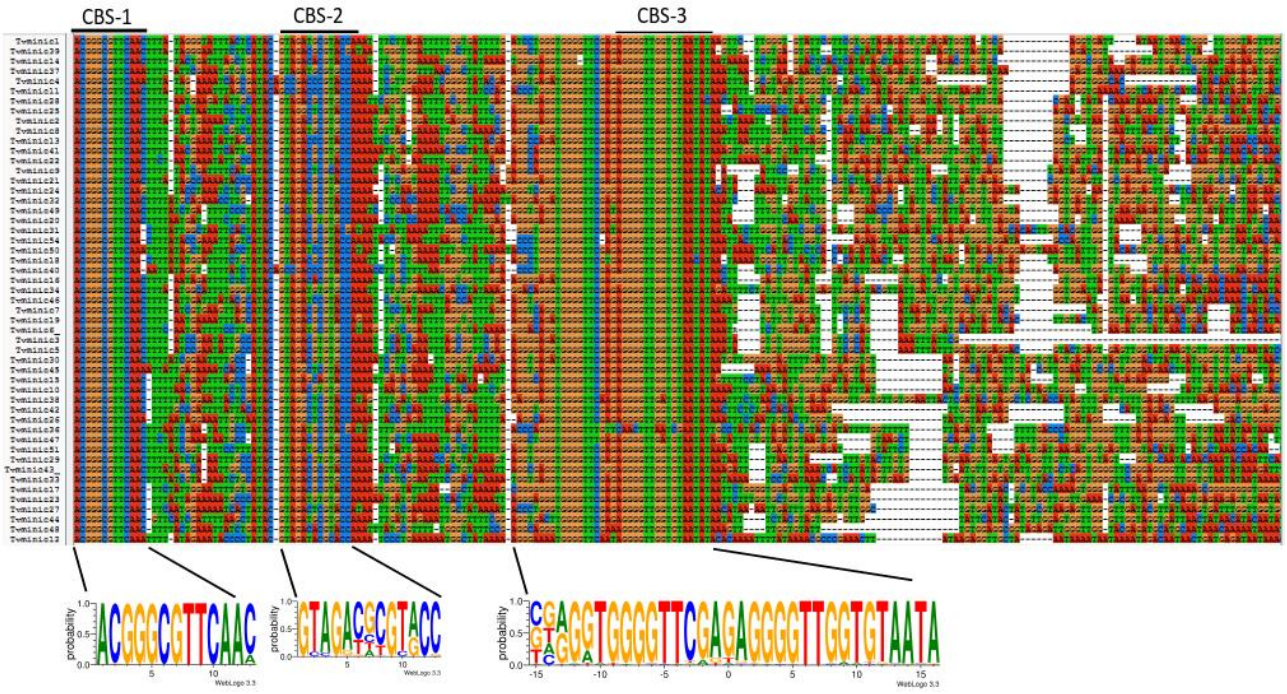

Figura 19.- Alineamiento de la región conservada que contiene los bloques CSB en 24 minicírculos. En la parte inferior de la imagen se muestra la representación como logo de cada uno de los tres bloques conservados.

Debido a que las secuencias CSB son muy cortas se utilizó una doble estrategia para su identificación; se mapeó la secuencia conservada CBS de *T. brucei* contra los contigs ensamblado de *T. vivax* y además se estudiaron las características estadísticas propias de la secuencia nucleotídica. Se realizaron análisis de componentes principales utilizando la frecuencia de dinucleótidos con el fin de separar los minicírculos de otras secuencias. Combinando ambas fuentes de datos fue posible identificar 54 variantes de minicírculos, con diferentes gRNAs, en MT1 y 46 en

Liem-176. Todos los de Liem-176 tienen homólogos en MT1, mientras que MT1 presenta algunos minicírculos propios.

Los resultados de ambas cepas americanas muestran que tienen un reducido conjunto de minicírculos considerando lo extensa que es la edición en el genoma mitocondrial.

También se investigó la abundancia relativa de cada tipo de minicírculo evaluando la profundidad de los mapeos de reads de DNAseq sobre los contigs de los minicírculos ensamblados. La profundidad relativa de la secuenciación de DNAseq permite calcular la frecuencia con la que se encuentra cierta clase de minicírculos en la población del genoma mitocondrial y la profundidad de RNAseq la abundancia relativa de dichos RNAs en las muestras secuenciadas.

El resultado observado fue una alta correlación en la abundancia a nivel de DNA de los distintos tipos de minicírculos en las cepas americanas (figura 20A). Además, cuando se estudia la profundidad de los secuenciados de DNAseq y RNAseq en Liem-176 se puede observar (figura 20B) que existe correlación entre la frecuencia con la que se encuentra un minicírculo a nivel de DNA y los niveles de transcripción.

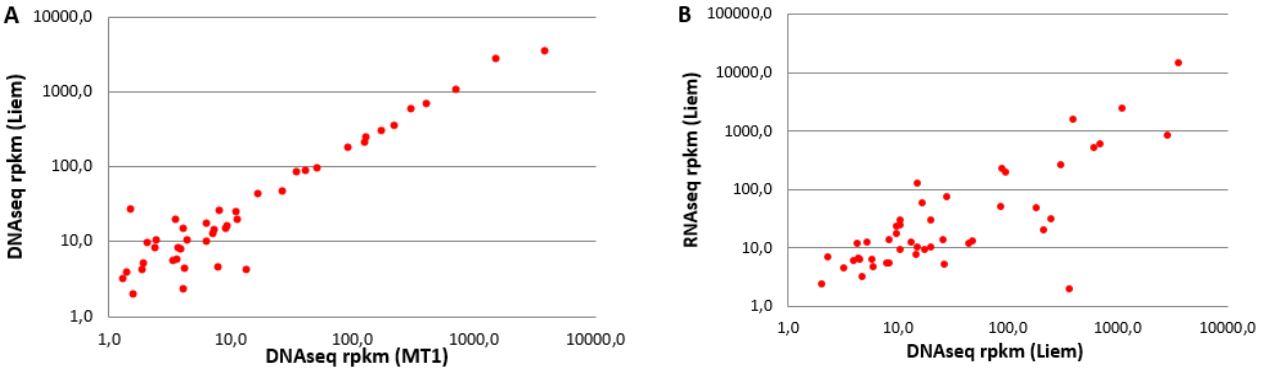

Figura 20.- En la figura A se muestra la fuerte correlación entre la cantidad de reads obtenidos en el secuenciado de DNA de MT1 y el de Liem-176 indicando que las poblaciones de los distintos tipos de minicírculos son similares. En la figura B se muestra que existe una correlación entre la frecuencia de un tipo de minicírculo y su abundancia a nivel de RNA.

Una observación interesante a partir del mapeo de los reads de RNAseq sobre los minicírculos ensamblados es que se transcribe la totalidad del minicírculo como un policistrón, incluyendo la región CSB, esto implicaría que al igual que lo que ocurre en el genoma nuclear debe existir un proceso de maduración post-transcripcional de los genes codificados por los minicírculos (ver publicación anexa).

Además se caracterizó la población de gRNAs de los minicírculos utilizando el software de alineamiento wu-blast con una matriz de sustitución modificada. En concreto se permitieron

uniones G-U lo que implica extender el tipo de alineamientos válidos. Esta modificación está justificada por el hecho de que las interacciones entre el gRNA y el mRNA en proceso de maduración ocurren normalmente.

En la cepa americana Liem-176 se obtuvieron la mayoría de los gRNAs necesarios para la edición de los genes de la subunidad A6 de la ATPasa y la proteína ribosomal RPS12. Sin embargo se encontraron muy pocos gRNAs adicionales. De hecho la enorme mayoría de los gRNAs que serían necesario para el proceso de maduración de los restantes genes mitocondriales está ausente. Los pocos gRNAs que eventualmente participarían en la maduración de estos otros genes fueron encontrados en minicírculos que también tenían gRNAs para los genes que se editan correctamente. Es decir la presencia de unos pocos gRNAs involucrados en la edición de otros genes se debe tan solo a que se encuentran como "polizones" acompañando a otros gRNAs imprescindibles para la correcta edición de A6 y RPS12.

Está perdida de la capacidad codificante de los genomas mitocondriales parece estar asociada a los cambios en el ciclo de vida de las cepas americanas. Es decir, el proceso de acumulación de mutaciones de pérdida de función de varios genes mitocondriales junto a la pérdida de minicírculos sugiere que las cepas americanas de *T. vivax* se encuentran sufriendo un proceso evolutivo similar al que ocurre en cepas de *T. brucei* que se han expandido por fuera de la región con moscas tsé-tsé y se mantienen exclusivamente en el hospedero mamífero transmitiéndose de forma mecánica e incluso han perdido la capacidad de multiplicarse en la mosca tsé-tsé (Lun et al. 2010).

Cuando el parásito se encuentra en la sangre de un hospedero mamífero depende totalmente de la glucosa del hospedero ya que utiliza exclusivamente la glicólisis para satisfacer su demanda de ATP. En este estadio no cuenta con un ciclo de Krebs funcionando ni con fosforilación oxidativa. Las reacciones que catalizan la glicólisis tienen lugar a una elevada tasa y difieren en varios aspectos de la glicólisis en otros eucariotas (Clayton & Michels 1996).

Las enzimas que catalizan los primeros siete pasos de la vía glicolítica se encuentran localizadas en alta concentración en organelos especializados relacionados con los peroxisomas y glioxisomas de otros eucariotas, a los cuales se les denomina glicosomas (Opperdoes & Borst 1977).

En *T.brucei* la diferenciación a la forma procíclica, en el insecto, está acompañada por cambios morfológicos, como el aumento en el tamaño y actividad de la mitocondria, y a cambios bioquímicos, como el cambio en la generación de ATP que en la forma procíclica depende de la cadena respiratoria (El-Sayed et al. 2000).

La forma sanguínea del parásito al final de la glicólisis excreta piruvato como producto final, mientras que en el insecto el piruvato no es excretado sino metabolizado en la mitocondria en donde puede seguir dos caminos para generar energía; entrar en el ciclo de Krebs o ser convertido dentro de la mitocondria a acetato (van Weelden et al. 2003).

En *T. equiperdum* y *T. evansi* se han observado cambios en algunos aminoácidos en el gen nuclear que codifica para la subunidad gamma de la ATP sintasa, la cual se encuentra muy conservada en otros tripanosomas. Esto permitiría compensar la pérdida de la porción F0 de la ATP sintasa codificada por el gen A6-ATPasa mitocondrial (Schnaufer et al. 2005). Se consideró la posibilidad que se observase lo mismo en las cepas americanas de *T. vivax* ya que parece estar en camino a perder el genoma mitocondrial, sin embargo al alinear las secuencias de la subunidad gamma de la ATP sintasa de varios tripanosomas junto a los *T. vivax* americanos no se encontraron estos cambios; algo consistente con el hecho de que aún mantienen funcional al gen de A6-ATPasa.

Ensamblado y anotación de los genomas nucleares

Para ensamblar los genomas nucleares de las cepas americanas se evaluaron los resultados obtenidos con diferentes ensambladores. Los mejores resultados, evaluando N50 y alineamientos con Mummer contra la cepa Y486, se obtuvieron con Abyss para el ensamblado del genoma de la cepa MT1 y Spades para el ensamblado del genoma de Liem-176.

Cepa	Número contigs	N50 (pb)	Max. contig (pb)	Total (Mpb)
MT1	1600	46631	345974	26.5
Liem	4762	8040	72432	24.5
Y486 (Genbank)	9736	7367	60075	41.3

Tabla 3.- Estadísticas del ensamblado de *T. vivax* para las cepas MT1 y Liem (ensamblado propio) e Y486 depositado en el GenBank.

Luego se utilizó el software getORF para extraer todos los ORFs mayores o iguales a 500 nucleótidos de longitud que comienzan con un codón de iniciación y terminan en un codón stop. Estas secuencias traducidas a aminoácidos fueron los utilizados inicialmente en la anotación genómica de ambas cepas.

Los ORFs se mapearon utilizando blastp y un e-value de 1e-10 contra una base de datos de genomas de Kinetoplástidos. Además se utilizó el software Blast2GO para extraer términos de ontología de las bases de datos GO y para mapear las secuencias contra la base de datos de dominios y familias de proteínas InterPro. Se tomó el mejor hit de blastp, los términos GO y los resultados de InterPro para realizar la anotación funcional de estas secuencias. Con estos datos se

crearon archivos en formato GFF que pueden ser utilizados en navegadores de genoma como Artemis y visualizadores de mapeos como Tablet.

Clasificación del espacio genómico

En trabajos anteriores de nuestro grupo de trabajo se observó en el genoma de la cepa Y486 dos grandes regiones genómicas que pueden ser caracterizadas en base a las propiedades estadísticas de la frecuencia de nucleótidos. Al realizar un análisis de componentes principales de la frecuencia de trinucleótidos para contigs del ensamblado genómico se pueden identificar de forma precisa dos tipos de secuencias con propiedades nucleotídicas diferentes.

Utilizando datos del ensamblado de estos contigs de Y486 disponibles de forma pública en el GenBank y ensamblados propios de las cepas MT1 y Liem se procedió a repetir estos estudios utilizando una estrategia similar donde además se aplicaron filtros en regiones de alta repetitividad que podrían aportar sesgos y se fragmentaron los contigs a largos similares.

Para caracterizar el espacio genómico en las diferentes cepas de *T.vivax* el primer paso fue fragmentar los scaffolds ensamblados a longitudes similares para evitar sesgos y a su vez poder contar con más puntos muestrales para analizar.

Para realizar la fragmentación se creó un script en Python que divide los scaffolds eliminando las regiones con Ns, luego a los contigs resultantes los divide en fragmentos de 5Kb descartando aquellas secuencias menores a 3Kb. Además se identificaron en las secuencias las regiones repetitivas para excluirlas del análisis ya que resultan en valores extremos y que dificultan la interpretación y la visualización gráfica de los resultados. Para este fin se utilizó la herramienta etandem de EMBOSS con sus parámetros por defecto que da por salida las coordenadas en los fragmentos de secuencias repetidas de entre 10 y 300 bases de largo. Las coordenadas de los repetidos se utilizaron con la finalidad de no considerar los eventuales ORFs presentes en estas regiones, ya que son característicos de la naturaleza repetitiva de la secuencia. Se excluyeron de posteriores análisis los fragmentos de contigs compuestos en más de un 75% de su largo por secuencias repetitivas.

Utilizando el paquete Biostrings en R se espejaron las secuencias, es decir se generó para cada secuencia su reverso complementario, con el fin de evitar el sesgo de cada hebra y se generó una tabla con las frecuencias de trinucleótidos para cada uno de estos fragmentos. Estos datos se utilizaron como entrada para realizar un análisis de componentes principales, en donde se observó que los dos primeros componentes explicaban entre el 54 y el 60% de la varianza observada en las secuencias de las diferentes cepas.

Para la cepa Y486, en la figura 21 se muestran los resultados de graficar los dos primeros componentes y se pueden observar claramente dos nubes de puntos que se corresponden con dos grupos de secuencias con características de composición nucleotídica diferente.

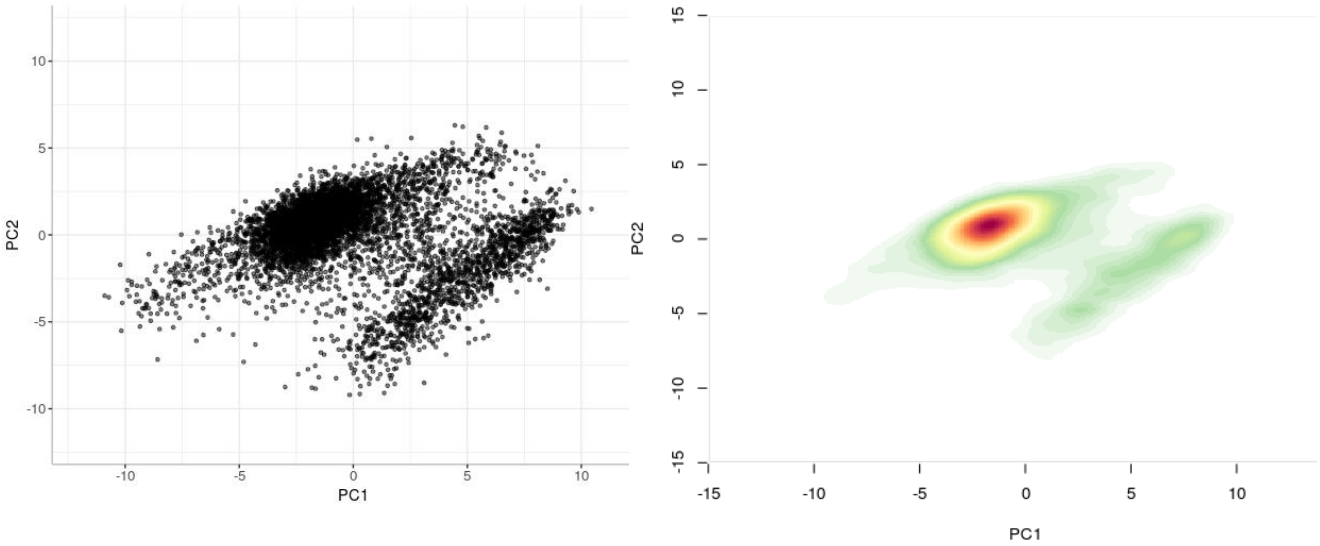

Figura 21.- PCA de la frecuencia de trinucleótidos de *T. vivax* Y486. Cada punto representa un fragmento de contig de entre 3 a 5 Kb. A la derecha un mapa de densidad de secuencias que muestra una clara división en dos poblaciones de secuencias.

Para averiguar las características funcionales de las secuencias asociadas a ambas nubes de puntos se muestrearon cien fragmentos al azar de la nube superior y otros cien de la inferior, marcados en rojo y verde en la figura 22. Se mapearon con Blast las secuencias codificantes anotadas de *T. vivax* Y486 contra estas secuencias elegidas al azar en cada nube. Los resultados de estos mapeos se resumen en las tablas 4 y 5.

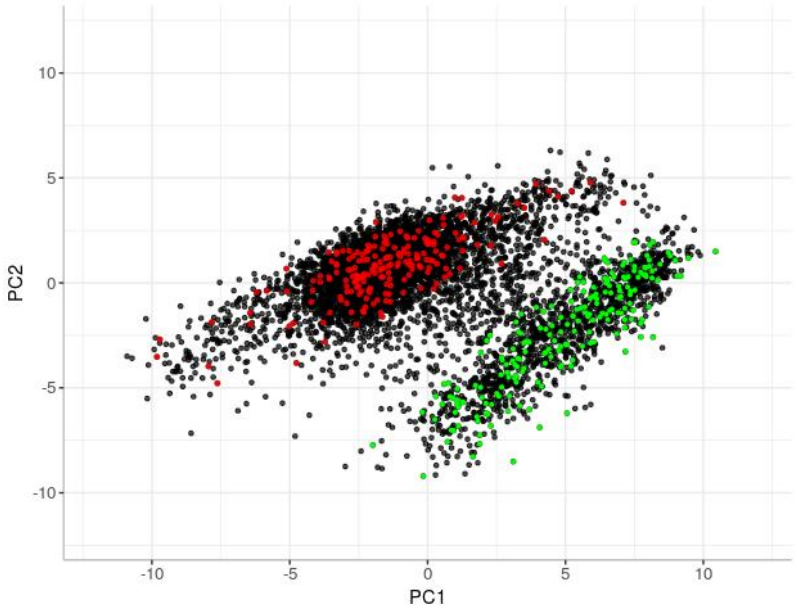

Figura 22.- PCA de la frecuencia de trinucleótidos de *T. vivax* Y486. Cada punto negro es un fragmento de entre 3 a 5 Kb, los puntos rojos y verdes son cien fragmentos, de cada nube de puntos, seleccionados al azar a los cuales se investigó la presencia de genes.

Cantidad	Anotación
61	variant surface glycoprotein, (VSG), putative
36	reverse transcriptase (RNA-dependent DNA polymerase)
15	SLACS reverse transcriptase, putative
6	retrotransposon hot spot protein (RHS), putative
4	proteophosphoglycan ppg4

Tabla 4.- Resultados obtenidos de mapear con Blast los CDS anotados de *T.vivax* Y486 contra los fragmentos seleccionados en color verde, en la figura 22.

Cantidad	Anotación
245	reverse transcriptase (RNA-dependent DNA polymerase)
200	retrotransposon hot spot protein (RHS), putative
8	ADP-ribosylation factor GTPase activating protein, putative
7	cysteine peptidase
6	DNA topoisomerase III, putative
5	cysteine peptidase, precursor
5	nucleic acid binding protein, putative
4	Zinc finger DHHC domain containing transmembrane protein

4	metacaspase, putative
3	40S ribosomal protein
3	cyclophilin, putative
3	RNA binding protein, putative

Tabla 5.- Resultados obtenidos de mapear con Blast los CDS anotados de *T.vivax* Y486 contra los fragmentos seleccionados en color rojo, en la figura 22.

Los resultados obtenidos del mapeo (tablas 4 y 5), muestran que las secuencias de la nube superior (figura 22), la más densamente poblada corresponde a secuencias asociadas a genes 'clásicos' como los de housekeeping. A esta región se le denominó ‘core’.

La nube de puntos inferior (figura 22), y de menor densidad, corresponde a regiones genómicas que contienen los genes asociados a la generación de variedad antigénica, aunque abundan secuencias codificantes descritas como proteínas hipotéticas de función aún no identificada. A esta región genómica le hemos denominado región ‘orfogénica’ porque, como veremos más adelante, contiene abundantes marcos de lectura muy largos.

Cuando se realiza el mismo tipo de análisis, un PCA de frecuencia de trinucleótidos en la cepa MT1 si bien puede observarse claramente la nube de puntos correspondiente a la región que contiene la mayoría de los genes ‘clásicos’, el espacio orfogénico que en Y486 se distinguía claramente (figura 22), aquí pasa a ser una región de puntos más dispersa (figura 23).

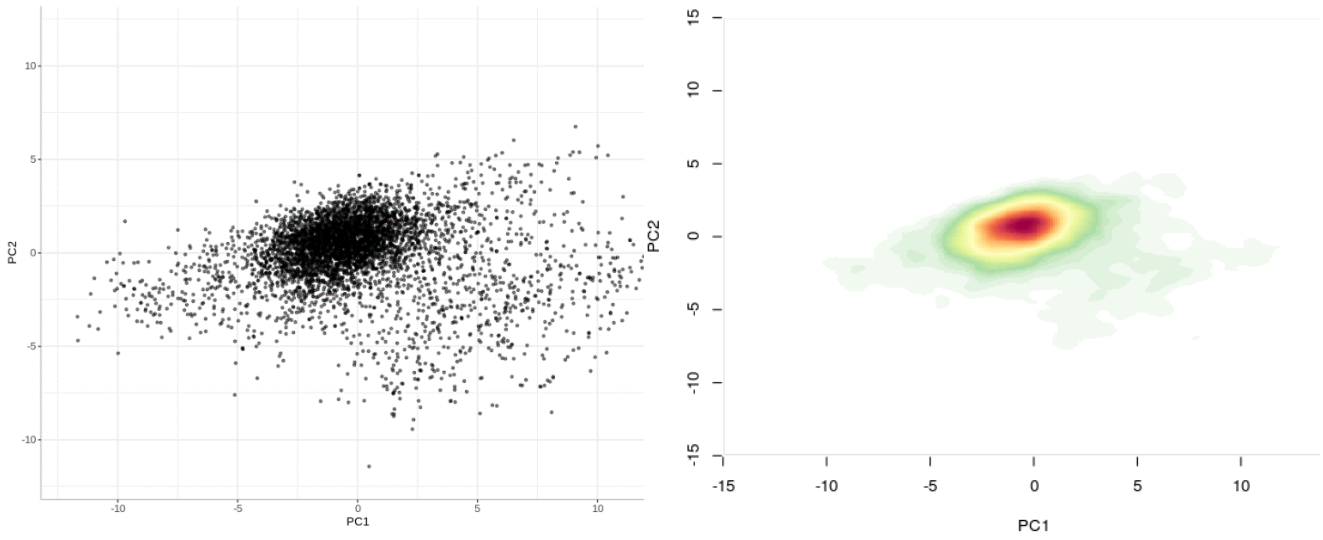

Figura 23.- PCA de la frecuencia de trinucleótidos de *T. vivax* MT1. Cada punto representa un fragmento de contig de entre 3-5 Kb. A la derecha un mapa de densidad de estos puntos. Puede observarse una clara pérdida de la población de secuencias asociada al espacio orfogénico.

En Liem-176, la otra cepa americana muy cercana a MT1 se observa el mismo patrón, (figura 24) el espacio del core se observa claramente, mientras que la región orfogénica se encuentra menos poblada.

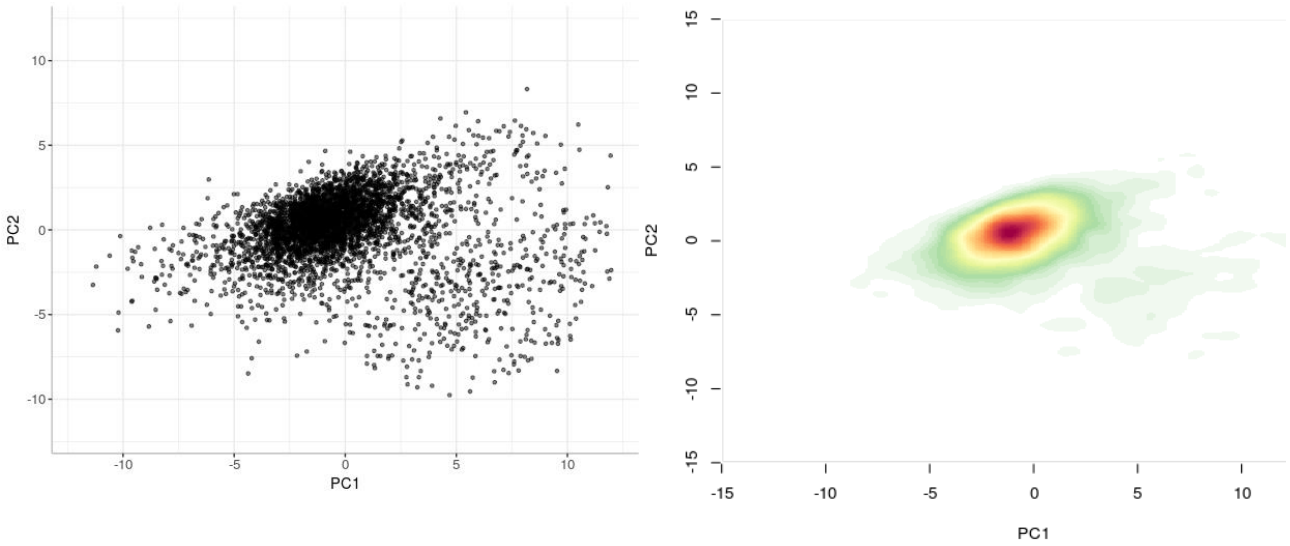

Figura 24.- PCA de la frecuencia de trinucleótidos de *T. vivax* Liem-176. A la derecha un mapa de densidad de los puntos. Cada punto representa un fragmento de contig. Se observa una clara pérdida de la población de contigs asociada al espacio orfogénico.

Aunque las poblaciones de secuencias se separan en todas las cepas, esta diferencia puede observarse claramente solamente en Y486, por lo que aparentemente las cepas americanas han visto reducido su genoma en las regiones asociadas con el espacio orfogénico.

Orfogenicidad del espacio genómico

Una característica de ciertas regiones del genoma es la presencia de marcos abiertos de lectura (ORFs) largos solapados y presentes en ambas hebras; dicha característica fue definida como “orfogenicidad”.

Utilizando el software Artemis (Rutherford et al, 2000) se puede visualizar la distribución de los marcos abierto de lectura mayores a un largo dado en una secuencia de DNA. En la figura 25 se señalan en color cian los ORFs de una longitud mayor o igual a 500 nucleótidos o 166 codones (un umbral arbitrario considerado como bastante mayor al largo esperado por azar en una secuencia con nucleótidos equiprobables, ver “Materiales y métodos”). En la figura 25A vemos como en los contigs del core estos ORFs se localizan en una misma hebra de forma consecutiva en un fragmento

genómico.En la figura 25B, en cambio se observa una distribución muy diferente donde los ORFs se encuentran en ambas hebras y solapados. Esta región genómica puede considerarse de elevada orfogenicidad.

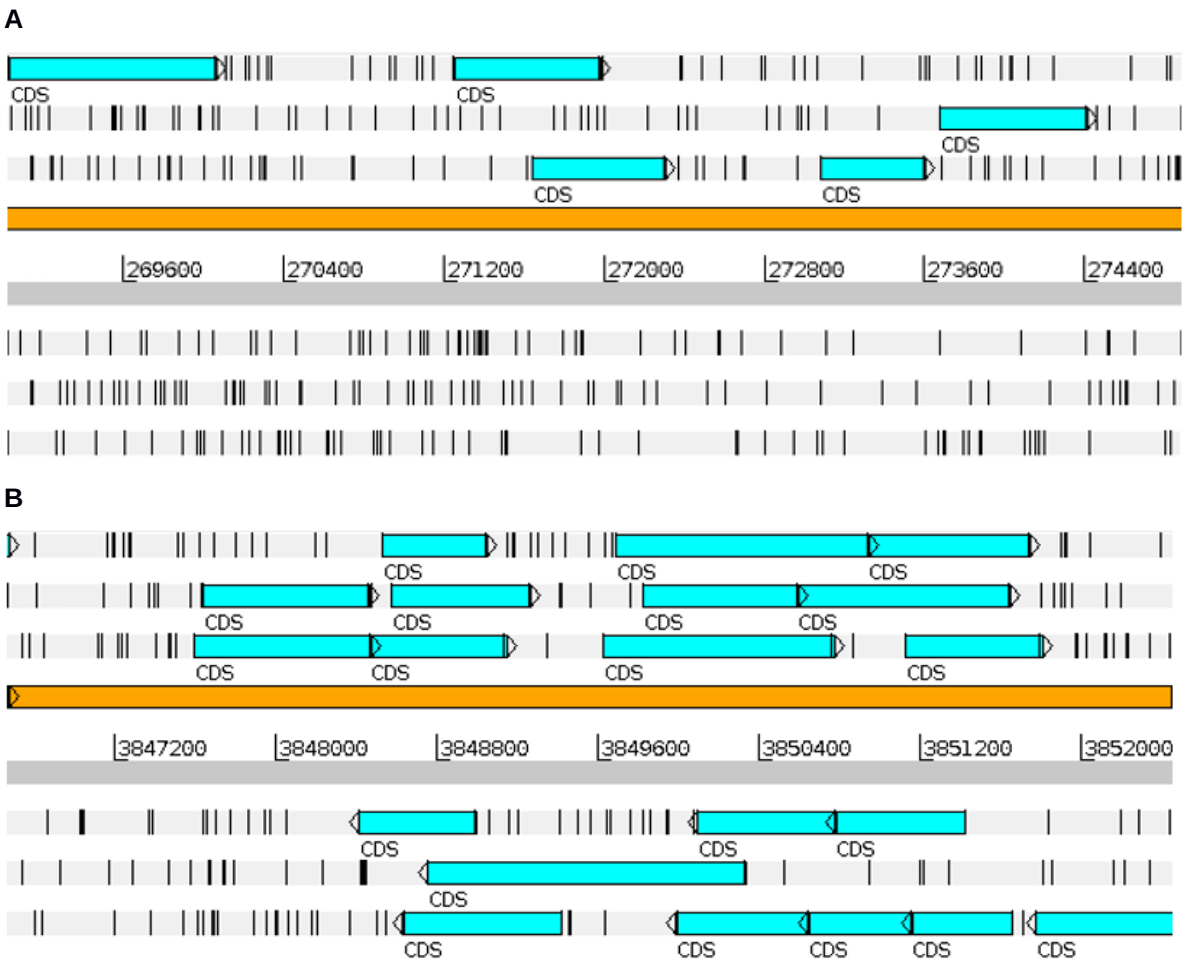

Figura 25: Captura de pantalla del software Artemis en donde se indican en cian los ORFs con una longitud mayor a 166 codones o 500 nucleótidos La línea naranja representa un contig, encima están los tres marcos de lectura de una hebra y debajo los tres de la hebra complementaria. En (A) se muestra una región del genoma asociada a genes clásicos y en (B) en una región del genoma de elevada orfogenicidad.

Si se toman del PCA de Y486 (figura 22) las secuencias correspondientes a la nube de puntos superior, o región ‘core’ y la de la nube inferior, región ‘orfogénica’ y se realiza un conteo del número de ORFs mayores o iguales a 166 codones se observa una gran diferencia en ambas poblaciones. En las regiones del core, el número promedio de ORFs cada 10 Kb es de 6.6, mientras que en la región orfogénica el promedio es de 23.8 (figura 26).

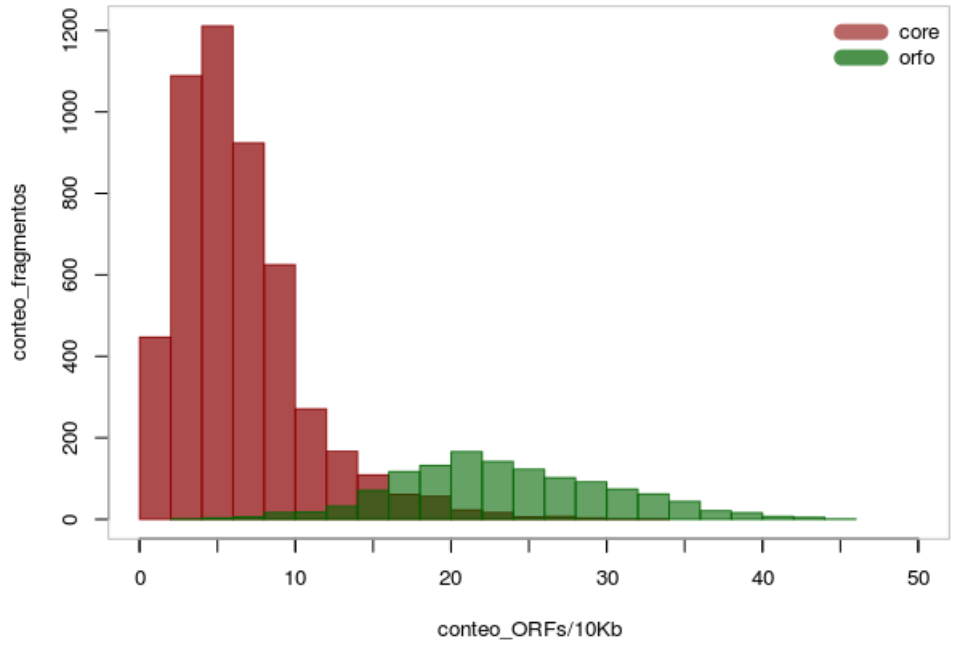

Figura 26.- Histograma del conteo de ORFs mayores o iguales a 166 codones de longitud en *T.vivax* Y486 en la región del genoma asociada al core (en rojo) y a la región orfogénica (en verde).

Con la finalidad de estudiar estas propiedades en el genoma se desarrollaron los índices descritos anteriormente en “Materiales y métodos”.

Para visualizar la densidad de ORFs en los contigs sobre el PCA se utilizó el índice de orfogenicidad (orfindex). Este como ya fue definido, indica en cuantos ORFs (mayores al umbral) está cada nucleótido y se calcula como la sumatoria de la longitud de todos los marcos abiertos de lectura de tamaño mayor o igual a 166 codones, normalizado por el largo de la secuencia. Un valor de 3 indica que cada nucleótido pertenece (en promedio) a 3 ORFs mayores al umbral, por lo tanto es una medida del grado de solapamiento. El rango de posibles valores de este índice se extiende entre 0, en caso de que no exista ningún ORF mayor al umbral y 6 si toda la secuencia analizada está compuesta por ORFs largos en los seis marcos.

En los gráficos siguientes se añade a los valores de los dos primeros componentes del PCA el valor de orfindex representado por el color de cada punto. En las tres cepas es posible observar que los valores de orfogenicidad son más altos en la región que es separada por el PCA, observándose más claramente en Y486 (figuras 27, 28 y A1).

Este resultado resalta las características distintivas de la región orfogénica y que las extensas regiones con marcos abiertos de lectura solapados observadas a simple vista, pueden ser estimadas en forma más precisa a través de métodos cuantitativos.

En Y486 puede observarse que las secuencias identificadas con el espacio orfogénico tienen valores para el índice de orfogenicidad unas tres veces mayor que las secuencias del espacio genómico correspondiente al core (figura 27). Las secuencias del core presentan un valor promedio

de alrededor de 0.5, que puede interpretarse como un marco abierto a lo largo de la mitad de la secuencia (color verde), mientras que las secuencias orfogénicas muchas de ellas presentan valores superiores a 2 (color naranja a rojo oscuro), lo cual podría interpretarse como zonas con dos marcos abiertos de lectura largos solapados a lo largo de toda secuencia.

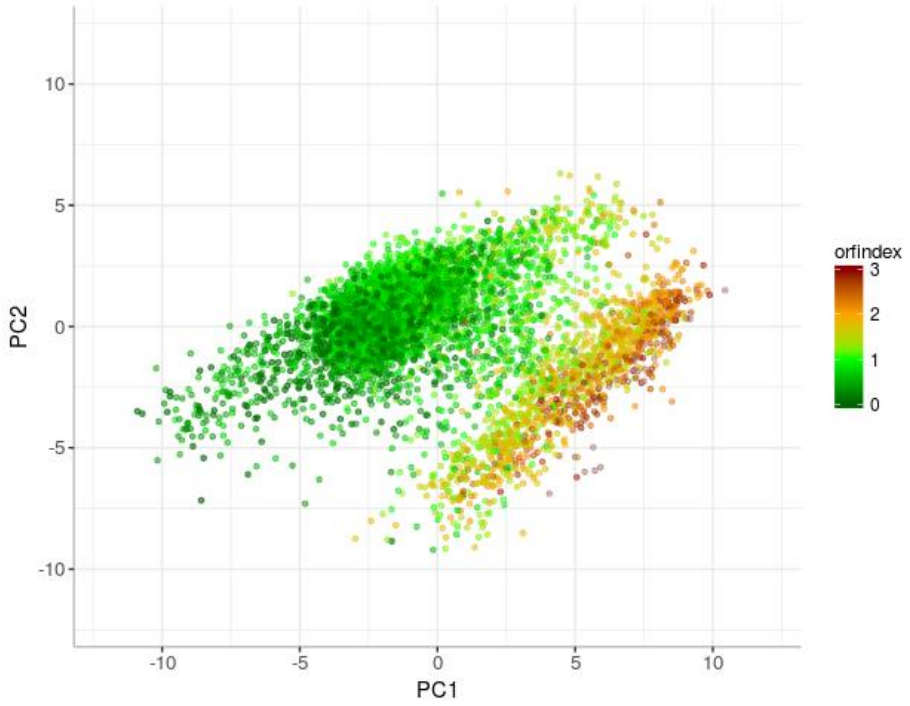

Figura 27.- Gráfica del PCA de frecuencia de trinucleótidos en *T.vivax* Y486. Cada punto representa una secuencia de entre 3 a 5 Kbp y la escala de color refleja el valor del índice de orfogenicidad (orfindex).

En las cepas MT1 y Liem (figuras 29 y A1) se observa un comportamiento similar de ambos tipos de secuencias a lo descrito para Y486 (figura 27). A pesar de que la población de secuencias que conforman el espacio orfogénico es mucho menor, estas presentan características de orfogenicidad similares, demostrando que la definición del índice de orfogenicidad es útil para discriminar entre ambos espacios genómicos aunque las diferencias entre ambas poblaciones de secuencias sean más difíciles de determinar si se considera solamente la distribución de la nube de puntos del PCA.

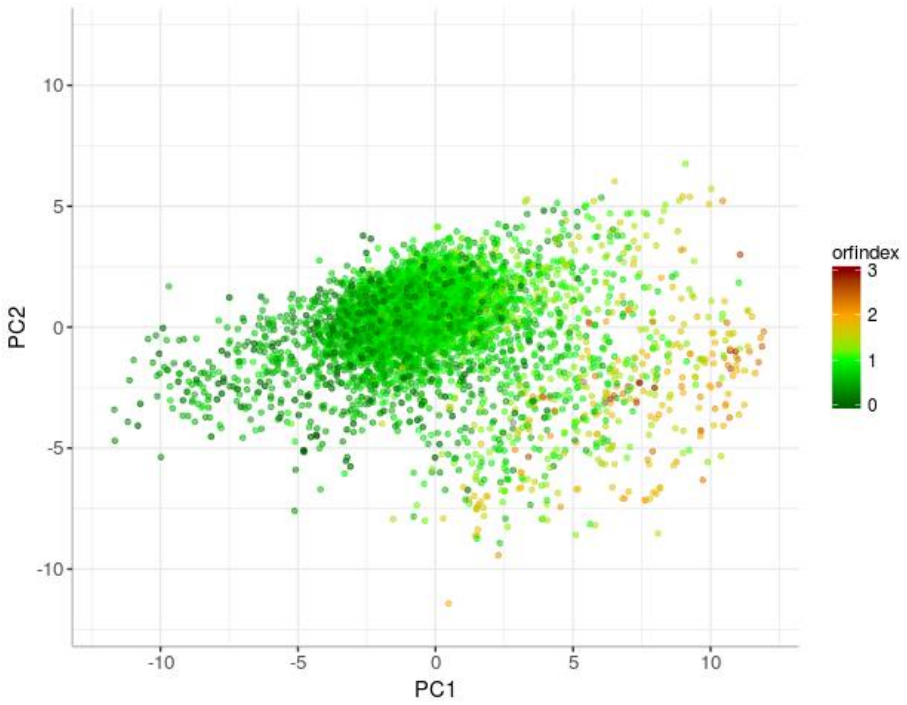

Figura 28.- Gráfica del PCA de frecuencia de trinucleótidos en *T.vivax* MT1. Si bien en el PCA la nube de puntos no separa de forma clara los compartimentos genómicos, los valores del índice de orfogenicidad los distingue claramente

Fuentes de variación en los componentes principales

Ante la observación de que ciertas regiones del genoma poseen elevados niveles de orfogenicidad, surge la interrogante sobre cuáles son las características de secuencia que pueden explicarlo.

Uno de las características de composición nucleotídica típicamente estudiada es el contenido GC, que puede influir en la presencia o ausencia de ORFs en las secuencias. Esto es debido a que los codones de terminación (TAA, TAG y TGA) son pobres en GC. En consecuencia es de esperar que las secuencias ricas en GC contengan menos stops y por tanto mayor orfogenicidad. La pregunta que surge inmediatamente es si los distintos niveles de orfogenicidad observados en estos genomas se deben simplemente a la distinta riqueza de GC.

Otro de los factores a estudiar es la distribución de codones stop en los contigs; si el número de stops observados se encuentra relacionado con el contenido GC o si existen otros factores que están influyendo en el elevado número de ORFs largos. Con la idea de buscar explicaciones a estos fenómenos observados es que se desarrollaron los índices descritos en "Materiales y métodos".

Además es importante estudiar el aporte de las variables originales a la varianza, para analizar cómo es su aporte a los componentes En la figura 29 se pueden realizar varias observaciones; en primer lugar se observa claramente que los tripletes con alto contenido GC se alinean a lo largo del

eje horizontal correspondiente al primer componente. Es decir la discriminación de secuencias a lo largo de este eje se encuentra relacionado con el contenido GC de las secuencias.

Asignando a cada punto del PCA los valores del contenido GC puede observarse claramente como aumentan estos valores a medida que aumenta el valor de PC1 (figura 29).

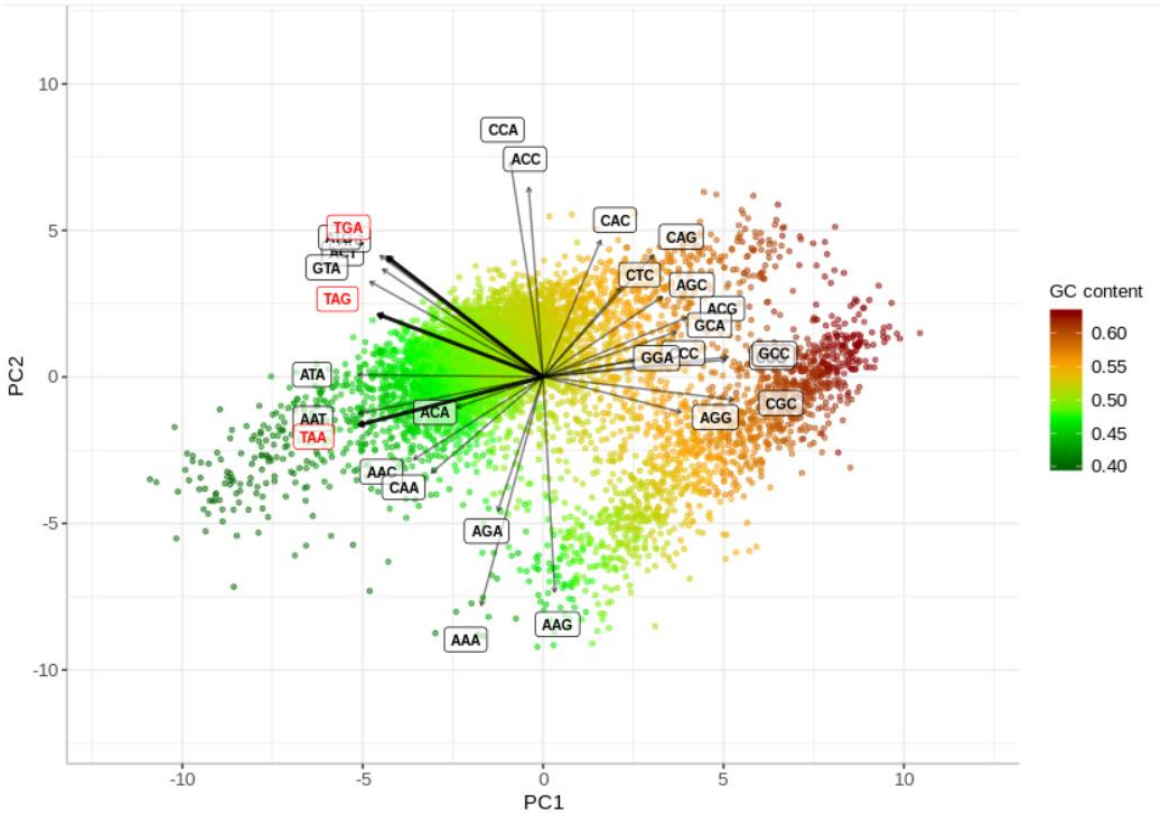

Figura 29.- PCA de la frecuencia de trinucleótidos en Y486, coloreado según el contenido GC. Las flechas indican el aporte de los valores originales a la varianza. En rojo se señalan los codones stop.

Para visualizar mejor la correlación entre valor de PC1 y el contenido GC se graficaron ambos valores en la figura 30. La correlación no solo es muy alta, sino que además pueden identificarse los dos espacios genómicos con claridad ya que presentan correlaciones con pendientes y en rangos diferentes. En la gráfica se indican las secuencias del core (en rojo) y las secuencias orfogénicas (en verde). Para un mismo valor de PC1 las secuencias del core tienen un contenido GC mayor que las secuencias orfogénicas. En resumen, a pesar de que el espacio orfogénico está ligeramente corrido hacia mayores valores de GC, esta no parece ser la explicación de su mayor capacidad de presentar marcos abierto de lectura.

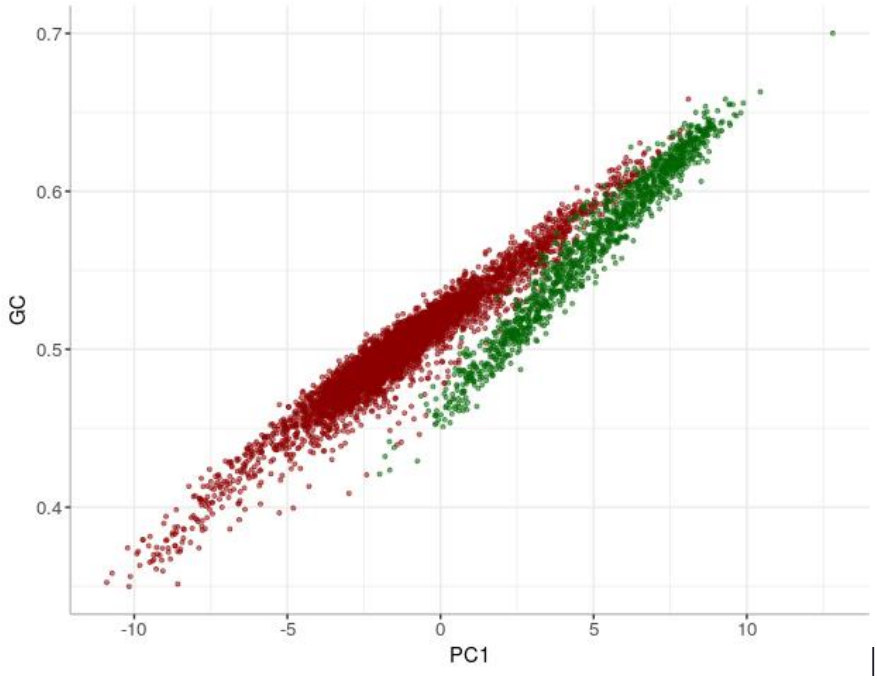

Figura 30.- Gráfica del contenido GC y el valor de PC1 para cada fragmento de contig de Y486. En rojo se señalan las secuencias del core, y en verde las secuencias del espacio orfogénico.

El aporte de las variables originales al segundo componente es algo más difícil de observar. De la figura 29 se desprende que los tripletes (y sus reversos complementarios) que más aportan a la separación en el eje de las ordenadas son AAG/CTT, AAA/TTT, AGA/TCT, cuya frecuencia es mayor en los contigs que presentan menores valores de PC2. Si en estos tripletes renombramos a todas las purinas como R y a las pirimidinas como Y, todos ellos serían de la forma RRR/YYY.

En cambio los tripletes que aumentan el valor de PC2 son CCA/TGG, ACC/GGT y CAC/GTG los cuales también pueden representarse como YYR/YRR, RYY/RRY y YRY/RYR.

Con estas observaciones puede intuirse que el segundo componente del PCA parece discriminar las secuencias genómicas según el nivel de orden o desorden que presenten la sucesión de purinas y pirimidinas en los tripletes.

Para corroborar si el orden en la sucesión de purinas y pirimidinas tiene un aporte importante al segundo componente principal, se realizó la sumatoria de todos los tractos de pirimidinas y purinas de una extensión mayor o igual a cuatro. Se obtuvo el valor de la proporción que conforman estos tractos en el total de las secuencias y se observó su distribución en el PCA de Y486 (figura 31).

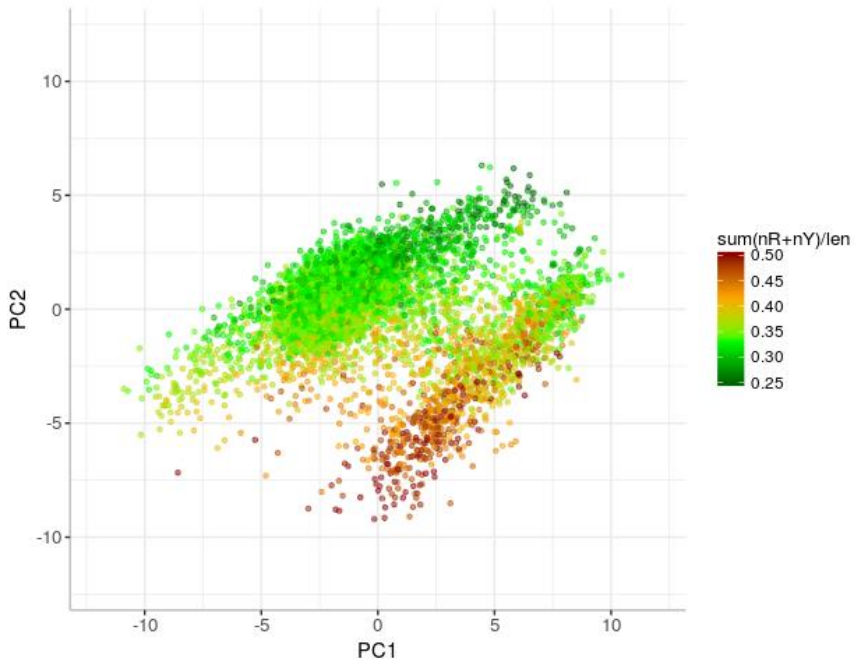

Figura 31.- PCA de las secuencias de *T.vivax* Y486 donde la escala de color representa la sumatoria total de los tractos donde se suceden cuatro o más purinas o pirimidinas.

En la figura 32 se grafican los valores de PC2 con los resultados de la proporción de tractos de purinas y pirimidinas en las secuencias y puede observarse una elevada correlación. Los puntos de color rojo corresponden a secuencias del core, y los de color verde con secuencias del espacio orfogénico.

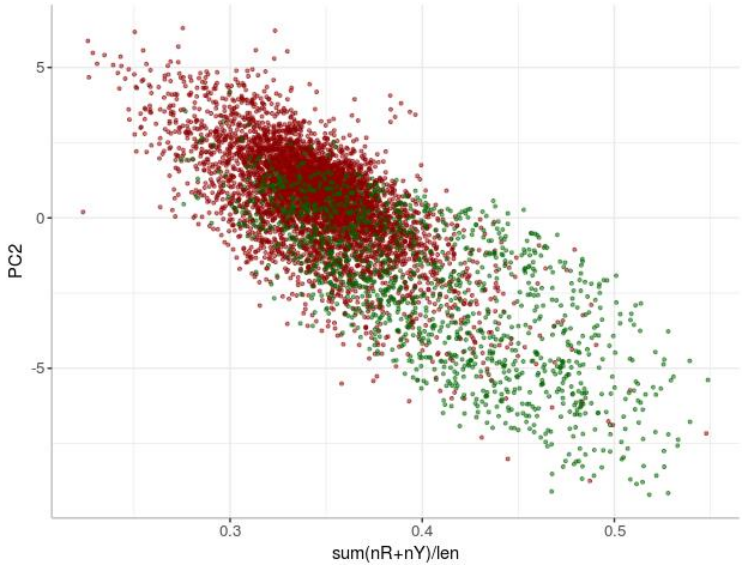

Figura 32.- Gráfica para las secuencias de Y486, de la proporción del genoma con una sucesión de cuatro o más purinas o pirimidinas y PC2. En rojo las secuencias del core y en verde las del espacio orfogénico.

Los codones de terminación están sub-representados en el espacio orfogénico

El siguiente paso consistió en investigar si la orfogenicidad está relacionada con la escasez de codones stop para lo cual se estimó la probabilidad de encontrar un codón stop al azar considerando la composición nucleotídica de una secuencia.

En concreto se calcula la probabilidad de encontrar las secuencias TAG, TAA y TGA dada la frecuencia de bases presentes en cada fragmento de contig (asumiendo independencia y distribución aleatoria de nucleótidos) y luego se compara este valor con los conteos reales de codones stop observados en ambas hebras normalizando por el largo de la secuencia.

Combinando estos dos valores se obtiene un índice que refleja sobreabundancia o escasez de codones de terminación. Este índice fue denominado stopOE y es una medida sobre la potencialidad orfogénica de las diferentes regiones del genoma.

Valores altos de stopOE, cercanos a uno, indican que la proporción de codones observados y esperados por azar es similar; los valores más bajos indican que el conteo de stops es menor al esperado por la frecuencia nucleotídica de la secuencia.

Como se observa en las figuras 33 (y en A2 del Anexo) las regiones orfogénicas poseen valores de stopOE menores que las regiones del core de genes, indicando que en estas regiones existen menos codones stop a los que se espera por azar, y este índice distingue claramente ambas regiones genómicas. Concluimos que la escasez de codones stop (en relación a lo esperado) es uno de los factores que contribuyen a la orfogenicidad.

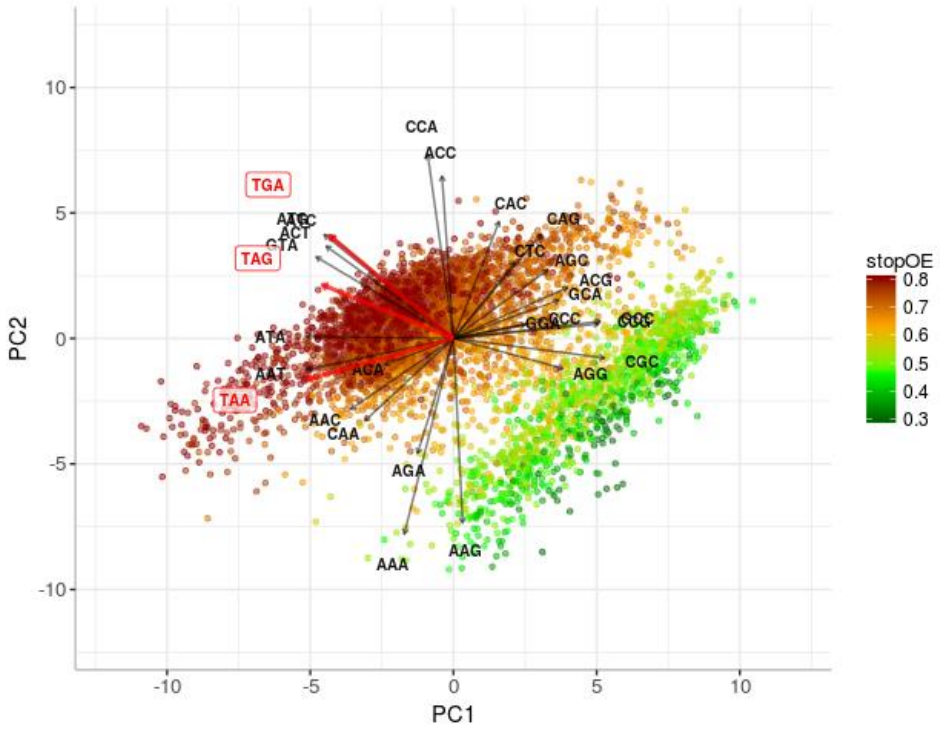

Figura 33.- Gráfica del PCA en *T.vivax* Y486. Puede observarse que la nube de puntos inferior que representa a la región orfogénica presenta valores de stopOE bajos lo que indica que se esperan por azar más codones stop de los que efectivamente se observan. Las flechas indican el aporte de los valores originales a la varianza. En rojo se señalan los codones stop.

Distribuciones esperadas y observadas de ORFs

Considerando que las regiones orfogénicas son pobres en codones de terminación, ¿es esto suficiente para explicar su gran capacidad codificante? ¿O hay otros factores en juego? Por ejemplo la distribución espacial de los codones de terminación puede afectar el largo de los ORFs. Se realizaron diferentes análisis con el fin de comparar los resultados de orfogenicidad obtenidos a través de los diferentes índices y las mediciones del número real de ORFs observados.

Para ello se estimó el número esperado de ORFs mayores o iguales a 166 codones de longitud utilizando como parámetro de probabilidad de encontrar codones stop basado en la frecuencia nucleotídica. Como se puede observar en la figura 34A (y A3 del Anexo), el número de ORFs estimado bajo estas condiciones es cercano a 0.5 por cada 10 Kb y presenta poca diferencia entre las regiones core y orfogénica, reafirmando la noción de que la diferencia de orfogenicidad no está asociada al contenido GC.

En la figura 34B se estima el número de ORFs (>166 codones), pero se utiliza como medida de la probabilidad de encontrar un codón stop la frecuencia de stops observada. En este caso el número de ORFs esperados por cada 10 Kb es claramente mayor en la región orfogénica donde se esperan

entre 5 a 10 ORFs, mientras que en la región del core se esperan entre 0 y 5.

En la figura 34C se muestran los resultados del conteo real de ORFs mayores o iguales a 166 codones de longitud. De la figura 34B se desprende que al utilizar la frecuencia de stops observados para estimar la distribución de los largos de ORFs, los resultados se aproximan mejor para describir lo observado en ambos regiones. Sin embargo el número de ORFs esperado en ambas regiones está por debajo del realmente observado (figura 34C), siendo está deficiencia especialmente marcada en la región orfogénica.

Estas observaciones permiten concluir que el número de ORFs observado en la región orfogénica no pueden explicarse como una diferencia en el contenido GC. La frecuencia de codones stops en estas secuencias explica mejor la diferencia en la orfogenicidad. Sin embargo esta deficiencia de codones stops en el espacio orfogénico no es la única causa, dado que cuando se usa la frecuencia real de stops (P2, ver "Materiales y métodos") como parámetro de la distribución geométrica se sigue subestimando en forma severa el número esperado de ORFs largos en relación a lo que realmente se observa en estas secuencias.

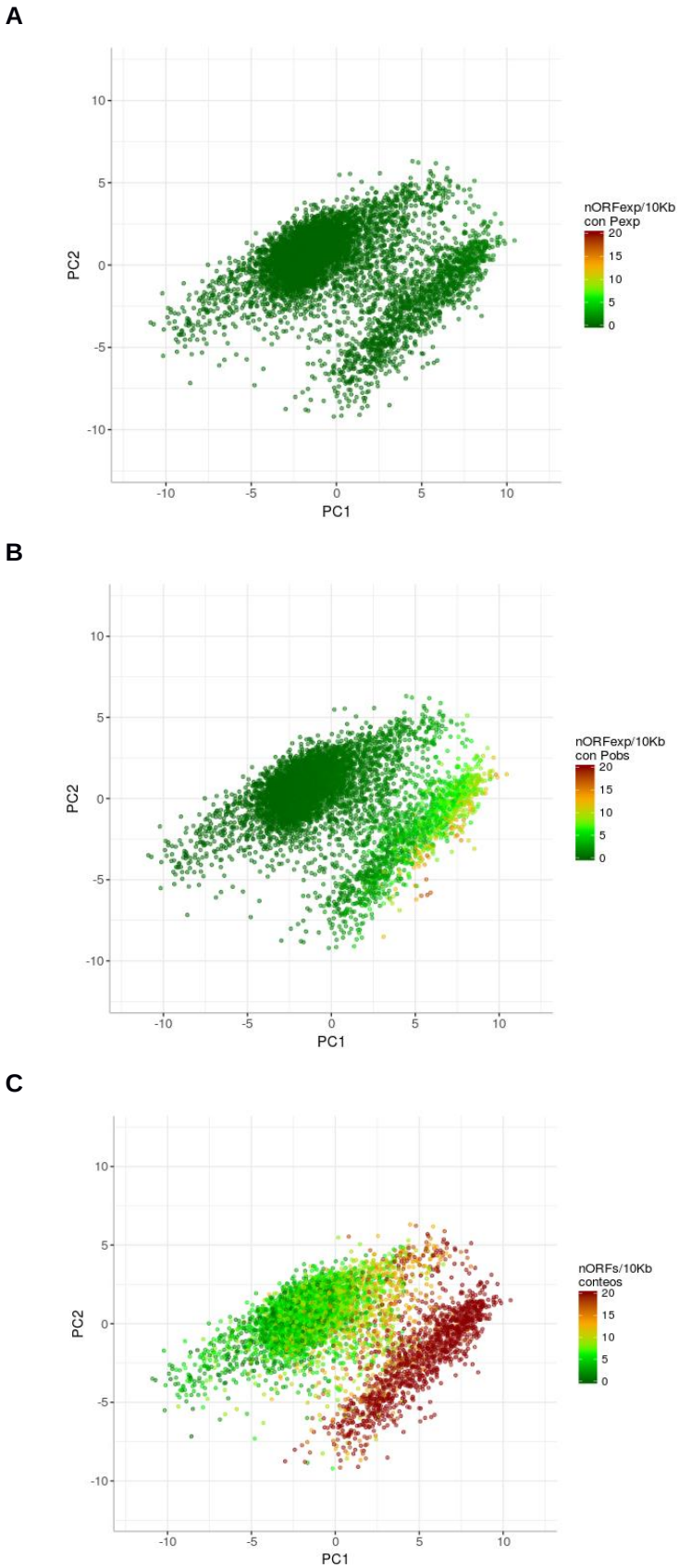

Figura 34.- Gráficas con el número de ORFs de 166 o más codones esperados por 10 Kb. En la (A) la estimación fue realizada utilizando la probabilidad de encontrar stops basándose en la frecuencia nucleotídica. En (B) utilizando la frecuencia de stops observada, y en (C) se muestra el conteo real de estos ORFs. En todas las gráficas se utilizó la misma escala de color con la finalidad de ilustrar las diferencias en los resultados obtenidos cuando se emplean diferentes estrategias.

Reducción del espacio orfogénico en las cepas americanas

En las figuras 23 y 24 puede observarse que la nube de puntos correspondiente a la región orfogénica de las cepas americanas MT1 y Liem-176 es mucho menor que la observada en la cepa africana Y486 (figura 21).

La aparente ausencia de buena parte de la región orfogénica fue analizada utilizando datos de secuenciado genómico de MT1 y Liem y además se corroboró su ausencia utilizando datos de secuenciado de RNA de Liem.

Utilizando bowtie2 se mapearon los reads del secuenciado de DNAseq de las cepas americanas sobre una referencia de Y486. Este mapeo cruzado entre cepas americanas y la africana tiene la finalidad determinar si las observaciones realizadas sobre la reducción del espacio orfogénico son correctas. Si realmente estas regiones se han visto reducidas, los reads de las cepas americanas mapearan con una menor cobertura o directamente no mapearan sobre muchos de los fragmentos de secuencias de la cepa africana asociados a regiones orfogénicas.

Utilizando las herramientas de samtools, y el programa mpileup se estimó el porcentaje de cobertura total para cada fragmento. Estos datos fueron utilizados para asignar un valor de cobertura a cada punto, o sea a cada fragmento de contig. Los valores de cobertura se extienden entre 0 y 1. El 0 (puntos en color gris) corresponde a secuencias sin reads, y entre 0.01 (verde oscuro) que corresponde a un 10% de cobertura a 100% (rojo oscuro), fragmento cubierto en su totalidad por reads de MT1.

En el gráfico de la figura 35 puede observarse claramente que la región de Y486 asociada al core se encuentra cubierta en su totalidad por los reads de DNAseq de MT1, por lo cual estas secuencias también se encuentra en estas cepa.

Sin embargo en la región orfogénica puede observarse que la cobertura es muy escasa, incluso un gran cantidad de fragmentos en donde no mapea ningún read de MT1, es decir estas secuencias se han perdido en la cepa americana (figura 35).

Debido a las similitudes entre las cepas americanas los resultados obtenidos con Liem-176 son prácticamente idénticos y no se muestran las gráficas.

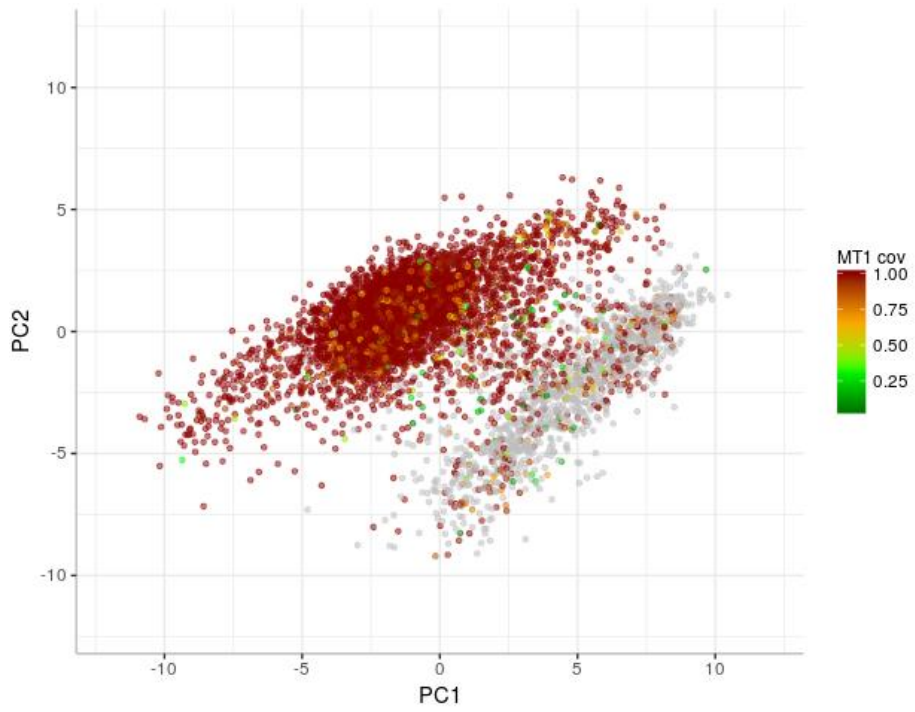

Figura 35.- PCA de *T.vivax* Y486 indicando la cobertura de los reads de DNAseq de MT1 sobre los fragmentos de 3-5 kb de Y486. En rojo oscuro se encuentran las secuencias con una cobertura completa, en amarillos una cobertura del 50% y en verde las secuencias de muy baja cobertura. En gris se señalan las secuencias que no reciben ningún read procedente de MT1.

Con estos resultados puede aseverarse sin lugar a dudas que las cepas americanas han visto reducido su espacio orfogénico respecto a la variante africana.

También es de interés estudiar si en las cepas americanas han surgido secuencias de este espacio genómico que no estén presentes en la cepa africana. Para corroborar esto se mapearon reads de Y486 sobre el ensamblado genómico de MT1.

En la figura 36 puede observarse que a pesar de la enorme reducción del espacio orfogénico en MT1 también se pueden encontrar secuencias que se ensamblaron por los reads de MT1 pero que no son mapeadas por ninguno de Y486.

Es interesante notar que a pesar de que se redujo la diversidad del espacio orfogénico, también surgieron nuevas secuencias que se encuentran presentes exclusivamente en las cepas americanas pero no que no cuentan con un equivalente en su contraparte africana.

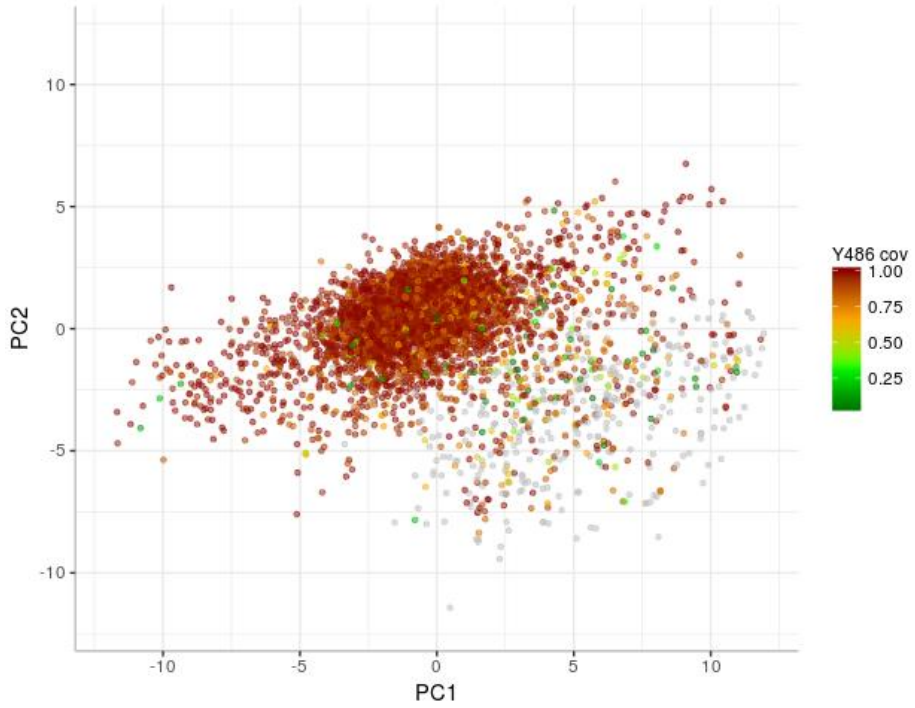

Figura 36.- PCA de *T.vivax* MT1 indicando la cobertura de los reads de DNAseq de Y486 sobre los fragmentos de MT1. En rojo oscuro se encuentran las secuencias con una cobertura completa, en amarillos una cobertura del 50% y en verde las secuencias de muy baja cobertura. En gris se señalan las secuencias que no reciben ningún read procedente de Y486. Nótese que a pesar de la disminución del espacio genómico orfogénico en MT1, aun así surgieron secuencias nuevas que no se encuentran presentes en Y486.

Exploración y visualización web de los resultados

Con la finalidad de facilitar la exploración de los resultados del análisis de la orfogenicidad del genoma nuclear en *T.vivax*, comparar los diferentes parámetros utilizados y explorar la anotación genómica, se desarrolló una página web con gráficos interactivos donde se muestran los principales resultados obtenidos.

La página web fue desarrollada utilizando la librería D3 de JavaScript cuyo fin es la representación de datos y creación de gráficos fácilmente manejables utilizando un canvas SVG que permite dibujar figuras y utilizar escalas de color que responden de manera interactiva con el usuario.

Los resultados de los análisis de datos fueron condensados en una base de datos en formato JSON que permiten cargar la información de manera asíncrona, y facilitan el procesado de los datos utilizando parsers estándar de JavaScript, además de ser fácilmente extensible con mínimas

modificaciones en los scripts.

En la página web se pueden cargar los resultados de los PCA para las tres cepas estudiadas y se dispone de un menú desplegable que permite colorear, los puntos de la gráfica según la variable de interés.

Las variables mostradas en la página son el índice de orfogenicidad, la proporción de codones stops observados sobre los esperados, el contenido GC, la proporción de secuencia cubierta por reads de otras cepas, es decir las zonas conservadas entre cepas, la profundidad de cobertura de los reads para cada cepa secuenciada, número de ORFs contados y proporción de la secuencia cubierta por tracto de purinas o pirimidinas.

Para facilitar la visualización de la nube de puntos se dispone de un slider para valores mínimos y máximos que permite ocultar los puntos por fuera del rango de valores de la variable seleccionada.

Los genes anotados para cada cepa fueron mapeados sobre las secuencias fragmentadas, y en caso de que un gen se encuentre en más de un 50% en un fragmento se asigna la anotación del gen a dicho fragmento. Esta anotación se encuentra disponible para cada punto de las gráficas y se visualiza, junto a otras estadísticas, como un tooltip que se despliega al mantener el cursor del ratón sobre un punto de la gráfica.

Además la página cuenta con una caja de búsqueda que permite buscar palabras claves en la anotación y selecciona, agrandando y con borde negro, los puntos que comparte anotación con la expresión buscada.

Las opciones mostradas como checkbox que se encuentran debajo del campo de búsqueda permiten ocultar o mostrar los puntos seleccionados.

Cada punto de la gráfica puede clickearse para ser cargado o descargado de una tabla que se genera bajo los controles de la gráfica y que aparte de contener la anotación, permite descargar el archivo fasta de la secuencia.

La página web se encuentra disponible en bioinformatica.fcien.edu.uy/vivax .

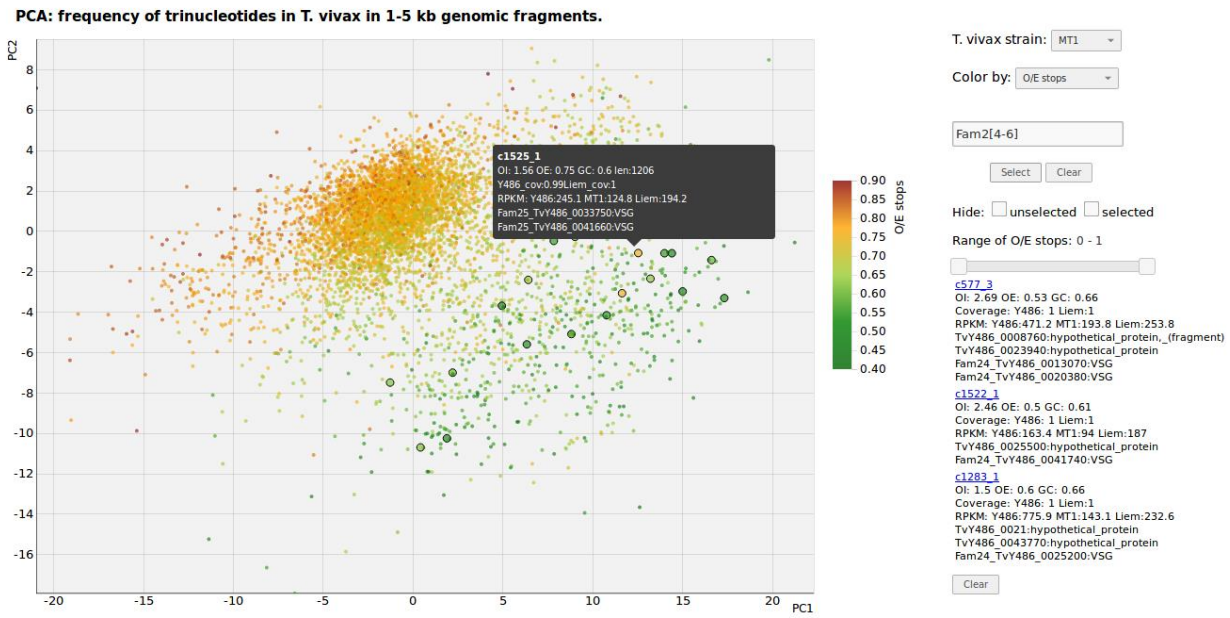

Figura 37.- Captura de pantalla del sitio web desarrollado. En este caso ilustrando los resultados del PCA de MT1 coloreados por la frecuencia de stops observada sobre esperada. Además indica con círculos de mayor tamaño en que secuencias encontró información asociada a la anotación de la expresión Fam2[4-6] . Esta búsqueda da como resultado las familias 24 a 26 de VSG.

Conclusiones

Genoma mitocondrial de *T.vivax*

Los tripanosomas africanos enfrentan dos ambientes muy diferentes durante su ciclo de vida y esto se ve reflejado principalmente en su metabolismo energético.

Cuando el parásito se encuentra en su hospedero mamífero la mitocondria se encuentra en un nivel de actividad mínimo, sin un ciclo de Krebs funcional. La generación de energía, o sea la producción de ATP se produce por fosforilación a nivel de sustrato a través de la vía glicolítica, utilizando la glucosa que provee el hospedero.

Sin embargo en el estadio procíclico, en el insecto, el parásito se encuentra en un medio muy pobre, por lo cual debe poner en marcha la maquinaria mitocondrial para generar energía a través de la fosforilación oxidativa. De este modo puede realizar parte del ciclo de vida en el sistema digestivo del vector, entrar en una fase proliferativa y prepararse para ser transmitido a un hospedero mamífero.

Las cepas americanas de *T.vivax* estudiadas se encuentran estrechamente relacionadas con cepas de África Occidental, de la cual la cepa Y486 también estudiada, es representativa (Garcia et al. 2014).

La principal diferencia entre las cepas americanas y africanas de *T.vivax* se encuentra en cómo se desarrollan las etapas de su ciclo de vida. En África el parásito es transmitido por insectos del género *Glossina*, y realiza en ellos parte del ciclo de vida. En América es transmitido por tabánidos a los cuales el parásito no se encuentra adaptado y en donde no realiza su etapa procíclica.

Al no transitar por el estadio procíclico en las cepas americanas, y la mitocondria no ser utilizada para realizar la fosforilación oxidativa en el estadio sanguíneo, no existe presión selectiva para mantener un genoma mitocondrial intacto.

Se ha observado que variantes de tripanosomas africanos similares a *T.brucei* que han migrado a regiones fuera del área de distribución de la mosca tsé-tsé como *T.evansi* y *T.equiperdum*, han perdido total o parcialmente el genoma mitocondrial.

Se debe destacar que en estas variantes de *T.brucei* todos los genes mitocondriales no son funcionales o han sido eliminados, y en algunos casos reemplazados en sus funciones por genes nucleares, como es el caso de la ATPasa-A6 cuya función es complementada por la subunidad gamma de la ATPasa nuclear (Dean et al. 2013).

Se ha postulado que mutaciones compensatorias en la ATPasa gamma nuclear eliminan la necesidad de la ATPasa-A6 mitocondrial (Jensen et al. 2008) y que esta es una precondición para la pérdida total del maxicírculo (Lun et al. 2010).

En las cepas de *T.vivax* americanas no observamos esta mutación y la ATPasa-A6 continúa siendo funcional en el genoma mitocondrial, por lo cual puede considerarse que se encuentra en una etapa intermedia o en un camino evolutivo diferente.

Regiones orfogénicas del genoma de *T.vivax*

El uso de métodos de análisis multivariados como el análisis de componentes principales puede ser utilizado para estudiar la heterogeneidad de secuencias genómicas tomando en cuenta características como la frecuencia de codones o de k-mers nucleotídicos (Grantham et al. 1980).

Este tipo de análisis permite identificar regiones o compartimentos del genoma que presentan características en cuanto a su frecuencia nucleotídica, y está relacionado a diferentes tipos de compartimentos genómicos.

Un PCA de frecuencias de trinucleótidos del genoma de la cepa africana Y486 permite observar que el genoma nuclear puede dividirse en dos grandes compartimentos genómicos, uno asociado a genes del core y otro asociado a regiones orfogénicas, donde se encuentran los genes involucrados en la variación antigénica (figuras 21 y 22). Esto es más difícil de visualizar en los genomas de cepas americanas por la reducción del espacio orfogénico en estas cepas (figura 23 y 24).

Se desarrollaron varios índices con la finalidad de cuantificar la orfogenicidad del genoma de *T.vivax* y poder estimar la contribución de los diferentes factores.

El índice 'orfindex' es una medida experimental de orfogenicidad, ya que se basa en datos obtenidos directamente de los ORFs observados, y muestra valores elevados de orfogenicidad en la región genómica asociada a la variación antigénica, en comparación con la región genómica del core (figuras 27 y 28). Este resultado es compatible con la abundancia de ORFs solapados en la región orfogénica.

Se analizaron las causas de la división del genoma en estos dos grandes compartimentos y cuáles son los elementos que contribuyen a los dos primeros componentes del PCA. La característica que más aporte hace al primer componente del PCA es el contenido GC, el cual es levemente más elevado en la región orfogénica que en el core (figura 29).

La importancia de segundo componente en la concentración diferente de codones stops en ambos compartimentos genómicos puede apreciarse al estudiar el aporte de las variables originales a los componentes (figura 29). Aquí se observa una elevada correlación entre los tractos de purinas y pirimidinas y el segundo componente (figura 32). El orden en la sucesión de purinas y pirimidinas es lo que determina la separación en este eje. Las regiones orfogénicas resultan más 'ordenadas', es decir se observan con más frecuencia tractos de purinas o de pirimidinas consecutivas, mientras que

las regiones del core son las que presentan un mayor ‘desorden’ (figura 31). Teniendo en cuenta que los tres codones de terminación están conformados por sucesiones que contienen una pirimidina y dos purinas (TAA, TAG y TGA), esta característica del DNA de las regiones orfogénicas probablemente sea el factor determinante (superior al aporte del contenido GC) en la concentración diferente de codones stops.

El otro de los índices que permite comprender los diferentes grados de orfogenicidad observados, es la relación entre el número de codones stop esperados (P1, ver “Materiales y métodos”) y la frecuencia real de codones stop (P2). Este índice al que llamamos ‘stopOE’ (figura 33), muestra que la región orfogénica se diferencia claramente de la región del core, por tener un número de stops mucho menor al esperado según la composición de bases.

Este resultado es interesante, ya que se ha observado que en la mayoría de los organismos ocurren naturalmente codones stop fuera de marco en las regiones codificantes y se les atribuye un rol de ahorro energético para evitar el costo metabólico que implica una traducción de un péptido con frameshifts (Tse et al. 2010).

Sin embargo la diferente concentración de codones de teminación no parece ser el único factor que determina la capacidad codificante incrementada del espacio orfogénico. Otro índice desarrollado permite ver con claridad este aspecto. El mismo predice el número esperado de ORFs mayores o iguales a 166 codones dado una determinada frecuencia de codones de terminación. Es interesante notar que incluso cuando se utiliza como parámetro la frecuencia observada de codones de terminación, se subestima el número de ORFs que son realmente observados en la región orfogenica. Esto indicaría que los codones stop no se encuentran distribuidos al azar, lo que refleja una presión evolutiva para mantener esta región genómica con grandes regiones de ORFs en múltiples marcos.

La extrema orfogenicidad observada en de *T.vivax* puede deberse a varios motivos. Por un lado se debe tener en cuenta que la optimización energética no se encuentra entre las prioridades del parásito, por lo cual podría carecer de presión selectiva para mantener stops fuera de marco.

Por otro lado la máxima orfogenicidad se encuentra en la región genómica asociada a los genes que generan las proteínas de superficie VSG, responsables de la variación antigénica. Dado que el gen activo de la VSG tiene una elevadísima tasa de transcripción, es razonable pensar que la eficiencia traduccional prime sobre la exactitud de la traducción, mientras no afecte la viabilidad del parásito.

Además es frecuente que las secuencias de los genes VSG se generen por mosaicismo y la presencia de codones stop podría generar secuencias de proteínas truncas.

Reducción del espacio orfogénico

En las cepas americanas de *T. vivax* (MT1 y Liem) puede observarse una clara reducción del espacio orfogénico del genoma.

Esto puede observarse cuando se realizan mapeos cruzados de reads de una cepa americana, sobre los contigs ensamblados de la cepa africana (figura 35). Ahí queda claramente demostrado como las cepas americanas, tanto MT1 como Liem, cubren por completo las secuencias de contigs africanos en las regiones genómicas asociadas a genes 'clásicos' y sin embargo queda sin cobertura gran parte de la región orfogénica del genoma.

Esta pérdida del repertorio de genes VSG y de otras familias multigénicas asociadas ha sido observada también en otros tripanosomas y parece ser la mayor fuente de variación del tamaño genómico entre cepas de una misma especie (Cross et al, 2014).

En trabajos de genómica comparativa en *T. brucei* se ha observado una enorme variación en el tamaño de los cromosomas homólogos dentro y entre cepas (El-Sayed et al. 2000).

Se ha reportado hasta un 30% de variación en el total del genoma entre diferentes cepas. La mayor parte de esta variación puede ser atribuida a la presencia de VSG en tándem en las regiones subteloméricas (Cross et al, 2014).

Referencias

Akiyoshi, B., & Gull, K. (2013). Evolutionary cell biology of chromosome segregation: Insights from trypanosomes. Open Biology, 3(5).

Allsopp, B., & Newton, S. (1985). Characterization of Trypanosoma (Duttonella) vivax by isoenzyme analysis. International Journal for Parasitology, 15(3), 265-270.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990) Basic local alignment search tool. J. Mol. Biol. 215:403-410.

Andrews S. (2010). FastQC: a quality control tool for high throughput sequence data. Disponible en: www.bioinformatics.babraham.ac.uk/projects/fastqc

Asburner M., et al. (2000) Gene Ontology: tool for the unification of biology. Nat Genet.; 25(1): 25–29.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A., Lesin, V., Pevzner, P. (2012). SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. Journal of Computational Biology, 19(5), 455-477.

Barrett, M. P., Burchmore, R. J. S., Stich, A., Lazzari, J. O., Frasch, A. C., Cazzulo, J. J., & Krishna, S. (2003). The trypanosomiases. The Lancet, 362, 1469–1480.

Becker, M. Aitcheson,N. Byles, E., Wickstead, B., Louis E., Rudenko, G. (2004). Isolation of the repertoire of VSG expression site containing telomeres of Trypanosoma brucei 427 using transformation-associated recombination in yeast. Genome Research, 14(11), 2319-2329.

Berná, L., Rodriguez, M., Chiribao, M. L., Parodi-Talice, A., Pita, S., Rijo, G., Alvarez-Valin, F., Robello, C. (2018). Expanding an expanded genome: Long-read sequencing of Trypanosoma cruzi. Microbial Genomics, 4(5)

Berriman, M., Ghedin, E., Hertz-Fowler, C., Blandin, G., Renauld, H., Bartholomeu, D.C., et al. (2005) The genome of the African trypanosome Trypanosoma brucei. Science 309: 416–422

Bitter, W., Gerrits, H., Kieft, R., and Borst, P. (1998) The role of transferrin-receptor variation in the host range of Trypanosoma brucei. Nature 391: 499–502.

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

Blum, M. L., Down, J. A., Gurnett, A. M., Carrington, M., Turner, M. J., & Wiley, D. C. (1993). A structural motif in the variant surface glycoproteins of Trypanosoma brucei. Nature, 362(6421), 603–9.

Carrington, M., Miller, N., Blum, M., Roditi, I., Wiley, D., and Turner, M. (1991). Variant specific glycoprotein of Trypanosoma brucei consists of two domains each having an independently conserved pattern of cysteine residues. J. Mol. Biol. 221: 823-835

Chevreux et al. (1999) Genome Sequence Assembly Using Trace Signals and Additional Sequence Information Computer Science and Biology: Proceedings of the German Conference on Bioinformatics (GCB) 99, pp. 45-56.

Clayton, C., & Michels, P. (1996). Metabolic compartmentation in African trypanosomes. Parasitology Today, 12(12), 465-471.

Clayton, C., & Shapira, M. (2007). Post-transcriptional regulation of gene expression in trypanosomes and leishmanias. *Molecular and Biochemical Parasitology, 156*(2), 93-101.

Clayton, C. (2014). Networks of gene expression regulation in Trypanosoma brucei. *Molecular and Biochemical Parasitology, 195*(2), 96-106.

Cock PA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B and de Hoon MJL (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics, 25, 1422-1423

Conesa, A., & Götz, S. (2008). Blast2GO: A Comprehensive Suite for Functional Analysis in Plant Genomics. International Journal of Plant Genomics, 2008, 1-12.

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

Cox, F. E. G. (2004). History of sleeping sickness (African trypanosomiasis). Infect Dis Clin North Am, 18(2), 231–245.

Cross, G. A., Kim, H., & Wickstead, B. (2014). Capturing the variant surface glycoprotein repertoire (the VSGnome) of Trypanosoma brucei Lister 427. Molecular and Biochemical Parasitology, 195(1), 59-73.

Dean, S., Gould, M. K., Dewar, C. E., & Schnaufer, A. C. (2013). Single point mutations in ATP synthase compensate for mitochondrial genome loss in trypanosomes. Proceedings of the National Academy of Sciences, 110(36), 14741-14746.

Delcher, A. L.; Kasif, S.; Fleischmann, R. D.; Peterson, J.; White, O.; Salzberg, S. L. (1999). "Alignment of whole genomes". Nucleic Acids Research. 27 (11): 2369–2376

Desquesnes, M. (2004). Livestock Trypanosomoses and their Vectors in Latin America. OIE.

Desquesnes M, Dia M.L. (2003) Trypanosoma vivax: mechanical transmission in cattle by one of the most common African tabanids, Atylotus agrestis. Experimental Parasitology (103) 35-43

Douris, V., Telford, M. J., & Averof, M. (2010). Evidence for Multiple Independent Origins of trans-Splicing in Metazoa. Molecular Biology and Evolution, 27(3), 684–693.

El-Sayed, N. M., Hegde, P., Quackenbush, J., Melville, S. E., & Donelson, J. E. (2000). The African trypanosome genome. *International Journal for Parasitology, 30*(4), 329-345.

El-Sayed, N. M. (2005). Comparative Genomics of Trypanosomatid Parasitic Protozoa. Science, 309(5733), 404–409.

Engstler, M., Pfohl, T., Herminghaus, S., Boshart, M., Wiegertjes, G., Heddergott, N., & Overath, P. (2007). Hydrodynamic Flow-Mediated Protein Sorting on the Cell Surface of Trypanosomes. *Cell, 131*(3), 505-515.

Everitt & Hothorn; (2011); An Introduction to Applied Multivariate Analysis with R; Springer

Fadda, A., Ryten, M., Droll, D., Rojas, F., Färber, V., Haanstra, J. R., Merce, C., Bakker, B., Mathews, K., Clayton, C. (2014). Transcriptome-wide analysis of trypanosome mRNA decay reveals complex degradation kinetics and suggests a role for co-transcriptional degradation in determining mRNA levels. Molecular Microbiology, 94(2), 307-326.

Ferragina, P. Manzini, G. (2000) "Opportunistic data structures with applications." Foundations of Computer Science. Proceedings. 41st Annual Symposium IEEE 2000

Fritz, M., Vanselow, J., Sauer, N., Lamer, S., Goos, C., Siegel, T., Subota, I., Schlosser, A., Carrington, M., Kramer, S. (2015). Novel insights into RNP granules by employing the trypanosomes microtubule skeleton as a molecular sieve. Nucleic Acids Research, 43(16), 8013-8032.

Garcia, H. A., Rodrigues, A. C., Rodrigues, C. M., Bengaly, Z., Minervino, A. H., Riet-Correa, F., . . . Teixeira, M. M. (2014). Microsatellite analysis supports clonal propagation and reduced divergence of Trypanosoma vivax from asymptomatic to fatally infected livestock in South America compared to West Africa. Parasites & Vectors,7(1), 210.

Ghedin, E., Bringaud, F., Peterson, J., Myler, P., Berriman, M., Ivens, A., . . . El-Sayed, N. M. (2004). Gene synteny and evolution of genome architecture in trypanosomatids. Molecular and Biochemical Parasitology, 134(2), 183-191.

Giddings, O.K., Eickhoff, C.S, Smith, T.J., Bryant, L.A., Hoft, D.F. (2006) Anatomical Route of Invasion and Protective Mucosal Immunity in Trypanosoma cruzi Conjunctival Infection. Infection and Immunity. 74 (10), 5549–5560

Glover, L., & Horn, D. (2006). Repression of polymerase I-mediated gene expression at Trypanosoma brucei telomeres. *EMBO Reports, 7*(1), 93-99.

Götz et al. (2008) "High-throughput functional annotation and data mining with the Blast2GO suite", Nucleic Acids Research, vol. 36, pp. 3420-3435.

Grantham, R., Gautier, C., Gouy, M., Mercier, R., & Pavé, A. (1980). Codon catalog usage and the genome hypothesis. Nucleic Acids Research, 8(1), 197-197.

Green, P. (1996). Documentation for PHRAP. Genome Center, University of Washington, Seattle.

Greif, G., Rodriguez, M., Reyna-Bello, A., Robello, C., & Alvarez-Valin, F. (2015). Kinetoplast adaptations in American strains from Trypanosoma vivax. Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis, 773, 69–82.

Günzl A., Bruderer T., Laufer G., Schimanski B., Tu L.C., Chung H.M., Lee P.T., Lee M.G. (2003) RNA Polymerase I Transcribes Procyclin Genes and Variant Surface Glycoprotein Gene Expression Sites in Trypanosoma brucei. Eukaryotic Cell 2 (3) 542–551

Haag, J., O'hUigin, C., & Overath, P. (1998). The molecular phylogeny of trypanosomes: evidence for an early divergence of the Salivaria. Molecular and Biochemical Parasitology, 91(1), 37–49.

Hoft, D.F., Farrar, P.L., Kratz-Owens, K., Shaffer, D. (1996). Gastric Invasion by Trypanosoma cruzi and Induction of Protective Mucosal Immune Responses. Infection and Immunity. 64 (9) 3800-3810.

Husson, Le & Pages; (2010); Exploratory Multivariate Analysis by Example Using R ; CRC Press

Jensen, R. E., Simpson, L., & Englund, P. T. (2008). What happens when Trypanosoma brucei leaves Africa. Trends in Parasitology, 24(10), 428-431.

Kramer, S. (2011). Developmental regulation of gene expression in the absence of transcriptional control: The case of kinetoplastids. Molecular and Biochemical Parasitology, 181(2), 61–72.

Kramer, S., & Carrington, M. (2011). Trans-acting proteins regulating mRNA maturation, stability and translation in trypanosomatids. Trends in Parasitology, 27(1), 23-30.

Langousis, G., & Hill, K. L. (2014). Motility and more: the flagellum of Trypanosoma brucei. Nature Reviews. Microbiology, 12(7), 505–18.

Langmead B. (2013) Introduction to the Burrows-Wheeler Transform and FM Index, Department of Computer Science, JHU

Langmead, B & Salzberg, S. (2012). Fast gapped-read alignment with Bowtie 2. Nature Methods, 9(4), 357–359.

Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). "The Sequence Alignment/Map format and SAMtools". Bioinformatics. 25 (16): 2078–2079.

Lozano R, et al. (2012) Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010. The Lancet 380, 2095-2128

Lukeš, J., Skalický, T., Týč, J., Votýpka, J., & Yurchenko, V. (2014). Evolution of parasitism in kinetoplastid flagellates. *Molecular and Biochemical Parasitology, 195*(2), 115-122.

Lukeš, J. (2010). The Remarkable Mitochondrion of Trypanosomes and Related Flagellates. (W. de Souza, Ed.) (Vol. 17). Berlin, Heidelberg: Springer Berlin Heidelberg.

Lun, Z.-R., Lai, D.-H., Li, F.-J., Lukes, J., & Ayala, F. J. (2010). Trypanosoma brucei: two steps to spread out from Africa. Trends in Parasitology, 26(9), 424–7.

Luo et al. (2012) SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. GigaScience 1:18.

Lythgoe, K. A., Morrison, L. J., Read, A. F., & Barry, J. D. (2007). Parasite-intrinsic factors can explain ordered progression of trypanosome antigenic variation. Proceedings of the National Academy of Sciences, 104(19), 8095-8100.

McKelvey, J. J. (1973). Man against tsetse: Struggle for Africa. Ithaca: Cornell University Press.

Miller, J. R., Koren, S., Sutton, G. (2010) "Assembly algorithms for next-generation sequencing data." Genomics 95(6), 315-27.

Morrison, L. J., Marcello, L., & Mcculloch, R. (2009). Antigenic variation in the African trypanosome: molecular mechanisms and phenotypic complexity. Cellular Microbiology, 11(12), 1724-1734.

Myers et al. (2000) A Whole-Genome Assembly of Drosophila. Science 287 2196-2204

Noe, L., Kucherov,G. YASS: enhancing the sensitivity of DNA similarity search, 2005, Nucleic Acids Research, 33(2):W540-W543.

Nolan D., Jackson D., Biggs M., Brabazon E., Pays, A., van Laethem F., Paturiaux-Hanocq F., Elliot J., Voorheis P., Etienne Pays E., (2000) Characterization of a Novel Alanine-rich Protein Located in Surface Microdomains in Trypanosoma brucei. Journal of Biological Chemistry. 275 (6) 4072-4080

Opperdoes, F. R., & Borst, P. (1977). Localization of nine glycolytic enzymes in a microbody-like organelle inTrypanosoma brucei: The glycosome. FEBS Letters, 80(2), 360-364.

Osório A. L. , Madruga, C. R., Desquesnes, M., Soares, C. O., Raquel, L., & Ribeiro, R. (2008). Trypanosoma (Duttonella) vivax : its biology, epidemiology, pathogenesis, and introduction in the New World - A Review. Mem Inst Oswaldo Cruz, 103, 1–13.

Pagès H, Aboyoun P, Gentleman R and DebRoy S (2017). Biostrings: Efficient manipulation of biological strings. R package version 2.46.0.

Palenchar, J. B., & Bellofatto, V. (2006). Gene transcription in trypanosomes. Molecular and Biochemical Parasitology, 146(2), 135–41.

Pays, E. (1989). Pseudogenes, chimaeric genes and the timing of antigen variation in African trypanosomes. *Trends in Genetics, 5*, 389-391.

Peacock, L., Ferris, V., Sharma, R., Sunter, J., Bailey, M., Carrington, M., & Gibson, W. (2011). Identification of the meiotic life cycle stage of Trypanosoma brucei in the tsetse fly. Proceedings of the

National Academy of Sciences, 108(9), 3671-3676.

Pevzner, A. P., Tang, H., Waterman, M.S. (2001). An Eulerian path approach to DNA fragment assembly. PNAS. 98 (17) 9748-9753

Ray D.S. (1989) Conserved sequence blocks in kinetoplast minicircles from diverse species of trypanosomes. Molecular & Cellular Biology 9 (3) 1365–1367.

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

Robinson, N. P., Burman, N., Melville, S. E., & Barry, J. D. (1999). Predominance of Duplicative VSG Gene Conversion in Antigenic Variation in African Trypanosomes. *Molecular and Cellular Biology*, *19*(9), 5839-5846.

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

Schafer da Silva, A., Machado Costa, M., Flores Polenz, M., Polenz, C.H., Geraldes Teixeira, M. M, Dos Anjos Lopes, S.T., Gonzalez Monteiro, S. (2009). Primeiro registro de "Trypanosoma vivax" em bovinos no Estado do Rio Grande do Sul, Brasil. Ciência Rural 39 (8) 2550-2554

Schnaufer, A., Clark-Walker, G. D., Steinberg, A. G., & Stuart, K. (2005). The F1-ATP synthase complex in bloodstream stage trypanosomes has an unusual and essential function. The EMBO Journal, 24(23), 4029-40.

Shapiro S., Naessens J., Liesegang B., Moloo K., Mogandu J., (1984) Analysis by flow cytometry of DNA synthesis during the life cycle of African trypanosomes. Acta Tropica 41 313-323

Shapiro T., Englund P., (1995) The structure and replication of kinetoplast DNA. Annu. Rev. Microbiol. 49 117-43.

Shaw, J. M., Feagin, J. E., Stuart, K., & Simpson, L. (1988). Editing of kinetoplastid mitochondrial mRNAs by uridine addition and deletion generates conserved amino acid sequences and AUG initiation codons. Cell, 53(3), 401-411.

Simpson, A. G., Stevens, J. R., & Lukeš, J. (2006). The evolution and diversity of kinetoplastid flagellates. Trends in Parasitology, 22(4), 168-174.

Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J., & Birol, I. (2009). ABySS: a parallel assembler for short read sequence data. Genome research, 19(6), 1117-1123.

Smith, D. H., Pepin, J., & Stich, A. H. (1998). Human African trypanosomiasis: An emerging public health crisis. British Medical Bulletin, 54(2), 341-355.

Stevens, J. R., Noyes, H. A, Dover, G. A, & Gibson, W. C. (1999). The ancient and divergent origins of the human pathogenic trypanosomes, T. brucei and T. cruzi. Parasitology, 118, Pt 1, 107–116.

Stevens, J., & Rambaut, A. (2001). Evolutionary rate differences in trypanosomes. Infection, Genetics and Evolution, 1(2), 143–150.

Stich, A. (2002). Human African trypanosomiasis. BMJ, 325(7357), 203-206.

Stuart, K. D., Schnaufer, A., Ernst, N. L., & Panigrahi, A. K. (2005). Complex management: RNA editing in trypanosomes. Trends in Biochemical Sciences, 30(2), 97–105.

Taylor, J., & Rudenko, G. (2006). Switching trypanosome coats: whats in the wardrobe? Trends in Genetics, 22(11), 614-620.

Thon, G., Baltz, T., Giroud, C., & Eisen, H. (1990). Trypanosome variable surface glycoproteins: Composite genes and order of expression. *Genes & Development, 4*(8), 1374-1383.

Tse, H., Cai, J. J., Tsoi, H., Lam, E. P., & Yuen, K. (2010). Natural selection retains overrepresented out-of-frame stop codons against frameshift peptides in prokaryotes. BMC Genomics, 11(1), 491.

Vickerman K. (1965) Polymorphism and mitochondrial activity in sleeping sickness trypanosomes. Nature, 208(5012):762-6

van Weelden, S., Fast, B., Vogt, A., van der Meer, P., Saas, J., & van Hellemond, J. et al. (2003). ProcyclicTrypanosoma brucei do not use Krebs Cycle activity for energy generation. Journal Of Biological Chemistry, 278(15), 12854-12863.

Wickstead, B., Ersfeld, K., and Gull, K. (2004) The small chromosomes of Trypanosoma brucei involved in antigenic variation are constructed around repetitive palindromes. Genome Res 14: 1014–1024

Zerbino, D. R.; Birney, E. (2008). "Velvet: Algorithms for de novo short read assembly using de Bruijn graphs". Genome Research. 18 (5): 821–829.

Ziegelbauer K., Quinten M., Schwarz H., Pearson T.W., Overath P. (1990) Synchronous differentiation of Trypanosoma brucei from bloodstream to procyclic forms in vitro. Eur J Biochem 192 (2) 373-8

Anexo

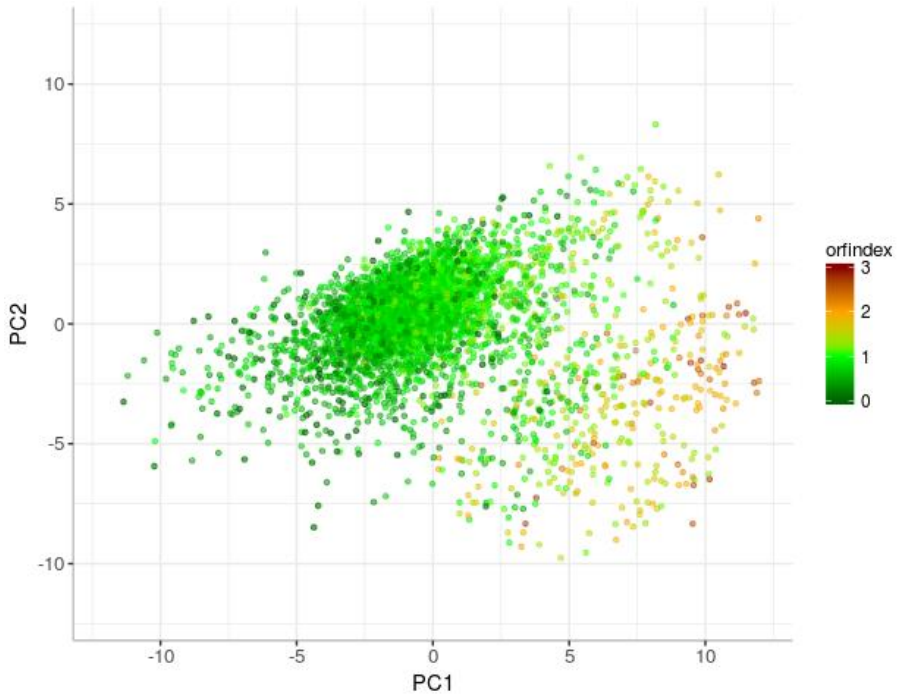

Figura A1.- Gráfica del PCA de frecuencia de trinucleótidos en *T.vivax* Liem-176. En la cepa Liem-176, los resultados obtenidos son muy similares a la otra cepa americana, MT1.

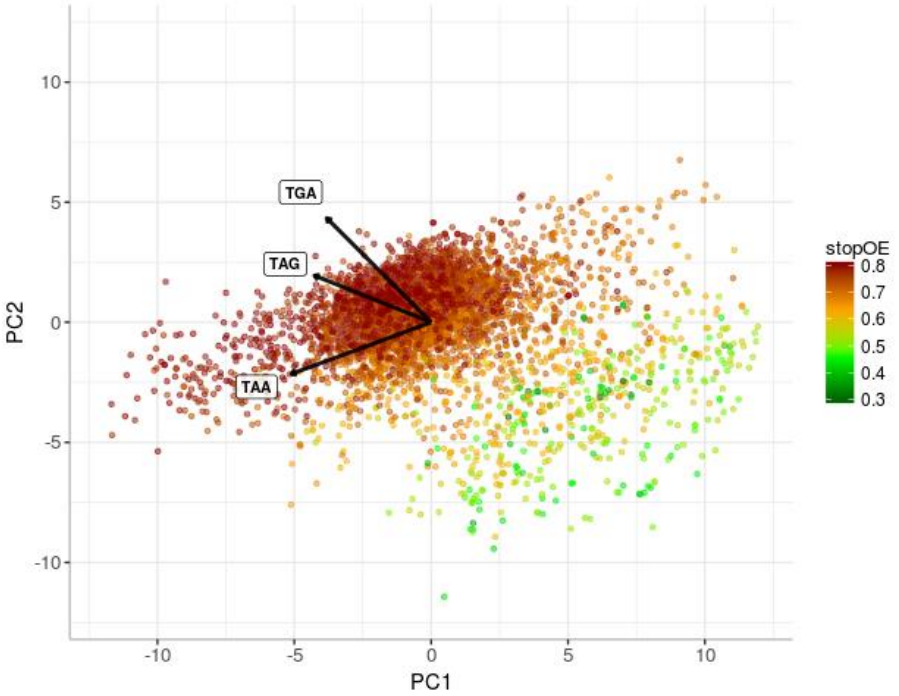

Figura A2.- Gráfica del PCA de *T.vivax* MT1. Se observa que la región orfogénica, la de puntos más dispersos presenta valores de stopOE más bajos que la región de genes 'tradicionales'. Las flechas indican el aporte de los valores originales de los codones stop a la varianza.

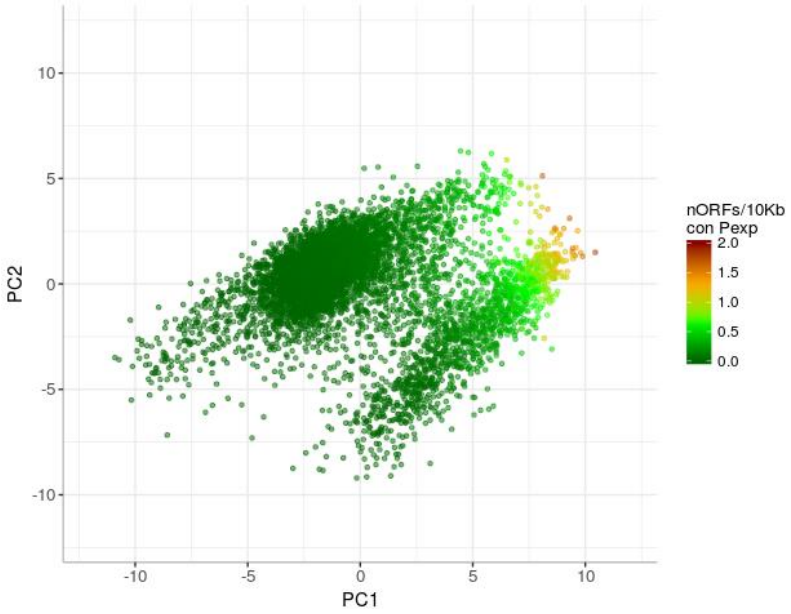

Figura A3.- Gráfica del PCA de *T.vivax* Y486 que estima el número de ORFs esperados utilizando la probabilidad de stops calculada como en base a la frecuencia nucleotídica.

Mutation Research 773 (2015) 69–82

Contents lists available at ScienceDirect

Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis

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

Kinetoplast adaptations in American strains from *Trypanosoma vivax*

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

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

ARTICLE INFO

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

Keywords:
Fast evolution
Mechanical transmission
Genome degradation
Editing

ABSTRACT

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

1. Introduction

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

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

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

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

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

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

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

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

2. Materials and methods

2.1. Parasites

2.1.1. Experimental infection and parasite purification

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

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

2.1.2. RNA and DNA purification and quality control

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

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

2.1.3. Library construction and sequencing

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

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

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

2.1.4. PCR and Sanger sequencing confirmation

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

2.1.5. Data handling and analysis

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

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

3. Results

3.1. Maxicircle genome: gene degradation in American strains

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

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

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

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

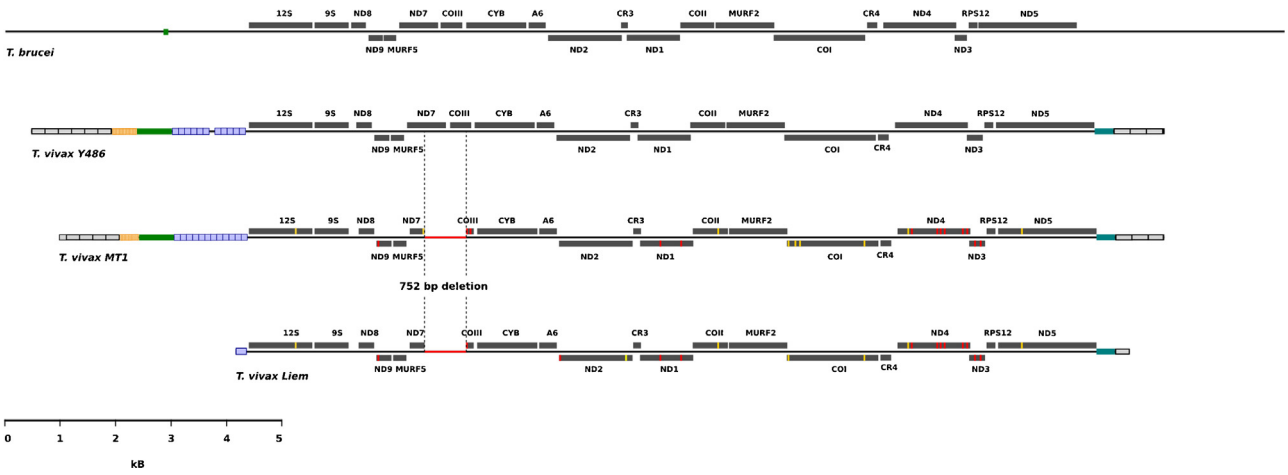

Fig. 1. Alignment of complete maxicircle sequences in the three *T. vivax* strains analyzed in this work. *T. brucei* maxicircle was included as a reference. Genes are indicated by gray boxes and the respective names indicated. Those genes placed above the line are located in the forward strand whereas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

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

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

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

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

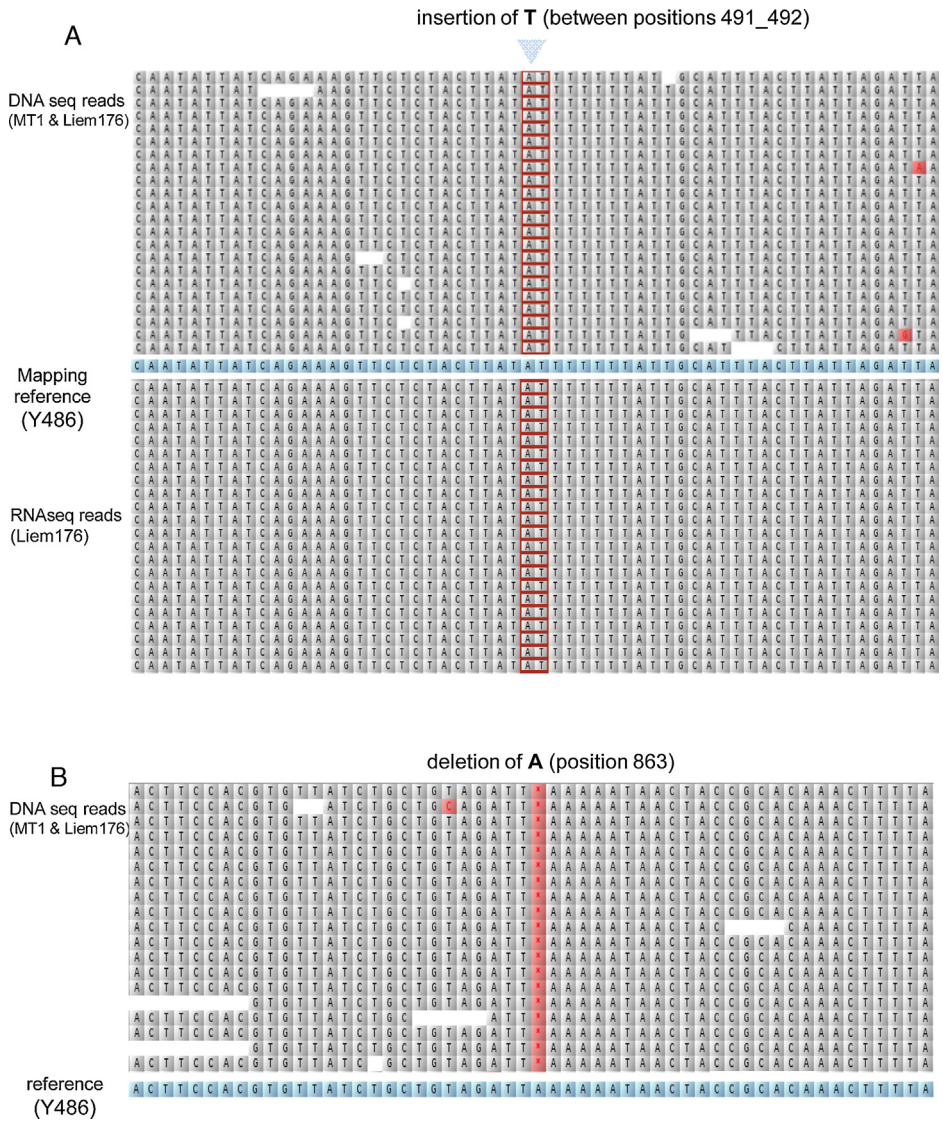

Fig. 2. Testing heteroplasmsy and novel editing sites in the ND1 gene from American strains. (A) Insertion of a **T** between positions 491 and 492. In the upper half of figure (above the mapping reference) DNA sequencing derived reads (from MT1 and Liem176) were mapped against the ND1 coding sequence.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

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

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

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

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

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

3.2. Editing of maxicircle genes

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

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

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

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

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

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

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

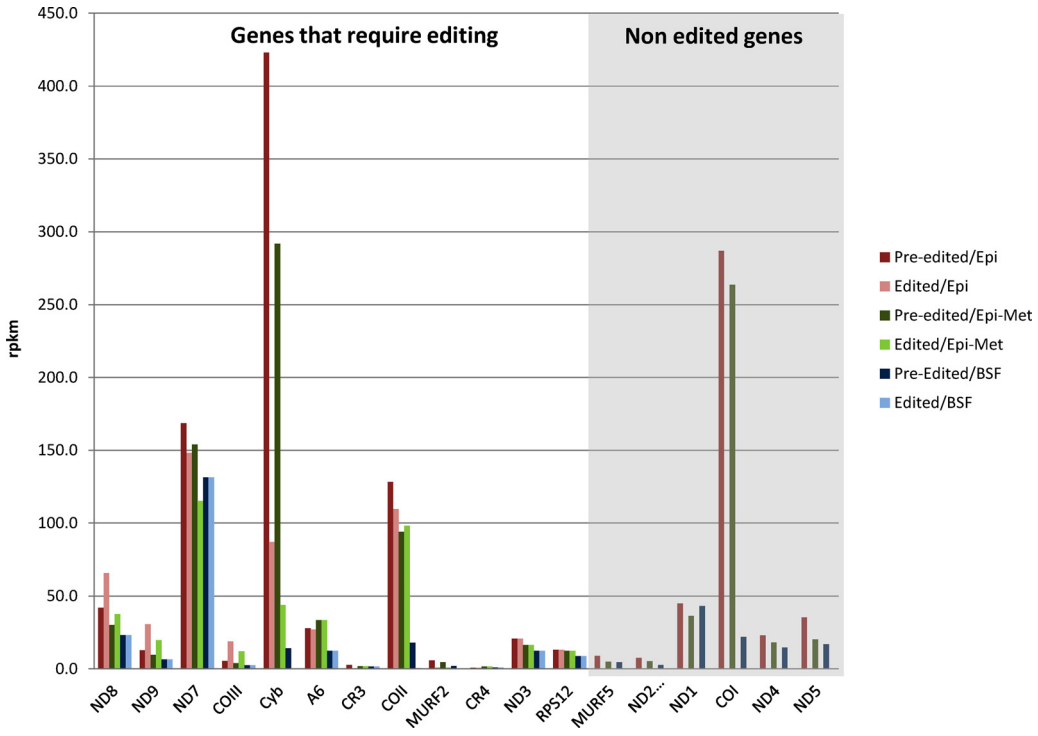

Fig. 3. Transcript levels of edited (mature mRNA) and pre-edited maxicircle genes in three different parasitic life stages (epimastigotes, bloodstream trypomastigotes and metacylic epimastigotes) in the strain Y486.

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

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

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

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

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

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

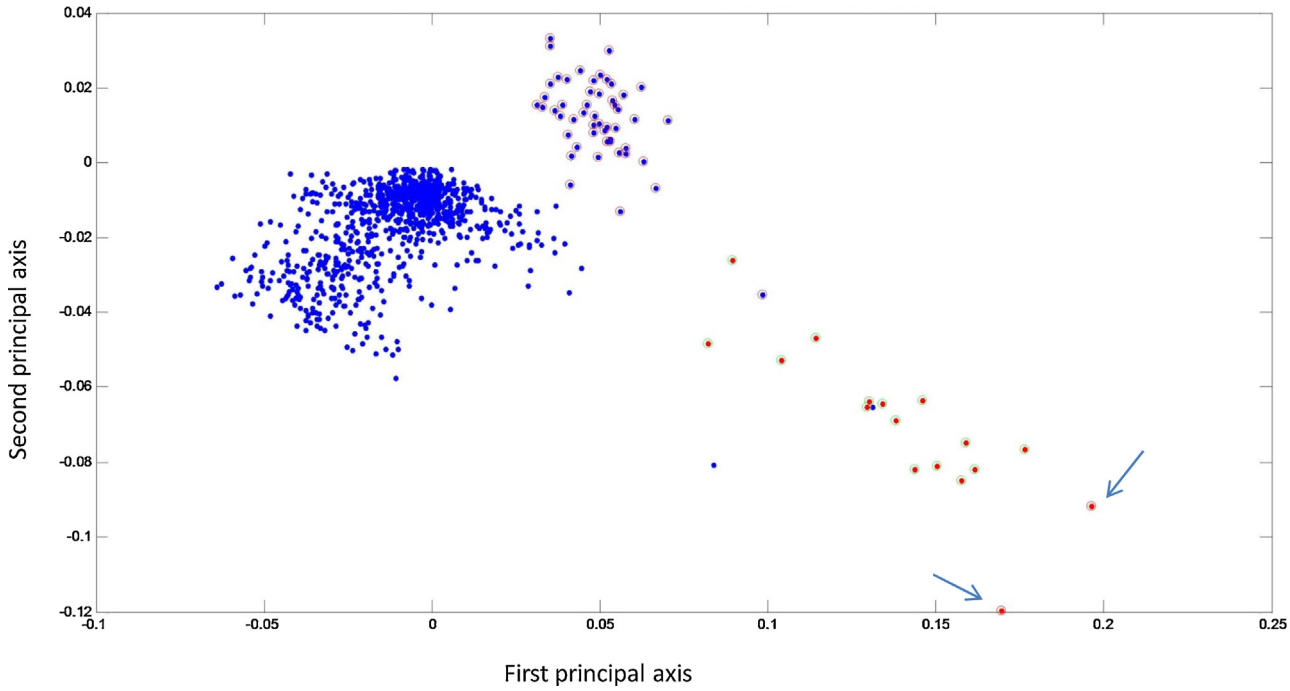

Fig. 4. Principal component analysis of dinucleotide frequencies. First and second axis plotted. Blue dots represent nuclear genome sequences (contigs), blue dots with red circles: putative minicircle sequences. Red dots inside green circles: maxicircle sequences. Maxicircle was splitted in segments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

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

3.4. Minicircle abundance and expression

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

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

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

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

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

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

3.5. Analysis of γ subunit of ATP synthase

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

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

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

Table 4 (*Continued*)

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

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

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

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

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

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

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

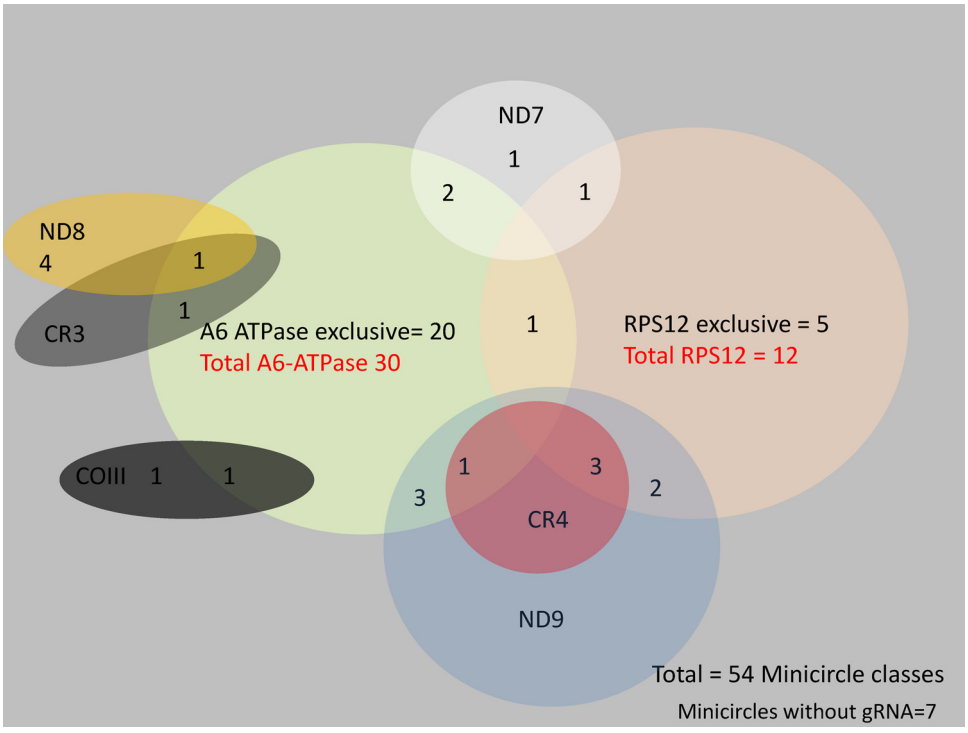

Fig. 5. Venn diagram showing the number of minicircle classes containing gRNA for each gene. Only six minicircle classes have gRNAs that do not participate in the editing of A6-ATPase and/or RPS12, whereas in seven minicircle classes no gRNAs were detected.

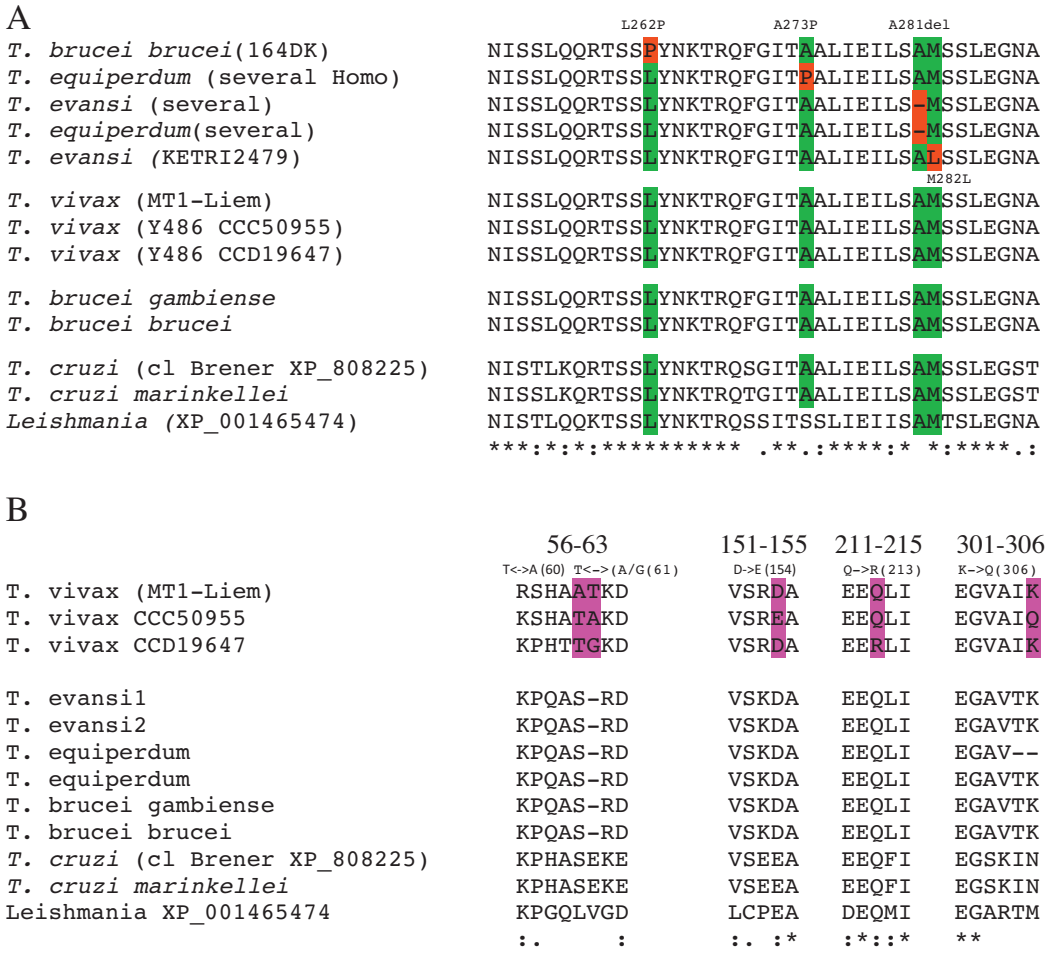

Fig. 6. Multiple alignment in relevant γ ATPase regions. The alignment includes representative trypanosomatid species, the African and American strains from *T. vivax* and several *T. evansi* and *T. equiperdum* strains that contain amino acid changes capable of compensating the loss of mitochondrial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

4. Discussion

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

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

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

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

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

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

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

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

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

Funding

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

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

Conflict of interest statement

There are no conflicts of interest to declare.

Acknowledgement

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

Appendix A. Supplementary data

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

References

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

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