Tesis de
Maestría en Bioinformática
PEDECIBA

Caracterización de la variabilidad composicional y estructural de los cromosomas de *Plasmodium vivax*

María Noël Irazoqui

Orientadores: Dr. Fernando Alvarez y Dr. Gustavo Guerberoff

Abril, 2018

Indice

RESUMEN

La malaria es una de las enfermedades infecciosas con mayor índice de mortalidad en el mundo, es endémica en más de la mitad de los países del mundo, principalmente en América del Sur, el sudeste de Asia, Oceanía y África. Si bien existen varias especies de *Plasmodium* que afectan humanos, *Plasmodium vivax,* cuyo vector es el mosquito hembra del género *Anopheles,* es el causante de la malaria humana más ampliamente distribuido y la principal causa de la enfermedad fuera de África. Debido a que la mayoría de las muertes relacionadas con la malaria son causadas por *P*. *falciparum,* repercutiendo en una escasa cantidad de recursos destinados a la malaria causada por *P. vivax*.

Los agentes causantes de la malaria presentan una notoria diversidad composicional intragenómica, aunque existe una notoria diversidad entre especies. La situación es muy llamativa en el parásito de la malaria *P. vivax,* el cual presenta un genoma que cubre un amplio espectro de composición (0.25 - 0.55 G+C) La gráfica del contenido G+C de los fragmentos obtenidos al segmentar la secuencia genómica, muestra una distribución bimodal con una campana centrada en 0.3 y otra en 0.48.

En esta tesis se analiza la variabilidad en el contenido G+C genómico de *P. vivax*, los posibles factores subyacentes de la misma y la eventual existencia de dicha diversidad composicional en otras especies del género.

La segmentación de las secuencias cromosómicas del genoma en fragmentos de 10 Kpb. revela que la distribución espacial de los segmentos con un contenido G+C diferente al contenido G+C genómico no es uniforme a lo largo de los cromosomas. Los segmentos más pobres en contenido G+C están localizados en las regiones teloméricas, aumentando gradualmente el valor del contenido G+C a medida que nos acercamos a las regiones centrales de los cromosomas.

La clasificación de las secuencias cromosómicas del genoma en categorías funcionales/génicas (proteínas, regiones intergénicas, etc.) permite apreciar que cuando nos acercamos a regiones teloméricas, el valor del contenido G+C es pobre independientemente de la categoría génica considerada. Esto indica que el bajo o alto contenido G+C de las regiones cromosómicas en cuestión no puede atribuirse a la sobre-representación de algún tipo particular de secuencias. Por el contrario, esta observación indicaría que fueron los

componentes génicos de cada una de estas regiones las que se adecuaron a la composición de la zona que las contiene.

Estudios adicionales que incluyen análisis de componentes principales refuerzan la idea de que el contenido G+C diferencial en cada zona es debido a una propiedad intrínseca de la zona y no a una particularidad de alguno de sus componentes. El análisis estructural del ADN en busca de relaciones entre la secuencia, el contenido G+C y la curvatura muestra una alta correlación inversa entre curvatura y contenido G+C revelando que el motivo del bajo contenido G+C telomérico es debido a una alta curvatura de la zona.

Los análisis realizados en *P. vivax* y especies emparentadas permiten concluir que existe una clara relación entre el contenido G+C y la curvatura del ADN tanto en *P. vivax* como en las especies emparentadas.

INTRODUCCIÓN

ASPECTOS BIOLÓGICOS Y EPIDEMIOLÓGICOS DE LA MALARIA

La malaria es una de las enfermedades infecciosas con mayor índice de mortalidad en todo el mundo, cuya magnitud es incuestionable. Ésta produce de 1 a 3 millones de muertes y aproximadamente 515 millones de casos clínicos al año. Es endémica en más de la mitad de los países del mundo, principalmente en América del Sur, el sudeste de Asia, Oceanía y África. Diversas especies del género *Plasmodium* (familia Plasmodiidae, filo Apicomplexa) son causantes de la malaria. Se conocen más de 175 especies de las cuales seis son capaces de infectar a los humanos, *P. falciparum, P. vivax, P. ovale, P. malariae, P. knowlesi* y *P. cynomolgi* [Barnwell et al. 2007]. Si bien la mayoría de las muertes relacionadas con la malaria son causadas por *P. falciparum*, *P. vivax* es el parásito de la malaria humana más ampliamente distribuido y la principal causa de la enfermedad fuera de África, especialmente en los países de Asia y América. La población en riesgo es de aproximadamente 2.500 millones de personas, y aunque la magnitud exacta de la enfermedad causada por *P. vivax* sigue siendo un tema de debate es probable que se haya subestimado y que entre 100 y 300 millones de casos clínicos cada año se deban a este parásito. Además, factores como la aparición de cepas de *P. vivax* resistente a la cloroquina,

el incremento en la gravedad clínica (incluyendo la mortalidad) asociadas exclusivamente con *P. vivax* y el calentamiento global indican que la incidencia de *P. vivax* aumentará en los próximos años [Mendis et al. 2001]. Dado que las infecciones por *P. vivax* tienen una tasa de mortalidad más baja que *P. falciparum,* y que algunos métodos experimentales habitualmente utilizados como el cultivo celular *in vitro* no han sido exitosos, este parásito causante de la malaria sigue siendo relativamente descuidado a la sombra de *P. falciparum*. Este escenario, combinado con el hecho de que estos parásitos son especie específicos y simpátricos [Snounou y White,2004] permite especular que una vacuna eficaz contra *P. falciparum* no protegerá contra *P. vivax*. Por el contrario, probablemente creará nuevas oportunidades para las infecciones por este último [Fernandez-Becerra et al. 2008]. Esto refuerza la necesidad de desarrollar vacunas específicas contra *P. vivax* o incluir formulaciones que ofrezcan protección contra ambas especies. En la actualidad el creciente volumen de información genómica de estos organismos combinada con datos evolutivos y funcionales permite abordar su estudio desde diversos ángulos, lo cual abre nuevas perspectivas para combatir la que es considerada la más mortal de las parasitemias [Kochar et al. 2005].

CICLO DE VIDA DEL PLASMODIUM E INFECCIÓN DE LA MALARIA

Estos parásitos requieren de dos tipos de hospederos: un invertebrado (mosquito) que actúa como vector y un vertebrado (reptil, ave o mamífero). Al invertebrado se le considera el hospedero definitivo dado que en él ocurre la reproducción sexual del parásito. La reproducción asexual ocurre en los tejidos del vertebrado, que es considerado como hospedero intermediario (figura 1) [De Koning-Ward et al. 2016].

Los *Plasmodium* son transmitidos de un vertebrado a otro por el mosquito hembra del género *Anopheles* que pueden contener dentro de sus glándulas salivales la forma infectiva del *Plasmodium* (el esporozoíto). Los machos no transmiten la enfermedad dado que ellos sólo se alimentan de néctares de plantas. En su forma infectiva, los esporozoítos son muy móviles y son trasportados rápidamente al hígado donde invaden a los hepatocitos. En estos ocurre una división asexual llamada esquizogonia hepática o ciclo extra-eritrocítico que da origen al estadio llamado merozoíto. En *P. vivax* se ha descrito otro estadio adicional llamado hipnozoíto, el cual queda latente en los hepatocitos y puede activar la infección en

un periodo relativamente corto entre 10 a 30 semanas, siendo responsable de las recidivas de la enfermedad. El merozoíto es liberado al torrente sanguíneo con consiguiente invasión de los eritrocitos (ciclo eritrocítico), donde pueden transmitirse del vertebrado al mosquito cuando este se alimenta reiniciando el ciclo del parásito. [Spencer et al. 2016]

PATRONES DE COMPOSICIÓN EN LOS GENOMAS DE EUCARIOTAS UNICELULARES

Varias investigaciones a nivel de secuencia mostraron que los genomas de los vertebrados están compartimentados en segmentos de algunos cientos de Kpb**.** caracterizados por niveles relativamente homogéneos de contenido G+C llamados isocoros. Éstos a su vez se encuentran organizados en varias familias discretas las cuales presentan distintos niveles de contenido G+C. Las principales propiedades estructurales y funcionales asociadas a estas familias de isocoros son, además de su contenido G+C, la distribución de genes (densidad codificante), la estructura de la cromatina, las frecuencias de secuencia cortas (di y trinucleótidos, así como palabras de mayor largo), el nivel de metilación del ADN, la expresión génica, la sincronización de la replicación y la recombinación.

Diversos resultados disponibles hasta el momento apoyan la idea de que los isocoros son un nivel fundamental de organización del genoma, no sólo en los vertebrados [Costantini y Musto, 2017] sino también en el resto de los eucariotas multicelulares analizados hasta el momento [Cammarano et al. 2009] [Costantini et al. 2013] [Lamolle et al. 2016]

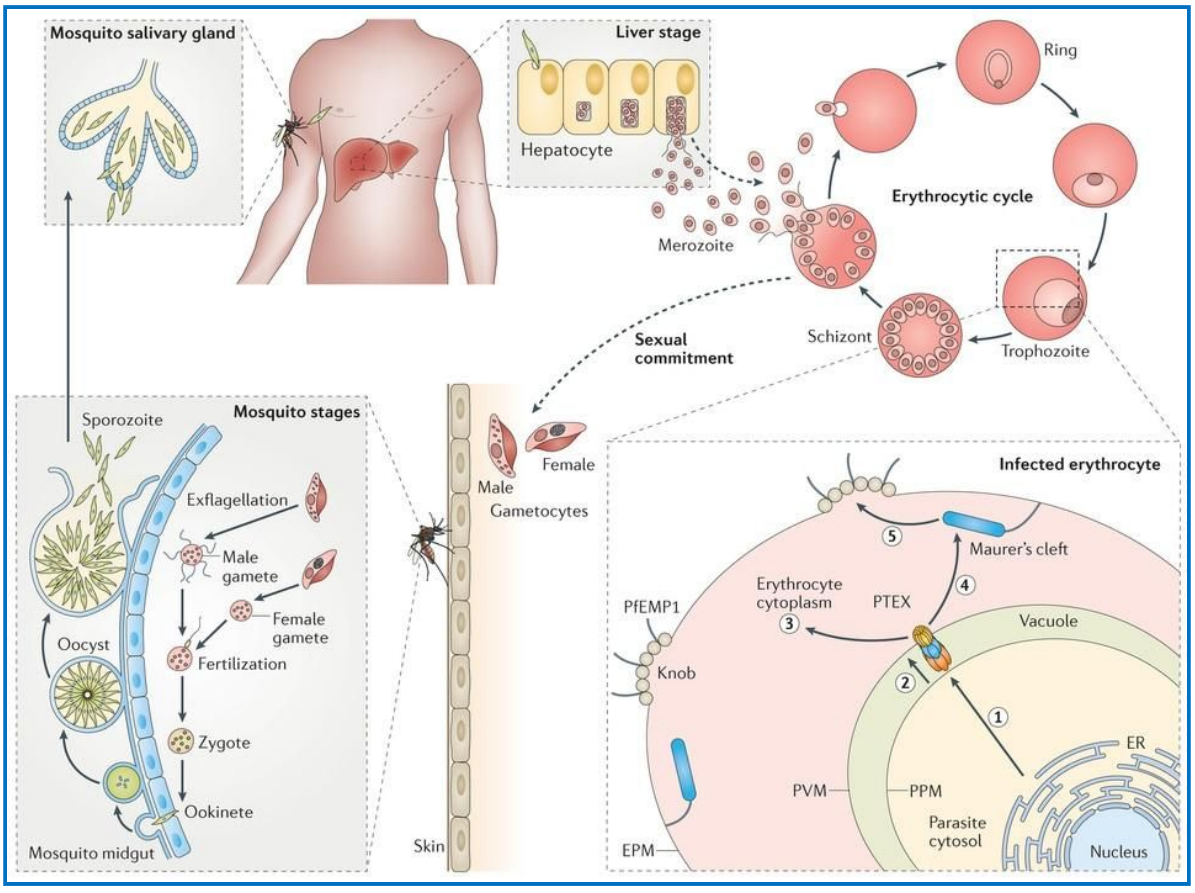


Figura 1: Ciclo de vida *Plasmodium* [De Koning-Ward et al. 2016].

En el caso de los vertebrados, donde el estudio de los isocoros ha sido exhaustivo, se ha propuesto que las regiones con alto contenido G+C son las que han evolucionado más recientemente y por tanto, han aumentado la heterogeneidad del genoma como un todo. [Costantini et al. 2013].

De acuerdo al punto de vista neutralista, éstos se formaron por variación del sesgo mutacional (en un sentido amplio, incluyendo la conversión génica) entre diferentes regiones genómicas. De acuerdo a esta visión, en las regiones ricas en G+C, las mutaciones surgen preferentemente en el sentido AT->GC, mientras que lo opuesto ocurriría en los isocoros pobres en G+C [Costantini et al. 2009]. Por otro lado la hipótesis seleccionista propone que su aparición es el resultado de la necesidad de una mayor termoestabilidad de la estructura helicoidal del ADN en los organismos homeotermos (de sangre caliente), donde los isocoros ricos en G+C precisamente han surgido [Musto et al 2004, 2005]. Sin embargo, en el caso de procariotas la correlación entre el contenido G+C y la temperatura del hábitat donde estos viven ha sido un tema de debate. Por un lado, varios trabajos han

propuesto la ausencia de correlación, lo cual debilita la hipótesis térmica. Esta conclusión fue respaldada por estudios en vertebrados ectotermos donde no se observa una correlación entre el contenido de G+C y la temperatura de adaptación [Vinogradov, 2003]. Se ha demostrado que en los genomas de vertebrados de sangre caliente el aumento del contenido G+C génico está asociado con un aumento absoluto en la flexibilidad de la hélice del ADN, así como con un aumento relativo en comparación con secuencias aleatorias. Esta tendencia tiene lugar tanto en los exones como en los intrones, siendo más pronunciada en estos últimos. A su vez, la energía libre de fusión (delta G) de exones e intrones aumenta absolutamente con la elevación del contenido G+C, pero disminuye en comparación con las secuencias aleatorias (de nuevo, esta tendencia es más fuerte en los intrones). En los genes de animales de sangre fría, plantas y organismos unicelulares, estas correlaciones son más débiles y a menudo no consistentes [Vinogradov, 2001].

Esto sugiere que, si la aparición de isocoros ricos en contenido G+C fue producto de la selección, la misma se produjo en función de la flexibilidad de la molécula de ADN no en función de la termoestabilidad, lo cual puede estar relacionado con una optimización para la transcripción activa en estas regiones genómicas ricas en genes [Vinogradov, 2003] [Richmond y Davey, 2003].

Estudios relativamente recientes demostraron que también en levaduras existe estructuración genómica de tipo “isocoro”. Estas regiones presentan diferencias no solamente en su contenido G+C, sino también en cuanto a la conformación de la cromatina, la modificación de las histonas y la transcripción [Costantini et al. 2013]. Más precisamente, los “isocoros” ricos en contenido G+C tienen una conformación de cromatina más laxa, diferentes niveles de acetilación de histonas y genes más ricos en contenido G+C con un mayor nivel de expresión. La existencia de dicha estructuración genómica en estos eucariotas unicelulares llevó a investigar si la organización tipo isocoro también existía en otros eucariotas unicelulares [Costantini et al. 2013]. Dichos estudios muestran que efectivamente varios grupos presentan diversidad composicional estructurada. Por ejemplo los kinetoplástidos presentan diversidad composicional al interior de sus genomas, tal como ocurre en *Trypanosoma brucei* y *T. equiperdum* (dos tripanosomas estrechamente relacionados) donde el contenido G+C muestra una distribución bimodal. En los agentes causantes de la malaria también existe una amplia diversidad composicional. Un ejemplo es *P. falciparum* que tiene el genoma nuclear más pobre en G+C (0.19) conocido hasta el

momento. Por otro lado, en *P. cynomolgi*, se observa compartimentación composicional. La situación es también muy llamativa en el parásito de la malaria *P. vivax,* el cual presenta un genoma que cubre un amplio espectro de composición (0.25 - 0.55 G+C) con dos secciones centradas en 0.3 y 0.48 respectivamente. (ver Figura 2)

En conclusión, estas observaciones en levaduras, *Plasmodium* y tripanosomas indicarían que la compartimentación composicional no sólo está restringida a los genomas de metazoos y plantas, sino que existe también en los de eucariotas unicelulares [Constantini et al. 2013]. Estos resultados generan el interés en entender la estructura y la evolución de los patrones de composición en eucariotas unicelulares.

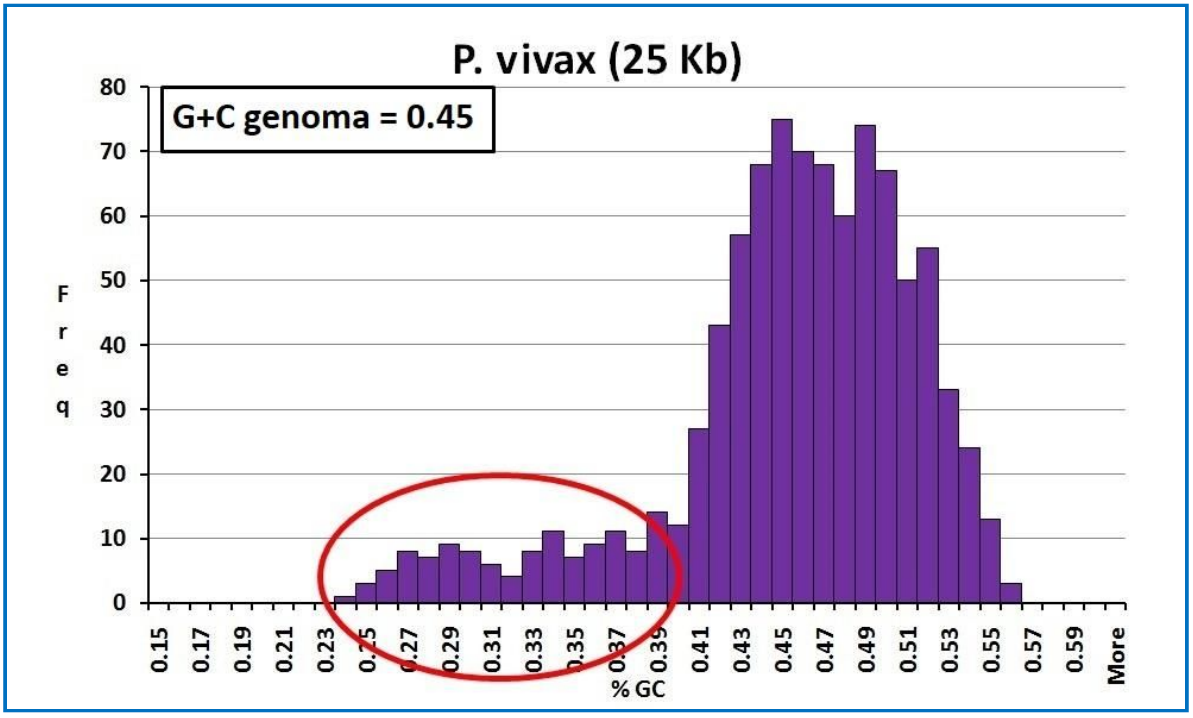


Figura 2: Contenido G+C de segmentos genómicos de 25 Kb de *P. vivax*

Los parásitos causantes de la malaria contienen varias familias multigénicas relacionadas con la evasión a la respuesta al sistema inmune del hospedero por lo tanto las mismas tienen un rol importante en la conservación de la cronicidad en la enfermedad. Estas familias multigénicas se ubican en general en las regiones subteloméricas donde las altas tasas de recombinación crean un ambiente que promueve la expansión de estas familias y la rápida generación de variantes con fenotipos antigénicos diferentes [Merino et al. 2006]. Si

bien tienen muchas características en común, la cantidad de familias por especie, el número de genes de las mismas y la organización varían de una especie a la otra. Así, *P. falciparum* contiene, entre otras, las familias var, rif, Stevor, clag, Pf60 y las más recientemente descubiertas Pfmc-2tm y surf. En contraste, *P. vivax* contiene una única y gran familia multigénica subtelomérica denominada vir (*P. vivax* Interspersed Repeats), que corresponde a aproximadamente el 10% de los genes codificantes. Esta superfamilia está clasificada en diferentes familias (denominadas A a L) que fueron determinadas por similitud de secuencias [Del Portillo et al. 2001][Frech y Chen, 2013] [Fernandez-Becerra et al. 2008].

Otras especies del género *Plasmodium* también contienen familias de genes homólogos a vir, los que en su mayoría también tienen ubicación subtelomérica (tabla 1). De hecho, recientemente se propuso que vir y el resto de las familias junto con rif / stevor de *P. falciparum* sean incluidas dentro de una nueva superfamilia de genes denominada pir (Plasmodium Interspersed Repeats) [Fernandez-Becerra et al. 2008].

Hospedero	Especie	Familia genes	Ubicación
Roedores	*P. chabaudi*	Cir	Subtelomérica
	P. berghei	Bir	Subtelomérica
	P. yoelli	Yir	Subtelomérica
Monos/Humanos	*P. knowlesi*	Kir	Subtelomérica e interna
Humanos	*P. vivax*	Vir	Subtelomérica
Humanos/simios africanos	*P. falciparum*	Rif, Stevor, clag, Pf60, Pfmc-2tm, surf	Subtelomérica e interna

Tabla 1: Superfamilia PIR constituida por familias génicas inmunovariantes en los distintos hospederos.

CARACTERÍSTICAS DEL GENOMA DE *P. VIVAX*

En muchos aspectos, los genomas de *Plasmodium* que infectan mamíferos *(P. falciparum, P. knowlesi, P. vivax, P. yoelli*) son similares, con tamaños que van de 23 a 27 MB, 14 cromosomas, y contienen aproximadamente 5,500 genes de los cuales la mayoría contiene al menos un intrón. No obstante las diferencias en el sesgo de nucleótidos pueden ser extremas: por ejemplo, *P. vivax* tiene un contenido G+C promedio de 0.43 mientras que en *P. falciparum* es de 0.19). Una importante cantidad (77%) de los genes son ortólogos entre las cuatro especies y aproximadamente la mitad de ellos codifican proteínas muy conservadas cuya función se desconoce (denominadas hipotéticas) [Carlton et al. 2008] [Costantini et al. 2013].

El genoma de *P. vivax* tienen un tamaño aproximado de 28 Mb distribuído en 14 cromosomas que varían en tamaño de 0.8 a 3.5 Mb. A la fecha el banco de datos NCBI contiene la secuencia de 8 cepas de *P. vivax* de las cuales 3 están secuenciadas a nivel cromosómico y las 5 restantes a nivel "scaffold" [Genome Assembly and Annotation report, https://www.ncbi.nlm.nih.gov/genome/genomes/35] La secuenciación de estas cepas revela que las mismas presentan algunas diferencias entre si, variando el tamaño del genoma entre 27 y 29 Mb, el contenido G+C genómico entre 0.40 y 0.45 y la cantidad de genes entre 5500 y 6700. [Carlton, 2003][Genome Assembly and Annotation report, https://www.ncbi.nlm.nih.gov/genome/genomes/35]. Nuestros análisis se basan en la cepa *Salvador I* por ser la que contiene el mayor nivel de detalle.

LA SUPERFAMILIA MULTIGÉNICA VIR

El análisis de un extremo del cromosoma de *P. vivax "wild type*" contenido en un cromosoma artificial de levadura (YAC) condujo al descubrimiento de una superfamilia multigénica subtelomérica denominada VIR [Del Portillo et al. 2001].

Una característica de *P. vivax* es la variación antigénica, o sea la capacidad de variar las proteínas superficiales durante el curso de una infección para eludir la respuesta inmune del hospedero, y la familia multigénica VIR, ha sido implicada en la misma. Se identificaron 346 genes (incluyendo 80 fragmentos y/o pseudogenes). Estructuralmente, estos genes varían enormemente, con un tamaño que comprende entre 156 y 2.316 pb**.** y de 1 a 5 exones. En

su mayoría estos genes se encuentran localizados en regiones subteloméricas ricas en A+T, sin embargo la extrema diversidad y subestructuración de las proteínas VIR llevan a especular que existen diferentes localizaciones y funciones subcelulares de sus miembros. Análisis de motivos en el repertorio total VIR develaron que aproximadamente la mitad (171) contienen un dominio transmembrana, y la mitad (160) contienen un motivo similar a la secuencia PEXEL/VSP vinculada a la exportación de proteínas [Hiller et al.2004].

Las proteínas vir se clasificaron inicialmente en seis subfamilias (A-F) sobre la base de similitud de secuencia, y posteriormente la tipificación de los genes VIR en el genoma de la cepa Salvador I detectó 8 nuevas subfamilias (G-I). La expresión génica de varios de estos genes se ha confirmado en infecciones naturales [Carlton et al. 2008].

Las proteínas VIR representan una familia extremadamente diversa, cuyos miembros aparecen actualmente más divergentes que los miembros de otras familias PIR parcialmente caracterizadas, como las familias CIR en *P. chabaudi* (135 miembros) y BIR en *P. berghei* (245 miembros). Se han mostrado características estructurales compartidas entre las proteínas de la subfamilia D de VIR y la familia Pfmc-2tm de *P. falciparum* localizada en las hendiduras de Maurer y entre las proteínas de la subfamilia A de VIR y la familia SURFIN de *P. falciparum* encontrada en la superficie de eritrocitos infectados [Merino et al. 2006].

De particular interés son las integrantes de la subfamilia A (36 genes), que se ha demostrado que provocan una respuesta inmune humoral durante el curso de infecciones naturales y la subfamilia E (con 114 genes), de los cuales 36 copias se encuentran en dos loci a ambos lados del centrómero en el cromosoma 6, con un grupo de 10 genes presente en una región con contenido G+C de 0.47 y un segundo grupo de 26 genes presente en una región con contenido G+C de 0.36 [Carlton et al. 2008].

Los datos recolectados desde el descubrimiento de la superfamilia vir en *P. vivax* indican fuertemente que su función principal no sería la variación antigénica en sentido estricto. Más bien, se propone que las proteínas VIR podrían estar implicadas en un mecanismo de evasión inmunológica nuevo (no entendido aun) que permite a este parásito escapar la barrera del bazo y establecer infecciones crónicas. Finalmente, el descubrimiento de cómo *P. vivax* establece un parasitismo prolongado, además de determinar si las proteínas VIR desempeñan un papel importante en el mecanismo de evasión inmunológica, guiará un enfoque más fundado para las estrategias de control alternativas contra este parásito de la malaria humana, descuidado y no benigno. [Becerra et al. 2008]

INFLUENCIA DE LA SECUENCIA EN LA CURVATURA DEL ADN

Si bien el ADN es el portador de la información genética, también es un objeto físico que ocupa espacio en la célula y tiene propiedades mecánicas. Bajo condiciones fisiológicas, el ADN asume generalmente la forma de una hélice dextrógira con una distancia de 0,34 nm por par de bases (pb**.**) y un diámetro de 2 nm, lo cual es conocido como ADN-B. La mecánica del ADN-B puede ser bien descripta por una serie de parámetros que puntualizamos a continuación [Benham y Mielke, 2005].

- TILT La inclinación del par de bases, se refiere al ángulo de las bases con respecto al eje de la hélice. Un par de bases que es perfectamente plano, o sea perpendicular al eje de la hélice tiene un ángulo de inclinación 0. (Rotación alrededor del eje X)
- ROLL La rotación del par de bases se refiere al ángulo de desviación de un par de bases con respecto al eje de la hélice a lo largo de una línea trazada entre dos pares de bases adyacentes. (Rotación alrededor del eje Y)
- TWIST La torcedura del ADN se refiere al ángulo entre los planos de dos pares de bases adyacentes. (Rotación alrededor del eje Z)
- SHIFT Desplazamiento lateral a lo largo de un plano perpendicular al eje de la hélice. (deslizamiento a lo largo del eje X)
- SLIDE desplazamiento a lo largo de un plano perpendicular al eje de la hélice. (deslizamiento a lo largo del eje Y)
- AXIAL RISE La elevación axial es la distancia entre las bases adyacentes en la doble hélice del ADN. (deslizamiento a lo largo del eje Z) [Sinder. DNA structure and Function. Academic Press, INC., 1994. 21-24]

Estos conceptos se aprecian mejor observando la figura 3.

El doblamiento del ADN (bending) como se ilustra en la figura 4, describe la tendencia de que los pares de bases sucesivos sean no paralelos. Esto es producido comúnmente por un movimiento de pares de bases adyacentes sobre sus ejes. Por el contrario, curvatura representa la tendencia del eje de la hélice a seguir un camino no lineal

durante un tramo largo de manera que contribuya al comportamiento macroscópico de la estructura del ADN (por ejemplo la formación de minicírculos) [Goodsell y Dickerson, 1994].

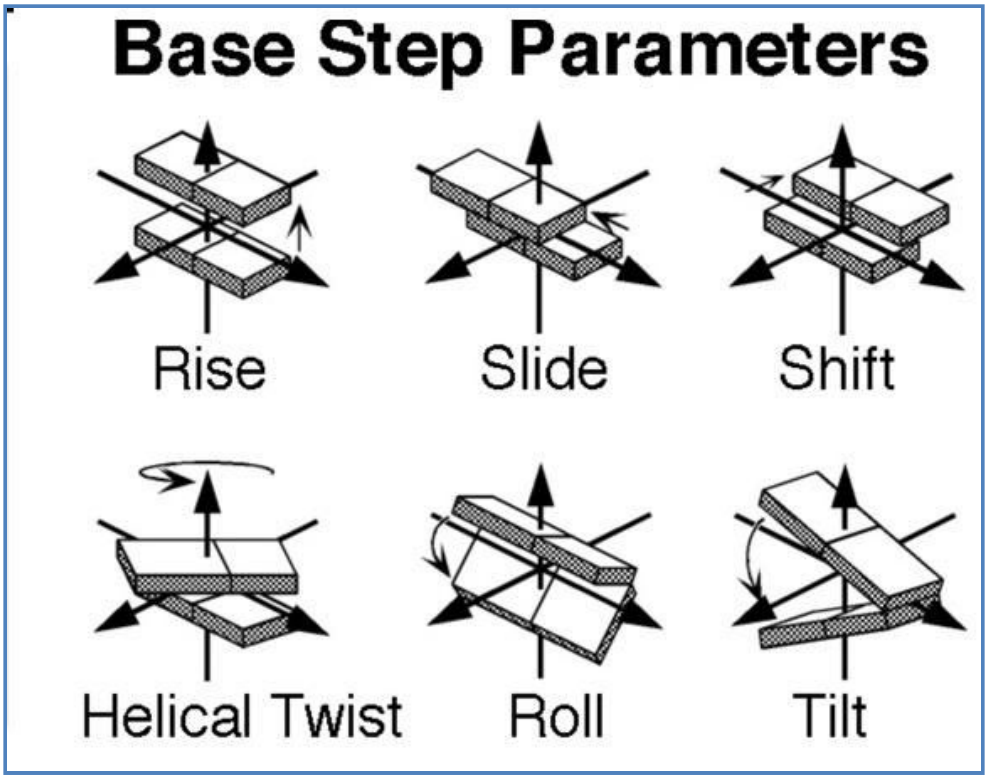


Figura 3: Esquema de los parámetros relacionados al movimiento entre pares de bases de la hélice de ADN obtenido de [Ho et al., 2011]

- El doblamiento del ADN juega un rol primordial en varios procesos esenciales de la célula como la regulación de los genes, la transcripción y la replicación, y hay varios estudios que relacionan la curvatura, el plegamiento y el comportamiento de los nucleosomas con la secuencia [Sinder. DNA structure and Function. Academic Press, INC., 1994. 21-24].

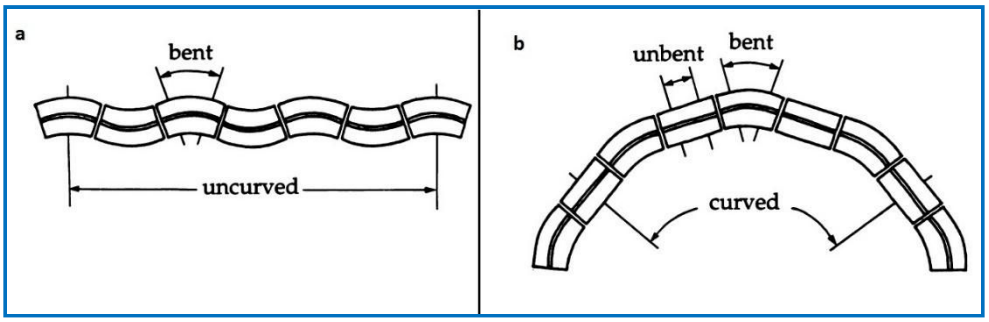


Figura 4: a*)* Representación del doblamiento(bend) del ADN b*)* Representación de la curvatura del ADN. Obtenido de [Goodsell y Dickerson, 1994].

Como los genomas pueden extenderse hasta varios Gpb., el ADN es uno de los polímeros más grandes en la célula y ocuparía en su estructura laxa un espacio mucho más grande que la propia célula. Para poder comprimir el ADN, la naturaleza ha desarrollado estructuras complejas de plegamiento y proteínas, que condensan el genoma en varios órdenes de magnitud (cromatina).

En eucariotas, el ADN se organiza en nucleosomas en los cuales éste es enrollado unas 2 vueltas alrededor de un octámero de histonas. El centro del nucleosoma consiste en dos subunidades de cada una de las cuatro histonas (H2A, H2B, H3 y H4). Las histonas se acoplan al ADN formando complejos sumamente estables que permiten mantener la estructura compacta de la cromatina. [Sinder. DNA structure and function. Academic Press, INC., 1994. 333-334].

La repetición de los nucleosomas, aproximadamente cada 200 pb y las histonas linker son algunos de los factores que permiten que el genoma tenga una estructura dinámica pero compacta que es activamente mantenida por la célula. [Sinder. DNA structure and function. Academic Press, INC., 1994. 335]

Si bien la cromatina permite la compactación en un espacio tan pequeño como el núcleo de la célula, también crea barreras físicas que limitan el acceso al ADN para procesos esenciales como transcripción, replicación y reparación. Sin embargo, los nucleosomas son estructuras muy dinámicas con procesos de apertura y cierre como la respiración nucleosómica (figura 5), desplazamiento, plegamiento y estiramiento que permiten acceso al ADN [Kornberg, 1974].

Las modificaciones (epigenéticas) de los residuos laterales en los aminoácidos de las colas y los núcleos de las histonas, así como del propio ADN, transforman la estructura (la

compactación) de los nucleosomas. Esto se debe a que al alterarse las cargas eléctricas de los residuos laterales disminuye la afinidad del nucleosoma por el ADN, desempeñando un papel fundamental en la regulación de procesos como la transcripción. Estas modificaciones pueden cambiar características como la respiración nucleosómica, o el plegamiento de las estructuras de órdenes superiores y podrían proporcionar un sitio de reconocimiento para factores de transcripción u otros elementos involucrados en diversas vías de regulación. Estos cambios a su vez están relacionados con la secuencia de ADN de los nucleosomas involucrados [Eslami-Mossallama et al. 2016].

La estructura cristalográfica de 1,9-Å del centro nucleosómico que contiene 147 pb. de ADN revela la conformación del ADN nucleosómico con una precisión sin precedentes. La estructura del ADN enrollado alrededor del nucleosoma es marcadamente diferente de la de los oligonucleótidos y de la de los complejos de ADN /proteína no histona. La curvatura de este ADN es en general el doble de la curvatura necesaria para acomodar la superhélice (esta tendencia es más fuerte en los intrones). En los genes de animales de sangre fría, plantas y organismos unicelulares, estas correlaciones son más débiles y a menudo no consistentes [Vinogradov, 2001]. El análisis de la estructura cristalográfica del ADN humano centromérico (α-satélite) en el nucleosoma ha identificado características comunes en la estructura nucleosomal a nivel de pares de bases.

Como era de esperar, los parámetros tilt y roll muestran tendencias oscilatorias típicas, con un período igual al de la vuelta completa de la hélice del ADN y una diferencia de fase de casi 2,5 pb, lo que es una indicación de la conformación de superhélice del ADN en el nucleosoma.

Sin embargo, mientras que el tilt contribuye casi en su totalidad a la formación de la superhélice, el valor de roll es el doble del valor esperado para la superhélice ideal. Esto redunda en una curvatura excesiva del ADN que se manifiesta como un alto valor negativo de roll en el surco menor compensado con un alto valor positivo en el surco mayor. Esta oscilación en los valores de roll es acompañado por oscilaciones en los valores de twist y slide. Entonces las regiones del surco mayor muestran un desenrollamiento sistemático del ADN mientras que lo contrario ocurre en las regiones del surco menor [Eslami-Mossallama et al. 2016].

Dependiendo de la secuencia del ADN, el perfil de curvatura en estas regiones puede ser suave, o puede concentrarse en un par de bases generando un cambio de curvatura brusco y pronunciado.

Dado que la molécula de ADN se deforma significativamente en un nucleosoma y la deformabilidad del ADN cambia con su secuencia, es de esperar que la secuencia de la molécula de ADN afecte tanto su afinidad por la formación de nucleosomas como su estructura dentro de los mismos [Eslami-Mossallama et al. 2016]**.**

Ejemplos de estos efectos dependientes de la secuencia incluyen la presencia de dinucleótidos muy flexibles como TA en las posiciones internas del surco menor, donde la distorsión de ADN es energéticamente más demandante y los elementos ricos en G+C en las posiciones internas del surco mayor lo cual permite una mayor predisposición a la curvatura y compresión del mismo.

También aparecen secuencias TTTAA ubicadas en las zonas internas del surco menor situadas a 1,5 vueltas de la doble hélice desde el centro del nucleosoma (SHL ± 1,5). Esto coincide con la señal de posicionamiento más rigurosa en el nucleosoma, por lo cual se requiere un gran estrechamiento del surco menor [Chua et al. 2012].

Recientemente se ha destacado el rol de la secuencia en estas oscilaciones en los valores de roll y slide y la afinidad por la formación de nucleosomas así como la correlación del contenido G+C de la secuencia con los patrones alternados de desplazamiento en regiones del surco menor [Chua et al., 2012].

La reciente disponibilidad de mapas nucleosómicos a nivel genómico ha permitido entender los mecanismos subyacentes a la organización de los nucleosomas [Eslami-Mossallama et al., 2016][Lu et al., 2016]. Luego de extensos esfuerzos de investigación, se ha establecido que la formación de nucleosomas está influenciada por los efectos combinados de múltiples factores, incluyendo la preferencia de secuencias de ADN, las proteínas de unión al ADN, los remodeladores del nucleosoma, la ARN polimerasa II, la maquinaria de transcripción, la metilación del ADN, las variaciones en las histonas y las modificaciones post traduccionales de las mismas. Particularmente, se ha intentado describir los modelos de formación de nucleosomas basados en señales de secuencia de ADN. [Lu et al., 2016]. En cuanto a cómo las se han realizado extensos estudios experimentales y bioinformáticos [Lu et al., 2016].

Los enfoques bioinformáticos para realizar dichas predicciones se basan típicamente en utilizar información de nucleosomas conocidos como juego de datos de prueba y utilizarlos

para predecir características estadísticas que pueden ser aplicadas en otras secuencias de propiedades dependientes de la secuencia contribuyen a la organización de los nucleosomas,

ADN. Este tipo de análisis genera reglas genéricas sobre secuencias, que coinciden fuertemente con características conocidas del ADN en los nucleosomas, tales como evitar los tractos A, una fuerte preferencia por los dinucleótidos TA, TT y AA en las posiciones internas del surco menor y los dinucleótidos GC en las posiciones internas del surco mayor (figura 6). Alternativamente, los enfoques basados en las propiedades físicas del ADN antes descritas comienzan con la obtención de parámetros mecánicos de estructuras de ADN y nucleosomas obtenidas o bien mediante técnicas de espectroscopía, o con simulaciones de dinámica molecular.

De esta forma se derivan reglas de secuencia "ab initio" para el posicionamiento de nucleosomas a lo largo del ADN. En termodinámica, también es necesario considerar otros mecanismos físicos, como el posicionamiento estadístico de los nucleosomas alrededor de partes inaccesibles del genoma, factores de transcripción, maquinaria de transcripción y / o elementos estructurales de las fibras de cromatina densamente compactadas [Chereji y Morozov, 2011].

La respiración de los nucleosomas, o la exposición del sitio, es un mecanismo donde un segmento de ADN se desenrolla de un extremo del nucleosoma mientras que el resto del ADN nucleosómico permanece envuelto. Este mecanismo se produce espontáneamente como resultado de fluctuaciones térmicas [Polach y Widom, 1995], y permite el acoplamiento de las proteínas de unión al ADN en sus lugares diana dentro del nucleosoma. Se ha demostrado experimentalmente que existe una fuerte dependencia entre la respiración nucleosómica y la secuencia. La interpretación de estos experimentos es, sin embargo, difícil [Eslami-Mossallama et al., 2016].

Los modelos computacionales existentes son de uso limitado, ya que aún no son capaces de tomar en cuenta los efectos de la secuencia. Una excepción es el enfoque computacional de Chereji y Morozov [Chereji y Morozov, 2014], que se ha desarrollado principalmente para interpretar el mapa nucleosómico de *S. cerevisiae*, pero es aplicable solamente a experimentos con un solo nucleosoma y requiere un gran número de parámetros de ajuste.

Para llegar a una comprensión clara de la respiración nucleosómica y cómo se ve afectada por la secuencia, se deben combinar experimentos bien diseñados con simulaciones *in-silico*.

El deslizamiento de los nucleosomas es un mecanismo mediante el cual un nucleosoma cambia su posición en una molécula de ADN sin disolver el octámero de histonas. Los primeros experimentos cuantitativos en condiciones bien controladas fueron presentados por Pennings, Meersseman y Bradbury [Pennings et al. 1991] Los autores idearon métodos para medir el reposicionamiento de nucleosomas mediante electroforesis en gel bidimensional. Demostraron que en repeticiones en tándem de secuencias posicionales 5s rDNA de erizo de mar de la longitud del nucleosoma (cada una de longitud 207 pb), los octámeros de histonas se ensamblan en una posición dominante rodeada de posiciones menores separadas 10 pb (cantidad de bases de una vuelta helicoidal del ADN). En estos experimentos se produjo una redistribución sustancial cuando la muestra se incubó durante 1 h a 37ºC, pero no a 4ºC. Se observó un conjunto de posiciones preferidas, todas a una distancia entre si múltiplo de 10 pb. Los resultados de estas técnicas indican que el grupo de posiciones del octámero está en equilibrio dinámico, en condiciones iónicas bajas, lo que sugiere que las posiciones menores reflejan fluctuaciones alrededor del sitio nucleosomático principal. También parecería que la movilidad del octámero de histonas es dependiente de la temperatura [Pennings et al. 1991] [Chua et al. 2012].

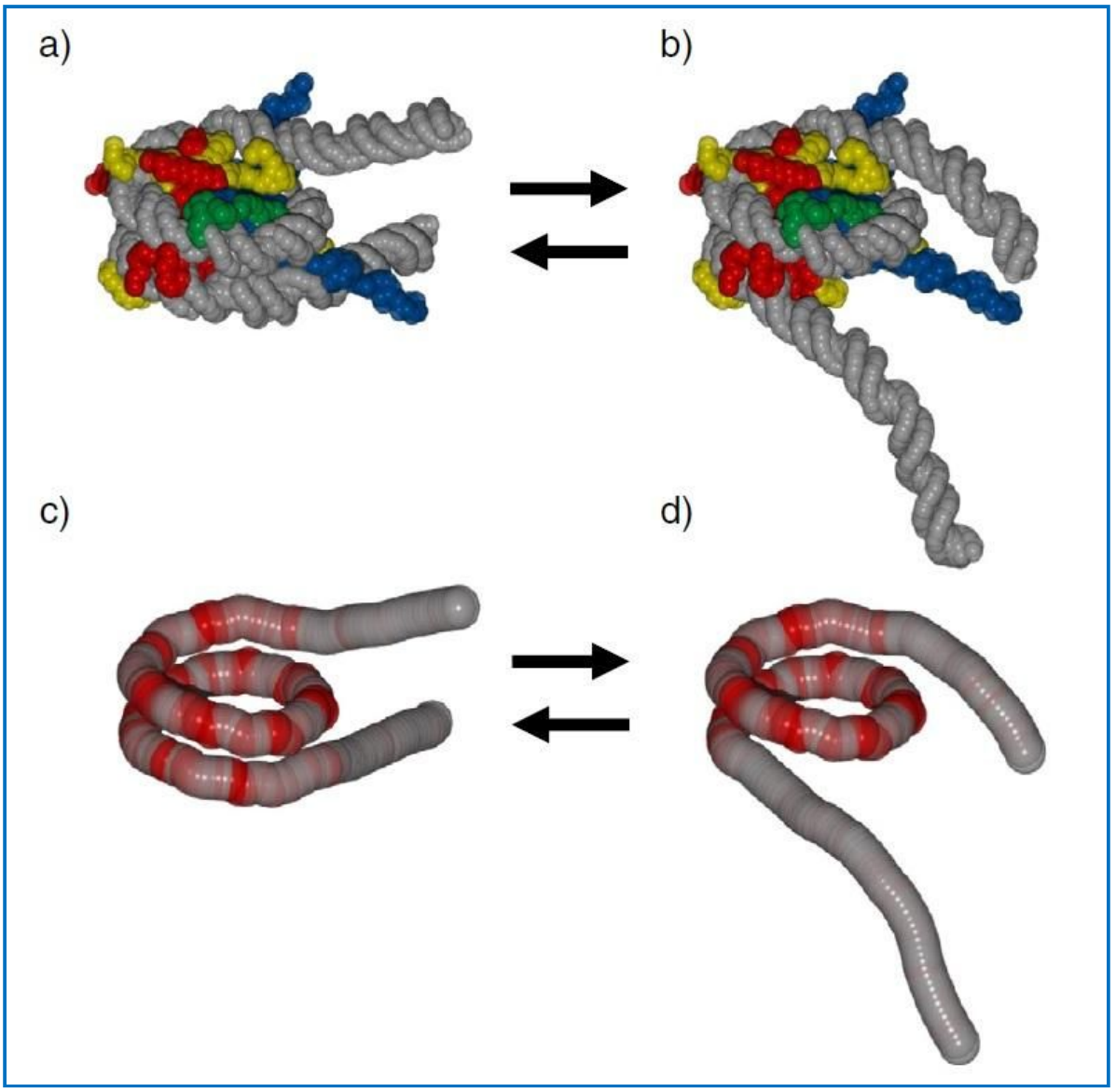


Figura 5: La respiración de los nucleosomas alivia el estrés mecánico en el ADN. La penalización energética del ADN plegado en un nucleosoma de alta curvatura se equilibra mediante interacciones electrostáticas entre el ADN y las histonas, dando como resultado un equilibrio dinámico entre los nucleosomas completamente envueltos (a y c) y los nucleosomas parcialmente desempaquetados (b y d) En las imágenes a y b representamos en color gris el ADN, en amarillo la histona H2A, en rojo H2B, en azul H3, en verde H4. En las imágenes c y d el color rojo muestra las desviaciones del aumento, giro y giro medios, distribuidos a lo largo del ADN. Modificado de [Eslami-Mossallama et al., 2016]

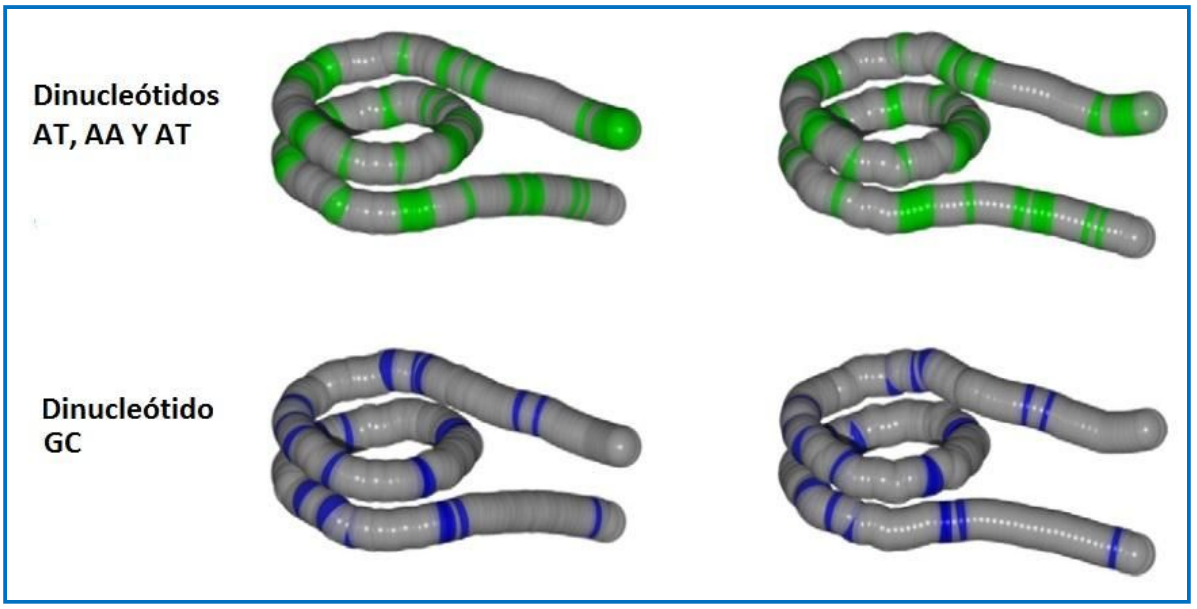


Figura 6: Distribución de dinucleótidos AT, AA y TT en verde y GC en azul. Estos correlacionan en gran medida con los puntos de tensión en el nucleosoma, enfatizando la relación mecánica entre estructura de nucleosomas y la secuencia de ADN. Modificado de [Eslami-Mossallama et al., 2016]

Flaus et al. Desarrollaron una estrategia para seguir el posicionamiento y reposicionamiento de nucleosomas con resolución a nivel de secuencia utilizando histonas H4 químicamente modificadas que inducen, después de la adición de radicales OH, una escisión (cleavage) cerca del "dyad axis". Utilizando este método en secuencias del virus MMTV, Flaus y Richmond estudiaron la dinámica del nucleosoma revelando claramente varias características de reposicionamiento. El fragmento más largo (438 pb) de esta cadena tenía dos secuencias de posicionamiento con dos nucleosomas ensamblados, cada uno en una posición única. Los autores determinaron el grado de reposicionamiento de los nucleosomas individuales en fragmentos más cortos (nucleosoma A en un fragmento de 242 pb y nucleosoma B en un fragmento de 219 pb) en función del tiempo y la temperatura de calentamiento. Se pudo observar que si bien las tasas de reposicionamiento aumentan sustancialmente con la temperatura, también dependen de la secuencia de posicionamiento y la longitud del fragmento [Lu et al. 2016]

La diferencia en el reposicionamiento de las dos secuencias es notable. Para el nucleosoma B más lento, la separación en el conjunto de nuevas posiciones fue en todos los casos múltiplo de 10 pb, mientras que el nucleosoma A, más móvil, no mostró una clara

preferencia por el posicionamiento rotacional. La secuencia de ADN del nucleosoma B contiene dinucleótidos AA / AT / TA / TT con una periodicidad de 10 pb, mientras que la secuencia del nucleosoma A contiene tramos de mononucleótidos. Los autores aducen que estas diferencias reflejan características específicas de las secuencias de ADN subyacentes [Lu et al. 2016].

En algunas especies de otros grupos de parásitos como Tripanosomas o Leishmania se ha reportado la existencia de regiones en las que los valores de curvatura se correlacionan inversamente con el contenido G+C evidenciando un potencial rol funcional para aquellos ADN cuya curvatura es dependiente de la secuencia. En estas regiones se han observado altos contenidos de A+T los cuales pueden ser debidos a tramos repetitivos de A que a su vez pueden inducir curvatura en el ADN (aunque esta característica no es el único determinante de la presencia de regiones altamente curvadas). La curvatura, junto con estructuras alternativas de ADN, están involucradas en la regulación de la iniciación de la transcripción, tanto en procariotas como en eucariotas, en el plegamiento de la cromatina, y la unión de factores de transcripción y/o factores reguladores que interactúan con la maquinaria de transcripción [Smircich et al. 2013].

Por último se desea resaltar que las regiones ricas en contenido A+T, que como vimos anteriormente inciden en la curvatura del ADN, tienen además la capacidad de estimular la recombinación en las zonas linderas ricas en bases repetidas y esto también está relacionado con la conformación estructural de la zona [Timchenko et al., 2002]. La recombinación genética es un proceso que puede generar transformaciones en los genomas, incrementando la variabilidad genética en las poblaciones, y es un factor fundamental de la selección natural pues permite combinar en “cis” alelos favorables de distintos genes, impactando fuertemente en la evolución. La recombinación es también un mecanismo clave a la hora de generar diversidad en parásitos como estrategias de evasión de la respuesta inmunitaria del hospedero. Por lo tanto la investigación de los mecanismos moleculares responsables de la recombinación y de sus factores inductores a nivel genómico (patrones en el ADN, plegamiento de la cromatina) son de sumo interés para profundizar la comprensión del proceso evolutivo en *P. vivax*.

OBJETIVOS

OBJETIVO GENERAL

El objetivo general de esta tesis consiste en caracterizar la variabilidad en el contenido G+C genómico de *P. vivax,* así como investigar los factores subyacentes de la misma. También se plantea investigar si dicha diversidad composicional es exclusiva de esta especie o afecta a otros *Plasmodium*.

OBJETIVOS ESPECÍFICOS

- Identificar la distribución espacial y comportamiento a nivel cromosómico de los segmentos de bajo contenido G+C.
- Explorar la o las causas funcionales que producen esta diferenciación en el contendido G+C.
- Investigar si esta peculiaridad es el subproducto de una distribución desigual de familias multigénicas, secuencias repetitivas, etc., cuyas características composicionales intrínsecas afecten la distribución del contenido G+C de las regiones que las contienen.
- Investigar si éste es un comportamiento exclusivo de *P. vivax* o si es una característica presentada por otras especies de *Plasmodium.*

MATERIALES Y MÉTODOS

OBTENCIÓN DE LAS SECUENCIAS GENÓMICAS DE ESPECIES DE *PLASMODIUM*

En base a las relaciones filogenéticas del género *Plasmodium* (figura 7) seleccionamos aquellas especies que fueran representativas de la variabilidad filogenética del mismo y cuya secuencia genómica estuviera completa (a nivel de ensamblaje cromosómico). Las especies seleccionadas y los números de acceso de sus secuencias genómicas se detallan en la tabla 3. Sus secuencias y archivos de anotación fueron descargadas de NCBI (https://www.ncbi.nlm.nih.gov) junto con los de *P. vivax.*

Especie	**Número de acceso**
Plasmodium vivax isolate Salvador I	NC_009906.. NC_009919
Plasmodium berghei ANKA	NC_036159..NC_035172
Plasmodium chabaudi chabaudi	NC_030101..NC_030114
Plasmodium cynomolgi	NC_020396..NC_020407
Plasmodium knowlesi	NC_011902..NC_011915
Plasmodium falciparum	NC_004325, NC_000521, NC_000910, NC_004314..NC_004331

Tabla 3: Números de acceso en NCBI de especies de *Plasmodium* estudiadas

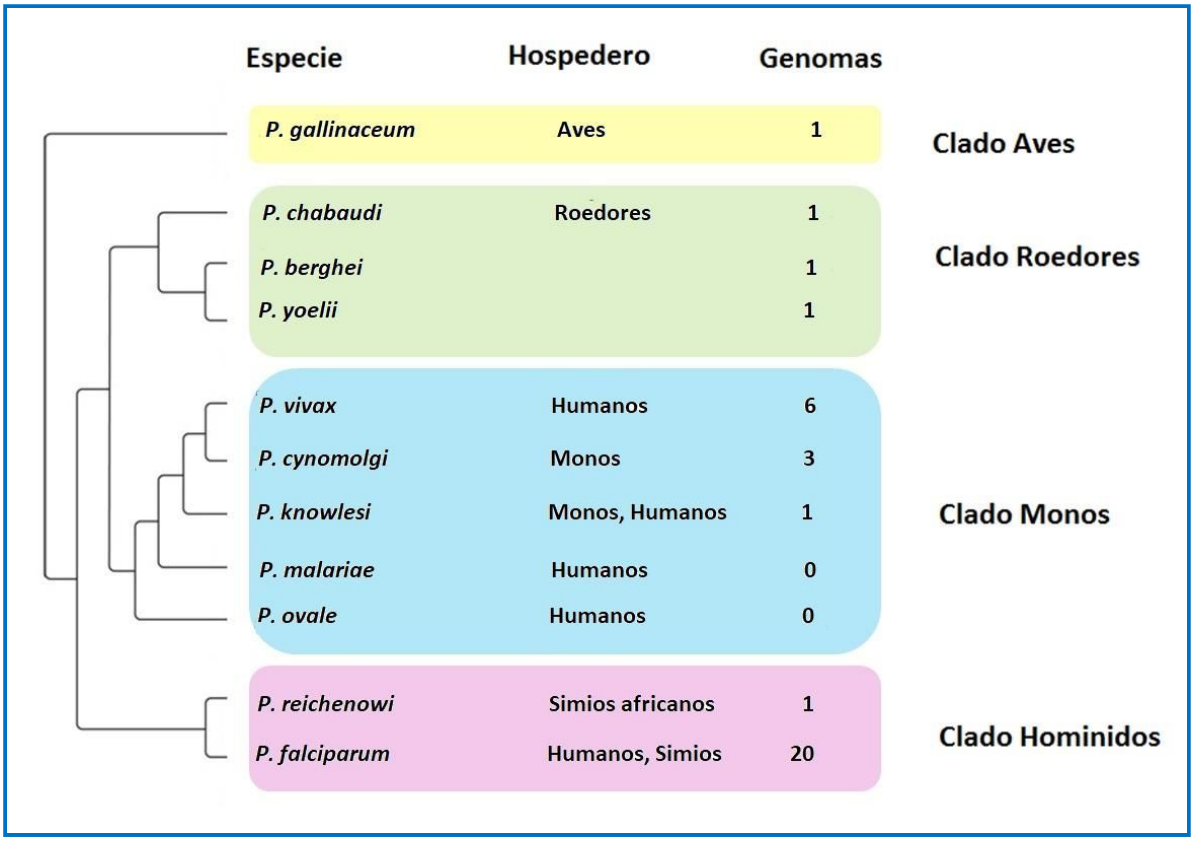


Figura 7: Árbol filogenético del género Plasmodium

ESTUDIO DE LA VARIABILIDAD COMPOSICIONAL Y SU DISTRIBUCIÓN ESPACIAL EN LOS CROMOSOMAS

Los genomas de las diferentes especies de *Plasmodium* fueron fragmentados usando segmentos de distinto tamaño, con el fin de identificar el más adecuado para nuestros datos Testeamos segmentos de 100 Kpb., 50 Kpb., 25 Kpb., y 15 Kpb. Optamos por segmentos de 25 Kpb. puesto que son los que mejor capturan la variabilidad estudiada.

El fraccionamiento a escala cromosómico se realizó con el mismo criterio. Luego de testar segmentos de 15 Kpb., 10 Kpb., 5 Kpb. y 3 Kpb. encontramos que los segmentos de 10 Kpb. eran los que mejor mostraban la variabilidad estudiada.

Es importante resaltar que a nivel cromosómico, graficamos el contenido G+C de los segmentos respetando su distribución espacial dentro del cromosoma. Esto permite

identificar visualmente las zonas cromosómicas donde se ubican aquellos segmentos con contenido G+C diferenciado.

La identificación de zonas cromosómicas con bajo contenido G+C o alto contenido G+C permite realizar estudios adicionales a los efectos de buscar indicios que permitan echar luz sobre las razones funcionales/evolutivas subyacentes a esta eventual compartimentalización genómica.

Un primer nivel de análisis consiste en indagar dentro de los eventuales segmentos cromosómicos de interés (con diversidad en contenido G+C) la información ya disponible en la anotación de los genomas analizados. Esta aproximación simple permite, por ejemplo, determinar si la diferenciación en contenido G+C podría atribuirse al incremento "inusual" de alguna categoría funcional. Concretamente considerando la secuencia nucleotídica y el archivo de anotación, podemos clasificar los datos en categorías funcionales. Identificamos las siguientes 6 categorías:

- Proteínas vir.
- Proteínas no vir
- ARN transferencia
- ARN ribosomal
- Micro ARN
- Regiones intergénicas

A su vez podemos también establecer regiones o zonas dentro de un cromosoma. Vamos a utilizar 3 zonas: zona inicial, zona interna y zona final. Entonces dada una secuencia cromosómica, en base al archivo de anotación le podemos asociar una categoría funcional y una zona (por ejemplo proteína vir en zona interna) como se ilustra en la figura 8.

Denominamos clase a una categoría funcional y una zona dentro del cromosoma. Calculamos el contenido G+C de cada una de estas clases y ponderamos este último teniendo en cuenta el tamaño de la clase dentro del tamaño total del cromosoma.

Existen diversas herramientas estadísticas para el estudio de asociación entre variabilidad composicional del genoma y clases funcionales. Pueden ser métodos univariados como las tablas de contingencia, que permiten determinar si existe asociación entre las formas de categorizar los datos o métodos multivariados como análisis de componentes principales o herramientas de clasificación. Por otro lado las frecuencias de trinucleótidos son buenos indicadores de las características de los distintos compartimentos genómicos o clases, por lo tanto son las variables de entrada utilizadas en el análisis de los datos aquí abordado y cuyos aspectos metodológicos se detallan en el contexto de la presentación de resultados.

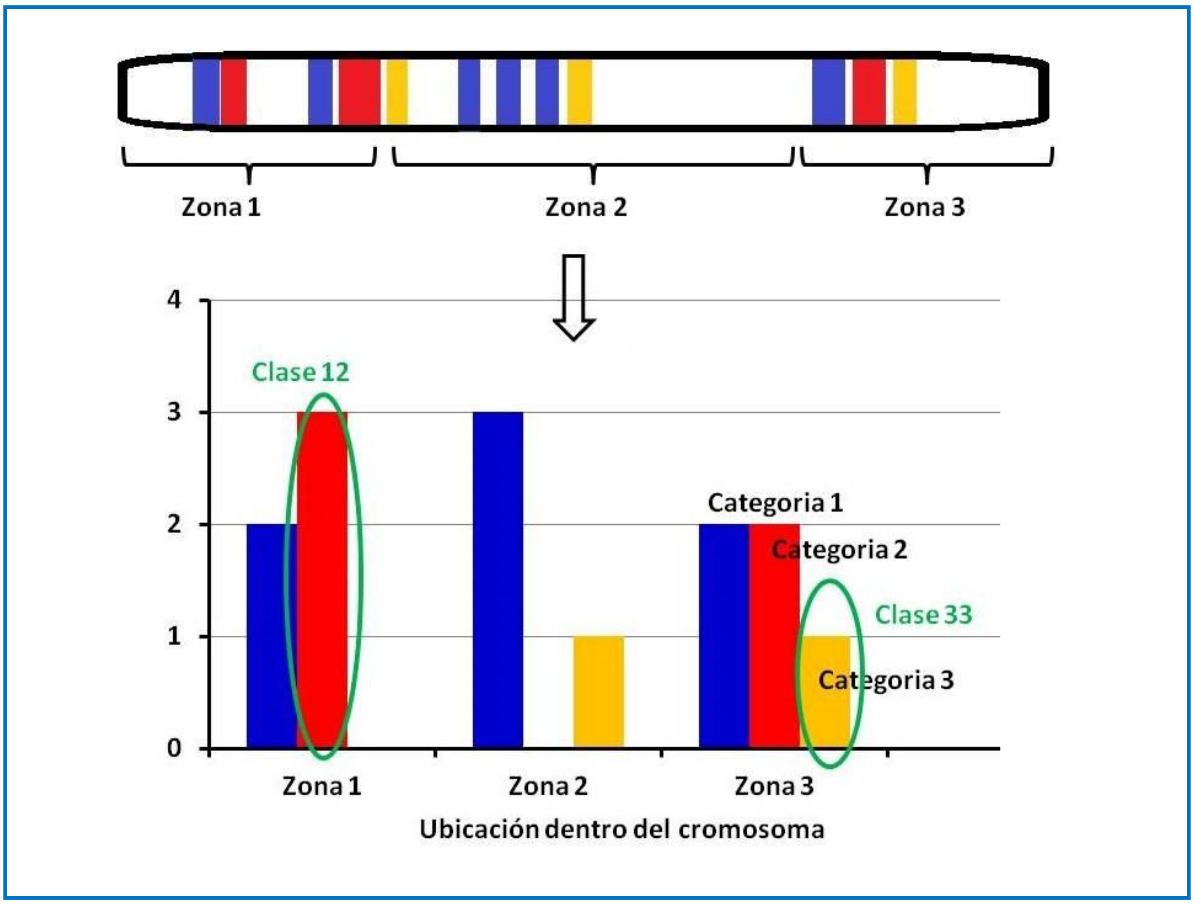


Figura 8: Clasificación de secuencia nucleotídica en clases

ESTUDIO DE LA CURVATURA DEL ADN Y SU RELACIÓN CON LA VARIABILIDAD COMPOSICIONAL

Basados en estudios anteriores [Goodsell y Dickerson, 1994], decidimos estudiar la curvatura de los cromosomas de *P. vivax* y su relación con el contenido G+C.

Existen diversas aplicaciones diseñadas para predecir la curvatura global (curvature) y local (bending) de una secuencia de ADN. Estas están basadas en las aproximaciones analíticas descriptas en la sección "Herramientas de predicción de curvatura".

Como material de calibración usamos secuencias de minicírculos de kinetoplastos. Un kinetoplasto es una masa de ADN circular extra nuclear y corresponde al genoma de la única gran mitocondria de los tripanosomátidos y otras especies de la clase Kinetoplastida. Este es uno de los ADN con mayor curvatura que existen en la naturaleza, por lo tanto puede usarse

como referencia. La comparación de los valores de curvatura de kinetoplastos con los valores aquí analizados permite determinar si estos últimos pueden o no considerarse altos.

HERRAMIENTAS DE PREDICCIÓN DE CURVATURA

DNA Curvature Analysis

Como primera aproximación utilizamos la herramienta on-line "DNA Curvature Analysis" para investigar y comparar la curvatura de diversos tipos de secuencias cortas. Es una aplicación on-line desarrollada por Christoph Gohlke del Laboratorio de Dinámica de Fluorescencia de la Universidad de California (https://www.lfd.uci.edu/~gohlke/dnacurve/). Ésta estima la estructura en tercera dimensión de una molécula de ADN a partir de su secuencia de nucleótidos. Acepta varios modelos de datos descriptos más adelante. Calcula los ángulos de curvatura así como también una aproximación gráfica que ilustra como luce la curvatura puntual y la curvatura global. La limitante es que acepta solamente secuencias con un largo máximo de 256 nucleótidos, y nuestras secuencias están el rango de 0.8 a 3 Mpb. A pesar de esta limitación es de gran utilidad a la hora de comparar la compatibilidad de resultados arrojada por los diferentes algoritmos.

BEND

También testeamos la herramienta BEND [Goodsell y Dickerson, 1994] la cual, aunque de sencilla utilización y versátil (permite varios modelos de curvatura), es limitada, pues acepta secuencias de un largo menor a las secuencias estudiadas.

Es una aplicación desarrollada en lenguaje FORTRAN por el Instituto de Biología Molecular de la Universidad de California [Goodsell y Dickerson, 1994] que permite calcular la magnitud de curvatura local y curvatura macroscópica en cada punto a lo largo de una secuencia arbitraria de ADN B, usando las propiedades intrínsecas de la misma, es decir cualquier modelo de curvatura deseado que especifique valores de inclinación (tilt), rotación (roll) y torsión (twist) en función de la secuencia.

Una de las características del ADN B es que sus pares de bases son prácticamente perpendiculares al eje global de la hélice, por lo tanto, el vector normal a cada par de bases puede tomarse como representación del eje local de la hélice en ese punto.

El programa BEND lee una secuencia y una matriz de ángulos de torsión (twist), rotación (roll) e inclinación (tilt) para cada posible par de bases. Este programa aplica los tres parámetros indicados en cada paso a lo largo de la secuencia, y calcula el vector normal al par de bases resultante. El primer par de bases está alineado de forma normal con el eje z, con un valor de torsión de 0,0 º. El valor de twist especificado es aplicado al segundo par de bases, y los valores de rotación e inclinación se utilizan para calcular su vector normal con respecto al primero.

Si el valor de rotación o el de inclinación son distintos de cero, el nuevo vector se alejará del eje z, produciendo la primera "curva". Este proceso es continuado a lo largo de la secuencia, aplicando los parámetros correspondientes a cada nuevo par de base en relación con su predecesor. El resultado es una lista de vectores normales para todos los pares de bases en la secuencia. Las curvas locales se calculan luego a partir de los vectores normales [Goodsell y Dickerson, 1994].

BANANA

Por último testeamos la aplicación BANANA de la suite EMBOSS (http://emboss.sourceforge.net/apps/release/6.6/emboss/apps/banana.html) la cual, al igual que BEND, es de fácil utilización y permite seleccionar entre diversos modelos de curvatura y acepta secuencias de gran largo como las estudiadas (entre 0.8 y 3 Mpb.). Esta predice tanto la curvatura local en cada punto (bending) así como la curvatura global (curvature) de una hebra de ADN B observando el comportamiento de los vectores normales a lo largo de la hélice de un par de bases con respecto al siguiente en la secuencia. Al igual que BEND se basa en el método de Goodsel y Dickerson [Goodsell y Dickerson, 1994] y acepta como modelo de curvatura todos los modelos testados en Goodsel y Dickerson y cualquier otro modelo que especifique los ángulos de inclinación, rotación y torsión. Luego de testear DNA Curvature Analysis, BEND y BANANA, seleccionamos esta última puesto que a diferencia de las otras acepta secuencias del largo de las estudiadas (entre 0.8 a 3 Mpb.).

Modelos de curvatura

- Satchwell et al (1986)

Los parámetros de giro-inclinación de este modelo se derivan puramente de observaciones experimentales de preferencias de ubicación de secuencias de ADN organizadas en torno a centros de nucleosomas, y en círculos cerrados de ADN de doble hélice de tamaño comparable, sin referencia a técnicas de solución que miden la curvatura *per se*. Por esta razón, podrían ser los parámetros más objetivos e imparciales de todos.

- Calladine et al (1988)

Este modelo fue formulado para explicar los resultados de la migración anómala en gel en 25 secuencias de ensayo, basándose en las características estructurales observadas en análisis de estructura mediante cristalografía de rayos X, en particular la falta de rotación o inclinación de 12-mers tales como C - GC - AAAAAAGCG.

- Bolshoy et al (1991)

Los parámetros de tilt, roll y twist de este modelo son elegidos para explicar la movilidad en gel y los datos de circularización de 54 oligonucleótidos sintéticos diferentes. Este modelo tiene los mayores valores de tilt y twist que cualquiera de las otras alternativas.

- Cacchione et al (1989)

Este modelo fija los valores twist roll y tilt en base a cálculos de energía conformacional. Los resultados son compatibles tanto con resultados de cristalografía como de movilidad electroforética de 62 secuencias diferentes.

- Koo & Crothers (1988)

Este modelo fue propuesto específicamente para explicar la migración electroforética anómala a través de geles. Se asignan valores de tilt y roll distintos de cero a los pares de bases en cualquier extremo de una serie de tres o más adeninas sucesivas y valor cero en cualquier otro lado.

Estos modelos de curvatura fueron testeados por los autores de la aplicación BEND [Goodsell y Dickerson, 1994] concluyéndose que el más preciso es Satchwell et al (1986) debido a que los parámetros de roll, tilt y twist de este modelo son los parámetros más objetivos e imparciales de todos ya que se derivan puramente de observaciones experimentales, sin referencia a técnicas de solución que miden la curvatura per se.

Por ese motivo fue el seleccionado para calcular la curvatura de nuestros datos.

RESULTADOS

VARIABILIDAD COMPOSICIONAL EN LOS GENOMAS DE PLASMODIO

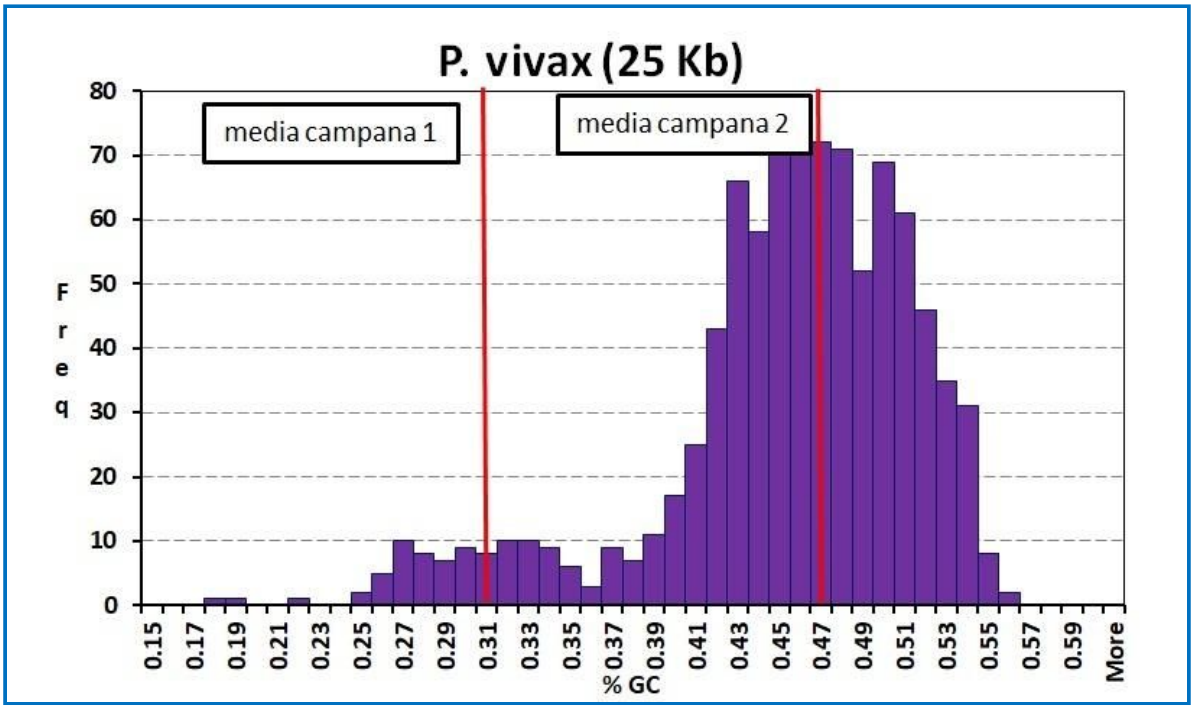


Figura 9: Distribución de contenido G+C de segmentos genómicos de 25 Kb en P. v*ivax*. Las líneas rojas indican la media de la primer y segunda campana respectivamente de la distribución.

Como mencionamos en la Introducción, el genoma de *P. vivax* está compuesto por segmentos de diverso contenido G+C (figura 9). Los mismos comprenden un amplio rango de valores que varía desde 0.25 a 0.57 y se distribuyen generando un histograma con distribución bimodal. La media de una de las campanas es de 0.31, y la segunda campana se agrupa en torno a un valor de 0.47, cercana al valor del contenido G+C genómico global (0.45).

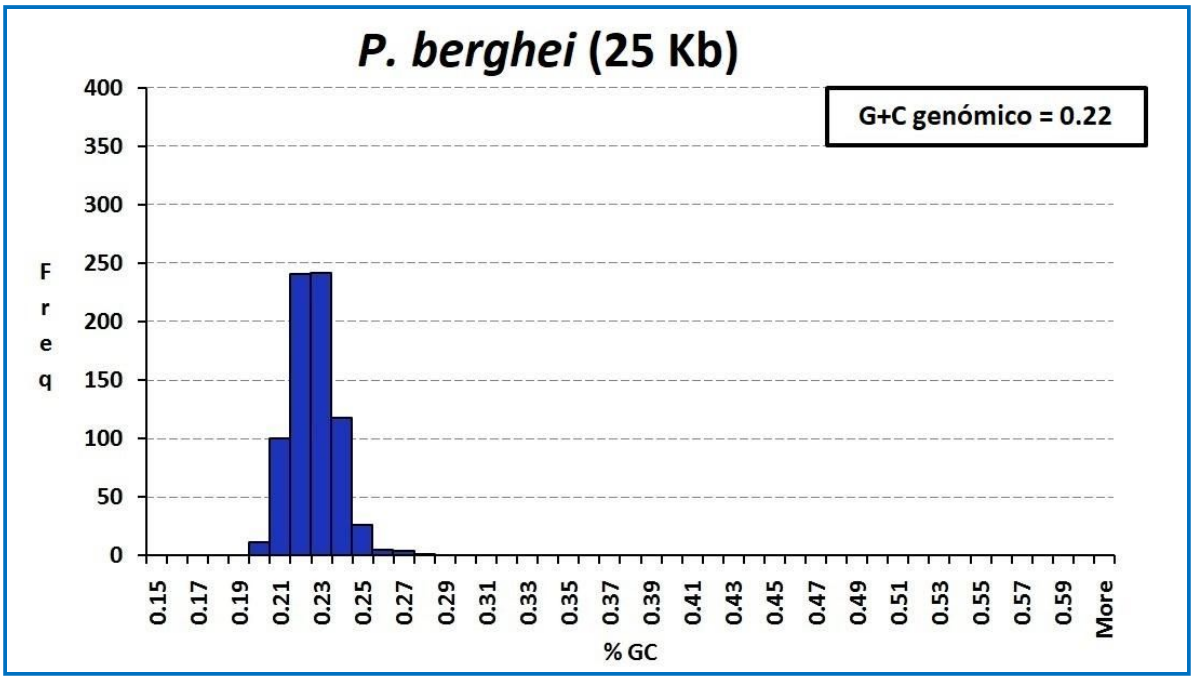


Figura 10: Distribución de contenido G+C de segmentos genómicos de 25 Kb de *P. berghei*

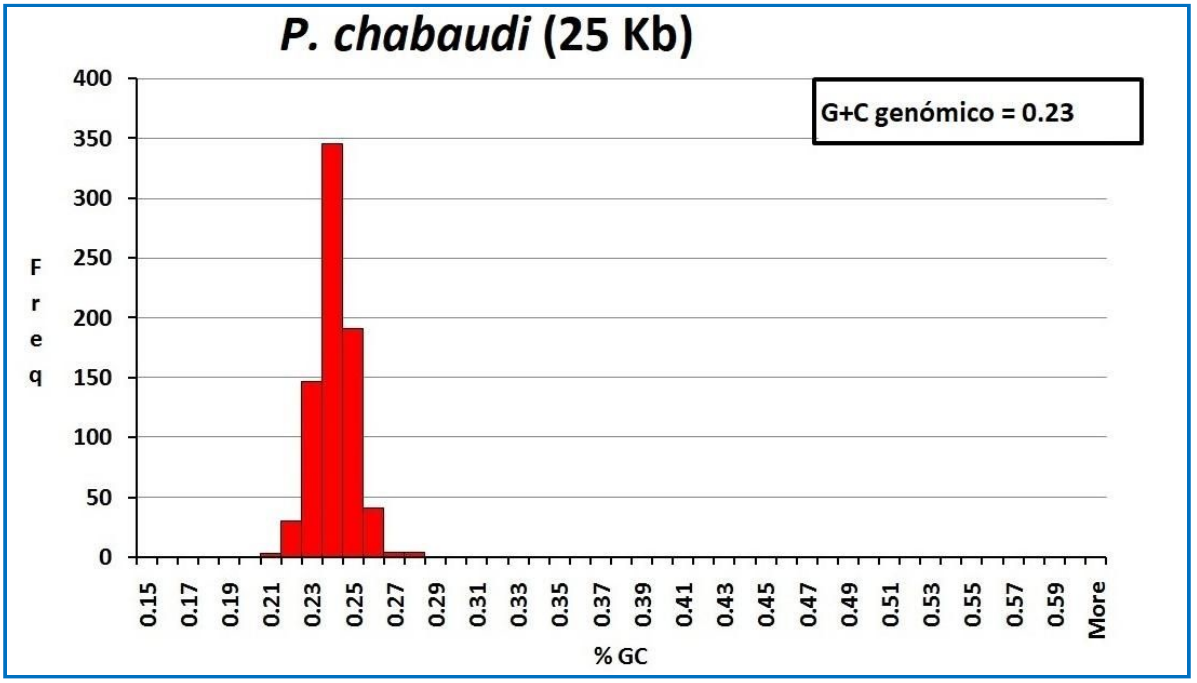


Figura 11: Distribución de contenido G+C de segmentos genómicos de 25 Kb de *P. chabaudi*

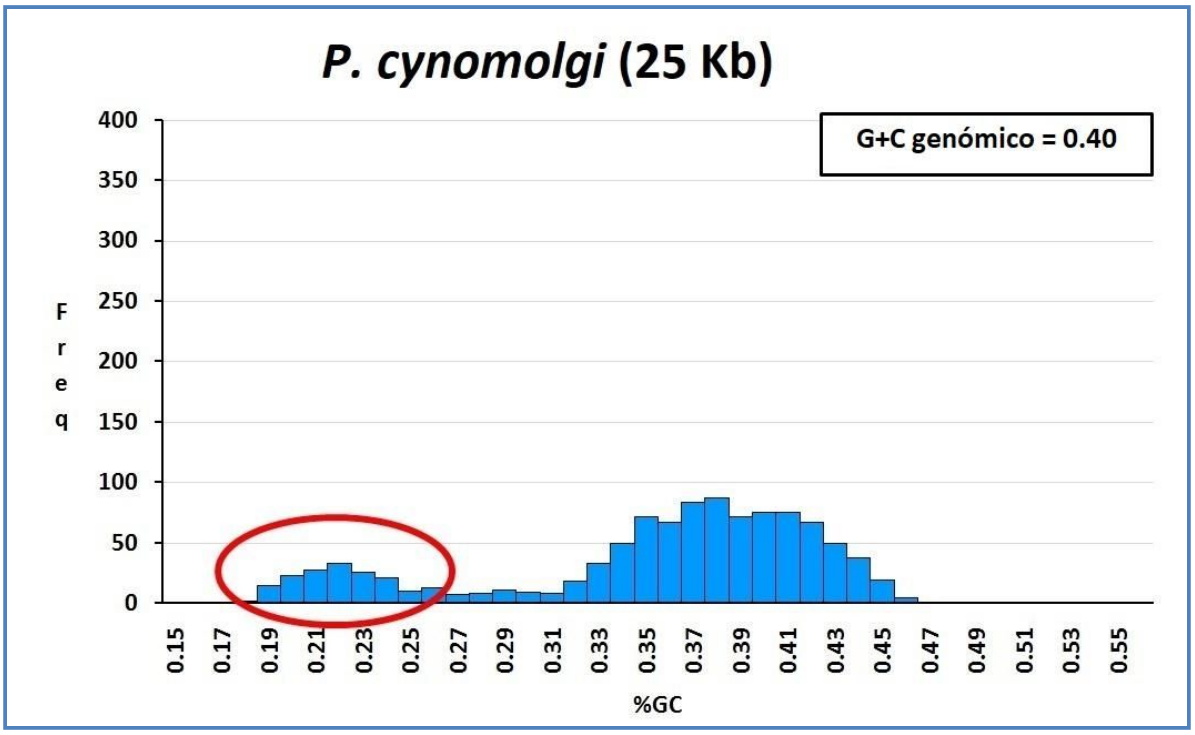


Figura 12: Distribución de contenido G+C de segmentos genómicos de 25 Kb de *P. cynomolgi*

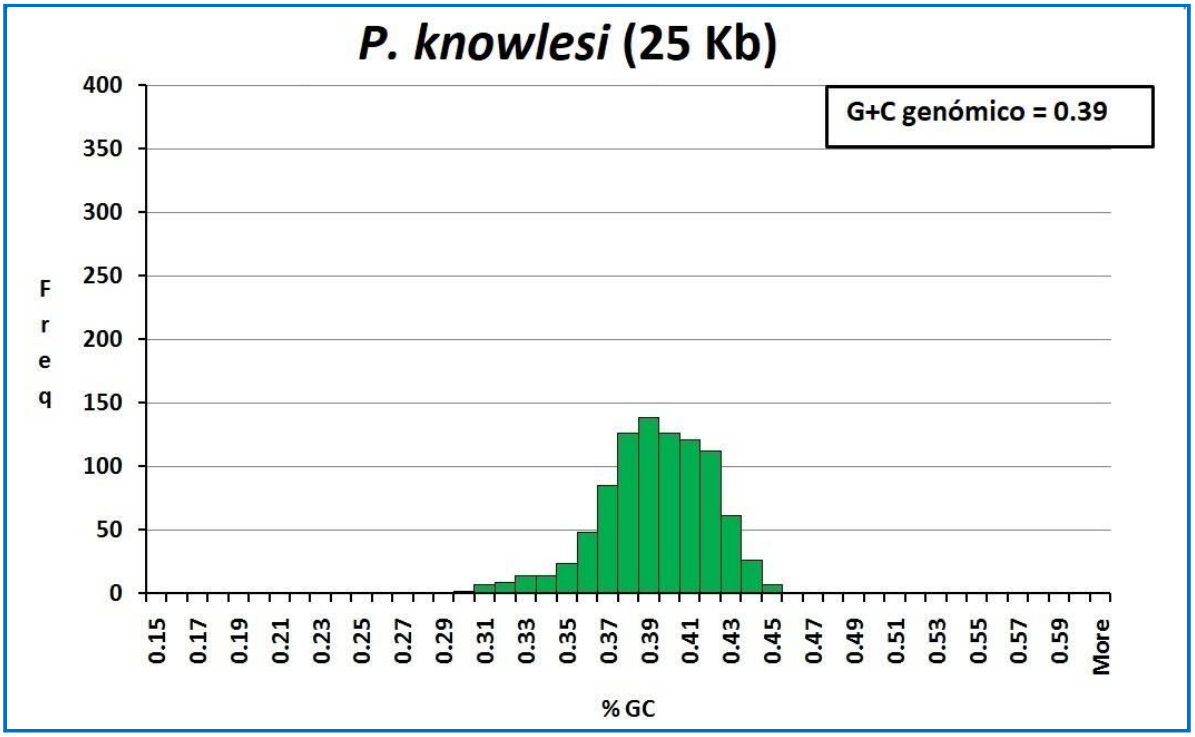


Figura 13: Distribución de contenido G+C de segmentos genómicos de 25 Kb de *P. knowlesi*

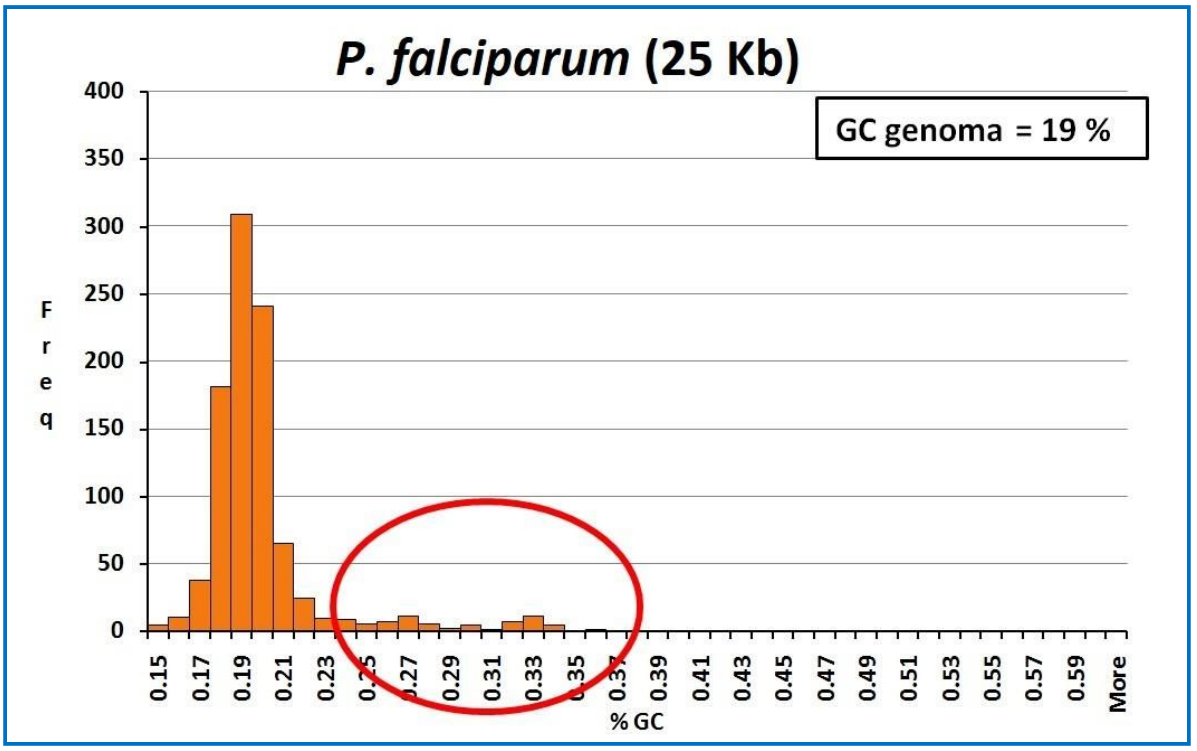


Figura 14: Distribución de contenido G+C de segmentos genómicos de 25 Kb de *P. falciparum*

Luego de replicar el procedimiento de segmentación realizado para *P. vivax* en otras especies de *Plasmodium*, observamos que los histogramas de los segmentos genómicos de bajo contenido G+C muestran comportamientos disímiles como podemos apreciar en la figuras 10, 11, 12,13 y 14. Algunas especies presentan un comportamiento similar a *P. vivax* con histogramas de distribución bimodal que agrupa en una campana segmentos con un contenido G+C cercanos al contenido G+C genómico global y una segunda campana correspondiente a segmentos con un contenido G+C notoriamente diferente como resumimos en la tabla 4.

Este es el caso de *P.cynomolgi,* la especie filogenéticamente más cercana a *P. vivax.* Su histograma es claramente bimodal con 2 campanas aún más definidas que en P. v*ivax,* pero con valores de G+C sustancialmente menores. Una de las campanas se distribuye alrededor de G+C=0.39, y la segunda en torno a G+C=0.22. *P. knowlesi*, especie también cercana a *P. vivax* muestra en cambio distribución unimodal, pero con un amplio rango de variación (figura 13)

P. berghei y P. chabaudi presentan una menor variabilidad, con segmentos con un contenido G+C similar al genómico y una distribución unimodal agrupada en torno a la media del

mismo**.** *P. falciparum* (la especie filogenéticamente más lejana dentro de las seleccionadas) presenta una distribución unimodal con una ligera anomalía en la cola derecha de la curva con un contenido G+C mayor al genómico, lo cual representa también un comportamiento diferencial de algunos segmentos comparados con el contenido G+C genómico.

Esto nos señala que la existencia de regiones con contenido G+C diferente al contenido G+C genómico total no es un comportamiento exclusivo de *P. vivax,* sino que por el contrario se repite en varias de las especies del mismo género.

Especie	**Distribución**	**Media campana/s**	**G+C genómico**
P. berghei	Unimodal	0.22	0.22
P. chabaudi	Unimodal	0.24	0.23
P. cynomolgi	Bimodal	0.22 y 0.39	0.4
P. knowlesi	Unimodal	0.40	0.39
P. falciparum	Unimodal	0.19	0.19
P. vivax	Bimodal	0.31 y 0.47	0.45

Table 4: Características de la distribución de segmentos genómicos de 25 Kpb. De Plasmodium

DISTRIBUCIÓN ESPACIAL POR CROMOSOMA DE LOS SEGMENTOS CON BAJO CONTENIDO G+C

La distribución espacial del contenido G+C a la interna de los cromosomas de *P. vivax* permite visualizar una alta variabilidad con un rango que va de 0.3 a 0.6. Como podemos apreciar en las figuras 15 a 18, la distribución del contenido G+C a lo largo de los cromosomas no es uniforme presentando una distribución con forma de campana achatada, en la que las zonas más pobres en G+C están localizadas en las regiones teloméricas y aumentan gradualmente hacia regiones centrales dibujando una meseta con segmentos de contenido G+C más alto.

De la observación de las figuras 9 y 15 a 18 se desprende que 0.39 es un valor razonable como límite entre las zonas consideradas como bajo contenido G+C o alto contenido G+C, pues éste es el punto medio entre las medianas de ambas distribuciones y además el análisis

por ventanas muestra que en las regiones ricas en G+C rara vez se pasa por debajo de esta barrera. Utilizando entonces este valor determinamos para cada cromosoma las fronteras de la región de bajo contenido G+C, la cual por su ubicación denominaremos “telomérica. Los valores con las coordenadas de inicio y fin de estas regiones se resumen en la tabla 5 y serán utilizados en varios de los estudios posteriores.

Es de destacar que esta distribución en forma de “campana achatada” se repite en todos los cromosomas excepto el 6, el cual como se puede observar en la figura 19, muestra una distribución llamativamente diferente presentando además de una zona de bajo contenido G+C en el telómero inicial, una bajada abrupta en contenido G+C entre los 300 y 450 kb, asimilables a lo que consideramos “zonas teloméricas” para luego retomar gradualmente el ascenso hacia la meseta (este cromosoma además no presenta descenso hacia el telómero final). Examinando el archivo de anotación de este cromosoma comprobamos que esta zona interna de bajo contenido G+C tiene una composición de elementos génicos equiparable a su zona telomérica inicial y a las zonas teloméricas del resto de los cromosomas. Esto significa que el bajo contenido G+C de la zona no es la única característica tipo telómero de esta zona. Las proteínas VIR son particularmente conspicuas en esta región, que como veremos a continuación es similar a lo que ocurre en las regiones teloméricas de los restantes cromosomas. Estos elementos tomados en conjunto indicarían que se trataría de una inversión cromosómica o una fusión de dos cromosomas.

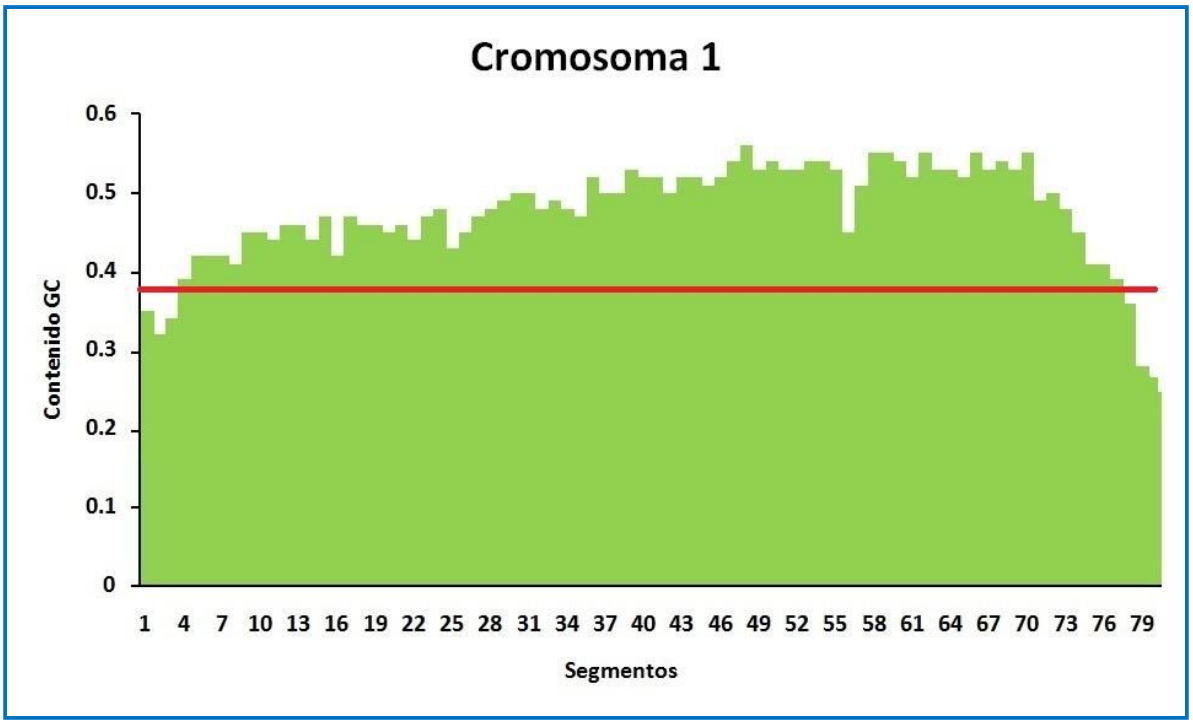


Figura 15: Distribución espacial del contenido G+C (segmentos de 10 Kpb.) En cromosoma 1 de *P. vivax*. La línea roja delimita los segmentos con bajo contenido G+C

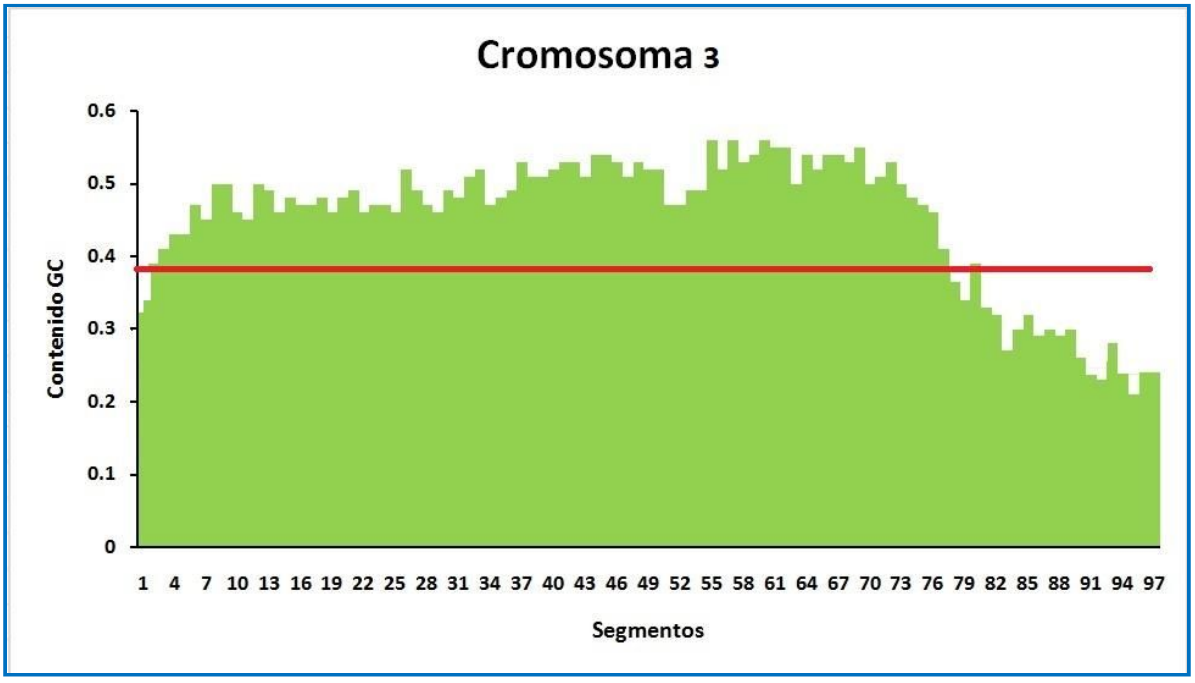


Figura 16: Distribución espacial del contenido G+C (segmentos de 10 Kpb.) En cromosoma 3 de *P. vivax*. La línea roja delimita los segmentos con bajo contenido G+C

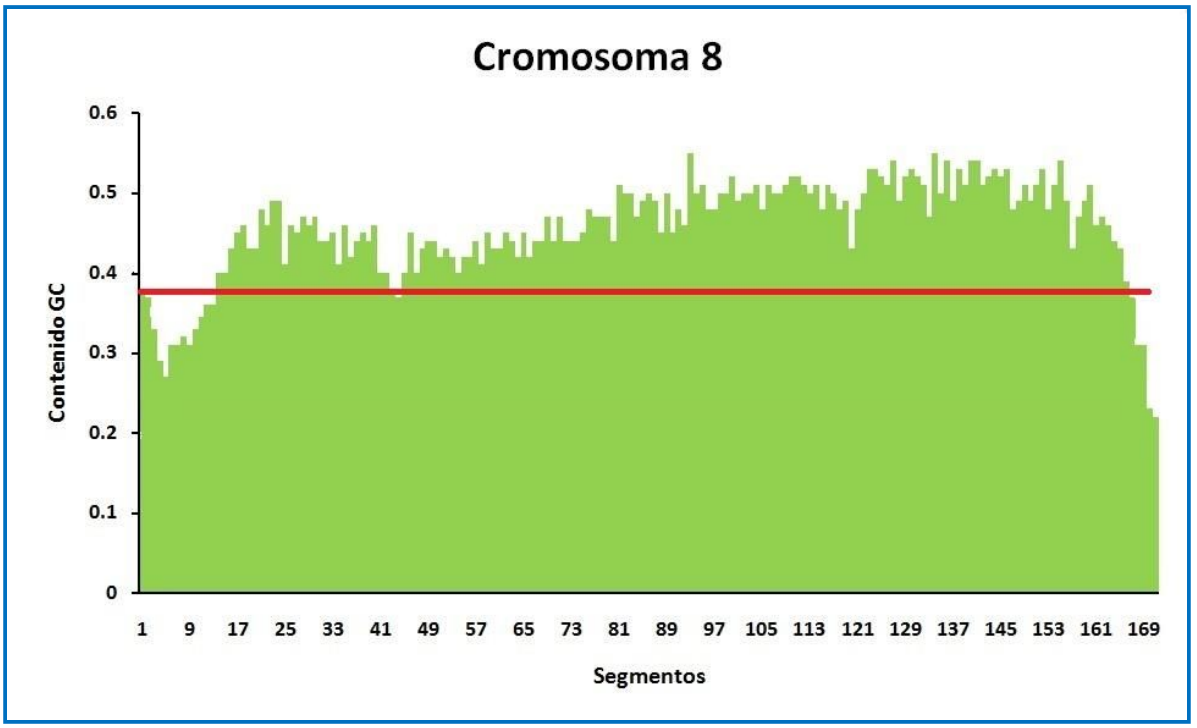


Figura 17: Distribución espacial del contenido G+C (segmentos de 10 Kpb.) En cromosoma 8 de *P. vivax*. La línea roja delimita los segmentos con bajo contenido G+C

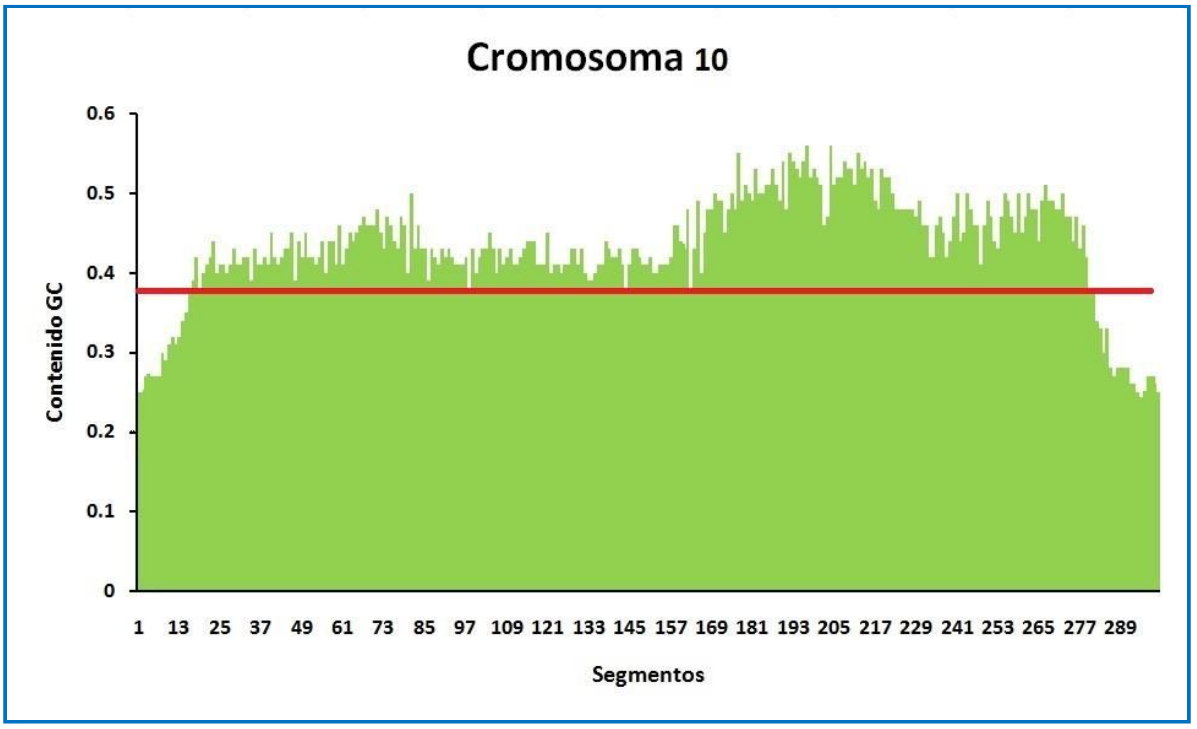


Figura 18: Distribución espacial del contenido G+C (segmentos de 10 Kpb.) En cromosoma 10 de *P. vivax*. La línea roja delimita los segmentos con bajo contenido G+C

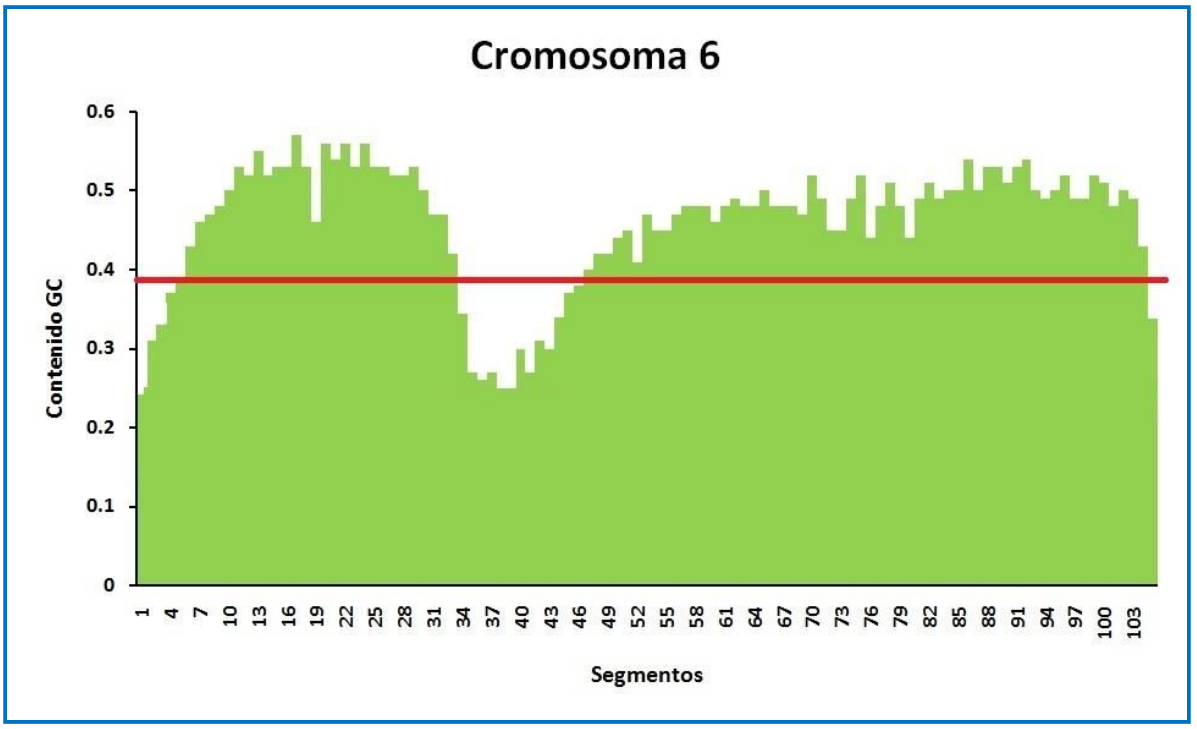


Figura 19: Distribución espacial del Contenido G+C de los segmentos de 10 Kpb. en el cromosoma 6 de *P. vivax*. La línea roja delimita los segmentos con bajo contenido G+C

Crom.	Tel. Ini		Tel final	
	Seg. Ini	Seg fin	Seg Ini	Seg fin
C1	0	3	80	83
C2	0	9	72	76
C3	0	1	81	100
C4	0	8	79	89
C5	0	11	130	139
C6	0	4	34	47
C7	0	3	143	152
C8	0	12	165	170
C9	0	3	192	195
C10	0	5	131	144
C11	0	5	200	209
C12	0	3	293	305
C13	0	3	204	206
C14	0	18	297	317

Zona telomérica inicial

Zona telomérica final

Tabla 5: Fronteras de las zonas de bajo contenido GC (teloméricas) de cada cromosoma

La secuencia de cada cromosoma puede ser clasificada en categorías funcionales/génicas en base a la información disponible en los archivos de anotación. En principio dividimos las secuencias génicas en 2 categorías: genes codificantes de proteínas y regiones intergénicas. Las regiones codificantes de proteínas a su vez las subdividimos en tipo vir y no vir utilizando la tipificación previamente definida en el trabajo de Carmen Becerra et al.(2006). Este tipo de proteína juega un rol fundamental en la evasión de la respuesta inmune del parásito y se abordará más detalladamente en estudios posteriores. También están presentes otros tipos de secuencias tales como ARNs de transferencia y micro ARNs. Sin embargo debido a su escasa representatividad, los mismos no fueron considerados en los análisis posteriores puesto que no es factible que estos afecten la composición de la región donde se ubican. A continuación hicimos una segunda clasificación agrupando las categorías por zona en base a su localización espacial dentro del cromosoma (telomérica, no-telomérica) definida anteriormente, resultando las clases detalladas en la tabla 6

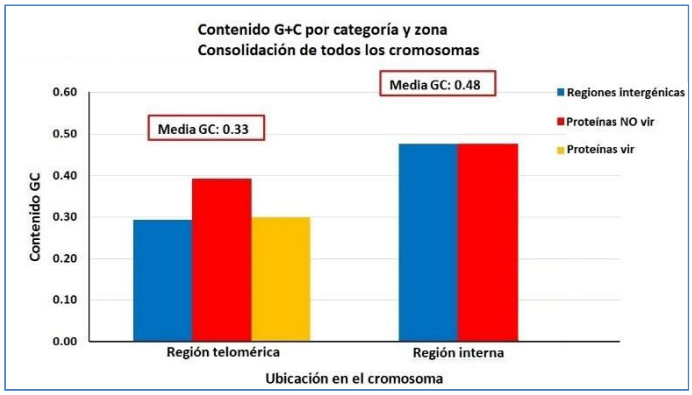
• Proteínas vir teloméricas • Proteínas no vir teloméricas • Regiones intergénicas teloméricas	• Proteínas vir zona interna • Proteínas no vir zona interna • Regiones intergénicas zona interna

Tabla 6: Clasificación de secuencias en clases (categoría funcional y zona)

Calculamos el contenido G+C para cada clase y la contribución proporcional de éstas a la zona que las contiene (o sea el largo porcentual de cada categoría con respecto al largo de cada zona). Como se observa en las figuras 20 y 21, al graficar el contenido G+C de cada clase observamos que en la zona telomérica el contenido G+C está entre 0.3 y 0.4 y en la zona media es uniformemente 0.48.

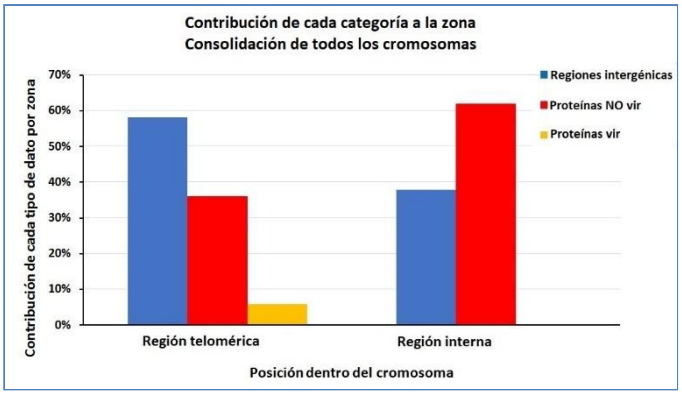
En ambos casos estos valores coinciden con la media de cada una de las campanas de la distribución bimodal en el histograma de contenido G+C de segmentos genómicos (figura 9). Surgen 3 consideraciones interesantes.

- El contenido G+C de los elementos génicos considerados es menor en las zonas teloméricas que en la zona central, y este comportamiento es independiente de la categoría considerada.
- Las proteínas vir se ubican exclusivamente en zonas teloméricas.
- Las regiones intergénicas son las mayores contribuyentes en la zona telomérica.



Categoría	GC por zona	
	Zona Telom.	Zona Interna
Regiones intergénicas	0.29	0.48
Proteínas no vir	0.39	0.48
Proteínas vir	0.30	
Media	0.33	0.48

Figura 20: Contenido G+C por categoría y zona. Las zonas son telomérica e interna y cada una de las barras representa una categoría (zonas intergénicas, proteínas no vir, proteínas vir)



Categoría	Tamaño(pb) por zona	
	Zona Telom.	Zona Interna
Total	2,212,173	19,885,108
Regiones intergénicas	1,286,887	7,533,964
Proteínas no vir	797,996	12,344,359
Proteínas vir	127,290	6,785

Figura 21: Contribución de cada categoría a la zona. Cada región de la gráfica representa una de las zonas consideradas (telomérica e interna) y las barras en cada zona determinan la relación porcentual del tamaño de la categoría con respecto al tamaño de la zona. La tabla a la derecha muestra los tamaños en pares de bases (pb) de cada categoría.

Si bien el contenido G+C de las proteínas vir y no vir localizadas en la zona telomérica es notoriamente más bajo que el contenido G+C de las proteínas localizadas en las zonas internas (fig. 21), la suma de la representatividad de proteínas vir y no vir en la zona telomérica (42%) es menor a la representatividad de las regiones intergénicas en la misma zona (58%). Sumado a esto, el contenido G+C de las regiones intergénicas también es más bajo en la zona telomérica que en la zona interna. Esto nos lleva a concluir que el bajo valor de contenido G+C telomérico no es un subproducto de la sobrerrepresentación de alguno de las categorías en particular, sino que por el contrario parece ser un comportamiento generalizado de la zona. Esto indicaría que fueron los componentes de estas regiones (proteínas, regiones intergénicas, etc.) los que adecuaron su composición a la composición de la zona que las contiene generándose de esta forma zonas de menor contenido G+C relativamente homogéneas.

ANÁLISIS DE COMPONENTES PRINCIPALES

El análisis de componentes principales (PCA) es una técnica de reducción de la dimensión que se utiliza regularmente en diversas formas de análisis, desde la neurociencia hasta el reconocimiento facial y la compresión de imágenes.

PCA intenta describir la información de un conjunto de variables observadas mediante un conjunto de variables más pequeño (las componentes principales). Estos componentes o nuevas variables no están correlacionadas y son combinaciones lineales de las variables de partida, y esperamos que solo unas pocas de esas componentes recojan la mayor parte de la información de los datos o sea que perdamos la menor cantidad de información posible.

Tenemos entonces un conjunto de m individuos a los que se les han medido n variables. Esto lo podemos representar como una matriz de datos m x n o visto de otra forma como un conjunto de m puntos en un hiperespacio de n dimensiones o ejes (X1, X2, ... Xn). El objetivo de PCA es realizar una representación de esa nube de m puntos en n dimensiones en un espacio reducido de p dimensiones ($p < n$) perdiendo la menor cantidad de información posible. En estadística, información es equivalente a dispersión, a varianza: una variable que no varía no tiene información, mientras que una variable que varía es muy informativa.

El procedimiento se basa en proyectar los puntos de la nube sobre nuevos ejes (Y1, Y2, ...Yn) de manera tal que la varianza sea lo mayor posible en los primeros p ejes. Las variables originales se han transformado en nuevas variables (las componentes), las cuales tienen desigualdad en cuanto a la información que aportan, lo que significa que tenemos unas componentes muy informativas (principales) y otras que no. Esta desigualdad generada al crear las componentes nos permite elegir, entre ellas, las p principales y eliminar las poco importantes, cosa que no sucedía con las originales porque todas eran igualmente importantes.

Para lograr estas nuevas variables o componentes principales, calculamos la matriz de covarianza de los datos originales y a esta matriz le calculamos los valores y vectores propios. El producto entre la matriz de vectores propios y los datos arroja la matriz de componentes principales.

El primer componente principal o sea el que recoge la mayor parte de la información es el que se deriva del vector propio con el mayor autovalor y así sucesivamente. El valor propio asociado a cada vector es proporcional a la varianza asociada al mismo

Ordenando los componentes de acuerdo a los valores propios de los autovalores que los generaron obtenemos un nuevo set de variables que aportan distintos grados de información. Quedándonos solamente con aquellos p vectores que aportan mayor información logramos reducir la dimensión del set de datos.

Como mencionamos en la sección “materiales y métodos”, las variables utilizadas para este análisis fueron las frecuencias de trinucleótidos. Tenemos 64 posibles trinucleótidos, lo cual hace muy complejo su análisis, por lo tanto la reducción de la dimensionalidad de estas variables aplicando PCA es una estrategia adecuada para el estudio de su comportamiento.

Con el fin de estimar la contribución de cada una de las categorías funcionales del genoma (proteínas vir, proteínas no vir, regiones intergénicas) a cada una de las regiones y determinar además si esta es homogénea, llevamos a cabo un análisis equivalente. Por tanto para cada una de las categorías de datos calculamos las frecuencias de trinucleótidos para ser usadas como variables de entrada en un PCA. Estas frecuencias fueron preprocesadas basados en las categorías funcionales especificadas en la tabla 6 (proteínas vir, proteínas no vir y regiones intergénicas) y las zonas especificadas en la tabla 5 (zona

telomérica y zona interna), de manera tal que una vez realizado el pca fuera posible identificar a que categoría y zona corresponde cada valor resultante.

Como se observa en la figura 22 se distinguen 2 clusters de puntos correspondientes a zonas telomérica e interna, que aunque tienen cierto grado de solapamiento, están diferenciados. Los puntos pertenecientes al área telomérica tienden a colocarse más hacia la izquierda de la gráfica, con un centro aproximado en las coordenadas 0.0, -0.05 Cuando discriminamos las categorías funcionales dentro de los clusters, se puede apreciar que el comportamiento es relativamente homogéneo, es decir en todas las categorías, mientras que los clusters que agrupan los puntos del área cromosómica central se sitúan a la derecha.

Esto no nos proporciona ninguna información concluyente acerca del comportamiento de cada categoría, pero es indicio de que algo está sucediendo que diferencia las zonas teloméricas de las zonas cromosómicas.

Si el carácter “contenido G+C telomérico” es un comportamiento generalizado de la zona telomérica, es de esperar que los trinucleótidos pobres en G+C tengan preferencia por las zonas teloméricas.

Para obtener respuestas más precisas decidimos hacer dos estudios estadísticos extra. Un histograma de los valores del primer componente principal de frecuencia de trinucleótidos discriminado por categoría funcional y zona, y una correlación entre primer componente de PCA de trinucleótidos y el contenido G+C por clase. Los histogramas del primer componente discriminando de acuerdo a las categorías de la tabla 5, permiten visualizar la distribución de sus valores y ver si aparecen más patrones. En las figuras 25, 26 y 27 observamos que hay un corrimiento de la curva de distribución hacia la izquierda en las zonas teloméricas independientemente de la categoría, coincidiendo con el patrón observado en las figuras 23 y 24.

Este corrimiento hacia la izquierda de la campana del conjunto de las zonas teloméricas, independiente de la categoría funcional, refuerza la idea de que existe alguna restricción funcional o estructural, que subyace a esta diferencia composicional que está asociada al comportamiento de la zona.

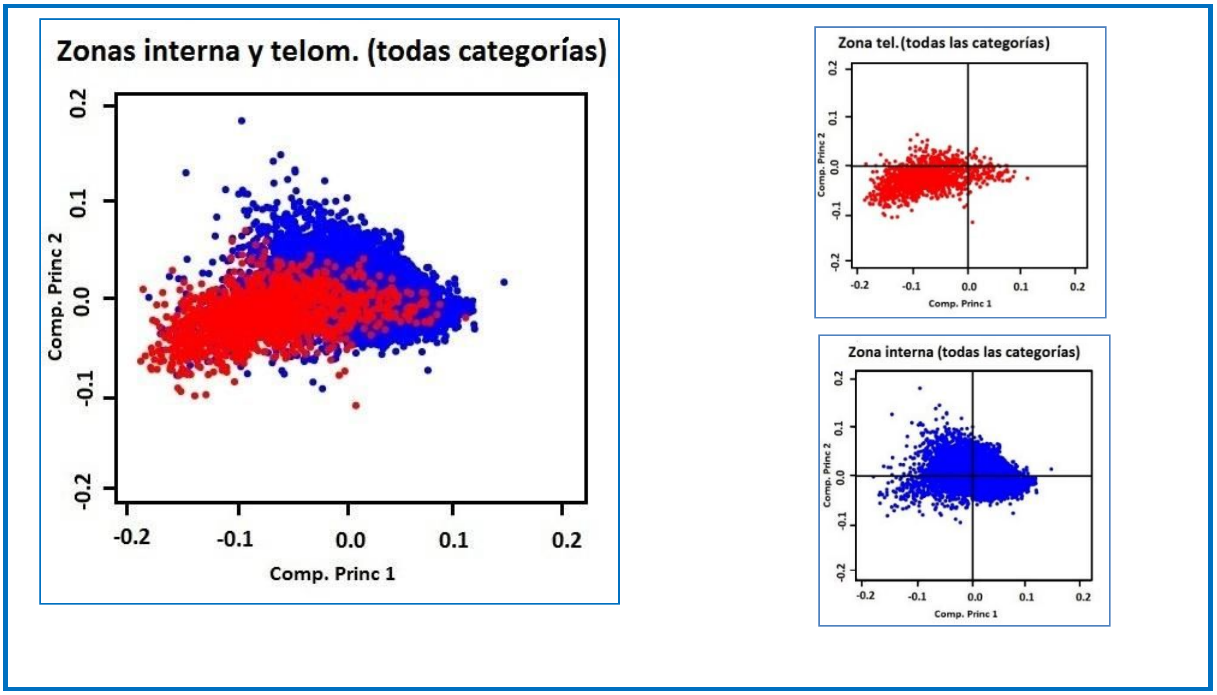


Figura 22: Gráfica de dispersión de los dos primeros componentes principales de frecuencias de trinucleótidos. Estamos diferenciando solamente por zona por lo tanto los puntos rojos pueden corresponder a cualquiera de las categorías funcionales (proteínas vir, proteínas no vir, región intergénicas) ubicadas en la zona telomérica y los puntos azules a cualquiera de las categorías en la zona interna

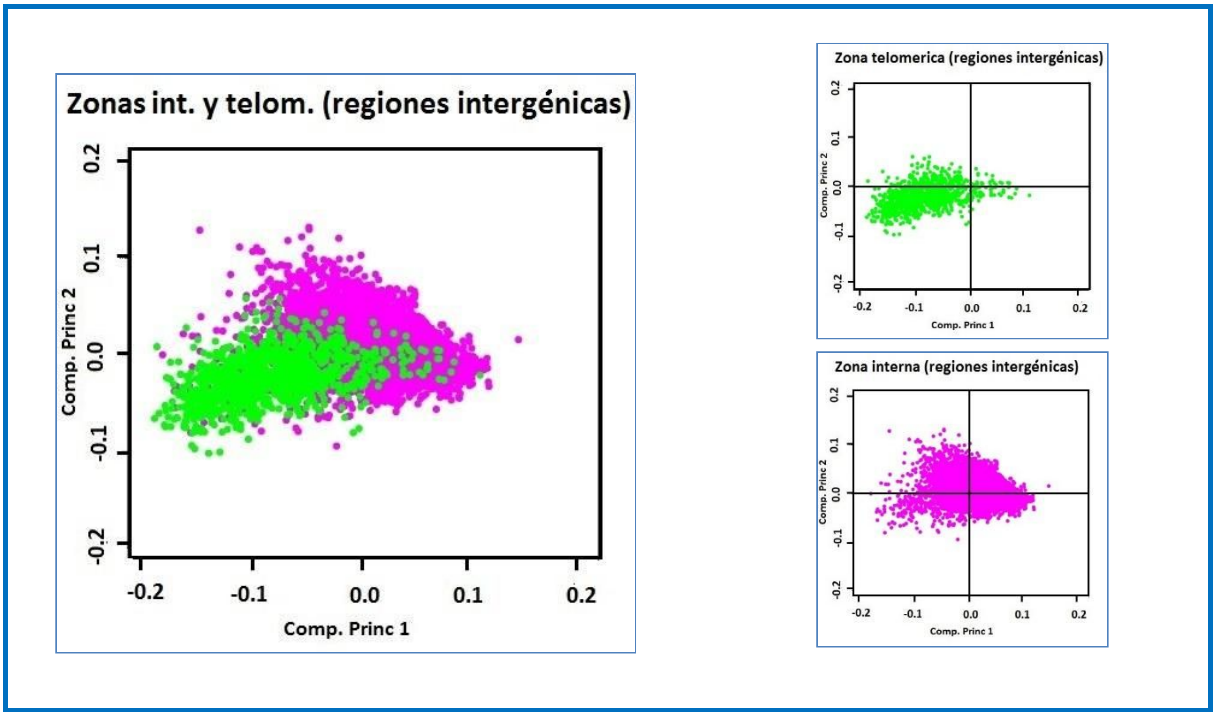


Figura 23: Gráfica de dispersión de los dos primeros componentes principales de frecuencias de trinucleótidos de zonas intergénicas Estamos diferenciando categoría y zona, por ende los puntos verdes corresponden a regiones intergénicas localizadas en la zona telomérica y los puntos rosa a zonas intergénicas localizadas en la zona interna.

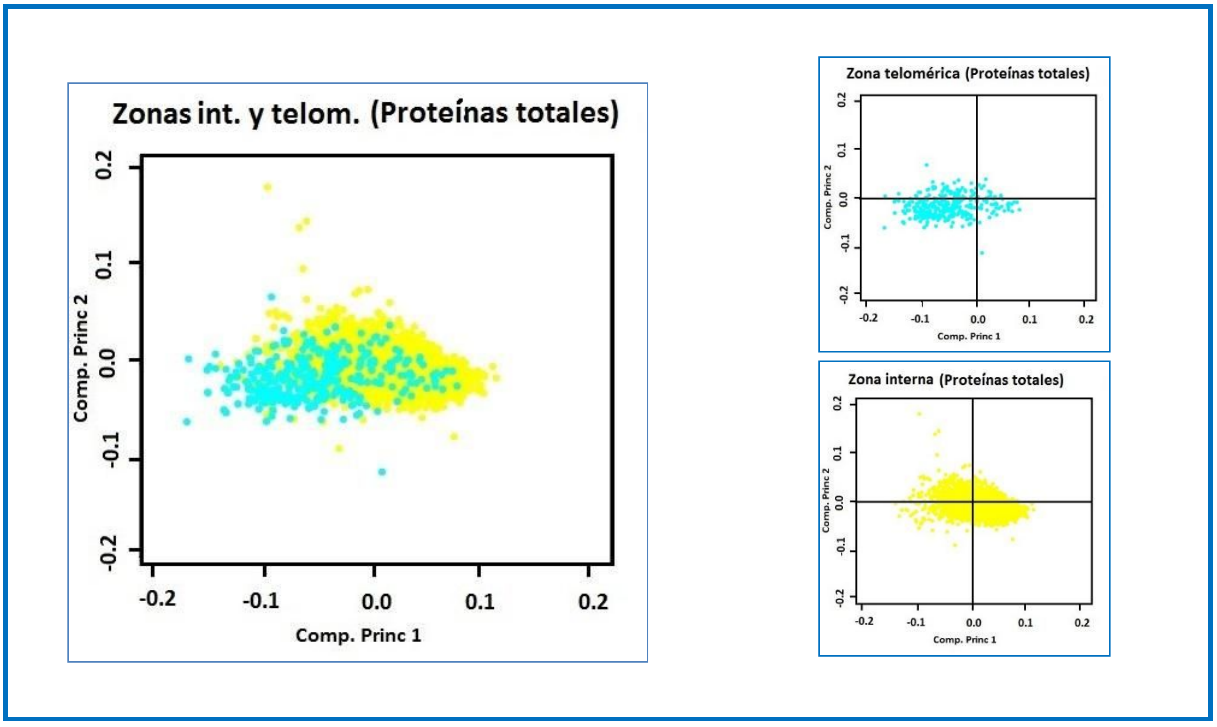


Figura 24: Gráfica de dispersión de los dos primeros componentes principales de frecuencias de trinucleótidos de proteínas totales discriminando por zonas. Estamos diferenciando categoría y zona. Los puntos celestes corresponden a proteínas localizadas en la zona telomérica y los puntos amarillos a proteínas localizadas en la zona interna.

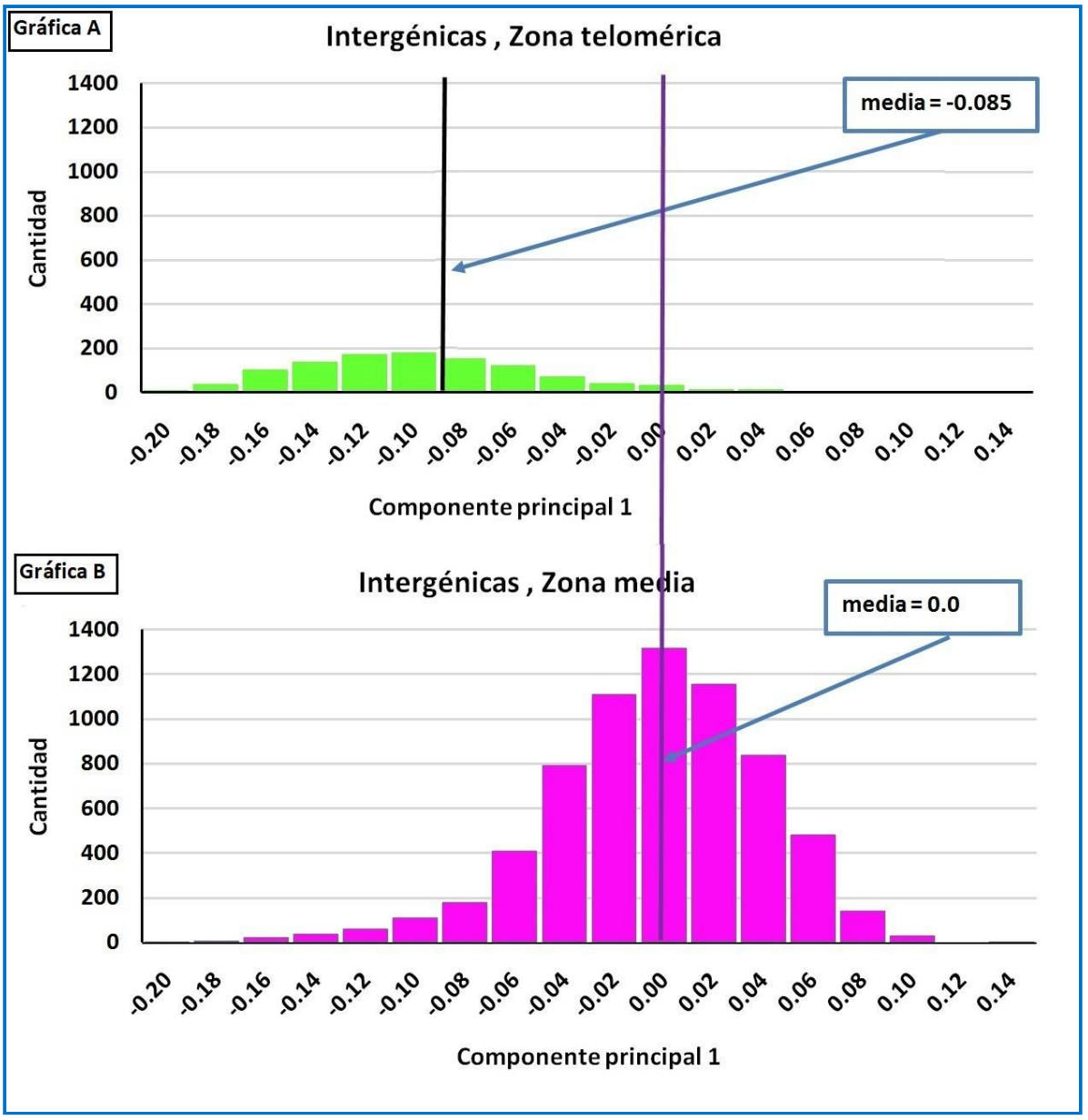


Figura 25: Distribución de los valores del primer componente principal de frecuencia de trinucleótidos, categoría Regiones Intergénicas. Las gráficas se discriminan por categoría y zona. La gráfica A corresponde a zonas teloméricas y la gráfica B a zonas cromosómicas internas. La línea violeta que atraviesa las gráficas en forma vertical parte de la media del primer componente en la zona interna para esa categoría y las líneas negras representan la media en las zonas teloméricas.

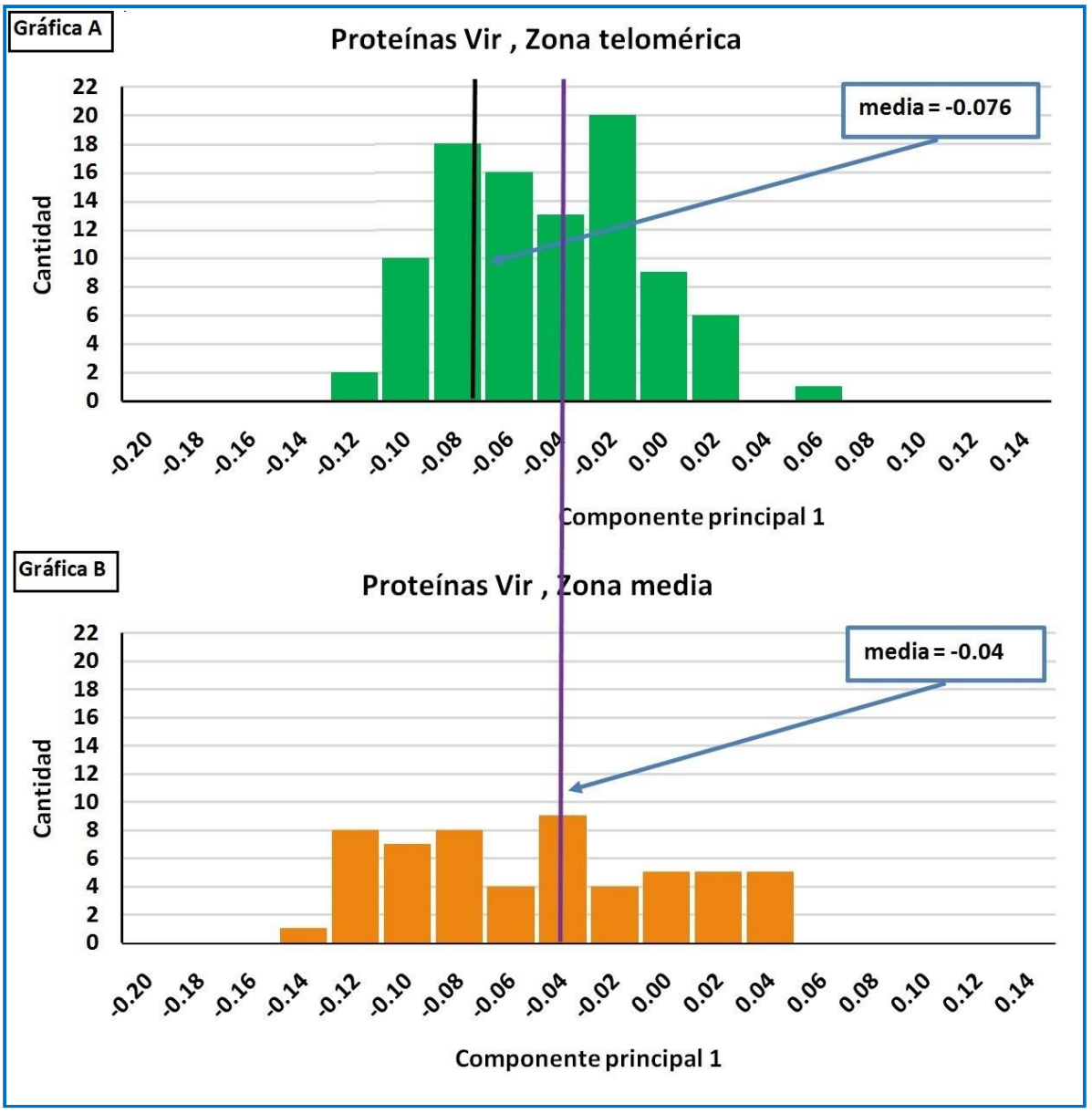


Figura 26: Distribución de los valores del primer componente principal de frecuencia de trinucleótidos, categoría Proteínas VIR. Las gráficas se discriminan por categoría y zona. La gráfica A corresponde a zonas teloméricas y la gráfica B a zonas cromosómicas internas. La línea violeta que atraviesa las gráficas en forma vertical parte de la media del primer componente en la zona interna para esa categoría y las líneas negras representan la media en las zonas teloméricas.

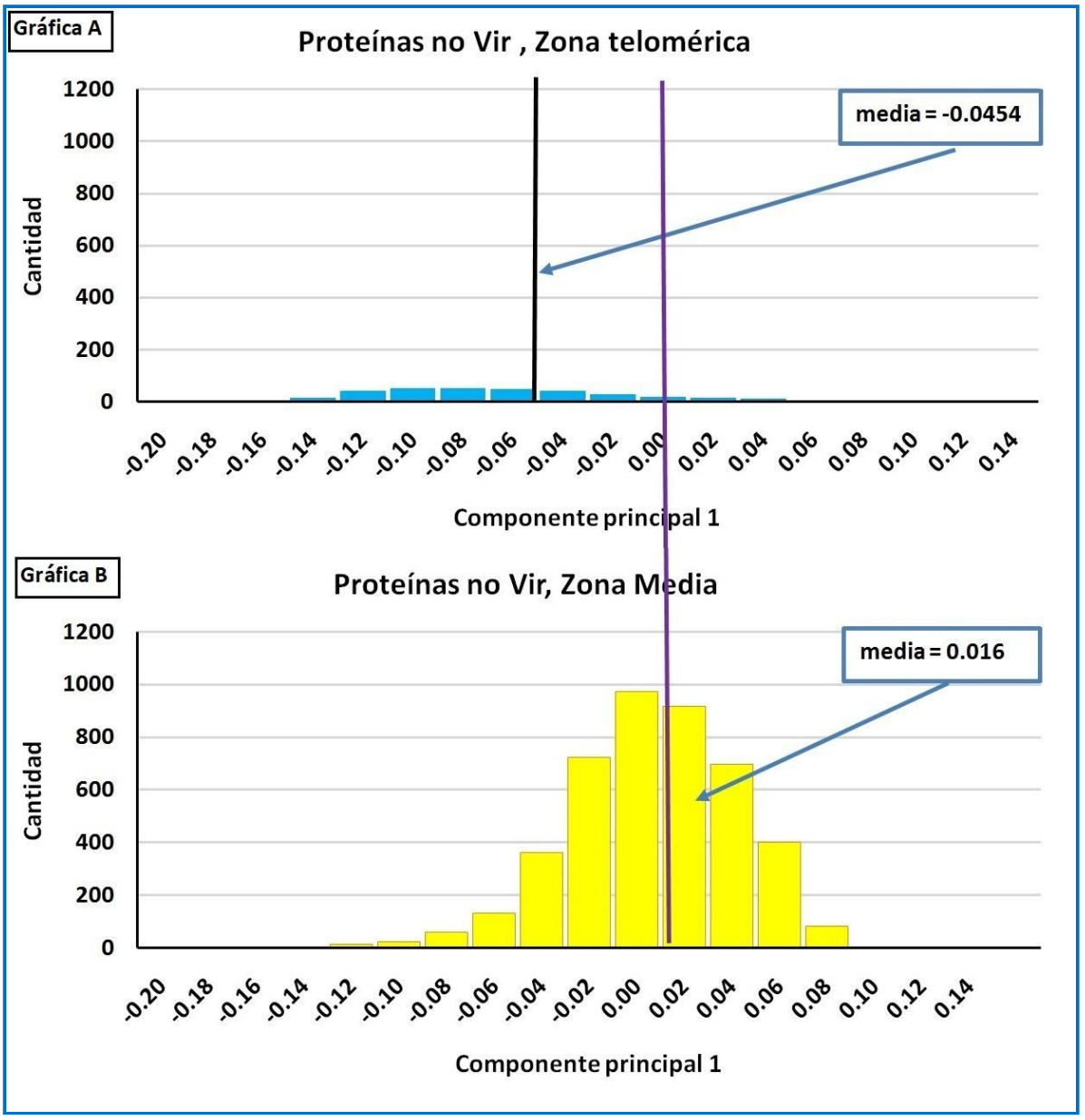


Figura 27: Distribución de los valores del primer componente principal de frecuencia de trinucleótidos, categoría Proteínas no VIR. Las gráficas se discriminan por categoría y zona. La gráfica A corresponde a zonas teloméricas y la gráfica B a zonas cromosómicas internas. La línea violeta que atraviesa las gráficas en forma vertical parte de la media del primer componente en la zona interna para esa categoría y las líneas negras representan la media en las zonas teloméricas.

Como nuestro interés se centra en descubrir si esta conducta diferente de los nucleótidos en las diferentes zonas está asociada al contenido G+C de las mismas, graficamos la relación entre el primer componente principal y el contenido G+C de cada clase. Como se refleja en las figuras 28, 29 y 30 y se detalla en la tabla 7, estos están fuertemente correlacionados. Esto refuerza la hipótesis formulada en el capítulo Resultados sección "distribución espacial

por cromosoma de los segmentos con bajo contenido G+C" de que el contenido G+C diferencial en cada zona es debido a una propiedad intrínseca de la zona y no a una característica o particularidad de alguno de los datos.

Categoría funcional	**Zona telomérica**	**Zona interna**
Regiones intergénicas	0.992823	0.989626
Proteínas no vir	0.991327	0.988575
Proteínas vir	0.987721	0.992198

Tabla 7: Coeficientes de correlación entre 1er componente de PCA de trinucleótidos y contenido G+C discriminado por categoría funcional y zona

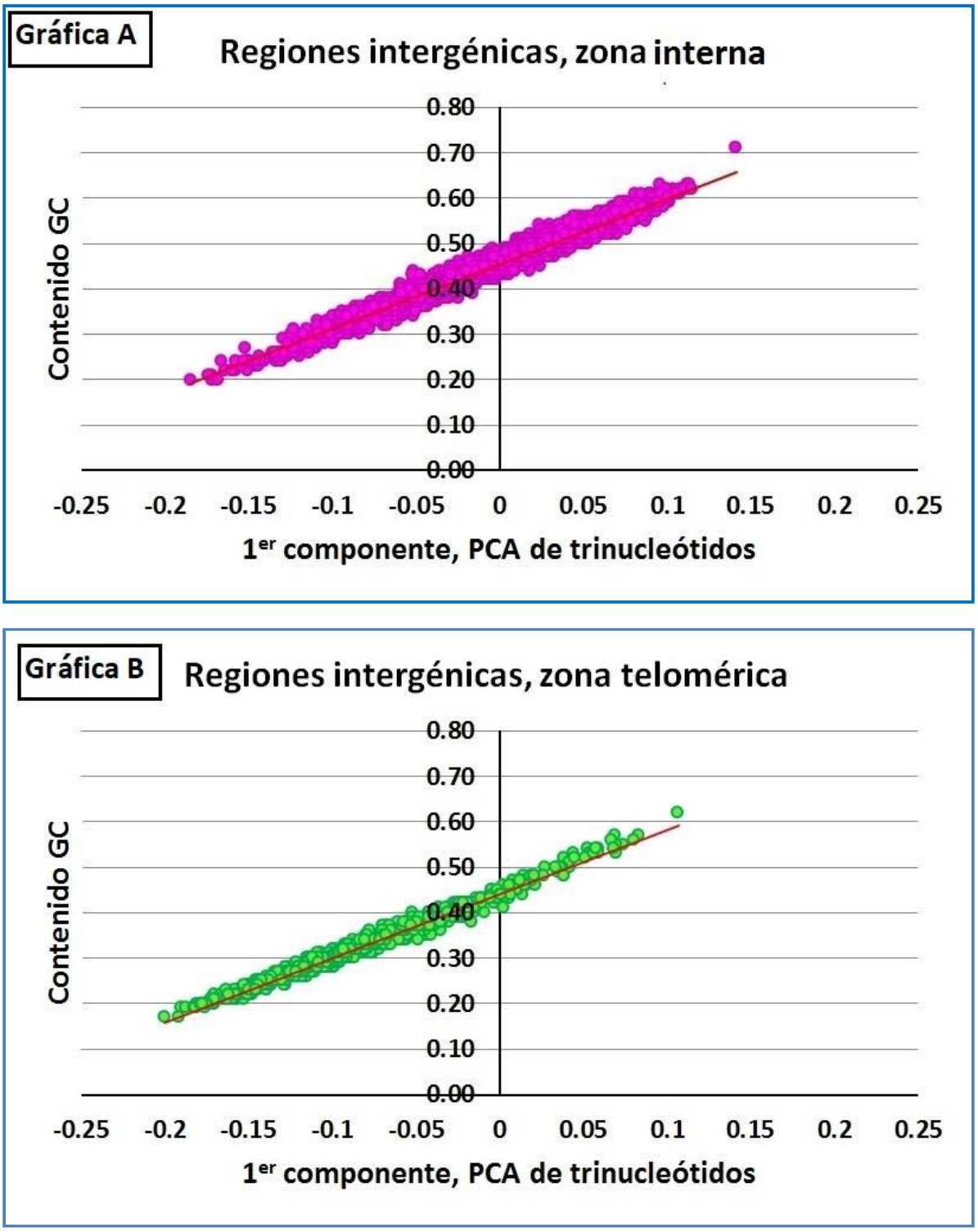


Figura 28: Correlación entre primer componente de PCA de trinucleótidos y contenido G+C por clase para la categoría funcional "Regiones intergénicas". La gráfica A corresponde a zona interna al cromosoma. La gráfica B corresponde a zona telomérica.

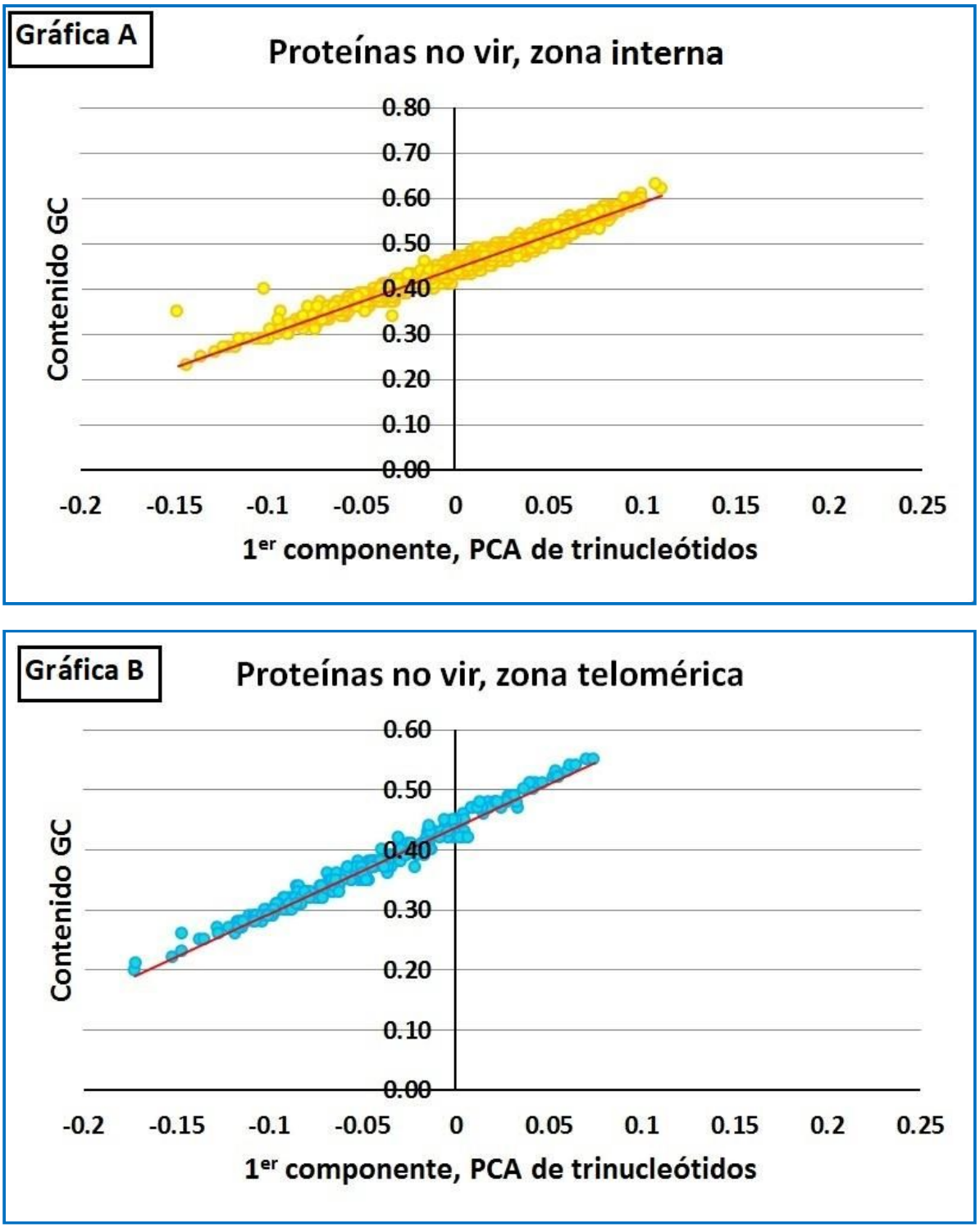
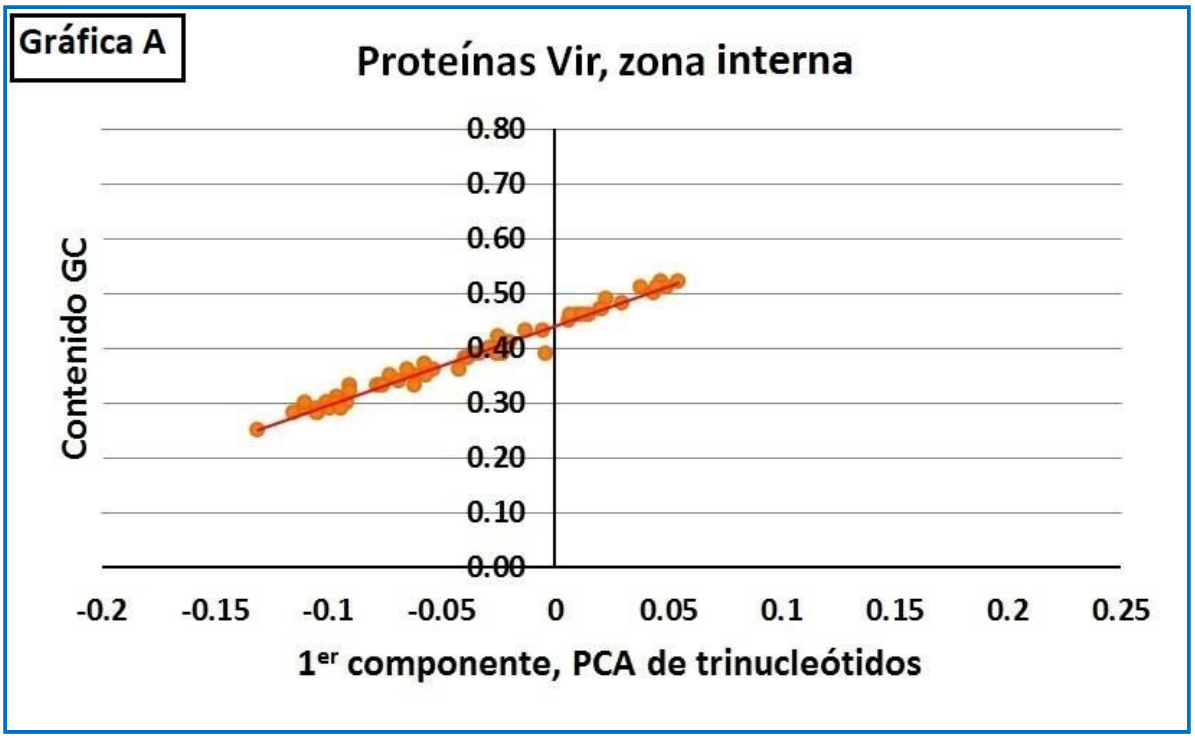


Figura 29: Correlación entre primer componente de PCA de trinucleótidos y contenido G+C por clase para la categoría funcional "Proteínas no VIR". La gráfica A corresponde a zona interna al cromosoma. La gráfica B corresponde a zona telomérica



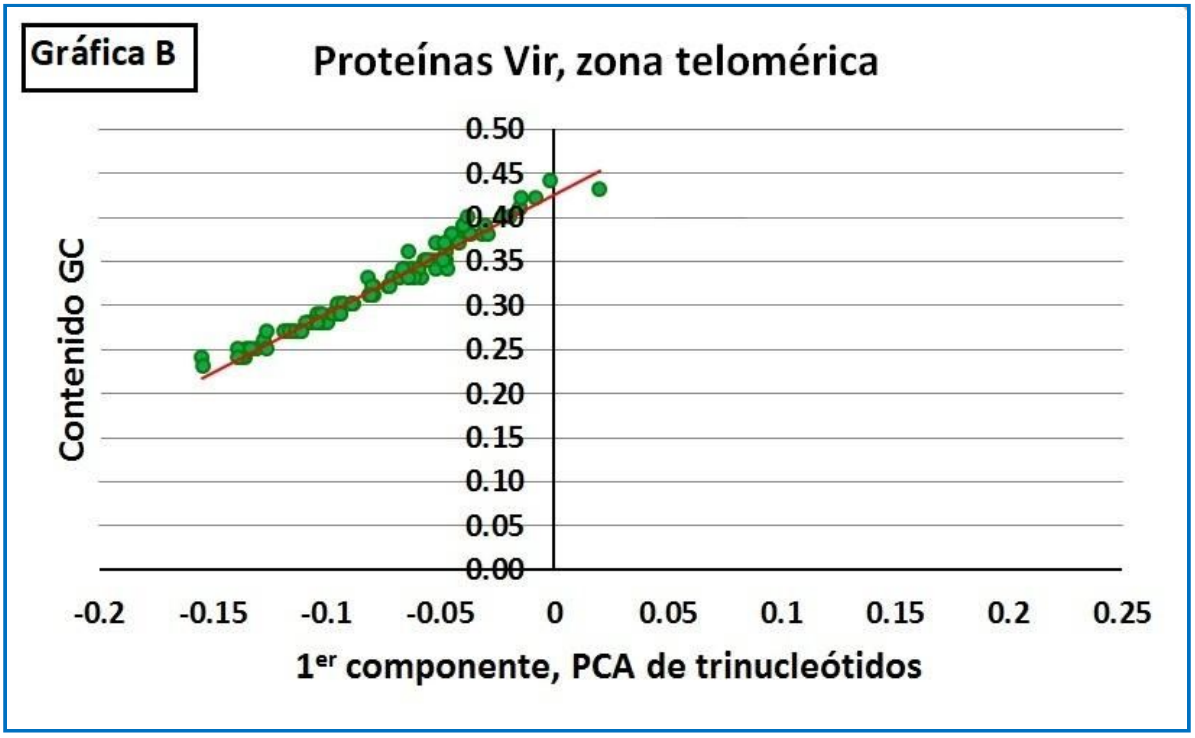


Figura 30: Correlación entre primer componente de PCA de trinucleótidos y contenido G+C por clase para la categoría funcional "Proteínas no VIR". La gráfica A corresponde a zona interna al cromosoma. La gráfica B corresponde a zona telomérica

DISTRIBUCIÓN ESPACIAL EN PROTEÍNAS CON BAJO CONTENIDO G+C DE PLASMODIOS

Resulta interesante investigar si existe un patrón de comportamiento similar en cuanto a la distribución espacial en las proteínas pertenecientes a sectores de bajo contenido G+C en *P. vivax* y sus homólogas en las especies de *Plasmodium* emparentadas

Tanto en *P. vivax* como en el resto de las especies emparentadas podemos conocer la ubicación espacial de las proteínas usando archivos de anotación. Luego de identificar por medio de blastp los genes ortólogos a los de *P. vivax* en las restantes especies analizadas (ver detalle en tabla 8), chequeamos si la ubicación es equivalente o no en ambas especies comparadas.

Para estas 5 especies analizadas (detalladas en la Tabla 8), definimos la zona telomérica como la región inicial o final del cromosoma, con un largo correspondiente al 15% del largo del genoma. Las proteínas homólogas identificadas serán declaradas teloméricas o no en base a su pertenencia a estas zonas.

	Cantidad de proteínas en especies relacionadas que son homólogas a proteínas en *P. vivax*	
Especie	Hit contra prot. telom. en *P. vivax*	Hit contra prot. No telom. en *P. vivax*
P. berghei	160	4429
P. chabaudi	172	4549
P. cynomolgi	343	4690
P. falciparum	221	4657
P. knowlesi	311	4613

Tabla 8: Cantidad de proteínas homólogas a proteínas en *P. vivax* luego de realizar blastp comparando *P.* vivax contra especies relacionadas de *Plasmodium.* La columna 2 contiene la cantidad de proteínas homólogas a proteínas teloméricas en *P. vivax*. La columna 3 contiene la cantidad de proteínas homólogas a proteínas no teloméricas en *P. vivax*

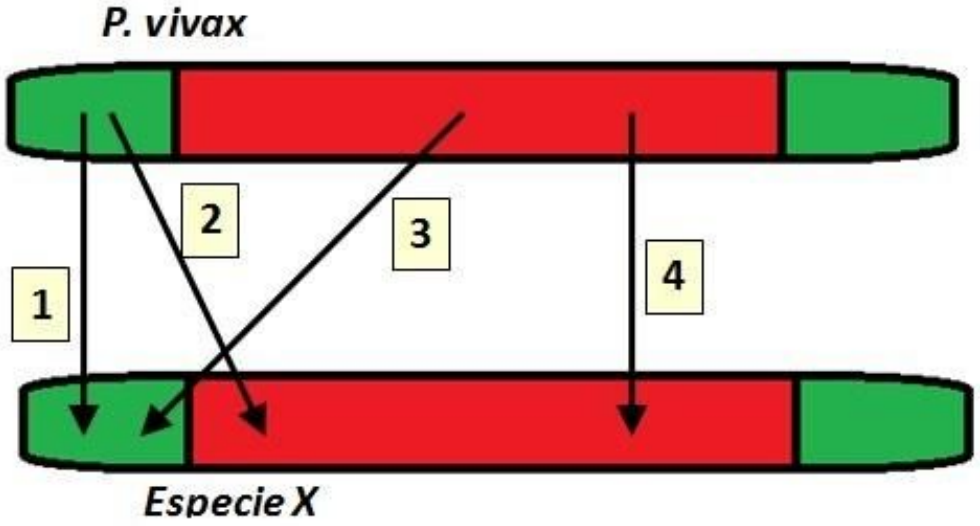
Se puede entonces establecer una clasificación de estas proteínas homólogas en las siguientes 4 categorías (figura 31)

- Proteínas teloméricas en especie X que son homólogas a proteínas teloméricas en *P. vivax*
- Proteínas teloméricas en especie X que son homólogas a proteínas NO teloméricas en *P. vivax*
- Proteínas NO teloméricas en especie X que son homólogas a proteínas teloméricas en *P. vivax*
- Proteínas NO teloméricas en especie X que son homólogas a proteínas NO teloméricas en *P. vivax*

Como se puede observar en la figura 32, hay una correlación entre la ubicación espacial de las proteínas teloméricas pertenecientes a *P. vivax y* sus homólogas en las especies emparentadas. Por el contrario, esta reciprocidad no se aprecia de forma muy clara entre las proteínas no teloméricas de *P. vivax* y sus homólogas en otras especies, ya que las mismas exhiben una distribución espacial más uniforme dentro del cromosoma.

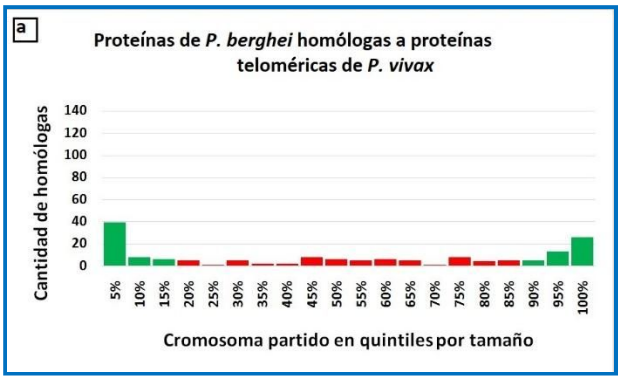
Se testó la asociación visual de las gráficas en la figura 32 usando una tabla de contingencia 2x2 test χ^2. Como podemos ver en la figura 33, los valores calculados para χ^2 superan en todos los casos los valores críticos de χ^2 para 1 grado de libertad, confirmando lo que se aprecia visualmente de correlación entre la ubicación espacial de las proteínas teloméricas pertenecientes a *P. vivax y* sus homólogas.

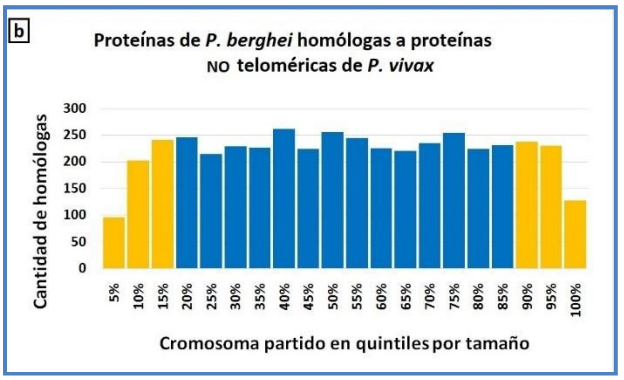
Un detalle a observar es la cantidad de proteínas con ubicación teloméricas en *P. vivax* cuyas homólogas pertenecen a zonas no teloméricas de los otros *Plasmodium*. Pensamos que esto se debe a errores introducidos por la falta de precisión al determinar las fronteras teloméricas de las especies emparentadas. Estas fronteras son más difusas y por lo tanto probablemente estemos catalogando como no teloméricas muchas proteínas que sí lo son.

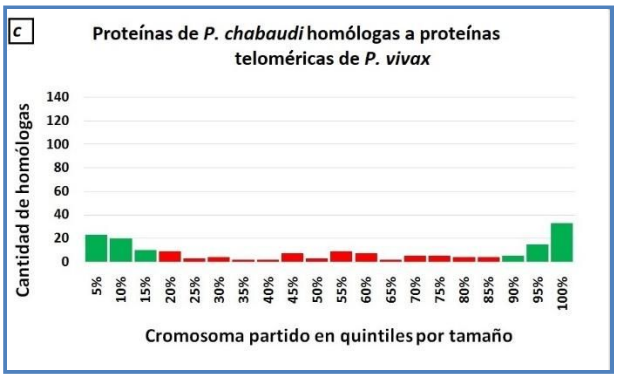


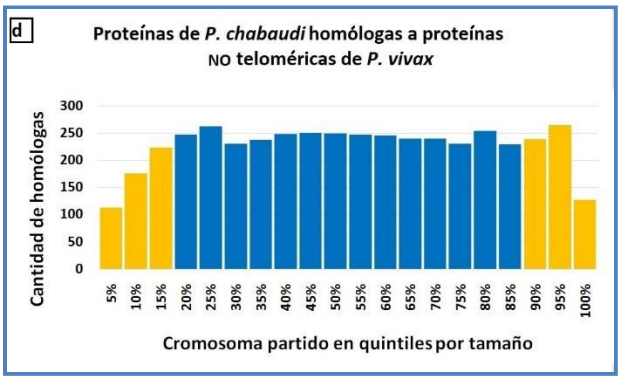
Ubicación de las proteínas en *P. vivax* (origen)	Ubicación de las proteínas Homólogas (Destino)	
	telomérica	No telomérica
telomérica	Homólogas teloméricas de proteínas *P. vivax* teloméricas (1)	Homólogas no teloméricas de proteínas *P. vivax* teloméricas (2)
No telomérica	Homólogas teloméricas de proteínas *P. vivax* no teloméricas (3)	Homólogas no teloméricas de proteínas *P. vivax* no teloméricas (4)

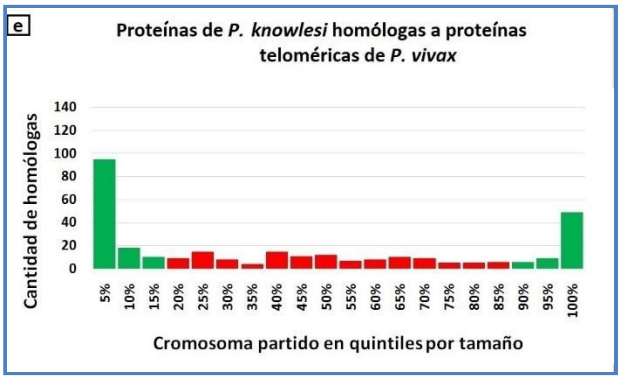
Figura 31: Clasificación de las proteínas homólogas a *P. vivax* en especies emparentadas de acuerdo con la ubicación espacial de la proteína en *P. vivax* y la homóloga en la especie (X) que está siendo considerada.

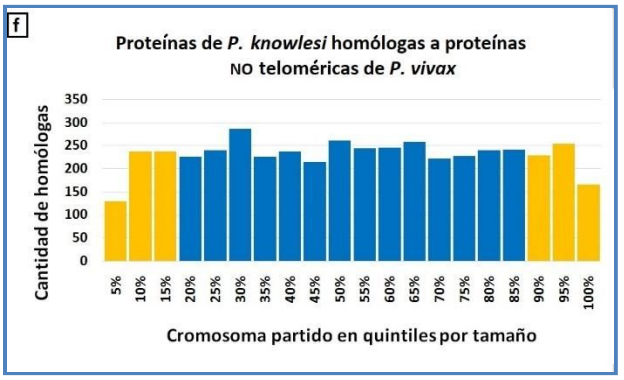

a
Proteínas de P. berghei homólogas a proteínas teloméricas de P. vivax
Cantidad de homólogas
Cromosoma partido en quintiles por tamaño

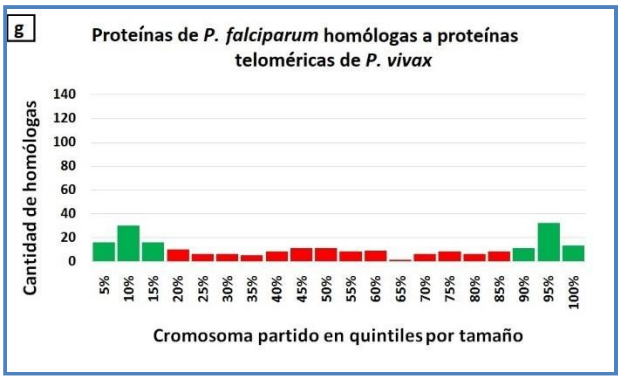

b
Proteínas de P. berghei homólogas a proteínas NO teloméricas de P. vivax
Cantidad de homólogas
Cromosoma partido en quintiles por tamaño

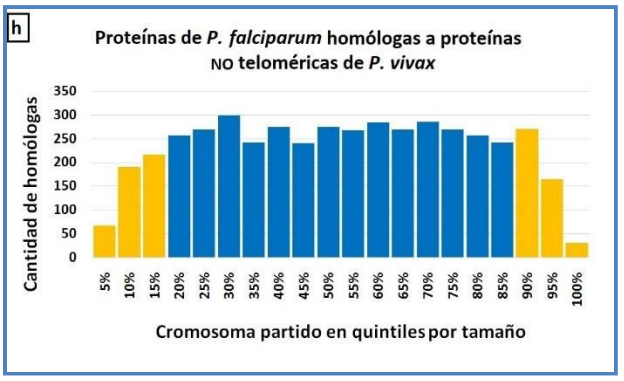

c
Proteínas de P. chabaudi homólogas a proteínas teloméricas de P. vivax
Cantidad de homólogas
Cromosoma partido en quintiles por tamaño


d
Proteínas de P. chabaudi homólogas a proteínas NO teloméricas de P. vivax
Cantidad de homólogas
Cromosoma partido en quintiles por tamaño


e
Proteínas de P. knowlesi homólogas a proteínas teloméricas de P. vivax
Cantidad de homólogas
Cromosoma partido en quintiles por tamaño


f
Proteínas de P. knowlesi homólogas a proteínas NO teloméricas de P. vivax
Cantidad de homólogas
Cromosoma partido en quintiles por tamaño


g
Proteínas de P. falciparum homólogas a proteínas teloméricas de P. vivax
Cantidad de homólogas
Cromosoma partido en quintiles por tamaño


h
Proteínas de P. falciparum homólogas a proteínas NO teloméricas de P. vivax
Cantidad de homólogas
Cromosoma partido en quintiles por tamaño

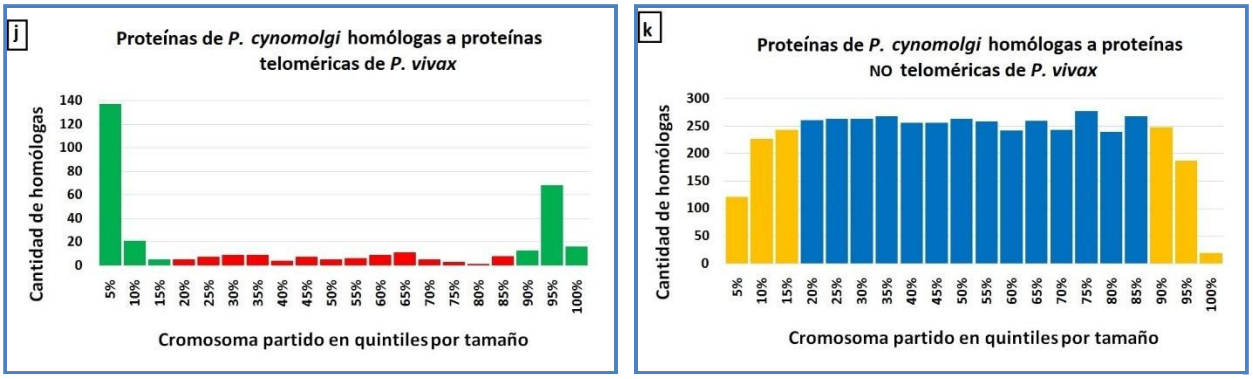


Figura 32: Distribución espacial de proteínas homólogas a las de *P. vivax*. Las gráficas en el cuadrante izquierdo representan homólogas a proteínas teloméricas. Están coloreadas en verde aquellas regiones consideradas teloméricas en el destino, y en rojo las no teloméricas.

P. berghei

Observaciones

Origen \ Destino	PB tel	PB notel	Marginal
PV tel	97	63	160
PV notel	1,134	3,292	4,426
Marginal	1,231	3,355	4,586

Frecuencias absolutas teóricas

Origen \ Destino	PB tel	PB notel
PV tel	43	117
PV notel	1,188	3,238

X^2 = 96.35

P. chabaudi

Origen \ Destino	PC tel	PC notel	Marginal
PV tel	106	66	172
PV notel	1,143	3,406	4,549
Marginal	1,249	3,472	4,721

Origen \ Destino	PB tel	PB notel
PV tel	46	126
PV notel	1,203	3,346

X^2 = 113.49

P. falciparum

Origen \ Destino	PF tel	PF notel	Marginal
PV tel	118	103	221
PV notel	937	3,720	4,657
Marginal	1,055	3,823	4,878

Origen \ Destino	PB tel	PB notel
PV tel	48	173
PV notel	1,007	3,650

X^2 = 137.81

P. knowlesi

Origen \ Destino	PK tel	PK notel	Marginal
PV tel	187	124	311
PV notel	1,251	3,362	4,613
Marginal	1,438	3,486	4,924

Origen \ Destino	PB tel	PB notel
PV tel	91	220
PV notel	1,347	3,266

X^2 = 153.55

P. cynomolgi

Origen \ Destino	PCy tel	PCy notel	Marginal
PV tel	260	89	349
PV notel	1,045	3,609	4,654
Marginal	1,305	3,698	5,003

Origen \ Destino	PCy tel	PCy notel
PV tel	91	258
PV notel	1,214	3,440

X^2 = 456.10

Figura 33: Test independencia (Contingencia usando X2 de Pearson) de distribución de proteínas homólogas a *P. vivax*.

El valor de χ^2 para gl=1 y alfa = 1e-10 es 42 por lo que todos los valores son significativos a este nivel.

POSIBLE ROL DE LA ESTRUCTURA DEL ADN Y LA CROMATINA EN LA COMPARTIMENTALIZACIÓN GENÓMICA

La observación de que el bajo contenido G+C se asocia con las zonas teloméricas nos lleva a postular que el motivo del mismo podría estar relacionado a la conformación estructural de la zona telomérica y al plegamiento del ADN en esa zona. Teniendo en cuenta estudios previos que evidencian la existencia de posibles implicancias de la secuencia de ADN y el plegamiento del mismo [Goodsell y Dickerson, 1994] decidimos continuar la investigación por ese camino y analizar el ADN en busca de relaciones entre la secuencia, el contenido G+C y la curvatura. Como ya fuera mencionado el ADN kinetoplástico fue usado con fines comparativos. En el cálculo de curvatura del ADN de los minicírculos de los kinetoplastos utilizando BANANA, la media de los valores de curvatura es 20, por lo tanto consideramos curvatura muy alta a aquellos valores superiores a 20.

Como se puede apreciar en la figura 34, que muestra la curvatura del ADN del cromosoma 1 en la zona de transición de la región telomérica a la región central del cromosoma, las zonas de bajo contenido G+C (información resumida en el cuadro ubicado en el extremo superior derecho de la gráfica) se corresponden con una mayor cantidad de picos (puntos con curvatura local > 20) y por lo tanto de mayor curvatura global. Siguiendo el mismo patrón las zonas con un contenido más alto de G+C se corresponden a una menor curvatura y menor cantidad de picos.

Como podemos apreciar en la gráfica 34, debido al volumen de datos y la alta variabilidad de un punto al otro, los datos graficados resultan difíciles de visualizar, por lo tanto tomamos ventanas de 10 Kpb. y silenciamos las zonas con curvatura menor que 20 reemplazando el valor de la misma por 0.

En estos gráficos más "limpios" superpusimos curvatura global, picos de curvatura mayores a 20 y contenido G+C de ventanas de 10 Kpb. y se observa claramente que cada máximo local en la curvatura se corresponde con mucha fidelidad con un mínimo local en el contenido G+C. Para todos los cromosomas, los coeficientes de correlación son negativos con valores en el rango de -0.6 a -0.8 para un valor p de 2 e-16 (tabla 9). Esto define una correlación inversamente proporcional entre curvatura y contenido G+C con una muy buena

significación estadística. Estos valores confirman la hipótesis formulada de que el bajo contenido G+C telomérico está relacionado con una alta curvatura de la zona (figuras 35, 36 y 37).

RELACIÓN ENTRE CURVATURA Y CONTENIDO G+C EN OTRAS ESPECIES DEL GÉNERO PLASMODIUM

En las especies emparentadas consideradas *(P. berghei, P. chabaudi, P. cynomolgi, P. knowlesi y P. falciparum)* se realizaron estudios relacionados a curvatura de ADN equivalentes a los hechos en *P. vivax* (cálculo de curvatura y comparación con contenido G+C). En términos generales los resultados son muy similares para todos los cromosomas de las especies emparentadas A modo de ejemplo discutimos los resultados en 3 casos tomados aleatoriamente. Como se puede visualizar en la figura 38, *P. cynomolgi,* la especie filogenéticamente más cercana y la que exhibe una variabilidad composicional con una distribución bimodal muy similar a *P. vivax* (figura 12**)** presenta una correlación inversa muy fuerte con un coeficiente de correlación de -0.77 y un valor p de 4.67 e-17. (tabla 10). *P. knowlesi,* cuya gráfica de variabilidad composicional muestra una distribución unimodal con una campana centrada en torno del valor de contenido G+C genómico (figura 13), muestra una correlación inversa con un coeficiente de correlación de -0.39. (figura 39). Este es un valor de correlación notoriamente más bajo que el correspondiente a los cromosomas de *P. vivax* o *P. cynomolgi* aunque el valor p también es estadísticamente significativo.

En *P. falciparum*, la especie filogenéticamente más alejada, el coeficiente es de 0.3 con un valor p de 0.90 e-4 (tabla 10). Esto indica una correlación positiva poco acentuada y con una significación estadística más baja que en los otros casos (figura 40). En conclusión, podemos afirmar que en mayor o menor grado parece haber una relación entre el contenido G+C y la curvatura del ADN tanto en *P. vivax* como en las especies emparentadas.

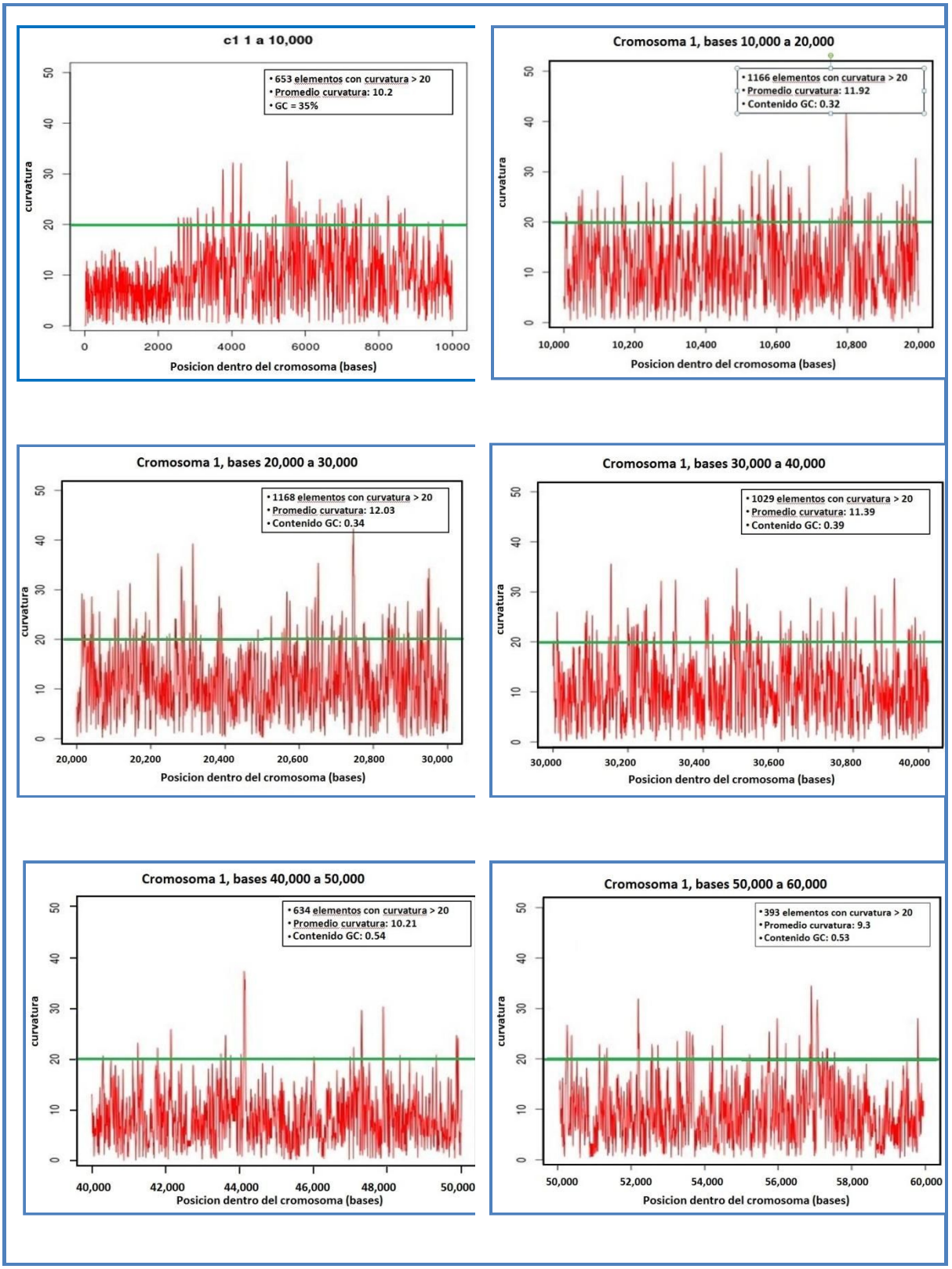


Figura 34: Curvatura de segmentos de 10 Kpb. del cromosoma 1 de *P. vivax*. La línea verde marca el límite que consideramos "curvatura alta" en comparación con la curvatura promedio de un minicírculo kinetoplástico. En el cuadro, en el extremo superior derecho de la gráfica, se indica la cantidad de puntos con curvatura local > 20, el promedio de curvatura y el contenido G+C del segmento.

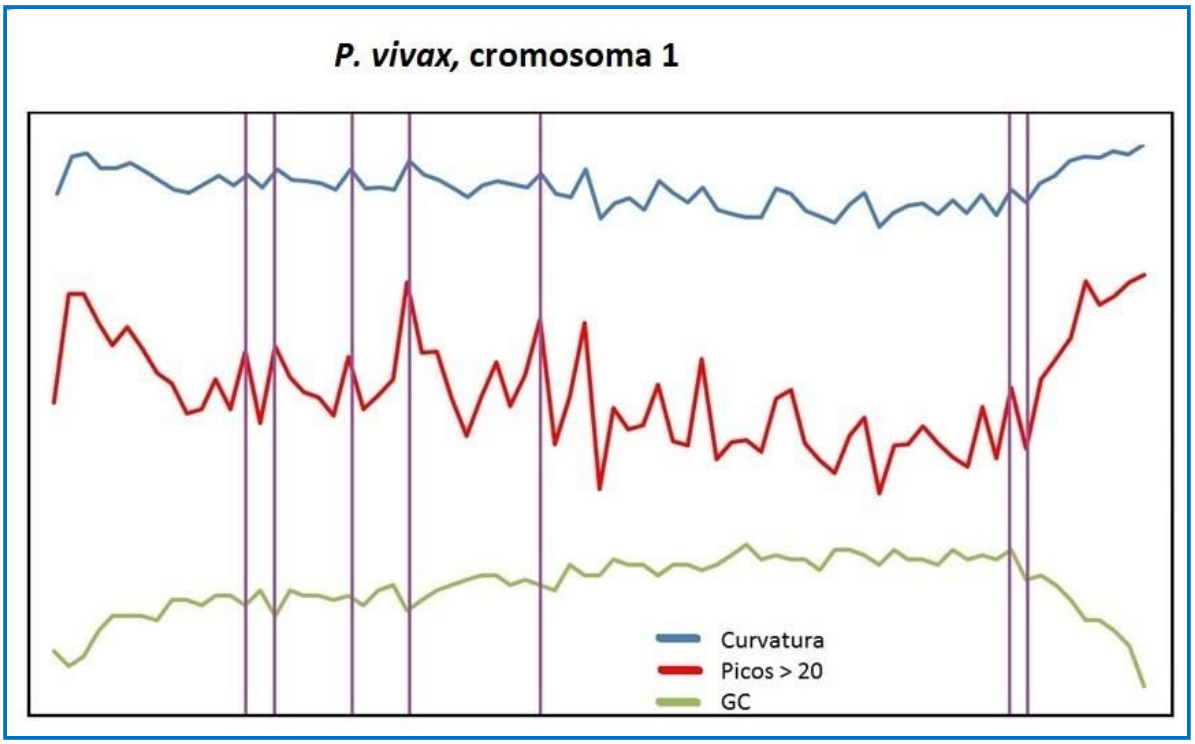


Figura 35: Superposición de gráfica de curvatura global, picos de curvatura local mayor a 20 y contenido G+C en cromosoma 1 de *P. vivax.* Las líneas violeta que atraviesan la gráfica marcan puntos de correspondencia que facilitan la visualización de la correlación entre las 3 variables

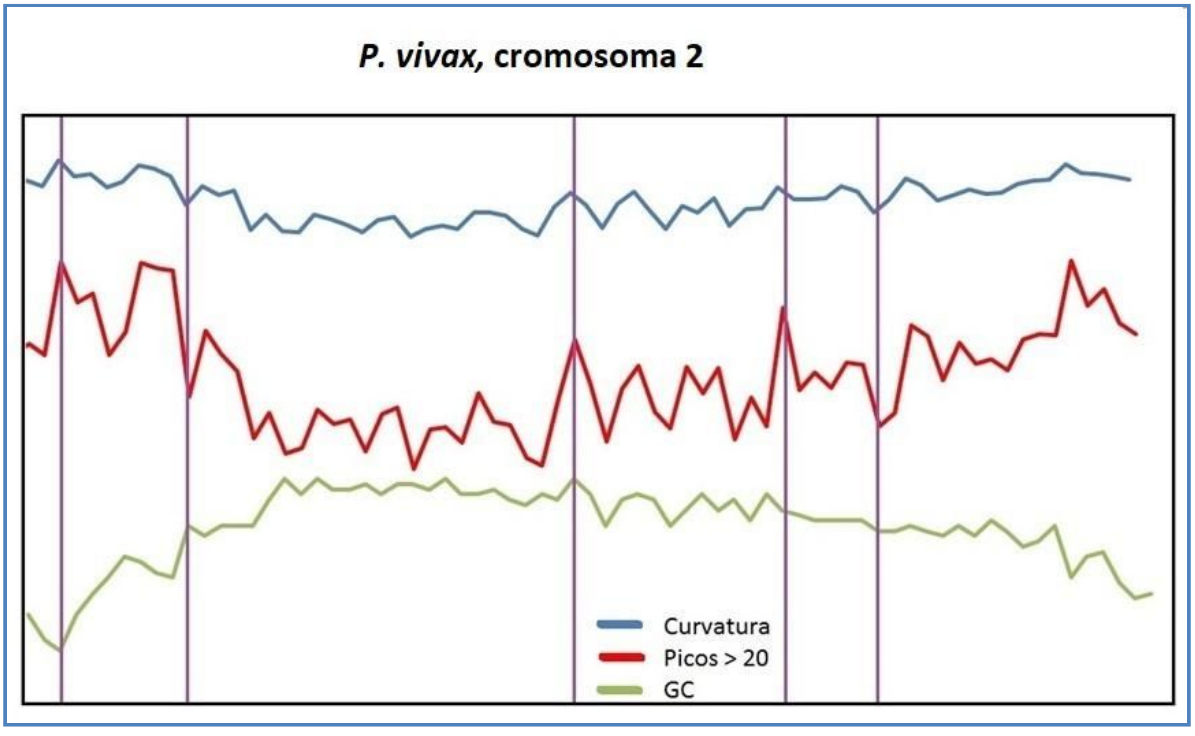


Figura 36: Superposición de gráfica de curvatura global, picos de curvatura local mayor a 20 y contenido G+C en cromosoma 2de *P. vivax.* Las líneas violeta que atraviesan la gráfica marcan puntos de correspondencia que facilitan la visualización de la correlación entre las 3 variables

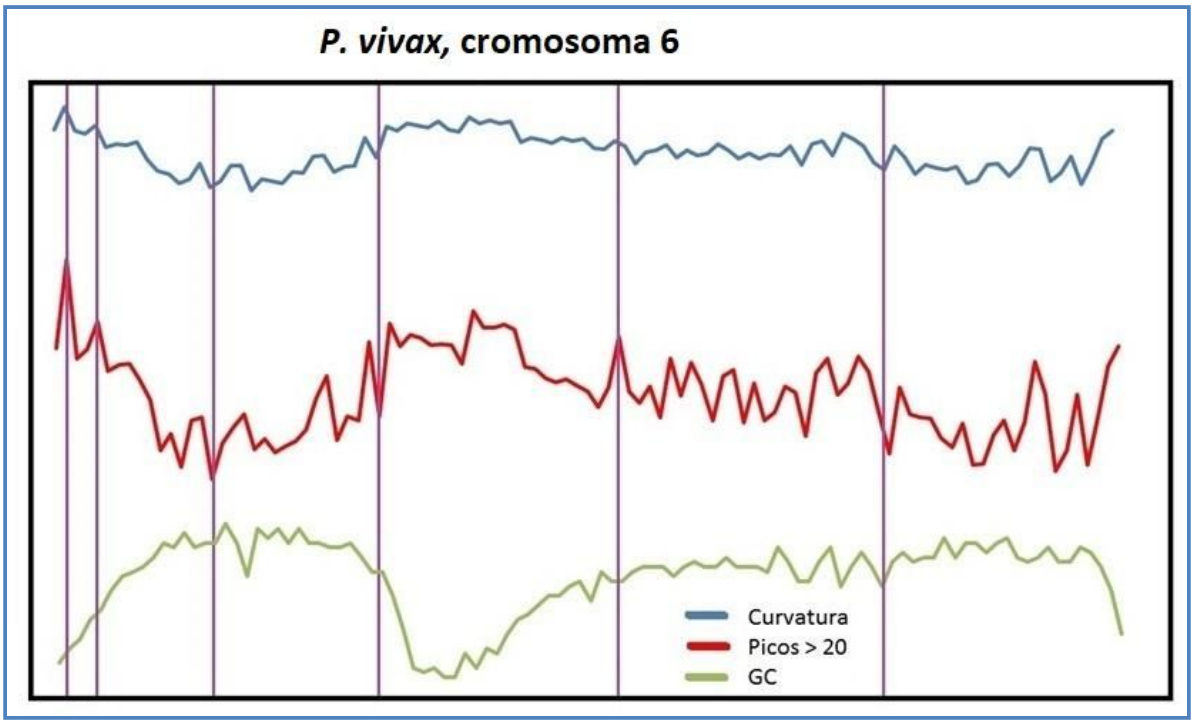


Figura 37: Superposición de gráfica de curvatura global, picos de curvatura local mayor a 20 y contenido G+C en cromosoma 6de *P. vivax.* Las líneas violeta que atraviesan la gráfica marcan puntos de correspondencia que facilitan la visualización de la correlación entre las 3 variables

Cromosoma	Coeficiente de correlación
P. vivax, cromosoma 1	-0.79
P. vivax, cromosoma 2	-0.75
P. vivax, cromosoma 3	-0.71
P. vivax, cromosoma 4	-0.72
P. vivax, cromosoma 5	-0.68
P. vivax, cromosoma 6	-0.85
P. vivax, cromosoma 7	-0.62
P. vivax, cromosoma 8	-0.67
P. vivax, cromosoma 9	-0.54
P. vivax, cromosoma 10	-0.71
P. vivax, cromosoma 11	-0.53
P. vivax, cromosoma 12	-0.60
P. vivax, cromosoma 13	-0.59
P. vivax, cromosoma 14	-0.63

Tabla 9: Coeficientes de correlación entre contenido G+C y curvatura global en cromosomas de *P.* vivax (valor p 2 e-16)

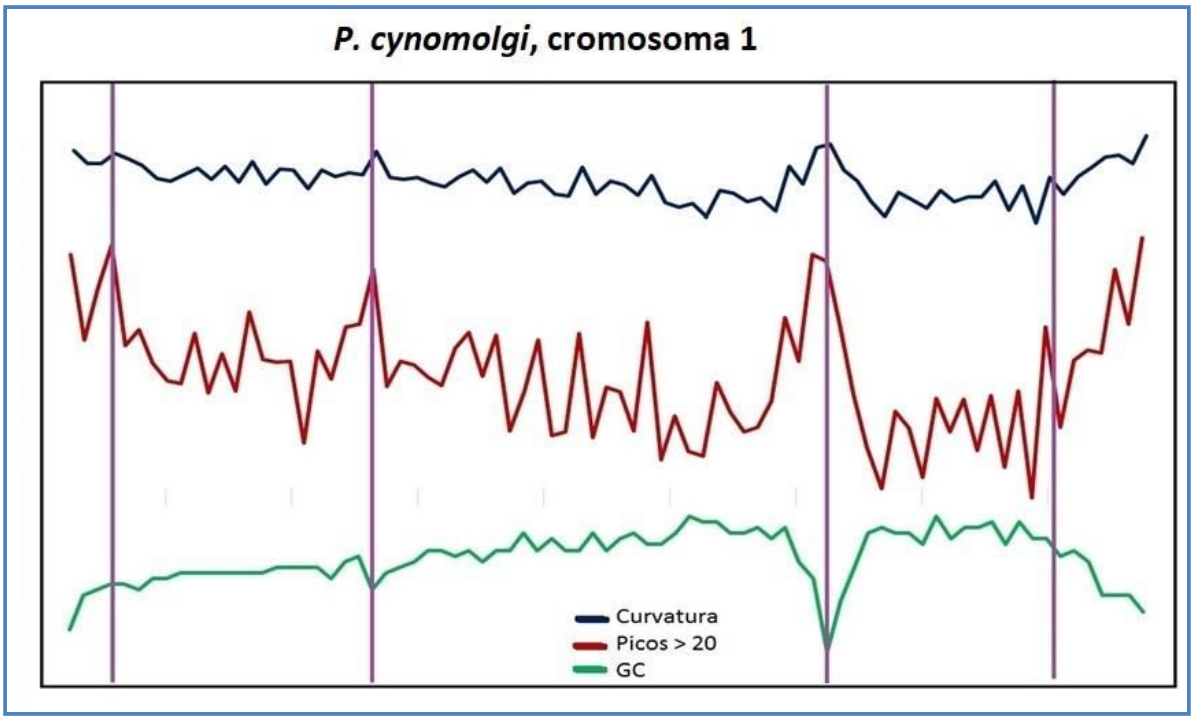


Figura 38: Superposición de gráfica de curvatura global, picos de curvatura local > 20 y contenido G+C en cromosoma 1 de *P. cynomolgi*, especie emparentada *con P. vivax.*

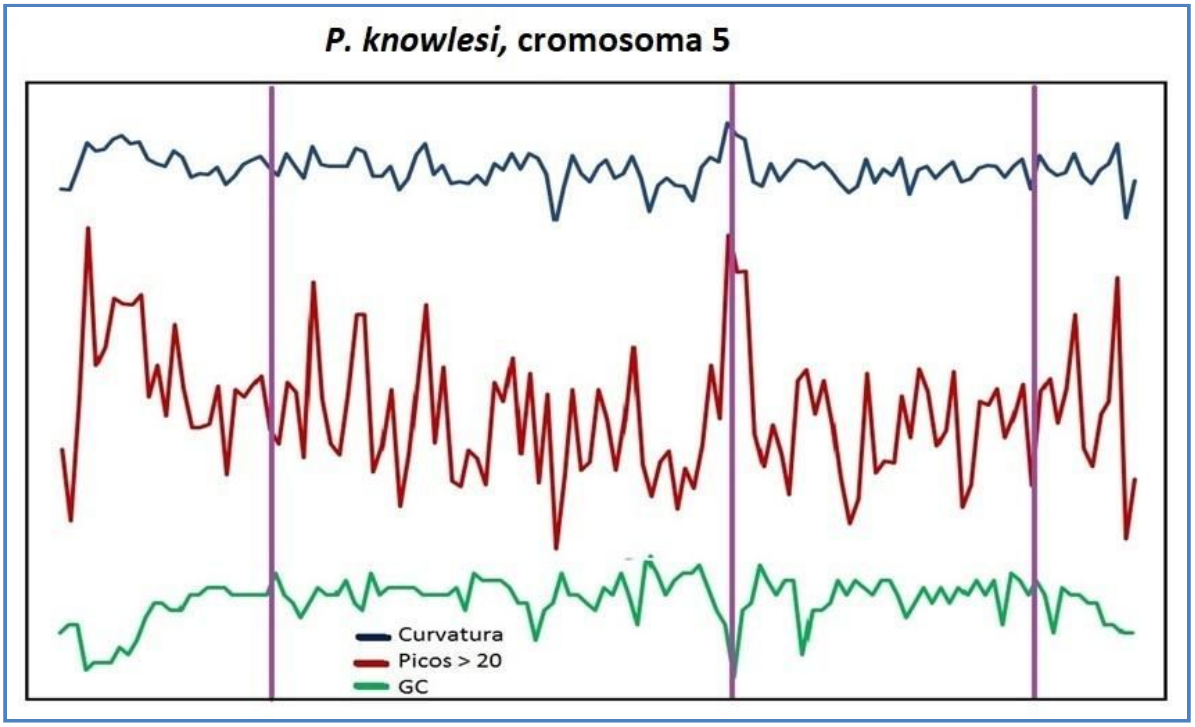


Figura 39: Superposición de gráfica de curvatura global, picos de curvatura local > 20 y contenido G+C en cromosoma 5 de *P. knowlesi*, especie emparentada *con P. vivax.*

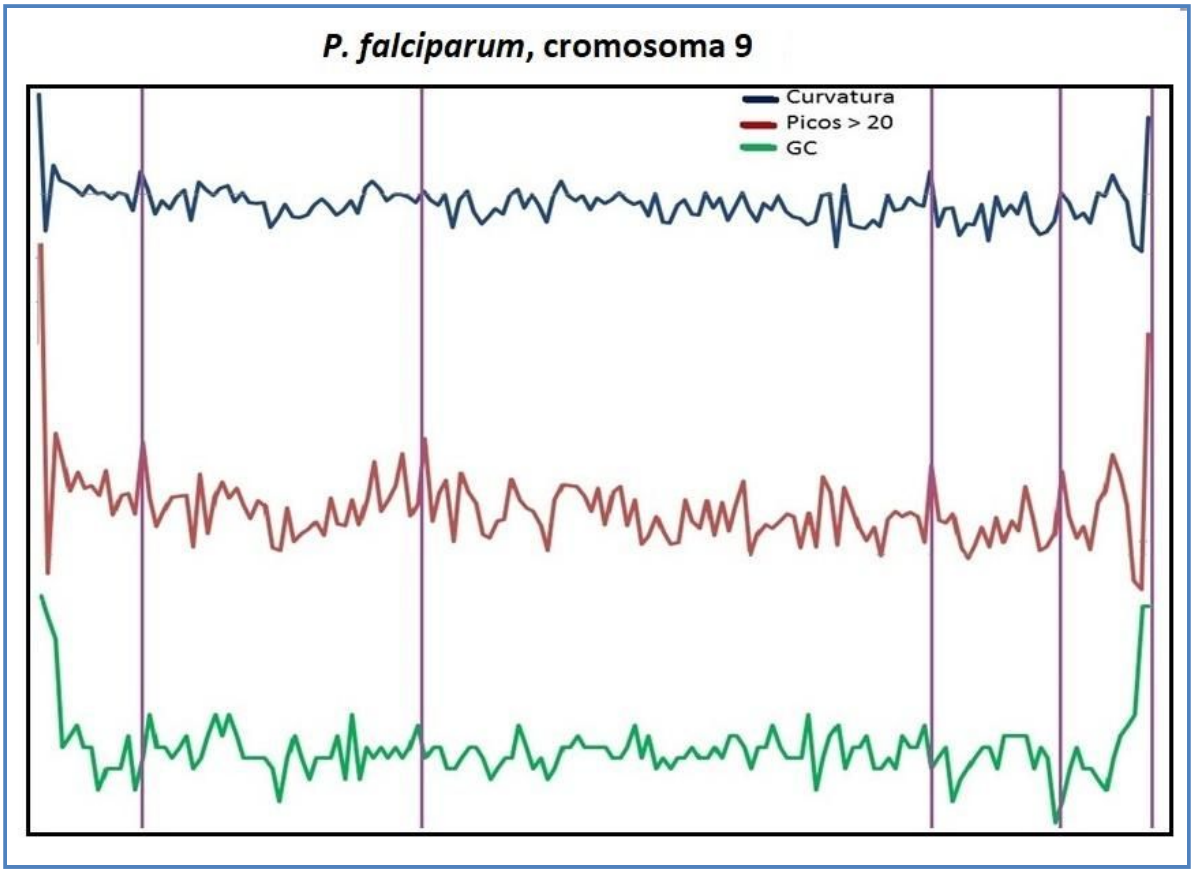


Figura 40: Superposición de gráfica de curvatura global, picos de curvatura local > 20 y contenido G+C en cromosoma 9 de *P. falciparum*, especie poco emparentada *con P. vivax*.

Especie/Cromosoma	Coeficiente de correlación	Valor p
P. cynomolgi, cromosoma 1	-0.77	4.67 e-17
P. knowlesi, cromosoma 5	-039	6.60 e-6
P. falciparum, cromosoma 9	0.30	0.90 e-4

Tabla 10: Coeficientes y de correlación entre contenido G+C y curvatura global en cromosomas de *Plasmodium* emparentados con *P. vivax*

CONCLUSIONES

Como mencionamos en la introducción, el contenido G+C de segmentos de 25Kpb del genoma de *P. vivax* comprende un amplio espectro de valores con una distribución bimodal con medias en 0.33 y 0.4, cuando lo esperable es que su distribución sea unimodal con una media cercana al contenido G+C genómico.

Esto nos llevó a profundizar el análisis a escala cromosómica revelando que la distribución espacial de los segmentos cromosómicos con contenido G+C diferencial al genómico muestra un comportamiento repetido en todos los cromosomas.

Los segmentos cuyo contenido G+C es más bajo que el contenido G+C genómico están localizados en las zonas teloméricas y los segmentos con un contenido G+C equiparable al genómico se ubican en las áreas internas del cromosoma.

Al discriminar las secuencias por categorías funcionales (proteínas vir, proteínas no vir, zonas intergénicas), vemos que tanto las proteínas vir como las proteínas no vir y regiones intergénicas teloméricas tienen un contenido G+C pobre, y las proteínas vir, no vir y regiones intergénicas localizadas en las zonas internas tienen un contenido G+C similar al genómico. O sea que independientemente de la categoría funcional y contenido génico considerado el contenido G+C acompaña la tendencia de la zona.

Esto es más notorio en zonas intergénicas, pero un factor a considerar es que las mismas no codifican y por lo tanto su secuencia es más libre que las zonas codificantes cuya secuencia forzosamente debe estar sujeta al código genético, por lo tanto pueden ajustar su estructura libremente a las restricciones de la zona.

Una característica interesante de las proteínas vir es que se localizan exclusivamente en las zonas teloméricas, con excepción de aquellas pertenecientes al cromosoma 6 que muestra indicios de haber tenido una inversión cromosómica o fusión entre cromosomas y en el proceso todas las proteínas o regiones intergénicas teloméricas se reubicaron en la zona interna. Otra característica interesante es que el mayor contribuyente a la región telomérica

son las regiones intergénicas, lo que nos lleva a concluir que el bajo valor de contenido G+C telomérico no es debido a la sobrerrepresentación de alguna de las categorías en particular, sino que por el contrario parece ser un comportamiento generalizado de la zona, indicando que fueron los distintos componentes funcionales las que adecuaron su composición generándose de esta forma zonas de menor contenido G+C relativamente homogéneas.

Los estudios estadísticos complementarios realizados (análisis de componentes principales de trinucleótidos, histograma de los valores del primer componte y la correlación entre primer componente de PCA y el contenido G+C de cada clase) refuerzan el comportamiento diferencial del contenido G+C dependiendo de la zona cromosómica a la que nos referenciemos.

En busca de las características funcionales que produjeran este comportamiento, pudimos detectar la existencia de una correlación inversa entre curvatura y contenido G+C en que cada máximo local en la curvatura se corresponde con mucha fidelidad con un mínimo local en el contenido G+C con coeficientes de correlación altamente significativos desde el punto de vista estadístico.

Esto confirma que el motivo de la asociación entre el bajo contenido G+C y la zona podría estar relacionado a la conformación estructural de la zona telomérica y al plegamiento del ADN en esa zona.

El estudio del contenido G+C de segmentos genómicos del mismo largo en otras especies relacionadas de *Plasmodium* muestra resultados disímiles. Mientras que *P. cynomolgi,* la especie filogenéticamente más cercana a *P. vivax* se comporta de una forma muy similar a *P. vivax* mostrando una distribución bimodal con 2 campanas aún más definidas, *P. berghei y P. chabaudi* presentan una menor variabilidad, con segmentos con un contenido G+C similar al genómico y una distribución unimodal agrupada en torno a la media del mismo**.**

A su vez *P. falciparum, l*a especie filogenéticamente más lejana dentro de las seleccionadas presenta una distribución de contenido G+C unimodal con una ligera anomalía en la cola derecha de la curva.

Esto nos señala que la existencia de regiones con contenido G+C diferente al contenido G+C genómico total no es un comportamiento exclusivo de *P. vivax,* sino que por el contrario se repite en varias de las especies del mismo género.

La investigación de proteínas homólogas a las proteínas teloméricas de *P. vivax* en otros *Plasmodium* muestra que estas también parecen localizarse en zonas teloméricas. Esto

indica que la distribución espacial de los segmentos de contenido G+C diferencial no es exclusiva de *P. vivax*.

Por último el estudio de la correlación entre la curvatura y el contenido G+C en otros *Plasmodium* muestra resultados muy similares. En todas las especies estudiadas, en *P. vivax* como en las especies emparentadas, existe una correlación en mayor o menor grado entre el contenido G+C y la curvatura del ADN.

REFERENCIAS

1. John Barnwell J, Carlton J, Collins W, Escalante A, Mullikin J, Saul A: **Neglected Burden of Human *vivax* Malaria: Comparative Analysis of *Plasmodium vivax* and Key Related Species** Am J Trop Med Hyg. 2001 Jan-Feb;64(1-2 Suppl):97-106

2. Hall N, Karras M, Raine JD, Carlton JM, Kooij TW, Berriman M, Florens L, Janssen CS, Pain A, Christophides GK, James K, Rutherford K, Harris B, Harris D, Churcher C, Quail MA, Ormond D, Doggett J, Trueman HE, Mendoza J, Bidwell SL, Rajandream MA, Carucci DJ, Yates JR 3rd, Kafatos FC, Janse CJ, Barrell B, Turner CM, Waters AP, Sinden RE.: **A Comprehensive Survey of the Plasmodium Life Cycle by Genomic Transcriptomic, and Proteomic Analyses**. Science 2005,307(5706): 82-86

3. Carlton J: **The Plasmodium vivax genome sequencing project**. TRENDS in Parasitology 2003, 19(5): 227-31.

4. Del Portillo H, Fernandez-Becerra C, Bowman S, Oliver K, Preuss M, Sanchez C, Schneider N, Villalobos J, Rajandream M, Harris D, Pereira da Silva L, Barrell B, Lanzer M. **A superfamily of variant genes encoded in the subtelomeric region of Plasmodium vivax.** Nature 2001, 410(6830): 839-42.

5. Christoph S. Janssen C, Phillips RS, Turner CM, Barrett MP**.: Plasmodium interspersed repeats: the major multigene superfamily of malaria parasites**. Nucleic Acids Res. 2004, 32(19): 5712-20.

6. Merino EF, Fernandez-Becerra C, Durham AM, Ferreira JE, Tumilasci VF, d'Arc-Neves J, da Silva-Nunes M, Ferreira MU, Wickramarachchi T, Udagama-Randeniya P, Handunnetti SM, Del Portillo HA. **Multi-character population study of the vir subtelomeric multigene superfamily of Plasmodium vivax, a major human malaria parasite**. Mol Biochem Parasitol. 2006, 149(1): 10-6

7. Fernandez-Becerra C1, Yamamoto MM, Vencio RZ, Lacerda M, Rosanas-Urgell A,del Portillo HA **Plasmodium vivax and the importance of the subtelomeric multigene vir superfamily**. Trends Parasitol. 2008, 25(1): 44-51.

8. Carlton JM,Adams JH, Silva JC, Bidwell SL, Lorenzi H, Caler E, Crabtree J, Angiuoli SV, Merino EF, Amedeo P, Cheng Q, Coulson RM, Crabb BS, Del Portillo HA, Essien K, Feldblyum TV, Fernandez-Becerra C, Gilson PR, Gueye AH, Guo X, Kang'a S, Kooij TW, Korsinczky M, Meyer EV, Nene V, Paulsen I, White O, Ralph SA, Ren Q, Sargeant TJ, Salzberg SL, Stoeckert CJ, Sullivan SA, Yamamoto MM, Hoffman SL, Wortman JR, Gardner MJ, Galinski MR, Barnwell JW, Fraser-Liggett CM**. Comparative genomics of the neglected human malaria parasite Plasmodium vivax**. Nature. 2008, 455(7214): 757-63

9. De Koning-Ward TF, Dixon MW, Tilley L, Gilson PR **Plasmodium species: master renovators of their host cells**. Nat Rev Microbiol. 2016, 14(8): 494-507.

10. Spencer L, Gómez A, Collovini E. **Mecanismos de invasion del esporozoíto y merozoíto de Plasmodium**. Revista Bionatura. 2016, vol 1, numero 2

11. Chernoff H, Lehmann EL. The **Use of Maximum Likelihood Estimates in χ 2 Tests for Goodness of Fit**. Ann. Math. Statist.1954, 25(3): 579-586.

12. Eslami-Mossallama B, Schiessel H, van Noort J**. Nucleosome dynamics: Sequence matters Advances in Colloid and Interface Science** 2016, 232: 101–113

13. Lu Y, Gan Y, Guana J, Zhou S. **An integrative analysis of nucleosome occupancy and positioning using diverse sequence dependent properties** . Neurocomputing 2016, 206 35–41

14. Smircich P, Forteza D, El-Sayed NJ, Garat B**. Genomic Analysis of Sequence-Dependent DNA Curvature in Leishmania**. Plos One 2013

15. Timchenko T.V., Vasyunina E.A., Scheglova T.V., Rogozin I.B., Sinitsyna O.I. **Computational prediction and experimental analysis of the curved dnas as the hot spots of recombination**. BGRS, 2002

16. Croll D, Lendenmann MH, Stewart E, mcdonald BA**. Impact of Recombination Hotspots on Genome Evolution of a Fungal Plant Pathogen. Genetics**. 2015; 201(3): 1213–1228.

17. Goodsell DS, Dickerson RE **Bending and curvature calculations in B-DNA. Nucleic Acids Research** 1994, 22(24) 5497-5503

18. Costantini M, Musto H. **The Isochores as a Fundamental Level of Genome Structure and Organization: A General Overview**. J Mol Evol. 2017, 84(2-3):93-103

19. Lamolle G, Protasio AV, Iriarte A, Jara E, Simon D, Musto H **An Isochore-Like Structure in the Genome of the Flatworm Schistosoma mansoni** Genome Biology and Evolution 2016, 8(8): 2312–2318

20. Vinogradov AE. **DNA helix: the importance of being GC-rich**. Nucleic Acids Research. 2003;31(7):1838-1844.

21. Mendis K, Sina BJ, Marchesini P, Carter R. **The neglected burden of Plasmodium vivax malaria** Am J Trop Med Hyg. 2001, 64(1-2):97-106

22. Snounou G, White NJ**. The co-existence of Plasmodium: sidelights from falciparum and vivax malaria in Thailand Trends** Parasitol 2004,. 20: 333–339.

23. Frech C, Chen N **Variant surface antigens of malaria parasites: functional and evolutionary insights from comparative gene family classification and analysis** BMC Genomics, 2013 14:427

24. Hiller NL, Bhattacharjee S, van Ooij C, Liolios K, Harrison T, Lopez-Estraño C, Haldar K **A host-targeting signal in virulence proteins reveals a secretome in malarial infection**. Science. 2004, 306(5703):1934-7

25. Merino EF, Fernandez-Becerra C, Durham AM, Ferreira JE, Tumilasci VF, d'Arc-Neves J, da Silva-Nunes M, Ferreira MU, Wickramarachchi T, Udagama-Randeniya P, Handunnetti SM, Del Portillo HA. **Multi-character population study of the vir subtelomeric multigene superfamily of Plasmodium vivax, a major human malaria parasite**. Mol.Biochem. Parasitol.2006, 149: 10–16.

26. Benham CJ,Mielke SP. **DNA mechanics** Annu Rev Biomed Eng. 2005, 7:21-53

27. Kornberg, R.D. 26 **Chromatin structure: A repeating unit of histones and DNA**. Science 1974, 184 (4139): 868-871

28. Lavery R, Moakher M, Maddocks JH, Petkeviciute D, Zakrzewska K **Conformational analysis of nucleic acids revisited: Curves+ . Nucleic Acids** Research 2009, 37(17):5917–5929

29. Chua EY, Vasudevan D, Davey GE, Wu B, Davey CA. **The mechanics behind DNA sequence-dependent properties of the nucleosome.** Nucleic Acids Research 2012, 40(13): 6338–6352

30. Chereji RV, Morozov AV**. Statistical Mechanics of Nucleosomes Constrained by Higher-Order Chromatin Structure** .Journal of Statistical Physics 2011, 144(2):379–404

31. Polach K.J. and Widom J. Mechanism of Protein Access to Specific DNA Sequences in Chromatin: A Dynamic Equilibrium Model for Gene Regulation. Journal of Molecular Biology 1995;254(2):130-49.

32. Răzvan V. Cherejia RV, Morozov AV. **Ubiquitous nucleosome crowding in the yeast genome.** Proc Natl Acad Sci. 2014;111(14):5236-41

33. Flaus A1, Luger K, Tan S, Richmond TJ. Mapping nucleosome position at single base-pair resolution by using site-directed hydroxyl radicals.Proc Natl Acad Sci 1996,93(4):1370-5.

34. Kochar DK, Saxena V, Singh N, Kochar SK, Kumar SV, Das A. **Plasmodium vivax malaria.** Emerging Infectious Diseases 2005 ; 11(1): 132–134.

35. Cammarano, R. ; Costantini, M, Bernardi G. **The isochore patterns of invertebrate genome.** BMC Genomics. 2009 :;10:538

36. Costantini M, Alvarez-Valin F, Costantini S, Cammarano R, Bernardi G. **Compositional patterns in the genomes of unicellular eukaryotes.** BMC Genomics. 2013 Nov 5;14:755

37. Costantini M, Cammarano R, Bernardi G. **The evolution of isochore patterns in vertebrate genomes** BMC Genomics. 2009;10:146.

38. Vinogradov AE. **Bendable genes of warm-blooded vertebrates.** Mol Biol Evol. 2001 18(12):2195-200

39. Timothy J. Richmond and Curt A. Davey. **The structure of DNA in the nucleosome core.** Nature 2003 423, 145-150.

40. Eugene Y. D. Chua, Dileep Vasudevan, Gabriela E. Davey Binwucurt A. Davey The mechanics behind DNA sequence-dependent properties of the nucleosome Nucleic Acids Research 2012, 40(13): 6338–6352.

41. Pennings S, Meersseman GE, Bradbury M **Mobility of positioned nucleosomes on 5 S rDNA** Journal of Molecular Biology 1991 220(5): 101-110

42. Moreira, D., López-García, P., & Vickerman, K. **An updated view of kinetoplastid phylogeny using environmental sequences and a closer outgroup: proposal for a new classification of the clasinetoplastea**. International journal of systematic and evolutionary microbiology 200**s K**4,54(5): 1861-1875.

43. Musto H, Naya H, Zavala A, Romero H, Alvarez-Valin F, Bernardi G. **Correlation between genomic G+C and optimal growth temperature in prokaryotes** FEBS Letters 573 (2004) 73–77

44. Musto H, Naya H, Zavala A, Romero H, Alvarez-Valin F, Bernardi G. **The correlation between genomic G+C and optimal growth temperature of prokaryotes is robust: a reply to Marashi and Ghalanbor.** Biochem Biophys Res Commun. 2005, 6;330(2):357-60

45. Ho, Pui & Carter, Megan**. DNA Structure: Alphabet Soup for the Cellular Soul**. (2011) 10.5772/18536.

46. Yun Lu, Yanglan Gan, Jihong Guan, Shuigeng Zhou. **An integrative analysis of nucleosome occupancy and positioning using diverse sequence dependent properties**. Elsevier Neurocomputing. 2016 35–41