Identificación de señales de selección natural en genes de *Plasmodium vivax* que codifican proteínas involucradas en el proceso de invasión para determinar su potencial uso en una vacuna antimalárica.

Diego Edison Garzón Ospina

Tesis Doctoral presentada como requisito para optar al título de Doctor en Ciencias Biomédicas y Biológicas de la Universidad del Rosario

Bogotá, 2018

Identificación de señales de selección natural en genes de *Plasmodium vivax* que codifican proteínas involucradas en el proceso de invasión para determinar su potencial uso en una vacuna antimalárica.

Estudiante

Diego Edison Garzón Ospina

Biólogo, Universidad INCCA de Colombia

Magister en Ciencias-Microbiología, Universidad Nacional de Colombia

Director

Manuel Alfonso Patarroyo Gutiérrez M.D., Dr.Sc.

Jefe del Departamento de Biología Molecular e Inmunología

Fundación Instituto de Inmunología de Colombia (FIDIC)

Profesor Titular, Escuela de Medicina y Ciencias de la Salud

Universidad del Rosario

DOCTORADO EN CIENCIAS BIOMÉDICAS Y BIOLÓGICAS

UNIVERSIDAD DEL ROSARIO

Bogotá, 2018

AGRADECIMIENTOS

Quiero dar mis agradecimientos y dedicar este trabajo a mi madre: **Martha Ospina Vargas**, y a mis hermanos: **Yamile Garzón Ospina** y **Julián David Escobar Ospina**, quienes me han acompañado y apoyado en todo momento. A **Sindy Paola Buitrago Puentes** por su acompañamiento y apoyo durante este tiempo.

También agradecer al Dr. **Manuel Alfonso Patarroyo Gutiérrez**, por haberme acogido en su equipo de trabajo y por apoyarme a lo largo de todos estos años.

Quiero agradecer y reconocer la labor de: **Andrea Estefanía Ramos, Darwin Andrés Moreno Pérez, Elizabeth Gutiérrez Vásquez, Heidy Daniela Ortiz Suarez, Lady Johanna Forero Rodríguez, Laura Alejandra Ricaurte Contreras, Leidy Paola Reyes, Luis Alfredo Baquero, Paola Andrea Camargo Ayala, Sindy Paola Buitrago Puentes, Ricardo De León Montero** y **Yimara Grosso Paz**, quienes aportaron su tiempo, dedicación y esfuerzo, permitiendo la culminación de este trabajo.

Y finalmente, mis agradecimientos al Doctorado, a la Universidad del Rosario y a los Jurados de tesis, quienes amablemente aceptaron la revisión de este trabajo, por sus enriquecedores comentarios y por su tiempo.

CONTENIDO

LISTA DE TABLAS

LISTA DE FIGURAS

LISTA DE ANEXOS

LISTA DE PUBLICACIONES

Publicación 1: **Garzón-Ospina D**, Forero-Rodríguez J, Patarroyo MA. *Inferring natural selection signals in Plasmodium vivax-encoded proteins having a potential role in merozoite invasion*. Infect Genet Evol. 2015 Jul; 33:182-8. doi: 10.1016/j.meegid.2015.05.001.

Publicación 2: **Garzón-Ospina D**, Forero-Rodríguez J, Patarroyo MA. *Evidence of functional divergence in MSP7 paralogous proteins: a molecular-evolutionary and phylogenetic analysis*. BMC Evol Biol. 2016 Nov 28;16 (1):256. doi: 10.1186/s12862-016-0830-x.

Publicación 3: Forero-Rodríguez J, **Garzón-Ospina D**, Patarroyo MA. *Low genetic diversity and functional constraint in loci encoding Plasmodium vivax P12 and P38 proteins in the Colombian population*. Malar J. 2014 Feb 18;13:58. doi: 10.1186/1475-2875-13-58.

Publicación 4: Forero-Rodríguez J, **Garzón-Ospina D**, Patarroyo MA. *Low genetic diversity in the locus encoding the Plasmodium vivax P41 protein in Colombia's parasite population*. Malar J. 2014 Sep 30;13:388. doi: 10.1186/1475-2875-13-388.

Publicación 5: **Garzón-Ospina D**, Forero-Rodríguez J, Patarroyo MA. *Heterogeneous genetic diversity pattern in Plasmodium vivax genes encoding merozoite surface proteins (MSP) -7E, -7F and -7L*. Malar J. 2014 Dec 13;13:495. doi: 10.1186/1475-2875-13-495.

Publicación 6: Buitrago SP, **Garzón-Ospina D**, Patarroyo MA. *Size polymorphism and low sequence diversity in the locus encoding the Plasmodium vivax rhoptry neck protein 4 (PvRON4) in Colombian isolates*. Malar J. 2016 Oct 18;15(1):501. doi: 10.1186/s12936-016-1563-4.

Publicación 7: Baquero LA, Moreno-Pérez DA, **Garzón-Ospina D**, Forero-Rodríguez J, Ortiz-Suárez HD, Patarroyo MA. *PvGAMA reticulocyte binding activity: predicting conserved functional regions by natural selection analysis*. Parasit Vectors. 2017 May 19;10(1):251. doi: 10.1186/s13071-017-2183-8.

Publicación 8: **Garzón-Ospina D**, Buitrago SP, Ramos AE, Patarroyo MA. *Identifying Potential Plasmodium vivax Sporozoite Stage Vaccine Candidates: An Analysis of Genetic Diversity and Natural Selection*. Front Genet. 2018 Jan 25;9:10. doi:10.3389/fgene.2018.00010.

Publicación 9: Camargo-Ayala PA, **Garzón-Ospina D**, Moreno-Pérez DA, Ricaurte-Contreras LA, Noya O and Patarroyo MA. *On the evolution and function of Plasmodium vivax reticulocyte binding surface antigen (pvrbsa)*. Front. Genet. 2018 Sep . 9:372. doi: 10.3389/fgene.2018.00372

ABREVIATURAS

π: diversidad nucleotídica por sitio

K: diversidad nucleotídica

d_N: tasas de sustitución no-sinónima por sitio no-sinónimo

d_S: la tasa de sustitución sinónima por sitio sinónimo

K_N: divergencia no-sinónima por sitio no-sinónimo

K_S: divergencia sinónima por sitio sinónimo

ω: tasa evolutiva (omega) = d_N/d_S o K_N/K_S

n: Numero se secuencias analizadas

Sitios: total de sitios analizados, excluyendo gaps

Ss: número de sitios segregantes (polimórficos)

S: número de sitios singleton

Ps: número de sitios parsimoniosos

SD: desviación estándar

NI: índice de neutralidad de la prueba McDonald-Kreitman (NI)

ORF: marco abierto de lectura

MK: prueba de McDonald-Kreitman

Pn: polimorfismo no-sinónimo

Ps: polimorfismo sinónimo

Dn: divergencia no-sinónima

Ds: divergencia sinónima

ML: máxima verosimilitud

BY: bayesiano

malERA: *Malaria Eradication Research Agenda*

Identificación de señales de selección natural en genes de *Plasmodium vivax* que codifican proteínas involucradas en el proceso de invasión para determinar su potencial uso en una vacuna antimalárica.

RESUMEN

Plasmodium vivax, es un endoparásito de origen Asiático que se dispersó alrededor del mundo y, actualmente, es predominante en las regiones tropicales y subtropicales del planeta que se encuentran entre los 15 y 16 °C. *P. vivax,* presenta una importancia biomédica dado que es el segundo agente responsable de malaria en humanos. Aunque numerosos esfuerzos se han realizado para disminuir el impacto de esta enfermedad, las condiciones sociales y económicas de los lugares más afectados, sumado a los conflictos sociales y políticos en varias áreas endémicas, hacen del control y eliminación de este parásito una tarea nada fácil. Como si esto no fuera suficiente, la aparición de resistencia a los insecticidas por parte del vector trasmisor, así como de parásitos resistentes a los antimaláricos, podría provocar una recurrencia de esta enfermedad. Teniendo en cuenta lo anterior, nuevas estrategias que permitan disminuir la incidencia de malaria por *P. vivax* se hacen prioritarias.

Una de las alternativas que podría ayudar al control de la malaria es el desarrollo de una vacuna contra los patógenos que la causan. Sin embargo, la elevada diversidad genética que *P. vivax* presenta, ha generado uno de los retos a afrontar para el diseño de una vacuna completamente efectiva. Actualmente, se han descrito varios antígenos potenciales candidatos a vacuna contra *P. vivax.* No obstante, la diversidad genética de un reducido número de estos antígenos ha sido evaluada, debido a los requerimientos de tiempo y recursos económicos. Adicionalmente, poco se sabe acerca del rol real de estos antígenos durante el proceso de invasión de este parasito a las células hospederas. Es por esto, que nuevos enfoques, que permitan en un menor tiempo y a bajo costo identificar las regiones de estos antígenos conservadas por selección natural (usualmente asociadas con regiones funcionales) se hacen necesarios. Estos nuevos enfoques podrían entonces servir como

punto de partida para la identificación o priorización de nuevos y promisorios candidatos vacunales.

Este trabajo presenta los resultados un enfoque alternativo que permite hacer un acercamiento a la diversidad genética de antígenos de *P. vivax*, determinando regiones bajo selección negativa, para ser consideradas durante el diseño de una vacuna completamente efectiva. Aunque este enfoque se utilizó para *P. vivax*, este podría también ser aplicado en otros organismos.

Identificación de señales de selección natural en genes de *Plasmodium vivax* que codifican proteínas involucradas en el proceso de invasión para determinar su potencial uso en una vacuna antimalárica.

ABSTRACT

Plasmodium vivax, an endoparasite that arose in Asia and spread around the world, has biomedical importance given that it is the second most important human-malaria parasite. Although efforts have been made to reduce the impact of this disease, the social and economic conditions of the most affected places, together with social and political conflicts in several endemic areas, make the parasite control and elimination a laborious task. The emergence of insecticide resistance by the transmitting vector as well as antimalarial-resistant parasites worsen the problem. Therefore, new alternatives to allow reducing the incidence of the disease have become a priority.

An antimalarial vaccine development against the causal pathogens has been proposed as a cost-effective intervention which would help in controlling malaria. However, the high *P. vivax* genetic diversity remains as one of the challenges to overcome for the design of a fully effective vaccine. Currently, several potential *P. vivax* vaccine candidates have been described. Nevertheless, the genetic diversity of a small number of them has been assessed, due to the high amount of time and economic resources required. Additionally, there is a modest knowledge about the real role of these antigens during the invasion process of target cells since maintaining an *in vitro* culture of this parasite species is particularly difficult. Therefore, new approaches that allow identifying conserved regions by natural selection which are frequently associated with functional importance, might be used as a starting point for the identification or prioritization of new potential vaccine candidates.

This work presents a new approach to assess the genetic diversity of potential candidate antigens, determining negatively selected regions that can then be considered for designing a fully effective vaccine. This approach is not limited to *P. vivax* and could be useful in other microorganisms.

INTRODUCCIÓN GENERAL

Los parásitos representan una gran proporción de los organismos vivos del planeta y son de importancia biomédica debido a su impacto en la salud humana (1). De todas las enfermedades parasitarias que afectan al ser humano, la malaria es la que produce más muertes alrededor del mundo (2, 3) (Tabla 1).

Tabla 1. Total de muertes causadas por agentes parasitarios en 2013.

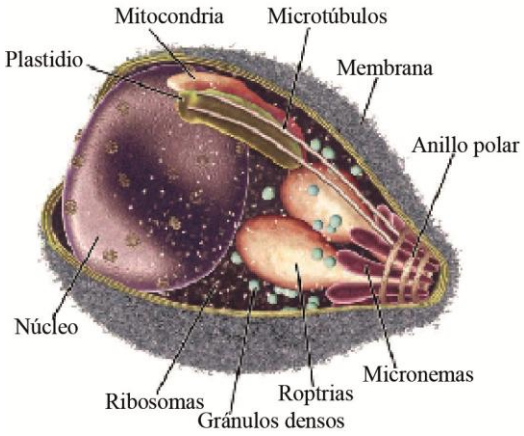
Parasitosis	Muertes globales en 2013
Malaria	854.600
Leishmaniasis	62.500
Criptosporidiosis	41.900
Amebiasis	11.300
Enfermedad de Chagas	10.600
Tripanosomiasis Africana	6.900
Esquistosomiasis	5.500
Ascariasis	4.500
Equinococosis	2.200
Cisticercosis	700
Total de muertes por parasitosis	**1.000.700**

Modificada de: Hotez, P., Herricks, J., 2015. PLOS Medicine Pathogens Neglected Tropical Diseases (2).

Esta enfermedad es causada por parásitos intracelulares obligados que pertenecen al *Phylum Apicomplexa,* orden *Haemosporidia* y género *Plasmodium*. Miembros de este *Phylum* se caracterizan por la presencia de organelos apicales (roptrias, micronemas y gránulos densos, Figura 1) importantes durante el proceso de invasión a las células hospederas.

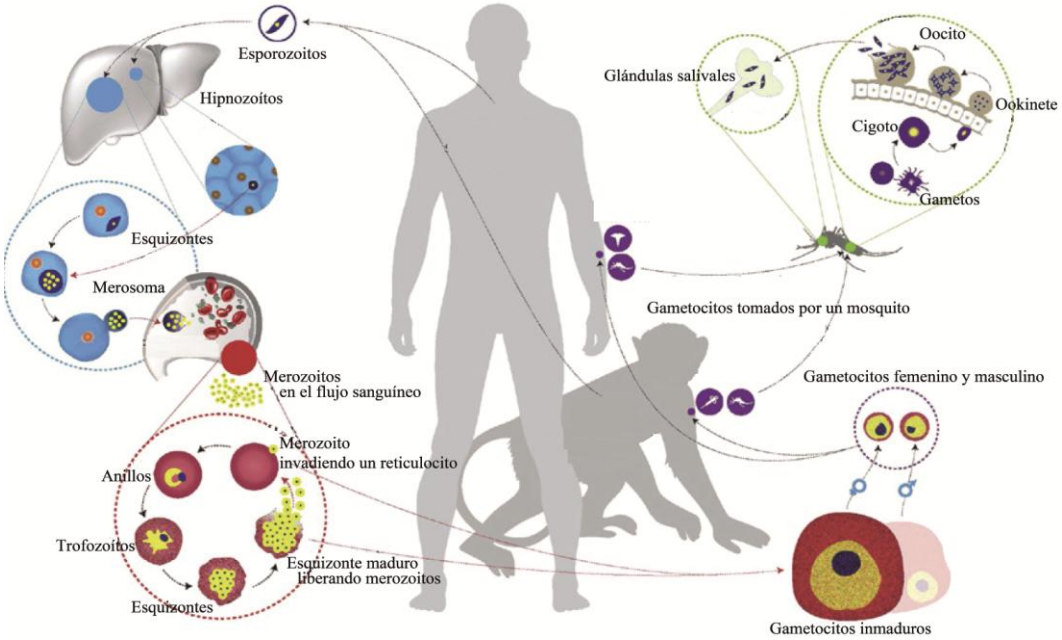
Plasmodium spp tiene un ciclo de vida complejo (Figura 2), desarrollándose en dos hospederos diferentes; uno intermediario (vertebrados terrestres como reptiles, aves y mamíferos (4-6)), donde se desarrolla la fase asexual y un hospedero definitivo y trasmisor del parásito (mosquitos del genero *Anopheles*), donde se lleva a cabo la fase sexual.

Figura 1. El merozoíto y su estructura interna.



Modificada de: Cowman A F y Crabb B S., 2006. Cell (7).

Figura 2. Ciclo de vida de *Plasmodium* spp.

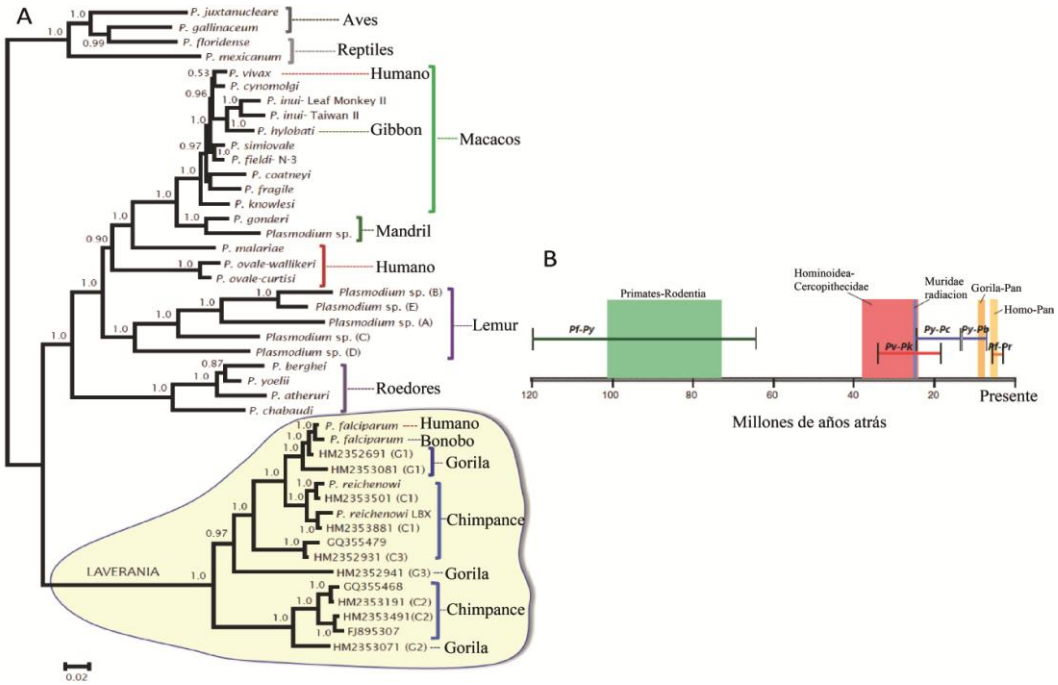


Modificada de: Galinski M, Meyer E, Barnwell J., 2013. Adv Parasitol (8).

A la fecha, solo cinco especies de *Plasmodium* son capaces de infectar al ser humano: *P. falciparum*, *P. vivax*, *P. ovale*, *P. malariae* (4, 5, 9) y *P. knowlesi* (10). *P. falciparum*, filogenéticamente relacionado con *P. reichenowi* (11, 12), divergió aproximadamente al mismo tiempo en que lo hicieron los humanos de los grandes simios Africanos (13) (Figura 3B). Por su parte, *P. vivax*, de origen Asiático (Figura 3A, aunque un origen Africano es

probable (14)), divergió de parásitos que infectan primates no humanos, hace unos 2 – 2,5 millones de años (11), mediante un evento conocido como cambio de hospedero (15). Por lo tanto, *P. falciparum* y *P. vivax*, a pesar de infectar seres humanos, presentan diferentes historias evolutivas.

Figura 3. Filogenia (A) y tiempos de divergencia (B) de algunas especies del género *Plasmodium*.



Modificado de: Pacheco, 2011. BMC Evol Biol (11) y Silva, 2015. Mol Biol Evol (13).

La incidencia de *P. vivax* ha disminuido desde 2010. Cerca de 6 a 11 millones de casos se atribuyeron a este parásito en 2016, alrededor de 7 millones menos que en 2010 (16). A pesar de esto, la carga social y económica causada por la malaria en los países endémicos (17) sigue siendo grande. Aunque las medidas de control contra este parásito parecen ser útiles, la eliminación de *P. vivax* (así como la eliminación de la malaria) no es una tarea sencilla. Las condiciones sociales y económicas de los lugares más afectados por esta enfermedad, así como los conflictos sociales y políticos en varias áreas endémicas, representan los principales desafíos para eliminar la malaria (16). Esto, junto con la dispersión de mosquitos resistentes a los insecticidas y de parásitos resistentes a los medicamentos, podría empeorar el panorama, generando una recurrencia de la enfermedad

(18-20). Por lo tanto, nuevas intervenciones como el diseño de una vacuna, se hacen necesarios (21). Así, las intervenciones existentes junto con las nuevas estrategias podrían contribuir a un mundo libre de malaria (21, 22).

Dado que *P. falciparum* es el agente más letal de la enfermedad, el diseño de una vacuna se centró durante mucho tiempo en este parásito. Sin embargo, dado que en los últimos años la infección por *P. vivax* está presentando complicaciones clínicas y fallas terapéuticas, la malERA (del inglés, *Malaria Eradication Research Agenda*), en 2013, reconoció la necesidad de desarrollar una vacuna contra este parásito (23). A pesar de esto, el desarrollo de una vacuna contra *P. vivax* se encuentra retrasado con respecto a *P. falciparum* (24), esto debido a que los aspectos de la biología de *P. vivax* son generalmente difíciles de estudiar de forma directa, dada la predilección de este parásito por los reticulocitos (9). Sin embargo, datos genómicos, trascriptómicos, proteómicos, de marcadores moleculares o estudios de genética de poblaciones, podrían ser cruciales para la predicción y evaluación de drogas y vacunas (1, 25).

En los últimos años, utilizando este tipo de datos (ómicos), se han caracterizado nuevos antígenos que podrían ser considerados como candidatos a vacuna (24, 26-28). No obstante, esta caracterización es tan sólo el primer paso. Debido a la alta diversidad (variabilidad) genética que exhibe *P. vivax* (25, 29), una vacuna diseñada a partir de estos antígenos podría ser poco eficaz, dadas las respuestas inmunes alelo-específicas que el polimorfismo dentro de estos antígenos podría generar (30-34). Así, con el fin de diseñar una vacuna completamente efectiva, se hace necesario centrar su desarrollo en regiones conservadas o dominios funcionales (35). No obstante, los dominios funcionales (por ejemplo, aquellos involucrados en la interacción patógeno-hospedero) de antígenos de *P. vivax*, no han podido ser dilucidados sino en un muy reducido grupo de ellos (36-38).

Las regiones funcionales de las proteínas, por ser críticas para llevar a cabo diferentes procesos celulares, evolucionan lentamente, observándose una correlación entre la conservación de un fragmento de la molécula con su importancia funcional (39, 40). Por lo tanto, determinar estas regiones podría ser un punto de partida para diseñar una vacuna

completamente efectiva, ya que se evitarían así, las respuestas inmunes alelo-específicas. Recientemente, se ha producido un incremento de los datos genómicos (29, 41), trascriptómicos (42) y proteómicos (43, 44) de *P. vivax* y especies filogenéticamente relacionadas (10, 45). Éstos, sumados a los estudios de variación genética en las poblaciones naturales, pueden dar acceso a información clave sobre la biología, ecología y evolución del parásito, usando un conjunto cada vez mayor de herramientas de genética de poblaciones y evolución molecular. Así, la generación o uso de este tipo de datos y herramientas, podría ser el punto de partida para la priorización o selección de los mejores antígenos para diseñar una vacuna completamente eficaz contra *P. vivax*.

***Candidatos a vacuna*.**

El esporozoíto y el merozoíto son las formas extracelulares infectivas de *Plasmodium* spp, por lo tanto, son blancos del sistema inmune. Numerosas proteínas, tanto de superficie como de los organelos apicales del parásito, participan activamente durante la invasión a la célula hospedera. Por lo tanto, estas moléculas son los principales blancos a incluir en una vacuna, ya que respuestas inmunes dirigidas hacia ellas podrían inhibir la invasión (46-48). A pesar de la existencia de múltiples antígenos, no todos son considerados candidatos promisorios. Para su inclusión en una vacuna, los candidatos deben cumplir con ciertas características (35, 49):

1) estar involucrados en el proceso de invasión a la célula hospedera
2) deben ser accesibles al sistema inmune del hospedero
3) inducir respuestas inmunes protectoras, y
4) carecer de variabilidad genética o por lo menos tener una variabilidad baja (5).

A pesar de que ya se cuenta con un número significativo de antígenos en *P. vivax* descritos (24, 50-58), la diversidad genética de solo algunos de estos antígenos ha sido evaluada. Esto debido a que los estudios usados para tal fin implican, no solo tiempo, sino también la disponibilidad de recursos económicos, convirtiéndose en una limitante para el desarrollo de una vacuna. Aún más, dada la predilección de *P. vivax* por los reticulocitos, el papel que

estos antígenos tienen durante la invasión no ha podido ser del todo descrito, dadas las dificultades técnicas para evaluar la interacción patógeno (*P. vivax*) - hospedero (reticulocitos).

Genética y evolución, herramientas para el diseño de vacunas.

La diversidad genética en los antígenos parasitarios puede ser, en parte, consecuencia de presiones selectivas impuestas por el sistema inmune del hospedero, lo que permite la fijación o acumulación de variantes (mutaciones) en la población parasitaria. Esta presión selectiva, puede entonces, incrementar la diversidad del parásito, lo cual genera el problema de las respuestas inmunes alelo-específicas (49, 59) que disminuyen la eficacia de las vacunas (34, 60).

Estudios poblaciones han permitido, no solo evaluar la diversidad de los antígenos, sino también identificar las fuerzas evolutivas responsables de generar el patrón de variación observado (61-71). Usualmente, los antígenos sobre la superficie del parásito están expuestos al sistema inmune del hospedero, lo que genera una presión selectiva sobre el antígeno. Teniendo en cuenta que las mutaciones en este grupo de parásitos se producen a un ritmo mayor que la tasa de mutación de línea base (49), la aparición de una variante, que permita la evasión de la respuesta inmune, se puede presentar a una tasa elevada. Esa mutación (o algunas ya presentes) se fijará en la población, dada la ventaja adaptativa que la variante tendría, y en este caso, la selección actuará hacia una dirección, manteniendo el mutante ventajoso (selección positiva). Sin embargo, muchas mutaciones no alcanzan a fijarse en la población. Éstas se mantienen circulando en la población, manteniendo frecuencias intermedias o alternando sus frecuencias con el tiempo (selección balanceante y selección dependiente de frecuencia). En este caso, la selección actúa manteniendo las diferentes variantes, ya que un alto número de alelos circulando, se convierte en un mecanismo de evasión de la respuesta inmune. Por lo tanto, se ha sugerido que antígenos o regiones con estos tipos de selección, podrían ser dianas de la respuesta inmune y por consiguiente ser evaluados como candidatos a vacuna (72, 73). No obstante, estos candidatos no necesariamente son las dianas ideales, debido a su alta diversidad genética.

Por otro lado, algunas variantes (mutaciones que alteren la secuencia de la proteína) podrían tener un efecto contrario a las variantes fijadas/mantenidas por selección positiva o balanceante. Por ejemplo, si una variante es reconocida fácilmente por el sistema inmune, ésta será eliminada de la población parasitaria. Así mismo, muchas de las mutaciones que se generan suelen ser eliminadas "rápidamente" debido a que alteran la función o estructura de la proteína. Esta eliminación se debe a la selección negativa (a veces llamada purificante), que elimina todas aquellas variantes que tengan un efecto negativo en el organismo (74). Dado que la selección negativa mantendrá conservadas porciones de la proteína donde esta está actuando, la detección de este tipo de selección podría entonces ser utilizado para la identificación de potenciales candidatos a vacunas, ya que estos, al ser conservados, evitarán respuestas inmunes alelo-específicas. (68, 74-76). Aún más, la inferencia de este tipo de selección podría ayudar a delimitar regiones funcionales dentro de los antígenos, debido a que los dominios funcionales suelen estar restringidos funcional o estructuralmente (39, 40), evolucionando más lentamente. En consecuencia, la región o antígeno bajo selección negativa se mantendrá conservado (a nivel de la secuencia de proteína) y así, si las respuestas inmunes se dirigen hacia estas regiones (y son protectoras), se obtendría una respuesta completamente efectiva.

Ya que los antígenos involucrados en el proceso de invasión podrían ser reconocidos por el sistema inmune de hospedero, estos podrían presentar una alta diversidad genética. Sin embargo, las partes funcionalmente importantes de la proteína (por ejemplo, las regiones implicadas en la interacción parásito-hospedero) deben estar restringidas funcionalmente, manteniéndose conservadas por selección negativa, dentro y entre las especies (40). En consecuencia, inferir este tipo de selección en antígenos de *P. vivax* (u otro patógeno) podría usarse para predecir regiones bajo restricción funcional, que luego podrían usarse como potenciales candidatas a vacunas.

El presente trabajo está dividido en tres partes, la primera (Capitulo 1 y 2) donde se analizó los datos genómicos disponibles en bases de datos con el fin de generar un flujo de trabajo que permitiese hacer una selección preliminar de los antígenos parasitarios más promisorios para el diseño de una vacuna contra *P. vivax* que evite las respuestas inmunes alelo-

específicas. La segunda parte (Capitulo 3) analizó por análisis poblaciones, la diversidad genética de 8 antígenos parasitarios, contrastando los resultados obtenidos en los Capítulos 1 y 3. La última parte (Capitulo 4), demuestra el potencial uso de esta metodología para la selección de antígenos (o regiones dentro de éstos) promisorios para el diseño de una vacuna contra *P. vivax*.

HIPOTESIS DE INVESTIGACIÓN

La diversidad genética observada en *P. vivax* es mayor a la de *P. falciparum* (25, 29), posiblemente debido a que *P. vivax* no ha experimentado cuellos de botella durante su historia evolutiva como *P. falciparum* (77-80). Así, gran parte del polimorfismo observado en esta especie podría ser ancestral y muchas de las variantes alélicas podrían haber surgido antes de la dispersión del parasito por el mundo (81). Por lo tanto, el análisis de un número reducido de secuencias de cepas parasitarias de diferentes regiones del mundo podría reflejar la diversidad genética un determinado antígeno. Si el análisis de pocas secuencias provenientes de diferentes regiones del mundo refleja la diversidad genética de un determinado antígeno, entonces estos resultados serán concordantes a los obtenidos analizando un mayor número de secuencias, y así, podrían seleccionase antígenos con baja diversidad en un menor tiempo, postulándolos como potenciales antígenos a ser incluidos en una vacuna contra *P. vivax*.

Por otra parte, dado que los antígenos parasitarios involucrados en el proceso de invasión podrían ser blancos del sistema inmune de hospedero, éstos pueden presentan una alta diversidad genética (61, 64, 65, 67, 70, 71, 90, 128, 129). Sin embargo, las partes funcionalmente importantes de la proteína (por ejemplo, aquellas regiones implicadas en la interacción parásito-hospedero) podrían estar restringidas funcionalmente, manteniéndose conservadas dentro y entre especies por la acción de la selección negativa. Por lo tanto, si las regiones conservadas entre especies de un determinado antígeno son el resultado de la acción de la selección negativa, entonces estas regiones serían funcionalmente importantes y podrían estar involucradas, por ejemplo, en la interacción patógeno-hospedero y así ser consideradas como promisorias para ser incluidas durante el diseño de una vacuna contra *P. vivax*.

OBJETIVOS

Objetivo general

Inferir señales de selección natural en antígenos de *Plasmodium vivax* para identificar y seleccionar los candidatos/regiones a vacuna más promisorios para el diseño de una vacuna completamente efectiva contra *P. vivax*.

Objetivos específicos

- Analizar 59 genes de *P. vivax* previamente reportados como potenciales candidatos a vacuna a partir de los datos genómicos de 5 aislados parasitarios y 2 especies filogenéticamente relacionas depositados en el GenBank.

- Calcular varios parámetros de diversidad genética y detectar señales de selección natural para cada uno de los 59 genes a partir de los datos genómicos obtenidos.

- Seleccionar los antígenos más promisorios a incluir en el diseño de vacuna antimalárica contra *P. vivax* a partir de los datos obtenidos.

- Identificar la presencia de infecciones únicas y la presencia de diferentes genotipos de *P. vivax* en muestras de ADN de aislados naturales obtenidas en diferentes regiones de Colombia.

- Evaluar la diversidad genética de 8 antígenos de *P. vivax* y determinar las fuerzas evolutivas que generan el patrón de variación observado, a partir de secuencias obtenidas de aislados parasitarios de la población Colombiana.

- Implementar la metodología propuesta en este trabajo durante la caracterización de nuevos antígenos de *P. vivax.*

CAPITULO 1

Un enfoque alternativo para la selección de candidatos promisorios a vacuna

Introducción

El desarrollo de una vacuna contra *P. vivax* se encuentra retrasado con respecto a *P. falciparum*, y pocos candidatos a vacuna han sido propuestos hasta la fecha (21). En *P. vivax*, la caracterización de nuevos potenciales candidatos involucra la búsqueda de secuencias con un alto porcentaje de similitud con los antígenos previamente descritos en *P. falciparum* (24). Adicionalmente, otras alternativas (82) han permitido por métodos bioinformáticos predecir proteínas de *P. vivax* que podrían tener un papel durante el proceso de invasión.

Estos estudios han permitido describir cerca de 60 antígenos de *P. vivax* que podrían ser considerados durante el diseño de una vacuna. Sin embargo, la diversidad genética y las fuerzas evolutivas de estas moléculas deben ser evaluadas para diseñar una vacuna completamente efectiva (21, 83). Análisis poblacionales y de evolución molecular son las principales estrategias para tal fin. Estas alternativas requieren la secuenciación de varias muestras parasitarias, por lo tanto, realizar tales estudios para todos estos genes implicaría tiempo y recursos económicos. No obstante, Cornejo *et al.* (84), utilizando un tamaño de muestra limitado (los genomas de 5 aislamientos) han identificado genes con señales de selección natural. Este tipo de análisis podría ser un punto de partida para detectar posibles nuevos candidatos a vacuna (73), similar a previos enfoques adoptados en *P. falciparum* (72, 85).

La primera parte de este trabajo utilizó tres pruebas diferentes para detectar señales de selección en 59 antígenos del merozoíto de *P. vivax* previamente caracterizados. Para este fin, la información disponible en *GenBank* de cinco genomas de aislados de *P. vivax* de diferentes regiones del mundo y de dos especies estrechamente relacionadas (*P. cynomolgi* y *P. knowlesi*) fue analizada. Los resultados obtenidos pueden ser utilizados para determinar qué antígenos deberían ser priorizados y evaluados en estudios adicionales dirigidos al diseño de una vacuna completamente efectiva contra *P. vivax*.

Metodología

Obtención de secuencias y alineamientos.

Secuencias de 59 genes (de la cepa adaptada a primates no humanos Sal-I) previamente sugeridos como potenciales candidatos a vacuna (24, 26-28, 50-58, 82, 86-89), fueron descargadas de la base de datos PlasmoDB (www.plasmodb.org/plasmo/). Cuarenta y ocho de estos genes no habían sido analizados previamente por análisis poblacionales, mientras 11, ya han sido evaluados por esta metodología (61-63, 68, 90). Posteriormente, estas secuencias fueron utilizadas con el fin de obtener las homólogas de estos 59 genes de cuatro cepas de *P. vivax* adaptadas a primates no humanos (Brasil-I, India-VII, Mauritania-I y Corea del Norte (29)) y de dos especies filogenéticamente relacionadas (*Plasmodium cynomolgi* (45) y *Plasmodium knowlesi* (10)), mediante una búsqueda por tBlasn (29). Adicionalmente, se realizó una búsqueda en la base de datos *GenBank* para obtener secuencias reportadas para otras cepas de *P. vivax* (VCG-I, Belem o Corea del Sur).

Análisis de las relaciones filogenéticas de los genes pertenecientes a las familias multigénicas.

Algunos genes incluidos en el análisis pertenecen a familias multigénicas, por lo tanto, previo a los análisis de diversidad genética y de selección natural, se realizó la identificación de las relaciones de ortología dentro de cada familia multigénica. Una combinación de 3 criterios fue utilizada para esta identificación, estos incluyen:

a. una señal filogenética (topología de un árbol filogenético)
b. similitud de secuencia (distancia genética), y
c. sintenia (posición genómica similar)

como ha sido previamente descrito (91-94).

Las secuencias de *P. vivax*, *P. cynomolgi* y *P. knowlesi* de las familias *sera*, *msp3*, *msp7*, *clag*, *PfamA* y *Pfamd*, fueron alineadas independientemente con todos los miembros de sus respectivas familias utilizando el método MUSCLE (95). El mejor modelo evolutivo para

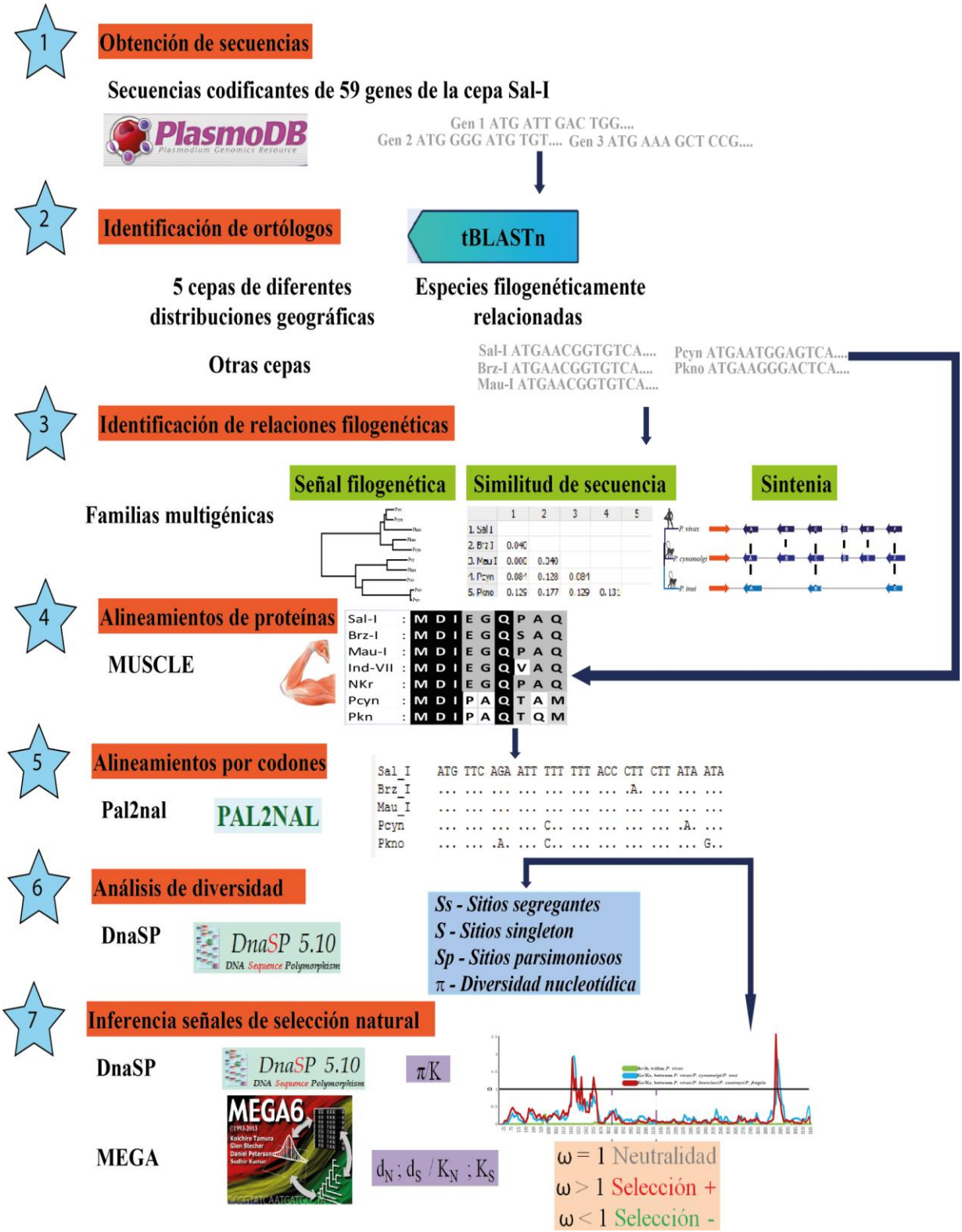
cada familia fue seleccionado por el criterio de información bayesiano, utilizando el programa MEGA v.5 (96) y fue utilizado para inferir árboles filogenéticos mediante el método de Máxima verosimilitud (ML). *Gaps* y regiones repetitivas no fueron tenidos en cuenta para los análisis. La confiabilidad de la topología fue evaluada por *bootstrap* (1.000 repeticiones). Las distancias genéticas fueron calculadas usando el programa MEGA v.5 y la sintenia fue determinada teniendo en cuenta estudios previos (91, 92, 94) y la información disponible en PlasmoDB (www.plasmodb.org/plasmo/).

Análisis de diversidad genética e inferencia de señales de selección natural.

Ya con la identificación de los genes ortólogos putativos de las familias multigénicas, se realizaron alineamientos múltiples para todos los 59 genes, usando las secuencias disponibles de cepas de *P. vivax* adaptadas a primates y las secuencias ortólogas de dos especies filogenéticamente relacionadas. A partir de estas secuencias de ADN, se infirió la secuencia de proteínas, las cuales fueron alineadas con el método MUSCLE (95). A partir de este alineamiento, se infirió el alineamiento de ADN (por codones) usando el programa Pal2Nal (97). Para todos los análisis, se utilizó la anotación de la cepa Sal-I como referencia. A partir de los alineamientos obtenidos, se estimaron varias medidas de la diversidad genética, utilizando el programa DnaSP v.5 (98).

Recientemente, Cornejo *et al*., (84) identificaron patrones consistentes con selección natural en el genoma de *P. vivax* usando la prueba de Hudson, Kreitman y Aguade, la versión genómica de la prueba de McDonald-Kreitman (MK) y el estimador D de Tajima (84); sin embargo, no se reportaron señales de selección positiva en los genes del merozoíto implicados en la invasión. En el presente trabajo, se evaluaron señales de la selección natural (positiva, balanceante y negativa) mediante la utilización de pruebas de comparación de datos de diferentes especies (99) tales como: la prueba convencional de MK (100) y la razón π/K (diversidad nucleotídica/divergencia nucleotídica) (72, 85). La prueba MK se realizó teniendo en cuenta la corrección de divergencia de Jukes-Cantor (101), mediante el uso del servidor web http://mkt.uab.es/mkt/MKT.asp (102). La razón π/K fue evaluada para la identificación de genes con valores mayores a 0,1, lo cual es esperado en genes bajo selección balanceante (72, 85).

Figura 4. Metodología del Capítulo 1



El programa MEGA v.5 (103) fue utilizado para evaluar señales de selección natural dentro de *P. vivax*, mediante el cálculo de las tasas de sustitución no-sinónima por sitio no-sinónimo (d_N) y la tasa de sustitución sinónima por sitio sinónimo (d_S), utilizando el método modificado de Nei-Gojobori (104). Así mismo, para inferir patrones de selección natural que podrían haber prevalecido durante la historia evolutiva de *Plasmodium*, se calcularon la diferencia entre el número promedio de la divergencia no-sinónima por sitio no-sinónimo y el número promedio de la divergencia sinónima por sitio sinónimo (K_N-K_S, adecuado para detectar selección pasada (99)), utilizando el método modificado de Nei-Gojobori con la corrección de Jukes-Cantor. Diferencias significativas entre d_N y d_S (o K_N y K_S) fueron evaluadas por la prueba *Z-test* o la prueba exacta de Fisher. Adicionalmente, se realizó una ventana deslizante de la tasa ω (omega = d_N/d_S o K_N/K_S) para determinar regiones específicas afectadas por tipos particulares de selección natural. Regiones con *gaps* o regiones repetitivas, no fueron tenidas en cuenta para los análisis. La Figura 4 esquematiza la metodología realizada.

Resultados y Discusión

Secuencias de 59 genes de *P. vivax* (obtenidas de al menos 5 cepas adaptadas a primates), que codifican potenciales candidatos a vacuna, fueron analizados con respecto a su diversidad genética y señales de selección natural (Publicación 1). Los cálculos de los estimadores de diversidad mostraron 16 genes altamente polimórficos ($\pi > 0{,}01$), 35 con un polimorfismo intermedio ($0{,}009 < \pi > 0{,}001$) y 8 que mostraron una baja diversidad genética ($\pi < 0{,}001$) (Tabla 2). La Figura 5, muestra la distribución de los polimorfismos dentro de cada uno de los 59 genes.

Tabla 2. Estimadores de la diversidad genética en 59 genes de *P. vivax*.

#	ID	Gen	n	Sitios	Ss	S	Ps	π (SD)
Genes sin estudios previos de genética de poblaciones								
1	PVX_000945	*pvron-1*	6	2340	5	3	2	0,0096 (0,0001)
2	PVX_000995	*pv41*	6	1086	14	5	9	0,0064 (0,0013)
3	PVX_002510	nucleosomal binding protein 1	5	750	2	1	1	0,0013 (0,0003)
4	PVX_003800	*Pvsera*	5	3033	4	4	0	0,0005 (0,0005)

5	PVX_003805	*pvsera*, putative	5	3507	436	195	241	0,0728 (0,0100)
6	PVX_003815	*pvsera*, truncated, putative	5	1335	15	12	3	0,0049 (0,0014)
7	PVX_003825	*pvsera-4*	4	2814	144	113	31	0,0282 (0,0072)
8	PVX_003830	*pvsera-5*	6	3090	752	528	224	0,1038 (0,0222)
9	PVX_003850	*pvsera-2*	6	3042	12	9	3	0,0016 (0,0004)
10	PVX_080305	hypothetical protein, conserved	5	804	1	1	0	0,0005 (0,0003)
11	PVX_081810	hypothetical protein, conserved	5	3921	18	15	3	0,0020 (0,0004)
12	PVX_081845	hypothetical protein	5	1044	3	1	2	0,0015 (0,0003)
13	PVX_084720	hypothetical protein, conserved	5	2720	9	5	4	0,0016 (0,0003)
14	PVX_086850	*pvvir-35*, putative	4	662	40	19	21	0,0360 (0,0090)
15	PVX_086930	*pvrhopH1/clag*	5	3978	38	30	8	0,0043 (0,0010)
16	PVX_090075	*pv34*	6	1092	1	1	0	0,0003 (0,0003)
17	PVX_090210	*Pvarp*	7	82	6	5	1	0,0029 (0,0007)
18	PVX_091434	*pvron-4*	6	2097	14	5	9	0,0033 (0,0007)
19	PVX_092425	hypothetical protein, conserved	4	1950	25	25	0	0,0070 (0,0032)
20	PVX_092975	erythrocyte binding protein 1	6	3440	36	31	5	0,0038 (0,0009)
21	PVX_092995	tryptophan-rich antigen	5	1059	25	19	6	0,0105 (0,0036)
22	PVX_094425	hypothetical protein, conserved	5	3045	3	2	1	0,0005 (0,0001)
23	PVX_096990	Pv-fam-d protein	5	1136	22	12	10	0,0095 (0,0022)
24	PVX_097565	Plasmodium exported protein	5	1311	6	6	0	0,0018 (0,0003)
25	PVX_097670	*pvmsp-3γ*, putative	6	1743	524	282	242	0,1408 (0,0154)
26	PVX_097675	*pvmsp-3γ*, putative	5	1758	445	281	164	0,1268 (0,0229)
27	PVX_097695	*pvmsp-3α*, putative	5	2613	424	240	184	0,0817 (0,0126)
28	PVX_097700	*pvmsp-3*, putative	4	3306	770	577	201	0,1332 (0,0245)
29	PVX_097705	*pvmsp-3α*, putative	5	2607	440	264	176	0,0841 (0,0110)
30	PVX_097710	*pvmsp-3*, putative	5	3607	867	506	361	0,1229 (0,0185)
31	PVX_097715	*pvmsp-3*, putative	5	1286	49	42	7	0,0163 (0,0034)
32	PVX_097960	*pv38*	6	1065	4	0	4	0,0022 (0,0004)
33	PVX_098585	*pvrbp-1*, putative	6	8451	39	22	17	0,0020 (0,0003)
34	PVX_098712	*pvrhopH3*	6	2673	4	3	1	0,0006 (0,0002)
35	PVX_101505	Pv-fam-d protein	5	1263	5	2	3	0,0021 (0,0004)
36	PVX_101555	hypothetical protein	5	2586	167	98	69	0,0337 (0,0081)
37	PVX_101605	hypothetical protein	5	582	3	0	3	0,0031 (0,0030)
38	PVX_109280	*pvfam-a*	7	753	1	1	0	0,0004 (0,0003)
39	PVX_112665	tryptophan-rich antigen	5	867	3	1	2	0,0018 (0,0004)
40	PVX_113775	*pv12*	7	942	4	2	1	0,0014 (0,0006)
41	PVX_117230	*pvser/thr*	5	4122	5	3	2	0,0006 (0,0001)
42	PVX_117880	*pvron-2*	7	6495	31	23	8	0,0016 (0,0003)
43	PVX_118525	hypothetical protein, conserved	5	5082	19	13	6	0,0017 (0,0004)
44	PVX_121885	*pvclag*, putative	5	4239	63	44	19	0,0069 (0,0010)
45	PVX_121920	*pvrbp-2*, like	5	7461	31	15	16	0,0021 (0,0003)
46	PVX_123105	hypothetical protein, conserved	5	2114	1	1	0	0,0002 (0,0001)

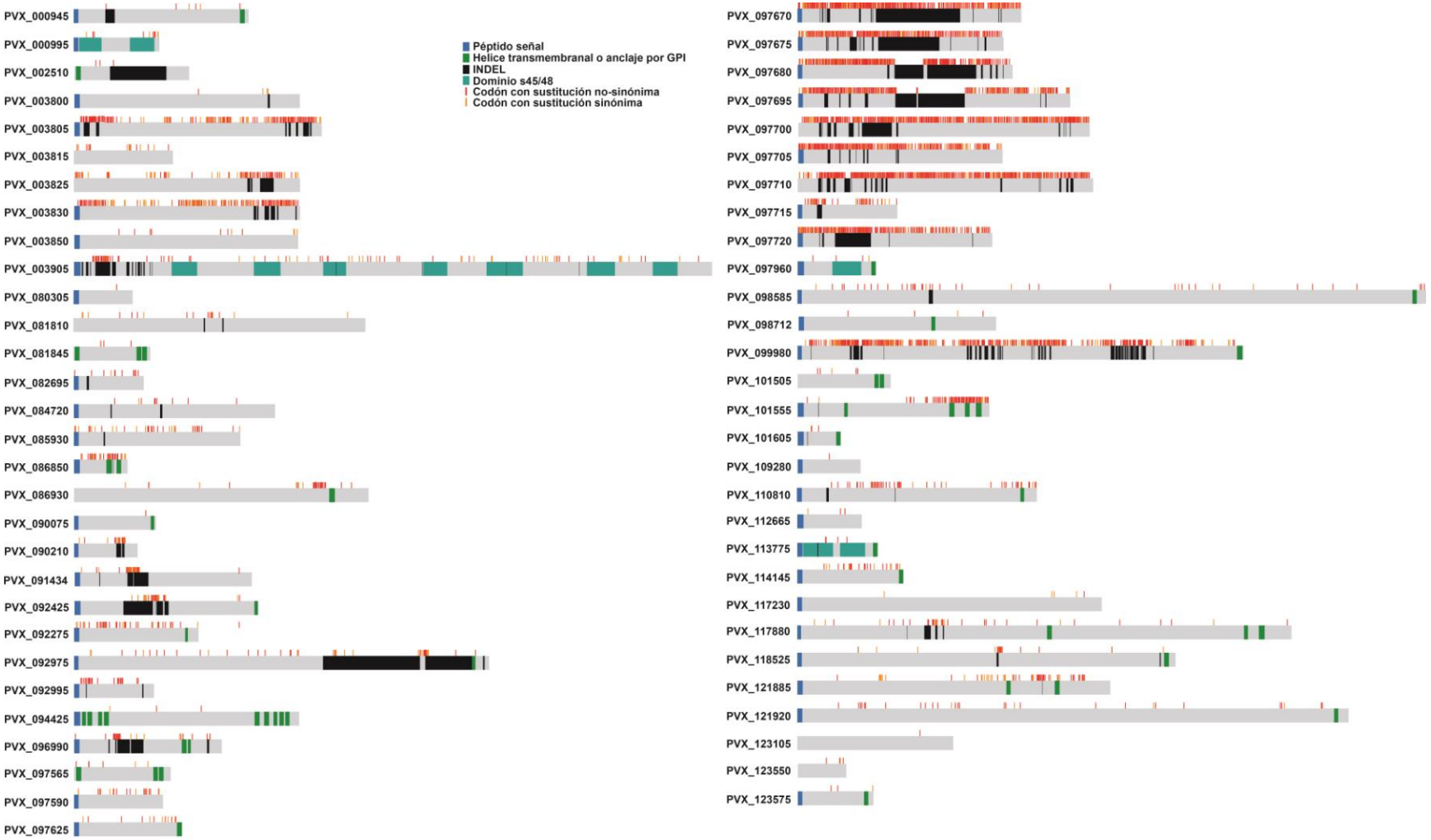
47	PVX_123550	hypothetical protein, conserved	5	647	4	3	1	0,0027 (0,0006)
48	PVX_123575	thrombospondin-related protein 3	6	966	3	2	1	0,0012 (0,0004)
Genes con estudios previos de genética de poblaciones								
1	PVX_003905	*pv230*, putative	5	8199	22	14	8	0,0013 (0,0002)
2	PVX_082695	*pvmsp-7K*, putative	6	849	7	5	2	0,0033 (0,0006)
3	PVX_085930	*pvrap-1*, putative	5	2223	2	2	0	0,0003 (0,0002)
4	PVX_092275	*pvama-1*	6	1683	31	18	13	0,0080 (0,0010)
5	PVX_097590	*pvrap-2*, putative	5	1203	0	0	0	0,000 (0,000)
6	PVX_097625	*pvmsp-8*	6	1320	6	3	3	0,0021 (0,0006)
7	PVX_097720	*pvmsp-3α*	6	2016	154	72	82	0,0349 (0,0158)
8	PVX_097680	*pvmsp-3β*	6	2007	329	165	164	0,0747 (0,0094)
9	PVX_099980	*pvmsp-1*	6	5058	532	170	362	0,0527 (0,0068)
10	PVX_110810	*Pvdbp*	6	3210	45	23	22	0,0063 (0,0010)
11	PVX_114145	*pvmsp-10*	6	1288	5	0	5	0,0021 (0,0003)

n: Número se secuencias analizadas; Sitios: total de sitios analizados, excluyendo gaps; Ss: número de sitios segregantes (polimórficos); S: número de sitios singleton; Ps: número de sitios parsimoniosos; π: diversidad nucleotídica por sitio; SD: desviación estándar.

Previo a la detección de señales de selección natural en estos 59 genes, las relaciones filogenéticas de miembros de familias multigénicas (identificación de ortólogos putativos) fueron establecidas (Anexo 1).

Una vez establecidas las relaciones de ortología, las señales de selección fueron detectadas mediante tres enfoques diferentes. Con la prueba MK se detectaron 16 genes bajo selección balanceante (índice de neutralidad (NI) > 1) y uno bajo selección positiva (NI < 1). La relación π/K identificó 11 genes bajo selección balanceante (Tabla 3). Estos genes podrían ser blancos del sistema inmune y, por lo tanto, podrían ser considerados como potenciales candidatos a vacuna. Sin embargo, el alto polimorfismo dentro de algunos de ellos, reduciría la eficacia de la vacuna debido a las respuestas inmunes alelo-específicas que se presentarían.

Figura 5. Distribución del polimorfismo dentro de los 59 genes *Plasmodium vivax* estudiados.



Los codones con mutaciones sinónimas (naranja) y no sinónimas (rojo) se muestran con líneas verticales encima de cada gen. En azul se representan las regiones que codifican el péptido señal, en verde las regiones predichas como transmembranales o de anclaje a GPI, en cian oscuro se representan los dominios s48/45 y en negro las regiones con INDELs (INserciones o DELeciones).

Sólo 7 genes mostraron señales de selección balanceante por ambas pruebas (π/K y MK). Una baja correlación entre estas pruebas ha sido previamente demostrada (72). Esto podría ser debido al efecto de la selección negativa débil, que puede producir un exceso de polimorfismo, pero a bajas frecuencias. Sin embargo, el valor del NI > 1 (o Pn/Ps > Dn/Ds) encontrado en este estudio, parece deberse al alto número de sustituciones sinónimas entre las especies (Ds), resultando en un Pn/Ps > Dn/Ds. El alto número de sustituciones sinónimas entre las especies, sugiere que la evolución ha mantenido las secuencias de proteínas conservadas en ambas, por lo que una restricción funcional/estructural parece ser probable. Esto sugiere que algunos genes podrían no estar bajo selección balanceante, sino bajo una selección negativa. Esta interpretación fue soportada por el hecho de que se encontró valores bajos de ω y valores negativos estadísticamente significativos en la prueba K_N-K_S (Tabla 3 y Anexo 2).

Tabla 3. Razón diversidad nucleotídica/divergencia (π/K), índice de neutralidad de la prueba McDonald-Kreitman (NI) y tasas de sustitución no sinónima (d_N) y sinónima (d_S) para algunos genes de *P. vivax* evaluados.

#	ID	Gen	Pcyn	Pkno	Pcyn		Pkno		d_N (SE)	d_S (SE)
			π/K	π/K	NI	p-value	NI	p-value		
Genes sin estudios previos de genética de poblaciones										
2	PVX_000995	*pv41*	0,045	0,033	3,218	0,031	2,960	0,044	0,0052 (0,0022)	0,0098 (0,0041)*
4	PVX_003800	*pvsera*	0,003	0,002	0,781	0,831	2,476	0,507	0,0002 (0,0002)	0,0014 (0,0008)‡
5	PVX_003805	*pvsera*, putative	0,325	-	2,592	0,000	-	-	0,0729 (0,0046)	0,0725 (0,0060)
7	PVX_003825	*pvsera-4*	0,164	-	2,949	0,000	-	-	0,0257 (0,0027)	0,0347 (0,0044)†
8	PVX_003830	*pvsera-5*	-	-	-	-	-	-	0,1028 (0,0048)•	0,0791 (0,0057)
11	PVX_081810	hypothetical protein, conserved	0,024	0,014	2,014	0,158	2,069	0,165	0,0015 (0,0005)	0,0032 (0,0011)‡
13	PVX_084720	hypothetical protein, conserved	0,008	0,006	10,919	0,000	8,928	0,000	0,0020 (0,0007)	0,0006 (0,0005)
14	PVX_086850	*pvvir-35*, putative	0,265	-	1,335	0,552	-	-	0,0407 (0,0071)ǂ	0,0240 (0,0081)
15	PVX_086930	*pvrhopH1/clag*	0,032	0,023	9,252	0,000	10,703	0,000	0,0048 (0,0010)	0,0029 (0,0011)
18	PVX_091434	*pvron-4*	0,021	0,016	3,262	0,024	0,649	0,465	0,0032 (0,0014)	0,0037 (0,0016)
19	PVX_092425	hypothetical protein, conserved	0,066	0,048	0,591	0,238	0,709	0,420	0,0045 (0,0015)	0,0104 (0,0028)◊
21	PVX_092995	tryptophan-rich antigen	0,057	-	Null	0,027	-	-	0,0142 (0,0032)•	0,0000 (0,0000)
25	PVX_097670	*pvmsp-3γ*, putative	0,756	-	1,499	0,092	-	-	0,1601 (0,0074)•	0,0906 (0,0080)
26	PVX_097675	*pvmsp-3γ*, putative	0,615	-	1,556	0,031	-	-	0,1412 (0,0072)•	0,0880 (0,0082)
27	PVX_097695	*pvmsp-3α*, putative	-	-	-	-	-	-	0,0885 (0,0048)◊	0,0689 (0,0059)
28	PVX_097700	*pvmsp-3*, putative	-	-	-	-	-	-	0,1397 (0,0054)◊	0,1160 (0,0074)
29	PVX_097705	*pvmsp-3α*, putative	-	-	-	-	-	-	0,0864 (0,0047)‡	0,0705 (0,0062)

30	PVX_097710	*pvmsp-3*, putative	-	-	-	-	-	-	0,1359 (0,0055)•	0,0956 (0,0059)
31	PVX_097715	*pvmsp-3*, putative	0,077	0,049	0,647	0,337	0,973	0,956	0,0136 (0,0024)	0,0225 (0,0050)†
33	PVX_098585	*pvrbp-1*, putative	0,014	-	2,441	0,056	-	-	0,0024 (0,0004)†	0,0008 (0,0004)
36	PVX_101555	hypothetical protein	0,263	-	2,115	0,003	-	-	0,0401 (0,0040)•	0,0169 (0,0031)
40	PVX_113775	*pv12*	0,006	0,005	Null	0,002	Null	0,000	0,0020 (0,0010)	0,0000 (0,0000)
41	PVX_117230	*pvser/thr*	0,007	0,004	0,389	0,384	0,430	0,438	0,0001 (0,0001)	0,0018 (0,0009)◊
42	PVX_117880	*pvron-2*	0,011	0,008	4,859	0,000	4,507	0,000	0,0016 (0,0004)	0,0016 (0,0006)
43	PVX_118525	hypothetical protein, conserved	0,014	0,009	1,125	0,877	1,479	0,395	0,0013 (0,0004)	0,0030 (0,0010)◊
44	PVX_121885	*pvclag*, putative	0,096	-	0,966	0,904	-	-	0,0059 (0,0011)	0,0095 (0,0019)◊
47	PVX_123550	hypothetical protein, conserved	0,048	0,029	9062	0,025	2,437	0,376	0,0016 (0,0011)	0,0059 (0,0041)
Genes con estudios previos de genética de poblaciones										
1	PVX_003905	*pv230*	0,005	0,004	2,7	0,025	2	0,093	0,0012 (0,0004)	0,0016 (0,0006)
4	PVX_092275	*pvama-1*	0,056	0,045	6,923	0,000	4,876	0,000	0,0066 (0,0017)	0,0085 (0,0029)
7	PVX_097720	*pvmsp-3α*	0,259	-	0,571	0,030	-	-	0,0324 (0,0035)	0,0416 (0,0054)†
8	PVX_097680	*pvmsp-3β*	0,533	-	1,312	0,254	-	-	0,0819 (0,0051)†	0,0558 (0,0063)
9	PVX_099980	*pvmsp-1*	0,274	0,203	2,300	0,000	2,744	0,000	0,0485 (0,0028)	0,0464 (0,0037)
10	PVX_110810	*Pvdbp*	0,050	0,015	2,924	0,013	3,228	0,005	0,0074 (0,0012)	0,0032 (0,0013)
11	PVX_114145	*pvmsp-10*	0,010	0,006	0,289	0,252	0,801	0,856	0,0011 (0,0008)	0,0046 (0,0025)*

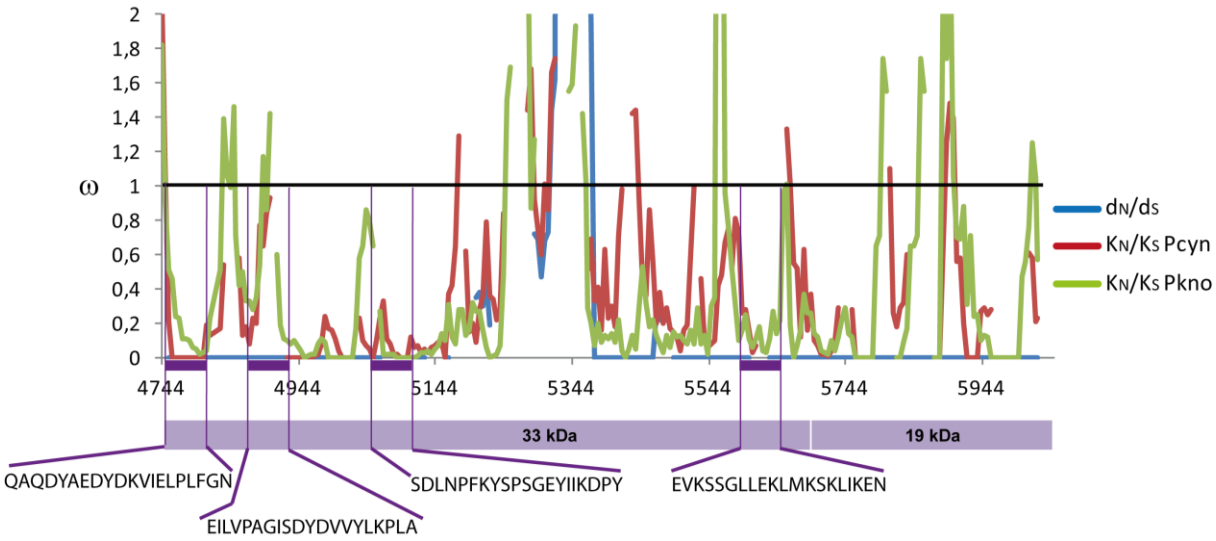
*-: El valor de la prueba no pudo ser estimado ya que no se encontraron secuencias ortólogas; Null: El índice de neutralidad (NI) no pudo ser estimado debido a que el polimorfismo neutral (o no neutral) es igual a 0; SE: error estándar; Pcyn: Valor obtenido al comparar las secuencias de P. vivax con P. cynomolgi; Pkno: Valor obtenido al comparar las secuencias de P. vivax con P. knowlesi. El valor de p para el NI de pv230, pvmsp-1 y pvama-1 no pudo ser estimado en el servidor web; por lo tanto, se calcularon con el programa DnaSP, que no considera la corrección de Jukes-Cantor. *: p < 0,06; ɾ: p < 0,05; ‡; p < 0,04; ◊: p < 0,01; †: p < 0,002; •: p < 0,0001. Se muestran sólo genes en los que se detectó una señal de selección.*

Los resultados de la prueba d_N y d_S mostraron 12 genes con señales de selección positiva, 9 con señales de selección negativa (Tabla 3), mientras que para los restantes, la neutralidad no pudo ser descartada. Esto sugiere que genes como *pvsera5*, *pvvir35*, los miembros de la familia *pvmsp3*, *pvrbp1*, PVX_092995 y PVX_101555 podrían estar bajo presión del sistema inmune, fijando o acumulando mutaciones de tipo no-sinónimo como mecanismo de evasión. Por lo tanto, una vacuna basada en las proteínas codificadas por estos genes, podría tener una baja efectividad debido a su alto polimorfismo. Por otra parte, las proteínas codificadas por genes bajo selección negativa podrían ser tenidas en cuenta para el desarrollo de vacunas (68, 74-76). Genes tales como *pv41*, PVX_092425, *pvclag* (PVX_121885) o *pvmsp10*, podrían ser buenos candidatos a vacuna, ya que la selección

negativa podría ser consecuencia de restricciones funcionales y, por lo tanto, las respuestas inmunes alelo-específicas serían abolidas.

Las presiones selectivas (o la restricción funcional/estructural) no son iguales a lo largo de toda la molécula, por lo tanto, diferentes tipos de selección podría actuar sobre ésta (90). Para determinar cómo se han acumulado las mutaciones sinónimas y no sinónimas en estos genes, se realizó una ventana deslizante de las tasas ω (d_N/d_S o K_N/K_S) permitiendo identificar dominios/regiones bajo selección positiva (ω > 1) o negativa (ω < 1). El análisis de ventana deslizante, mostró que varios genes que carecen de valores significativos de d_N o d_S, tienen dominios con un valor ω > 1, mientras otras regiones presentaban valores de ω < 1 (Anexo 2). Interesantemente, las regiones con un valor de ω < 1, coinciden con aquellas regiones donde dominios funcionales han sido descritos, sugiriendo que estos dominios tienen restricciones funcionales, debido a su rol en la invasión del parásito a la célula hospedera (Figura 6 y Anexo 2).

Figura 6. Ventana deslizante de la tasa ω en la región C-terminal de PvMSP-1



El fragmento de PvMSP-1 de 33 kDa está involucrado en el reconocimiento de la célula hospedera (38). Las regiones específicas que interactúan con los reticulocitos (resaltadas en cuadros púrpura) presentan una baja divergencia (ω < 1), mientras las regiones no funcionales presentan una alta divergencia (ω > 1). El péptido 2 muestra una alta divergencia que podría ser el resultado de adaptaciones especie-específicas (105, 106), lo que eleva el valor ω. Este resultado concuerda con análisis poblaciones previos de este fragmento (107).

Teniendo en cuenta lo anterior, se buscaron este tipo de regiones (con $\omega < 1$) en genes con señales de selección negativa. Por ejemplo, los genes *pvron2* y *pvron4* presentaron regiones con valores $\omega < 1$ hacia los extremos 3'. Consecuentemente, la región C-terminal de las proteínas PvRON2 y PvRON4 podrían ser las regiones funcionales. Así mismo, miembros de la familia 6-Cys (*pv12, pv41* y *pv38*) con dominios s48/46, mostraron una baja diversidad genética, donde los dominios s48/46 mostraron un exceso de sustituciones sinónimas entre especies, generando un NI > 1 y valores ω (K_N/K_S) < 1, lo cual es consistente con la acción de la selección negativa. Por lo tanto, estos dominios parecen estar bajo restricciones funcionales, convirtiéndolos en potenciales candidatos a vacuna que eviten las respuestas inmunes alelo-específicas. Otras proteínas con este tipo de patrones son *pvrhopH1*/*clag, pvser*/Thr, PVX_081810 y PVX_092425, por lo que estos podrían también ser considerados durante el diseño de una vacuna (Publicación 1).

Conclusiones

El uso de tres pruebas diferentes permitió la identificación de señales consistentes con selección natural. Miembros de las familias multigénicas *pvsera* y *pvmsp3* mostraron señales de selección positiva, probablemente debido a que las proteínas codificadas son reconocidas por el sistema inmune, sin embargo, no parecen ser los candidatos más apropiados para el desarrollo de vacunas, debido a su alto polimorfismo.

Las proteínas codificadas por *pvclag*, *pvser*/thr, *pvrhopH1*/*clag7*, *pvron2*, *pvron4*, *pv12*, *pv38*, *pv41*, PVX_081810 y PVX_092425 (o dominios dentro de ellas), mostraron patrones esperados en regiones bajo restricciones funcionales, por lo tanto, podrían ser consideradas como los candidatos más adecuados y así, ser priorizadas para estudios posteriores (análisis poblacionales y funcionales) para el desarrollo de una vacuna contra *P. vivax* que evite las respuestas inmunes alelo-específicas.

El enfoque acá utilizado (Figura 4) podría entonces ser una herramienta útil para la selección de candidatos promisorios a vacuna. Los antígenos para tener en cuenta deben presentar:

a. una diversidad genética limitada o, al menos, un dominio con este patrón
b. una señal de selección negativa ($d_S > d_N$ o $K_S > K_N$), así como un $\omega < 1$.

Sin embargo, los genes bajo selección positiva ($\omega > 1$), podrían tomarse en cuenta si estos presentan dominios con una diversidad genética limitada y valores bajos en la tasa ω. Los resultados de este capítulo fueron publicados y están disponibles bajo la siguiente referencia:

Publicación 1: **Garzón-Ospina D**, Forero-Rodríguez J, Patarroyo MA. *Inferring natural selection signals in Plasmodium vivax-encoded proteins having a potential role in merozoite invasion*. Infect Genet Evol. 2015 Jul; 33:182-8. doi: 10.1016/j.meegid.2015.05.001.

CAPITULO 2

Adaptaciones linaje-específicas, una limitante para la selección de antígenos basados en el enfoque propuesto.

Introducción

Estudios previos han mostrado que la región C-terminal de miembros de la familia multigénica MSP7 de *P. vivax* (*pvmsp7*) son altamente conservados (63, 90), inclusive entre parálogos (90). Por esta razón, el dominio C-terminal de los antígenos codificados por los genes *pvmsp7* podrían ser buenos candidatos a vacuna (63, 90). Entre el grupo de genes analizados en el Capítulo 1, se encontraba el gen *msp7K*, miembro altamente conservado en *P. vivax* (63), quien no mostró patrones que nos permitieran seleccionarlo como potencial candidato, a pesar de haberlo propuesto previamente (63). Los resultados obtenidos para este gen en el Capítulo 1, no mostraron señales de selección negativa entre especies y por el contrario mostraron que *pvmsp7K* es altamente divergente entre *P. vivax* y *P. cynomolgi*.

Trabajos previos en el gen *msp1* (del inglés, *Merozoite Surface Protein 1*) han sugerido que éste presenta señales de selección positiva linaje-específicas (105, 106), observándose una alta divergencia en este gen entre *P. vivax* y especies de *Plasmodium* que infectan cercopitécidos (primates del viejo mundo). Esta divergencia por selección positiva coincide con la radiación de los macacos asiáticos hace 3 a 6 millones de años [26, 65], sugiriendo que esta selección, podría ser el resultado de adaptaciones a nuevos hospederos disponibles hace 3 a 6 millones de años atrás. Teniendo en cuenta que en *P. falciparum* las proteínas MSP1 y MSP7 forman un complejo implicado en la invasión [9, 10] y dada la presencia de ambas proteínas en *P. vivax*, ¿podrían entonces los miembros de la familia *msp7* de *P. vivax*, estar bajo las mismas presiones selectivas de *msp1*?

Si los genes *pvmsp7* divergieron por selección positiva como consecuencia de la adaptación a los seres humanos durante el cambio de hospedero (15), esto podría incrementar la tasa ω, incrementando la divergencia entre especies, y por ende, *msp7K* sería descartado como potencial candidato a vacuna. Con el fin de evaluar esta hipótesis se reconstruyeron las relaciones filogenéticas de la familia *msp7* en *Plasmodium* y se evaluó la presencia de señales de selección linaje-específica.

Metodología

Usando una aproximación similar a la previamente reportada (92), se identificaron miembros de la familia *msp7* en 13 especies de *Plasmodium* y se establecieron las relaciones de ortología. Las secuencias de ADN de miembros de *msp7* identificados fueron usadas para deducir su respectiva secuencia de aminoácidos, para posteriormente, ser alineadas usando el método MUSCLE (95). El método DLRTS (108) se utilizó con el fin de inferir el árbol de los genes *msp7* teniendo en cuenta la filogenia de *Plasmodium* spp. Adicionalmente, se infirieron árboles filogenéticos utilizando métodos de máxima verosimilitud (ML, con el programa RAxML (109)) y Bayesiano (BY con el programa MrBayes (110)) con el modelo JTT +G +F (seleccionado como el mejor modelo evolutivo por el método ProtTest (111)). Estos análisis se realizaron en la plataforma *CIPRES Science Gateway* (112, 113).

Identificación de selección episódica linaje-específica

El modelo *Branch-site REL* se utilizó para evaluar la presencia de linajes (ramas) en la filogenia de la familia *msp7*, donde un porcentaje de sitios presenten una selección diversificada episódica (114). El método MUSCLE se usó para alinear independientemente las secuencias de aminoácidos de cada grupo ortólogo (Figuras 8 y 9), para posteriormente inferir el alineamiento por codones de cada grupo ortólogo con el programa PAL2NAL (97). El mejor modelo evolutivo para cada alineamiento (para ADN y proteínas) fue determinado usando los programas jModelTest (115) y ProtTest (111), respectivamente. Los alineamientos y filogenias inferidas por ML para cada alineamiento fueron utilizadas como entrada para correr el método *Branch-site REL* en el paquete HyPhy (116). Adicionalmente, el servidor web *Datamonkey* (117) fue utilizado para este mismo análisis. Finalmente, el método MEME (118) se utilizó para inferir qué sitios estaban bajo selección positiva episódica en cada grupo de ortólogos.

Figura 7. Metodología del Capítulo 2

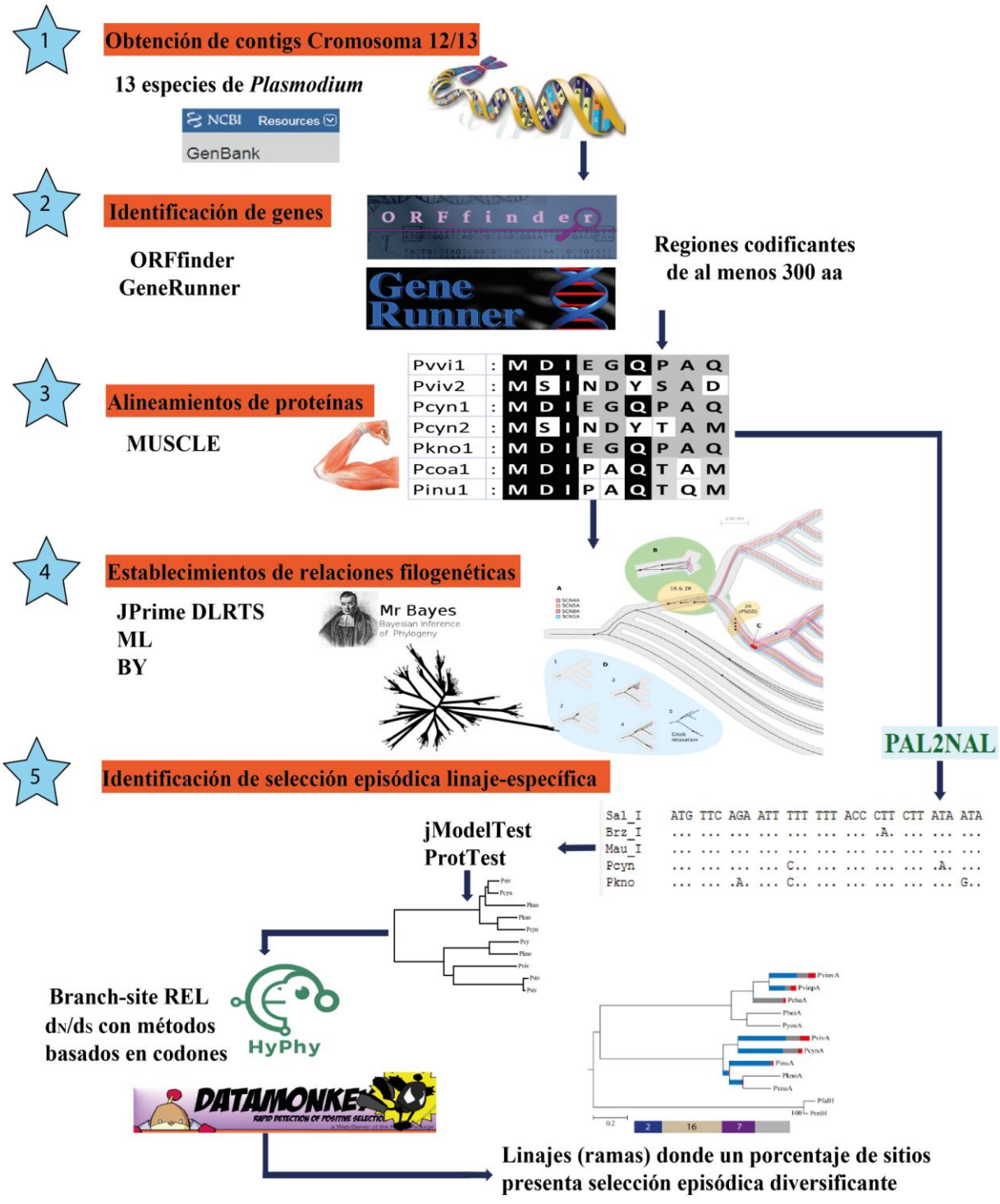
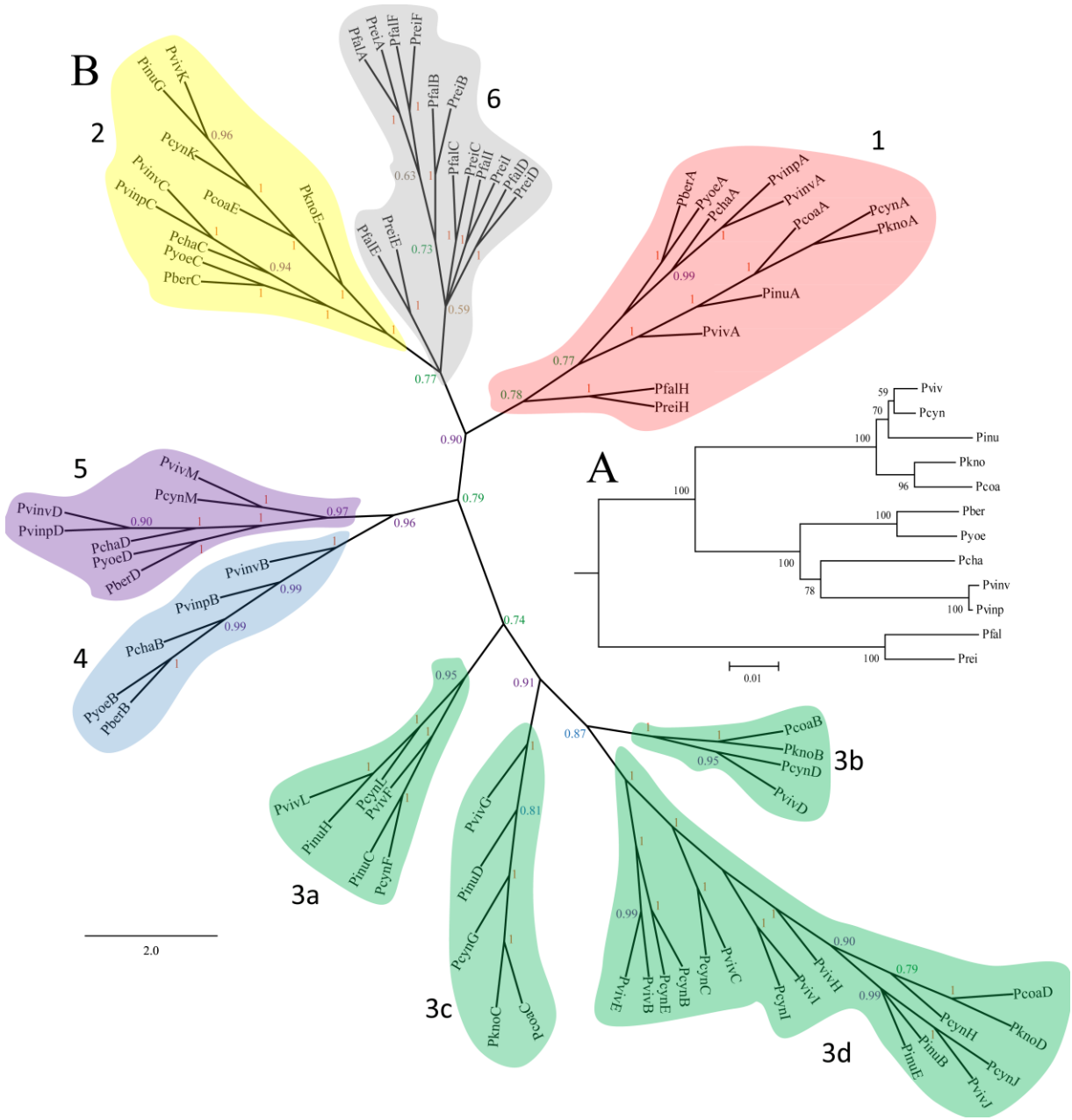
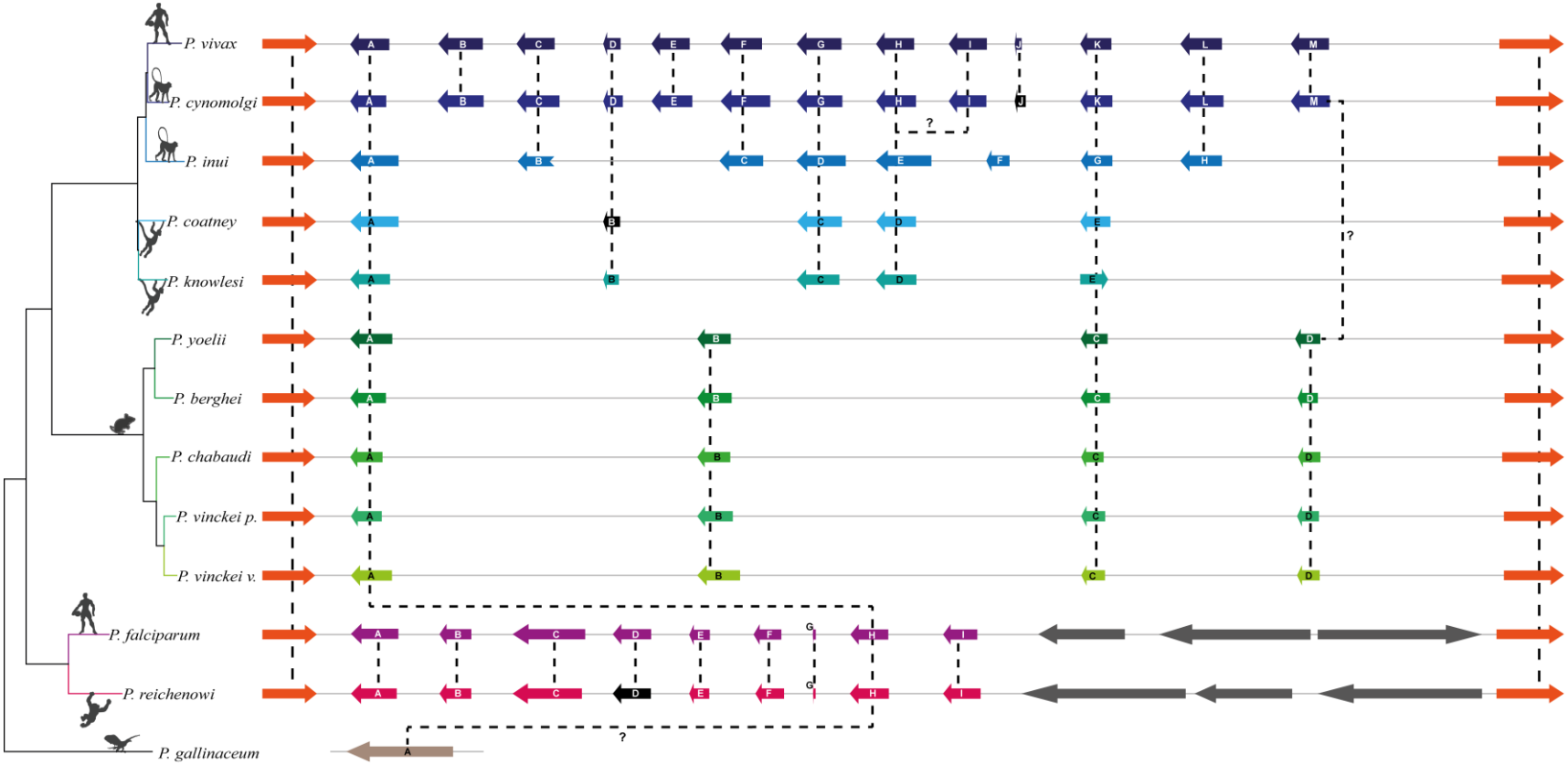


Figura 8. Filogenia de la familia de genes *msp7* inferida por el modelo evolutivo DLTRS



A. Filogenia del género Plasmodium utilizada para inferir el árbol MSP7 por el método DLTRS. B. Árbol MSP7 inferido a partir del árbol del género Plasmodium. Los números representan clados diferentes, mientras que los números en las ramas son valores de probabilidad posteriores. Nueve clados principales fueron identificados en el árbol. Las proteínas se agrupan de acuerdo con las relaciones filogenéticas del parásito, los clados 1 (rojo) y 2 (amarillo) son los más ancestrales y el clado 5 (violeta) también podría ser ancestral. Las secuencias del linaje de parásitos que infectan cercopitécidos se representan en verde, las proteínas del linaje de parásitos que infectan roedores en azul y el linaje de parásitos que infectan homínidos en gris.

Figura 9. Representación esquemática de los loci *msp7* en 13 genomas de *Plasmodium*



Los genes que flanquean la región que contiene la familia msp7 en las especies de Plasmodium están representados por recuadros de color naranja. Los recuadros de colores dentro de las secuencias flanqueantes representan los genes msp7 en cada especie, mientras que los recuadros negros simbolizan los pseudogenes. Los genes se nombraron en orden alfabético de izquierda a derecha. Las líneas punteadas conectan genes ortólogos. Todos los genes están representados a escala, pero la distancia entre ellos no es representativa. Los signos de interrogación indican relaciones de ortología que no fueron claramente establecidas. A diferencia de los parásitos que infectan homínidos, las especies de linaje de parásitos que infectan cercopitécidos y roedores parecen tener historias evolutivas similares con respecto a la expansión de msp7. Los recuadros grises son genes específicos del linaje y que no pertenecen a la familia msp7 (las representaciones de estos últimos genes no están a escala).

Tabla 4. Caracterización *in-silico* de las proteínas MSP7

		Genes *msp7*											
		A	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**	**K**	**L**	**M**
P. vivax	**SP**	y	y	y	y	y	y	y	y	y	-	y	y	y
	MSP7_C	y	y	y	-	y	y	y	y	y	y	y	y	-
P. cynomolgi	**SP**	y	y	y	y	-	y	y	y	y	-	y	y	y
	MSP7_C	y	y	y	-	y	y	y	y	y	y	y	y	-
P. inui	**SP**	y	-	y	y	y	-	y	y					
	MSP7_C	y	y	y	y	y	-	y	y					
P. knowlesi	**SP**	y	y	y	y	y								
	MSP7_C	y	-	y	y	y								
P. coatneyi	**SP**	y	-	y	y	y								
	MSP7_C	y	-	y	y	y								
P. chabaudi	**SP**	y	y	y	y									
	MSP7_C	y	y	y	-									
P. vinckei v.	**SP**	y	y	y	y									
	MSP7_C	y	y	y	-									
P. vinckei p.	**SP**	y	y	y	y									
	MSP7_C	y	y	y	-									
P. berghei	**SP**	y	y	y	y									
	MSP7_C	y	y	y	-									
P. yoelii	**SP**	y	y	y	y									
	MSP7_C	y	y	y	-									
P. falciparum	**SP**	y	y	-	y	y	y	y						
	MSP7_C	y	y	y	y	y	y	y	y	y				
P. reichenowi	**SP**	y	-	y	y	y	y	y						
	MSP7_C	y	y	y	y	y	y	y	y	y				
P. gallinaceum	**SP**	y												
	MSP7_C	y												

Las proteínas MSP7 fueron evaluadas con respecto a la presencia de péptido señal (SP) y del dominio MSP_7C (número de acceso en Pfam: PF12948), característico de los miembros de esta familia. Con la letra "y" se indica la presencia del péptido señal o del dominio. Aquellas proteínas que no presentaron estas estructuras se indican con el símbolo "-".

Resultados y discusión

Estructura genética y relaciones filogenéticas de la región msp7 en Plasmodium spp.

La región cromosómica en *P. vivax*, donde la familia *msp7* está ubicada, es delimitada por los genes PVX_082640 y PVX_082715 (92, 93, 119), por lo tanto, ortólogos a éstos fueron identificados en las demás especies (Figura 9). Dado que las proteínas MSP7 parecen ser codificadas por un único exón (54), las regiones cromosómicas entre los genes que delimitan la familia, fueron analizadas con la herramienta *ORFFinder* y *GeneRunner*. Este análisis identifico 79 marcos abiertos de lectura (ORFs por sus siglas en inglés, Publicación 2) con la misma orientación de trascripción (Figura 9). Estos ORF fueron nombrados en orden alfabético con respecto a su aparición después del gen PVX_082640 (o sus ortólogos en cada especie). Posteriormente, se determinó la presencia del péptido señal y del dominio MSP_7C (Tabla 4). El número de genes *msp7* varía de especie a especie, sugiriendo que eventos de duplicación (o deleción) han ocurrido de manera especie-específica. Esto confirma resultados previos (92, 119), donde se ha sugerido que esta familia evoluciona bajo el modelo de nacimiento y muerte de genes (Figura 9).

Relaciones filogenéticas de los miembros de msp7 en Plasmodium

El método DLTRS (108, 120) fue utilizado para reconciliar el árbol de los genes *msp7* con la filogenia del género. En el árbol inferido por este método (Figura 8) se observan 9 grupos principales. Los grupos 1, 2, 3a, 3b, 3c, 4 y 5 representan grupos ortólogos, mientras en los grupos 3d y 6 se agrupan tanto ortólogos como parálogos. Este agrupamiento fue similar al observado en las filogenias inferidas por ML y BY (Anexo 3). El grupo 1 (MSP7A/H) fue el gen más ancestral, seguido por el grupo 2, el cual incluye secuencias de parásitos que infectan cercopitécidos y roedores. En *P. gallinaceum*, se observó un gen con una longitud mayor a los otros *msp7*, pero aún no es claro si este gen es ortólogo al gen más ancestral (*msp7A/H*). Como sólo se encontró un gen *msp7* en el genoma de *P. gallinaceum*, la expansión de la familia *msp7* debió haber ocurrido después de la radiación de parásitos de mamíferos hace 40 millones de años (11).

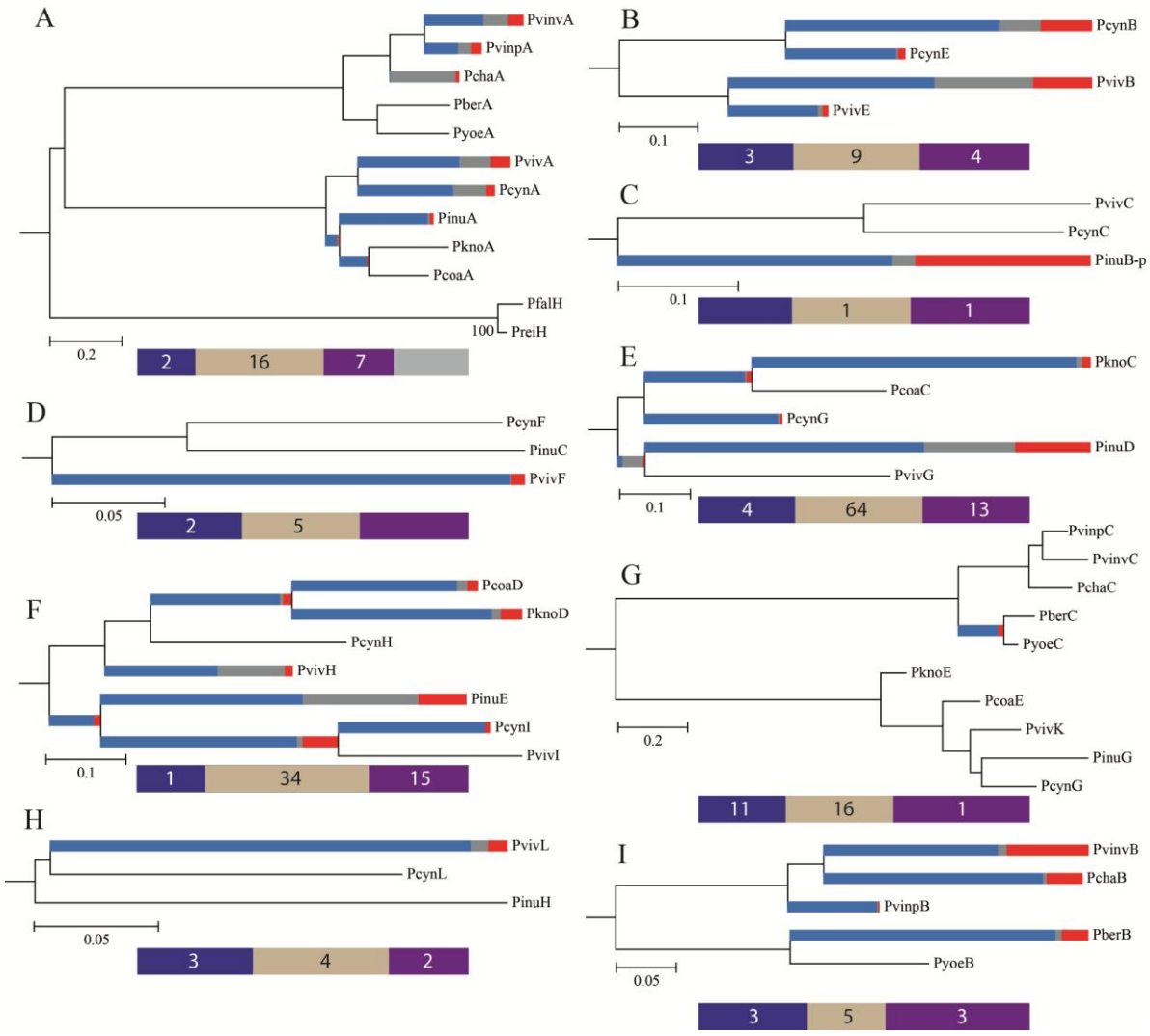
Selección positiva linaje-especifica en msp7

MSP1 es la proteína que media la interacción inicial entre los merozoítos y los glóbulos rojos (121-124). Estudios previos han mostrado que el gen *msp1* presenta señales de selección positiva. Esta selección ocurrió en algún momento y dentro de sólo algunos linajes ancestrales (selección positiva episódica linaje-específica) dentro del grupo de parásitos de cercopitécidos (105, 125) y esto sería el resultado de la adaptación de poblaciones ancestrales del parásito a las nuevas especies de macacos que surgieron hace 3,7 a 5,1 millones de años (105) o durante el cambio de hospedero que dio origen a *P. vivax* (15). Dado que MSP7 forma un complejo con MSP1 (121, 123, 124), MSP7 debería estar bajo las mismas presiones selectivas y por lo tanto, señales de selección linaje-específica podrían estar presentes en los miembros de esta familia. Varias ramas (linajes) mostraron señales de selección positiva en miembros de la familia *msp7*. Como en MSP1 (105, 125), muchas de estas ramas eran ramas internas (Figura 10A, E, F and G). La mayor cantidad de codones bajo selección positiva se localizó en las regiones centrales de *msp7* y pocos en los extremos 5' o 3'. Estudios previos han mostrado que la región central es la más polimórfica y parece estar involucrada en la evasión de la respuesta inmune (90). Los sitios seleccionados positivamente en las regiones centrales de las proteínas MSP7, podrían ser la consecuencia de la evasión a la respuesta inmune del hospedero, lo que permitiría la adaptación del parásito. Por otro lado, los sitios bajo selección positiva en la región C-terminal, podrían ser el resultado de la coevolución entre el receptor del hospedero y los ligandos MSP7 del parásito o del resultado de la adquisición de nuevas funciones (Publicación 2).

Si la selección positiva ha actuado de manera linaje-específica, se observará una alta divergencia entre las secuencias de especies relacionadas. Mientras un determinado aminoácido se fijó en un linaje (o especie) por selección positiva en una determinada posición, en los otros linajes (o especies), aminoácidos diferentes podrían presentarse en esa misma posición. Por lo tanto, esto generará un incremento en la tasa ω cuando se comparan las secuencias. Así, aunque se trate de regiones funcionalmente importantes, la presencia de la selección positiva linaje-específica en esta región, hará que la tasa ω sea alta

y, por lo tanto, regiones funcionales con este patrón, serán descartadas si se sigue el enfoque presentado en el Capítulo 1.

Figura 10. Filogenias de *msp7s* analizadas por el método *Branch-site REL*



Cada grupo ortólogo identificado en la Figura 9 fue analizado por el método Branch-site REL. El color sobre las ramas indica el tipo de selección (rojo $\omega > 13$, selección positiva; azul $\omega \leq 1$, selección negativa) mientras el color gris representa un $\omega = 1$ (neutralidad). La longitud de cada color representa el porcentaje de sitios en la clase correspondiente encontrada por Branch-site REL. A. Clado 1; B. pviv/pcynmsp7B y 7E; C. pviv/pcynmsp7C y pinumsp7B; D. pviv/pcynmsp7F y pinumsp7C; E. pviv/pcynmsp7G, pkno/pcoamsp7C y pinumsp7D; F. pviv/pcynmsp7H/7I, pkno/pcoamsp7D y pinumsp7E; G. Clado 2; H. pviv/pcynmsp7L y pinumsp7H y I. Clado 4. En la parte inferior de cada filogenia hay una representación a escala del gene msp7 correspondiente. Los cuadros azul oscuro dentro de este esquema representan la región N-terminal, los marrones claros simbolizan la región central y los cuadros

morados, el dominio MSP_7C. Los números dentro de las casillas representan el número de codones bajo selección positiva inferida por los métodos MEME, SLAC, FEL, REL y FUBAR utilizando el servidor web Datamonkey.

Conclusiones

La familia *msp7* es el resultado de procesos de duplicación diferencial en *Plasmodium* spp, lo que genera que el número de copias varíe de una especie a otra. Estos genes han fijado mutaciones por selección positiva de manera linaje-específica. Este tipo de selección genera un incremento en las tasas evolutivas, a pesar de que se trate de regiones funcionalmente importantes, por lo tanto, señales de selección positiva linaje-específicas podrían representar una limitante cuando se seleccionen genes como candidatos a vacuna siguiendo el enfoque presentado en el Capítulo 1.

Los resultados de este capítulo fueron publicados y están disponibles bajo la siguiente referencia:

Publicación 2: **Garzón-Ospina D**, Forero-Rodríguez J, Patarroyo MA. *Evidence of functional divergence in MSP7 paralogous proteins: a molecular-evolutionary and phylogenetic analysis*. BMC Evol Biol. 2016 Nov 28;16 (1):256. doi: 10.1186/s12862-016-0830-x.

CAPITULO 3

Validación del enfoque alternativo presentado en el Capítulo 1 para la selección de candidatos a vacuna.

Introducción

El desarrollo de una vacuna antimalárica es una de las estrategias propuestas para el control de la malaria. Sin embargo, en *P. vivax*, pocos antígenos han sido caracterizados como candidatos promisorios a ser incluidos dentro de una vacuna (126, 127). Adicionalmente, la alta diversidad genética presente en algunos de los antígenos parasitarios (61, 64, 65, 67, 70, 71, 90, 128, 129) ha obstaculizado el desarrollo de una vacuna debido a las respuestas inmunes alelo-específicas que esta diversidad podría generar (32, 34, 59). Por tal motivo, la evaluación de la diversidad genética de los antígenos parasitarios es esencial durante el diseño de una vacuna completamente efectiva. No obstante, llevar a cabo estos estudios para todos y cada uno de los genes candidatos, implicaría tiempo y recursos económicos, convirtiéndose en una nueva limitante para el desarrollo de una vacuna contra este patógeno. En el Capítulo 1 de este proyecto, se propuso un flujo de trabajo que permite realizar una selección de antígenos promisorios en un menor tiempo. Sin embargo, con el fin de validar este flujo de trabajo, se realizaron análisis poblacionales para determinar si lo observado a partir de 5 secuencias de asilados de diversas partes del mundo, coincide con los análisis realizados a partir de un número mayor de secuencias.

Metodología

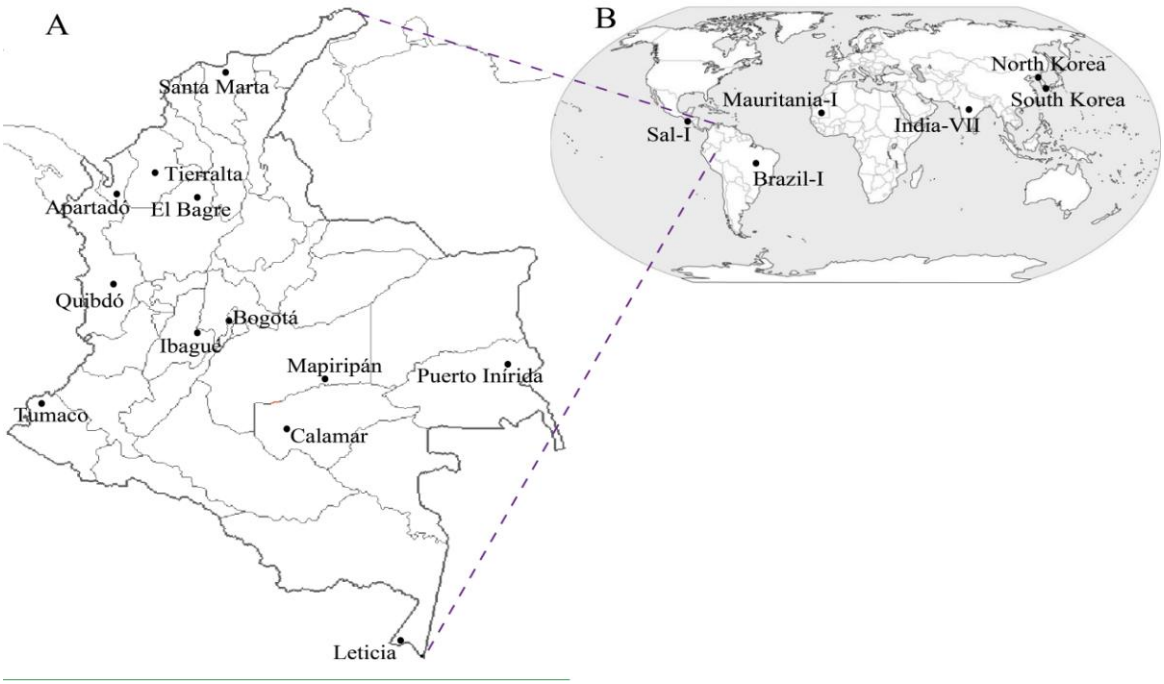
Entre los genes seleccionados en el Capítulo 1 como candidatos promisorios, se destacan tres miembros de la familia del dominio de 6-cisteinas (6-Cys), tres proteínas de roptrias, algunos antígenos de membrana y una proteína hipotética. Algunos de estos fueron seleccionados para ser analizados mediante análisis poblacionales y así validar los datos obtenidos en el Capítulo 1. Adicionalmente, miembros de la familia *msp7* fueron también seleccionados para estos análisis.

Verificación de la integridad del ADN.

El ADN parasitario utilizado fue extraído a partir de muestras de sangre total provenientes de diferentes regiones del país (Figura 11) obtenidas desde el 2007 hasta el 2016. La presencia e integridad del ADN parasitario (almacenado a -20 ^{o}C) fue verificado por la amplificación por PCR de la subunidad ribosomal 18S de *P. vivax*. Los productos de PCR

fueron evaluados por electroforesis en gel de agarosa al 1,5%, con el marcador de peso Hyperladder I.

Figura 11. Lugar de procedencia de las muestras de ADN parasitario en Colombia



A. Procedencia de las muestras de ADN genómico obtenido entre el 2007 y el 2016. B. Procedencia de las secuencias de las cepas de referencia

Determinación de la presencia de infecciones únicas por P. vivax y de diferentes genotipos en las muestras utilizadas.

La técnica de PCR-RPLFs fue implementada con el fin de determinar si las muestras positivas para *P. vivax* correspondían a infecciones únicas. Adicionalmente, esto permitió evaluar la presencia de diferentes genotipos en las muestras. Un fragmento del gen *msp1* (o *msp3*) fue amplificado por PCR con iniciadores previamente reportados (130, 131). Los amplímeros fueron entonces digeridos con las enzimas de restricción Alu-I y Mnl-I (Alu-I y Hha-I para *msp3*). Los productos de esta digestión fueron evaluados por electroforesis en gel de agarosa al 3%, con el marcador de peso Hyperladder V. Adicionalmente, algunos de los amplímeros de *msp3* fueron secuenciados para confirmar la presencia de infecciones únicas.

Diseño de iniciadores y amplificación por PCR

El diseño de los iniciadores para la amplificación del genes seleccionados, se realizó con el programa *GeneRunner* (Hastings software, Inc.) y el servidor en-línea *Oligoanalizer* (IDT® *Integrated DNA Technologies*, https://www.idtdna.com/calc/analyzer), utilizando para ésto, los datos genómicos de la cepa Sal-I de *P. vivax*. Se diseñó un juego de iniciadores para la amplificación de cada gen, y de ser necesario, otros para la secuenciación. La optimización de las condiciones de PCR para la amplificación de los genes, se realizó utilizando los iniciadores diseñados previamente. Para ésta, se utilizó la enzima de alta fidelidad *KAPA-HiFi HotStart Readymix* (kapabiosystems®) y ADN genómico de la cepa VCG-I como control positivo. Una vez optimizadas las condiciones de PCR, se amplificó el ADN parasitario de al menos 30 muestras procedentes de Colombia (Publicaciones 3 - 6).

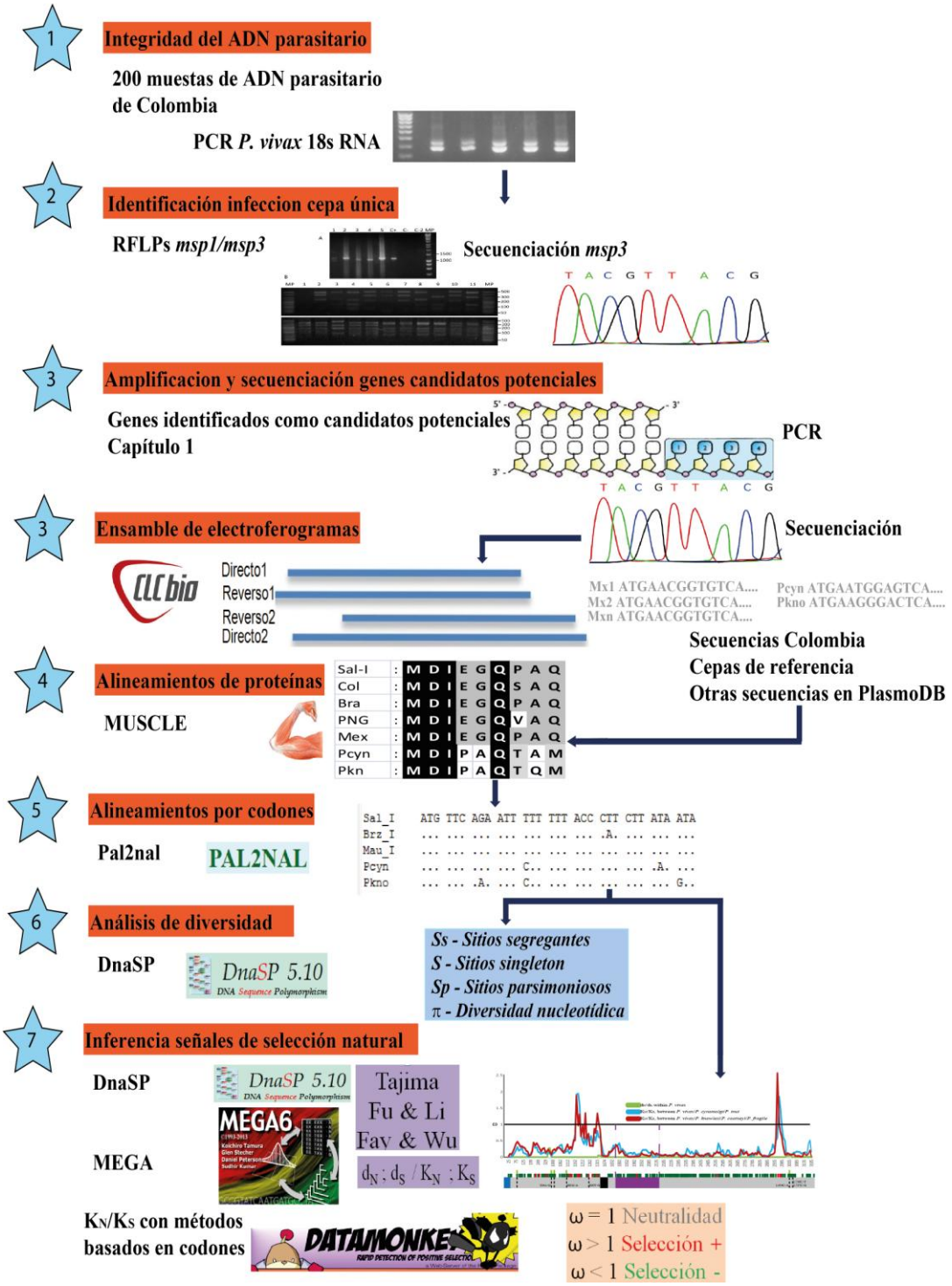
Purificación y Secuenciación de los fragmentos amplificados.

Los productos de PCR fueron purificados con el *kit* de purificación *UltraClean PCR Clean-up* (MOBIO®), siguiendo las recomendaciones del fabricante. Los fragmentos purificados fueron verificados por medio de electroforesis en gel de agarosa al 1,5%, con el marcador de peso Hyperladder I. Para el gen *pvron4* se realizó la clonación de este fragmento en el vector pGEM-T (Publicación 6). Los productos de PCR purificados o el ADN plasmídico fueron secuenciados bidireccionalmente con los iniciadores de amplificación o los iniciadores universales T7 y SP6. La secuenciación se realizó utilizando el *kit BigDye Terminator* en Macrogen, Seúl, Corea del Sur. Al menos dos productos de PCRs independientes o dos plásmidos por muestra, fueron enviados a secuenciación.

Análisis de la diversidad genética y de las fuerzas evolutivas en los loci seleccionados.

Se analizaron y ensamblaron los electroferogramas obtenidos por secuenciación usando el programa *CLC DNA workbench* (CLC bio, Cambridge, MA, USA). Una vez corregidas las secuencias, éstas fueron comparadas y analizadas frente a secuencias de referencia y de especies filogenéticamente relacionadas disponibles en bases de datos.

Figura 12. Metodología del Capítulo 3



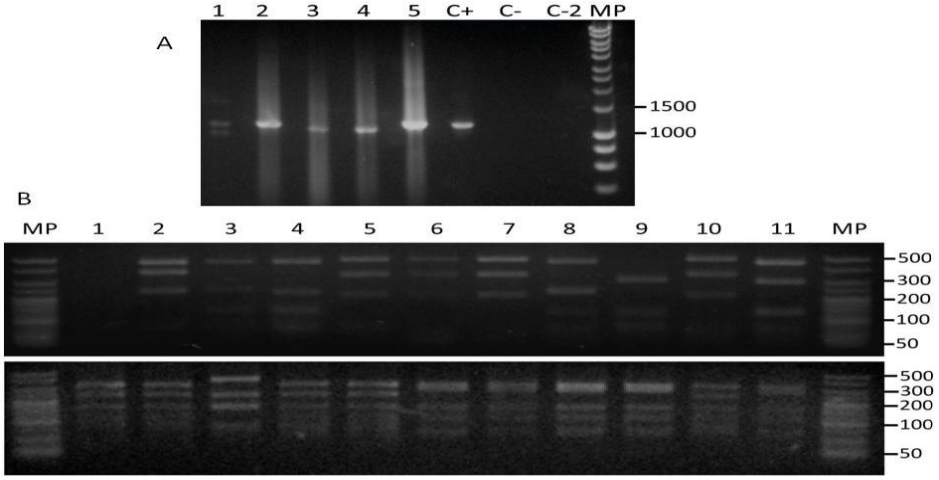
Con las secuencias de la población Colombiana, se realizaron alineamientos de las secuencias derivadas de aminoácidos con el método MUSCLE (95) y a partir de éste, se infirió el alineamiento de ADN usando el programa Pal2Nal (97). A partir de este último alineamiento y usando el programa DnaSP v.5 (98), se calcularon el número de sitios segregantes y la diversidad nucleotídica por sitio.

Se evaluaron señales de selección natural utilizando métodos basados en el cálculo de sustituciones de tipo sinónimo y no-sinónimo. Usando el programa MEGA v.5 (103), se calcularon el número promedio de sustituciones sinónimas por sitio sinónimo (d_S) y el número promedio de sustituciones no-sinónimas por sitio no-sinónimo (d_N) empleando el método modificado de Nei-Gojobori (104). Por otra parte, para determinar señales de selección natural entre especies, se calculó el número promedio de divergencias sinónimas por sitio sinónimo (K_S) y el número promedio de divergencias no-sinónimas por sitio no-sinónimo (K_N), empleando el método modificado de Nei-Gojobori, teniendo en cuenta la corrección de Jukes-Cantor (101). Las diferencias estadísticas entre estas tasas se determinaron aplicando la prueba *Z-test*, incorporada en el programa MEGA. Presiones selectivas por sitio (codón) fueron inferidas calculando las tasas de sustituciones sinónimas y no-sinónimas mediante los métodos SLAC, FEL, REL, IFEL (132), MEME (118) y FUBAR (133), implementados en el servidor en línea Datamonkey (117), con previa evaluación de recombinación por el método GARD (134).

Resultados y discusión

Muestras de ADN parasitario provenientes de diferentes regiones del país (Figura 11), fueron utilizadas para la amplificación de los genes *pv12*, *pv38*, *pv41*, *pvmsp7E*, *pvmsp7F*, *pvmsp7L*, *pvron4* y *pvceltos*. Previo a la amplificación de estos genes, se realizó una aproximación para la detección de infecciones por cepa única mediante PCR-RFLP. Adicionalmente, esta técnica permitió identificar la presencia de diferentes genotipos en las muestras utilizadas (Figura 13).

Figura 13. Patrones de PCR-RFLPs



A. *Amplificación del fragmento 2 de pvmsp1; Carril 1: Muestra con infección múltiple, carriles del 2-5: Algunas de las muestras utilizadas en este estudio, C+: ADN de la cepa VCG-I, C-: control negativo, C-2: ADN cepa FVO de P. falciparum, MP: marcador de peso. B. Carriles del 1-11: Algunas de las muestras utilizadas en este estudio, panel superior fragmentos de digestión con Alu-I, panel inferior fragmentos de digestión con Mnl-I, MP: marcador de peso.*

Fragmentos de los genes *pv12*, *pv38* (Publicación 3), *pv41* (Publicación 4), *pvmsp7E*, *pvmsp7F*, *pvmsp7L* (Publicación 5), *pvron4* (Publicación 6) y *pvceltos* (manuscrito en preparación) fueron amplificados y secuenciados. A partir de las secuencias obtenidas, varios parámetros de diversidad para cada gen fueron calculados (Tabla 5). Estos parámetros fueron entonces comparados con los obtenidos al analizar sólo secuencias de cepas de referencia (Tabla 5). Posteriormente, estos mismos parámetros fueron calculados a partir de alineamientos con todas las secuencias disponibles a la fecha para cada gen.

Ventanas deslizantes de la diversidad nucleotídica por sitio (π), fueron obtenidas a partir de estos 3 sets de datos (Figura 14). Adicionalmente, a partir de estos 3 sets de datos, fueron calculadas las diferencias entre K_N y K_S (Tablas 6), así como ventanas deslizantes para la tasa evolutiva ω (K_N/K_S, Figura 15). Los resultados observados en las Figuras 14 y 15, así como las tablas 5 y 6, muestran una concordancia entre los datos obtenidos a partir de 5 cepas de *P. vivax* de diferentes regiones del mundo y los análisis realizados a partir de un número mayor de secuencias.

Tabla 5. Estimadores de diversidad genética

	Datos Capitulo 1				Datos Capitulo 3				Datos con secuencias disponibles			
Gen	**n**	**Sitios**	**Ss**	**π**	**n**	**Sitios**	**Ss**	**π**	**n**	**Sitios**	**Ss**	**π**
pv12	7	942	4	0,0014 (0,0006)	70	1047	1	0,0003 (0,0001)	156	930	19	0,0011 (0,0001)
pv38	6	1065	4	0,0022 (0,0004)	46	1062	8	0,0024 (0,0002)	132	1035	19	0,0027 (0,0002)
pv41	6	1086	14	0,0064 (0,0013)	30	1115	10	0,0028 (0,0005)	119	1116	33	0,0047 (0,0003)
pvmsp7E	5	1047	147	0,0651 (0,0167)	31	1044	164	0,0558 (0,0004)	44	1047	171	0,0548 (0,0044)
pvmsp7F	6	1164	2	0,0007 (0,0002)	36	1176	2	0,0007 (0,0004)	120	1146	8	0,0009 (0,0001)
pvmsp7L	6	1212	3	0,0008 (0,0004)	31	1212	4	0,0006 (0,0001)	122	1212	13	0,0010 (0,0001)
pvron4	6	2097	14	0,0033 (0,0007)	73	2464	5	0,0004 (0,0009)	159	2172	29	0,0006 (0,0001)
pvceltos	5	588	3	0,0024 (0,0010)	62	546	1	0,0008 (0,0001)	209	546	7	0,0018 (0,0001)

Los estimadores de la diversidad genética se calcularon utilizando las secuencias obtenidas de cepas de referencia (datos Capitulo 1), con secuencias obtenidas en Colombia (datos Capítulo 3) y a partir de las secuencias de cepas de referencia, de las secuencias colombianas y de todas las secuencias disponibles a la fecha obtenidas de PlasmoDB o GenBank como un único set de datos. n: número de secuencias analizadas, sitios: total de sitios analizados (sin incluir gaps), Ss: número de sitios segregados, π: diversidad de nucleótidos por sitio.

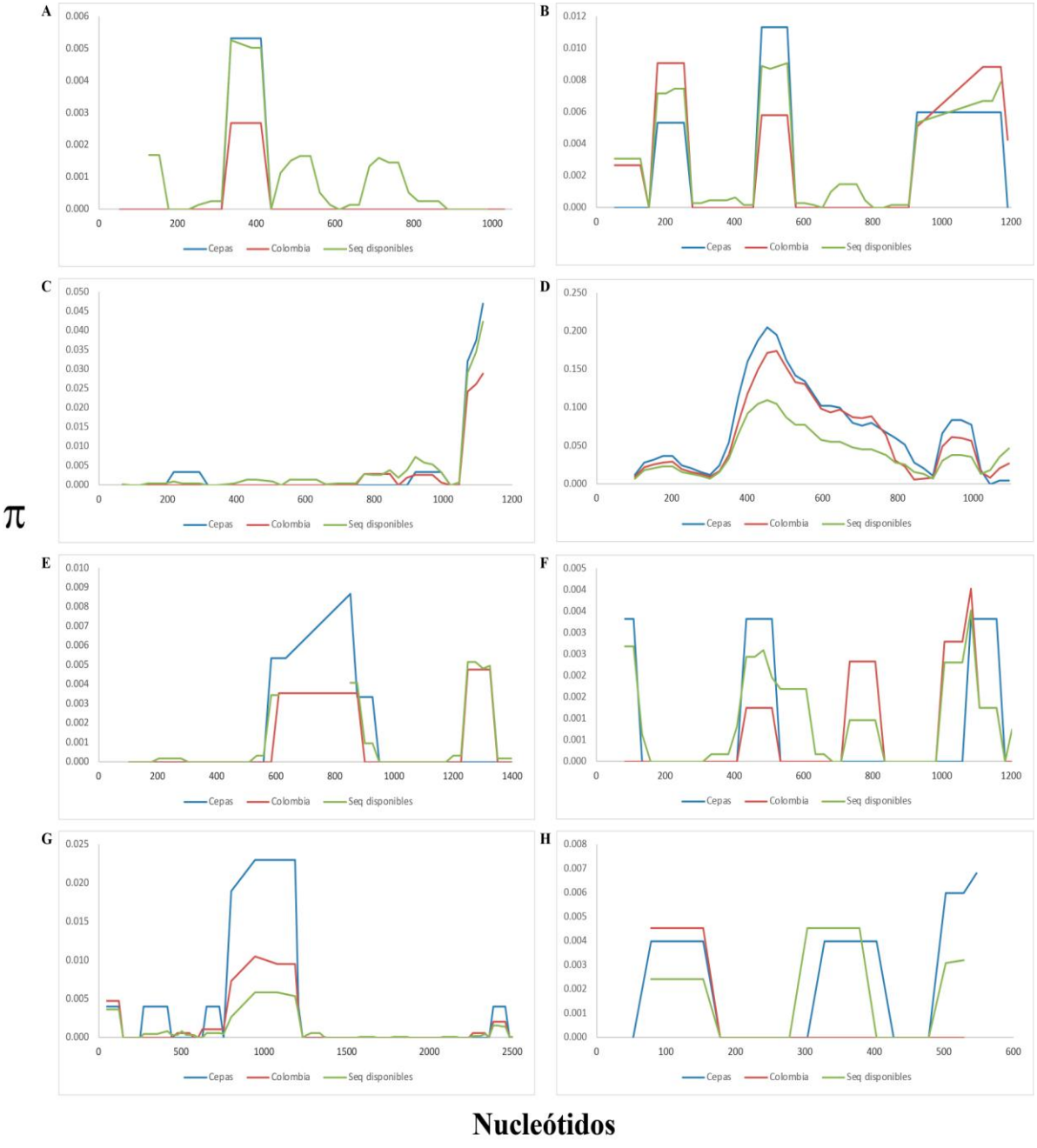
Tabla 6. Tasas de divergencia no-sinónima (K_N) y sinónima (K_S)

	Datos Capitulo 1		Datos Capitulo 3		Datos con secuencias disponibles	
Gen	**n**	**K_N-K_S**	**n**	**K_N-K_S**	**n**	**K_N-K_S**
pv12	7	-0,268 (0,031)*	70	-0,029 (0,003)*	156	-0,013 (0,002)*
pv38	6	-0,288 (0,033)*	46	-0,054 (0,007)*	132	-0,022 (0,003)*
pv41	6	-0,149 (0,019)*	30	-0,040 (0,005)*	119	-0,012 (0,003)*
pvmsp7E	5	-0,073 (0,023)*	31	-0,023 (0,012)**	44	-0,020 (0,011)**
pvmsp7F	6	-0,014 (0,010)	36	-0,002 (0,002)	120	0,000 (0,001)
pvmsp7L	6	-0,018 (0,012)	31	-0,003 (0,003)	122	-0,001 (0,001)
pvron4	6	-0,187 (0,015)*	73	-0,022 (0,002)*	159	-0,011 (0,001)*
pvceltos	5	-0,019 (0,012)	62	-0,006 (0,003)**	209	-0,002 (0,002)

**: p <0,001*

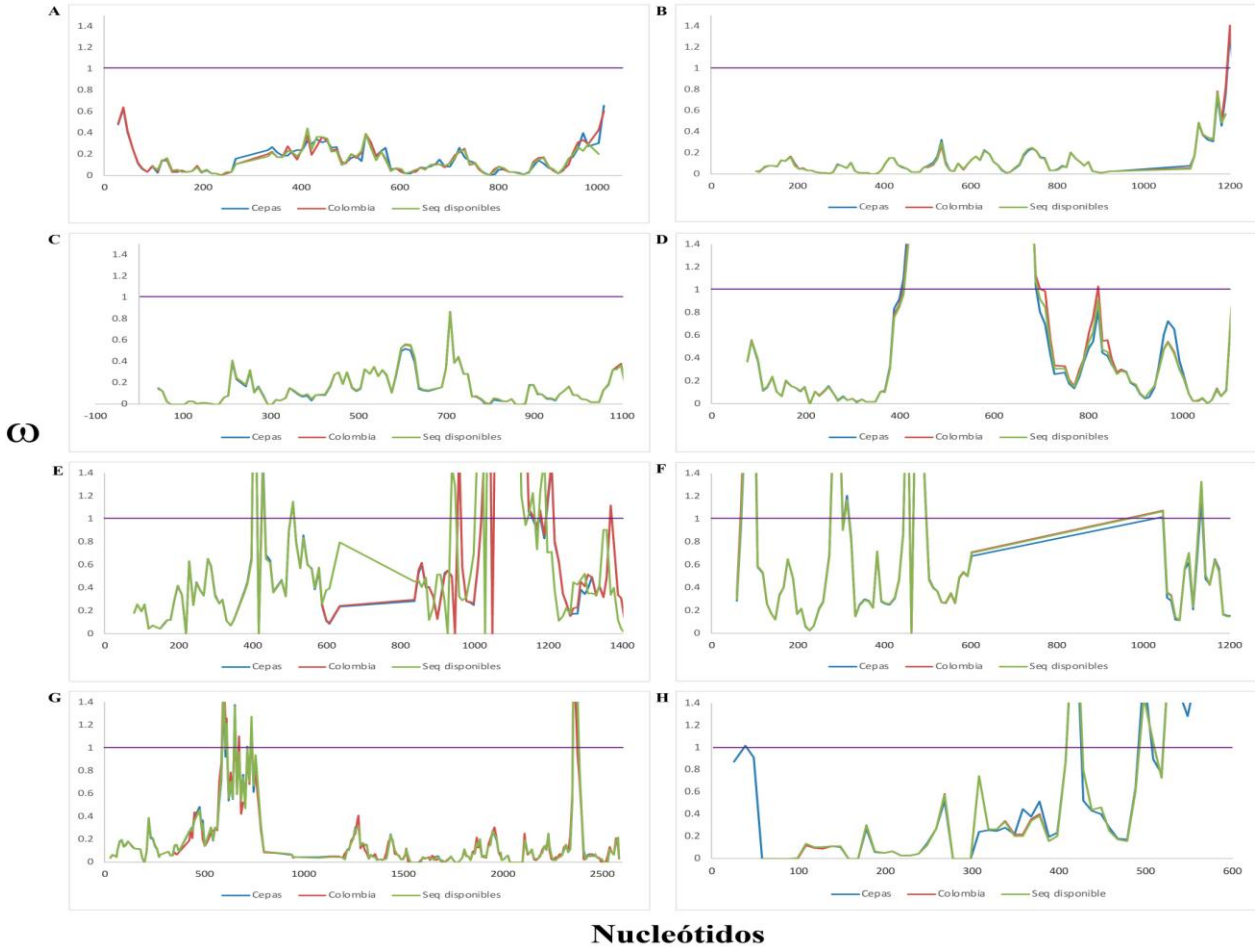
***: p < 0,03*

Figura 14. Ventanas deslizantes de π para 8 antígenos de *P. vivax*



La diversidad nucleotídica por sitio fue calculada mediante ventanas deslizantes para 8 antígenos de P. vivax para comparar los resultados de los tres sets de datos (datos Capitulo 1 [azul, incluye solo las cepas de referencia], datos Capitulo 2 [rojo, incluye las secuencias obtenidas de la población parasitaria de Colombia] y secuencias (seq) disponibles [verde, incluye las secuencias de las cepas, las de parásitos colombianos y otras secuencias disponibles en bases de datos]). A. pv12, B. pv38, C. pv41. D. pvmsp7E, E. pvmsp7F, F. pvmsp7L, G. pvron4 y H. pvceltos.

Figura 15. Ventanas deslizantes de la tasa ω entre especies



Ventanas deslizantes de la tase ω entre especies (K_N/K_S) para 8 antígenos de P. vivax fueron calculadas para comparar los resultados de los tres sets de datos (datos Capitulo 1 [azul, incluye solo las cepas de referencia], datos Capitulo 2 [rojo, incluye las secuencias obtenidas de la población parasitaria de Colombia] y secuencias (seq) disponibles [verde, incluye las secuencias de las cepas, las de parásitos colombianos y otras secuencias disponibles en bases de datos]). A. pv12, B. pv38, C. pv41. D. pvmsp7E, E. pvmsp7F, F. pvmsp7L, G. pvron4 y H. pvceltos.

Conclusiones

A partir de al menos 30 muestras de ADN parasitario obtenido de regiones endémicas de malaria en Colombia, se amplificaron y secuenciaron los genes *pv12, pv38, pv41, pvmsp7E, pvmsp7F, pvmsp7L, pvron4* y *pvceltos*. A partir de estas secuencias, se obtuvieron datos de diferentes estimadores de diversidad genética, así como de la tasa evolutiva ω. Estos resultados mostraron una concordancia con aquellos obtenidos en el Capítulo 1 de este trabajo. Por lo tanto, el enfoque implementado en el Capítulo 1, se convierte en una aproximación alternativa para la selección de candidatos promisorios a vacuna, desde el punto de vista que los candidatos deben ser altamente conservados para evitar las respuestas alelo-específicas que disminuyen la eficacia de las vacunas. Adicionalmente, la evaluación de la tasa evolutiva ω podría ayudar a la identificación de regiones funcionalmente restringidas dentro de los antígenos parasitarios.

Los resultados de este capítulo fueron publicados y están disponibles bajo las siguientes referencias:

Publicación 3: Forero-Rodríguez J, **Garzón-Ospina D**, Patarroyo MA. *Low genetic diversity and functional constraint in loci encoding Plasmodium vivax P12 and P38 proteins in the Colombian population*. Malar J. 2014 Feb 18;13:58. doi: 10.1186/1475-2875-13-58.

Publicación 4: Forero-Rodríguez J, **Garzón-Ospina D**, Patarroyo MA. *Low genetic diversity in the locus encoding the Plasmodium vivax P41 protein in Colombia's parasite population*. Malar J. 2014 Sep 30;13:388. doi: 10.1186/1475-2875-13-388.

Publicación 5: **Garzón-Ospina D**, Forero-Rodríguez J, Patarroyo MA. *Heterogeneous genetic diversity pattern in Plasmodium vivax genes encoding merozoite surface proteins (MSP) -7E, -7F and -7L*. Malar J. 2014 Dec 13;13:495. doi: 10.1186/1475-2875-13-495.

Publicación 6: Buitrago SP, **Garzón-Ospina D**, Patarroyo MA. *Size polymorphism and low sequence diversity in the locus encoding the Plasmodium vivax rhoptry neck protein 4 (PvRON4) in Colombian isolates*. Malar J. 2016 Oct 18;15(1):501. doi: 10.1186/s12936-016-1563-4.

CAPITULO 4

Implementación del enfoque del Capítulo 1 durante la caracterización de nuevos antígenos de *P. vivax*.

Introducción

En 2013, la malERA (del inglés *Malaria Eradication Research Agenda*) reconoció la necesidad de desarrollar de una vacuna contra *P. vivax* debido a las complicaciones que la infección por este parásito está generando en diferentes zonas endémicas (23). Con respecto a *P. falciparum*, la investigación en *P. vivax* se encuentra retrasada debido a la compleja biología de este parásito (135), y a la fecha pocos candidatos a vacuna han sido evaluados en fases clínicas (21). Uno de los principales factores de este retraso es la predilección de *P. vivax* por los reticulocitos, los cuales son células rojas inmaduras que no suelen superar el 1% de las células circulantes en sangre periférica (136, 137). Debido a que *P. vivax* invade exclusivamente este tipo celular, los cultivos *in vitro* de este parásito son de corta duración y, por lo tanto, no ha sido posible realizar las mismas aproximaciones hechas en *P. falciparum* (138, 139). En *P. vivax* se han descrito más de 50 antígenos expresados en el estadio sanguíneo (el merozoíto) (24, 26, 28, 87, 140, 141), pero las regiones de interacción con los reticulocitos sólo han sido establecidas para un reducido número de ellos (36-38, 142). Adicionalmente, *P. vivax* presenta casi el doble de la diversidad genética de *P. falciparum* (29), y ésto, es un problema adicional para el desarrollo de una vacuna.

Para diseñar una vacuna contra *P. vivax*, siguiendo la misma estrategia utilizada en *P. falciparum* (138, 139), se hace entonces necesario, no sólo la identificación de antígenos parasitarios, sino además, se deben establecer las regiones de interacción patógeno-hospedero, así como la diversidad de esos antígenos (u regiones específicas). Sin embargo, para evaluar estos aspectos, sería necesario, primero, contar una fuente permanente de reticulocitos que permitiera realizar ensayos de unión con la proteína completa y diferentes fragmentos de la molécula, siendo ésta, precisamente una de las limitantes que no ha permitido avanzar con los ensayos de unión. Segundo, la diversidad de estos antígenos se hace en determinadas poblaciones, amplificando y secuenciando al menos 30 muestras parasitarias (21).

El enfoque presentado en este trabajo está basado en la premisa que es posible hacer un acercamiento a la diversidad de un antígeno a partir de pocos asilados de diferentes

regiones del mundo. Los resultados del Capítulo 1 de este trabajo, muestran que dominios involucrados en la interacción parasito-glóbulo rojo son conservados entre especies y mostraron señales de selección negativa ($\omega < 1$). Por lo tanto, este enfoque, además de proporcionar información de las regiones ideales para el diseño de una vacuna (conservadas por selección negativa), podría también delimitar qué regiones de un antígeno podrían ser evaluadas en ensayos funcionales y así, en lugar de utilizar toda la molécula (36-38), sólo una parte de ella sería utilizada, optimizando los recursos disponibles. Así, para determinar la utilidad del enfoque presentado en este trabajo durante el desarrollo de una vacuna altamente efectiva contra *P. vivax*, éste fue implementado durante la caracterización de nuevos antígenos de *P. vivax* (publicaciones 7-9).

Metodología

Con la colaboración del grupo de Biología molecular de la FIDIC, quienes describen nuevos antígenos de *P. vivax*, se seleccionaron tres antígenos: PvGAMA, PvRBSA y PvCelTOS. Paralelamente a la caracterización de estos antígenos (143), se implementó la metodología de la Figura 4, con el fin de tener una aproximación a la diversidad genética y las fuerzas evolutivas de estos loci.

Secuencias de ADN que codifican para *pvgama*, *pvrbsa* y *pvceltos* de 5 cepas de *P. vivax* (Sal-I, Brazil-I, India-VII, Mauritania-I y Corea del Norte (29)) y ortólogos de al menos dos especies filogenéticamente relacionadas (*P. cynomolgi* y *P. knowlesi*) (144) fueron obtenidas y alineadas (siguiendo los mismos pasos mencionados en anteriores Capítulos). A partir de las secuencias de *P. vivax*, la diversidad nucleotídica por sitio (π) fue estimada. Usando las secuencias de *P. vivax* junto con las de especies relacionadas, se estimó una ventana deslizante de la tasa ω (K_N/K_S). Sitios bajo selección entre especies fueron inferidos usando los algoritmos SLAC, FEL, REL (132), MEME (118) y FUBAR (133) en el servidor Datamonkey, con previa evaluación de recombinación por el método GARD (134).

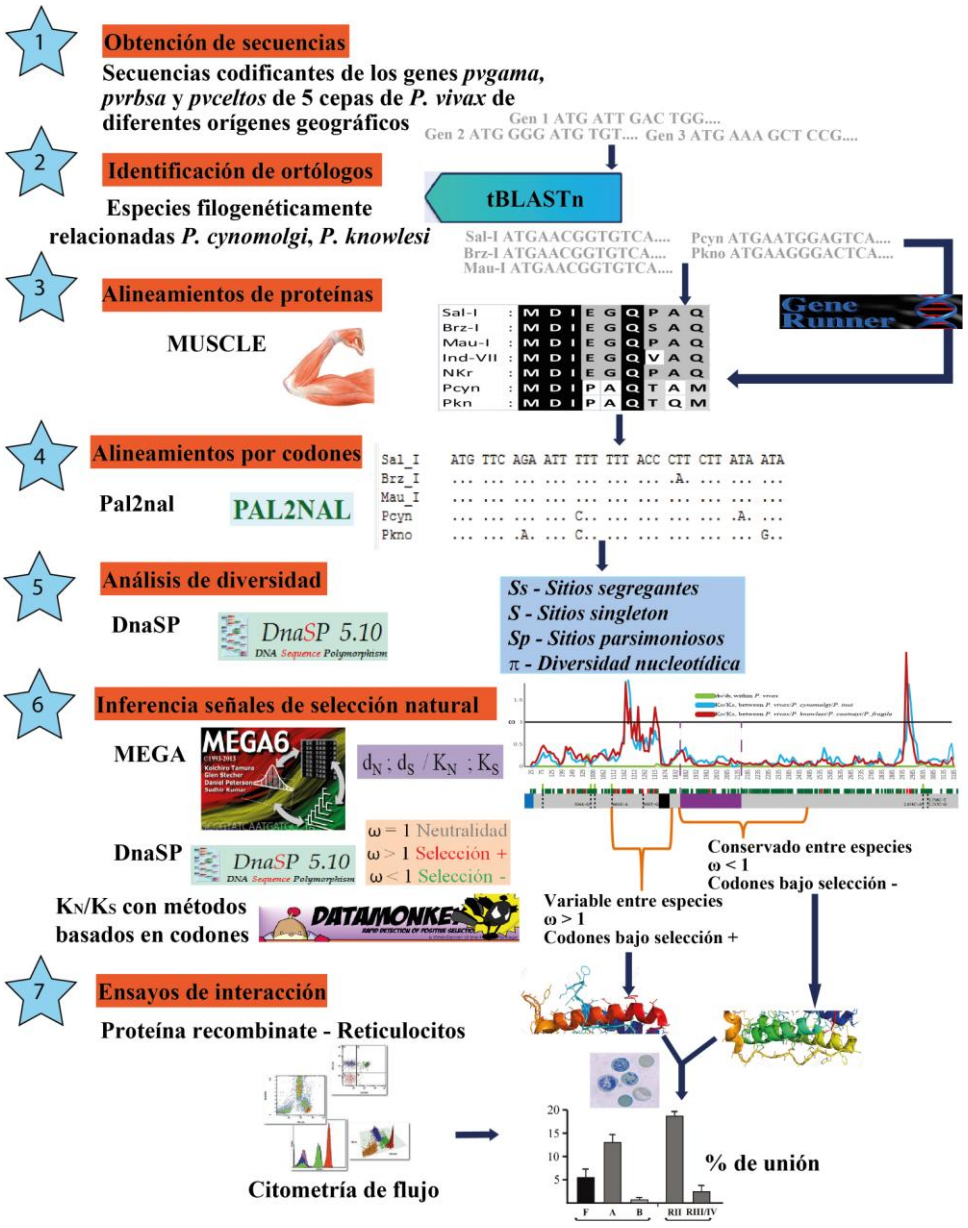
A partir de los resultados de los análisis anteriores, se seleccionaron al menos dos fragmentos de los antígenos *pvgama* (CR1, CR2, VR1 y VR2) y *pvrbsa* (*pvrbsa*-A y *pvrbsa*-B), los cuales fueron evaluados en ensayos de unión, para determinar su participación en la interacción patógeno-hospedero. Estos experimentos fueron realizados por el grupo de Biología Molecular de la FIDIC. Brevemente, los fragmentos seleccionados fueron amplificados por PCR, purificados y ligados a un vector de expresión. El plásmido recombinante fue utilizado para transformar células *E. coli* BL21-DE3. Las células recombinantes fueron utilizadas para la expresión cada fragmento de forma soluble, como ha sido descrito previamente (140). Posteriormente, los fragmentos recombinantes fueron purificados por cromatografía de afinidad en fase sólida, utilizando una resina Ni^{+2}-NTA (Qiagen, Valencia, CA, USA).

La unión de estos fragmentos a reticulocitos humanos (obtenidos de sangre de cordón umbilical (SCU) de niños recién nacidos), fue analizada por triplicado a través de citometría de flujo. $2x10^7$ de reticulocitos fueron incubados con 25µg de cada proteína recombinante por 16h a 4°C y 4 RPM en un agitador de tubos. Posteriormente, las muestras se incubaron con el anticuerpo monoclonal anti-His-PE (dilución 1:40, MACSmolecular-Miltenyi Biotec, San Diego, CA, USA) en oscuridad y sin agitación. Los reticulocitos fueron marcados con los anticuerpos monoclonales anti-CD45-APC (dilución 1:80, Becton Dickinson, Franklin Lakes, NJ, USA) y anti-CD71-APC-H7, respectivamente. Posteriormente, la unión de las moléculas a reticulocitos (CD71+CD45-) fue cuantificada analizando 100.000 eventos en un citómetro FACS Canto II (BD, San Diego, CA, USA) y el programa FACS Diva (145). Las regiones II y III/IV de DBP fueron usadas como control de unión positivo y negativo, respectivamente.

Para PvCelTOS, se realizó un ensayo de competición (inhibición de la unión de la proteína recombínate), el cual fue llevado a cabo por el grupo de Receptor-Ligando de la FIDIC. Brevemente, en experimentos iniciales, la proteína PvCelTOS recombinante mostró ser capaz de interactuar con células HeLA. Para determinar las regiones dentro de la proteína responsables de esta interacción, péptidos de 20 aminoácidos cubriendo la totalidad de la

secuencia, fueron sintetizados. Estos fueron pre-incubados en un radio molar de 1:20 (proteína:péptido) con $2x10^7$ células de la linea HeLA por una hora a 4°C y 4 RPM en un agitador de tubos. Posteriormente, esta mezcla fue incubada con la proteína PvCelTOS recombinante y el porcentaje de inhibición de la actividad de unión de PvCelTOS se cuantificó mediante el análisis de 100.000 eventos usando un citómetro FACS Canto II (Biosciences) y el software FACSDiva (BD).

Figura 16. Metodología del Capítulo 4



Resultados y discusión

El análisis de las secuencias de los loci *pvgama*, *pvrbsa* y *pvceltos* a partir de 5 cepas de *P. vivax*, mostró que estos son genes con una limitada diversidad genética (Tabla 6). Estos datos fueron concordantes cuando se analizó un mayor número de secuencias (Tabla 6 y Publicaciones 8 y 9).

Tabla 7. Estimadores de diversidad

	Datos Cepas				Datos con secuencias disponibles			
Gen	**n**	**Sitios**	**Ss**	**π**	**n**	**Sitios**	**Ss**	**π**
pvgama	6	2013	5	0,0006 (0,0004)	75	2067	25	0,0009 (0,0001)
pvrbsa	5	1290	22	0,0073 (0,0022)	232	1354	47	0,0085 (0,0002)
pvceltos	5	588	3	0,0024 (0,0010)	209	546	7	0,0018 (0,0001)

n: número de secuencia utilizadas. Sitios: número total de sitios analizados. Ss: número de sitios segregrantes. π: diversidad nucleotídica por sitio.

De acuerdo con los resultados de la ventana deslizante de las tasas ω (Figuras 17 y 18), se seleccionaron al menos 2 regiones por cada gen: las primeras (*pvgama* CR1, CR2 y *pvrbsa*-A) en donde se observó una baja diversidad entre especies y señales de selección negativa (ω < 1 y codones bajo selección negativa). Las segundas regiones (*pvgama* VR1, VR2 y *pvrbsa*-B) fueron regiones donde se observó un valor de ω > 1, algunos sitios bajo selección positiva y fueron variables entre especies de *Plasmodium*. Estos fragmentos fueron expresados de forma recombinante en un sistema procariótico y evaluados por citometría de flujo para determinar si estas regiones permiten la interacción entre el patógeno y el hospedero.

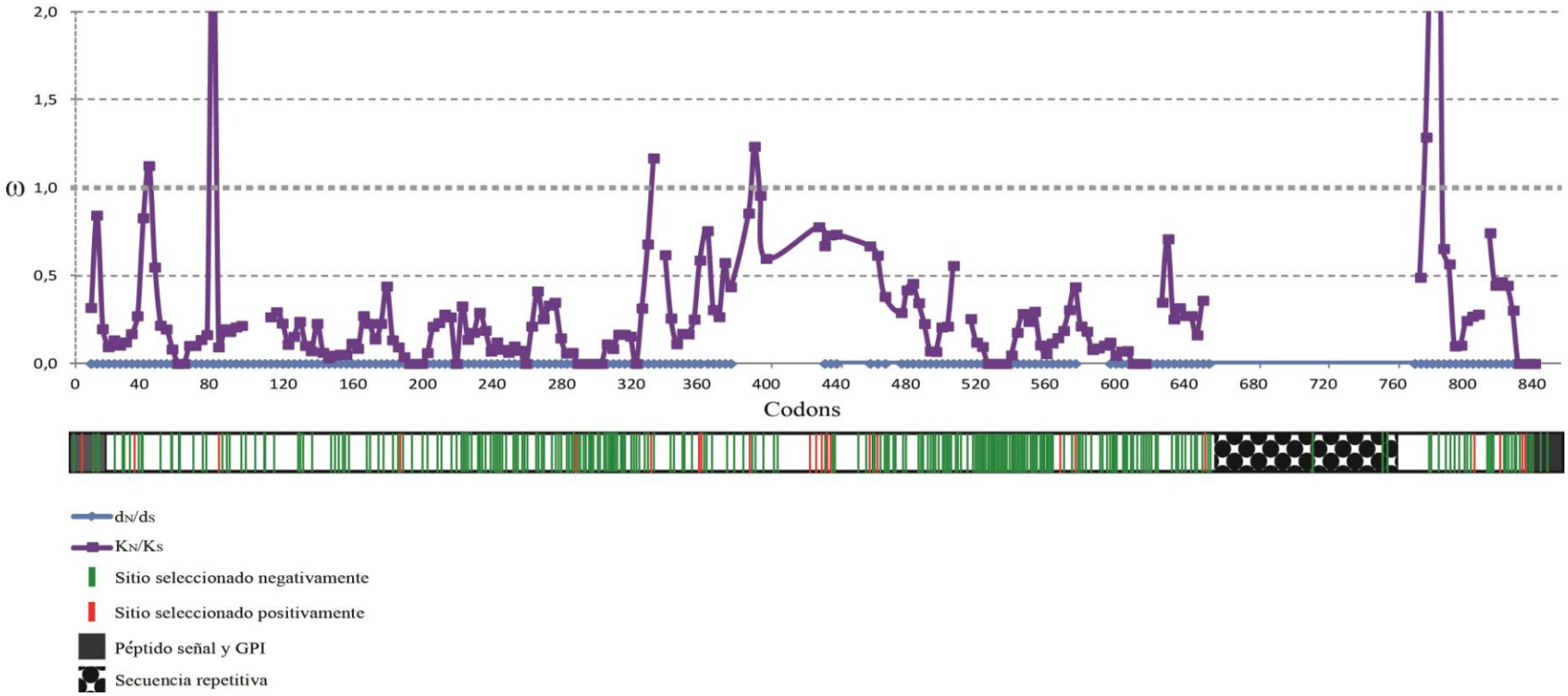
Tanto para PvGAMA como para PvRBSA, las regiones de estas proteínas que mostraron ser conservadas entre especies y con señales de selección negativa (PvGAMA CR1, CR2 y PvRBSA-A) mostraron una unión a los reticulocitos humanos que fue estadísticamente significativa con respecto al control negativo (región DBP-III/IV, que no interactúa con estas células) y a las regiones sin señales de selección negativa (PvGAMA VR2 y PvRBSA-B) (Figura 19). Esto sugiere que las regiones conservadas entre especies y con

señales de selección negativa de PvGAMA y PvRBSA son regiones que permiten la interacción entre el parásito y los reticulocitos humanos, por lo que podrían ser consideradas durante el diseño de una vacuna contra *P. vivax*.

Por otra parte, para PvCelTOS, debido a su corta longitud, se sintetizaron péptidos no sobrepuestos de 20 aminoácidos, los cuales fueron utilizados para ensayos de inhibición de la unión. Los resultados de estos ensayos (Figura 20), muestran que los péptidos que inhiben la unión de PvCelTOS a las células diana, concuerdan con las regiones donde se observó un $\omega < 1$ y donde codones bajo selección negativa fueron localizados (Figura 21 y Publicación 8). Un ω mayor a 1 fue observado en la región C-terminal de la proteína, cerca de la región que está inhibiendo la unión de PvCelTOS a la célula diana. En *P. vivax*, esta región es predicha por estar expuesta al sistema inmune del hospedero (Publicación 8 y (146)) y estudios previos han mostrado que esta región es antigénica en infecciones naturales en *P. vivax* (146). Por lo tanto, esta proteína es un potencial candidato a vacuna, debido a que es reconocida por el sistema inmune del hospedero y a que es la región funcional de la proteína.

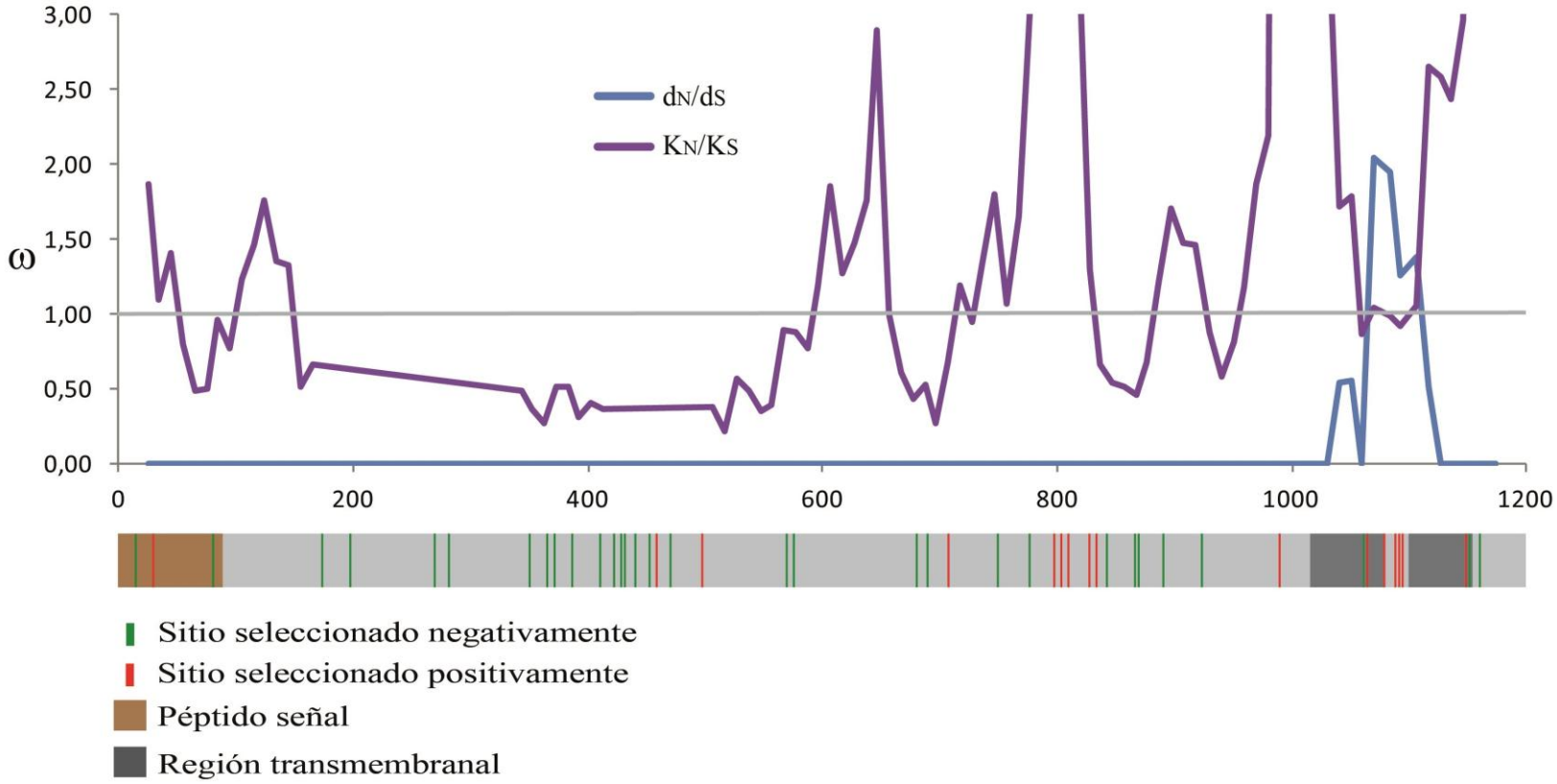
Todos estos resultados muestran que el enfoque propuesto en el Capítulo 1 puede ser utilizado para tener una aproximación de la diversidad genética de un determinado antígeno pero, además, también permite establecer las posibles regiones involucradas en la interacción patógeno-hospedero. Otros resultados obtenidos por diferentes grupos soportan el uso de este enfoque. PvGAMA fue descrito en 2016 por un grupo coreano (147). La proteína fue dividida en 7 fragmentos diferentes y cada fragmento fue evaluado con respecto a su habilidad de unión. Los fragmentos que mostraron unión (147), coinciden con el fragmento identificado en el presente trabajo como conservado entre especies, con un $\omega < 1$ y varios codones seleccionados negativamente.

Figura 17. Ventana deslizante para el locus *pvgama*



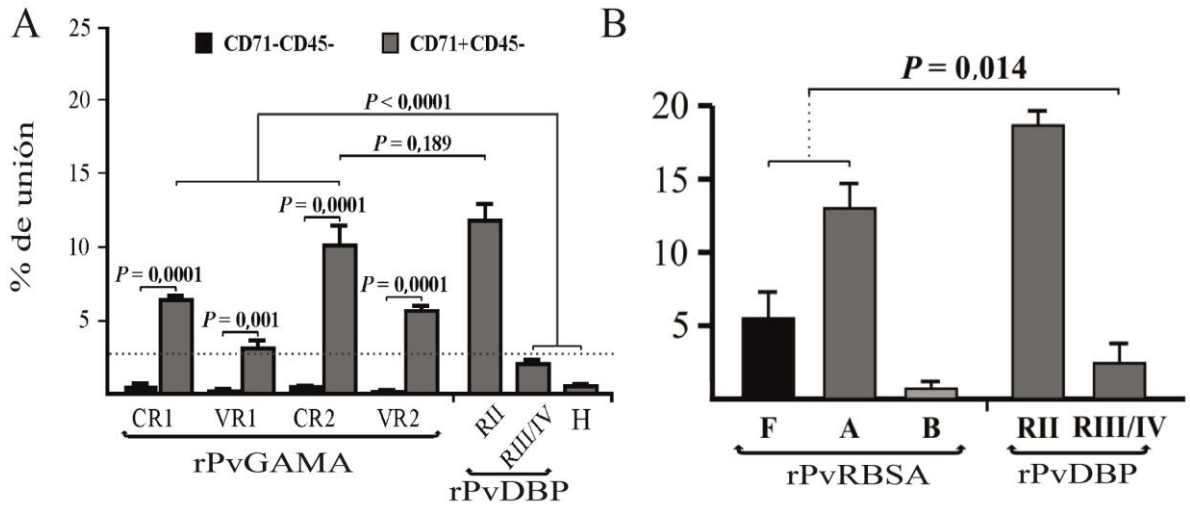
Los valores ω de gama (d_N/d_S) dentro de P. vivax se muestran en azul, mientras la tasa de divergencia ω (K_N / K_S) entre las especies que infectan primates y P. vivax se muestran en morado. Una representación del gen gama se presenta debajo de la ventana deslizante que indica el péptido señal, la región repetida y el anclaje a GPI. Los sitios bajo selección negativa están representados por líneas verdes y los sitios seleccionados positivamente entre especies se muestran por líneas rojas.

Figura 18. Ventana deslizante para el locus *pvrbsa*



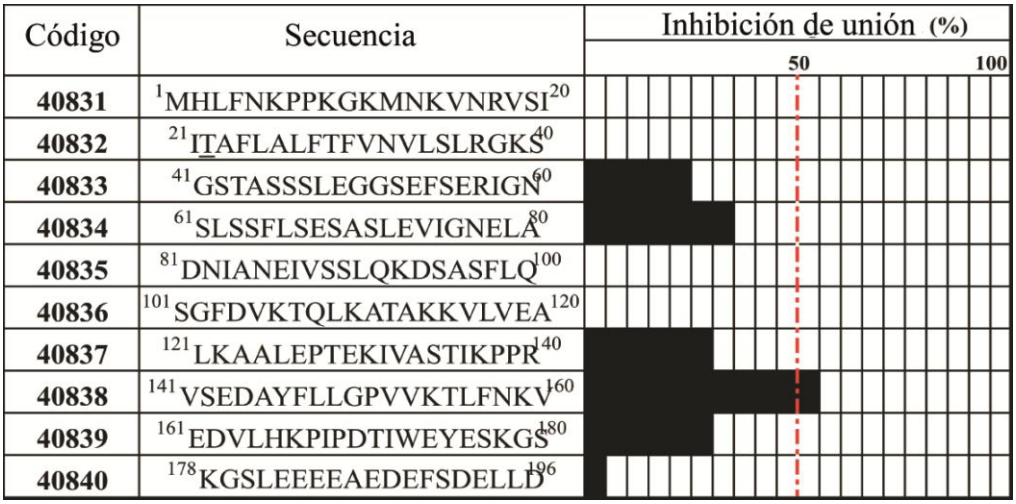
Los valores ω de rbsa (d_N/d_S) dentro de P. vivax se muestran en azul, mientras la tasa de divergencia ω (K_N / K_S) entre las especies que infectan primates y P. vivax se muestran en morado. Una representación del gen rbsa se presenta debajo de la ventana deslizante que indica el péptido señal y las hélices transmembranales. Los sitios bajo selección negativa están representados por líneas verdes y los sitios seleccionados positivamente entre especies se muestran por líneas rojas.

Figura 19. Actividad de unión de fragmentos de PvGAMA y PvRBSA



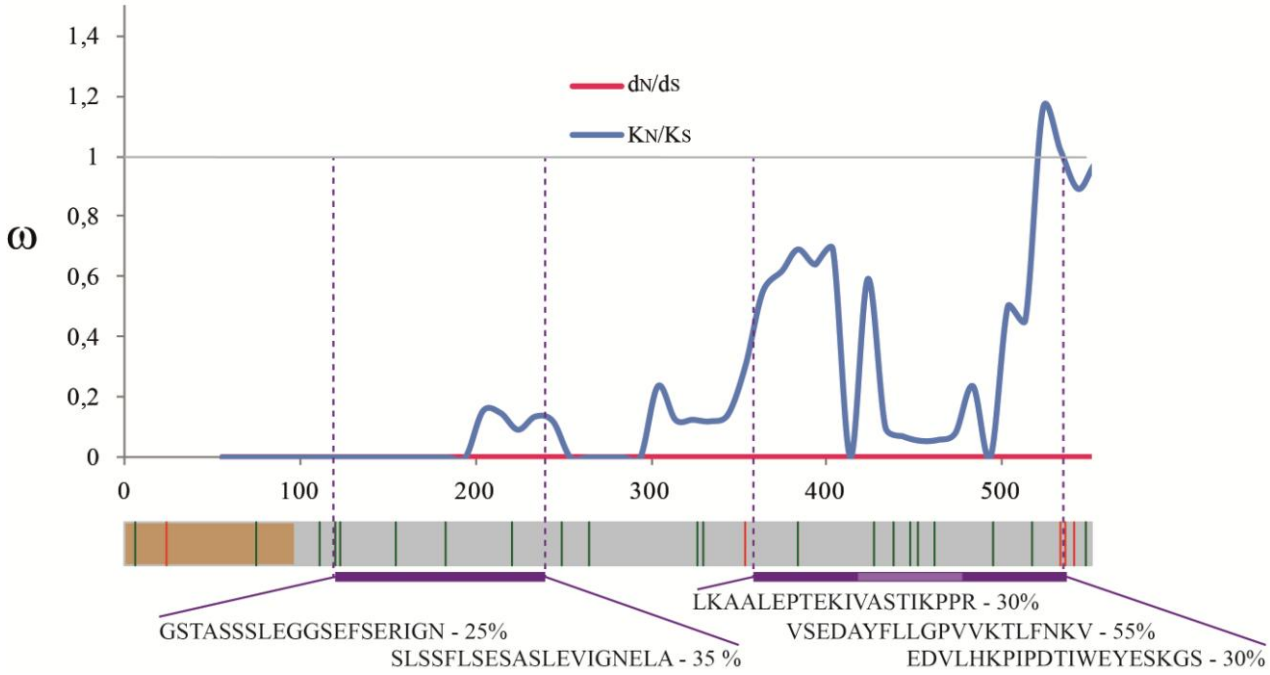
Análisis de citometría de flujo mostrando el porcentaje de unión de los fragmentos recombinantes de las proteínas PvGAMA (A) y PvRBSA (B) a reticulocitos humanos. Como control de unión (positivo), se utilizó la región II de DBP que previamente se ha reportado se une a los reticulocitos, y como control negativo, la región III y IV (37). Células CD71-CD45-: Normocitos; CD71+ CD45-: Reticulocitos. Panel A: CR1 y CR2 son las regiones con patrón de restricción funcional en PvGAMA. VR1 y VR2 son regiones variables de PvGAMA. H: control de histidinas. Panel B: F indica la proteína PvRBSA completa, A, la región identificada bajo restricción funcional y B, la región variable.

Figura 20. Ensayos de inhibición de la unión con péptidos de PvCelTOS



Secuencias de 20 aminoácidos (péptidos) correspondientes a la proteína PvCelTOS. Se observa el porcentaje de inhibición de la unión de la proteína completa que cada péptido presentó.

Figura 21. Ventana deslizante para el locus *pvceltos*



Los valores de ω (d_N/d_S) para pvceltos se muestran en magenta, mientras la divergencia ω (K_N/K_S) entre P. vivax y P. cynomolgi es mostrada en azul. Un diagrama del gen se presenta debajo de la ventana deslizante. La región que codifica el péptido señal se indica en café y los sitios bajo selección positiva y negativa se indican en rojo y verde, respectivamente. Las secuencias de los péptidos que mostraron una inhibición de la unión de la proteína recombinante son mostradas en morado. Estos péptidos coinciden con las regiones donde se observaron sitios bajo selección negativa.

Recientemente, ensayos funcionales para PvRON4 fueron realizados (148), identificando que la región C-terminal de esta proteína interactúa con los reticulocitos humanos. Estos resultados (148) concuerdan con los obtenidos en este trabajo (Anexo 4 y Publicación 6) que sugirió que la región C-terminal de PvRON4 está restringida funcional/estructuralmente (región conservada entre especies, con un ω < 1 y codones bajo selección negativa).

Conclusiones

De acuerdo a la teoría neutral de evolución molecular (39), dentro de una proteína, las regiones funcionales suelen evolucionar más lentamente y, por lo tanto, son conservadas. Esa conservación suele ser mantenida por la selección negativa, debido a que cualquier

alteración en la secuencia de aminoácidos podría afectar la función de la proteína. Por lo tanto, la predicción de regiones bajo restricción funcional podría ser utilizada como una alternativa para determinar las regiones involucradas en la interacción patógeno-hospedero. Así, en vez de analizar múltiples y pequeños fragmentos de la proteína, este enfoque podría utilizarse para ubicar regiones funcionales potenciales en *P. vivax* y así realizar ensayos de unión sólo con aquellas regiones con señales de selección negativa, optimizando los reticulocitos, que son la limitante para realizar estos estudios. Adicionalmente, al ser estas regiones conservadas, se podrían usar durante el desarrollo de una vacuna antimalárica para así evitar las respuestas alelo-específicas de disminuyen la eficacia de las vacunas.

Los resultados de este capítulo fueron publicados y están disponibles bajo las siguientes referencias:

Publicación 7: Baquero LA, Moreno-Pérez DA, **Garzón-Ospina D**, Forero-Rodríguez J, Ortiz-Suárez HD, Patarroyo MA. *PvGAMA reticulocyte binding activity: predicting conserved functional regions by natural selection analysis*. Parasit Vectors. 2017 May 19;10(1):251. doi: 10.1186/s13071-017-2183-8.

Publicación 8: **Garzón-Ospina D**, Buitrago SP, Ramos AE, Patarroyo MA. *Identifying Potential Plasmodium vivax Sporozoite Stage Vaccine Candidates: An Analysis of Genetic Diversity and Natural Selection*. Front Genet. 2018 Jan 25;9:10. doi:10.3389/fgene.2018.00010.

Publicación 9: Camargo-Ayala PA, **Garzón-Ospina D**, Moreno-Pérez DA, Ricaurte-Contreras LA, Noya O and Patarroyo MA. *On the evolution and function of Plasmodium vivax reticulocyte binding surface antigen (pvrbsa)*. Front. Genet. 2018 Sep . 9:372 doi: 10.3389/fgene.2018.00372.

CONCLUSIONES GENERALES

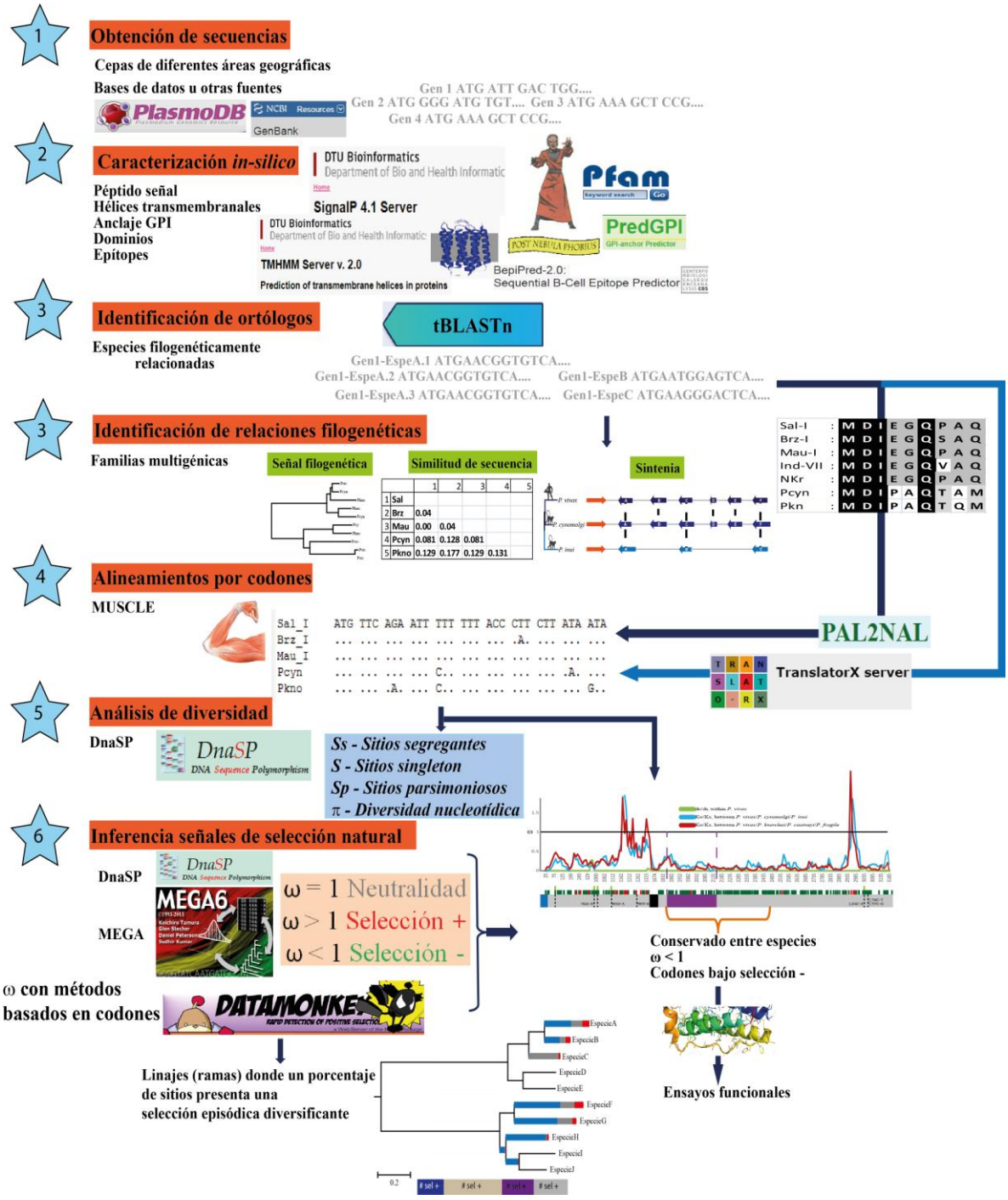
Este trabajo presentó y validó el potencial de un enfoque alternativo para la selección de antígenos promisorios para el diseño de una vacuna altamente efectiva contra *P. vivax*. A partir de los datos (secuencias genómicas) de 5 cepas de *P. vivax*, fue posible hacer una aproximación de la diversidad genética de varios antígenos parasitarios. Estos datos son similares a los obtenidos si se analizan un mayor número de secuencias. Adicionalmente, a partir de estos datos, fue posible determinar regiones que presentan un patrón consistente con restricciones funcionales (región conservada entre especies, con un $\omega < 1$ y codones bajo selección negativa). Al analizar aquellas regiones con un patrón de restricción funcional mediante ensayos de unión, se pudo establecer que estas regiones juegan un papel durante la interacción patógeno-hospedero, lo que soporta la utilización de este enfoque durante la selección de candidatos.

Aunque este enfoque presentó buenos resultados, también tiene limitaciones. La presencia de señales de selección positiva linaje-específica, puede incrementar la tasa ω, lo que descartaría el antígeno (u una región particular de éste) como candidato promisorio. No obstante, la identificación de este tipo de señales, junto con la identificación de codones bajo selección negativa, podría ser utilizada para hacer frente a esta limitación. Teniendo en cuenta toda la información presentada en este trabajo, junto con las herramientas bioinformáticas disponibles, se diseñó la Figura 22, que muestra una alternativa a seguir para la selección de antígenos promisorios a vacuna.

Este trabajo evaluó la diversidad genética e infirió señales consistentes con selección natural de diferentes antígenos de *P. vivax*. De acuerdo con los resultados obtenidos, los antígenos (o regiones bajo restricción funcional dentro de estos): *pvclag*, *pvser*/*thr*, *pvrhopH1*/*clag7*, *pvron2*, PVX_092425 (Publicación 1), *pv12*, *pv38* (Publicación 1 y 3), *pv41* (Publicación 1 y 4), *pvron4* (Publicación 1 y 6), *pvgama* (Publicación 7), *pvrbsa* (Publicación 9) y *pvceltos* (así como los antígenos del estadio esporozoíto *pvp52*, *pvp36*,

pvspatr, *pvplp1*, *pvmcp1*, *pvtlp*, y *pvmb2*, Publicación 8) son antígenos promisorios a evaluar durante el diseño de una vacuna contra *P. vivax* altamente efectiva.

Figura 22. Flujo de trabajo a seguir para seleccionar antígenos promisorios a incluir en el diseño de vacunas contra agentes infecciosos



PERSPECTIVAS GENERALES

En este trabajo de describió y valido un enfoque alternativo para la selección de antígenos promisorios a ser considerados durante el diseño de una vacuna contra *P. vivax* que evite las respuestas inmunes alelo-específicas. Cerca de 20 antígenos fueron sugeridos como promisorios. Sin embargo, debido a los alcances de este trabajo, sólo las regiones predichas como funcionalmente importantes de 3 de estos fueron corroboradas experimentalmente. Por lo tanto, futuros estudios funcionales en los restantes 16 antígenos, así como, de otros antígenos no incluidos en este trabajo, podrían ser llevados a cabo tomando como base los resultados obtenidos y así determinar el rol de éstos durante el proceso de invasión de *P. vivax* a los reticulocitos humanos.

Por otra parte, estudios de antigenicidad, inmunogenicidad y protección, podrían ser llevados a cabo con los antígenos acá descritos como promisorios (o regiones restringidas funcionalmente dentro de ellos), para así evaluar si estos pueden ser considerados candidatos a vacuna y ser incluidos en ensayos preclínicos de una vacuna contra *P. vivax*.

Finalmente, el enfoque presentado en este trabajo podría ser aplicado a otros patógenos, durante la búsqueda de potenciales candidatos a vacuna.

REFERENCIAS

1. de Meeus T, McCoy KD, Prugnolle F, Chevillon C, Durand P, Hurtrez-Bousses S, et al. Population genetics and molecular epidemiology or how to "debusquer la bete". Infect Genet Evol. 2007 Mar;7(2):308-32.
2. Hotez P, Herricks J. One Million Deaths by Parasites. USA. http://blogs.plos.org/speakingofmedicine/2015/01/16/one-million-deaths-parasites/: PLOS Medicine Pathogens Neglected Tropical Diseases; 2015 [cited 2015 July 13 2015].
3. Global, regional, and national age-sex specific all-cause and cause-specific mortality for 240 causes of death, 1990-2013: a systematic analysis for the Global Burden of Disease Study 2013. Lancet. 2015 Jan 10;385(9963):117-71.
4. Escalante AA, Ayala FJ. Phylogeny of the malarial genus Plasmodium, derived from rRNA gene sequences. Proceedings of the National Academy of Sciences of the United States of America. 1994 Nov 22;91(24):11373-7.
5. Rich SM, Ayala FJ. Progress in malaria research: the case for phylogenetics. Advances in parasitology. 2003;54:255-80.
6. Schaer J, Perkins SL, Decher J, Leendertz FH, Fahr J, Weber N, et al. High diversity of West African bat malaria parasites and a tight link with rodent Plasmodium taxa. Proceedings of the National Academy of Sciences of the United States of America. 2013 Oct 22;110(43):17415-9.
7. Cowman AF, Crabb BS. Invasion of red blood cells by malaria parasites. Cell. 2006 Feb 24;124(4):755-66.
8. Galinski MR, Meyer EV, Barnwell JW. Plasmodium vivax: modern strategies to study a persistent parasite's life cycle. Advances in parasitology. 2013;81:1-26.
9. Coatney GR, National Institute of Allergy and Infectious Diseases (U.S.). The primate malarias. Bethesda, Md.,: U.S. National Institute of Allergy and Infectious Diseases; for sale by the Supt. of Docs., U.S. Govt. Print. Off., Washington; 1971.
10. Pain A, Bohme U, Berry AE, Mungall K, Finn RD, Jackson AP, et al. The genome of the simian and human malaria parasite Plasmodium knowlesi. Nature. 2008 Oct 9;455(7214):799-803.
11. Pacheco MA, Battistuzzi FU, Junge RE, Cornejo OE, Williams CV, Landau I, et al. Timing the origin of human malarias: the lemur puzzle. BMC Evol Biol. 2011;11:299.
12. Hayakawa T, Culleton R, Otani H, Horii T, Tanabe K. Big bang in the evolution of extant malaria parasites. Mol Biol Evol. 2008 Oct;25(10):2233-9.
13. Silva JC, Egan A, Arze C, Spouge JL, Harris DG. A new method for estimating species age supports the coexistence of malaria parasites and their Mammalian hosts. Mol Biol Evol. 2015 May;32(5):1354-64.
14. Liu W, Li Y, Shaw KS, Learn GH, Plenderleith LJ, Malenke JA, et al. African origin of the malaria parasite Plasmodium vivax. Nat Commun. 2014;5:3346.
15. Mu J, Joy DA, Duan J, Huang Y, Carlton J, Walker J, et al. Host switch leads to emergence of Plasmodium vivax malaria in humans. Mol Biol Evol. 2005 Aug;22(8):1686-93.
16. WHO. World malaria report 2017: Geneva: World Health Organization; 2017. Licence: CC BY-NC-SA 3.0 IGO; 2017 [cited 2018 05/04/2018]. Available from:

http://apps.who.int/iris/bitstream/handle/10665/259492/9789241565523-eng.pdf?sequence=1.
17. Suh KN, Kain KC, Keystone JS. Malaria. CMAJ. 2004 May 25;170(11):1693-702.
18. Maxmen A. Malaria surge feared. Nature. 2012 May 15;485(7398):293.
19. Huijben S, Paaijmans KP. Putting evolution in elimination: Winning our ongoing battle with evolving malaria mosquitoes and parasites. Evol Appl. 2018 Apr;11(4):415-30.
20. Price RN, von Seidlein L, Valecha N, Nosten F, Baird JK, White NJ. Global extent of chloroquine-resistant Plasmodium vivax: a systematic review and meta-analysis. Lancet Infect Dis. 2014 Oct;14(10):982-91.
21. Barry AE, Arnott A. Strategies for designing and monitoring malaria vaccines targeting diverse antigens. Front Immunol. 2014;5:359.
22. White NJ, Pukrittayakamee S, Hien TT, Faiz MA, Mokuolu OA, Dondorp AM. Malaria. Lancet. 2014 Feb 22;383(9918):723-35.
23. Garrido-Cardenas JA, Mesa-Valle C, Manzano-Agugliaro F. Genetic approach towards a vaccine against malaria. Eur J Clin Microbiol Infect Dis. 2018 Jun 28.
24. Patarroyo MA, Calderon D, Moreno-Perez DA. Vaccines against Plasmodium vivax: a research challenge. Expert Rev Vaccines. 2012 Oct;11(10):1249-60.
25. Luo Z, Sullivan SA, Carlton JM. The biology of Plasmodium vivax explored through genomics. Ann N Y Acad Sci. 2015 Apr;1342:53-61.
26. Arevalo-Pinzon G, Curtidor H, Abril J, Patarroyo MA. Annotation and characterization of the Plasmodium vivax rhoptry neck protein 4 (PvRON4). Malar J. 2013 Oct 5;12:356.
27. Arevalo-Pinzon G, Curtidor H, Patino LC, Patarroyo MA. PvRON2, a new Plasmodium vivax rhoptry neck antigen. Malar J. 2011;10:60.
28. Moreno-Perez DA, Saldarriaga A, Patarroyo MA. Characterizing PvARP, a novel Plasmodium vivax antigen. Malar J. 2013;12:165.
29. Neafsey DE, Galinsky K, Jiang RH, Young L, Sykes SM, Saif S, et al. The malaria parasite Plasmodium vivax exhibits greater genetic diversity than Plasmodium falciparum. Nat Genet. 2012 Sep;44(9):1046-50.
30. Ellis RD, Sagara I, Doumbo O, Wu Y. Blood stage vaccines for Plasmodium falciparum: current status and the way forward. Hum Vaccin. 2010 Aug;6(8):627-34.
31. Fluck C, Smith T, Beck HP, Irion A, Betuela I, Alpers MP, et al. Strain-specific humoral response to a polymorphic malaria vaccine. Infect Immun. 2004 Nov;72(11):6300-5.
32. Genton B, Reed ZH. Asexual blood-stage malaria vaccine development: facing the challenges. Curr Opin Infect Dis. 2007 Oct;20(5):467-75.
33. Ouattara A, Takala-Harrison S, Thera MA, Coulibaly D, Niangaly A, Saye R, et al. Molecular basis of allele-specific efficacy of a blood-stage malaria vaccine: vaccine development implications. J Infect Dis. 2013 Feb 1;207(3):511-9.
34. Takala SL, Plowe CV. Genetic diversity and malaria vaccine design, testing and efficacy: preventing and overcoming 'vaccine resistant malaria'. Parasite Immunol. 2009 Sep;31(9):560-73.
35. Richie TL, Saul A. Progress and challenges for malaria vaccines. Nature. 2002 Feb 7;415(6872):694-701.

36. Urquiza M, Patarroyo MA, Mari V, Ocampo M, Suarez J, Lopez R, et al. Identification and polymorphism of Plasmodium vivax RBP-1 peptides which bind specifically to reticulocytes. Peptides. 2002 Dec;23(12):2265-77.
37. Ocampo M, Vera R, Eduardo Rodriguez L, Curtidor H, Urquiza M, Suarez J, et al. Plasmodium vivax Duffy binding protein peptides specifically bind to reticulocytes. Peptides. 2002 Jan;23(1):13-22.
38. Rodriguez LE, Urquiza M, Ocampo M, Curtidor H, Suarez J, Garcia J, et al. Plasmodium vivax MSP-1 peptides have high specific binding activity to human reticulocytes. Vaccine. 2002 Jan 31;20(9-10):1331-9.
39. Kimura M. The neutral theory of molecular evolution. Cambridge Cambridgeshire ; New York: Cambridge University Press; 1983.
40. Graur D, Zheng Y, Price N, Azevedo RB, Zufall RA, Elhaik E. On the immortality of television sets: "function" in the human genome according to the evolution-free gospel of ENCODE. Genome Biol Evol. 2013;5(3):578-90.
41. Carlton JM, Adams JH, Silva JC, Bidwell SL, Lorenzi H, Caler E, et al. Comparative genomics of the neglected human malaria parasite Plasmodium vivax. Nature. 2008 Oct 9;455(7214):757-63.
42. Bozdech Z, Mok S, Hu G, Imwong M, Jaidee A, Russell B, et al. The transcriptome of Plasmodium vivax reveals divergence and diversity of transcriptional regulation in malaria parasites. Proceedings of the National Academy of Sciences of the United States of America. 2008 Oct 21;105(42):16290-5.
43. Chen JH, Jung JW, Wang Y, Ha KS, Lu F, Lim CS, et al. Immunoproteomics profiling of blood stage *Plasmodium vivax* infection by high-throughput screening assays. Journal of proteome research. 2011 Dec 3;9(12):6479-89.
44. Moreno-Perez DA, Degano R, Ibarrola N, Muro A, Patarroyo MA. Determining the *Plasmodium vivax* VCG-1 strain blood stage proteome. Journal of proteomics. 2014 Oct 11.
45. Tachibana S, Sullivan SA, Kawai S, Nakamura S, Kim HR, Goto N, et al. Plasmodium cynomolgi genome sequences provide insight into Plasmodium vivax and the monkey malaria clade. Nat Genet. 2012 Sep;44(9):1051-5.
46. Good MF. Towards a blood-stage vaccine for malaria: are we following all the leads? Nature reviews. 2001 Nov;1(2):117-25.
47. O'Donnell RA, de Koning-Ward TF, Burt RA, Bockarie M, Reeder JC, Cowman AF, et al. Antibodies against merozoite surface protein (MSP)-1(19) are a major component of the invasion-inhibitory response in individuals immune to malaria. The Journal of experimental medicine. 2001 Jun 18;193(12):1403-12.
48. Rojas-Caraballo J, Mongui A, Giraldo MA, Delgado G, Granados D, Millan-Cortes D, et al. Immunogenicity and protection-inducing ability of recombinant Plasmodium vivax rhoptry-associated protein 2 in Aotus monkeys: a potential vaccine candidate. Vaccine. 2009 May 11;27(21):2870-6.
49. Farooq F, Bergmann-Leitner ES. Immune Escape Mechanisms are Plasmodium's Secret Weapons Foiling the Success of Potent and Persistently Efficacious Malaria Vaccines. Clin Immunol. 2015 Sep 2.
50. Angel DI, Mongui A, Ardila J, Vanegas M, Patarroyo MA. The Plasmodium vivax Pv41 surface protein: identification and characterization. Biochem Biophys Res Commun. 2008 Dec 26;377(4):1113-7.

51. Mongui A, Angel DI, Gallego G, Reyes C, Martinez P, Guhl F, et al. Characterization and antigenicity of the promising vaccine candidate Plasmodium vivax 34kDa rhoptry antigen (Pv34). Vaccine. 2009 Dec 11;28(2):415-21.
52. Mongui A, Angel DI, Moreno-Perez DA, Villarreal-Gonzalez S, Almonacid H, Vanegas M, et al. Identification and characterization of the Plasmodium vivax thrombospondin-related apical merozoite protein. Malar J. 2010;9:283.
53. Mongui A, Perez-Leal O, Rojas-Caraballo J, Angel DI, Cortes J, Patarroyo MA. Identifying and characterising the Plasmodium falciparum RhopH3 Plasmodium vivax homologue. Biochem Biophys Res Commun. 2007 Jul 6;358(3):861-6.
54. Mongui A, Perez-Leal O, Soto SC, Cortes J, Patarroyo MA. Cloning, expression, and characterisation of a Plasmodium vivax MSP7 family merozoite surface protein. Biochem Biophys Res Commun. 2006 Dec 22;351(3):639-44.
55. Moreno-Perez DA, Mongui A, Soler LN, Sanchez-Ladino M, Patarroyo MA. Identifying and characterizing a member of the RhopH1/Clag family in Plasmodium vivax. Gene. 2011 Jul 15;481(1):17-23.
56. Perez-Leal O, Mongui A, Cortes J, Yepes G, Leiton J, Patarroyo MA. The Plasmodium vivax rhoptry-associated protein 1. Biochem Biophys Res Commun. 2006 Mar 24;341(4):1053-8.
57. Perez-Leal O, Sierra AY, Barrero CA, Moncada C, Martinez P, Cortes J, et al. Identifying and characterising the Plasmodium falciparum merozoite surface protein 10 Plasmodium vivax homologue. Biochem Biophys Res Commun. 2005 Jun 17;331(4):1178-84.
58. Perez-Leal O, Sierra AY, Barrero CA, Moncada C, Martinez P, Cortes J, et al. Plasmodium vivax merozoite surface protein 8 cloning, expression, and characterisation. Biochem Biophys Res Commun. 2004 Nov 26;324(4):1393-9.
59. Zambrano-Villa S, Rosales-Borjas D, Carrero JC, Ortiz-Ortiz L. How protozoan parasites evade the immune response. Trends Parasitol. 2002 Jun;18(6):272-8.
60. Thera MA, Doumbo OK, Coulibaly D, Laurens MB, Ouattara A, Kone AK, et al. A field trial to assess a blood-stage malaria vaccine. N Engl J Med. 2011 Sep 15;365(11):1004-13.
61. Figtree M, Pasay CJ, Slade R, Cheng Q, Cloonan N, Walker J, et al. Plasmodium vivax synonymous substitution frequencies, evolution and population structure deduced from diversity in AMA 1 and MSP 1 genes. Molecular and biochemical parasitology. 2000 Apr 30;108(1):53-66.
62. Garzon-Ospina D, Romero-Murillo L, Patarroyo MA. Limited genetic polymorphism of the Plasmodium vivax low molecular weight rhoptry protein complex in the Colombian population. Infect Genet Evol. 2010 Mar;10(2):261-7.
63. Garzon-Ospina D, Romero-Murillo L, Tobon LF, Patarroyo MA. Low genetic polymorphism of merozoite surface proteins 7 and 10 in Colombian Plasmodium vivax isolates. Infect Genet Evol. 2011 Mar;11(2):528-31.
64. Gomez A, Suarez CF, Martinez P, Saravia C, Patarroyo MA. High polymorphism in Plasmodium vivax merozoite surface protein-5 (MSP5). Parasitology. 2006 Dec;133(Pt 6):661-72.
65. Martinez P, Suarez CF, Cardenas PP, Patarroyo MA. Plasmodium vivax Duffy binding protein: a modular evolutionary proposal. Parasitology. 2004 Apr;128(Pt 4):353-66.

66. Martinez P, Suarez CF, Gomez A, Cardenas PP, Guerrero JE, Patarroyo MA. High level of conservation in Plasmodium vivax merozoite surface protein 4 (PvMSP4). Infect Genet Evol. 2005 Oct;5(4):354-61.
67. Mascorro CN, Zhao K, Khuntirat B, Sattabongkot J, Yan G, Escalante AA, et al. Molecular evolution and intragenic recombination of the merozoite surface protein MSP-3alpha from the malaria parasite Plasmodium vivax in Thailand. Parasitology. 2005 Jul;131(Pt 1):25-35.
68. Pacheco MA, Elango AP, Rahman AA, Fisher D, Collins WE, Barnwell JW, et al. Evidence of purifying selection on merozoite surface protein 8 (MSP8) and 10 (MSP10) in Plasmodium spp. Infect Genet Evol. 2012 Jul;12(5):978-86.
69. Putaporntip C, Jongwutiwes S, Ferreira MU, Kanbara H, Udomsangpetch R, Cui L. Limited global diversity of the Plasmodium vivax merozoite surface protein 4 gene. Infect Genet Evol. 2009 Sep;9(5):821-6.
70. Putaporntip C, Jongwutiwes S, Seethamchai S, Kanbara H, Tanabe K. Intragenic recombination in the 3' portion of the merozoite surface protein 1 gene of Plasmodium vivax. Molecular and biochemical parasitology. 2000 Jul;109(2):111-9.
71. Putaporntip C, Udomsangpetch R, Pattanawong U, Cui L, Jongwutiwes S. Genetic diversity of the Plasmodium vivax merozoite surface protein-5 locus from diverse geographic origins. Gene. 2010 May 15;456(1-2):24-35.
72. Tetteh KK, Stewart LB, Ochola LI, Amambua-Ngwa A, Thomas AW, Marsh K, et al. Prospective identification of malaria parasite genes under balancing selection. PLoS One. 2009;4(5):e5568.
73. Weedall GD, Conway DJ. Detecting signatures of balancing selection to identify targets of anti-parasite immunity. Trends Parasitol. 2010 Jul;26(7):363-9.
74. Suzuki Y. Natural selection on the influenza virus genome. Mol Biol Evol. 2006 Oct;23(10):1902-11.
75. Mazumder R, Hu ZZ, Vinayaka CR, Sagripanti JL, Frost SD, Kosakovsky Pond SL, et al. Computational analysis and identification of amino acid sites in dengue E proteins relevant to development of diagnostics and vaccines. Virus Genes. 2007 Oct;35(2):175-86.
76. Suzuki Y. Negative selection on neutralization epitopes of poliovirus surface proteins: implications for prediction of candidate epitopes for immunization. Gene. 2004 Mar 17;328:127-33.
77. Rich SM, Licht MC, Hudson RR, Ayala FJ. Malaria's Eve: evidence of a recent population bottleneck throughout the world populations of Plasmodium falciparum. Proceedings of the National Academy of Sciences of the United States of America. 1998 Apr 14;95(8):4425-30.
78. Hughes AL, Verra F. Ancient polymorphism and the hypothesis of a recent bottleneck in the malaria parasite Plasmodium falciparum. Genetics. 1998 Sep;150(1):511-3.
79. Griffing SM, Viana GM, Mixson-Hayden T, Sridaran S, Alam MT, de Oliveira AM, et al. Historical shifts in Brazilian P. falciparum population structure and drug resistance alleles. PLoS One. 2013;8(3):e58984.
80. McCollum AM, Mueller K, Villegas L, Udhayakumar V, Escalante AA. Common origin and fixation of Plasmodium falciparum dhfr and dhps mutations associated with sulfadoxine-pyrimethamine resistance in a low-transmission area in South America. Antimicrob Agents Chemother. 2007 Jun;51(6):2085-91.

81. Rodrigues PT, Valdivia HO, de Oliveira TC, Alves JMP, Duarte A, Cerutti-Junior C, et al. Human migration and the spread of malaria parasites to the New World. Sci Rep. 2018 Jan 31;8(1):1993.
82. Restrepo-Montoya D, Becerra D, Carvajal-Patino JG, Mongui A, Nino LF, Patarroyo ME, et al. Identification of Plasmodium vivax proteins with potential role in invasion using sequence redundancy reduction and profile hidden Markov models. PLoS One. 2011;6(10):e25189.
83. Arnott A, Barry AE, Reeder JC. Understanding the population genetics of Plasmodium vivax is essential for malaria control and elimination. Malar J. 2012;11:14.
84. Cornejo OE, Fisher D, Escalante AA. Genome-Wide Patterns of Genetic Polymorphism and Signatures of Selection in Plasmodium vivax. Genome Biol Evol. 2014;7(1):106-19.
85. Ochola LI, Tetteh KK, Stewart LB, Riitho V, Marsh K, Conway DJ. Allele frequency-based and polymorphism-versus-divergence indices of balancing selection in a new filtered set of polymorphic genes in Plasmodium falciparum. Mol Biol Evol. 2010 Oct;27(10):2344-51.
86. Mongui A, Angel DI, Guzman C, Vanegas M, Patarroyo MA. Characterisation of the Plasmodium vivax Pv38 antigen. Biochem Biophys Res Commun. 2008 Nov 14;376(2):326-30.
87. Moreno-Perez DA, Areiza-Rojas R, Florez-Buitrago X, Silva Y, Patarroyo ME, Patarroyo MA. The GPI-anchored 6-Cys protein Pv12 is present in detergent-resistant microdomains of Plasmodium vivax blood stage schizonts. Protist. 2013 Jan;164(1):37-48.
88. Moreno-Perez DA, Montenegro M, Patarroyo ME, Patarroyo MA. Identification, characterization and antigenicity of the Plasmodium vivax rhoptry neck protein 1 (PvRON1). Malar J. 2011;10:314.
89. Patarroyo MA, Perez-Leal O, Lopez Y, Cortes J, Rojas-Caraballo J, Gomez A, et al. Identification and characterisation of the Plasmodium vivax rhoptry-associated protein 2. Biochem Biophys Res Commun. 2005 Nov 25;337(3):853-9.
90. Garzon-Ospina D, Lopez C, Forero-Rodriguez J, Patarroyo MA. Genetic diversity and selection in three Plasmodium vivax merozoite surface protein 7 (Pvmsp-7) genes in a Colombian population. PLoS One. 2012;7(9):e45962.
91. Arisue N, Kawai S, Hirai M, Palacpac NM, Jia M, Kaneko A, et al. Clues to evolution of the SERA multigene family in 18 Plasmodium species. PLoS One. 2011;6(3):e17775.
92. Garzon-Ospina D, Cadavid LF, Patarroyo MA. Differential expansion of the merozoite surface protein (msp)-7 gene family in Plasmodium species under a birth-and-death model of evolution. Mol Phylogenet Evol. 2010 May;55(2):399-408.
93. Garzon-Ospina D, Forero-Rodriguez J, Patarroyo MA. Heterogeneous genetic diversity pattern in Plasmodium vivax genes encoding merozoite surface proteins (MSP) -7E, -7F and -7L. Malar J. 2014;13:495.
94. Rice BL, Acosta MM, Pacheco MA, Carlton JM, Barnwell JW, Escalante AA. The origin and diversification of the merozoite surface protein 3 (msp3) multi-gene family in Plasmodium vivax and related parasites. Mol Phylogenet Evol. 2014 Sep;78:172-84.
95. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32(5):1792-7.

96. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol. 2011 Oct;28(10):2731-9.
97. Suyama M, Torrents D, Bork P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W609-12.
98. Librado P, Rozas J. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics. 2009 Jun 1;25(11):1451-2.
99. Nielsen R. Molecular signatures of natural selection. Annu Rev Genet. 2005;39:197-218.
100. McDonald JH, Kreitman M. Adaptive protein evolution at the Adh locus in Drosophila. Nature. 1991 Jun 20;351(6328):652-4.
101. Jukes THaCRC. Evolution of protein molecules. In H. N. Munro, ed., Mammalian Protein Metabolism. New York: Academic Press; 1969.
102. Egea R, Casillas S, Barbadilla A. Standard and generalized McDonald-Kreitman test: a website to detect selection by comparing different classes of DNA sites. Nucleic Acids Res. 2008 Jul 1;36(Web Server issue):W157-62.
103. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. Mol Biol Evol. 2013 Dec;30(12):2725-9.
104. Zhang J, Rosenberg HF, Nei M. Positive Darwinian selection after gene duplication in primate ribonuclease genes. Proceedings of the National Academy of Sciences of the United States of America. 1998 Mar 31;95(7):3708-13.
105. Sawai H, Otani H, Arisue N, Palacpac N, de Oliveira Martins L, Pathirana S, et al. Lineage-specific positive selection at the merozoite surface protein 1 (msp1) locus of Plasmodium vivax and related simian malaria parasites. BMC Evol Biol. 2010;10:52.
106. Tanabe K, Escalante A, Sakihama N, Honda M, Arisue N, Horii T, et al. Recent independent evolution of msp1 polymorphism in Plasmodium vivax and related simian malaria parasites. Molecular and biochemical parasitology. 2007 Nov;156(1):74-9.
107. Parobek CM, Bailey JA, Hathaway NJ, Socheat D, Rogers WO, Juliano JJ. Differing patterns of selection and geospatial genetic diversity within two leading Plasmodium vivax candidate vaccine antigens. PLoS Negl Trop Dis. 2014 Apr;8(4):e2796.
108. Sjostrand J, Tofigh A, Daubin V, Arvestad L, Sennblad B, Lagergren J. A Bayesian method for analyzing lateral gene transfer. Syst Biol. 2014 May;63(3):409-20.
109. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014 May 1;30(9):1312-3.
110. Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Hohna S, et al. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst Biol. 2012 May;61(3):539-42.
111. Abascal F, Zardoya R, Posada D. ProtTest: selection of best-fit models of protein evolution. Bioinformatics. 2005 May 1;21(9):2104-5.
112. Miller MA, Pfeiffer, W., and Schwartz, T. Creating the CIPRES Science Gateway for inference of large phylogenetic trees. Proceedings of the Gateway Computing Environments Workshop (GCE), 14 Nov 2010, New Orleans, LA pp 1 - 8. 2010.
113. Miller MA, Schwartz T, Pickett BE, He S, Klem EB, Scheuermann RH, et al. A RESTful API for Access to Phylogenetic Tools via the CIPRES Science Gateway. Evol Bioinform Online. 2015;11:43-8.

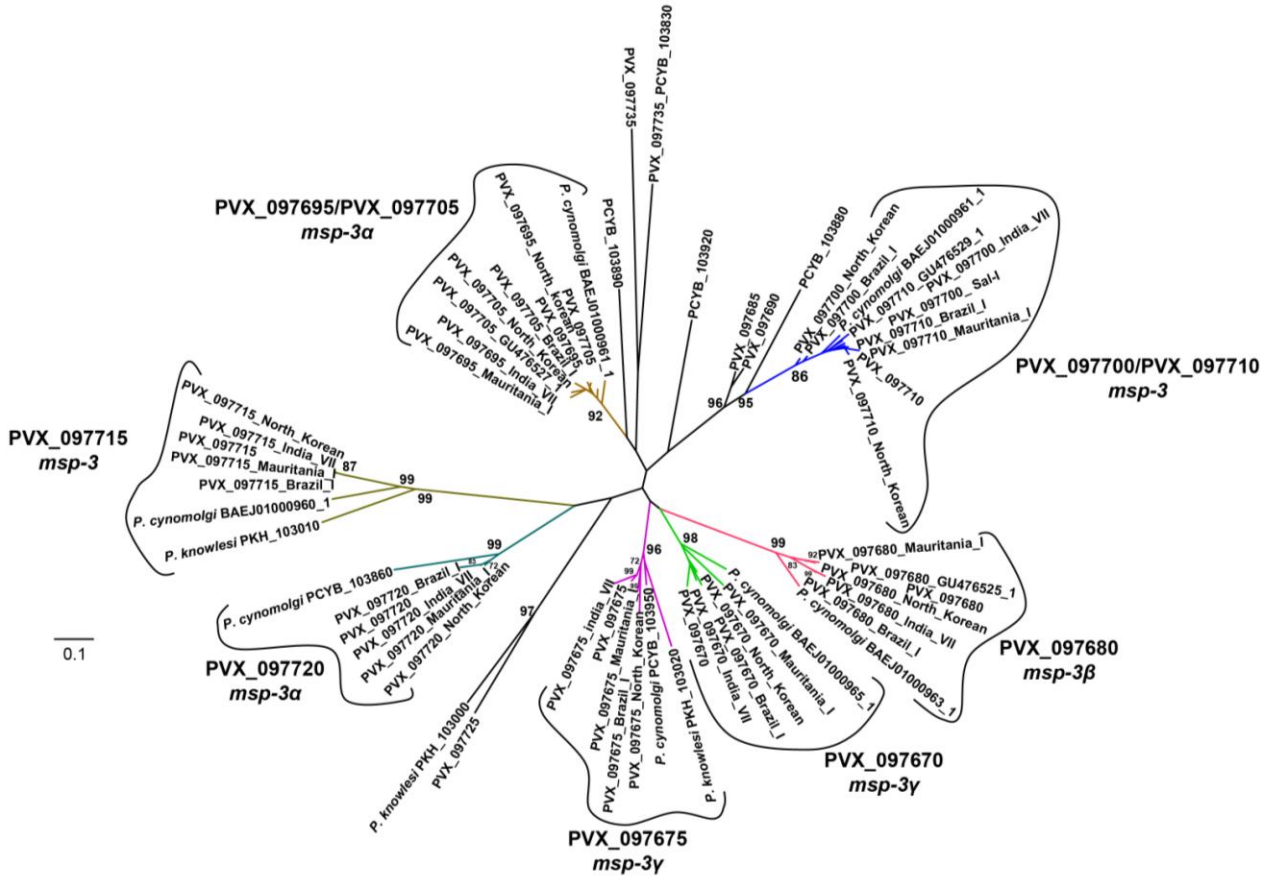
114. Kosakovsky Pond SL, Murrell B, Fourment M, Frost SD, Delport W, Scheffler K. A random effects branch-site model for detecting episodic diversifying selection. Mol Biol Evol. 2011 Nov;28(11):3033-43.
115. Darriba D, Taboada GL, Doallo R, Posada D. jModelTest 2: more models, new heuristics and parallel computing. Nat Methods. 2012 Aug;9(8):772.
116. Pond SL, Frost SD, Muse SV. HyPhy: hypothesis testing using phylogenies. Bioinformatics. 2005 Mar 1;21(5):676-9.
117. Delport W, Poon AF, Frost SD, Kosakovsky Pond SL. Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology. Bioinformatics. 2010 Oct 1;26(19):2455-7.
118. Murrell B, Wertheim JO, Moola S, Weighill T, Scheffler K, Kosakovsky Pond SL. Detecting individual sites subject to episodic diversifying selection. PLoS Genet. 2012;8(7):e1002764.
119. Kadekoppala M, Holder AA. Merozoite surface proteins of the malaria parasite: the MSP1 complex and the MSP7 family. Int J Parasitol. 2010 Aug 15;40(10):1155-61.
120. Sjostrand J, Sennblad B, Arvestad L, Lagergren J. DLRS: gene tree evolution in light of a species tree. Bioinformatics. 2012 Nov 15;28(22):2994-5.
121. Kauth CW, Woehlbier U, Kern M, Mekonnen Z, Lutz R, Mucke N, et al. Interactions between merozoite surface proteins 1, 6, and 7 of the malaria parasite Plasmodium falciparum. J Biol Chem. 2006 Oct 20;281(42):31517-27.
122. Garcia Y, Puentes A, Curtidor H, Cifuentes G, Reyes C, Barreto J, et al. Identifying merozoite surface protein 4 and merozoite surface protein 7 Plasmodium falciparum protein family members specifically binding to human erythrocytes suggests a new malarial parasite-redundant survival mechanism. J Med Chem. 2007 Nov 15;50(23):5665-75.
123. Pachebat JA, Ling IT, Grainger M, Trucco C, Howell S, Fernandez-Reyes D, et al. The 22 kDa component of the protein complex on the surface of Plasmodium falciparum merozoites is derived from a larger precursor, merozoite surface protein 7. Molecular and biochemical parasitology. 2001 Sep 28;117(1):83-9.
124. Mello K, Daly TM, Morrisey J, Vaidya AB, Long CA, Bergman LW. A multigene family that interacts with the amino terminus of plasmodium MSP-1 identified using the yeast two-hybrid system. Eukaryot Cell. 2002 Dec;1(6):915-25.
125. Muehlenbein MP, Pacheco MA, Taylor JE, Prall SP, Ambu L, Nathan S, et al. Accelerated diversification of nonhuman primate malarias in Southeast Asia: adaptive radiation or geographic speciation? Mol Biol Evol. 2015 Feb;32(2):422-39.
126. Carvalho LJ, Daniel-Ribeiro CT, Goto H. Malaria vaccine: candidate antigens, mechanisms, constraints and prospects. Scand J Immunol. 2002 Oct;56(4):327-43.
127. Jones TR, Hoffman SL. Malaria vaccine development. Clin Microbiol Rev. 1994 Jul;7(3):303-10.
128. Escalante AA, Lal AA, Ayala FJ. Genetic polymorphism and natural selection in the malaria parasite Plasmodium falciparum. Genetics. 1998 May;149(1):189-202.
129. Chenet SM, Tapia LL, Escalante AA, Durand S, Lucas C, Bacon DJ. Genetic diversity and population structure of genes encoding vaccine candidate antigens of Plasmodium vivax. Malar J. 2012;11:68.

130. Imwong M, Pukrittayakamee S, Gruner AC, Renia L, Letourneur F, Looareesuwan S, et al. Practical PCR genotyping protocols for Plasmodium vivax using Pvcs and Pvmsp1. Malar J. 2005 Apr 27;4:20.
131. Bruce MC, Galinski MR, Barnwell JW, Snounou G, Day KP. Polymorphism at the merozoite surface protein-3alpha locus of Plasmodium vivax: global and local diversity. Am J Trop Med Hyg. 1999 Oct;61(4):518-25.
132. Kosakovsky Pond SL, Frost SD. Not so different after all: a comparison of methods for detecting amino acid sites under selection. Mol Biol Evol. 2005 May;22(5):1208-22.
133. Murrell B, Moola S, Mabona A, Weighill T, Sheward D, Kosakovsky Pond SL, et al. FUBAR: a fast, unconstrained bayesian approximation for inferring selection. Mol Biol Evol. 2013 May;30(5):1196-205.
134. Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SD. Automated phylogenetic detection of recombination using a genetic algorithm. Mol Biol Evol. 2006 Oct;23(10):1891-901.
135. Bourgard C, Albrecht L, Kayano A, Sunnerhagen P, Costa FTM. Plasmodium vivax Biology: Insights Provided by Genomics, Transcriptomics and Proteomics. Front Cell Infect Microbiol. 2018;8:34.
136. Moreno-Perez DA, Ruiz JA, Patarroyo MA. Reticulocytes: Plasmodium vivax target cells. Biol Cell. 2013 Jun;105(6):251-60.
137. Malleret B, Li A, Zhang R, Tan KS, Suwanarusk R, Claser C, et al. Plasmodium vivax: restricted tropism and rapid remodeling of CD71-positive reticulocytes. Blood. 2015 Feb 19;125(8):1314-24.
138. Patarroyo ME, Patarroyo MA. Emerging rules for subunit-based, multiantigenic, multistage chemically synthesized vaccines. Acc Chem Res. 2008 Mar;41(3):377-86.
139. Rodriguez LE, Curtidor H, Urquiza M, Cifuentes G, Reyes C, Patarroyo ME. Intimate molecular interactions of P. falciparum merozoite proteins involved in invasion of red blood cells and their implications for vaccine design. Chem Rev. 2008 Sep;108(9):3656-705.
140. Moreno-Perez DA, Baquero LA, Chitiva-Ardila DM, Patarroyo MA. Characterising PvRBSA: an exclusive protein from Plasmodium species infecting reticulocytes. Parasit Vectors. 2017 May 18;10(1):243.
141. Arevalo-Pinzon G, Bermudez M, Curtidor H, Patarroyo MA. The Plasmodium vivax rhoptry neck protein 5 is expressed in the apical pole of Plasmodium vivax VCG-1 strain schizonts and binds to human reticulocytes. Malar J. 2015 Mar 7;14:106.
142. Arevalo-Pinzon G, Bermudez M, Hernandez D, Curtidor H, Patarroyo MA. Plasmodium vivax ligand-receptor interaction: PvAMA-1 domain I contains the minimal regions for specific interaction with CD71+ reticulocytes. Sci Rep. 2017 Aug 30;7(1):9616.
143. Moreno-Pérez DA. Determinación del proteoma de la cepa VCG1 de Plasmodium vivax y caracterización de moléculas candidatas para su inclusión en el desarrollo de una vacuna. Bogotá: Universidad del Rosario y Universidad de Salamanca. 2017.
144. Escalante AA, Cornejo OE, Freeland DE, Poe AC, Durrego E, Collins WE, et al. A monkey's tale: the origin of Plasmodium vivax as a human malaria parasite. Proc Natl Acad Sci U S A. 2005 Feb 8;102(6):1980-5.
145. Baquero LA, Moreno-Perez DA, Garzon-Ospina D, Forero-Rodriguez J, Ortiz-Suarez HD, Patarroyo MA. PvGAMA reticulocyte binding activity: predicting conserved

functional regions by natural selection analysis. Parasites & vectors. 2017 May 19;10(1):251.
146. Rodrigues-da-Silva RN, Soares IF, Lopez-Camacho C, Martins da Silva JH, Perce-da-Silva DS, Teva A, et al. Plasmodium vivax Cell-Traversal Protein for Ookinetes and Sporozoites: Naturally Acquired Humoral Immune Response and B-Cell Epitope Mapping in Brazilian Amazon Inhabitants. Front Immunol. 2017;8:77.
147. Cheng Y, Lu F, Wang B, Li J, Han JH, Ito D, et al. Plasmodium vivax GPI-anchored micronemal antigen (PvGAMA) binds human erythrocytes independent of Duffy antigen status. Sci Rep. 2016 Oct 19;6:35581.
148. Bermudez M, Arevalo-Pinzon G, Rubio L, Chaloin O, Muller S, Curtidor H, et al. Receptor-ligand and parasite protein-protein interactions in Plasmodium vivax: Analysing rhoptry neck proteins 2 and 4. Cell Microbiol. 2018 Jul;20(7):e12835.

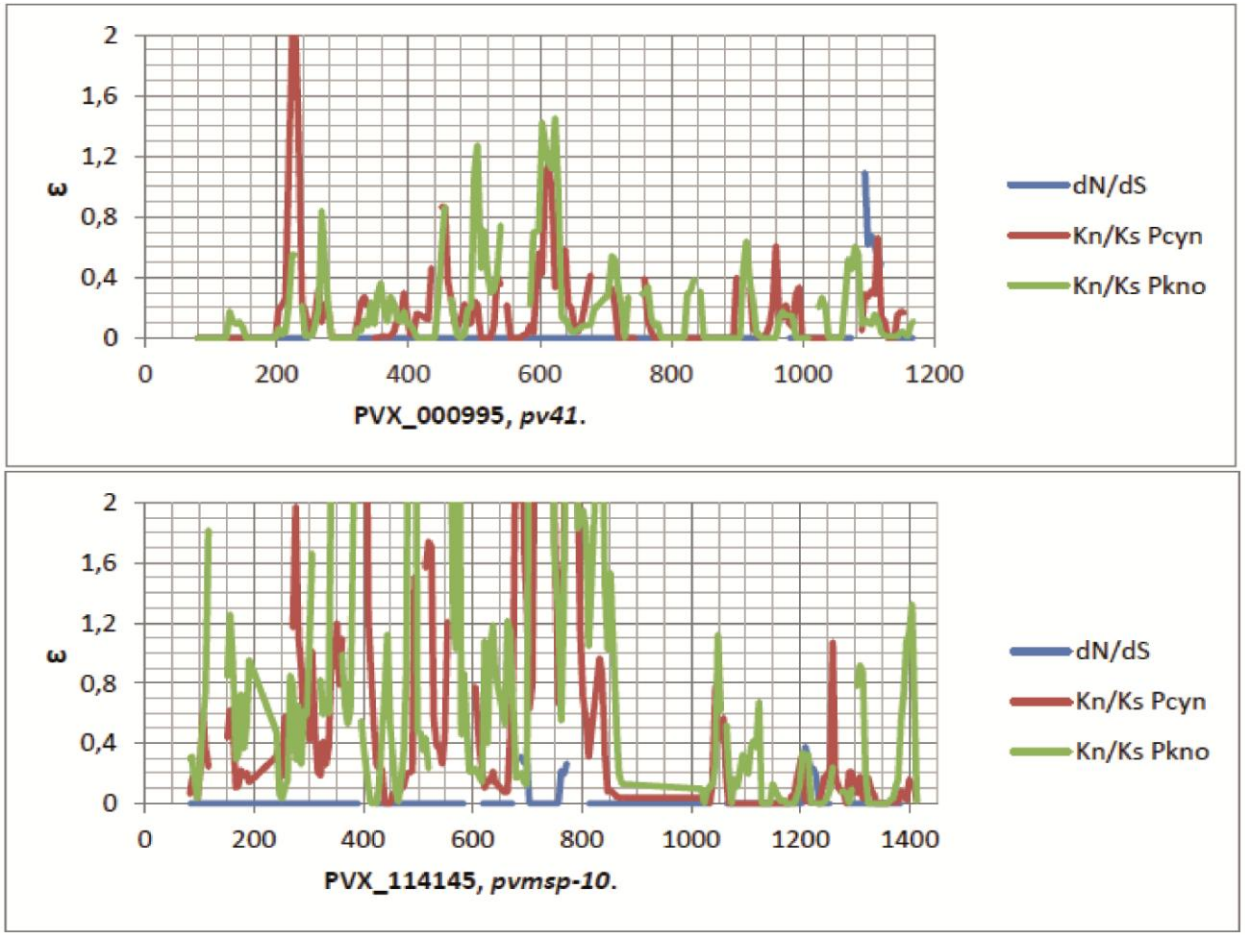
ANEXOS

Anexo 1. Árbol filogenético de la familia msp3



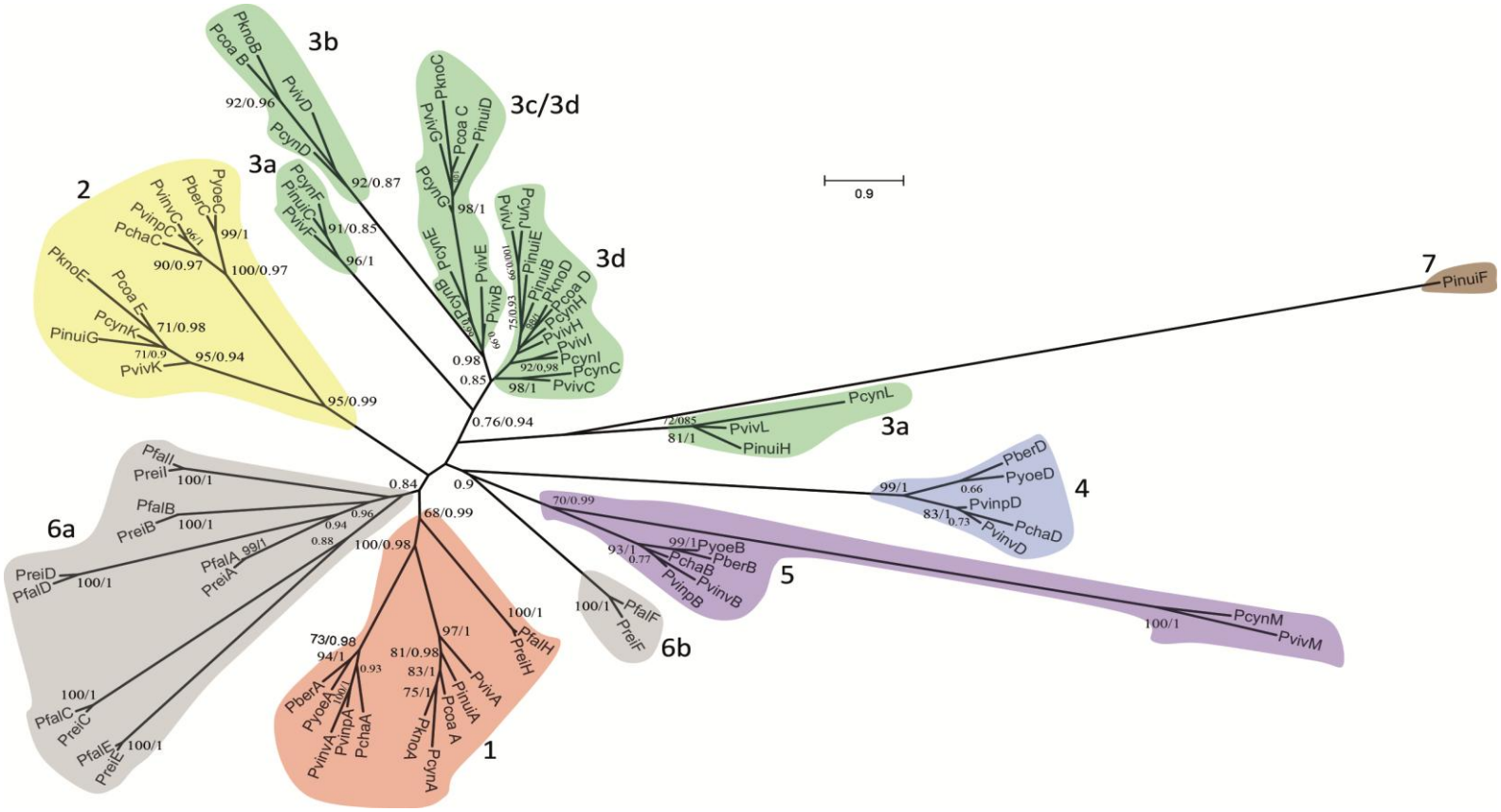
Los genes ortólogos deben agruparse en una relación uno a uno. Se muestra sólo el árbol de la familia msp3, los árboles restantes pueden ser consultados en el Suplemento 2 de la publicación 1 (https://www.sciencedirect.com/science/article/pii/S156713481500163X).

Anexo 2. Ventanas deslizantes de la tasa ω (d_N/d_S o K_N/K_S)



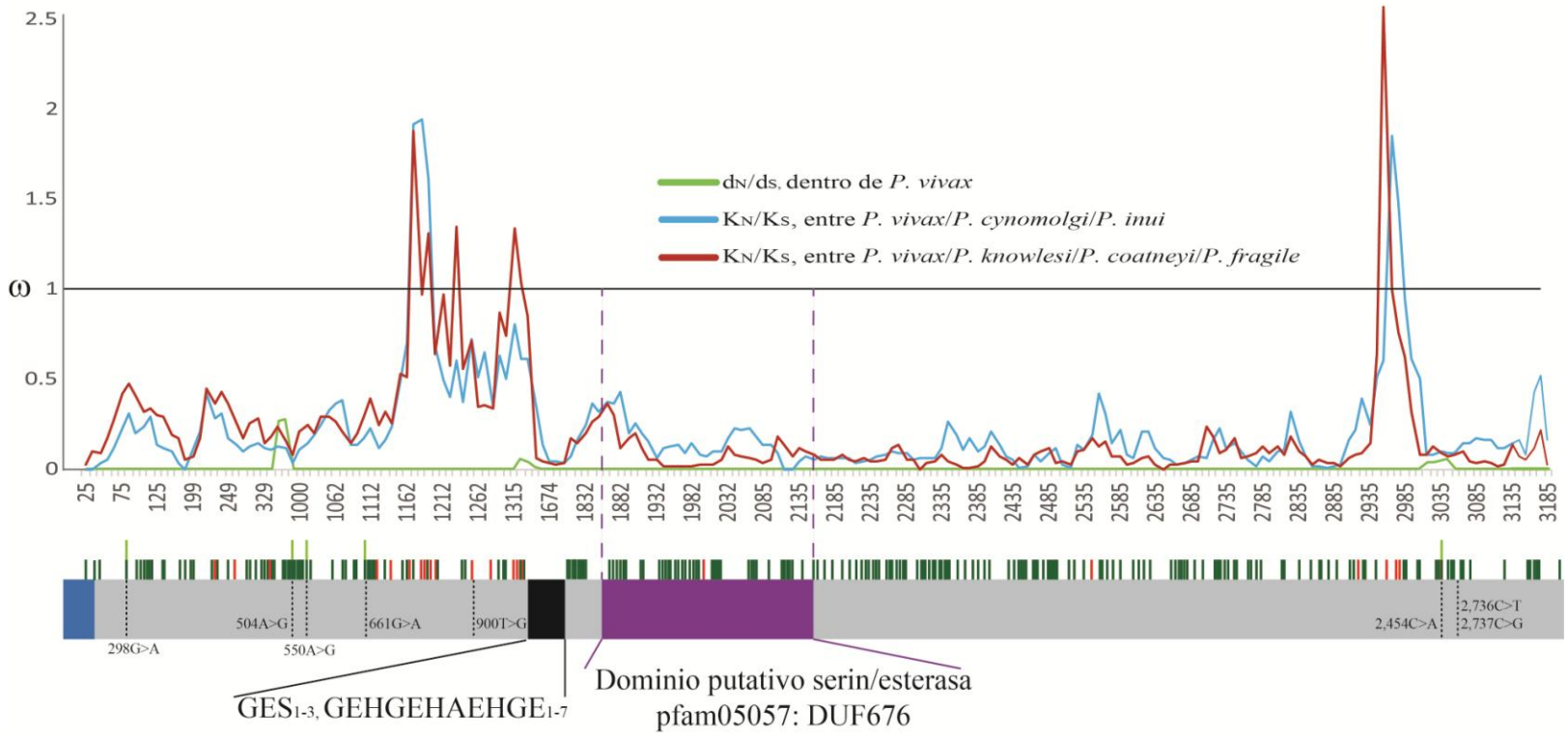
La razón d_N/d_S dentro P. vivax es mostrada en azul, mientras la razón de la divergencia K_N/K_S entre P. vivax y P. cynomolgi, y entre P. vivax and P. knowlesi, es mostrada en rojo y verde, respectivamente. Sólo se muestran las ventanas de los genes pv41 y msp10, las ventanas deslizantes restantes pueden ser consultadas en el suplemento 3 de la publicación 1 (https://www.sciencedirect.com/science/article/pii/S156713481500163X). En msp10, la región 3' que mostró un ω < 1 codifica para los dominios EGF-like, involucrados en interacciones proteína-proteína.

Anexo 3. Filogenia de la familia msp7 inferida por el método de máxima verosimilitud



Los números (basados en los números de la Figura 7) representan los diferentes clados, mientras que los números en las ramas son los valores del bootstrap y de la probabilidad posterior. Trece clados fueron identificados en el árbol; sin embargo, los clados 1 y 6 no fueron claramente respaldados por los valores del bootstrap o los valores de probabilidad posteriores. Las proteínas se agrupan de acuerdo con las relaciones filogenéticas del parásito, siendo los clados 1 (rojo) y 2 (amarillo) los más ancestrales; el clado 5 también podría ser un clúster ancestral. Las secuencias de los clados del linaje de parásitos que infectan cercopitécidos se representan en verde, las proteínas del linaje de parásitos que infectan roedores en azul y el linaje de parásitos que infectan homínidos en gris. El grupo de proteínas en marrón es una duplicación observada sólo en la especie P. inui. Tanto los árboles ML como BY mostraron topologías similares, por lo que sólo se muestra el árbol ML.

Anexo 4. Ventana deslizante del gen pvron4



Los valores ω de ron4 (d_N/d_S) dentro de P. vivax se muestran en verde, mientras la tasa de divergencia ω (K_N / K_S) entre las especies que infectan primates y P. vivax se muestran en azul (para las especies evolutivamente más cercanas a P. vivax) y rojas (más distantes). Una representación del gen ron4 se presenta debajo de la ventana deslizante que indica el péptido señal (azul), la región repetida (negra) y el dominio putativo de esterasa /lipasa (púrpura). Los SNPs identificados en P. vivax se muestran como líneas discontinuas. La numeración está basada en el alineamiento del archivo adicional 2 de la Publicación 6. Los sitios bajo selección negativa están representados por líneas verdes claras (dentro de P. vivax) y verde oscuro (entre las especies), mientras que los sitios seleccionados positivamente entre especies se muestran por líneas rojas. Detalles de esta figura pueden ser consultados en la Publicación 6 (https://malariajournal.biomedcentral.com/articles/10.1186/s12936-016-1563-4).

PUBLICACIONES

Publicación 1: **Garzón-Ospina D**, Forero-Rodríguez J, Patarroyo MA. *Inferring natural selection signals in Plasmodium vivax-encoded proteins having a potential role in merozoite invasion*. Infect Genet Evol. 2015 Jul; 33:182-8. doi: 10.1016/j.meegid.2015.05.001.

Publicación 2: **Garzón-Ospina D**, Forero-Rodríguez J, Patarroyo MA. *Evidence of functional divergence in MSP7 paralogous proteins: a molecular-evolutionary and phylogenetic analysis*. BMC Evol Biol. 2016 Nov 28;16 (1):256. doi: 10.1186/s12862-016-0830-x.

Publicación 3: Forero-Rodríguez J, **Garzón-Ospina D**, Patarroyo MA. *Low genetic diversity and functional constraint in loci encoding Plasmodium vivax P12 and P38 proteins in the Colombian population*. Malar J. 2014 Feb 18;13:58. doi: 10.1186/1475-2875-13-58.

Publicación 4: Forero-Rodríguez J, **Garzón-Ospina D**, Patarroyo MA. *Low genetic diversity in the locus encoding the Plasmodium vivax P41 protein in Colombia's parasite population*. Malar J. 2014 Sep 30;13:388. doi: 10.1186/1475-2875-13-388.

Publicación 5: **Garzón-Ospina D**, Forero-Rodríguez J, Patarroyo MA. *Heterogeneous genetic diversity pattern in Plasmodium vivax genes encoding merozoite surface proteins (MSP) -7E, -7F and -7L*. Malar J. 2014 Dec 13;13:495. doi: 10.1186/1475-2875-13-495.

Publicación 6: Buitrago SP, **Garzón-Ospina D**, Patarroyo MA. *Size polymorphism and low sequence diversity in the locus encoding the Plasmodium vivax rhoptry neck protein 4 (PvRON4) in Colombian isolates*. Malar J. 2016 Oct 18;15(1):501. doi: 10.1186/s12936-016-1563-4.

Publicación 7: Baquero LA, Moreno-Pérez DA, **Garzón-Ospina D**, Forero-Rodríguez J, Ortiz-Suárez HD, Patarroyo MA. *PvGAMA reticulocyte binding activity: predicting conserved functional regions by natural selection analysis*. Parasit Vectors. 2017 May 19;10(1):251. doi: 10.1186/s13071-017-2183-8.

Publicación 8: **Garzón-Ospina D**, Buitrago SP, Ramos AE, Patarroyo MA. *Identifying Potential Plasmodium vivax Sporozoite Stage Vaccine Candidates: An Analysis of Genetic Diversity and Natural Selection*. Front Genet. 2018 Jan 25;9:10. doi:10.3389/fgene.2018.00010.

Publicación 9: Camargo-Ayala PA, **Garzón-Ospina D**, Moreno-Pérez DA, Ricaurte-Contreras LA, Noya O and Patarroyo MA. *On the evolution and function of Plasmodium vivax reticulocyte binding surface antigen (pvrbsa)*. Front. Genet. 2018 Sep . 9:372 doi: 10.3389/fgene.2018.00372

Infection, Genetics and Evolution 33 (2015) 182–188

Contents lists available at ScienceDirect

Infection, Genetics and Evolution

journal homepage: www.elsevier.com/locate/meegid

Short communication

Inferring natural selection signals in *Plasmodium vivax*-encoded proteins having a potential role in merozoite invasion

Diego Garzón-Ospina [a,b], Johanna Forero-Rodríguez [a], Manuel A. Patarroyo [a,b,*]

[a] Molecular Biology and Immunology Department, Fundación Instituto de Inmunología de Colombia (FIDIC), Carrera 50 No. 26-20, Bogotá DC, Colombia
[b] Basic Sciences Department, School of Medicine and Health Sciences, Universidad del Rosario, Carrera 24 No. 63C-69, Bogotá DC, Colombia

ARTICLE INFO

Article history:
Received 5 February 2015
Received in revised form 30 April 2015
Accepted 2 May 2015
Available online 2 May 2015

Keywords:
Plasmodium vivax
Anti-malarial vaccine
Natural selection signal
Allele-specific response

ABSTRACT

Detecting natural selection signals in *Plasmodium* parasites antigens might be used for identifying potential new vaccine candidates. Fifty-nine *Plasmodium vivax*-Sal-I genes encoding proteins having a potential role in invasion were used as query for identifying them in recent *P. vivax* strain genome sequences and two closely-related *Plasmodium* species. Several measures of DNA sequence variation were then calculated and selection signatures were detected by using different approaches. Our results may be used for determining which genes expressed during *P. vivax* merozoite stage could be prioritised for further population genetics or functional studies for designing a *P. vivax* vaccine which would avoid allele-specific immune responses.

1. Introduction

Malaria is a disease caused by *Plasmodium* parasites (Cox, 2010). *Plasmodium falciparum* is the best characterised species, whereas research into *Plasmodium vivax* has been more limited (Arnott et al., 2012; Patarroyo et al., 2012). Likewise, anti-*P. vivax* vaccine development is behindhand and few vaccine candidates have been proposed to date. Characterising potential new candidates involved the search for sequences having a high level of identity with *P. falciparum* antigens (Patarroyo et al., 2012). Recently, Restrepo-Montoya et al. (2011) has led to categorising several *P. vivax* proteins having a potential role in invasion by bioinformatics approaches.

The proteins characterised in the aforementioned studies could be used for *P. vivax* asexual-blood vaccine development since they seem to be implicated in invasion; however, these molecules' genetic diversity and evolutionary forces must be ascertained by population genetics analysis to design a completely effective vaccine (Arnott et al., 2012; Barry and Arnott, 2014). The most commonly used tests in population genetics are based on the allele frequency spectrum and require the sequencing of many isolates; therefore, performing such studies for all these genes would involve much time and resources. However, Cornejo et al. (2014), using a limited sample size (Genomes from 5 isolates) have identified genes having signatures consistent with selection. This kind of analysis could be a starting point for detecting potential new vaccine candidates (Weedall and Conway, 2010), similar to the approach adopted for *P. falciparum* (Ochola et al., 2010; Tetteh et al., 2009).

The present study has used three different approaches for detecting selection signals within 59 previously-characterised merozoite antigens using the sequences from five *P. vivax* isolates and two closely-related species. The results may be used for determining which antigens might be prioritised and evaluated in further studies aimed at designing a completely effective vaccine.

2. Material and methods

2.1. Target sequences and alignments

Sequences were obtained for 59 protein-encoding genes from the Salvador I isolate (Sal-I); these genes had been previously characterised by adopting a molecular approach (Arevalo-Pinzon et al., 2011, 2013; Moreno-Perez et al., 2013b; Patarroyo et al., 2012) or suggested as promising vaccine candidates by having a potential role in invasion (Restrepo-Montoya et al., 2011) (Supplementary data 1). Forty-eight genes had not been subjected to previous population genetic analysis and 11 have been previously evaluated.

* Corresponding author at: Molecular Biology and Immunology Department, Fundación Instituto de Inmunología de Colombia (FIDIC), Carrera 50 No. 26-20, Bogotá DC, Colombia.
E-mail addresses: degarzon@gmail.com (D. Garzón-Ospina), lady2007_10@hotmail.com (J. Forero-Rodríguez), mapatarr.fidic@gmail.com (M.A. Patarroyo).

http://dx.doi.org/10.1016/j.meegid.2015.05.001
1567-1348/

These sequences were used as query for searching for them in the available genomic *P. vivax* isolate sequences (Neafsey et al., 2012) and two closely-related species (*Plasmodium cynomolgi* and *Plasmodium knowlesi*) (Pain et al., 2008; Tachibana et al., 2012) using the tBlastn tool from the protozoa genomic NCBI database. A tBlastn search in GenBank database was made regarding sequences reported for other stains (VCG-I, Belen or South Korea).

Some genes belong to multigene families; therefore orthologous identification should be performed. A combination of criteria was used for identifying putative orthologues, including a phylogenetic signal (tree topology), sequence similarity (genetic distance) and synteny (similar genomic position), as previously described (Arisue et al., 2011; Garzon-Ospina et al., 2010, 2014; Rice et al., 2014). *P. vivax*, *P. cynomolgi* and *P. knowlesi sera*, *msp-3*, *msp-7*, *clag*, *pfam-a* and *pfam-d* genes were aligned with all members of their families, respectively using MUSCLE (Edgar, 2004), followed by manual edition. The best evolutionary model was selected for each alignment by Bayesian Information Criterion, using MEGA software (Tamura et al., 2011). Maximum likelihood phylogenetic trees were then inferred using the respective model; all gaps and ambiguously-aligned regions were removed. Topology reliability was evaluated by bootstrapping (1000 iterations). Multiple alignments were then made (by MUSCLE) for single-copy genes using sequences from isolates, together with *P. cynomolgi* and *P. knowlesi* orthologous sequences.

2.2. Genetic diversity and natural selection analysis

DnaSP software (Librado and Rozas, 2009) was used for estimating several measures regarding DNA sequence variation. Cornejo et al. (2014) had previously identified patterns consistent with natural selection acting across the *P. vivax* genome by using the two-dimensional Hudson, Kreitman and Aguade (HKA) test, the genome-wide version of the McDonald–Kreitman (MK) test and Tajima D estimator; however, natural selection signals were not found for several genes involved in merozoite invasion. We assessed natural selection by conventional MK test (McDonald and Kreitman, 1991) and π/K ratio. The MK test was performed taking the Jukes–Cantor divergence correction into account (Jukes, 1969) by using a web server (Egea et al., 2008). The π/K ratio was evaluated for identifying genes having a high value correlated with balancing selection (Ochola et al., 2010; Tetteh et al., 2009). MEGA software was used to assess selection signals within *P. vivax* by calculating the non-synonymous substitution per site rate (d_N) and synonymous substitution per site rate (d_S) by the modified Nei–Gojobori method (Zhang et al., 1998). Likewise, to infer natural selection signatures which could have prevailed during *Plasmodium* evolutionary history (using *P. vivax*, *P. cynomolgi* and *P. knowlesi* sequences as data set) the difference between the average number of non-synonymous divergence substitutions per non-synonymous site rate and of synonymous divergence substitutions per synonymous site rate (K_N–K_S) was inferred using the modified Nei–Gojobori method with Jukes–Cantor correction. Significant differences were evaluated by *Z*-test (when non-synonymous and synonymous substitutions > 10) or Fisher's exact test (when non-synonymous and synonymous substitutions < 10). Furthermore, a sliding window for d_N/d_S and K_N/K_S ratios (ω) was performed; gaps and ambiguously aligned regions were removed for analysis. Genetic diversity and selection were assessed by Sal-I annotation as reference.

The most suitable antigens for vaccine development regarding our approach should have limited diversity or at least a domain having this pattern. Such genes/domains should have a natural negative selection signal (and $\omega < 1$). However, genes under positive selection might be taken into account if provided with domains having both limited diversity and low ω values.

3. Results

3.1. Diversity analysis

Sequences from 59 previously identified *P. vivax* protein-encoding genes having a potential role in invasion were analysed here. Sequence analysis revealed premature stop codons in PVX_096990 and PVX_097710 genes in Mauritania-I and Brazil-I isolates, respectively. Few genes were absent or incomplete in some isolates; i.e. the PVX_003825 gene was not found in a North Korean isolate whereas PVX_092425 was incomplete at the 5′-end, PVX_086850 and PVX_086930 were missing in the Brazil-I isolate, PVX_097700 was absent in the Mauritania-I isolate and PVX_097710 appeared not to be present in the India-VII isolate in which the PVX_096990 gene was incomplete at the 5′-end.

Genetic diversity measurement revealed 16 highly polymorphic genes ($\pi > 0.01$), 35 with intermediate polymorphism ($0.009 < \pi > 0.001$) and 8 having low genetic diversity ($\pi < 0.001$) (Table 1 and Supplementary data 1). Fig. 1 shows the nucleotide polymorphism distribution within the aforementioned 59 genes.

Phylogenetic trees were then inferred to determinate putative orthologous relationships for the multigene families (Supplementary data 2). Putative orthologues had to be clustered in a clade in a one-to-one relationship and had to have a similar genomic position. The clades formed in some families agreed with previous reports (Arisue et al., 2011; Garzon-Ospina et al., 2010; Rice et al., 2014). Hence, 34 of these 59 genes were found in both *P. cynomolgi* and *P. knowlesi* species, whereas another 15 genes were only present in *P. cynomolgi*; 10 genes appeared to be exclusive to *P. vivax*.

3.2. Natural selection signatures in P. vivax genes

Three different approaches were used for screening natural selection signals. The neutral index (NI) from the MK test showed that 16 genes had excess polymorphism regarding divergence (NI > 1) and only one had NI < 1, while the π/K ratio showed 11 genes which might be under balancing selection (Table 2 and Supplementary data 1).

A statistically significant $d_N > d_S$ was found in 12 genes while 9 had significant $d_N < d_S$ values (Table 2 and Supplementary data 1). The K_N–K_S difference gave negative selection between species for 35 genes whereas another 10 displayed positive selection (Supplementary data 1). Some genes had a different natural selection signal from that previously reported (Tachibana et al., 2012), probably since we used sequences from 5 isolates, unlike Tachibana et al., who only used the Sal-I isolate. Some genes evaluated here were not assessed in the aforementioned report.

Since the Nei–Gojobori method is a conservative test, we performed a sliding window for the ω rate (d_N/d_S and/or K_N/K_S) for identifying specific domains within genes having a determined selective signal (Supplementary data 3). Several genes lacking significant d_N or d_S rates displayed a particular domain having $d_N/d_S \pm 1$ but also $K_N/K_S < 1$ throughout sequences. Members of *pvmsp-3* and *pvsera* multigene families had $K_N/K_S > 1$ values throughout all genes, suggesting high divergence between *P. vivax* and related species.

4. Discussion

A vaccine focusing on *P. vivax* is urgently needed for malaria control; however, its design has been delayed, mainly due to slow

Table 1
Measurement of DNA sequence variation for the 59 *P. vivax* genes.

#	ID	Name	*n*	Sites	*Ss*	*S*	*Ps*	π (SD)
Genes lacking previous population genetics analysis								
1	PVX_000945	*pvron-1*	6	2340	5	3	2	0.0096 (0.0001)
2	PVX_000995	*pv41*	6	1086	14	5	9	0.0064 (0.0013)
3	PVX_002510	Nucleosomal binding protein 1	5	750	2	1	1	0.0013 (0.0003)
4	PVX_003800	*pvsera*	5	3033	4	4	0	0.0005 (0.0005)
5	PVX_003805	*pvsera*, putative	5	3507	436	195	241	0.0728 (0.0100)
6	PVX_003815	*pvsera*, truncated, putative	5	1335	15	12	3	0.0049 (0.0014)
7	PVX_003825	*pvsera-4*	4	2814	144	113	31	0.0282 (0.0072)
8	PVX_003830	*pvsera-5*	6	3090	752	528	224	0.1038 (0.0222)
9	PVX_003850	*pvsera-2*	6	3042	12	9	3	0.0016 (0.0004)
10	PVX_080305	Hypothetical protein, conserved	5	804	1	1	0	0.0005 (0.0003)
11	PVX_081810	Hypothetical protein, conserved	5	3921	18	15	3	0.0020 (0.0004)
12	PVX_081845	Hypothetical protein	5	1044	3	1	2	0.0015 (0.0003)
13	PVX_084720	Hypothetical protein, conserved	5	2720	9	5	4	0.0016 (0.0003)
14	PVX_086850	*pvvir-35*, putative	4	662	40	19	21	0.0360 (0.0090)
15	PVX_086930	*pvrhopH1/clag*	5	3978	38	30	8	0.0043 (0.0010)
16	PVX_090075	*pv34*	6	1092	1	1	0	0.0003 (0.0003)
17	PVX_090210	*pvarp*	7	682	6	5	1	0.0029 (0.0007)
18	PVX_091434	*pvron-4*	6	2097	14	5	9	0.0033 (0.0007)
19	PVX_092425	Hypothetical protein, conserved	4	1950	25	25	0	0.0070 (0.0032)
20	PVX_092975	Erythrocyte binding protein 1	6	3440	36	31	5	0.0038 (0.0009)
21	PVX_092995	Tryptophan-rich antigen	5	1059	25	19	6	0.0105 (0.0036)
22	PVX_094425	Hypothetical protein, conserved	5	3045	3	2	1	0.0005 (0.0001)
23	PVX_096990	Pv-fam-d protein	5	1136	22	12	10	0.0095 (0.0022)
24	PVX_097565	*Plasmodium* exported protein	5	1311	6	6	0	0.0018 (0.0003)
25	PVX_097670	*pvmsp-3γ*, putative	6	1743	524	282	242	0.1408 (0.0154)
26	PVX_097675	*pvmsp-3γ*, putative	5	1758	445	281	164	0.1268 (0.0229)
27	PVX_097695	*pvmsp-3α*, putative	5	2613	424	240	184	0.0817 (0.0126)
28	PVX_097700	*pvmsp-3*, putative	4	3306	770	577	201	0.1332 (0.0245)
29	PVX_097705	*pvmsp-3α*, putative	5	2607	440	264	176	0.0841 (0.0110)
30	PVX_097710	*pvmsp-3*, putative	5	3607	867	506	361	0.1229 (0.0185)
31	PVX_097715	*pvmsp-3*, putative	5	1286	49	42	7	0.0163 (0.0034)
32	PVX_097960	*pv38*	6	1065	4	0	4	0.0022 (0.0004)
33	PVX_098585	*pvrbp-1*, putative	6	8451	39	22	17	0.0020 (0.0003)
34	PVX_098712	*pvrhopH3*	6	2673	4	3	1	0.0006 (0.0002)
35	PVX_101505	Pv-fam-d protein	5	1263	5	2	3	0.0021 (0.0004)
36	PVX_101555	Hypothetical protein	5	2586	167	98	69	0.0337 (0.0081)
37	PVX_101605	Hypothetical protein	5	582	3	0	3	0.0031 (0.0030)
38	PVX_109280	*pvfam-a*	7	753	1	1	0	0.0004 (0.0003)
39	PVX_112665	Tryptophan-rich antigen	5	867	3	1	2	0.0018 (0.0004)
40	PVX_113775	*pv12*	7	942	4	2	1	0.0014 (0.0006)
41	PVX_117230	*pvser/thr*	5	4122	5	3	2	0.0006 (0.0001)
42	PVX_117880	*pvron-2*	7	6495	31	23	8	0.0016 (0.0003)
43	PVX_118525	Hypothetical protein, conserved	5	5082	19	13	6	0.0017 (0.0004)
44	PVX_121885	*pvclag*, putative	5	4239	63	44	19	0.0069 (0.0010)
45	PVX_121920	*pvrbp-2*, like	5	7461	31	15	16	0.0021 (0.0003)
46	PVX_123105	Hypothetical protein, conserved	5	2114	1	1	0	0.0002 (0.0001)
47	PVX_123550	Hypothetical protein, conserved	5	647	4	3	1	0.0027 (0.0006)
48	PVX_123575	Thrombospondin-related protein 3	6	966	3	2	1	0.0012 (0.0004)
Genes having previous population genetics analysis								
1	PVX_003905	*pv230*, putative	5	8199	22	14	8	0.0013 (0.0002)
2	PVX_082695	*pvmsp-7K*, putative	6	849	7	5	2	0.0033 (0.0006)
3	PVX_085930	*pvrap-1*, putative	5	2223	2	2	0	0.0003 (0.0002)
4	PVX_092275	*pvama-1*	6	1683	31	18	13	0.0080 (0.0010)
5	PVX_097590	*pvrap-2*, putative	5	1203	0	0	0	0.0000 (0.0000)
6	PVX_097625	*pvmsp-8*	6	1320	6	3	3	0.0021 (0.0006)
7	PVX_097720	*pvmsp-3α*	6	2016	154	72	82	0.0349 (0.0158)
8	PVX_097680	*pvmsp-3β*	6	2007	329	165	164	0.0747 (0.0094)
9	PVX_099980	*pvmsp-1*	6	5058	532	170	362	0.0527 (0.0068)
10	PVX_110810	*pvdbp*	6	3210	45	23	22	0.0063 (0.0010)
11	PVX_114145	*pvmsp-10*	6	1288	5	0	5	0.0021 (0.0003)

n: number of sequences analysed; sites: total sites analysed, excluding gaps; *Ss*: number of segregating sites; *S*: number of singleton sites; *Ps*: number of informative-parsimonious sites; π: nucleotide diversity per site; SD: standard deviation. The data reported here showed that several genes had limited diversity; however, global *P. vivax* parasite populations are not equivalent and new allele variants could exist. Population genetics analysis of different populations should thus be performed for evaluating the extent of genetic diversity in natural isolates.

antigen characterisation. Natural selection signatures found in antigens could be a starting point for identifying potential new vaccine candidates (Arnott et al., 2012; Ochola et al., 2010; Suzuki, 2006; Tetteh et al., 2009; Weedall and Conway, 2010). The MK and HKA tests assume that most accumulated diversity follows the neutral model (mainly affected by demographic effects) and this would lead to identifying genes departing from this pattern as being under selection (Cornejo et al., 2014). Cornejo et al. (2014) found *P. vivax* genes under positive selection by using modified versions of the aforementioned tests; however, most genes found are not involved in host–merozoite interactions and could not therefore be used in vaccine development.

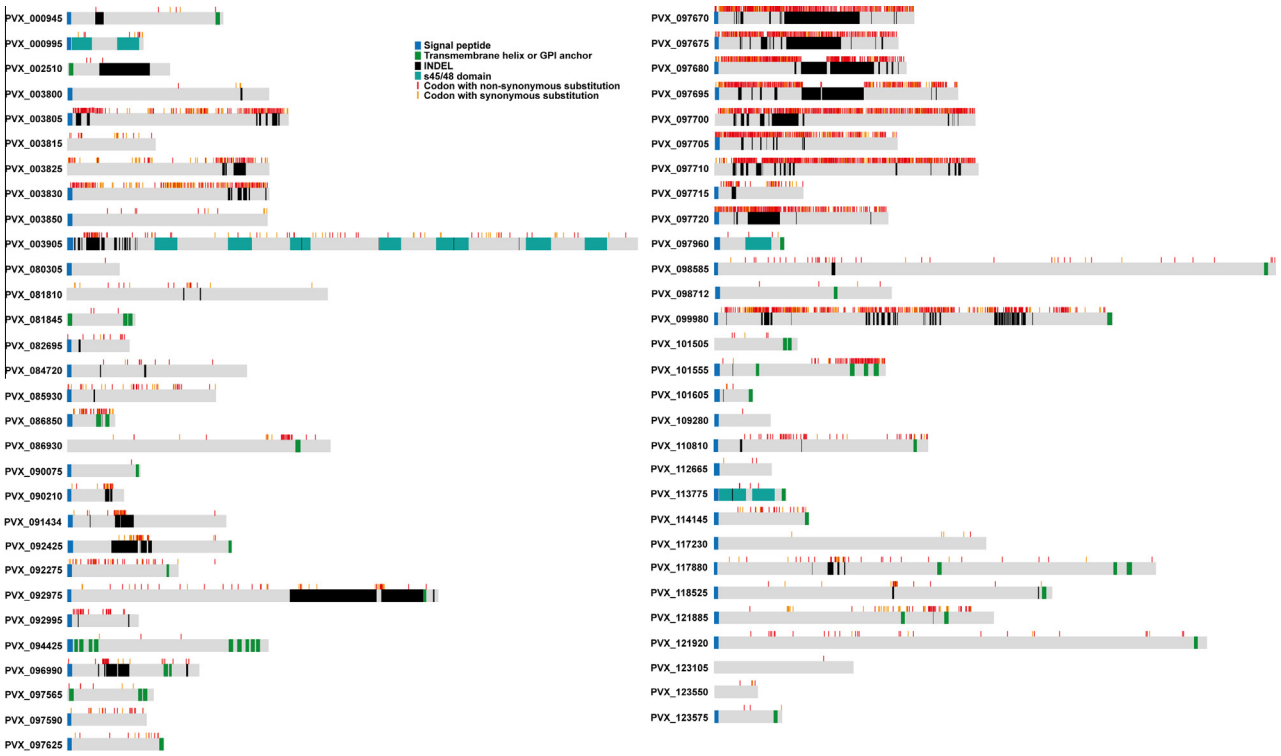


Fig. 1. Polymorphism distribution within the 59 *Plasmodium vivax* genes studied here. Codons having non-synonymous (red) and synonymous (orange) mutations are shown with vertical lines above each gene. Signal peptide-encoding regions (blue), predicted transmembrane helices or GPI anchors (green), s48/45 domains (dark cyan) and regions having INDELs (insertion and/or deletions (black)) are indicated.

The 59 genes evaluated here did not display selection signatures in the report mentioned above; however, according to the π/K ratio and conventional MK tests, 11 and 16 protein-encoding genes, respectively (Table 2 and Supplementary data 1) appeared be under balancing selection, suggesting that they are immune system targets and could thus be evaluated as vaccine candidates. However, high protein polymorphism within some of them would reduce vaccine effectiveness due allele-specific immune responses.

Only 7 genes showed balancing selection signals by both π/K and MK tests. Low correlation between them has previously been shown (Tetteh et al., 2009), suggesting that the MK test had low power for detecting balancing selection (Tetteh et al., 2009; Weedall and Conway, 2010). Such low power could be due to the effect of weak negative selection thereby leading to excess polymorphism, but just at low frequencies (Fay et al., 2002). However, our statistical results showed that the NI (or (Pn/Ps)/(Dn/Ds)) for several genes mainly reflected a high number of synonymous substitutions between species (Ds) rather than a high Pn, lowering Dn/Ds and thus increasing NI. This implied that the major factor in causing NI > 1 could was not balancing selection within species but long-term purifying selection (between species). This interpretation was supported by the fact that we also found low ω values and negative statistical significant values for the K_N–K_S test. The high synonymous substitution number found between species and the low ω values thus suggested that evolution had ensured that protein sequences remained conserved in both species by fixing substitutions (after speciation) which did not alter the amino acids; a functional/structural constraint would then have been likely. Therefore, some genes could not have been subject to balancing selection but rather negative selection. Both kinds of selection might be important for vaccine development. Balancing selection detected parasite antigens subject to immune pressure whereas negative selection identified genes having functional constraint which could avoid allele-specific responses.

According to d_N and d_S results, 12 genes had positive selection signals, 9 were under negative selection and neutrality could not be ruled out for the remaining genes. This suggested that genes such as *pvsera-5*, *pvvir-35*, members of *pvmsp-3* family, *pvrbp-1*, PVX_092995 and PVX_101555 appeared to be under immune pressure, fixing or accumulating several non-synonymous mutations as an evasion mechanism. Vaccine effectiveness would thus be reduced due to their high polymorphism.

The negatively selected genes found could be taken into account for vaccine development (Mazumder et al., 2007; Pacheco et al., 2012; Suzuki, 2004, 2006). Genes such as *pv41*, PVX_092425, *pvclag* (PVX_121885) or *pvmsp-10* could be good vaccine candidates since negative selection might be a consequence of functional constraints within them, conserving the encoded protein and therefore allele-specific immune responses could thus be avoided. Although some genes had negative selection signatures, such as *pvsera* (PVX_003825) and *pvmsp-3* (PVX_097715), they were high polymorphic at protein level, therefore making them unsuitable as candidates. This result could have been due to different selective pressure acting on the above genes. Some regions could be under negative selection but others accumulated non-synonymous substitution as an immune evasion mechanism.

Since selective pressures (or functional/structural constraint) are not the same throughout a full encoded protein sequence, several genes under selection might not have been found as the Nei–Gojobori method is a conservative test (as are the HKA and MK tests). Regardless of such difficulties, a sliding window for the d_N/d_S ratio would allow specific domains to be identified within genes where the non-synonymous and synonymous substitutions were accumulated at different rates and therefore different kinds

Table 2
Nucleotide diversity and divergence ratios (π/K), McDonald–Kreitman index (neutral index: NI) and non-synonymous (d_N) and synonymous (d_S) rates for 59 *P. vivax* genes.

#	ID	Name	Pcyn	Pkno	Pcyn		Pkno		d_N (SE)	d_S (SE)
			π/K	π/K	NI	*p*-Value	NI	*p*-Value		
Genes lacking previous population genetics analysis										
2	PVX_000995	*pv41*	0.045	0.033	3.218	0.031	2.960	0.044	0.0052 (0.0022)	0.0098 (0.0041)*
4	PVX_003800	*Pvsera*	0.003	0.002	0.781	0.831	2.476	0.507	0.0002 (0.0002)	0.0014 (0.0008)‡
5	PVX_003805	*pvsera*, putative	0.325	–	2.592	0.000	–	–	0.0729 (0.0046)	0.0725 (0.0060)
7	PVX_003825	*pvsera-4*	0.164	–	2.949	0.000	–	–	0.0257 (0.0027)	0.0347 (0.0044)†
8	PVX_003830	*pvsera-5*	–	–	–	–	–	–	0.1028 (0.0048)•	0.0791 (0.0057)
11	PVX_081810	Hypothetical protein, conserved	0.024	0.014	2.014	0.158	2.069	0.165	0.0015 (0.0005)	0.0032 (0.0011)‡
13	PVX_084720	Hypothetical protein, conserved	0.008	0.006	10.919	0.000	8.928	0.000	0.0020 (0.0007)	0.0006 (0.0005)
14	PVX_086850	*pvvir-35*, putative	0.265	–	1.335	0.552	–	–	0.0407 (0.0071)**	0.0240 (0.0081)
15	PVX_086930	*pvrhopH1/clag*	0.032	0.023	9.252	0.000	10.703	0.000	0.0048 (0.0010)	0.0029 (0.0011)
18	PVX_091434	*pvron-4*	0.021	0.016	3.262	0.024	0.649	0.465	0.0032 (0.0014)	0.0037 (0.0016)
19	PVX_092425	Hypothetical protein, conserved	0.066	0.048	0.591	0.238	0.709	0.420	0.0045 (0.0015)	0.0104 (0.0028)◇
21	PVX_092995	Tryptophan-rich antigen	0.057	–	Null	0.027	–	–	0.0142 (0.0032)•	0.0000 (0.0000)
25	PVX_097670	*pvmsp-3γ*, putative	0.756	–	1.499	0.092	–	–	0.1601 (0.0074)•	0.0906 (0.0080)
26	PVX_097675	*pvmsp-3γ*, putative	0.615	–	1.556	0.031	–	–	0.1412 (0.0072)•	0.0880 (0.0082)
27	PVX_097695	*pvmsp-3α*, putative	–	–	–	–	–	–	0.0885 (0.0048)◇	0.0689 (0.0059)
28	PVX_097700	*pvmsp-3*, putative	–	–	–	–	–	–	0.1397 (0.0054)◇	0.1160 (0.0074)
29	PVX_097705	*pvmsp-3α*, putative	–	–	–	–	–	–	0.0864 (0.0047)‡	0.0705 (0.0062)
30	PVX_097710	*pvmsp-3*, putative	–	–	–	–	–	–	0.1359 (0.0055)•	0.0956 (0.0059)
31	PVX_097715	*pvmsp-3*, putative	0.077	0.049	0.647	0.337	0.973	0.956	0.0136 (0.0024)	0.0225 (0.0050)†
33	PVX_098585	*pvrbp-1*, putative	0.014	–	2.441	0.056	–	–	0.0024 (0.0004)†	0.0008 (0.0004)
36	PVX_101555	Hypothetical protein	0.263	–	2.115	0.003	–	–	0.0401 (0.0040)•	0.0169 (0.0031)
40	PVX_113775	*pv12*	0.006	0.005	Null	0.002	Null	0.000	0.0020 (0.0010)	0.0000 (0.0000)
41	PVX_117230	*pvser/thr*	0.007	0.004	0.389	0.384	0.430	0.438	0.0001 (0.0001)	0.0018 (0.0009)◇
42	PVX_117880	*pvron-2*	0.011	0.008	4.859	0.000	4.507	0.000	0.0016 (0.0004)	0.0016 (0.0006)
43	PVX_118525	Hypothetical protein, conserved	0.014	0.009	1.125	0.877	1.479	0.395	0.0013 (0.0004)	0.0030 (0.0010)◇
44	PVX_121885	*pvclag*, putative	0.096	–	0.966	0.904	–	–	0.0059 (0.0011)	0.0095 (0.0019)◇
47	PVX_123550	Hypothetical protein, conserved	0.048	0.029	9.062	0.025	2.437	0.376	0.0016 (0.0011)	0.0059 (0.0041)
Genes having previous population genetics analysis										
1	PVX_003905	*pv230*	0.005	0.004	2.700	0.025	2.300	0.093	0.0012 (0.0004)	0.0016 (0.0006)
4	PVX_092275	*pvama-1*	0.056	0.045	6.923	0.000	4.876	0.000	0.0066 (0.0017)	0.0085 (0.0029)
7	PVX_097720	*pvmsp-3α*	0.259	–	0.571	0.030	–	–	0.0324 (0.0035)	0.0416 (0.0054)†
8	PVX_097680	*pvmsp-3β*	0.533	–	1.312	0.254	–	–	0.0819 (0.0051)†	0.0558 (0.0063)
9	PVX_099980	*pvmsp-1*	0.274	0.203	2.300	0.000	2.744	0.000	0.0485 (0.0028)	0.0464 (0.0037)
10	PVX_110810	*Pvdbp*	0.050	0.015	2.924	0.013	3.228	0.005	0.0074 (0.0012)	0.0032 (0.0013)
11	PVX_114145	*pvmsp-10*	0.010	0.006	0.289	0.252	0.801	0.856	0.0011 (0.0008)	0.0046 (0.0025)*

–: value could not be estimated because orthologous sequences were not found; null: neutral index (NI) could not be estimated due to neutral (or non-neutral) polymorphism being equal to 0; SE: standard error; Pcyn: value obtained by comparison with *Plasmodium cynomolgi*; Pkno: value obtained by comparison with *Plasmodium knowlesi*. The NI and *p*-value for *pv230*, *pvmsp-1* and *pvama-1* could not be estimated in the web server; therefore they were calculated with DnaSP, which did not consider the Jukes–Cantor divergence correction.
Only genes having a natural selection signal are shown.
* $p < 0.06$.
** $p < 0.05$.
‡ $p < 0.04$.
◇ $p < 0.01$.
† $p < 0.002$.
• $p < 0.0001$.

of selection could be assumed (e.g. positive: $\omega > 1$, negative: $\omega < 1$). Hence, although a particular gene might have either high diversity or/and positive selection, they could have functional domains under constraint and thus vaccine development should be focused on such domains (Richie and Saul, 2002).

Sliding window analysis revealed that several genes lacking significant d_N or d_S values had $\omega > 1$ domains whilst others had $\omega < 1$. According to our results, *pvama*-1 had $d_N = d_S$; however, some domains had a high d_N/d_S ratio (>2 from nucleotides 120 to 139, 370 to 434, 810 to 849, 1115 to 1174 and 1290 to 1329), thereby agreeing with a previous report (Gunasekera et al., 2007). Low K_N/K_S ratios were found throughout the entire sequence (Supplementary data 2); consequently, the non-synonymous substitutions fixed in domains having $d_N/d_S > 1$ could facilitate immune evasion while conserved regions (having low ω values) might have functional constrains due to their interaction with RON proteins (Vulliez-Le Normand et al., 2012) or host cells (Kato et al., 2005). *pv230* had neither significant d_N or d_S rates, in spite of negative selection having been reported (Doi et al., 2011). d_N/d_S values above 1 were found in this gene (nucleotides 509–605 and 1256–1280), suggesting that d_N was higher than d_S in these regions, whereas the remaining gene regions having $\omega < 1$ might have been under negative selection due to the presence of the s48/45 domains which are involved in invasion (Garcia et al., 2009).

Regarding *pvmsp-10*, the sliding window gave high divergence ($K_N/K_S > 1$) at the 5′-end unlike 3'-end which had low divergence ($K_N/K_S < 1$). It has been previously shown that polymorphism is mainly found at the 5′-end (Pacheco et al., 2012), whereas the 3′-end (encoding EGF-like domains) is high conserved within (Garzon-Ospina et al., 2011) and between (Pacheco et al., 2012) species, probably because this region is the functional one (Pacheco et al., 2012). *pvmsp-1* has similar behaviour; functional binding regions have been reported for its encoded protein (Rodriguez et al., 2002). This gene had $d_N = d_S$ and high divergence between *P. vivax* and closely-related species (Supplementary data 2). However, the peptides involved in binding to target cells mostly had low ω values (Supplementary data 4). Accord to these results (and the aforementioned ones for *pvama-1* and *pv230*), we thus hypothesise that functionally important regions (e.g. those involved in binding to target cells) would consist of fixed synonymous substitutions (after speciation) producing low ω (d_N/d_S and/or K_N/K_S) values. Thus, regions having low ω values would be highly conserved, likely due to them being under functional constraint. Consequently, these regions could be considered for vaccine development to avoid allele-specific immune responses.

The above genes have been previously evaluated by population genetics and some of them are considered vaccine candidates. We thus searched for similar patterns in genes which have not previously been studied. *pvron-2* and *pvron-4* had regions having $\omega > 1$ (1.4, nucleotides 1893–1932 and 2.4, nucleotides 766–877, respectively) suggesting that these could be targets for immune responses. However, the 3′-ends in both genes had low divergence ($\omega < 0.6$). Thus, the C-terminus of PvRON-2 and PvRON-4 proteins might have been subject to functional constraint and would thus make them potential vaccine candidates for avoiding allele-specific immune responses. Further analysis should be performed to evaluate whether these regions are involve in parasite interaction.

Genes such as *pv41*, *pvfam-d*, *pv38*, PVX_101605, *pvclag* and *pvrbp-2 like* had low diversity and had a region having d_N/d_S values around 0.4 or near to 1 (Supplementary data 3). Regarding the influenza virus, a d_N/d_S larger that 0.3 is associated with escape mutants (Suzuki, 2006); consequently, these particular domains could have been the outcome of immune pressure whereas the remaining gene sequences were conserved within and between species. An interesting pattern was observed in 6-cystein protein family members (*pv12*, *pv41* and *pv38*); they are immune system targets (Chen et al., 2010; Mongui et al., 2008; Moreno-Perez et al., 2013a) but have low genetic diversity and $d_S > d_N$ (not significant). Nevertheless, all these genes had an excess of synonymous substitutions between species, providing significant values by MK test and low K_N/K_S ratios. These results suggested that the encoded protein sequences might be subject to functional/structural constraint since there was low divergence between *P. vivax* and *P. cynomolgi* (or *P. knowlesi*); thus, proteins' biological structures encoded by these genes have been maintained in the long-term and, consequently, NI > 1 could have resulted from negative selection. Hence, similar to *pv230* (Doi et al., 2011), these 6-cystein protein family members might be subject to negative selection due to functional/structural constraint, since s48/45 domains are present, making them attractive vaccine candidates.

Proteins encoded by *pvrhopH1/clag*, *pvser/thr*, PVX_081810 and PVX_092425 genes could also be considered for a vaccine. These genes had limited diversity, displayed negative selection signatures and domains having low K_N/K_S ratios. *pvrhopH1/clag* had little divergence for almost all sequences at the 5′-end. *pvser/thr* domains having low K_N/K_S values covered conserved 323–773, 1173–1773 and 2184–3284 nucleotides whereas PVX_081810 and PVX_092425 genes had K_N/K_S values at the 3′-end. Consequently, these regions could be considered during vaccine development to avoid allele-specific immune responses. Further analysis should be performed regarding these domains.

A limitation of our approach is that specie-specific adaptation during *P. vivax* host-switch led to $K_N/K_S > 1$ and therefore functional domains could not be conserved between species (Garzon-Ospina et al., 2014) and these domains (genes) were consequently discarded by us.

5. Conclusions

Despite previous data (Cornejo et al., 2014) not having displayed selection signals in the 59 genes used here, we did identify some signatures consistent with natural selection. Members of the *pvsera* and *pvmsp-3* multigene families were subject to positive selection, likely due to the encoded proteins being targets for an immune response; however, they would not be the most appropriate ones for vaccine development due to their high polymorphism.

Proteins encoded by *pvclag*, *pvser/thr*, *pvrhopH1/clag*, *pvron-2*, *pvron-4*, *pv12*, *pv38*, *pv41*, PVX_081810 and PVX_092425 genes (or domains within them) had the patterns expected in regions having functional constrains; they could therefore be the most suitable candidates and may be prioritised for further studies (population genetics and/or functional ones) for developing a *P. vivax* vaccine which would avoid allele-specific immune responses.

Genetic diversity and evolutionary forces for *pv12*, *pv38* (Forero-Rodriguez et al., 2014a) and *pv41* (Forero-Rodriguez et al., 2014b) have been assessed recently; the data reported in such studies has agreed with the aforementioned results, suggesting that our approach provides a suitable platform for selecting potential vaccine candidates.

Acknowledgments

We would like to thank Jason Garry for translating and revising the manuscript. This work was financed by the "Colombian Science, Technology and Innovation Department (COLCIENCIAS)" through contract RC # 0309-2013.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.meegid.2015.05.001.

References

Arevalo-Pinzon, G., Curtidor, H., Abril, J., Patarroyo, M.A., 2013. Annotation and characterization of the *Plasmodium vivax* rhoptry neck protein 4 (PvRON4). Malar. J. 12, 356.

Arevalo-Pinzon, G., Curtidor, H., Patino, L.C., Patarroyo, M.A., 2011. PvRON2, a new *Plasmodium vivax* rhoptry neck antigen. Malar. J. 10, 60.

Arisue, N., Kawai, S., Hirai, M., Palacpac, N.M., Jia, M., Kaneko, A., Tanabe, K., Horii, T., 2011. Clues to evolution of the SERA multigene family in 18 Plasmodium species. PLoS ONE 6, e17775.

Arnott, A., Barry, A.E., Reeder, J.C., 2012. Understanding the population genetics of *Plasmodium vivax* is essential for malaria control and elimination. Malar. J. 11, 14.

Barry, A.E., Arnott, A., 2014. Strategies for designing and monitoring malaria vaccines targeting diverse antigens. Front. Immunol. 5, 359.

Cornejo, O.E., Fisher, D., Escalante, A.A., 2014. Genome-wide patterns of genetic polymorphism and signatures of selection in *Plasmodium vivax*. Genome Biol. Evol. 7, 106–119.

Cox, F.E., 2010. History of the discovery of the malaria parasites and their vectors. Parasit. Vectors 3, 5.

Chen, J.H., Jung, J.W., Wang, Y., Ha, K.S., Lu, F., Lim, C.S., Takeo, S., Tsuboi, T., Han, E.T., 2010. Immunoproteomics profiling of blood stage *Plasmodium vivax* infection by high-throughput screening assays. J. Proteome Res. 9, 6479–6489.

Doi, M., Tanabe, K., Tachibana, S., Hamai, M., Tachibana, M., Mita, T., Yagi, M., Zeyrek, F.Y., Ferreira, M.U., Ohmae, H., Kaneko, A., Randrianarivelojosia, M., Sattabongkot, J., Cao, Y.M., Horii, T., Torii, M., Tsuboi, T., 2011. Worldwide sequence conservation of transmission-blocking vaccine candidate Pvs230 in *Plasmodium vivax*. Vaccine 29, 4308–4315.

Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32, 1792–1797.

Egea, R., Casillas, S., Barbadilla, A., 2008. Standard and generalized McDonald–Kreitman test: a website to detect selection by comparing different classes of DNA sites. Nucleic Acids Res. 36, W157–W162.

Fay, J.C., Wyckoff, G.J., Wu, C.I., 2002. Testing the neutral theory of molecular evolution with genomic data from Drosophila. Nature 415, 1024–1026.

Forero-Rodriguez, J., Garzon-Ospina, D., Patarroyo, M.A., 2014a. Low genetic diversity and functional constraint in loci encoding *Plasmodium vivax* P12 and P38 proteins in the Colombian population. Malar. J. 13, 58.

Forero-Rodriguez, J., Garzon-Ospina, D., Patarroyo, M.A., 2014b. Low genetic diversity in the locus encoding the *Plasmodium vivax* P41 protein in Colombia's parasite population. Malar. J. 13, 388.

Garcia, J., Curtidor, H., Pinzon, C.G., Vanegas, M., Moreno, A., Patarroyo, M.E., 2009. Identification of conserved erythrocyte binding regions in members of the *Plasmodium falciparum* Cys6 lipid raft-associated protein family. Vaccine 27, 3953–3962.

Garzon-Ospina, D., Cadavid, L.F., Patarroyo, M.A., 2010. Differential expansion of the merozoite surface protein (msp)-7 gene family in Plasmodium species under a birth-and-death model of evolution. Mol. Phylogenet. Evol. 55, 399–408.

Garzon-Ospina, D., Forero-Rodriguez, J., Patarroyo, M.A., 2014. Heterogeneous genetic diversity pattern in *Plasmodium vivax* genes encoding merozoite surface proteins (MSP)-7E, -7F and -7L. Malar. J. 13, 495.

Garzon-Ospina, D., Romero-Murillo, L., Tobon, L.F., Patarroyo, M.A., 2011. Low genetic polymorphism of merozoite surface proteins 7 and 10 in Colombian *Plasmodium vivax* isolates. Infect. Genet. Evol. 11, 528–531.

Gunasekera, A.M., Wickramarachchi, T., Neafsey, D.E., Ganguli, I., Perera, L., Premaratne, P.H., Hartl, D., Handunnetti, S.M., Udagama-Randeniya, P.V., Wirth, D.F., 2007. Genetic diversity and selection at the *Plasmodium vivax* apical membrane antigen-1 (PvAMA-1) locus in a Sri Lankan population. Mol. Biol. Evol. 24, 939–947.

Jukes, T.H.a.C.R.C., 1969. Evolution of protein molecules. In: Munro, H.N. (Ed.), Mammalian Protein Metabolism. Academic Press, New York.

Kato, K., Mayer, D.C., Singh, S., Reid, M., Miller, L.H., 2005. Domain III of *Plasmodium falciparum* apical membrane antigen 1 binds to the erythrocyte membrane protein Kx. Proc. Natl. Acad. Sci. U.S.A. 102, 5552–5557.

Librado, P., Rozas, J., 2009. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics (Oxford, England) 25, 1451–1452.

Mazumder, R., Hu, Z.Z., Vinayaka, C.R., Sagripanti, J.L., Frost, S.D., Kosakovsky Pond, S.L., Wu, C.H., 2007. Computational analysis and identification of amino acid sites in dengue E proteins relevant to development of diagnostics and vaccines. Virus Genes 35, 175–186.

McDonald, J.H., Kreitman, M., 1991. Adaptive protein evolution at the Adh locus in Drosophila. Nature 351, 652–654.

Mongui, A., Angel, D.I., Guzman, C., Vanegas, M., Patarroyo, M.A., 2008. Characterisation of the *Plasmodium vivax* Pv38 antigen. Biochem. Biophys. Res. Commun. 376, 326–330.

Moreno-Perez, D.A., Areiza-Rojas, R., Florez-Buitrago, X., Silva, Y., Patarroyo, M.E., Patarroyo, M.A., 2013a. The GPI-anchored 6-Cys protein Pv12 is present in detergent–resistant microdomains of *Plasmodium vivax* blood stage schizonts. Protist 164, 37–48.

Moreno-Perez, D.A., Saldarriaga, A., Patarroyo, M.A., 2013b. Characterizing PvARP, a novel *Plasmodium vivax* antigen. Malar. J. 12, 165.

Neafsey, D.E., Galinsky, K., Jiang, R.H., Young, L., Sykes, S.M., Saif, S., Gujja, S., Goldberg, J.M., Young, S., Zeng, Q., Chapman, S.B., Dash, A.P., Anvikar, A.R., Sutton, P.L., Birren, B.W., Escalante, A.A., Barnwell, J.W., Carlton, J.M., 2012. The malaria parasite *Plasmodium vivax* exhibits greater genetic diversity than *Plasmodium falciparum*. Nat. Genet. 44, 1046–1050.

Ochola, L.I., Tetteh, K.K., Stewart, L.B., Riitho, V., Marsh, K., Conway, D.J., 2010. Allele frequency-based and polymorphism-versus-divergence indices of balancing selection in a new filtered set of polymorphic genes in *Plasmodium falciparum*. Mol. Biol. Evol. 27, 2344–2351.

Pacheco, M.A., Elango, A.P., Rahman, A.A., Fisher, D., Collins, W.E., Barnwell, J.W., Escalante, A.A., 2012. Evidence of purifying selection on merozoite surface protein 8 (MSP8) and 10 (MSP10) in Plasmodium spp. Infect. Genet. Evol. 12, 978–986.

Pain, A., Bohme, U., Berry, A.E., Mungall, K., Finn, R.D., Jackson, A.P., Mourier, T., Mistry, J., Pasini, E.M., Aslett, M.A., Balasubrammaniam, S., Borgwardt, K., Brooks, K., Carret, C., Carver, T.J., Cherevach, I., Chillingworth, T., Clark, T.G., Galinski, M.R., Hall, N., Harper, D., Harris, D., Hauser, H., Ivens, A., Janssen, C.S., Keane, T., Larke, N., Lapp, S., Marti, M., Moule, S., Meyer, I.M., Ormond, D., Peters, N., Sanders, M., Sanders, S., Sargeant, T.J., Simmonds, M., Smith, F., Squares, R., Thurston, S., Tivey, A.R., Walker, D., White, B., Zuiderwijk, E., Churcher, C., Quail, M.A., Cowman, A.F., Turner, C.M., Rajandream, M.A., Kocken, C.H., Thomas, A.W., Newbold, C.I., Barrell, B.G., Berriman, M., 2008. The genome of the simian and human malaria parasite *Plasmodium knowlesi*. Nature 455, 799–803.

Patarroyo, M.A., Calderon, D., Moreno-Perez, D.A., 2012. Vaccines against *Plasmodium vivax*: a research challenge. Expert Rev. Vaccines 11, 1249–1260.

Restrepo-Montoya, D., Becerra, D., Carvajal-Patino, J.G., Mongui, A., Nino, L.F., Patarroyo, M.E., Patarroyo, M.A., 2011. Identification of *Plasmodium vivax* proteins with potential role in invasion using sequence redundancy reduction and profile hidden Markov models. PLoS One 6, e25189.

Rice, B.L., Acosta, M.M., Pacheco, M.A., Carlton, J.M., Barnwell, J.W., Escalante, A.A., 2014. The origin and diversification of the merozoite surface protein 3 (msp3) multi-gene family in *Plasmodium vivax* and related parasites. Mol. Phylogenet. Evol. 78, 172–184.

Richie, T.L., Saul, A., 2002. Progress and challenges for malaria vaccines. Nature 415, 694–701.

Rodriguez, L.E., Urquiza, M., Ocampo, M., Curtidor, H., Suarez, J., Garcia, J., Vera, R., Puentes, A., Lopez, R., Pinto, M., Rivera, Z., Patarroyo, M.E., 2002. *Plasmodium vivax* MSP-1 peptides have high specific binding activity to human reticulocytes. Vaccine 20, 1331–1339.

Suzuki, Y., 2004. Negative selection on neutralization epitopes of poliovirus surface proteins: implications for prediction of candidate epitopes for immunization. Gene 328, 127–133.

Suzuki, Y., 2006. Natural selection on the influenza virus genome. Mol. Biol. Evol. 23, 1902–1911.

Tachibana, S., Sullivan, S.A., Kawai, S., Nakamura, S., Kim, H.R., Goto, N., Arisue, N., Palacpac, N.M., Honma, H., Yagi, M., Tougan, T., Katakai, Y., Kaneko, O., Mita, T., Kita, K., Yasutomi, Y., Sutton, P.L., Shakhbatyan, R., Horii, T., Yasunaga, T., Barnwell, J.W., Escalante, A.A., Carlton, J.M., Tanabe, K., 2012. *Plasmodium cynomolgi* genome sequences provide insight into *Plasmodium vivax* and the monkey malaria clade. Nat. Genet. 44, 1051–1055.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., Kumar, S., 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol. Biol. Evol. 28, 2731–2739.

Tetteh, K.K., Stewart, L.B., Ochola, L.I., Amambua-Ngwa, A., Thomas, A.W., Marsh, K., Weedall, G.D., Conway, D.J., 2009. Prospective identification of malaria parasite genes under balancing selection. PLoS One 4, e5568.

Vulliez-Le Normand, B., Tonkin, M.L., Lamarque, M.H., Langer, S., Hoos, S., Roques, M., Saul, F.A., Faber, B.W., Bentley, G.A., Boulanger, M.J., Lebrun, M., 2012. Structural and functional insights into the malaria parasite moving junction complex. PLoS Pathog. 8, e1002755.

Weedall, G.D., Conway, D.J., 2010. Detecting signatures of balancing selection to identify targets of anti-parasite immunity. Trends Parasitol. 26, 363–369.

Zhang, J., Rosenberg, H.F., Nei, M., 1998. Positive Darwinian selection after gene duplication in primate ribonuclease genes. Proc. Natl. Acad Sci. U.S.A. 95, 3708–3713.

Garzón-Ospina et al. BMC Evolutionary Biology (2016) 16:256
DOI 10.1186/s12862-016-0830-x

BMC Evolutionary Biology

RESEARCH ARTICLE Open Access

Evidence of functional divergence in MSP7 paralogous proteins: a molecular-evolutionary and phylogenetic analysis

Diego Garzón-Ospina[1,2], Johanna Forero-Rodríguez[1] and Manuel A. Patarroyo[1,3*]

Abstract

Background: The merozoite surface protein 7 (MSP7) is a *Plasmodium* protein which is involved in parasite invasion; the gene encoding it belongs to a multigene family. It has been proposed that MSP7 paralogues seem to be functionally redundant; however, recent experiments have suggested that they could have different roles.

Results: The *msp7* multigene family has been described in newly available *Plasmodium* genomes; phylogenetic relationships were established in 12 species by using different molecular evolutionary approaches for assessing functional divergence amongst MSP7 members. Gene expansion and contraction rule *msp7* family evolution; however, some members could have had concerted evolution. Molecular evolutionary analysis showed that relaxed and/or intensified selection modulated *Plasmodium msp7* paralogous evolution. Furthermore, episodic diversifying selection and changes in evolutionary rates suggested that some paralogous proteins have diverged functionally.

Conclusions: Even though *msp7* has mainly evolved in line with a birth-and-death evolutionary model, gene conversion has taken place between some paralogous genes allowing them to maintain their functional redundancy. On the other hand, the evolutionary rate of some MSP7 paralogs has become altered, as well as undergoing relaxed or intensified (positive) selection, suggesting functional divergence. This could mean that some MSP7s can form different parasite protein complexes and/or recognise different host receptors during parasite invasion. These results highlight the importance of this gene family in the *Plasmodium* genus.

Keywords: *Plasmodium*, Multigene family, *msp7*, Episodic positive selection, Functional divergence, Relaxed selection, Intensified selection

Background

DNA duplication is an important source of novelty regarding evolution, providing the basis for new molecular activities [1, 2]. The genomes from the three kingdoms of life have been modulated by this mechanism, having multiple copies of genes [2]. Multigene families might evolve in line with a concerted or birth-and-death evolutionary model [3]; paralogous genes keep the same function in the former due to gene conversion whilst paralogous genes could lose or acquire a new function in a birth-and-death model. Since functional importance is highly correlated with evolutionary conservation [4, 5], molecular biologists have used evolutionary approaches to infer functional changes in paralogous genes/proteins using DNA/amino acid sequences [5–8].

Gene duplication seems to be recurrent in *Plasmodium* genus. These parasites are able to infect several vertebrates such as birds, rodents and primates. More than 200 species have been described to date. Clustering in different lineages occurs according to host (bird/reptile-parasite, rodent-parasite, monkey-parasite and hominid-parasite lineages) [9, 10]. Several genes produced by gene duplication are involved in host-cell invasion [11–14]. MSP7 is a merozoite surface protein encoded by a gene belonging to a multigene family located in chromosome 13 in hominid- and rodent-parasites but in chromosome 12 in monkey-parasites. This family has a different copy number amongst *Plasmodium* species [15, 16]. These genes are

* Correspondence: mapatarr.fidic@gmail.com
[1]Molecular Biology and Immunology Department, Fundación Instituto de Inmunología de Colombia (FIDIC), Carrera 50#26-20, Bogotá, DC, Colombia
[3]School of Medicine and Health Sciences, Universidad del Rosario, Carrera 24#63C-69, Bogotá, DC, Colombia
Full list of author information is available at the end of the article

expressed simultaneously but they are independently regulated [17–20]. Functional assays have shown that *P. falciparum* MSP7 (MSP7I) is proteolytically processed; the resulting 22 kDa C-terminal region fragment is not covalently associated with MSP1 [7] and has cross-reactivity with other MSP7 proteins [17]. Furthermore, this fragment appears be involved in invasion by binding to red blood cells [21]. The C-terminal regions in *P. yoelii* from different MSP7s seem to be necessary to interact with the 83 kDa MSP1 fragment [17]. The MSP7 knockout reduces the normal growth rate of the mutant parasite in *P. berghei*; however, it becomes restored a few days later [22].

The *msp7* family appears to follow the birth-and-death evolutionary model; some gene copies have been maintained in the genome for a long time and others appear to be more recent [15]. This family has had a complex evolutionary history regarding *P. vivax*. The C-terminal region is involved in gene conversion [23]. Moreover, this region is highly conserved and under negative selection, suggesting functional/structural constraint [23–25]; by contrast, some *P. vivax* MSP7 proteins' central regions have high genetic diversity, maintained by balancing selection, possibly as an immune evasion mechanism [23, 24].

Recent protein-protein interaction assays have shown that MSP7 proteins do not appear to bind to the same host receptor [26]; moreover, these proteins seem to be forming different protein complexes in the parasite [7, 27–30], maybe to perform different parasite-host interactions. Such results flout the functional redundancy hypothesis [15, 18]; functional divergence in MSP7 paralogs thus appears to be probable. This study has analysed data concerning *msp7* multigene family evolution, including 13 available *Plasmodium* genomes by evaluating their phylogenetic relationships and adopting different and new molecular evolutionary approaches for assessing functional divergence amongst MSP7 proteins.

Methods

Sequence data, alignments and phylogenetic tree reconstruction

Genome sequences from 11 *Plasmodium* species (and one subspecies, GenBank access number: *P. reichenowi,* GCA_000723685.1; *P. falciparum,* GCA_000002765.1; *P. vivax,* GCA_000002415.2; *P. cynomolgi,* GCA_000321355.1; *P. inui,* GCA_000524495.1; *P. knowlesi,* GCA_000006355.1; *P. coatneyi,* GCA_000725905.1*; P. chabaudi,* GCA_000003075.2; *P. vinckei,* GCA_000709005.1 and GCA_000524515.1; *P. yoelii,* GCA_000003085.2 and *P. berghei,* GCA_000005395.1) as well as the partial genome sequences from *P. gallinaceum* (Wellcome Trust Sanger Institute, http://www.sanger.ac.uk/resources/downloads/protozoa/plasmodium-gallinaceum.html) were analysed to obtain *msp7* multigene family genomic regions. The *msp7* gene copy number for these 13 genomes was established, as reported previously [15, 24].

All gene sequences found were used to deduce amino acid sequences by using Gene Runner software; these sequences were then screened to distinguish the MSP_7C domain using the Pfam server [31] (domain access number: PF12948). All amino acid sequences were then aligned using the MUSCLE algorithm [32] and manually edited by GeneDoc software [33]. The best amino acid substitution model was selected by Akaike's information criterion using the ProtTest algorithm [34]; the JTT + G + F model was used to infer phylogenetic tress using maximum likelihood (ML) and Bayesian (BY) methods. RAxML was used for ML analysis [35] and topology reliability was evaluated by bootstrap, using 1000 replicates. A Metropolis-coupled Markov chain Monte Carlo (MCMC) algorithm was used for BY analysis [36] with MrBayes [37]. This analysis was run until reaching a standard deviation of split frequencies (ASDSF) value lower than 0.01; sump and sumt commands were used for tabulating posterior probabilities and building a consensus tree; in addition to ASDSF, the PSRF parameter was used for monitoring convergence. Both analyses were performed at CIPRES Science Gateway [38, 39]. A recent evolutionary multigene family model called DLRTS [40] was also performed; this method infers a gene tree by evolving down on a given species tree (with divergence times) by means of duplication, loss and transfer events according to a birth-death-like process [40, 41]. The species tree was inferred with a fragment of cytochrome C oxidase subunit 2 and divergence times were obtained from Pacheco et al. [42]. DLRTS was run using the MCMC algorithm for 10 million generations.

Gene conversion amongst *msp7* members

It has been shown that some *msp7* family members (*msp7H* and *msp7I*) have evolved by gene conversion, thereby contributing to these members' homogenisation [23]. *msp7* sequences were obtained from 6 *P. vivax* isolates (Salvador-I, Mauritania-I, India-VII, North Korean, Brazil-I and ctg isolate) [43, 44] and used to assess whether this pattern has also taken place in other *msp7* members. Betran's method was used for gene conversion amongst paralogous genes [45] as well as the GENECONV algorithm [46]. DnaSP software was used for the former method [47] where only conversion tracts larger than 10 nucleotides were considered whilst RDP3 v3.4 software was used for GENECONV [48], considering just conversion tracts having $p < 0.01$. The same approach was followed for *P. falciparum msp7* genes, using the 3D7, FCR3, RO33, 7G8, K1, T9/102 and w2mef isolates.

Identifying episodic diversifying selection on *msp7* tree branches

The random effects branch-site model (Branch-site REL) was used for assessing whether *msp7* multigene family lineages had been subject to episodic diversifying selection [49]. This method identifies branches where a percentage of sites have evolved under episodic diversifying selection. The MUSCLE algorithm was used for independently aligning each orthologous cluster's amino acid sequences (Figs. 1 and 2); PAL2NAL software [50] was then used for inferring codon alignments from the aligned amino acid sequences. The best evolutionary models for DNA and protein alignments were inferred by using jModelTest [51] and ProtTest [34], respectively. ML phylogenetic trees were obtained for DNA and protein alignments for each orthologous cluster and used as phylogenetic framework to perform the Branch-site REL method using the HyPhy software package [52]; additionally, the Datamonkey web server [53] was also used to perform this method. The MEME method [54] from Datamonkey server was used to infer which sites were under episodic positive selection in each cluster.

Evolutionary analysis for testing functional divergence

Functional redundancy was previously proposed for the MSP7 family by using sequences from seven species [15]. The sequence number in this research was increased and two different phylogeny-based approaches were used to assess functional divergence or redundancy between MSP7 members; one involved using DIVERGE v.3 software [55] to estimate type-I functional divergence [5] which is an indicator of functional changes between members of a multigene family [5, 6, 56–61]. This method is based on (site-specific) shifted evolutionary rates. It assesses whether there has been a significant change in evolution rate after duplication (or speciation) events by calculating the coefficient of divergence (θ_D) and determining (e.g. by a Likelihood ratio test [5]) whether it is statistically significant for rejecting the null hypothesis (no functional divergence). The software then computes a posterior probability for detecting amino acids responsible for such divergence. Taking into account that new functions in paralogous proteins might emerge after gene duplication whenever selective strength is relaxed or whether positive selection is intensified, a second approach was used with the RELAX method [62]. This method allows partitioning a phylogeny into two subsets of branches to determine whether selective strength was relaxed or intensified in one of these subsets (test branch) relative to the other (reference branch). The Datamonkey web server was used for this analysis.

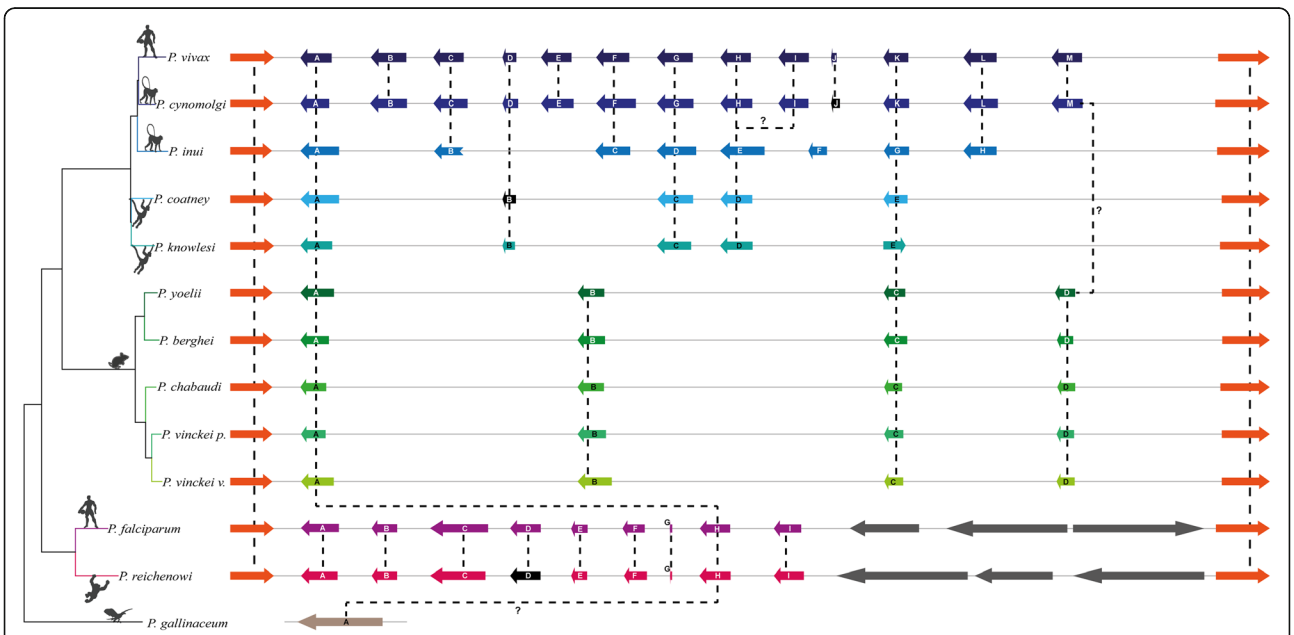


Fig. 1 Schematic representation of the chromosomal *msp7* loci in 13 *Plasmodium* genomes. The genes flanking the *msp7* chromosome region in *Plasmodium* species are represented by *orange boxes*. The *coloured boxes* within flanking sequences represent *msp7* genes in each species whilst *black boxes* symbolise pseudogenes. The genes are given in alphabetical order from left to right. The dashed lines connect orthologous genes. All genes are represented to scale, but the distance between them is not representative. Question marks refer to what were not clearly orthologous relationships. By contrast with hominid-parasites, species from monkey-parasite and rodent-parasite lineages seem to have similar evolutionary histories regarding *msp7* expansion. The *grey boxes* are lineage-specific genes only found in hominid-parasite species, as they do not belong to the *msp7* family (the latter gene representations are not to scale)

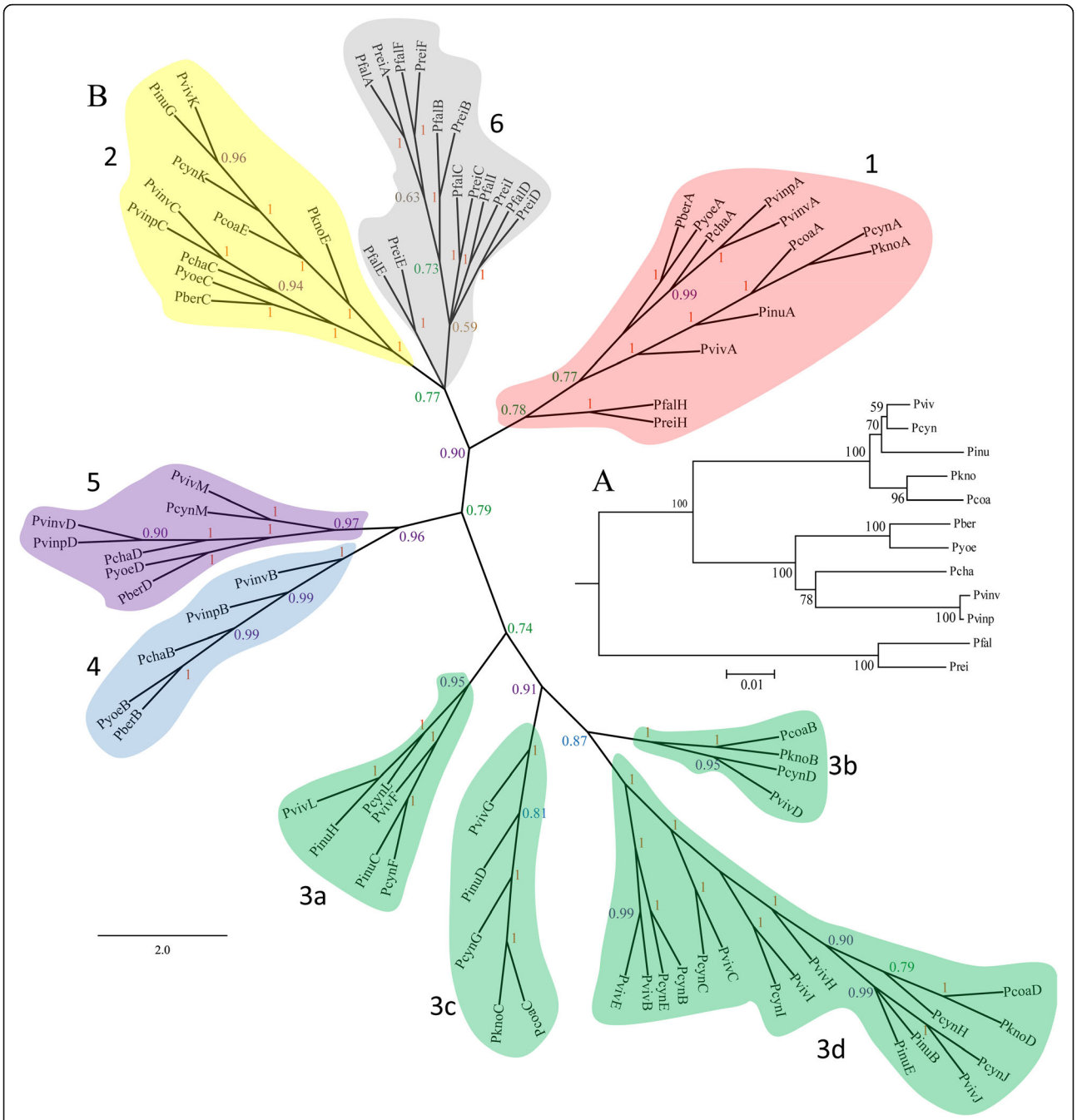


Fig. 2 *msp7* gene family phylogeny inferred by the DLTRS evolutionary model. **a** Species tree used for generating the MSP7 tree. **b** MSP7 tree created by evolving down the species tree. *Numbers* represent different clades whilst numbers on branches are posterior probability values. Nine major clades were identified on the tree. Proteins were clustered in agreement with parasite phylogenetic relationships, clades 1 (*red*), 2 (*yellow*) and 5 (*purple*) being the most ancestral ones. The clades clustering genes from monkey-parasite lineage are depicted in green, proteins from rodent-parasite lineage in blue and hominid-parasite lineage in grey. The *P. inui* specie-specific duplicate was not considered in this analysis. Due to the family's complex evolutionary history (which includes gene conversion, intragenic recombination, positive and/or balancing selection) the MCMC analysis did not converge and therefore the duplication/lost rates were not obtained even though a tree reconciliation similar to other topologies was inferred (BY and ML)

Results

msp7 chromosomal locus genetic structure in *Plasmodium* spp

Whole genome sequences from 11 *Plasmodium* species, 1 subspecies and 1 partial draft genome sequence were screened for describing the *msp7* chromosomal locus. The *msp7* locus is circumscribed by PVX_082640 and PVX_082715 genes in *P. vivax* [15, 16, 24]; genes sharing high similarity to them were thus searched in the remaining species. Since MSP7 proteins appear be encoded by a single exon [63], the contigs enclosing flanking genes were analysed using ORFFinder and Gene Runner software to identify open reading frames (ORFs) encoding proteins larger than 200 amino acids. Seventy-nine ORFs (Additional file 1) in 13 genomes having the same transcription orientation were found (Fig. 1). These ORFs had 0.6 to 1.3 kilobases (kb) but *P. gallinaceum msp7* had a 3.1 kb length. Like previous studies [15, 16, 24], the copy number was different in *Plasmodium* spp. *P. vivax* and *P. cynomolgi* had the largest copy number (12 ORFs) whilst the lowest copy number was found in *P. gallinaceum* (just one gene). Shorter fragments (having more than 30% similarity with the identified ORFs) were also found in *P. vivax*, *P. cynomolgi*, *P. falciparum* and *P. reichenowi*. These ORFs (and small fragments) were named in alphabetical order regarding PVX_082640 and its homologous genes (Fig. 1 and Additional file 1).

Data regarding *P. inui*, *msp7B* (*pinumsp7B*) was incomplete due to gaps in the contig whilst *pcynmsp7J*, *pcynmsp7L*, *pcoamspB*, *preimsp7D* and *pvinvmsp7A* had premature stop codons. However, *pcynmsp7L* could encode a full MSP7 protein since it was shown to have intron donor/acceptor sites by GeneScan [64] screening, as previously shown [24]. Despite GeneScan not showing an intron/exon structure for *pvinvmsp7A*, it has putative donor/acceptor sites (Additional file 2). The Phobius algorithm was used for determining the presence of signal peptides within ORFs and Pfam for the MSP_7C domain; some genes did not have a signal peptide or the characteristic MSP_7C domain in the C-terminal region (Table 1).

Table 1 In-silico characterisation of putative MSP7 proteins

		msp7 genes											
		A	B	C	D	E	F	G	H	I	J	K	L	M
P. vivax	SP	y	y	y	y	y	y	y	y	y	-	y	y	y
	MSP7_C	y	y	y	-	y	y	y	y	y	y	y	y	-
P. cynomolgi	SP	y	y	y	y	-	y	y	y	y	-	y	y	y
	MSP7_C	y	y	y	-	y	y	y	y	y	y	y	y	-
P. inui	SP	y	-	y	y	y	-	y	y					
	MSP7_C	y	y	y	y	y	-	y	y					
P. knowlesi	SP	y	y	y	y	y								
	MSP7_C	y	-	y	y	y								
P. coatneyi	SP	y	-	y	y	y								
	MSP7_C	y	-	y	y	y								
P. chabaudi	SP	y	y	y	y									
	MSP7_C	y	y	y	-									
P. vinckei v.	SP	y	y	y	y									
	MSP7_C	y	y	y	-									
P. vinckei p.	SP	y	y	y	y									
	MSP7_C	y	y	y	-									
P. berghei	SP	y	y	y	y									
	MSP7_C	y	y	y	-									
P. yoelii	SP	y	y	y	y									
	MSP7_C	y	y	y	-									
P. falciparum	SP	y	y	-	y	y	y	y						
	MSP7_C	y	y	y	y	y	y	y	y	y				
P. reichenowi	SP	y	-	y	y	y	y	y						
	MSP7_C	y	y	y	y	y	y	y	y	y				
P. gallinaceum	SP	y												
	MSP7_C	y												

Eighty-three sequences between flanking genes were screening for identifying a signal peptide and the characteristic MSP_7C domain (Pfam access number: PF12948). y: proteins having a signal peptide according to the Phobius algorithm or a MSP_7C domain in a Pfam search. -: proteins appeared not to have a signal peptide or MSP_7C domain

The *Plasmodium msp7* family's phylogenetic relationships

Phylogenetic relationships for this family were identified as previously described [15, 24, 65, 66]. A multiple alignment was performed for deducing 80 *msp7* genes' amino acid sequences (excluding the shorter gene fragments and *P. gallinaceum* gene). A phylogenetic tree was then inferred by using ML and BY methods with the JTT + G + F model; both topologies gave similar branch patterns, displaying 12 major clades (Additional file 3), with clade 1 clustering sequences from 12 *Plasmodium* species considered in this study, clade 2 another 10 of them and clade 5 clustering the last gene in the chromosomal region from rodent-parasites, *P. vivax* and *P. cynomolgi*. The remaining clades put together sequences according to *Plasmodium* lineages (i.e. MSP7 sequences from the monkey-parasite lineage were in clades 3, clade 4 clustered the sequences from rodent-parasite lineage and clades 6 clustered sequences from the hominid-parasite lineage). Clades 1, 2, 3a, 3b, 4 and 6b only contained orthologous proteins whilst clades 3c and 3d had orthologous and paralogous proteins from monkey-parasite lineage, the clade 6a clustered sequences from hominid-parasite lineage whilst the sequence in clade 7 appeared to be exclusive to *P. inui*. In addition to previous studies [15], the DLTRS model was implemented which reconciles the gene tree to the species tree [40, 41]. The fraction of sampled values discarded as burn-in during analysis

was 0.25 but the MCMC chain did not converge after more than 10 million generations, therefore, gene duplication, loss or transfer rates were not obtained. However, the reconciliation of the *msp7* gene tree to the *Plasmodium* species tree was obtained. The tree inferred by DLTRS had 9 major clades (Fig. 2), thereby agreeing with the ML and BY clades (Fig. 2 and Additional file 3). In all phylogenies (Fig. 2 and Additional file 3) the sequences without the MSP_7C domain (PvivMSP7D, PcynMSP7D, PknoMSP7B and PcoaMSP7B) appeared to be phylogenetically related to sequences having an MSP_7C domain. However, this relationship was only supported by posterior probabilities but not by bootstrapping (Fig. 2 and Additional file 3). Furthermore, these sequences, like others not containing an MSP_7C domain, had noticeable similarity (>48%) with MSP7 members at the N-terminal end (Additional file 4).

Since PgalMSP7A is larger than other MSP7s, we did not take it into account for the aforementioned analysis; then, only the MSP_7C domain was used for inferring their phylogenetic relationships to determine whether PgalMSP7A was orthologous to MSP7A/H. The branch pattern from this topology (Additional file 5) was similar to the phylogeny obtained by using all sequences (Fig. 2, Additional files 3 and 5). PgalMSP7A clustered with MSP7A; however, posterior probability and bootstrapping were low. PgalMSP7A did not cluster with any other MSP7 in DLTRS tree; instead it appeared as an outgroup.

Gene conversion amongst *msp7* genes

An atypical pattern in the clades clustering MSP7H/7I and MSP7B/7E/7G was observed in the phylogenies inferred above (Fig. 2, Additional files 3 and 5). Contrary to what would have been expected, PvivMSP7B was clustered with PvivMSP7E and PvivMSP7G and not with their respective orthologues; likewise, PvivMSP7H and PvivMSP7I seemed to share a common origin. It has previously been reported that gene conversion takes place between PvivMSP7H and 7I [23]. We obtained the *pvivmsp7*s sequences from 6 *P. vivax* isolates to assess whether gene conversion takes place in PvivMSP7B, 7E and 7G. Alignments for *pvivmsp7C*, *7H* and *7I* and another for *pvivmsp7B*, *7E* and *7G* were performed; Betran's method and the GENECONV algorithm were then used, displaying recombination tracks between isolates but also between paralogous genes. The Betran algorithm found 2 conversion tracks between *pvivmsp7C* and *7I*, whilst there were 3 conversion tracks between *pvivmsp7B* and *7E* and another two between *pvivmsp7E* and *7G*. Figure 3 shows the conversion tracks found by GENECONV. The same approach was followed for *P. falciparum* by using different reference isolates. By contrast with *P. vivax, pfmsp7* members did not seem to be affected by gene conversion since no conversion tracks were found amongst them (data not shown).

Episodic positive selection on *msp7* branches

Previous studies have shown ancestral positive selection regarding the *msp1* gene in the monkey-parasite lineage [67, 68]. Such episodic positive selection could have occurred in order to adapt to newly-appeared macaque species [67]. Since MSP1 and MSP7 form a protein complex [7, 18, 29] involved in parasite invasion, both proteins should have similar selective pressures. Since this has been assessed just for *msp7E* and *7L* [24], the Branch-site REL method was used here for assessing whether other lineages are subject to episodic diversifying selection in *msp7* evolutionary history; evidence was

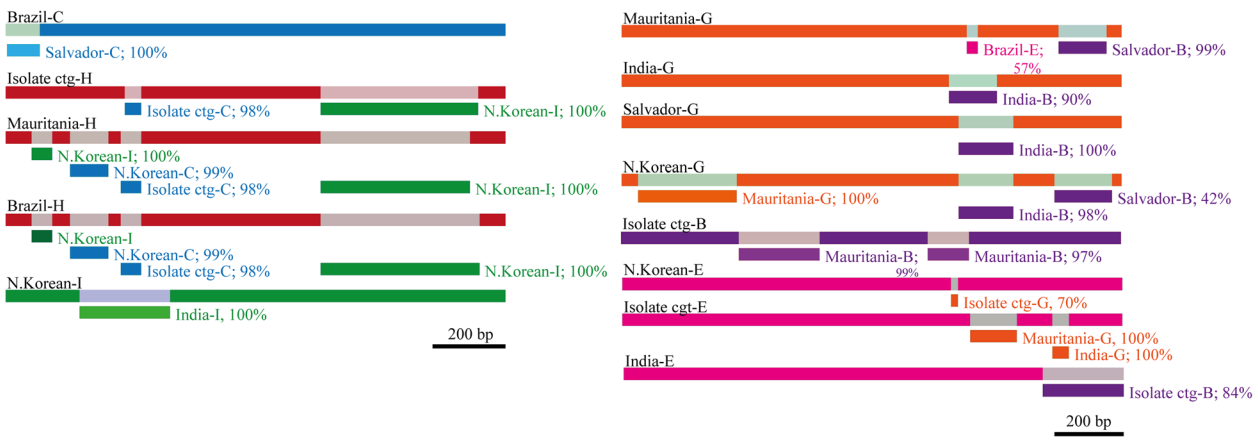


Fig. 3 Schematic representations of gene conversion tracks identified by the GENECONV method. Each gene is represented by a colour bar (*pvivmsp7C* (*blue*), *pvivmsp7H* (*red*), *pvivmsp7I* (*green*), *pvivmsp7B* (*purple*), *pvivmsp7E* (*fuchsia*), and *pvivmsp7G* (*orange*)); a different *coloured rectangle* is a graphical representation of sequence fragments potentially originating from gene conversion. Conversion tracks were mainly located in the 3′-ends. The % value refers to the similarity value for the sequence region involved in gene conversion (or intragenic recombination)

found of strong episodic diversifying selection in a few internal branches and in several external branches (Fig. 4 and Additional file 6). Regarding rodent-parasites, the lineages leading to MSP7A and 7B in *P. vinckei vinckei*; *P. vinckei petteri*, *P. chabaudi* (Fig. 4a) and *P. berghei* (Fig. 4i) were under selection. Just one internal branch (the *P. berghei*/*P. yoelii* MSP7C ancestor, Fig. 4g) displayed episodic selection. Concerning monkey-parasites, a percentage of sites regarding the lineages leading to MSP7A in *P. vivax*, *P. cynomolgi* and *P. inui* as well as the *P. inui*/*P. knowlesi*/*P. coatneyi* and *P. knowlesi*/*P. coatneyi* lineage ancestors (Fig. 4a) were under very strong episodic positive selection (ω > 33). Likewise, the lineages that gave rise to *P. cynomolgi* MSP7B, 7E, *P. vivax* MSP7B, 7E (Fig. 4b); *P. inui* 7B (Fig. 4c); *P. vivax* 7F (Fig. 4d); *P. knowlesi* 7C, *P. cynomolgi* 7G, *P. inui* 7D (Fig. 4e); *P. coatneyi* 7D, *P. knowlesi* 7D, *P. vivax* 7H, *P. inui* 7E, *P. cynomolgi* 7I (Fig. 4f) and *P. vivax* 7L (Fig. 4h) also were under

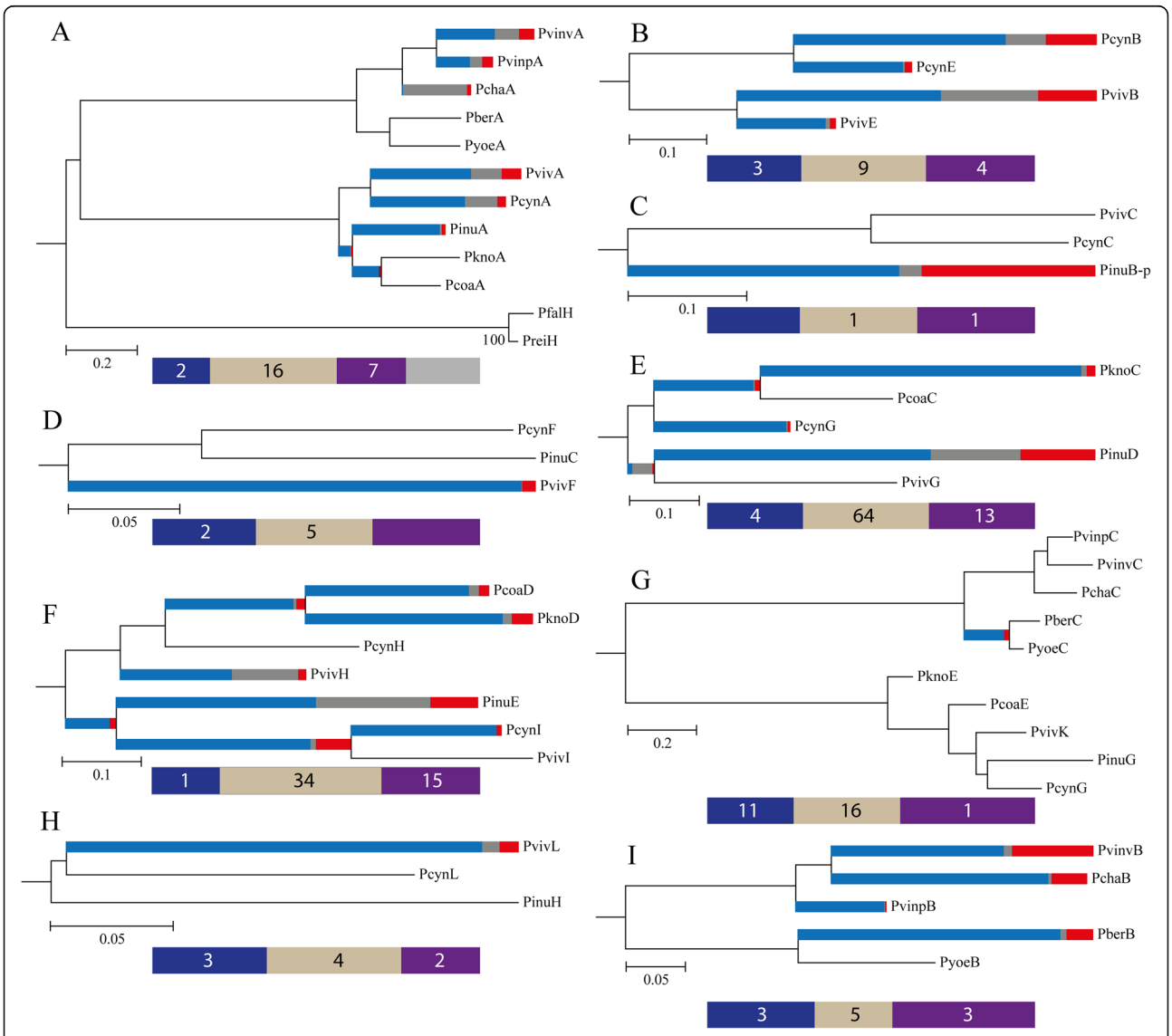


Fig. 4 Phylogenies analysed for episodic selection. Each orthologous cluster was analysed by the Branch-site REL method. The shade of each colour on branches indicates strength of selection (*red* shows ω >13, *blue* ω ≤1 and *grey* ω = 1). The size of each colour represents the percentage of sites in the corresponding class found by Branch-site REL. Branches have been classified as undergoing episodic diversifying selection by the *p*-value corrected for multiple testing using the Holm-Bonferroni method at $p < 0.05$. **a**. clade 1; **b**. *pviv/pcynmsp7B* and *7E*; **c**. *pviv/pcynmsp7C* and *pinumsp7B*; **d**. *pviv/pcynmsp7F* and *pinumsp7C*; **e**. *pviv/pcynmsp7G*, *pkno/pcoamsp7C* and *pinumsp7D*; **f**. *pviv/pcynmsp7H/7I*, *pkno/pcoamsp7D* and *pinumsp7E*; **g**. clade 2; **h**. *pviv/pcynmsp7L* and *pinumsp7H* and **i**. clade 4. At the bottom of each phylogeny there is a scale representation of *msp7s*. The *blue boxes* represent the encoded N-terminal region, the *light brown* ones symbolise the central region and the *purple boxes* the MSP_7C domain. Numbers within boxes represent the number of codons under positive selection inferred by MEME, SLAC, FEL, REL and FUBAR methods using the Datamonkey web server

positive selection. Very strong episodic positive selection was observed in the *P. knowlesi*/*P. coatneyi* 7C, *P. inui* 7D/*P. vivax* 7G, *P. coatneyi*/*P. knowlesi* 7D, *P. inui* 7E/*P. cynomolgi 7I*/*P. vivax* 7I and *P. cynomolgi*/*P. vivax* MSP7I ancestral branches (Fig. 4 and Additional file 6).

Evolutionary analysis for testing functional divergence

Gu's type-I functional divergence and RELAX methods were used to identify functional divergence amongst MSP7 proteins. Pairwise comparisons between paralogous proteins (e.g. from different clades) as well as between orthologues (e.g. proteins from different parasite lineage but within the same clade), showed high coefficient of divergence (θ_D) values (Table 2); between 10 and 20% of the sites had a significant change in their evolutionary rate (Additional file 7). Likewise, we found either a relaxed selection or an intensified selection amongst MSP7 paralogues (Table 2 and Additional file 8).

Discussion

It has been suggested that DNA duplication is the main source of evolutionary innovation [1, 2] since duplicate DNA fragments might evolve to new functions. However, acquiring new functions (neofunctionalisation) is not

Table 2 *In silico* assessment of functional divergence between paralogous and orthologous MSP7 proteins

#	Cluster A	Compared to	Cluster B	θ_D	$LRT_{\theta D}$	RELAX (*p*-value)
1	Clade 1 Primate-parasites		Clade 1 Rodent-parasites	0.54	4.74[a]	NP
2	Clade 1 Primate-parasites		Clade 3d Primate-parasites (B/E)	−0.21	0.00	Intensification (0.0038)
3	Clade 1 Rodent-parasites		Clade 3d Primate-parasites (B/E)	0.78	15.19[a]	NP
4	Clade 1 Primate-parasites		Clade 3d Primate-parasites (C/H/E/D/I)	0.34	4.25[a]	Intensification (0.00006)
5	Clade 1 Primate-parasites		Clade 3c Primate-parasites (G/C/D)	0.90	14.46[a]	Relaxation (1)
6	Clade 1 Rodent-parasites		Clade 3c Primate-parasites (G/C/D)	0.54	8.89[a]	NP
7	Clade 1 Rodent-parasites		Clade 3d Primate-parasites (C/H/E/D/I)	0.37	13.01[a]	NP
8	Clade 1 Primate-parasites		Clade 2 Primate-parasites	0.45	3.79	Relaxation (0,2)
9	Clade 1 Primate-parasites		Clade 2 Rodent-parasites	0.20	0.76	NP
10	Clade 1 Rodent-parasites		Clade 2 Primate-parasites	0.74	13.21[a]	NP
11	Clade 1 Rodent-parasites		Clade 2 Rodent-parasites	0.93	24.39[a]	Intensification (9.1e-7)
12	Clade 2 Primate-parasites		Clade 2 Rodent-parasites	0.87	12.60[a]	NP
13	Clade 1 Primate-parasites		Clade 4 Rodent-parasites	0.15	0.14	NP
14	Clade 1 Rodent-parasites		Clade 4 Rodent-parasites	0.69	22.20[a]	Relaxation (0.4)
15	Clade 3d Primate-parasites (B/E)		Clade 3d Primate-parasites (C/H/E/D/I)	0.03	0.14	Intensification (0.04)
16	Clade 3d Primate-parasites (B/E)		Clade 2 Primate-parasites	0.72	18.91[a]	Intensification
17	Clade 3d Primate-parasites (B/E)		Clade 2 Rodent-parasites	0.31	4.99[a]	NP
18	Clade 3d Primate-parasites (B/E)		Clade 4 Rodent-parasites	0.35	7.76[a]	NP
19	Clade 3d Primate-parasites (B/E)		Clade 3c Primate-parasites (G/C/D)	0.37	3.91[a]	Intensification (2.8e-7)
20	Clade 2 Primate-parasites		Clade 3d Primate-parasites (C/H/E/D/I)	0.93	37.86[a]	Intensification (0.0002)
21	Clade 3d Primate-parasites (C/H/E/D/I)		Clade 2 Rodent-parasites	0.81	20.90[a]	NP
22	Clade 3d Primate-parasites (C/H/E/D/I)		Clade 4 Rodent-parasites	0.72	15.04[a]	NP
23	Clade 2 Primate-parasites		Clade 4 Rodent-parasites	1.0	27.64[a]	NP
24	Clade 3c Primate-parasites (G/C/D)		Clade 3d Primate-parasites (C/H/E/D/I)	0.45	9.55[a]	Relaxation (0.000008)
25	Clade 2 Primate-parasites		Clade 3c Primate-parasites (G/C/D)	1.0	23.68[a]	Relaxation (0.5)
26	Clade 3c Primate-parasites (G/C/D)		Clade 2 Rodent-parasites	0.46	3.04	NP
27	Clade 3c Primate-parasites (G/C/D)		Clade 4 Rodent-parasites	0.44	3.94[a]	NP
28	Clade 4 Rodent-parasites		Clade 2 Rodent-parasites	0.57	8.56[a]	Intensification (0.03)

The coefficients of divergence (θ_D) and their LRT values from pairwise cluster comparisons in the *msp7* multigene family. $LRT_{\theta D}$ is the (log) score for the likelihood ratio test against the null hypothesis ($\theta_D = 0$) [5]. It is the output of DIVERGE and it follows a chi-square distribution with one degree of freedom; thus, values greater than or equal to 3.84 ([a]) indicate functional divergence between pairwise clusters. Selection intensity (relaxation or intensification) found by the RELAX method is shown for paralogous pairwise comparisons (see Additional file 8). Comparisons 2, 4 and 15 revealed fewer positive selected sites on test branches than on reference branches as well as an intensification of negative selected sites and non-significant θ_D. Comparisons 11, 19, 20 and 28 revealed an increased proportion of positive selected sites on the test branches, having an intensification of this kind of selection, while the proportion of negative selected sites stayed the same or decreased. The θ_D values were statistically significant. Comparisons 5 and 24 gave a statistically significant θ_D and relaxed selection (on test branches). NP: analysis was not performed because proteins came from different species

always the duplicate's outcome [69]. There are other fates for paralogous fragments such as non-functionalisation (or pseudogenisation), subfunctionalisation [69] or even functional redundancy. Encoded-protein gene duplication in parasitic organisms involved in host recognition could provide an advantage, regardless of whether such duplicates increase host recognition ability. *Plasmodium* genomes have a lot of genes as multigene families [11–14, 44], some of them are functionally equivalent whilst others have functionally diverged (they recognise different host receptors) [70–73].

The *msp7* family has been previously described in eight species [15, 16, 24], displaying different expansion. The present study has analysed 4 new species and completed analysis for *P. reichenowi*. An unequal copy gene number was found in *Plasmodium* species, suggesting lineage or specie-specific duplications or deletions. We also found genes which have become pseudogenes, thereby confirming the birth-and-death model of evolution for this family (Fig. 1). Then again, phylogenetic analysis was used for establishing phylogenetic relationships amongst *msp7* paralogues. According to the phylogenetic trees (Fig. 2 and Additional file 3), MSP7A/H (clade 1) was the most ancestral gene, followed by clade 2 which is shared by monkey- and rodent-parasites. Clade 5 was also an old clade since it is shared amongst monkey- and rodent-parasites; however, not all monkey-parasites had this copy and it thus became lost in *P. knowlesi*, *P. coatneyi* and *P. inui*. Moreover, these proteins in monkey- and rodent-parasites did not have the MSP_7C domain. The remaining clades were clustered in agreement with *Plasmodium* species' relationships (e.g. *P. vivax* genes clustered with *P. cynomolgi* genes) and they were syntenic, suggesting they are orthologues. These expansions reproduced the genus phylogenetic relationships. Species belonging to monkey-parasite lineage had the highest copy number. *P. vivax* and *P. cynomolgi* (sister taxa) shared the whole *msp7* repertory. *P. inui* is the phylogenetically closest species to the aforementioned ones, having 7 orthologues followed by *P. knowlesi* and *P. coatneyi* having 5 orthologues. The latter species are sister taxa sharing a common ancestor as well as the whole *msp7* repertory. The monkey-parasite lineage is a sister taxon to rodent-parasite lineage. At least two orthologous genes were found within these two lineages whilst only one gene was shared between these and the hominid-parasite lineage (Fig. 1). The most ancient *Plasmodium* lineage is the bird/reptile-parasite. We analysed *P. gallinaceum* and found just one large *msp7* gene. According to the MSP7 C-terminal phylogenetic tree (Additional file 5), it is still unclear whether this large gene is orthologous to the most ancestral gene (*msp7A*/*H*). As we did not find any more *msp7* genes in the *P. gallinaceum* partial genome, gene expansion should have taken place after mammal-parasite radiation 40 million year ago [42].

We found 83 sequences between bordering genes in the 13 genomes (Fig. 1 and Additional file 1); however, 11 protein sequences did not have the MSP_7C domain at the C-terminal end, though some of them did cluster with MSP7 proteins (those containing the MSP_7C domain). Proteins lacking such domain had high similarity with MSP7s at the N-terminal end (Additional file 4). This suggested that proteins lacking an MSP_7C domain (MSP7-like) are incomplete duplicates or have lost the domain throughout *Plasmodium* evolutionary history.

On the other hand, groups clustering paralogous proteins (3d clade) displayed an unusual pattern (Fig. 2 and Additional file 3). Proteins such as Pviv/PcynMSP7E and Pviv/PcynMSP7B seemed to be more similar within species than between species. A similar branch pattern was observed in the MSP_7C phylogenetic tree (Additional file 5), where PvivMSP7B was more similar to PvivMSP7E than PcynMSP7B. Likewise, PvivMSP7H was more similar to PvivMSP7I whilst PcynMSP7H clustered together with PcynMSP7I. A previous study has shown that gene conversion takes place in *pvivmsp7H* and *7I* genes [23]; such branching pattern is therefore a consequence of gene conversion. We also observed gene conversion tracks amongst *P. vivax* reference isolates in the aforementioned genes, suggesting that this mechanism also occurs in *pvivmsp7B*, *7E* and *7G*; however, such genes are not near each other (Fig. 1) and there was no complete gene homogenisation. Therefore, it is not clear whether this mechanism is taking place at present or they are ancient gene conversion events.

Parasite invasion involves several protein-protein interactions between parasite and host. MSP1 is the protein mediating initial interaction, this protein and MSP7 form a complex involved in parasite-host interaction [7, 18, 21, 29]. MSP1 has shown an episodic positive selection signal throughout its evolutionary history [67, 68]; such positive selection is shown at ancestral branches and is likely the result of adapting to newly-appeared macaque species 3.7–5.1 million years ago [67] or during human switching [74]. Since MSP7 is in a complex with MSP1 [7, 18, 29], the former should have similar selective pressures and therefore similar behaviour. We have found several lineages under strong diversifying selection ($\omega > 10$ [49], Additional file 6). As in MSP1 [67, 68], few internal branches were under episodic selection (Fig. 4a, e, f and g); this pattern could be the outcome of adaptation to new hosts during *Plasmodium* sympatric speciation, as has been suggested for other antigens [67, 68]. On the other hand, several external lineages (branches) were under selection. This could be the outcome of changes in evolutionary rates throughout *msp7* paralogous evolution which have been favoured by

selection since they may have promoted adaptation to a new host or acquiring new molecular activities.

Previous work did not recognise codons under positive selection [15]; however, here we identified codons under selection by using improved and/or newly developed methods (Fig. 4). The greatest amount of codons under positive selection was located in central *msp7* regions and a few at 5′ or 3′-ends. Population genetics studies have shown the central region to be the most polymorphic and it seems to be involved in immune evasion [23, 24]. The positive selected sites found amongst species in central regions could thus be the consequence of host adaptation to avoid host immune responses. On the other hand, positive selected sites at the encoding C-terminal region could be the outcome of coevolution between host receptor and parasite MSP7 ligands and also the result of the acquisition of new roles (Additional file 8).

Despite functional redundancy having been suggested for the MSP7 family [15, 18], some groups have shown that MSP7 proteins appear not to bind to the same host receptor [26]. Likewise, some PvivMSP7 proteins seem to be forming different protein complexes in the parasite [7, 27–30] which might allow different parasite-host interactions. We could not find evidence of functional divergence in MSP7 paralogues when comparing different clades (e.g. clade 1 against clade 2) in a previous in silico study [15]. This could have been because orthologous proteins might use different regions to interact with a host or with their own parasite proteins. The *P. falciparum* (hominid-parasite lineage) MSP1 region involved in host recognition is the 19 kDa fragment [75]; nevertheless, the 33 kDa fragment in *P. vivax* (monkey-parasite lineage) facilitates parasite-host interaction [76]. Therefore, even though they are orthologous, both proteins have differences in their evolutionary rates within functional regions [77]. Whether this behaviour also took place in MSP7, DIVERGE gave false negatives regarding functional divergence. We thus analysed MSP7 proteins from other *Plasmodium* species to assess functional divergence amongst MSP7 clades (paralogous) but also within clades (orthologous). Unlike a previous study [15], just the MSP7 C-terminal region was analysed here, taking into account that this region has the domain (MSP_7C) defining members of this family, it is the only MSP7 region in the protein complex [7], in *P. vivax* MSP7s C-terminal regions are highly conserved and under negative selection whilst other regions are highly polymorphic [23–25] and most regions binding to red blood cells are in the MSP7 C-terminal region [21]. We have found changes in evolutionary rates amongst paralogous proteins in this region. The coefficients of divergence (θ_D) were statistically significantly larger than 0 between some clades (Table 2 and Additional file 7), suggesting that some MSP7s have diverged functionally. Such divergence could mean that different MSP7 proteins could form different parasite protein complexes or that MPS7s could interact with different host receptors. We also found changes in evolutionary rates within clades (between orthologues) leading to large θ_D values (e.g. between monkey-parasite MSP7A and rodent-parasite MSP7A). This could have been due to functional divergence or different protein regions carrying out the function (as previously shown for MSP1 [75, 76]).

It has been demonstrated that duplicated genes experience a brief period of relaxed selection early in their history [69]. This relaxation could have led to pseudogenisation or, rarely, evolving to new functions [69]. Moreover, positive selection in duplicates could also lead to them acquiring new molecular activities [62]; relaxed selection or intensification of positive selection must thus be identified in MSP7s having functional divergence. Our DIVERGE results were consistent with RELAX results (Table 2). Proteins showing functional divergence also displayed functional constraint relaxation or intensification of positive selection. However, some clades having high θ_D did not show relaxation or intensification. This could have been due to episodic positive selection. This kind of selection acts very quickly and involves a switch from negative selection to positive selection and back to negative selection [62]. Episodic selection was found in all *msp7* paralogous clusters by two different approaches (Fig. 4, Additional files 6 and 7); consequently, episodic positive selection allowing functional divergence could not have been detected by RELAX.

On the other hand, some paralogous proteins had no statistical θ_D values; they also displayed intensification of negative selection, thereby suggesting that they are functionally equivalent. Furthermore, functional redundancy in MSP7H, 7I and/or 7B and 7E could be favoured by gene conversion; some MSP7 members could therefore evolve by "partial gene conversion" affecting some but not all MSP7 paralogous proteins.

Conclusion

We have described the *msp7* family in different *Plasmodium* species, using different phylogenetic and molecular evolution analyses. Although *msp7* evolved mainly in line with a birth-and-death evolutionary model, some members have evolved in a concerted way. Gene conversion has taken place between some paralogous genes allowing gene sequence homogenisation, these paralogous genes consequently keeping the same function. However, some gene conversion tracks could be ancient and thus the homogenisation has been lost. In addition, some paralogous proteins did not show changes in their evolutionary rates; thus, MSP7A, 7B, 7C, 7E, 7H and 7I in monkey-parasites seem to be functionally equivalent copies. Other MSP7 members showed alteration in their

evolutionary rate as well as relaxed or intensified (positive) selection; functional divergence may thus have occurred in them. Such functional divergence could enable MSP7F, 7G, 7K and 7L (from monkey-parasites) and MSP7A, 7B and 7C (from rodent-parasites) to form different parasite protein complexes and/or recognise different host receptors during invasion. In fact, protein-protein assays have shown that PvivMSP7A interacts with PvivTRAg56.2 [30] whilst PvivMSP7L has been found forming a complex with a member of the MSP3 family [27]. Moreover, PvivMSP7G (but not 7C or 7L) is able to bind to human P-selectin whilst PberMSP7C binds better to mouse P-selectin than PberMSP7A [26]. The results described here highlight this family's importance in the *Plasmodium* genus. Further functional assays should be performed based on these results to gain a deeper understanding of the biology of *Plasmodium* invasion.

Additional files

Additional file 1: Eighty-tree sequences from MSP7 family found in 13 *Plasmodium* genomes. The PlasmoDB accession numbers for *msp7* genes from eight species are shown. (TXT 88 kb)

Additional file 2: Putative donor/acceptor sites in *P. vinckei vinckei msp7A*. (PDF 42 kb)

Additional file 3: *msp7* gene family phylogeny inferred by the maximum likelihood method. Numbers represent different clades (based on Fig. 2 numbers from main text) whilst numbers on branches are bootstrap and posterior probability values. Thirteen major clades were identified on the tree; however, clades 1 and 6 were not clearly supported by bootstrap values. Proteins were clustered in agreement with parasite phylogenetic relationships, clades 1 (red) and 2 (yellow) being the most ancestral ones; clade 5 could also be an old cluster. The clades clustering genes from monkey-parasite lineage are depicted in green, proteins from rodent-parasite lineage in blue and hominid-parasite lineage in grey. The protein cluster in brown is a *P. inui* specie-specific duplicate. Both ML and BY trees showed similar topologies, but only the ML tree is shown. Some internal branches were inconsistent between BY and ML trees. Since *msp7* family has a complex evolutionary history (having gene conversion, intragenic recombination and/or natural selection) the tree inference is affected. However, external branches were consistent in both topologies as well as in Fig 2's tree. In BY analysis, the settings for prior and likelihood were: prset aamodel = fixed (JONES); prset statefreqpr = fixed (empirical); lset rates = gamma (JTT + G + F). More than 11 million generations were required for reaching an ASDSF value lower than 0.01. The PSRF parameter was also used for monitoring convergence (PSRF: 1.000). (TIF 3836 kb)

Additional file 4: Similarity values in the N-terminal region between MSP7 proteins and sequences lacking an MSP_7C domain. (PDF 24 kb)

Additional file 5: Phylogenetic tree inferred with the C-terminal (MSP_7C) domain. Mammal-parasite MSP7 sequences containing the MSP_7C domain and the PgalMSP7 C-terminal domain were aligned and phylogenetic trees were then inferred. PgalMSP7 clustered with the most ancestral MSP7 protein (clade 1, see Fig. 2 in the main text) though this group was not supported by bootstrap and/or posterior probability. The remaining sequences clustered according to host-parasite lineages. Numbers on branches show bootstrap and posterior probabilities values. Clades shown in red and yellow were the most ancestral ones. The clades clustering genes from monkey-parasite lineage are depicted in green, proteins from rodent-parasite lineage in blue and hominid-parasite lineage in grey. Both ML and BY trees showed similar topologies but only the ML tree is shown. Numbers outside the clades represent the number of clades in Fig. 2 from the main text. (TIF 3038 kb)

Additional file 6: Episodic positive selection on MSP7 branches. ω + values reflect the maximum likelihood estimate rate of positive selection. *p*-value obtained after Holm-Bonferroni multiple testing correction. The Branch-site REL method was performed by HyPhy software using both amino acid and DNA phylogenies. The Datamonkey web server was also used for calculating this method. The number of sites under episodic positive selection was identified by MEME using Datamonkey. The letters in the first panel correspond to the letters in Fig. 4 from the main text. (PDF 292 kb)

Additional file 7: Putative sites involved in functional divergence. DIVERGE computed a posterior probability for detecting putative amino acids responsible for functional divergence in pairwise comparisons having statistical significant θ_D values. Such putative sites were highlighted in green and sites in the C-terminal region (MSP_7C domain) under positive selection found by the MEME method were tagged with a red plus symbol (+). Since functional divergence could involve relaxation of functional constraint or could be due to positive selection, a perfect correlation between putative amino acids responsible for divergence and positive selection would not have been expected. Positive selection might be involved in the acquisition of a new role but could also be the outcome of adaptation to a new host during *Plasmodium*'s evolutionary history. Positive selected sites were inferred using the clade (e.g. sequences within clade 1) but not using the sequences from the comparison (e.g. Clade 1 Primate-parasites vs Clade 2 Primate-parasites). (PDF 478 kb)

Additional file 8: Selection intensity (functional constraint relaxation or selection intensification) throughout *msp7* paralogous genes using the partitioned descriptive model. Three ω parameters (positive, negative and neutral) and the relative proportion of sites are plotted for test (blue) and reference (red) branches. The grey vertical dashed line at ω = 1 represents neutral evolution. The numbers in parenthesis indicate the comparison number in Table 2 from the main text. Comparisons 2, 4 and 15 show that the percentage of positive selected sites decreased on test branches; they also showed an intensification of negative selected sites (on test branches), having non-significant θ_D. These results suggested functional redundancy. Comparisons 11, 19, 20 and 28 displayed an increased percentage of positive selected sites having an intensification of this kind of selection on the test branches whilst the percentage of negative selected sites remained equivalent or decreased. The θ_D values in these comparisons were statistically significant, suggesting functional divergence. Comparison 24 showed a statistically significant θ_D and relaxed selection, indicating functional divergence. (TIF 6617 kb)

Abbreviations

ASDSF: A standard deviation of split frequencies; Branch-site REL: Random effects branch-site method; BY: Bayesian method; DLTRS: Duplications, losses, transfers, rates & sequence evolution; HyPhy: Hypothesis testing using phylogenies; Kb: Kilobases; MCMC: Metropolis-coupled Markov chain Monte Carlo; MEME: Mixed effects model of evolution; ML: Maximum likelihood; MSP3: Merozoite surface protein 3; MSP7: Merozoite surface protein 7; ORFs: Open reading frames; PvivTRAg56.2: *P. vivax* tryptophan-rich antigen 56.2; RDP: Recombination detection program; θ_D: Coefficient of divergence; ω: Omega rate (d_N/d_S)

Acknowledgements

We would like to thank Jason Garry for translating and reviewing the manuscript.

Funding

This work was financed by the Departamento Administrativo de Ciencia, Tecnología e Innovación (COLCIENCIAS) through grant RC # 0309-2013.

Availability of data and materials

The datasets supporting this article's results are available in the Dryad Digital Repository (http://dx.doi.org/10.5061/dryad.3qk84) [78].

Authors' contributions

DG-O devised and designed the study, performed the molecular evolutionary analysis and wrote the manuscript. JF-R participated in designing the study, the molecular evolutionary analysis and writing the manuscript. MAP coordinated the study and helped to write the manuscript. All the authors have read and approved the final version of the manuscript.

Competing interests
The authors declare that they have no competing interests.

Consent for publication
Not applicable.

Ethics approval and consent to participate
Not applicable.

Author details
[1]Molecular Biology and Immunology Department, Fundación Instituto de Inmunología de Colombia (FIDIC), Carrera 50#26-20, Bogotá, DC, Colombia. [2]PhD Programme in Biomedical and Biological Sciences, Universidad del Rosario, Carrera 24#63C-69, Bogotá, DC, Colombia. [3]School of Medicine and Health Sciences, Universidad del Rosario, Carrera 24#63C-69, Bogotá, DC, Colombia.

Received: 15 September 2016 Accepted: 17 November 2016
Published online: 28 November 2016

References

1. Ohno S. Evolution by gene duplication. London, New York: Allen & Unwin; Springer-Verlag; 1970.
2. Zhang J. Evolution by gene duplication: an update. Trends Ecol Evol. 2003; 18(6):292–8.
3. Nei M, Rooney AP. Concerted and birth-and-death evolution of multigene families. Annu Rev Genet. 2005;39:121–52.
4. Kimura M. The neutral theory of molecular evolution. Cambridge Cambridgeshire, New York: Cambridge University Press; 1983.
5. Gu X. Statistical methods for testing functional divergence after gene duplication. Mol Biol Evol. 1999;16(12):1664–74.
6. Gaucher EA, Gu X, Miyamoto MM, Benner SA. Predicting functional divergence in protein evolution by site-specific rate shifts. Trends Biochem Sci. 2002;27(6):315–21.
7. Pachebat JA, Ling IT, Grainger M, Trucco C, Howell S, Fernandez-Reyes D, Gunaratne R, Holder AA. The 22 kDa component of the protein complex on the surface of Plasmodium falciparum merozoites is derived from a larger precursor, merozoite surface protein 7. Mol Biochem Parasitol. 2001;117(1):83–9.
8. Kondrashov FA, Rogozin IB, Wolf YI, Koonin EV. Selection in the evolution of gene duplications. Genome Biol. 2002;3(2):RESEARCH0008.
9. Duval L, Fourment M, Nerrienet E, Rousset D, Sadeuh SA, Goodman SM, Andriaholinirina NV, Randrianarivelojosia M, Paul RE, Robert V, et al. African apes as reservoirs of Plasmodium falciparum and the origin and diversification of the Laverania subgenus. Proc Natl Acad Sci U S A. 2010; 107(23):10561–6.
10. Hall N. Genomic insights into the other malaria. Nat Genet. 2012;44(9):962–3.
11. Tachibana S, Sullivan SA, Kawai S, Nakamura S, Kim HR, Goto N, Arisue N, Palacpac NM, Honma H, Yagi M, et al. Plasmodium cynomolgi genome sequences provide insight into Plasmodium vivax and the monkey malaria clade. Nat Genet. 2012;44(9):1051–5.
12. Singh V, Gupta P, Pande V. Revisiting the multigene families: Plasmodium var and vir genes. J Vector Borne Dis. 2014;51(2):75–81.
13. Sundararaman SA, Plenderleith LJ, Liu W, Loy DE, Learn GH, Li Y, Shaw KS, Ayouba A, Peeters M, Speede S, et al. Genomes of cryptic chimpanzee Plasmodium species reveal key evolutionary events leading to human malaria. Nat Commun. 2016;7:11078.
14. Gupta A, Thiruvengadam G, Desai SA. The conserved clag multigene family of malaria parasites: essential roles in host-pathogen interaction. Drug Resist Updat. 2015;18:47–54.
15. Garzon-Ospina D, Cadavid LF, Patarroyo MA. Differential expansion of the merozoite surface protein (msp)-7 gene family in Plasmodium species under a birth-and-death model of evolution. Mol Phylogenet Evol. 2010; 55(2):399–408.
16. Kadekoppala M, Holder AA. Merozoite surface proteins of the malaria parasite: the MSP1 complex and the MSP7 family. Int J Parasitol. 2010;40(10): 1155–61.
17. Mello K, Daly TM, Long CA, Burns JM, Bergman LW. Members of the merozoite surface protein 7 family with similar expression patterns differ in ability to protect against Plasmodium yoelii malaria. Infect Immun. 2004; 72(2):1010–8.
18. Mello K, Daly TM, Morrisey J, Vaidya AB, Long CA, Bergman LW. A multigene family that interacts with the amino terminus of plasmodium MSP-1 identified using the yeast two-hybrid system. Eukaryot Cell. 2002;1(6):915–25.
19. Bozdech Z, Mok S, Hu G, Imwong M, Jaidee A, Russell B, Ginsburg H, Nosten F, Day NP, White NJ, et al. The transcriptome of Plasmodium vivax reveals divergence and diversity of transcriptional regulation in malaria parasites. Proc Natl Acad Sci U S A. 2008;105(42):16290–5.
20. Zhu L, Mok S, Imwong M, Jaidee A, Russell B, Nosten F, Day NP, White NJ, Preiser PR, Bozdech Z. New insights into the Plasmodium vivax transcriptome using RNA-Seq. Sci Rep. 2016;6:20498.
21. Garcia Y, Puentes A, Curtidor H, Cifuentes G, Reyes C, Barreto J, Moreno A, Patarroyo ME. Identifying merozoite surface protein 4 and merozoite surface protein 7 Plasmodium falciparum protein family members specifically binding to human erythrocytes suggests a new malarial parasite-redundant survival mechanism. J Med Chem. 2007;50(23):5665–75.
22. Tewari R, Ogun SA, Gunaratne RS, Crisanti A, Holder AA. Disruption of Plasmodium berghei merozoite surface protein 7 gene modulates parasite growth in vivo. Blood. 2005;105(1):394–6.
23. Garzon-Ospina D, Lopez C, Forero-Rodriguez J, Patarroyo MA. Genetic diversity and selection in three Plasmodium vivax merozoite surface protein 7 (Pvmsp-7) genes in a Colombian population. PLoS One. 2012;7(9):e45962.
24. Garzon-Ospina D, Forero-Rodriguez J, Patarroyo MA. Heterogeneous genetic diversity pattern in Plasmodium vivax genes encoding merozoite surface proteins (MSP) -7E, −7 F and -7 L. Malar J. 2014;13:495.
25. Garzon-Ospina D, Romero-Murillo L, Tobon LF, Patarroyo MA. Low genetic polymorphism of merozoite surface proteins 7 and 10 in Colombian Plasmodium vivax isolates. Infect Genet Evol. 2011;11(2):528–31.
26. Perrin AJ, Bartholdson SJ, Wright GJ. P-selectin is a host receptor for Plasmodium MSP7 ligands. Malar J. 2015;14:238.
27. Hostetler JB, Sharma S, Bartholdson SJ, Wright GJ, Fairhurst RM, Rayner JC. A Library of Plasmodium vivax Recombinant Merozoite Proteins Reveals New Vaccine Candidates and Protein-Protein Interactions. PLoS Negl Trop Dis. 2015;9(12):e0004264.
28. Lin CS, Uboldi AD, Epp C, Bujard H, Tsuboi T, Czabotar PE, Cowman AF. Multiple Plasmodium falciparum Merozoite Surface Protein 1 Complexes Mediate Merozoite Binding to Human Erythrocytes. J Biol Chem. 2016; 291(14):7703–15.
29. Kauth CW, Woehlbier U, Kern M, Mekonnen Z, Lutz R, Mucke N, Langowski J, Bujard H. Interactions between merozoite surface proteins 1, 6, and 7 of the malaria parasite Plasmodium falciparum. J Biol Chem. 2006;281(42):31517–27.
30. Tyagi K, Hossain ME, Thakur V, Aggarwal P, Malhotra P, Mohmmed A, Sharma YD. Plasmodium vivax Tryptophan Rich Antigen PvTRAg36.6 Interacts with PvETRAMP and PvTRAg56.6 Interacts with PvMSP7 during Erythrocytic Stages of the Parasite. PLoS One. 2016;11(3):e0151065.
31. Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta M, Qureshi M, Sangrador-Vegas A, et al. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. 2016; 44(D1):D279–85.
32. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32(5):1792–7.
33. Nicholas KB, Nicholas HBJ. GeneDoc: A tool for editing and annotating multiple sequence alignments. 1997.
34. Abascal F, Zardoya R, Posada D. ProtTest: selection of best-fit models of protein evolution. Bioinformatics. 2005;21(9):2104–5.
35. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30(9):1312–3.
36. Altekar G, Dwarkadas S, Huelsenbeck JP, Ronquist F. Parallel Metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference. Bioinformatics. 2004;20(3):407–15.
37. Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Hohna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst Biol. 2012;61(3):539–42.
38. Miller MA, Pfeiffer W, Schwartz T. Creating the CIPRES Science Gateway for inference of large phylogenetic trees. Proceedings of the Gateway Computing Environments Workshop (GCE). New Orleans, LA; 2010. p. 1–8. http://www.phylo.org/sub_sections/portal/sc2010_paper.pdf.
39. Miller MA, Schwartz T, Pickett BE, He S, Klem EB, Scheuermann RH, Passarotti M, Kaufman S, O'Leary MA. A RESTful API for Access to Phylogenetic Tools via the CIPRES Science Gateway. Evol Bioinformatics Online. 2015;11:43–8.

40. Sjostrand J, Tofigh A, Daubin V, Arvestad L, Sennblad B, Lagergren J. A Bayesian method for analyzing lateral gene transfer. Syst Biol. 2014;63(3):409–20.
41. Sjostrand J, Sennblad B, Arvestad L, Lagergren J. DLRS: gene tree evolution in light of a species tree. Bioinformatics. 2012;28(22):2994–5.
42. Pacheco MA, Battistuzzi FU, Junge RE, Cornejo OE, Williams CV, Landau I, Rabetafika L, Snounou G, Jones-Engel L, Escalante AA. Timing the origin of human malarias: the lemur puzzle. BMC Evol Biol. 2011;11:299.
43. Carlton JM, Adams JH, Silva JC, Bidwell SL, Lorenzi H, Caler E, Crabtree J, Angiuoli SV, Merino EF, Amedeo P, et al. Comparative genomics of the neglected human malaria parasite Plasmodium vivax. Nature. 2008; 455(7214):757–63.
44. Neafsey DE, Galinsky K, Jiang RH, Young L, Sykes SM, Saif S, Gujja S, Goldberg JM, Young S, Zeng Q, et al. The malaria parasite Plasmodium vivax exhibits greater genetic diversity than Plasmodium falciparum. Nat Genet. 2012;44(9):1046–50.
45. Betran E, Rozas J, Navarro A, Barbadilla A. The estimation of the number and the length distribution of gene conversion tracts from population DNA sequence data. Genetics. 1997;146(1):89–99.
46. Sawyer S. Statistical tests for detecting gene conversion. Mol Biol Evol. 1989; 6(5):526–38.
47. Librado P, Rozas J. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics. 2009;25(11):1451–2.
48. Martin DP, Lemey P, Lott M, Moulton V, Posada D, Lefeuvre P. RDP3: a flexible and fast computer program for analyzing recombination. Bioinformatics. 2010; 26(19):2462–3.
49. Kosakovsky Pond SL, Murrell B, Fourment M, Frost SD, Delport W, Scheffler K. A random effects branch-site model for detecting episodic diversifying selection. Mol Biol Evol. 2011;28(11):3033–43.
50. Suyama M, Torrents D, Bork P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. 2006;34(Web Server issue):W609–12.
51. Darriba D, Taboada GL, Doallo R, Posada D. jModelTest 2: more models, new heuristics and parallel computing. Nat Methods. 2012;9(8):772.
52. Pond SL, Frost SD, Muse SV. HyPhy: hypothesis testing using phylogenies. Bioinformatics. 2005;21(5):676–9.
53. Delport W, Poon AF, Frost SD, Kosakovsky Pond SL. Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology. Bioinformatics. 2010;26(19):2455–7.
54. Murrell B, Wertheim JO, Moola S, Weighill T, Scheffler K, Kosakovsky Pond SL. Detecting individual sites subject to episodic diversifying selection. PLoS Genet. 2012;8(7):e1002764.
55. Gu X, Zou Y, Su Z, Huang W, Zhou Z, Arendsee Z, Zeng Y. An update of DIVERGE software for functional divergence analysis of protein family. Mol Biol Evol. 2013;30(7):1713–9.
56. Yan J, Cai Z. Molecular evolution and functional divergence of the cytochrome P450 3 (CYP3) Family in Actinopterygii (ray-finned fish). PLoS One. 2010;5(12):e14276.
57. Zhao Z, Liu H, Luo Y, Zhou S, An L, Wang C, Jin Q, Zhou M, Xu JR. Molecular evolution and functional divergence of tubulin superfamily in the fungal tree of life. Sci Rep. 2014;4:6746.
58. Wang Y, Gu X. Functional divergence in the caspase gene family and altered functional constraints: statistical analysis and prediction. Genetics. 2001;158(3):1311–20.
59. Zhou H, Gu J, Lamont SJ, Gu X. Evolutionary analysis for functional divergence of the toll-like receptor gene family and altered functional constraints. J Mol Evol. 2007;65(2):119–23.
60. McNally D, Fares MA. In silico identification of functional divergence between the multiple groEL gene paralogs in Chlamydiae. BMC Evol Biol. 2007;7:81.
61. Song W, Qin Y, Zhu Y, Yin G, Wu N, Li Y, Hu Y. Delineation of plant caleosin residues critical for functional divergence, positive selection and coevolution. BMC Evol Biol. 2014;14:124.
62. Wertheim JO, Murrell B, Smith MD, Kosakovsky Pond SL, Scheffler K. RELAX: detecting relaxed selection in a phylogenetic framework. Mol Biol Evol. 2015;32(3):820–32.
63. Mongui A, Perez-Leal O, Soto SC, Cortes J, Patarroyo MA. Cloning, expression, and characterisation of a Plasmodium vivax MSP7 family merozoite surface protein. Biochem Biophys Res Commun. 2006;351(3):639–44.
64. Burge C, Karlin S. Prediction of complete gene structures in human genomic DNA. J Mol Biol. 1997;268(1):78–94.
65. Arisue N, Kawai S, Hirai M, Palacpac NM, Jia M, Kaneko A, Tanabe K, Horii T. Clues to evolution of the SERA multigene family in 18 Plasmodium species. PLoS One. 2011;6(3):e17775.
66. Rice BL, Acosta MM, Pacheco MA, Carlton JM, Barnwell JW, Escalante AA. The origin and diversification of the merozoite surface protein 3 (msp3) multi-gene family in Plasmodium vivax and related parasites. Mol Phylogenet Evol. 2014;78:172–84.
67. Sawai H, Otani H, Arisue N, Palacpac N, de Oliveira Martins L, Pathirana S, Handunnetti S, Kawai S, Kishino H, Horii T, et al. Lineage-specific positive selection at the merozoite surface protein 1 (msp1) locus of Plasmodium vivax and related simian malaria parasites. BMC Evol Biol. 2010;10:52.
68. Muehlenbein MP, Pacheco MA, Taylor JE, Prall SP, Ambu L, Nathan S, Alsisto S, Ramirez D, Escalante AA. Accelerated diversification of nonhuman primate malarias in Southeast Asia: adaptive radiation or geographic speciation? Mol Biol Evol. 2015;32(2):422–39.
69. Lynch M, Conery JS. The evolutionary fate and consequences of duplicate genes. Science. 2000;290(5494):1151–5.
70. Orlandi PA, Klotz FW, Haynes JD. A malaria invasion receptor, the 175-kilodalton erythrocyte binding antigen of Plasmodium falciparum recognizes the terminal Neu5Ac(alpha 2–3)Gal- sequences of glycophorin A. J Cell Biol. 1992;116(4):901–9.
71. Maier AG, Duraisingh MT, Reeder JC, Patel SS, Kazura JW, Zimmerman PA, Cowman AF. Plasmodium falciparum erythrocyte invasion through glycophorin C and selection for Gerbich negativity in human populations. Nat Med. 2003;9(1):87–92.
72. Triglia T, Duraisingh MT, Good RT, Cowman AF. Reticulocyte-binding protein homologue 1 is required for sialic acid-dependent invasion into human erythrocytes by Plasmodium falciparum. Mol Microbiol. 2005;55(1):162–74.
73. Stubbs J, Simpson KM, Triglia T, Plouffe D, Tonkin CJ, Duraisingh MT, Maier AG, Winzeler EA, Cowman AF. Molecular mechanism for switching of P. falciparum invasion pathways into human erythrocytes. Science. 2005; 309(5739):1384–7.
74. Mu J, Joy DA, Duan J, Huang Y, Carlton J, Walker J, Barnwell J, Beerli P, Charleston MA, Pybus OG, et al. Host switch leads to emergence of Plasmodium vivax malaria in humans. Mol Biol Evol. 2005;22(8):1686–93.
75. Urquiza M, Rodriguez LE, Suarez JE, Guzman F, Ocampo M, Curtidor H, Segura C, Trujillo E, Patarroyo ME. Identification of Plasmodium falciparum MSP-1 peptides able to bind to human red blood cells. Parasite Immunol. 1996;18(10):515–26.
76. Rodriguez LE, Urquiza M, Ocampo M, Curtidor H, Suarez J, Garcia J, Vera R, Puentes A, Lopez R, Pinto M, et al. Plasmodium vivax MSP-1 peptides have high specific binding activity to human reticulocytes. Vaccine. 2002;20(9–10): 1331–9.
77. Parobek CM, Bailey JA, Hathaway NJ, Socheat D, Rogers WO, Juliano JJ. Differing patterns of selection and geospatial genetic diversity within two leading Plasmodium vivax candidate vaccine antigens. PLoS Negl Trop Dis. 2014;8(4):e2796.
78. Garzón-Ospina D, Forero-Rodríguez JA. PM: Data from: Evidence of functional divergence in MSP7 paralogous proteins: a molecular-evolutionary and phylogenetic analysis. *Dryad Digital Repository* 2016, http://dx.doi.org/10.5061/dryad.1q26f.

MALARIA JOURNAL

RESEARCH Open Access

Low genetic diversity and functional constraint in loci encoding *Plasmodium vivax* P12 and P38 proteins in the Colombian population

Johanna Forero-Rodríguez[1,2†], Diego Garzón-Ospina[1,3†] and Manuel A Patarroyo[1,3*]

Abstract

Background: *Plasmodium vivax* is one of the five species causing malaria in human beings, affecting around 391 million people annually. The development of an anti-malarial vaccine has been proposed as an alternative for controlling this disease. However, its development has been hampered by allele-specific responses produced by the high genetic diversity shown by some parasite antigens. Evaluating these antigens' genetic diversity is thus essential when designing a completely effective vaccine.

Methods: The gene sequences of *Plasmodium vivax p12* (*pv12*) and *p38* (*pv38*), obtained from field isolates in Colombia, were used for evaluating haplotype polymorphism and distribution by population genetics analysis. The evolutionary forces generating the variation pattern so observed were also determined.

Results: Both *pv12* and *pv38* were shown to have low genetic diversity. The neutral model for *pv12* could not be discarded, whilst polymorphism in *pv38* was maintained by balanced selection restricted to the gene's 5′ region. Both encoded proteins seemed to have functional/structural constraints due to the presence of s48/45 domains, which were seen to be highly conserved.

Conclusions: Due to the role that malaria parasite P12 and P38 proteins seem to play during invasion in *Plasmodium* species, added to the Pv12 and Pv38 antigenic characteristics and the low genetic diversity observed, these proteins might be good candidates to be evaluated in the design of a multistage/multi-antigen vaccine.

Keywords: 6-Cys, *pv12*, *pv38*, s48/45 domain, Functional constraint, *Plasmodium vivax*, Genetic diversity, Anti-malarial vaccine

Background

Malaria is a disease caused by protozoan parasites from the *Plasmodium* genus, five of which cause the disease in human beings (*Plasmodium falciparum, Plasmodium vivax, Plasmodium ovale, Plasmodium malariae and Plasmodium knowlesi*) [1,2]. This parasite is transmitted by the bite of an infected *Anopheles* female mosquito. Around 3.3 billon people are at risk of malaria annually, mainly in tropical and subtropical areas of the world, children aged less than five years and pregnant women being the most vulnerable [3]. *Plasmodium falciparum* is responsible for the disease's most lethal form, being predominantly found on the African continent whilst *P. vivax* is widely distributed around the world. Even though it has been thought that infection caused by the latter species was benign, recent studies have shown that *P. vivax* can cause clinical complications [4]. It has been found that 2,488 million people are at risk of becoming infected by *P. vivax* on the continents of Asia and America, 132 to 391 million cases occurring annually [5].

In spite of control strategies having been introduced in different countries, malaria continues to be a public health problem due to the parasite's resistance to anti-malarial treatments [6] and the vector's resistance to insecticides [7], among other causes. More effective measures have

* Correspondence: mapatarr.fidic@gmail.com
†Equal contributors
[1]Molecular Biology and Immunology Department, Fundación Instituto de Inmunología de Colombia (FIDIC), Carrera 50 No. 26-20, Bogotá, DC, Colombia
[3]School of Medicine and Health Sciences, Universidad del Rosario, Bogotá, DC, Colombia
Full list of author information is available at the end of the article

thus to be implemented for controlling such disease, including the development of an anti-malarial vaccine.

Several antigens have been characterized as promising candidates for inclusion in a vaccine [8,9], however, the genetic diversity of some of them [10-18] has hampered the development of such vaccine [19,20] as these genetic variations provoke allele-specific responses [21,22] making them become a mechanism for evading the immune system [23]. It has been necessary to focus vaccine development on conserved domains or antigens to avoid such responses [24], since these regions could have functional constraint and have had slower evolution [25].

Developing a multi-antigen vaccine against the parasite's blood stage has been focused on blocking all host-pathogen interactions to stop merozoite entry to red blood cells (RBC) [26]. A group of proteins anchored to the membrane via glycosylphosphatidylinositol (GPI) has been identified in *P. falciparum*, predominantly located in detergent-resistant membrane (DRM) domains [27,28]; they have been implicated in the parasite's initial interaction with RBC [29-33] and some have been considered as being candidates for being included in a vaccine [34,35]. One group of proteins belonging to the 6-cystein (6-Cys) family is particularly noteworthy among these DRMs (i.e., Pf12, Pf38, Pf41 and Pf92) as they have been characterized by having s48/45 domains (ID in PFAM: PF07422). Members of this family are expressed during different parasite stages [28,36] and some of them (e.g., Pf48/45, Pf230) have been considered as vaccine candidates for the sexual stage [36,37].

Pf12 and Pf38 are expressed during late stages of the intra-erythrocyte cycle, each having two high binding peptides, suggesting an active role during invasion of RBC [30]. Orthologous genes encoding these proteins have been characterized recently in *P. vivax* [38,39]. Both proteins have a signal peptide, a GPI anchor sequence and have been associated with DRMs [38,39]. Pv12 has two s48/45 domains [39] whilst Pv38 has a single domain located towards the C-terminal end [38]. These proteins have been shown to be antigenic [38-40], suggesting that they are exposed to the immune system, probably during *P. vivax* invasion of RBC.

The present study involved a population genetics analysis for evaluating the genetic diversity of *pv12* and *pv38* loci and the evolutionary processes generating this variation pattern; the results revealed these antigens' low genetic diversity in the Colombian population, possibly due to functional/structural constraints in s48/45 domains. Since the proteins encoded by these genes share structural characteristics with other vaccine candidates, added to the fact that Pv12 and Pv38 are targets for the immune response [38-40] and have conserved domains, they should be considered when designing a multistage/multi-antigen anti-malarial vaccine.

Methods

Ethics statement

The parasitized DNA used in this study was extracted from total blood collected from different Colombian areas (Antioquia, Atlántico, Bogotá, Caquetá, Cordoba, Chocó, Guainía, Guaviare, Magdalena, Meta, Nariño, and Tolima) from 2007 to 2010. All *P. vivax*-infected patients who provided blood samples were notified about the object of the study and signed an informed consent form if they agreed to participate. All procedures involved in taking blood samples were approved by Fundación Instituto de Inmunología de Colombia (FIDIC) ethics committee.

Parasitized DNA presence and integrity

Parasitized DNA presence and integrity in 100 samples stored at -20°C (2007-2010) at FIDIC (from different areas of Colombia) were evaluated by *18S* ribosomal RNA gene amplification using specific primers for *P. vivax* (SSU-F 5′-ATGAACGAGATCTTAACCTGC-3′ and SSU-R 5′-CATCACGATATGTA5TGATAAAGATTACC-3′) in a touchdown PCR [41]. The reaction contained: 1x Mango Taq reaction buffer (Bioline), 2.5 mM $MgCl_2$, 0.25 mM dNTPs, 0.5 mM of each primer, 0.1 U Mango Taq DNA polymerase (Bioline) and 10-40 ng gDNA in 10 mL final volume. The PCR thermal profile was: one initial denaturing cycle at 95°C (5 min), followed by ten cycles at 95°C (20 sec), annealing at 65°C (30 sec) and an extension step at 72°C (45 sec). Annealing temperature was reduced by 1°C in each cycle until reaching 55°C; 35 additional cycles were run at this temperature followed by a final extension cycle at 72°C (10 min). PCR products were visualized by electrophoresis on 1.5% agarose gel in 1× TAE, using 1 μL SYBR-Safe (Invitrogen).

Identifying infection caused by single *Plasmodium vivax* strain

Infection by the single *P. vivax* strain was identified by PCR-RFLP of the *pvmsp-1* polymorphic marker. The *pvmsp-1* gene fragment 2 (blocks 6, 7 and 8) was amplified using direct 5′-AAAATCGAGAGCATGATCGCC ACTGAGAAG-3′ and reverse 5′-AGCTTGTACTTTC CATAGTGGTCCAG-3′ primers [42]. The amplified fragments were digested with Alu I and Mnl I restriction enzymes, as described elsewhere [42]. The products were visualized by electrophoresis on 3% agarose gel in 1× TAE, using 1 μL SYBR-Safe (Invitrogen).

PCR amplification of *pv12* and *pv38* genes

A set of primers was designed for amplifying each of the genes based on Sal-I reference strain sequences (accession numbers in PlasmoDB: PVX_113775 for *pv12* and PVX_097960 for *pv38*). The following primers were used: for *pv12*, *pv12*-direct 5′-GTACCGCTTAACAC CGC-3′ and *pv12*-reverse 5′-GCACTACATTATAAAG

AAAAGGACC-3′ and for *pv38*, *pv38*-direct 5′-CGCT TCTTTCACCGCTTC-3′ and *pv38*-reverse 5′-CACAC ATTAACGCTGCTTCG-3′. The PCR reaction mixture contained 10 mM Tris HCL, 50 mM KCl (GeneAmp 10× PCR Buffer II [Applied Biosystems]), 1.5 mM $MgCl_2$, 0.2 mM of each dNTP, 0.5 μM of each primer, 0.76 U Amplitaq Gold DNA polymerase (Applied Biosystems) and 10-40 ng gDNA in a 50 μL final volume. The PCR thermal profile was as follows: one cycle at 95°C (7 min), 40 cycles at 95°C (20 sec), 56°C (30 sec), 72°C (1 min) and a final extension cycle at 72°C (10 min). PCR products were purified using a commercial UltraClean PCR Clean-up kit (MO BIO). The purified PCR products were sequenced in both directions with the amplification primers using the BigDye method with capillary electrophoresis, using ABI-3730 XL (MACROGEN, Seoul, South Korea). Two independent PCR products were sequenced to ensure that errors were ruled out.

Analysing genetic diversity

The electropherograms obtained by sequencing were analysed and forward and reverse sequences were assembled using CLC Main workbench software v.5 (CLC bio, Cambridge, MA, USA). The *pv12* and *pv38* genes were analysed and compared to reference sequences obtained from several sequencing projects [43,44] (accession numbers, *pv12*: XM_001616094.1, AFBK01001496.1, AFNI01000939.1, AFMK01001167.1 and AFNJ01001458.1; *pv38*: XM_001613202.1, AFNI01000834.1, AFNJ010000 90.1, AFMK01001057.1 and AFBK01001340.1) or those reported in the GenBank database (accession numbers for *pv12*: GU476521.1; and for *pv38*: JF427569.1 and JF427570.1). Gene Runner software was used for translating the sequences for deducing the amino acid sequences. These sequences were then aligned using the MUSCLE algorithm [45], and manually edited. Amino acid alignment was then used for inferring DNA using PAL2NAL software [46].

DnaSP software (v.5) [47] was used for evaluating intra-population genetic polymorphism by calculating: the number of polymorphic segregating sites (Ss), the number of singleton sites (s), the number of parsimony-informative sites (Ps), the number of haplotypes (H), haplotype diversity (Hd, which was multiplied by (n-1)/n according to Depaulis and Veuille [47,48]), the Watterson estimator (θw) and nucleotide diversity per site (π). DNA sequence variation was calculated using the sequences obtained from the aforementioned databases, plus the Colombian ones (worldwide isolates, global diversity) and just those obtained for the Colombian population (local diversity). The frequency for each Colombian haplotype was also estimated by count and year.

Two test families were used for evaluating the neutral molecular evolution model for the Colombian population: (1) frequency spectrum test, and (2) haplotype test. The former involved calculating Tajima's D statistics [49], Fu and Li's D* and F* [50] and Fay and Wu's H statistic [51]. Tajima's D statistic compares the difference between segregating sites and the average of nucleotide differences between two randomly taken sequences. Fu and Li's D* statistic takes the difference between the number of singleton sites and the total of mutations, whilst F* takes the difference between the number of singleton sites and the average of nucleotide differences between two randomly taken sequences. Fay and Wu's H statistic is based on the difference of the average number of nucleotide differences between pairs of sequences and the frequency of the derived variants. Fu's Fs statistic [52], K-test and H-test [48] are tests for calculating haplotype distribution. The Fs statistic compares the number of haplotypes observed to the expected number of haplotypes in a random sample. K-test and H-test [48] are based on haplotype number and haplotype diversity, respectively; these statistics are conditioned by sample size (n) and the number of segregating sites (Ss). Test significance was determined by coalescence simulations using DnaSP (v.5) [47] and ALLELIX software (kindly supplied by Dr Sylvain Mousset). Sites having gaps were not taken into account in any of the tests performed.

The effect of natural selection was evaluated regarding intra and interspecies; the average number of non-synonymous substitutions per non-synonymous site (d_N) and the average number of synonymous substitutions per synonymous site (d_S) were calculated for the former by using the modified Nei-Gojobori method [53]. The significant differences between the above were determined by using Fisher's exact test (suitable for d_N and $d_S < 10$) and codon-based Z-test incorporated in MEGA software (v.5) [54]. Differences between d_N and d_S per site were calculated by using SLAC, FEL, REL [55], IFEL [56], MEME [57], and FUBAR [58] methods. The average number of non-synonymous divergence substitutions per non-synonymous site (K_N) and the average number of synonymous divergence substitutions per synonymous site (K_S) were calculated using the modified Nei-Gojobori method [53], with Jukes-Cantor correction [59], to infer natural selection signals which may have prevailed during malarial parasite evolutionary history (interspecies; using *Plasmodium cynomolgi* (accession number BAEJ 01001076.1) and *P. knowlesi* (accession number NC_011 912.1) orthologous sequences). The significant differences between K_N and K_S were determined by using a codon-based Z-test incorporated in MEGA software (v.5) [54]. The McDonald-Kreitman test [60] was also calculated; this is based on a comparison of intraspecific polymorphism to interspecific divergence (using *Plasmodium cynomolgi* (accession number BAEJ01001076.1) and *P. knowlesi* (accession number NC_011912.1) orthologous

sequences). This test involved using a web server [61], which takes Jukes-Cantor divergence correction into account [59]. All the above tests were calculated using the sequences obtained from the databases plus the Colombian ones and just those obtained for the Colombian population.

Z_{nS} [62] and ZZ [63] statistics were calculated for evaluating the influence of linkage disequilibrium (LD) and intragenic recombination, respectively. The minimum number of recombination (Rm) events was also calculated; this included calculating effective population size and the probability of recombination between adjacent nucleotides per generation [64]. Additionally, the GARD method [65] available at the Datamonkey web server [66] was performed. These tests were performed using the sequences obtained from the Colombian population.

Results and discussion

The presence of genomic DNA (gDNA) and identification of single *Plasmodium vivax* strain infection

An *18S* subunit rRNA gene fragment was amplified from 100 samples of *P. vivax* collected from different areas of Colombia and stored from 2007 to 2010. Seventy-seven samples revealed an amplicon at the expected size, indicating the presence of *P. vivax* gDNA. A region of the *pvmsp-1* gene was then amplified and digested with restriction enzymes, showing that seven of the 77 samples proving positive for *P. vivax* had multiple infections. Only 70 samples were thus considered for later analysis. Due to the low number of samples collected from some areas, they were grouped according to geographical localisation and epidemiological conditions (South-west: Chocó, Nariño; South-east: Caquetá, Guainía, Guaviare, Meta; Midwest: Bogota, Tolima; North-west: Atlántico, Antioquia Cordoba, Magdalena).

Genetic diversity in *pv12*

Seventy samples amplified a 1,200 base pair (bp) fragment corresponding to the *pv12* gene (South-west n = 6; South-east: n = 20; Midwest: n = 8; North-west: n =36). These amplicons were purified and sequenced; the sequences were then analysed, compared to different reference sequences obtained from various sequencing projects [43,44] and those having a different haplotype were deposited in the GenBank database (accession numbers KF667328 and KF667329).

Four single nucleotide polymorphisms (SNP) were observed throughout the *pv12* gene sequence (Figure 1A) located in positions 375 (N125K), 379 (T127A), 539 (L180W) and 662 (N221S). Only one SNP (nucleotide 375) was found in the Colombian population. A repeat region was observed; it was formed by previously reported amino acids N[A/V][H/Q] [39], in which an insertion was observed in the North Korean sequence (Figure 1A, haplotype 1) and deletions in the Colombian sequences (Figure 1A, haplotypes 2 and 3).

Six haplotypes were found in *pv12* (Figure 1A and Table 1) around the world, four of which are present in Colombia at 8.7, 5.8, 10.1, and 75.4% frequency for haplotypes 2, 3, 5 and 6, respectively. Haplotypes 2, 5 and 6 were present in the different Colombian locations (Additional file 1), haplotype 6 being the most predominant per year (2007 n = 9; 2008 n = 17; 2009 n = 15; 2010 n = 29) and per location, having higher than 70% frequency (Figure 2A and Additional file 1). The remaining haplotypes were absent or had low frequency (Figure 2A and Additional file 1). Interestingly, haplotype 3 was present in Colombia during 2009 but absent in the other years studied (Figure 2A). The percentage of samples from the South-east area (some of them presenting haplotype 3) was greater than for other years, suggesting that haplotype 3 was restricted to a particular geographical area (Additional file 1) and/or that this had very low frequency in different Colombian subpopulations. Haplotype 2 was absent from 2007 to 2008 but present between 2009 and 2010 (Figure 2A); differently to haplotype 3, this haplotype was present everywhere, except in the South-west location (Additional file 1). This appeared to be consistent with previous studies which have reported numerous private haplotypes in American *Plasmodium vivax* populations [67]. These results suggested that the Colombian population had one predominant *pv12* haplotype and several low frequency alleles, which are geographically isolated or were not detected during some periods of time. Since *P. vivax* populations within countries seem to be strongly structured [67], new *pv12* haplotypes could appear in other parasite populations.

This gene had 0.0004 ± 0.0001 global nucleotide diversity (π) and 0.0003 ± 0.0001 for the Colombian population (Table 1). This value was about 2.5 times less than that reported for its orthologue in *P. falciparum* (π = 0.001) [68]; however, both values were low when compared to other membrane proteins [10-14,17], suggesting that this gene is highly conserved in different *Plasmodium* species. This value places *pv12* among the most conserved antigen-encoding genes characterized to date in *P. vivax*.

Mutations in *pv12* appear to be selectively neutral

Several tests for evaluating the hypothesis that mutations in *pv12* are neutral were performed. No significant values were found for the Tajima, Fu and Li, Fay and Wu or Fu tests (Table 2); likewise, the Colombian population's number of haplotypes (4) and haplotype diversity (0.406 ± 0.07) (Table 2) were as expected under neutrality according to the K-test and H-test. Since neutrality could not be ruled out, the mutations or haplotypes found in *pv12* could have been randomly fixed; this might explain the

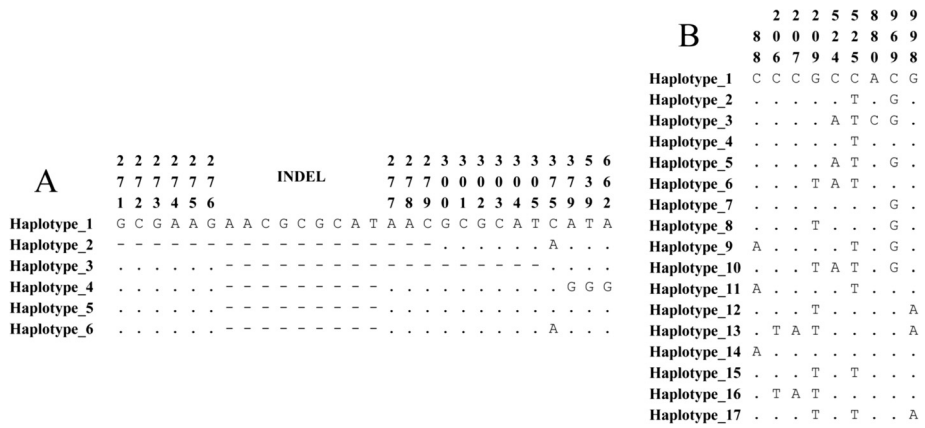


Figure 1 *pv12* and *pv38* haplotype alignment. A. Alignment of the six haplotypes found in the *pv12* gene. Haplotypes 2, 3, 5 and 6 were found in the Colombian population. **B.** Alignment of the 17 haplotypes found in *pv38*, 14 of them found in the Colombian population (haplotypes 1, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, and 17). The dots indicate nucleotide identity and dashes indicate nucleotide absence. Numbering is based on the Sal-I reference sequence.

possible geographical isolation of haplotype 3, since different alleles could have become fixed in different populations according to the neutral hypothesis. Alternatively, the geographical isolation of haplotype 3 could have resulted from the structured *P. vivax* population in America, where haplotypes may have diversified *in situ*.

Natural selection in *pv12*

The gene was split into two regions: region A, nucleotides 1-546 (amino acids 1-182 including one s48/45 domain) and region B, nucleotides 547-1,095 (amino acids 183-365 including the other s48/45 domain). Synonymous substitution per synonymous site (d_S) and non-synonymous substitution per non-synonymous site rates (d_N) were calculated using the gene's total length to evaluate whether natural selection had any effect on *pv12* evolution. Full length gene and split regions had non-significant values (Table 3); likewise, when d_N and d_S were estimated for s48/45 domains, no significant values were observed (Additional file 2), contrary to that suggested for *pf12*, where purifying selection action has been reported [69]. The Datamonkey server was used for calculating d_N and d_S rates for each codon; no selected sites were found, indicating (once more) that the gene did not appear to deviate from neutrality.

Table 1 Estimators for *pv12* and *pv38* global and local genetic diversity

n	Gene	Sites	Ss	S	Ps	H	θw (sd)	π (sd)
Worldwide isolates								
76	***pv12***	927	4	3	1	6	0.0009 (0.0005)	0.0004 (0.0001)
53	***pv38***	1,035	9	1	8	17	0.0019 (0.0006)	0.0026 (0.0002)
Colombian population								
70	***pv12***	1,047	1	0	1	4	0.0002 (0.0002)	0.0003 (0.0001)
46	***pv38***	1,062	8	0	8	14	0.0017 (0.0006)	0.0024 (0.0002)

Estimators of genetic diversity were calculated using the sequences obtained from databases plus the Colombian ones (worldwide isolates, global diversity) and just those obtained for the Colombian population (Colombian population, local diversity). n: number of isolates, sites: total of sites analysed excluding gaps, Ss: number of segregating sites, S: number of singleton sites, Ps: number of informative-parsimonious sites, H: number of haplotypes, θ_w: Watterson estimator, π: nucleotide diversity per site. sd: standard deviation.

However, assessing how natural selection acts on low genetic diversity antigens is not easy [70]; the fact that *Plasmodium vivax* shares its most recent common ancestor with parasites infecting primates (e.g. *P. cynomolgi* and *P. knowlesi*) led to inferring patterns which may have prevailed during their evolutionary history [70,71]. When synonymous divergence substitution per synonymous site (K_S) and non-synonymous divergence substitution per non-synonymous site (K_N) rates were calculated, a significantly higher K_S than K_N was found (Table 4). Moreover, a sliding window for ω (d_N/d_S and/or K_N/K_S) revealed < 1 values throughout the gene (Figure 3), which could have been a consequence of negative selection. Moreover, significant values were observed when the McDonald-Kreitman (MK) test was used for comparing intraspecific polymorphism and interspecific divergence (using all the haplotypes found for this gene): $P_N/P_S > D_N/D_S$ (Table 5), revealing (similar to K_S rates) a large accumulation of synonymous substitutions between species, which could be interpreted as negative selection. Such accumulation of interspecies synonymous substitutions suggested that evolution tried to maintain protein structure by eliminating all deleterious mutations. However, when the MK test was done with haplotypes found in Colombia (and in spite of the accumulation of synonymous substitutions between species), no significant values were observed in this population (Table 5). Although Pv12 is exposed to the immune system [39,40], it had a high level of conservation. This pattern could have been because *pv12* had diverged

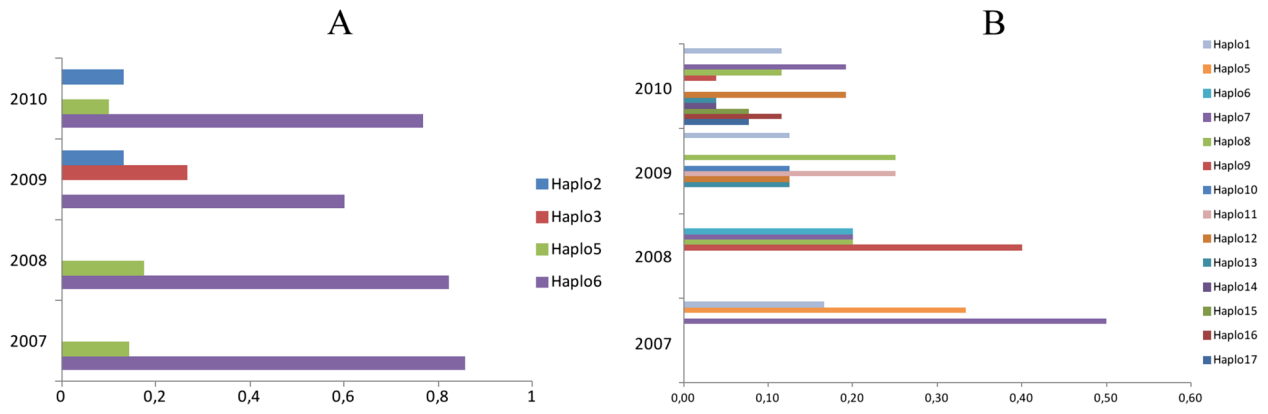


Figure 2 Frequency of the *pv12* and *pv38* haplotypes present in the Colombian population. A. Frequency per year for the four *pv12* haplotypes found in the Colombian population: n = 9 in 2007, n = 17 in 2008, n = 15 in 2009 and n = 29 in 2010. **B.** Frequency per year for the 14 *pv38* haplotypes found in the Colombian population. n = 6 in 2007, n = 6 in 2008, n = 8 in 2009 and n = 26 in 2010.

by negative selection, due to a possible functional/structural constraint imposed by the presence of s48/45 domains [72] which seem to play an important role during host cell recognition [30,69,72].

Genetic diversity in *pv38*

Only 46 out of 70 samples could be amplified for the *pv38* gene, giving a 1,121 bp fragment (South-west n = 6; South-east: n = 13; Midwest: n = 4; North-west: n = 23). The 46 sequences obtained from Colombian isolates were compared to and analysed regarding reference sequences obtained from different regions of the world [43,44]. Colombian sequences that have a different haplotype to that of previously reported ones can be found in GenBank (accession numbers KF667330-KF667340).

Nine SNPs were observed in the *pv38* gene (Figure 1B), most of which were no-synonymous (nucleotides: 88 (R30S), 206/207 (A69V), 209 (R70L), 524/525 (T175N), 880 (M294L), and 998 (S333N)), similar to that found in *Pf38* [73]. Positions 525 and 969 produced synonymous substitutions (a change in protein sequence was generated when the substitution in position 525 was accompanied with another one in position 524). The parasite population in Colombia has eight of these nine SNPs, all being informative-parsimonious sites. Similar to that reported for its orthologue in *P. falciparum* [73], most substitutions were found in the gene's 5′ region.

Seventeen haplotypes were identified from alignment (including sequences from different regions of the world) (Figure 1B), 14 of which were found in Colombia's parasite population at different frequencies: 11% haplotype 1, 4% haplotype 5, 2% haplotype 6, 20% haplotype 7, 15% haplotype 8, 7% haplotype 9, 2% haplotype 10, 4% haplotype 11, 13% haplotype 12, 4% haplotype 13, 2% haplotype 14, 4% haplotype 15, 7% haplotype 16, and 4% haplotype 17. Most haplotypes were found in intermediate frequencies per year (2007 n = 6; 2008 n = 6; 2009 n = 8; 2010 n = 26) and none exceeded 50% (Figure 2B). The absence of some haplotypes in determined years, or in some locations, could not just have been due to the low frequency which they might have had but also to the difference in the number of samples for each year (n = 6 in 2007, n = 6 in 2008, n = 8 in 2009 and n = 26 in 2010) or because American *P. vivax* populations appear to be structured and therefore several privative haplotypes might be found [67].

π in this gene was 0.0026 ± 0.0002 worldwide and 0.0024 ± 0.0002 in the Colombian population (Table 1), this being 1.3 times lower than that for its orthologue in *P. falciparum* (π = 0.0034) [68,73] showing that the *pv38*

Table 2 *pv12* and *pv38* neutrality, linkage disequilibrium and recombination tests for the Colombian population

N	Gene	Tajima	Fu and Li		Fay and Wu's H	Fu's Fs	K-test	H-test (sd)	Z_{ns}	ZZ	RM
		D	D*	F*							
70	***pv12***	0.365	0.516	0.548	0.000	0.902	4	0.406 (0.07)	ND	ND	0
46	***pv38***	1.147	1.304	1.473	−1.275	−4.451	14*	0.890 (0.02)*	0.107	0.125	2

n: number of isolates.
*: $p < 0.05$.
ND: not determined.
sd: standard deviation.

Table 3 Synonymous substitution per synonymous site rate (d_S) and non-synonymous substitution per non-synonymous site rate (d_N) for *pv12* and *pv38* genes

n	Gene	Region A		Region B		Full length gene	
		d_S (se)	d_N (se)	d_S (se)	d_N (se)	d_S (se)	d_N (se)
Worldwide isolates							
76	***pv12***	0.000 (0.000)	0.001 (0.001)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.001 (0.000)
53	***pv38***	0.001 (0.001)	0.003 (0.002)	0.006 (0.004)	0.001 (0.001)	0.004 (0.002)	0.002 (0.001)
Colombian population							
70	***pv12***	0.000 (0.000)	0.001 (0.001)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
46	***pv38***	0.001 (0.001)	0.003 (0.002)	0.005 (0.004)	0.001 (0.001)	0.004 (0.002)	0.002 (0.001)

d_N and d_S rates were estimated by using sequences obtained from databases together with Colombian ones (worldwide isolates) and just with those obtained in the Colombian population. n: number of isolates. *pv12*: region A, nucleotides 1-546 and region B, nucleotides 547-1,095. *pv38*: region A, nucleotides 1-459 and region B, nucleotides 460-1,065. se: standard error. No statistically significant differences were found.

gene had low diversity, at least in the two main species affecting human beings.

Deviation from the neutral model of molecular evolution in *pv38*

Tajima's D, Fu and Li's D* and F*, Fay and Wu's H and Fu's Fs neutrality tests did not reveal statistically significant values (Table 2), suggesting that the gene might follow the neutral evolution model. However, the presence of 14 haplotypes and 0.890 ± 0.02 haplotype diversity in the Colombian population was greater than that expected under neutrality according to K-test and H-test results (Table 2). This suggested balanced ancestral polymorphism [48], being similar to that reported for the *P. falciparum p38* gene which showed evidence of balanced selection in 5' region [73].

Natural selection in *pv38*

A modified Nei Gojobori method was used for calculating d_N and d_S rates for showing some type of selection in the *pv38* gene. Similar to that used regarding *pv12*, the *pv38* gene was divided into two regions: region A, covering position 1-459 (amino acids 1-153) and region B, nucleotides 460-1,065 (amino acids 154-355 including

Table 4 Synonymous divergence substitution per synonymous site (K_S) rate and non-synonymous divergence substitution per non-synonymous site (K_N) rate

		P. vivax/P. Cynomolgi					
n	Gene	s48/45 domain in region A		s48/45 domain in region B		Full-length gene	
		K_S (se)	K_N (se)	K_S (se)	K_N (se)	K_S (se)	K_N (se)
Worldwide isolates							
78	***pv12***	0.016 (0.003)†	0.005 (0.002)	0.019 (0.004)†	0.003 (0.001)	0.016 (0.002)*	0.004 (0.001)
54	***pv38***	-		0.030 (0.007)†	0.005 (0.001)	0.031 (0.004)*	0.007 (0.001)
Colombian isolates							
71	***pv12***	0.018 (0.003)†	0.005 (0.001)	0.021 (0.004)†	0.003 (0.001)	0.016 (0.002)*	0.005 (0.001)
47	***pv38***	-		0.033 (0.007)†	0.006 (0.001)	0.033 (0.004)*	0.008 (0.001)
		P. vivax/P. knowlesi					
n	Gene	s48/45 domain in region A		s48/45 domain in region B		Full-length gene	
		K_S (se)	K_N (se)	K_S (se)	K_N (se)	K_S (se)	K_N (se)
Worldwide isolates							
78	***pv12***	0.025 (0.005)†	0.006 (0.002)	0.020 (0.004)†	0.003 (0.001)	0.022 (0.003)*	0.005 (0.001)
54	***pv38***	-		0.028 (0.006)†	0.005 (0.001)	0.034 (0.004)*	0.007 (0.001)
Colombian isolates							
71	***pv12***	0.027 (0.006)†	0.006 (0.001)	0.022 (0.005)†	0.003 (0.001)	0.023 (0.002)*	0.005 (0.001)
47	***pv38***	-		0.031 (0.007)†	0.005 (0.001)	0.038 (0.005)*	0.008 (0.001)

K_N and K_S rates were estimated by using sequences obtained from databases (worldwide isolates) together with Colombian ones, and just with those obtained in the Colombian population. n: number of isolates. *pv12* s48/45 domain in region A: nucleotides 82-471; *pv12* s48/45 domain in region B: nucleotides 589-906; *pv38* s48/45 domain in region B: nucleotides 481-852; -: There are no s48/45 domains in *pv38* region A. Numbering is based on the Sal-I reference sequence.
*: $p < 0.000$, †: $p < 0.002$.

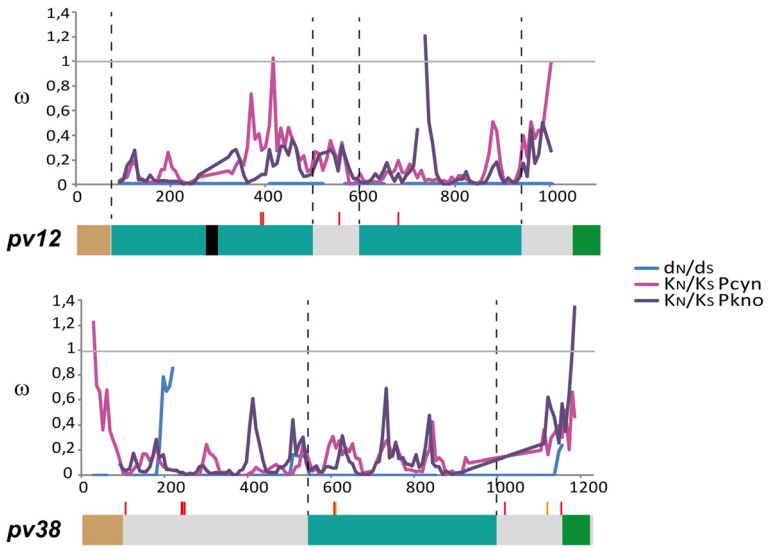


Figure 3 Sliding window analysis for ω rates. The ω (d_N/d_S) values for *Plasmodium vivax p12* and *p38* are shown in blue, whereas the divergence (ω: K_N/K_S) between *Plasmodium vivax* and *Plasmodium cynomolgi* (Pcyn) and *Plasmodium vivax* and *Plasmodium knowlesi* (Pkno) is displayed in magenta and purple, respectively. A gene diagram is shown below the sliding window. Regions encoding signal peptides (brown), GPI anchors (green), s48/45 domains (dark cyan) as well as the N[A/V][H/Q] repeat (black) are indicated. Non-synonymous (red) and synonymous (orange) substitutions are shown with vertical lines above each gene.

the s48/45 domain). There were more d_N substitutions in region A than d_S substitutions, whilst there were more d_S substitutions in region B than d_N ones, even though no significant values were observed (Table 4 and Additional file 2). Selection tests by codon revealed positive selection in codon 70 and negative selection in codons 175 and 323, suggesting that the gene was influenced by selection. When the long-term effect of natural selection was explored by comparing divergence rates (K_S and K_N), *pv38* had a higher statistically significant K_S rate than K_N (Table 4), revealing ω values below 1 throughout the gene (Figure 3), suggesting divergence by negative selection.

The McDonald-Kreitman test revealed statistically significant values (Table 4), when intraspecific polymorphism and interspecific divergence was compared, showing $P_N/P_S > D_N/D_S$ ($p < 0.02$). This result could have been the result of either a negative selection or a balanced selection [61,74]. K-test and H-test results (Table 2) and the presence of different haplotypes at intermediate frequencies (Figure 2B) suggested that it is most probable that *pv38*

Table 5 McDonald-Kreitman test for evaluating the action of natural selection

		P. vivax/P. cynomolgi			*P. vivax/P. knowlesi*		
Worldwide isolates							
		Fixed	**Polymorphic**	**$P_N/P_S > D_N/D_S$ p-values**	**Fixed**	**Polymorphic**	**$P_N/P_S > D_N/D_S$ p-values**
pv12	**Non-synonymous substitutions**	78.66	4	*0.002*	93.86	4	*0.000*
	Synonymous substitutions	190.23	0		340.47	0	
pv38	**Non-synonymous substitutions**	85.90	6	*0.004*	85.14	6	*0.003*
	Synonymous substitutions	257.31	3		265.94	3	
Colombian population							
pv12	**Non-synonymous substitutions**	93.05	1	0.146	115.54	1	0.083
	Synonymous substitutions	197.20	0		347.80	0	
pv38	**Non-synonymous substitutions**	89.22	5	*0.023*	88.50	5	*0.016*
	Synonymous substitutions	248.66	3		264.90	3	

The McDonald-Kreitman test was done using sequences obtained from databases (worldwide isolates) together with Colombian ones, and just with those obtained in the Colombian population. The interspecies divergence data were obtained from comparing *Plasmodium vivax* sequences with two related species: *Plasmodium cynomolgi* and *Plasmodium knowlesi*. Significant values are shown in italics.

was influenced by balanced selection, similar to that reported for *P. falciparum* [73]. Such selection seemed to be domain specific. Significant values were observed for region A ($p = 0.014$) when intraspecific polymorphism and interspecific divergence was calculated in each region (Additional file 3), this being where most of the substitutions found became accumulated, whilst neutrality could not be ruled out for region B ($p = 0.1$). Functional/structural constraint due to the presence of an s48/45 domain was also probable for *pv38*, given this region's low diversity, two negatively selected sites and a statistically significant $K_S > K_N$.

Linkage disequilibrium (LD) and recombination

Several statistics were calculated for determining possible associations between polymorphisms and/or the presence of recombination in *pv38*. Z_{nS} did not reveal statistically significant values, indicating that *pv38* polymorphisms were not associated. Lineal regression between linkage disequilibrium (LD) and nucleotide distance revealed a reduction in LD as nucleotide distance increased, indicating that intragenic recombination might have led to new variations being produced.

The ZZ statistic was calculated to confirm whether recombination affected *pv38* evolution, showing no significant values (Table 2); however, 2 RM (minimum recombination events) were found. The GARD method (in Datamonkey web server) gave a recombination breakpoint in position 524. Prior studies have suggested that new haplotypes could be produced through recombination in spite of functional constraints [73]. Intragenic recombination could thus be one of the factors promoting diversity in the *pv38* gene. Crosslinking during recombination could produce new combinations between the gene's 5′ (region A) and 3′ region (region B) as the breakpoint found in this gene was located upstream of the region encoding the s48/45 domain (region B). As only one polymorphic site was found in *pv12*, the aforementioned tests were not carried for this gene.

pv12 and *pv38* should be considered for an antimalarial vaccine

The lack of a totally effective vaccine against human malarial parasites is at least partly due to high genetic diversity found in proteins involved in red blood cell invasion. These molecules' constant exposure to the host's immune system allows the fixation of mutations generating an adaptive advantage preventing their recognition. Antigens such as *pvmsp-1*, *pvdbp*, *pvmsp-3α*, *pvmsp-5*, *pvmsp-7C*, *pvmsp-7H*, *pvmsp-7I* and *pvama-1* have shown high genetic diversity which appears to be maintained by positive-balancing selection [10-15,75-78]; however, other antigens are highly conserved despite being exposed to the host's immune system. Surface antigens such as *pvmsp-4*, *pvmsp-7A*, *pvmsp-7 K*, *pvmsp-8*, *pvmsp-10*, *pv230* or others in the rhoptries (*pvrap-1* and *pvrap-2*) appear to evolve more slowly due to a possible functional constraint in their encoded proteins [70,71,79-82]. Thus, most mutations have become eliminated from the population, maintaining a conserved protein structure, even throughout these parasites' evolutionary history [70,71]. The latter behaviour seems to have been directing *pv12* and *pv38* evolution, highlighting high conservation at both intra- and inter-species level due to the influence of negative selection exerted on s48/45 domains which are important for red blood cell recognition [30]. Although antigens having low genetic diversity are usually not immunogenic [83] nor do they induce protection-inducing responses [84], some limited polymorphism antigens have been shown to be able to induce immunogenicity and protection [85]. Therefore, *pv12* and *pv38* (or their s48/45 domains) should be evaluated regarding vaccine development because immune responses against 6-Cys family antigens appear to be directed against structural epitopes in s48/45 domains [86-88], blocking such domains should prevent invasion [30,88] and being highly conserved and having a functional constraint, allele-specific immune responses are thus avoided.

Conclusions

The *p12* and *p38* genes in *P. vivax* were seen to have low genetic diversity; the regions encoding the s48/45 domains seemed to be functionally or structurally constrained. Several members of the 6-Cys family are found on the surface of malaria parasites in every stage [28,36-39,69] and some of them (e g, P48/45, P230) are considered to be promising (transmission-blocking) vaccine candidates [36,37,87]. Epitopes identified by monoclonal antibodies against this type of protein are structural and have been localized within s48/45 domains [86,87] which seem to be involved in host-pathogen interaction [30,72]. Since *pv12* and *pv38* share structural characteristics with members of the 6-Cys family, added to their antigenic characteristics [38-40] and the low genetic diversity found in this study, the proteins encoded by these genes or their functionally/structurally constrained (conserved) regions could be born in mind when designing a multistage, multi-antigen subunit-based anti-malarial vaccine.

Additional files

Additional file 1: *pv12* and *pv38* haplotypes distribution in the Colombian population. Haplotype distribution found in *pv12* (A) and *pv38* (B) from 2007 to 2010.

Additional file 2: Synonymous substitution per synonymous site rate (d_S) and non-synonymous substitution per non-synonymous site rate (d_N) in s48/45 domains from *pv12* and *pv38* genes. No statistically significant differences were found by codon-based Z-test or Fisher's exact tests. se: Standard error. *pv12* s48/45 domain in region A: nucleotides 82-471; *pv12* s48/45 domain in region B: nucleotides 589-906;

pv38 s48/45 domain in region B: nucleotides 481-852 -: There is no s48/45 domain in the *pv38* region. Numbering is based on the Sal-I reference sequence.

Additional file 3: McDonald-Kreitman test for evaluating the action of natural selection in *pv12* and *pv38* gene regions A and B. The McDonald-Kreitman test was done using sequences obtained from databases (worldwide isolates) together with Colombian ones, and just with those obtained in the Colombian population. The interspecies divergence data was obtained from comparing *Plasmodium vivax* sequences with two related species: *Plasmodium cynomolgi* and *Plasmodium knowlesi*. Significant values are underlined. *pv12*: region A, nucleotides 1-546 and region B, nucleotides 547-1,095. *pv38*: region A, nucleotides 1-459 and region B, nucleotides 460-1,065.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
JF-R devised the study, participated in designing it, performed the experiments, made the population genetics analysis and wrote the manuscript. DG-O devised and designed the study, helped perform the experiments, carried out the population genetics analysis and wrote the manuscript. MAP devised and coordinated the study, and helped to write the manuscript. All the authors have read and approved the final manuscript.

Acknowledgements
We would especially like to thank Dr Sylvain Mousset who provided the ALLELIX software for our analysis. We would also like to thank Jason Garry for translating the manuscript and Professor Manuel E. Patarroyo for his comments and suggestions. This work was financed by the "Departamento Administrativo de Ciencia, Tecnología e Innovación (COLCIENCIAS)" through contract RC # 0309-2013. JF-R received financing through COLCIENCIAS cooperation agreement # 0757-2012.

Author details
[1]Molecular Biology and Immunology Department, Fundación Instituto de Inmunología de Colombia (FIDIC), Carrera 50 No. 26-20, Bogotá, DC, Colombia. [2]Microbiology postgraduate programme, Universidad Nacional de Colombia, Bogotá, DC, Colombia. [3]School of Medicine and Health Sciences, Universidad del Rosario, Bogotá, DC, Colombia.

Received: 14 September 2013 Accepted: 13 February 2014
Published: 18 February 2014

References
1. Rich SM, Ayala FJ: **Progress in malaria research: the case for phylogenetics.** *Adv Parasitol* 2003, **54:**255–280.
2. White NJ: ***Plasmodium knowlesi*: the fifth human malaria parasite.** *Clin Infect Dis* 2008, **46:**172–173.
3. WHO: *World Malaria Report 2012.* Geneva: World Health Organization; 2012. http://www.who.int/malaria/publications/world_malaria_report_2012/wmr2012_no_profiles.pdf.
4. Price RN, Tjitra E, Guerra CA, Yeung S, White NJ, Anstey NM: **Vivax malaria: neglected and not benign.** *Am J Trop Med Hyg* 2007, **77:**79–87.
5. Gething PW, Elyazar IR, Moyes CL, Smith DL, Battle KE, Guerra CA, Patil AP, Tatem AJ, Howes RE, Myers MF, George DB, Horby P, Wertheim HF, Price RN, Müeller I, Baird JK, Hay SI: **A long neglected world malaria map: *Plasmodium vivax* endemicity in 2010.** *PLoS Negl Trop Dis* 2012, **6:**e1814.
6. **Drug resistance in malaria.** http://www.who.int/csr/resources/publications/drugresist/malaria.pdf.
7. Ranson H, Rossiter L, Ortelli F, Jensen B, Wang X, Roth CW, Collins FH, Hemingway J: **Identification of a novel class of insect glutathione S-transferases involved in resistance to DDT in the malaria vector *Anopheles gambiae*.** *Biochem J* 2001, **359:**295–304.
8. Carvalho LJ, Daniel-Ribeiro CT, Goto H: **Malaria vaccine: candidate antigens, mechanisms, constraints and prospects.** *Scand J Immunol* 2002, **56:**327–343.
9. Jones TR, Hoffman SL: **Malaria vaccine development.** *Clin Microbiol Rev* 1994, **7:**303–310.
10. Gomez A, Suarez CF, Martinez P, Saravia C, Patarroyo MA: **High polymorphism in *Plasmodium vivax* merozoite surface protein-5 (MSP5).** *Parasitology* 2006, **133:**661–672.
11. Martinez P, Suarez CF, Cardenas PP, Patarroyo MA: ***Plasmodium vivax* Duffy binding protein: a modular evolutionary proposal.** *Parasitology* 2004, **128:**353–366.
12. Putaporntip C, Jongwutiwes S, Seethamchai S, Kanbara H, Tanabe K: **Intragenic recombination in the 3′ portion of the merozoite surface protein 1 gene of *Plasmodium vivax*.** *Mol Biochem Parasitol* 2000, **109:**111–119.
13. Putaporntip C, Udomsangpetch R, Pattanawong U, Cui L, Jongwutiwes S: **Genetic diversity of the *Plasmodium vivax* merozoite surface protein-5 locus from diverse geographic origins.** *Gene* 2010, **456:**24–35.
14. Garzon-Ospina D, Lopez C, Forero-Rodriguez J, Patarroyo MA: **Genetic diversity and selection in three *Plasmodium vivax* merozoite surface protein 7 (Pvmsp-7) genes in a Colombian population.** *PLoS One* 2012, **7:**e45962.
15. Figtree M, Pasay CJ, Slade R, Cheng Q, Cloonan N, Walker J, Saul A: ***Plasmodium vivax* synonymous substitution frequencies, evolution and population structure deduced from diversity in AMA 1 and MSP 1 genes.** *Mol Biochem Parasitol* 2000, **108:**53–66.
16. Mascorro CN, Zhao K, Khuntirat B, Sattabongkot J, Yan G, Escalante AA, Cui L: **Molecular evolution and intragenic recombination of the merozoite surface protein MSP-3alpha from the malaria parasite *Plasmodium vivax* in Thailand.** *Parasitology* 2005, **131:**25–35.
17. Escalante AA, Lal AA, Ayala FJ: **Genetic polymorphism and natural selection in the malaria parasite *Plasmodium falciparum*.** *Genetics* 1998, **149:**189–202.
18. Chenet SM, Tapia LL, Escalante AA, Durand S, Lucas C, Bacon DJ: **Genetic diversity and population structure of genes encoding vaccine candidate antigens of Plasmodium vivax.** *Malar J* 2012, **11:**68.
19. Takala SL, Plowe CV: **Genetic diversity and malaria vaccine design, testing and efficacy: preventing and overcoming 'vaccine resistant malaria'.** *Parasite Immunol* 2009, **31:**560–573.
20. Genton B, Reed ZH: **Asexual blood-stage malaria vaccine development: facing the challenges.** *Curr Opin Infect Dis* 2007, **20:**467–475.
21. Fluck C, Smith T, Beck HP, Irion A, Betuela I, Alpers MP, Anders R, Saul A, Genton B, Felger I: **Strain-specific humoral response to a polymorphic malaria vaccine.** *Infect Immun* 2004, **72:**6300–6305.
22. Ouattara A, Takala-Harrison S, Thera MA, Coulibaly D, Niangaly A, Saye R, Tolo Y, Dutta S, Heppner DG, Soisson L, Diggs CL, Vekemans J, Cohen J, Blackwelder WC, Dube T, Laurens MB, Doumbo OK, Plowe CV: **Molecular basis of allele-specific efficacy of a blood-stage malaria vaccine: vaccine development implications.** *J Infect Dis* 2013, **207:**511–519.
23. Zambrano-Villa S, Rosales-Borjas D, Carrero JC, Ortiz-Ortiz L: **How protozoan parasites evade the immune response.** *Trends Parasitol* 2002, **18:**272–278.
24. Richie TL, Saul A: **Progress and challenges for malaria vaccines.** *Nature* 2002, **415:**694–701.
25. Kimura M: *The neutral theory of molecular evolution. Cambridge Cambridgeshire.* New York: Cambridge University Press; 1983.
26. O'Donnell RA, De Koning-Ward TF, Burt RA, Bockarie M, Reeder JC, Cowman AF, Crabb BS: **Antibodies against merozoite surface protein (MSP)-1(19) are a major component of the invasion-inhibitory response in individuals immune to malaria.** *J Exp Med* 2001, **193:**1403–1412.
27. Nagao E, Seydel KB, Dvorak JA: **Detergent-resistant erythrocyte membrane rafts are modified by a *Plasmodium falciparum* infection.** *Exp Parasitol* 2002, **102:**57–59.
28. Sanders PR, Gilson PR, Cantin GT, Greenbaum DC, Nebl T, Carucci DJ, McConville MJ, Schofield L, Hodder AN, Yates JR 3rd, Crabb BS: **Distinct protein classes including novel merozoite surface antigens in Raft-like membranes of *Plasmodium falciparum*.** *J Biol Chem* 2005, **280:**40169–40176.
29. Arevalo-Pinzon G, Curtidor H, Vanegas M, Vizcaino C, Patarroyo MA, Patarroyo ME: **Conserved high activity binding peptides from the *Plasmodium falciparum* Pf34 rhoptry protein inhibit merozoites in vitro invasion of red blood cells.** *Peptides* 2010, **31:**1987–1994.
30. Garcia J, Curtidor H, Pinzon CG, Vanegas M, Moreno A, Patarroyo ME: **Identification of conserved erythrocyte binding regions in members of the *Plasmodium falciparum* Cys6 lipid raft-associated protein family.** *Vaccine* 2009, **27:**3953–3962.
31. Garcia Y, Puentes A, Curtidor H, Cifuentes G, Reyes C, Barreto J, Moreno A, Patarroyo ME: **Identifying merozoite surface protein 4 and merozoite surface protein 7 *Plasmodium falciparum* protein family members specifically binding to human erythrocytes suggests a new malarial parasite-redundant survival mechanism.** *J Med Chem* 2007, **50:**5665–5675.

32. Urquiza M, Rodriguez LE, Suarez JE, Guzman F, Ocampo M, Curtidor H, Segura C, Trujillo E, Patarroyo ME: **Identification of *Plasmodium falciparum* MSP-1 peptides able to bind to human red blood cells.** *Parasite Immunol* 1996, **18:**515–526.
33. Rodriguez LE, Curtidor H, Urquiza M, Cifuentes G, Reyes C, Patarroyo ME: **Intimate molecular interactions of *P. falciparum* merozoite proteins involved in invasion of red blood cells and their implications for vaccine design.** *Chem Rev* 2008, **108:**3656–3705.
34. Barrero CA, Delgado G, Sierra AY, Silva Y, Parra-Lopez C, Patarroyo MA: **Gamma interferon levels and antibody production induced by two PvMSP-1 recombinant polypeptides are associated with protective immunity against *P. vivax* in Aotus monkeys.** *Vaccine* 2005, **23:**4048–4053.
35. Richards JS, Beeson JG: **The future for blood-stage vaccines against malaria.** *Immunol Cell Biol* 2009, **87:**377–390.
36. Van Dijk MR, Van Schaijk BC, Khan SM, Van Dooren MW, Ramesar J, Kaczanowski S, Van Gemert GJ, Kroeze H, Stunnenberg HG, Eling WM, *et al*: **Three members of the 6-cys protein family of Plasmodium play a role in gamete fertility.** *PLoS Pathog* 2010, **6:**e1000853.
37. Williamson KC: **Pfs230: from malaria transmission-blocking vaccine candidate toward function.** *Parasite Immunol* 2003, **25:**351–359.
38. Mongui A, Angel DI, Guzman C, Vanegas M, Patarroyo MA: **Characterisation of the *Plasmodium vivax* Pv38 antigen.** *Biochem Biophys Res Commun* 2008, **376:**326–330.
39. Moreno-Perez DA, Areiza-Rojas R, Florez-Buitrago X, Silva Y, Patarroyo ME, Patarroyo MA: **The GPI-anchored 6-Cys protein Pv12 is present in detergent-resistant microdomains of *Plasmodium vivax* blood stage schizonts.** *Protist* 2013, **164:**37–48.
40. Chen JH, Jung JW, Wang Y, Ha KS, Lu F, Lim CS, Takeo S, Tsuboi T, Han ET: **Immunoproteomics profiling of blood stage *Plasmodium vivax* infection by high-throughput screening assays.** *J Proteome Res* 2010, **9:**6479–6489.
41. Prajapati SK, Joshi H, Shalini S, Patarroyo MA, Suwanarusk R, Kumar A, Sharma SK, Eapen A, Dev V, Bhatt RM, Valecha N, Nosten F, Rizvi MA, Dash AP: ***Plasmodium vivax* lineages: geographical distribution, tandem repeat polymorphism, and phylogenetic relationship.** *Malar J* 2011, **10:**374.
42. Imwong M, Pukrittayakamee S, Gruner AC, Renia L, Letourneur F, Looareesuwan S, White NJ, Snounou G: **Practical PCR genotyping protocols for *Plasmodium vivax* using Pvcs and Pvmsp1.** *Malar J* 2005, **4:**20.
43. Neafsey DE, Galinsky K, Jiang RH, Young L, Sykes SM, Saif S, Gujja S, Goldberg JM, Young S, Zeng Q, Chapman SB, Dash AP, Anvikar AR, Sutton PL, Birren BW, Escalante AA, Barnwell JW, Carlton JM: **The malaria parasite *Plasmodium vivax* exhibits greater genetic diversity than *Plasmodium falciparum*.** *Nat Genet* 2012, **44:**1046–1050.
44. 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, *et al:* **Comparative genomics of the neglected human malaria parasite *Plasmodium vivax*.** *Nature* 2008, **455:**757–763.
45. Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput.** *Nucleic Acids Res* 2004, **32:**1792–1797.
46. Suyama M, Torrents D, Bork P: **PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments.** *Nucleic Acids Res* 2006, **34:**W609–W612.
47. Librado P, Rozas J: **DnaSP v5: a software for comprehensive analysis of DNA polymorphism data.** *Bioinformatics* 2009, **25:**1451–1452.
48. Depaulis F, Veuille M: **Neutrality tests based on the distribution of haplotypes under an infinite-site model.** *Mol Biol Evol* 1998, **15:**1788–1790.
49. Tajima F: **Statistical method for testing the neutral mutation hypothesis by DNA polymorphism.** *Genetics* 1989, **123:**585–595.
50. Fu YX, Li WH: **Statistical tests of neutrality of mutations.** *Genetics* 1993, **133:**693–709.
51. Fay JC, Wu CI: **Hitchhiking under positive Darwinian selection.** *Genetics* 2000, **155:**1405–1413.
52. Fu YX: **Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection.** *Genetics* 1997, **147:**915–925.
53. Zhang J, Rosenberg HF, Nei M: **Positive Darwinian selection after gene duplication in primate ribonuclease genes.** *Proc Natl Acad Sci U S A* 1998, **95:**3708–3713.
54. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S: **MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods.** *Mol Biol Evol* 2011, **28:**2731–2739.
55. Kosakovsky Pond SL, Frost SD: **Not so different after all: a comparison of methods for detecting amino acid sites under selection.** *Mol Biol Evol* 2005, **22:**1208–1222.
56. Pond SL, Frost SD, Grossman Z, Gravenor MB, Richman DD, Brown AJ: **Adaptation to different human populations by HIV-1 revealed by codon-based analyses.** *PLoS Comput Biol* 2006, **2:**e62.
57. Murrell B, Wertheim JO, Moola S, Weighill T, Scheffler K, Kosakovsky Pond SL: **Detecting individual sites subject to episodic diversifying selection.** *PLoS Genet* 2012, **8:**e1002764.
58. Murrell B, Moola S, Mabona A, Weighill T, Sheward D, Kosakovsky Pond SL, Scheffler K: **FUBAR: a fast, unconstrained bayesian approximation for inferring selection.** *Mol Biol Evol* 2013, **30:**1196–1205.
59. Jukes TH, Cantor CR: **Evolution of protein molecules.** In *Mammalian Protein Metabolism*. Edited by Munro HN. New York: Academic Press; 1969.
60. McDonald JH, Kreitman M: **Adaptive protein evolution at the Adh locus in Drosophila.** *Nature* 1991, **351:**652–654.
61. Egea R, Casillas S, Barbadilla A: **Standard and generalized McDonald-Kreitman test: a website to detect selection by comparing different classes of DNA sites.** *Nucleic Acids Res* 2008, **36:**W157–W162.
62. Kelly JK: **A test of neutrality based on interlocus associations.** *Genetics* 1997, **146:**1197–1206.
63. Rozas J, Gullaud M, Blandin G, Aguade M: **DNA variation at the rp49 gene region of *Drosophila simulans*: evolutionary inferences from an unusual haplotype structure.** *Genetics* 2001, **158:**1147–1155.
64. Hudson RR, Kaplan NL: **Statistical properties of the number of recombination events in the history of a sample of DNA sequences.** *Genetics* 1985, **111:**147–164.
65. Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SD: **Automated phylogenetic detection of recombination using a genetic algorithm.** *Mol Biol Evol* 2006, **23:**1891–1901.
66. Delport W, Poon AF, Frost SD, Kosakovsky Pond SL: **Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology.** *Bioinformatics* 2010, **26:**2455–2457.
67. Taylor JE, Pacheco MA, Bacon DJ, Beg MA, Machado RL, Fairhurst RM, Herrera S, Kim JY, Menard D, Póvoa MM, Villegas L, Mulyanto, Snounou G, Cui L, Zeyrek FY, Escalante AA: **The evolutionary history of *Plasmodium vivax* as inferred from mitochondrial genomes: parasite genetic diversity in the Americas.** *Mol Biol Evol* 2013, **30:**2050–2064.
68. Tetteh KK, Stewart LB, Ochola LI, Amambua-Ngwa A, Thomas AW, Marsh K, Weedall GD, Conway DJ: **Prospective identification of malaria parasite genes under balancing selection.** *PLoS One* 2009, **4:**e5568.
69. Tonkin ML, Arredondo SA, Loveless BC, Serpa JJ, Makepeace KA, Sundar N, Petrotchenko EV, Miller LH, Grigg ME, Boulanger MJ: **Structural and biochemical characterization of *Plasmodium falciparum* 12 (Pf12) reveals a unique interdomain organization and the potential for an antiparallel arrangement with Pf41.** *J Biol Chem* 2013, **288:**12805–12817.
70. Pacheco MA, Ryan EM, Poe AC, Basco L, Udhayakumar V, Collins WE, Escalante AA: **Evidence for negative selection on the gene encoding rhoptry-associated protein 1 (RAP-1) in Plasmodium spp.** *Infect Genet Evol* 2010, **10:**655–661.
71. Pacheco MA, Elango AP, Rahman AA, Fisher D, Collins WE, Barnwell JW, Escalante AA: **Evidence of purifying selection on merozoite surface protein 8 (MSP8) and 10 (MSP10) in Plasmodium spp.** *Infect Genet Evol* 2012, **12:**978–986.
72. Arredondo SA, Cai M, Takayama Y, MacDonald NJ, Anderson DE, Aravind L, Clore GM, Miller LH: **Structure of the Plasmodium 6-cysteine s48/45 domain.** *Proc Natl Acad Sci U S A* 2012, **109:**6692–6697.
73. Reeder JC, Wapling J, Mueller I, Siba PM, Barry AE: **Population genetic analysis of the *Plasmodium falciparum* 6-cys protein Pf38 in Papua New Guinea reveals domain-specific balancing selection.** *Malar J* 2011, **10:**126.
74. Parsch J, Zhang Z, Baines JF: **The influence of demography and weak selection on the McDonald-Kreitman test: an empirical study in Drosophila.** *Mol Biol Evol* 2009, **26:**691–698.
75. Ord R, Polley S, Tami A, Sutherland CJ: **High sequence diversity and evidence of balancing selection in the Pvmsp3alpha gene of *Plasmodium vivax* in the Venezuelan Amazon.** *Mol Biochem Parasitol* 2005, **144:**86–93.

76. Kang JM, Ju HL, Kang YM, Lee DH, Moon SU, Sohn WM, Park JW, Kim TS, Na BK: **Genetic polymorphism and natural selection in the C-terminal 42 kDa region of merozoite surface protein-1 among *Plasmodium vivax* Korean isolates.** *Malar J* 2012, **11:**206.
77. Zakeri S, Sadeghi H, Mehrizi AA, Djadid ND: **Population genetic structure and polymorphism analysis of gene encoding apical membrane antigen-1 (AMA-1) of Iranian *Plasmodium vivax* wild isolates.** *Acta Trop* 2013, **126:**269–279.
78. Ju HL, Kang JM, Moon SU, Kim JY, Lee HW, Lin K, Sohn WM, Lee JS, Kim TS, Na BK: **Genetic polymorphism and natural selection of Duffy binding protein of *Plasmodium vivax* Myanmar isolates.** *Malar J* 2012, **11:**60.
79. Garzon-Ospina D, Romero-Murillo L, Tobon LF, Patarroyo MA: **Low genetic polymorphism of merozoite surface proteins 7 and 10 in Colombian Plasmodium vivax isolates.** *Infect Genet Evol* 2011, **11:**528–531.
80. Garzon-Ospina D, Romero-Murillo L, Patarroyo MA: **Limited genetic polymorphism of the *Plasmodium vivax* low molecular weight rhoptry protein complex in the Colombian population.** *Infect Genet Evol* 2010, **10:**261–267.
81. Martinez P, Suarez CF, Gomez A, Cardenas PP, Guerrero JE, Patarroyo MA: **High level of conservation in *Plasmodium vivax* merozoite surface protein 4 (PvMSP4).** *Infect Genet Evol* 2005, **5:**354–361.
82. Doi M, Tanabe K, Tachibana S, Hamai M, Tachibana M, Mita T, Yagi M, Zeyrek FY, Ferreira MU, Ohmae H, Kaneko A, Randrianarivelojosia M, Sattabongkot J, Cao YM, Horii T, Torii M, Tsuboi T: **Worldwide sequence conservation of transmission-blocking vaccine candidate Pvs230 in *Plasmodium vivax*.** *Vaccine* 2011, **29:**4308–4315.
83. Patarroyo MA, Calderon D, Moreno-Perez DA: **Vaccines against *Plasmodium vivax*: a research challenge.** *Expert Rev Vaccines* 2012, **11:**1249–1260.
84. Giraldo MA, Arevalo-Pinzon G, Rojas-Caraballo J, Mongui A, Rodriguez R, Patarroyo MA: **Vaccination with recombinant *Plasmodium vivax* MSP-10 formulated in different adjuvants induces strong immunogenicity but no protection.** *Vaccine* 2009, **28:**7–13.
85. Rojas-Caraballo J, Mongui A, Giraldo MA, Delgado G, Granados D, Millan-Cortes D, Martinez P, Rodriguez R, Patarroyo MA: **Immunogenicity and protection-inducing ability of recombinant *Plasmodium vivax* rhoptry-associated protein 2 in Aotus monkeys: a potential vaccine candidate.** *Vaccine* 2009, **27:**2870–2876.
86. Carter R, Coulson A, Bhatti S, Taylor BJ, Elliott JF: **Predicted disulfide-bonded structures for three uniquely related proteins of *Plasmodium falciparum*, Pfs230, Pfs48/45 and Pf12.** *Mol Biochem Parasitol* 1995, **71:**203–210.
87. Tachibana M, Sato C, Otsuki H, Sattabongkot J, Kaneko O, Torii M, Tsuboi T: ***Plasmodium vivax* gametocyte protein Pvs230 is a transmission-blocking vaccine candidate.** *Vaccine* 2012, **30:**1807–1812.
88. Feller T, Thom P, Koch N, Spiegel H, Addai-Mensah O, Fischer R, Reimann A, Pradel G, Fendel R, Schillberg S, Scheuermayer M, Schinkel H: **Plant-based production of recombinant Plasmodium surface protein pf38 and evaluation of its potential as a vaccine candidate.** *PLoS One* 2013, **8:**e79920.

doi:10.1186/1475-2875-13-58
Cite this article as: Forero-Rodríguez *et al.*: Low genetic diversity and functional constraint in loci encoding *Plasmodium vivax* P12 and P38 proteins in the Colombian population. *Malaria Journal* 2014 **13**:58.

MALARIA JOURNAL

RESEARCH Open Access

Low genetic diversity in the locus encoding the *Plasmodium vivax* P41 protein in Colombia's parasite population

Johanna Forero-Rodríguez[1,2], Diego Garzón-Ospina[1,3] and Manuel A Patarroyo[1,3*]

Abstract

Background: The development of malaria vaccine has been hindered by the allele-specific responses produced by some parasite antigens' high genetic diversity. Such antigen genetic diversity must thus be evaluated when designing a completely effective vaccine. *Plasmodium falciparum* P12, P38 and P41 proteins have red blood cell binding regions in the s48/45 domains and are located on merozoite surface, P41 forming a heteroduplex with P12. These three genes have been identified in *Plasmodium vivax* and share similar characteristics with their orthologues in *Plasmodium falciparum*. *Plasmodium vivax pv12* and *pv38* have low genetic diversity but *pv41* polymorphism has not been described.

Methods: The present study was aimed at evaluating the *P. vivax p41* (*pv41*) gene's polymorphism. DNA sequences from Colombian clinical isolates from *pv41* gene were analysed for characterising and studying the genetic diversity and the evolutionary forces that produced the variation pattern so observed.

Results: Similarly to other members of the 6-Cys family, *pv41* had low genetic polymorphism. *pv41* 3′-end displayed the highest nucleotide diversity value; several substitutions found there were under positive selection. Negatively selected codons at inter-species level were identified in the s48/45 domains; *p41* would thus seem to have functional/structural constraints due to the presence of these domains.

Conclusions: In spite of the functional constraints of Pv41 s48/45 domains, immune system pressure seems to have allowed non-synonymous substitutions to become fixed within them as an adaptation mechanism; including Pv41 s48/45 domains in a vaccine should thus be carefully evaluated due to these domains containing some allele variants.

Keywords: *Plasmodium vivax*, 6-Cys, *pv41*, s48/45 domains, Genetic variability, Functional constraint, Anti-malarial vaccine

Background

Of the five malaria parasites (*Plasmodium falciparum, Plasmodium vivax, Plasmodium malarie, Plasmodium ovale* and *Plasmodium knowlesi*) affecting human beings, *P. falciparum* is the species causing the most severe clinical manifestations, whilst *P. vivax* is the species most widely distributed throughout the world, mainly affecting the Asian and American continents and causing the highest morbidity outside of Africa. In spite of efforts to date for controlling malaria, it continues to be a serious public health problem; 18.9 million cases of *P. vivax* occurred in 2012, children under five years old and pregnant women being the most vulnerable populations [1].

An anti-malarial vaccine represents one of the alternative control measures regarding this disease; developing a multi-antigen vaccine against the parasite's blood stage is focused on blocking all interactions with a host cell, thereby avoiding recognition and subsequent invasion. Several antigens have been proposed as vaccine candidates [2-4]; however, as many of them have high genetic diversity [5-12], this is an obstacle regarding such proposal

* Correspondence: mapatarr.fidic@gmail.com
[1]Fundación Instituto de Inmunología de Colombia (FIDIC), Carrera 50 No. 26-20, Bogotá, DC, Colombia
[3]School of Medicine and Health Sciences, Universidad del Rosario, Bogotá, DC, Colombia
Full list of author information is available at the end of the article

[13,14] since they induce allele-specific immune responses [15]. The genetic diversity of candidate antigens must thus be evaluated [14,16] for selecting the most frequent variants or conserved domains [13,14].

Proteins involved in red blood cell (RBC) invasion have been characterized in merozoite surface regions known as detergent-resistant membranes (DRM) [17-19], many of these being potential vaccine candidates [4,20,21]. Such DRMs include a group of proteins belonging to the 6-Cys family (P12, P38, P41 and P92) which is characterised by the presence of domains containing six conserved cysteines called s48/45 [17,22-24]. The *P. falciparum* P41 (Pf41) protein has two high-activity binding peptides in the s48/45 domains [17], thereby suggesting a role in RBC invasion. This protein does not have GPI-anchored domains and its presence on merozoite membrane is due to the formation of an inverted heteroduplex with Pf12 [25,26]. The *pv41* gene has recently been characterised in *P. vivax* (*pv41*) [22,27]; this gene encodes a 385 residue-long membrane protein. Similar to its orthologue in *P. falciparum*, the protein has a signal peptide and two s48/45 domains but no GPI-anchor. The *P. vivax* P41 (Pv41) protein has been shown to be antigenic [27,28], suggesting that it is exposed to the host immune system, probably during invasion of the host cell.

Given that Pv41 has been located on merozoite surface and that it has no membrane anchoring domains [22,27], it could be interacting with another protein anchored to parasite surface. This protein's similarity with its orthologue in *P. falciparum* suggests that Pv41 might form a complex with Pv12, a protein which has been shown to be highly conserved [29]. The present study was therefore aimed at using population genetics analysis for evaluating the *pv41* gene's genetic diversity by determining the evolutionary processes producing the locus's variation pattern. The results showed that *pv41* had low genetic diversity, the gene's 3′-end region being the most diverse, fixing mutations by positive selection, probably as a mechanism for evading the immune system. Like other members of the 6-Cys family, this gene seemed to have functional constraints due to the presence of s48/45 domains.

Methods

Declaration of ethical considerations

This study involved using thirty *P. vivax*-infected samples collected between 2007 and 2010 (2007: 5 isolates, 2008: 3 isolates, 2009: 8 isolates, 2010: 14 isolates); they had been obtained from different regions of Colombia (Figure 1, South-west: Chocó, Nariño; South-east: Caquetá, Guainía, Guaviare, Meta; Midwest: Bogotá, Tolima; North-west:

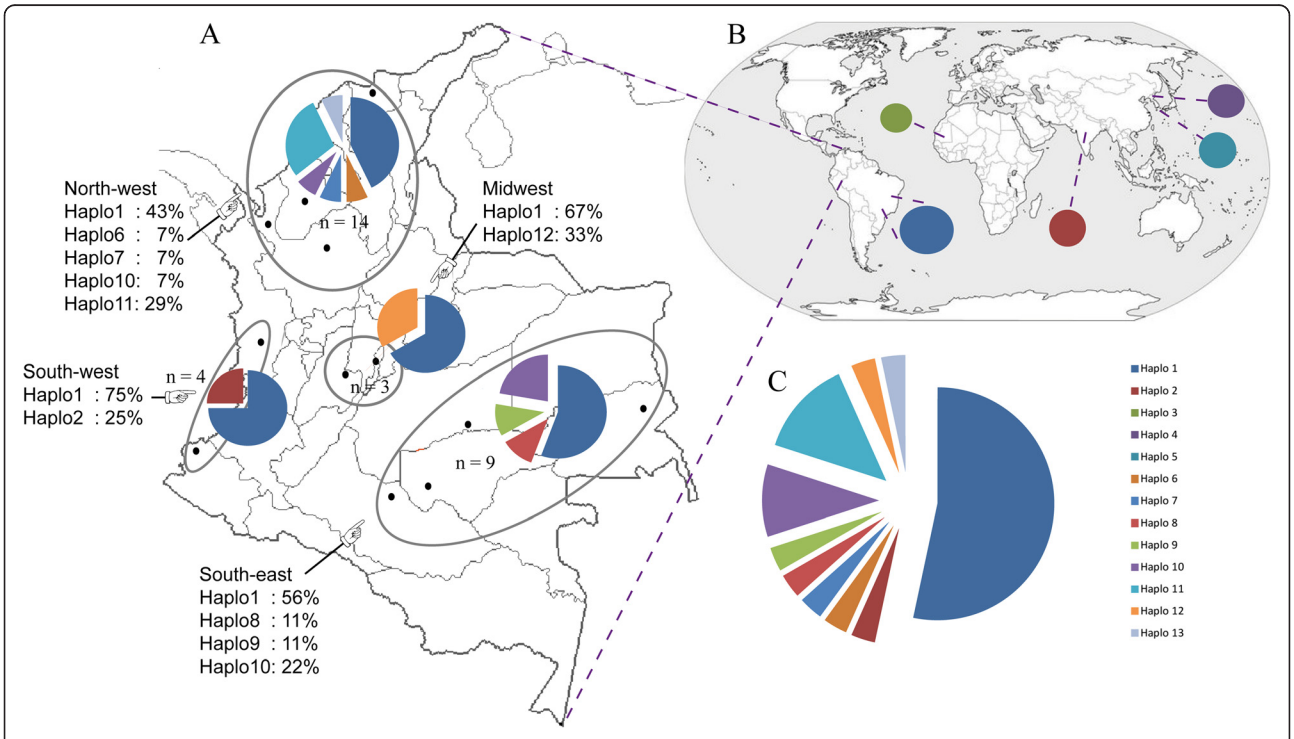


Figure 1 *pv41* haplotype distribution in the Colombian population and origin of the reference strain sequences. Panel **A** shows the haplotype distribution found in *pv41*. Panel **B** shows the origin of the reference strain sequences and panel **C** represents haplotype frequency in the Colombian population. Haplotype 1: Sal-I, Brazil-I, haplotype 2: India, haplotype 3: Mauritania, haplotype 4: North Korea, haplotype 5: South Korea, haplotype 1–2 and 6–13: Colombian isolates.

Atlántico, Antioquia, Córdoba). All *P. vivax*-infected patients who provided blood samples were notified of the study's objective and then signed an informed consent form. All the procedures involved in taking the samples had already been approved by the Fundación Instituto de Inmunología de Colombia's (FIDIC) ethics' committee.

Genotyping *Plasmodium vivax* samples

PCR-RFLP of the *pvmsp-1* polymorphic marker was used for identifying/analysing different genotypes in the samples and infection by a single *P. vivax* strain, as described previously [30]. Briefly, this gene's blocks 6, 7 and 8 were amplified with direct 5′-AAAATCGAGAGCATGATCGC CACTGAGAAG-3′ and reverse 5′-AGCTTGTACTTTC CATAGTGGTCCAG-3′ primers. The amplified fragments were digested with Alu I and Mnl I restriction enzymes.

PCR amplification of the *pv41* gene

Previously reported primers were used for amplifying *pv41* [22]. The PCR reaction mixture contained 10 mM Tris HCl, 50 mM KCl (GeneAmp 10X PCR Buffer II (Applied Biosystems)), 1.5 mM $MgCl_2$, 0.2 mM of each dNTP, 0.5 μM of each primer (direct 5′ ATGAAAAGG CTCCTCCTGC 3′ and reverse 5′ CTCCTGGAAGGA CTTGGC 3′), 0.76 U Amplitaq Gold DNA polymerase (Applied Biosystems) and 40 ng genomic DNA at 50 μL final volume. The PCR thermal profile was as follows: one cycle at 95°C (7 min), 40 cycles at 95°C (20 sec), 60°C (30 sec), 72°C (1 min) and a final extension cycle at 72°C (10 min). The amplification products were purified using an UltraClean PCR Clean-up kit (MO BIO). The purified PCR products were bidirectionally sequenced with the amplification primers using the BigDye method with capillary electrophoresis, using the ABI-3730 XL sequencer (MACROGEN, Seoul, South Korea). Two independent PCR products were sequenced per sample to rule out errors.

Analysing genetic diversity

CLC Main workbench software v.5 (CLC bio, Cambridge, MA, USA) was used for analysing and assembling the electropherograms obtained by sequencing, giving one sequence per sample. The 30 sequences obtained from Colombian isolates were compared to and analysed regarding reference sequences obtained from several sequencing projects [31,32] (PlasmoDB accession number: PVX_000995, GenBank accession number: AFNI01000110.1, AFNJ01000259.1, AFMK01000149.1 and AFBK01000223.1) or reported in databases (GenBank accession number: GU476495.1). These 36 sequences were then compared to *Plasmodium cynomolgi* (GenBank accession number: BAEJ01000104.1) and *P. knowlesi* orthologous sequences (PlasmoDB accession number: PKH_030970), two species which are phylogenetically close to *P. vivax* [33]. Gene Runner software was used for translating all the sequences for obtaining the deduced amino acid sequences; the MUSCLE algorithm was then used for aligning such sequences [34] and then edited manually. The PAL2NAL web-based tool [35] was then used for converting protein alignments into their respective nucleotide alignments.

DnaSP v.5 software [36] was used for quantifying *pv41* genetic polymorphism by calculating: the number of segregant sites (Ss), the number of singleton sites (s), the number of parsimony-informative sites (Ps), the number of haplotypes (H), haplotype diversity (Hd, multiplied by (n-1)/n, according to Depaulis and Veuille [36,37]), the Watterson estimator (θ^w), the average number of nucleotide differences (k) and nucleotide diversity per site (π). Data was obtained for the reference sequences plus the Colombian sequences (worldwide diversity), as well as for just the Colombian sequences (local diversity).

The Colombian parasite population sequences were used for evaluating the neutral model of molecular evolution using tests based on the frequency spectrum of nucleotide polymorphisms and haplotype distribution. Tajima's D test [38], Fu and Li's D* and F* tests [39], and Fay and Wu's H test [40] were calculated for the first group of tests. Fu's Fs test [41] and K-test and H-test [37] were calculated as part of the group of tests based on haplotype distribution. The significance of all tests was determined by coalescence simulations using DnaSP v.5 [36] and ALLELIX software (provided by Dr Sylvain Mousset). Sites having gaps were not taken into account for all tests.

The effect of natural selection was evaluated by calculating the difference between the average number of non-synonymous substitutions per non-synonymous site (d_N) and the average number of synonymous substitutions per synonymous site (d_S) using the modified Nei-Gojobori method [42]. Significance was determined by using Fisher's exact tests and the Z test incorporated in MEGA v.5 software [43]. SLAC, FEL, REL [44], IFEL [45], MEME [46] and FUBAR methods [47] were used for calculating the ω (d_N/d_S) value for each codon in the *pv41* alignment.

The McDonald-Kreitman test [48] was calculated for evaluating the effect of natural selection on *p41* during the evolutionary history of *P. vivax* and related species (*Plasmodium cynomolgi* and *P. knowlesi*); this test compared intraspecific polymorphism with interspecific divergence using a web server [49], which takes the Jukes-Cantor distance correction regarding divergence per site [50] into account. The Nei-Gojobori modified method [42] was also used for calculating the difference between non-synonymous (K_N) and synonymous (K_S) divergence rates using Jukes-Cantor divergence correction [50]. Significant values were determined by using the Z test incorporate in MEGA v.5 software [43]. SLAC, FEL, REL [44], MEME [46] and FUBAR [47] methods were used

for determining sites under interspecies selection using the *P. vivax*, *P. cynomolgi* and *P. knowlesi* sequences as data set.

Z_{nS} [51] and ZZ [52] tests were calculated for evaluating non-random associations between polymorphisms (linkage disequilibrium or LD) and the influence of intragenic recombination on *pv41*. The minimum number of recombination events (Rm) [53] was also calculated and the GARD method [54] available from Datamonkey [55] was used for evaluating recombination processes.

Results

Genetic diversity in *pv41*

Thirty *P. vivax*-infected samples, obtained from different parts of Colombia (Figure 1), were genotyped using the *pvmsp-1* polymorphic marker. The RFLP patterns produced from *pvmsp-1* blocks 6–8 suggested the presence of different genotypes in the aforementioned samples as well as single strain infections in each sample. Taking into account that all these samples have been previously used in other studies involving genes having high polymorphism [6], in which none of the electropherograms revealed overlapping peaks during the sequencing, we can ascertain the absence of multiple infections.

The 30 genotyped isolates had a 1,152 base pair (bp) fragment corresponding to the *pv41* gene. The sequences obtained from these 30 isolates (Additional file 1) were compared to and analysed together with sequences reported by several sequencing projects [31,32]. Sequences having a different haplotype were deposited in the GenBank database (accession numbers KM212268-KM212275).

Table 1 gives the values for the estimators of genetic diversity. Seventeen segregant sites were observed in the sequences from different parts of the world, 12 of them being parsimony-informative sites and five singleton sites; 13 haplotypes were found (Figure 2). Aligning the proteins from *P. vivax* isolates from different geographical locations revealed substitutions in ten amino acids: N88D, E89V, A258V, Q301H, K312N, M355R, S359H, Y361F, N363D and R373G (numeration based on the Sal-I reference sequence). Ten segregant sites were found in the Colombian population (nine of them being parsimony-informative sites), giving ten haplotypes (haplotypes 1, 2, 6–13) and 0.679 ± 0.083 haplotype diversity. Haplotype 1 had 50% frequency, followed by haplotype 11 (13% frequency) and haplotype 10 (10% frequency); the remaining haplotypes had low frequency (around 3%).

Table 1 Genetic diversity estimators for *pv41*

n	Sites	Ss	S	Ps	H	θ^w	*k*	π
Worldwide diversity								
36	1,068	17	5	12	13	0.0038 ± 0.0009	3.9	0.0037 ± 0.0006
Local diversity								
30	1,115	10	1	9	10	0.0023 ± 0.0007	3.1	0.0028 ± 0.0005

The estimators of genetic diversity were calculated by using the sequences obtained from the databases plus the Colombian ones (worldwide diversity) and just using those obtained in the Colombian population (local diversity). n: number of isolates, sites: total of sites analysed (excluding gaps), Ss: number of segregant sites, S: number of singleton sites, Ps: number of parsimony-informative sites, H: number of haplotypes, *k*: average number of nucleotide differences by sequence pairs, θ^w: Watterson estimator, π: nucleotide diversity per site.

	1	2	2	7	9	9	0	0	0	0	0	0	0	0	1	1	1
	7	6	6	7	0	3	1	6	7	7	7	8	8	8	1	1	1
	1	2	6	3	3	6	2	4	5	6	7	2	3	7	6	7	9
Haplotype_1	T	G	T	T	C	T	T	G	C	A	T	A	T	G	A	G	A
Haplotype_2	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.
Haplotype_3	.	A	.	.	.	.	.	T	C	G	T	C	A	C	.	C	
Haplotype_4	.	.	.	.	.	.	.	T	C	G	T	C	A	C	.	C	
Haplotype_5	C	.	A	.	.	A	C	T	T	C	G	T	C	A	C	.	C
Haplotype_6	.	.	.	.	A	.	.	.	T	C	G	T	C	A	.	.	.
Haplotype_7	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	A	.
Haplotype_8	.	.	.	.	A	.	.	.	T	C	G	T	C	A	.	A	.
Haplotype_9	.	.	.	.	.	.	.	.	T	C	G	T	C	A	.	A	.
Haplotype_10	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.
Haplotype_11	.	.	.	.	.	.	.	.	T	C	G	T	C	A	.	.	.
Haplotype_12	.	.	.	C	.	.	.	.	T	C	G	T	C	A	.	.	.
Haplotype_13	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.

Figure 2 Aligning the haplotypes found in the *pv41* gene. The numbers in the upper part indicate the nucleotide position where a substitution was observed; the dots indicate nucleotide identity.

The average number of nucleotide differences per pairs of sequences (*k*) was 3.9 when sequences from different parts of the world (worldwide diversity) were analysed and 3.1 for the Colombian population (Table 1). Low Watterson estimator (θ^w = 0.0038 ± 0.0009) and nucleotide diversity values (π = 0.0037 ± 0.0006) were observed when the available sequences obtained from the databases plus the Colombian ones were analysed; θ^w was 0.0023 ± 0.0007 and π 0.0028 ± 0.0005 for the Colombian population (Table 1). The nucleotide diversity analysis for Colombian locations showed that the Midwest was the most diverse at the *pv41* locus whilst the lowest value was found in Colombia's South-west area (Additional file 2). The gene region having the highest π value was found between nucleotides 1,064 to 1,130.

Evaluating the effect of natural selection on *pv41*

Tajima's D, Fu and Li's D* and F*, Fay and Wu's H, Fu's Fs and the K- and H-test neutrality tests did not give statistically significant values (Table 2); this meant that neutrality could not be ruled out. The differences between non-synonymous and synonymous (d_N - d_S) substitutions rates throughout the gene were evaluated for estimating the effect of natural selection in *pv41*, as well as in each s48/45 domain (s48/45 N-Terminal: nucleotide 76–351 and s48/45 C-Terminal: nucleotide 784–1,095);

Table 2 Tests based on the neutral model of molecular evolution, linkage disequilibrium and recombination for the *pv41* gene in the Colombian population

n	Tajima	Fu and Li		Fay and Wu H	Fu Fs	K-test	H-test	Zns	ZZ	RM
	D	D*	F*							
30	0.79023	0.86738	0.9868	−1.857	−1.267	10	0.679 ± 0.08	0.3627*	0.2073*	2

*p <0.05.

however, no significant values were found (Table 3). The sliding window (Figure 3) for the ω (d_N/d_S) rate gave a ω close to 1 at the 3′-end of *pv41*, indicating a number of non-synonymous substitutions fixed within *P. vivax* in this region at a higher rate than in the rest of the sequence. Tests estimating d_N/d_S for each site (codon) were then performed for identifying whether individual codons in *pv41* were under selection; seven codons were found to be under positive selection and one codon under negative selection (Figure 3). Substitutions V269A, H312Q and G384R were exclusive for the Colombian population. The K323N, H370S amino acid changes were found in Colombian isolates and some reference sequences, whilst the N88D and E89V substitutions were present in Mauritanian and South Korean sequences, respectively.

The McDonald-Kreitman test was calculated for evaluating how selection had acted throughout *p41*'s evolutionary history; it revealed significant values, thereby showing that polymorphism was greater than divergence ($p < 0.05$) (Table 4). A sliding window for ω divergence (K_N/K_S, non-synonymous divergence/synonymous divergence), obtained by comparing the *P. vivax* sequences to sequences from phylogenetically close species (*P. cynomolgi* and *P. knowlesi*), gave values less than 1 in the s48/45 domains, as well as in some areas between these domains, thereby indicating that K_S tended to be greater than K_N. Significant negative values ($p < 0.001$) were found when estimating the difference between non-synonymous and synonymous divergence (K_N - K_S) (Table 5). The codon-based selection tests found 13 positively selected codons and 77 negatively selected codons at inter-species level (Figure 3).

Linkage disequilibrium (LD) and recombination

The Z_{nS}, ZZ and RM tests were calculated for determining possible associations between polymorphism and/or the presence of recombination in *pv41* (Table 2). The Z_{nS} test gave 0.3627, this being statistically significant ($p < 0.05$). Lineal regression between LD and nucleotide distance gave a slight reduction in LD as nucleotide distance increased, suggesting recombination events. This was confirmed when the ZZ test was calculated, giving 0.2073 ($p < 0.05$); two minimum recombination sites were found (Table 2). The GARD method (available from the Datamonkey web server) gave a recombination breakpoint in position 936 (number based on Sal-I sequence) confirming than intragenic recombination was involved in generating new haplotypes in *pv41*.

Discussion

Merozoite-expressed members of the 6-Cys family in *P. falciparum* (Pf12, Pf38 and Pf41) have high RBC binding activity peptides [17], indicating that these play a role during recognition of a host cell. Previous studies have shown that members of this family are antigenic [23,24,27,28] and highly conserved (*p12* and *p38*) in both *P. falciparum* and *P. vivax* [26,29,56,57]. This means that they are promising candidates for inclusion in an anti-malarial vaccine, avoiding allele-specific immune responses. The *pv41* gene has been shown to be highly conserved when compared to other genes encoding antigens in *P. vivax* (e.g., *pvmsp-7* [6], *pvmsp-5* [7,12], *pvmsp-3* [9,10], *pvmsp-1* [5,8]).

The *pv41* nucleotide diversity was low in the Colombian population; however, π values and haplotype number were dissimilar for each Colombian locality, suggesting different evolutionary histories possibly due to a structured population. However, this pattern could have been due to few samples having been collected from some locations. The use of neutral markers could lead to confirming whether Colombia has a structured population.

pv41 nucleotide diversity was higher than that reported for *pv12*, but similar to that found in *pv38* [29];

Table 3 Difference between the non-synonymous substitutions per non-synonymous site (d_N) and synonymous substitution per synonymous site (d_S) rate

n	s48/45 N-terminal	s48/45 C-terminal	Complete sequences
Worldwide isolates	d_N - d_S	d_N - d_S	d_N - d_S
36	−0.0001 ± 0.0008	0.0018 ± 0.0015	−0.0005 ± 0.0013
Colombian isolate			
30	0.0000 ± 0.0000	0.0024 ± 0.0015	0.0007 ± 0.0010

No statistically significant values were found.

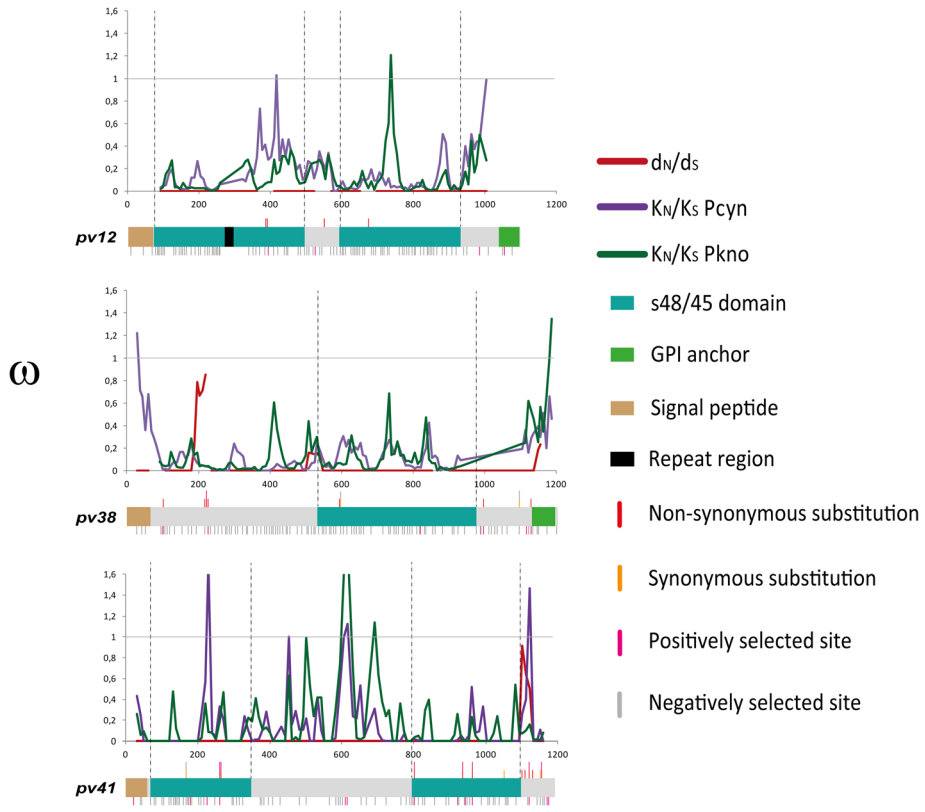


Figure 3 Sliding window of omega values ($\omega = d_N/d_S$ and K_N/K_S) for three members of the 6-Cys family expressed in merozoites. The ω values (d_N/d_S) for *Plasmodium vivax pv12*, *pv38* and *pv41* genes are shown in red, divergence ($\omega = K_N/K_S$) between *P. vivax* and *Plasmodium cynomolgi* (Pcyn) in purple and between *P. vivax* and *Plasmodium knowlesi* (Pkno) in green. The graphical representation of each gene is given below the respective sliding window, showing the position of the segregant sites and which of them were influenced by natural selection. Intraspecies selected sites can be seen in the upper part of each gene and inter-species selected sites are indicated in the lower part. The sites were identified by the Datamonkey server. The schemes for *pv12* and *pv38* have been modified from Forero-Rodríguez *et al.* [29].

however, fewer haplotypes were found in *pv41* compared to *pv38* (14 haplotypes have been reported for it in the Colombian population) [29]. Since the Pv41 protein has no membrane-anchoring domains, it could be interacting with proteins anchored to the merozoite surface. It has been shown that Pf12 and Pf41 proteins form an inverted heteroduplex on parasite membrane [25,26]. Due to these proteins' similarity, it is probable that Pv12 and Pv41 may also interact in *P. vivax*. This could explain the high degree of conservation found in Pv12 ($\pi = 0.0004 \pm 0.0001$ [29]). If Pv41 forms a protein complex with Pv12, the latter could be masked whilst Pv41 would be more exposed to a host's immune system, greater diversity thus being found in Pv41 ($\pi = 0.0037 \pm 0.0006$) regarding Pv12 ($\pi = 0.0004 \pm 0.0001$). Since such complex formation would be anti-parallel, the region

Table 4 McDonald-Kreitman test for evaluating the action of natural selection on the *p41* gene

	P. vivax/P. cynomolgi			*P. vivax/P. knowlesi*		
Worldwide isolates	**Fixed**	**Polymorphic**	**NI (p-values)**	**Fixed**	**Polymorphic**	**NI (p-values)**
Non-synonymous substitutions	45.62	11	4.45 (0.003)	61.95	11	4.12 (0.004)
Synonymous substitutions	110.71	6		138.81	6	
Colombian isolates						
Non-synonymous substitutions	46.69	8	9.65 (0.000)	63.06	8	8.80 (0.001)
Synonymous substitutions	112.65	2		138.81	2	

The McDonald-Kreitman test involved using the sequences obtained from the databases together with the Colombian ones (worldwide isolates), and just those obtained in the Colombian population (Colombian isolates). The data regarding divergence between species was obtained by comparing *P. vivax* sequences to that from two related species: *P. cynomolgi* and *P. knowlesi*. NI: neutral index.

Table 5 Difference between non-synonymous divergence per non-synonymous site (K_N) and synonymous divergence per synonymous site (K_S)

P. vivax/P. cynomolgi			
n	**s48/45 N-terminal**	**s48/45 C-terminal**	**Complete sequences**
Worldwide isolates	**K_N - K_S**	**K_N - K_S**	**K_N - K_S**
36	−0.0151 ± 0.0031*	−0.0107 ± 0.0032**	−0.0160 ± 0.0028*
Colombian isolates			
30	−0.0178 ± 0.0038*	- 0.0126 ± 0.0035**	−0.0174 ± 0.0030*
P. vivax/P. knowlesi			
n	**s48/45 N-terminal**	**s48/45 c-terminal**	**Complete sequences**
Worldwide isolates	**K_N - K_S**	**K_N - K_S**	**K_N - K_S**
36	−0.0196 ± 0.0036*	−0.0107 ± 0.0034**	−0.0185 ± 0.0031*
Colombian isolates			
30	−0.0233 ± 0.0042*	−0.0125 ± 0.0036**	−0.0217 ± 0.0035*

K_N - K_S difference was estimated using the sequences obtained from the databases together with the Colombian ones (worldwide isolates) and just with those obtained in the Colombian population (Colombian isolates).
n: number of isolates. *p <0.000; **p <0.001.

most exposed to Pv41 would be the C-terminal in which high fixation of non-synonymous substitutions was observed (Figure 3).

No significant values were found in the neutrality tests based on the polymorphism frequency spectrum or the haplotype-based tests (Table 2), meaning that the hypothesis regarding neutrality could not be ruled out. Such hypothesis stated that *pv41* haplotypes could be fixed in different populations thereby producing a population structure in this locus and new *pv41* haplotypes might thus appear if new parasites populations are evaluated.

No significant values were found when the effect of natural selection was evaluated by means of the difference between non-synonymous and synonymous substitutions (d_N - d_S) in either the whole gene or in each s48/45 domain (Table 3). However, the *pv41* sliding window gave a peak close to 1 at the 3′-end of the gene (Figure 3); several non-synonymous mutations would thus seem to be fixed in this region. The codon-based selection tests showed that seven out of the ten codons having mutations producing a change in the protein were positively selected (Figure 3). Three of these seven codons (V89E, H359S and G373R) produced radical substitutions (changing amino acid physical/chemical properties). The R355M substitution also produced a radical change but selection signals were not identified in this site. Such positively selected codons were predominantly found towards the gene's 3′-end (encoding the protein's C-terminal region) and could have been fixed to enable evading the immune system since this region would be more exposed due to the possible antiparallel formation of a Pv12/Pv41complex. Substitutions in codons 258, 301 and 312 located in the s48/45 domain could become deleterious due to them being able to alter the domain's structure; however, they had positive selection signals. Such substitutions were conservative and maintained the amino acids' physical-chemical characteristics, thereby enabling evasion of the immune system and maintaining the domain's structural conformation. Interspecies ω values were higher than 1 in some regions of *p41*, mainly outside s48/45 domains. Thirteen codons were positively selected at interspecies level; amino acid fixation would allow immune evasion of the respective host. Alternatively, positive sites found in s48/45 domains (which are involved in red blood cell invasion [17]) would be a P41 adaptation to the host receptor molecule.

The Z_{nS} test had significant values, indicating LD. The linked positions were found in the 3′-end of the gene. The mutations found there led to changes in protein sequence H359S, Y361F and D363N. The first substitution (H359S) produced a radical amino acid change, which was fixed by positive selection whilst the other two changes were conservative without selection signals. Since amino acid H359S was fixed by positive selection, this led to Y361F and D363N becoming fixed due to the short physical distance between them.

Genetic diversity in *pv41* was produced by point mutations (Figure 2); however, the recombination could also have been responsible for the genetic polymorphism found in this gene. The lineal regression between LD and nucleotide distance had a slight reduction in LD as nucleotide distance increased; this may have been a consequence of recombination processes. The ZZ test gave significant values, suggesting that recombination took place in this gene. Two minimum recombination sites were found and the GARD method (available from the Datamonkey web server) identified a recombination breakpoint in position 936, meaning that recombination produced new haplotypes in *pv41*.

The McDonald-Kreitman (MK) and omega divergence tests ($\omega = K_N/K_S$) were calculated for inferring natural

selection signals which might have influence the evolutionary history of *p41*. The latter was calculated for the gene's complete length and for each s48/45 domain. Significant values were found in the MK test throughout the whole gene (Table 5), polymorphism being greater than divergence; this could have resulted from weak negative selection or balancing positive selection. The latter is responsible for keeping allele variants (haplotypes) at intermediate frequencies as a mechanism for evading host immune responses; however, a major haplotype was found in the Colombian population whilst the rest occurred at low frequency. Due to the population structure reported in America [58], haplotype segregation could have led to different frequencies or new haplotypes could have diversified within American (or Colombian) subpopulations, meaning that if just one population is analysed, then balancing positive selection signals will not always be detected with population methods (Tajima, Fu and Li, Fay and Wu, Fu and K-test, and H-test). Alternatively, if balancing selection has resulted from frequency dependent selection, it would be expected that a haplotype would be presented as a major allele during a determined period of time and then become replaced by another less frequent one as an evasion mechanism. These haplotypes' frequency must therefore be evaluated during different intervals of time in several populations involving larger sampling.

On the other hand, the ω (K_N/K_S) rate sliding window showed that most values obtained throughout the gene were lower than 1, indicating high synonymous substitution fixation following *P. vivax*/*P. cynomolgi*/*P. knowlesi* divergence. The same pattern was observed in *pv12* and *pv38* (Figure 3 and [29]). The difference between non-synonymous and synonymous (K_N - K_S) divergence was estimated, giving significant negative values ($p < 0.001$) in *pv41* as well as in the s48/45 domains of this gene (Table 4). A large amount of negatively selected codons were identified which were preferentially located in the s48/45 domains (Figure 3). These results suggested that *p41* had diverged due to negative selection; such pattern was similar to that previously reported for other members of the *P. vivax* 6-Cys family [29,56]. *pv12* and *pv38*, like *pv41*, had various codons under negative selection at interspecies level which were preferentially located in the s48/45 domains (Figure 3). Such accumulation of interspecies synonymous substitutions suggested that evolution had tried to maintain domain structure in the different members of the 6-Cys family by eliminating all deleterious mutations due to the functional importance which these domains seem to have [17,59].

Conclusions

6-Cys family members seem to play a role during host cell recognition [17,59]. Due to the high degree of P12, P38 [29] and P41 protein conservation (at both intraspecies and interspecies level) given by the fixation of a large amount of synonymous substitutions, these three proteins may have evolved under strong functional constraints, possibly due to the presence of s48/45 domains which seem to have served as ligands for recognising the host cell [17,59,60]. Consequently, s48/45 domains should remain conserved as the resulting mutations could be deleterious; their evolution would thus have been slower regarding other functionally less important ones. Pv12, Pv38 and Pv41 thus warrant consideration as valuable candidates for developing a vaccine. However, a functional constraint does not imply that these regions may not vary. Pv41 s48/45 domains have been seen to have changes in their protein sequence, which seem to have been positively selected. Such changes conserve physical-chemical properties and thus structure/function may not become compromised, but could enable evasion of the immune response. Including Pv41 in a vaccine should thus be carefully evaluated due to the presence of variants in these regions.

This is also another aspect that must be taken into account when developing vaccines. It has been proposed that a completely effective vaccine requires the inclusion of both functional and conserved regions; however, vaccination could thus produce new selective pressure in these regions and parasites could fix mutations as an adaptation mechanism (in spite of their functional importance) and the appearance of new variants might thus reduce vaccine's efficacy.

Additional files

Additional file 1: DNA sequences from 30 isolates obtained in this study.

Additional file 2: Nucleotide diversity (π) values for subpopulations within Colombia. n: number of isolates, Ss: number of segregant sites, S: number of singleton sites, Ps: number of parsimony-informative sites, H: number of haplotypes, *k*: average number of nucleotide differences by sequence pairs, π: nucleotide diversity per site.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

JF-R devised the study, participated in designing it, performed the experiments, made the population genetics analysis and wrote the manuscript. DG-O devised and designed the study, helped perform the experiments, carried out the population genetics analysis and wrote the manuscript. MAP coordinated the study, and helped to write the manuscript. All the authors have read and approved the final manuscript.

Acknowledgements

We would like to thank Dr Sylvain Mousset for providing the Allelix software for our analysis, Professor María Isabel Chacón for her comments and suggestions and Jason Garry for translating this manuscript. This work was financed by the "Departamento Administrativo de Ciencia, Tecnología e Innovación (Colciencias)" through contracts RC #0309–2013 and 709–2013. JF-R received funding from COLCIENCIAS via cooperation agreement #0719–2013.

Author details
[1]Fundación Instituto de Inmunología de Colombia (FIDIC), Carrera 50 No. 26-20, Bogotá, DC, Colombia. [2]Microbiology Postgraduate Programme, Universidad Nacional de Colombia, Bogotá, DC, Colombia. [3]School of Medicine and Health Sciences, Universidad del Rosario, Bogotá, DC, Colombia.

Received: 5 August 2014 Accepted: 27 September 2014
Published: 30 September 2014

References

1. WHO: *World Malaria Report 2013.* Geneva: World Health Organization; 2013. http://www.who.int/malaria/publications/world_malaria_report_2013/wmr2013_no_profiles.pdf.
2. Carvalho LJ, Daniel-Ribeiro CT, Goto H: **Malaria vaccine: candidate antigens, mechanisms, constraints and prospects.** *Scand J Immunol* 2002, **56:**327–343.
3. Jones TR, Hoffman SL: **Malaria vaccine development.** *Clin Microbiol Rev* 1994, **7:**303–310.
4. Patarroyo MA, Calderon D, Moreno-Perez DA: **Vaccines against *Plasmodium vivax*: a research challenge.** *Expert Rev Vaccines* 2012, **11:**1249–1260.
5. Figtree M, Pasay CJ, Slade R, Cheng Q, Cloonan N, Walker J, Saul A: ***Plasmodium vivax* synonymous substitution frequencies, evolution and population structure deduced from diversity in AMA 1 and MSP 1 genes.** *Mol Biochem Parasitol* 2000, **108:**53–66.
6. Garzon-Ospina D, Lopez C, Forero-Rodriguez J, Patarroyo MA: **Genetic diversity and selection in three *Plasmodium vivax* merozoite surface protein 7 (Pvmsp-7) genes in a Colombian population.** *PLoS One* 2012, **7:**e45962.
7. Gomez A, Suarez CF, Martinez P, Saravia C, Patarroyo MA: **High polymorphism in Plasmodium vivax merozoite surface protein-5 (MSP5).** *Parasitology* 2006, **133:**661–672.
8. Kang JM, Ju HL, Kang YM, Lee DH, Moon SU, Sohn WM, Park JW, Kim TS, Na BK: **Genetic polymorphism and natural selection in the C-terminal 42 kDa region of merozoite surface protein-1 among *Plasmodium vivax* Korean isolates.** *Malar J* 2012, **11:**206.
9. Mascorro CN, Zhao K, Khuntirat B, Sattabongkot J, Yan G, Escalante AA, Cui L: **Molecular evolution and intragenic recombination of the merozoite surface protein MSP-3alpha from the malaria parasite *Plasmodium vivax* in Thailand.** *Parasitology* 2005, **131:**25–35.
10. Ord R, Polley S, Tami A, Sutherland CJ: **High sequence diversity and evidence of balancing selection in the Pvmsp3alpha gene of *Plasmodium vivax* in the Venezuelan Amazon.** *Mol Biochem Parasitol* 2005, **144:**86–93.
11. Putaporntip C, Jongwutiwes S, Seethamchai S, Kanbara H, Tanabe K: **Intragenic recombination in the 3′ portion of the merozoite surface protein 1 gene of *Plasmodium vivax*.** *Mol Biochem Parasitol* 2000, **109:**111–119.
12. Putaporntip C, Udomsangpetch R, Pattanawong U, Cui L, Jongwutiwes S: **Genetic diversity of the *Plasmodium vivax* merozoite surface protein-5 locus from diverse geographic origins.** *Gene* 2010, **456:**24–35.
13. Richie TL, Saul A: **Progress and challenges for malaria vaccines.** *Nature* 2002, **415:**694–701.
14. Takala SL, Plowe CV: **Genetic diversity and malaria vaccine design, testing and efficacy: preventing and overcoming 'vaccine resistant malaria'.** *Parasite Immunol* 2009, **31:**560–573.
15. Polley SD, Tetteh KK, Lloyd JM, Akpogheneta OJ, Greenwood BM, Bojang KA, Conway DJ: ***Plasmodium falciparum* merozoite surface protein 3 is a target of allele-specific immunity and alleles are maintained by natural selection.** *J Infect Dis* 2007, **195:**279–287.
16. Arnott A, Barry AE, Reeder JC: **Understanding the population genetics of *Plasmodium vivax* is essential for malaria control and elimination.** *Malar J* 2012, **11:**14.
17. Garcia J, Curtidor H, Pinzon CG, Vanegas M, Moreno A, Patarroyo ME: **Identification of conserved erythrocyte binding regions in members of the *Plasmodium falciparum* Cys6 lipid raft-associated protein family.** *Vaccine* 2009, **27:**3953–3962.
18. Nagao E, Seydel KB, Dvorak JA: **Detergent-resistant erythrocyte membrane rafts are modified by a *Plasmodium falciparum* infection.** *Exp Parasitol* 2002, **102:**57–59.
19. Sanders PR, Gilson PR, Cantin GT, Greenbaum DC, Nebl T, Carucci DJ, McConville MJ, Schofield L, Hodder AN, Yates JR 3rd, Crabb BS: **Distinct protein classes including novel merozoite surface antigens in Raft-like membranes of *Plasmodium falciparum*.** *J Biol Chem* 2005, **280:**40169–40176.
20. Barrero CA, Delgado G, Sierra AY, Silva Y, Parra-Lopez C, Patarroyo MA: **Gamma interferon levels and antibody production induced by two PvMSP-1 recombinant polypeptides are associated with protective immunity against *P. vivax* in Aotus monkeys.** *Vaccine* 2005, **23:**4048–4053.
21. Richards JS, Beeson JG: **The future for blood-stage vaccines against malaria.** *Immunol Cell Biol* 2009, **87:**377–390.
22. Angel DI, Mongui A, Ardila J, Vanegas M, Patarroyo MA: **The *Plasmodium vivax* Pv41 surface protein: identification and characterization.** *Biochem Biophys Res Commun* 2008, **377:**1113–1117.
23. Mongui A, Angel DI, Guzman C, Vanegas M, Patarroyo MA: **Characterisation of the *Plasmodium vivax* Pv38 antigen.** *Biochem Biophys Res Commun* 2008, **376:**326–330.
24. Moreno-Perez DA, Areiza-Rojas R, Florez-Buitrago X, Silva Y, Patarroyo ME, Patarroyo MA: **The GPI-anchored 6-Cys protein Pv12 is present in detergent-resistant microdomains of *Plasmodium vivax* blood stage schizonts.** *Protist* 2013, **164:**37–48.
25. Taechalertpaisarn T, Crosnier C, Bartholdson SJ, Hodder AN, Thompson J, Bustamante LY, Wilson DW, Sanders PR, Wright GJ, Rayner JC, Cowman AF, Gilson PR, Crabb BS: **Biochemical and functional analysis of two *Plasmodium falciparum* blood-stage 6-cys proteins: P12 and P41.** *PLoS One* 2012, **7:**e41937.
26. Tonkin ML, Arredondo SA, Loveless BC, Serpa JJ, Makepeace KA, Sundar N, Petrotchenko EV, Miller LH, Grigg ME, Boulanger MJ: **Structural and biochemical characterization of *Plasmodium falciparum* 12 (Pf12) reveals a unique interdomain organization and the potential for an antiparallel arrangement with Pf41.** *J Biol Chem* 2013, **288:**12805–12817.
27. Cheng Y, Lu F, Tsuboi T, Han ET: **Characterization of a novel merozoite surface protein of *Plasmodium vivax*, Pv41.** *Acta Trop* 2013, **126:**222–228.
28. Chen JH, Jung JW, Wang Y, Ha KS, Lu F, Lim CS, Takeo S, Tsuboi T, Han ET: **Immunoproteomics profiling of blood stage *Plasmodium vivax* infection by high-throughput screening assays.** *J Proteome Res* 2010, **9:**6479–6489.
29. Forero-Rodriguez J, Garzon-Ospina D, Patarroyo MA: **Low genetic diversity and functional constraint in loci encoding *Plasmodium vivax* P12 and P38 proteins in the Colombian population.** *Malar J* 2014, **13:**58.
30. Imwong M, Pukrittayakamee S, Gruner AC, Renia L, Letourneur F, Looareesuwan S, White NJ, Snounou G: **Practical PCR genotyping protocols for *Plasmodium vivax* using Pvcs and Pvmsp1.** *Malar J* 2005, **4:**20.
31. Neafsey DE, Galinsky K, Jiang RH, Young L, Sykes SM, Saif S, Gujja S, Goldberg JM, Young S, Zeng Q, Chapman SB, Dash AP, Anvikar AR, Sutton PL, Birren BW, Escalante AA, Barnwell JW, Carlton JM: **The malaria parasite *Plasmodium vivax* exhibits greater genetic diversity than *Plasmodium falciparum*.** *Nat Genet* 2012, **44:**1046–1050.
32. Bozdech Z, Mok S, Hu G, Imwong M, Jaidee A, Russell B, Ginsburg H, Nosten F, Day NP, White NJ, Carlton JM, Preiser PR: **The transcriptome of *Plasmodium vivax* reveals divergence and diversity of transcriptional regulation in malaria parasites.** *Proc Natl Acad Sci U S A* 2008, **105:**16290–16295.
33. Pacheco MA, Battistuzzi FU, Junge RE, Cornejo OE, Williams CV, Landau I, Rabetafika L, Snounou G, Jones-Engel L, Escalante AA: **Timing the origin of human malarias: the lemur puzzle.** *BMC Evol Biol* 2011, **11:**299.
34. Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput.** *Nucleic Acids Res* 2004, **32:**1792–1797.
35. Suyama M, Torrents D, Bork P: **PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments.** *Nucleic Acids Res* 2006, **34:**W609–W612.
36. Librado P, Rozas J: **DnaSP v5: a software for comprehensive analysis of DNA polymorphism data.** *Bioinformatics* 2009, **25:**1451–1452.
37. Depaulis F, Veuille M: **Neutrality tests based on the distribution of haplotypes under an infinite-site model.** *Mol Biol Evol* 1998, **15:**1788–1790.
38. Tajima F: **Statistical method for testing the neutral mutation hypothesis by DNA polymorphism.** *Genetics* 1989, **123:**585–595.
39. Fu YX, Li WH: **Statistical tests of neutrality of mutations.** *Genetics* 1993, **133:**693–709.
40. Fay JC, Wu CI: **Hitchhiking under positive Darwinian selection.** *Genetics* 2000, **155:**1405–1413.

41. Fu YX: **Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection.** *Genetics* 1997, **147:**915–925.
42. Zhang J, Rosenberg HF, Nei M: **Positive Darwinian selection after gene duplication in primate ribonuclease genes.** *Proc Natl Acad Sci U S A* 1998, **95:**3708–3713.
43. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S: **MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods.** *Mol Biol Evol* 2011, **28:**2731–2739.
44. Kosakovsky Pond SL, Frost SD: **Not so different after all: a comparison of methods for detecting amino acid sites under selection.** *Mol Biol Evol* 2005, **22:**1208–1222.
45. Pond SL, Frost SD, Grossman Z, Gravenor MB, Richman DD, Brown AJ: **Adaptation to different human populations by HIV-1 revealed by codon-based analyses.** *PLoS Comput Biol* 2006, **2:**e62.
46. Murrell B, Wertheim JO, Moola S, Weighill T, Scheffler K, Kosakovsky Pond SL: **Detecting individual sites subject to episodic diversifying selection.** *PLoS Genet* 2012, **8:**e1002764.
47. Murrell B, Moola S, Mabona A, Weighill T, Sheward D, Kdosakovsky Pond SL, Scheffler K: **FUBAR: a fast, unconstrained bayesian approximation for inferring selection.** *Mol Biol Evol* 2013, **30:**1196–1205.
48. McDonald JH, Kreitman M: **Adaptive protein evolution at the Adh locus in Drosophila.** *Nature* 1991, **351:**652–654.
49. Egea R, Casillas S, Barbadilla A: **Standard and generalized McDonald-Kreitman test: a website to detect selection by comparing different classes of DNA sites.** *Nucleic Acids Res* 2008, **36:**W157–W162.
50. Jukes THCRC: **Evolution of Protein Molecules.** In *Mammalian Protein Metabolism.* Edited by Munro HN. New York: Academic Press; 1969.
51. Kelly JK: **A test of neutrality based on interlocus associations.** *Genetics* 1997, **146:**1197–1206.
52. Rozas J, Gullaud M, Blandin G, Aguade M: **DNA variation at the rp49 gene region of *Drosophila simulans*: evolutionary inferences from an unusual haplotype structure.** *Genetics* 2001, **158:**1147–1155.
53. Hudson RR, Kaplan NL: **Statistical properties of the number of recombination events in the history of a sample of DNA sequences.** *Genetics* 1985, **111:**147–164.
54. Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SD: **Automated phylogenetic detection of recombination using a genetic algorithm.** *Mol Biol Evol* 2006, **23:**1891–1901.
55. Delport W, Poon AF, Frost SD, Kosakovsky Pond SL: **Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology.** *Bioinformatics* 2010, **26:**2455–2457.
56. Doi M, Tanabe K, Tachibana S, Hamai M, Tachibana M, Mita T, Yagi M, Zeyrek FY, Ferreira MU, Ohmae H, Kaneko A, Randrianarivelojosia M, Sattabongkot J, Cao YM, Horii T, Torii M, Tsuboi T: **Worldwide sequence conservation of transmission-blocking vaccine candidate Pvs230 in *Plasmodium vivax*.** *Vaccine* 2011, **29:**4308–4315.
57. Reeder JC, Wapling J, Mueller I, Siba PM, Barry AE: **Population genetic analysis of the *Plasmodium falciparum* 6-cys protein Pf38 in Papua New Guinea reveals domain-specific balancing selection.** *Malar J* 2011, **10:**126.
58. Taylor JE, Pacheco MA, Bacon DJ, Beg MA, Machado RL, Fairhurst RM, Herrera S, Kim JY, Menard D, Povoa MM, Villegas L, Mulyanto, Snounou G, Cui L, Zeyrek FY, Escalante AA: **The evolutionary history of *Plasmodium vivax* as inferred from mitochondrial genomes: parasite genetic diversity in the Americas.** *Mol Biol Evol* 2013, **30:**2050–2064.
59. Cowman AF, Crabb BS: **Invasion of red blood cells by malaria parasites.** *Cell* 2006, **124:**755–766.
60. Arredondo SA, Cai M, Takayama Y, MacDonald NJ, Anderson DE, Aravind L, Clore GM, Miller LH: **Structure of the Plasmodium 6-cysteine s48/45 domain.** *Proc Natl Acad Sci U S A* 2012, **109:**6692–6697.

doi:10.1186/1475-2875-13-388
Cite this article as: Forero-Rodríguez *et al.*: Low genetic diversity in the locus encoding the *Plasmodium vivax* P41 protein in Colombia's parasite population. *Malaria Journal* 2014 **13**:388.

MALARIA JOURNAL

RESEARCH Open Access

Heterogeneous genetic diversity pattern in *Plasmodium vivax* genes encoding merozoite surface proteins (MSP) -7E, –7F and -7L

Diego Garzón-Ospina[1,2†], Johanna Forero-Rodríguez[1†] and Manuel A Patarroyo[1,2*]

Abstract

Background: The *msp-7* gene has become differentially expanded in the *Plasmodium* genus; *Plasmodium vivax* has the highest copy number of this gene, several of which encode antigenic proteins in merozoites.

Methods: DNA sequences from thirty-six Colombian clinical isolates from *P. vivax* (*pv*) *msp-7E*, *–7F* and *-7L* genes were analysed for characterizing and studying the genetic diversity of these *pvmsp-7* members which are expressed during the intra-erythrocyte stage; natural selection signals producing the variation pattern so observed were evaluated.

Results: The *pvmsp-7E* gene was highly polymorphic compared to *pvmsp-7F* and *pvmsp-7L* which were seen to have limited genetic diversity; *pvmsp-7E* polymorphism was seen to have been maintained by different types of positive selection. Even though these copies seemed to be species-specific duplications, a search in the *Plasmodium cynomolgi* genome (*P. vivax* sister taxon) showed that both species shared the whole *msp-7* repertoire. This led to exploring the long-term effect of natural selection by comparing the orthologous sequences which led to finding signatures for lineage-specific positive selection.

Conclusions: The results confirmed that the *P. vivax msp-7* family has a heterogeneous genetic diversity pattern; some members are highly conserved whilst others are highly diverse. The results suggested that the 3'-end of these genes encode MSP-7 proteins' functional region whilst the central region of *pvmsp-7E* has evolved rapidly. The lineage-specific positive selection signals found suggested that mutations occurring in *msp-7s* genes during host switch may have succeeded in adapting the ancestral *P. vivax* parasite population to humans.

Keywords: Plasmodium vivax, msp-7, Genetic diversity, Natural selection, Selective sweep

Background

Malaria remains a major public health problem worldwide. *Plasmodium falciparum* is the parasite species causing the lethal form of the disease whilst *Plasmodium vivax* has long been considered a parasite causing mild disease, thereby diverting attention away from this species regarding research; however, recent studies have reported that this species also causes severe clinical syndromes [1,2]. Even though both species infect humans, they both emerged from different evolutionary lineages; whilst *P. vivax* shares a common ancestor with Asian non-human primate malaria, *P. falciparum* has diverged from parasites infecting great apes [3].

The different evolutionary paths leading to the appearance of *P. vivax* and *P. falciparum* have also led to important differences regarding hosts being invaded by both species [4,5]. In spite of such differences, initial interaction between the parasite and red blood cells (RBC) seems to be directed by the MSP-1 protein [6-8] which is present in all species from the genus. MSP-1 forms a complex with MSP-6 and MSP-7 in *P. falciparum* [9-11]; the latter protein is encoded by a gene forming part of a multigene family which has been differentially expanded amongst *Plasmodium* species [12]. Studies involving *msp-7* family members have shown that the resulting protein products are located on the parasite membrane and that a 22 kDa C-terminal fragment (derived

* Correspondence: mapatarr.fidic@gmail.com
†Equal contributors
[1]Molecular Biology and Immunology Department, Fundación Instituto de Inmunología de Colombia (FIDIC), Carrera 50 No. 26-20, Bogotá, DC, Colombia
[2]Basic Sciences Department, School of Medicine and Health Sciences, Universidad del Rosario, Carrera 24 No. 63C-69, Bogotá, DC, Colombia

from proteolytic processing during parasite development) [10] has regions interacting with RBC [13]. The *msp-7* knockout in *P. falciparum* (*pfmsp-7I*) and *Plasmodium berghei* (*pbmsp-7B*) has shown that even though its absence is not lethal, it does reduce mutant parasite invasion ability [14,15]. These results, together with prior *in silico* analysis, have suggested that the members of this family could have functional redundancy [12,15,16] and their protein products (or some of them) could thus be involved in invasion. On the other hand, antigenicity studies have shown that some of these genes' protein products are recognized by sera from infected patients [17,18]. Antibodies directed against these proteins can inhibit parasite invasion of RBC [19], whilst immunization with members of the *Plasmodium yoelii msp-7* family has shown that they can confer protection in vaccinated mice following experimental challenge [20].

The genetic variability patterns observed in *msp-7* family members have been different between *P. falciparum* and *P. vivax* [21-24]; whilst members of the former species have low polymorphism [23,24], some members of *P. vivax* (*pvmsp-7C*, *pvmsp-7H* and *pvmsp-7I*) are highly polymorphic [21]. However, other members, such as *pvmsp-7A* and *pvmsp-7K*, are amongst the most conserved *P. vivax* antigens [22]. There are thirteen *msp-7* genes in this species' chromosome 12; these have been named in alphabetical order according to their location regarding the PVX_082640 gene [12]. Eleven of these genes are transcribed, but only seven of them are transcribed during the last hours of the intra-erythrocyte stage [25]. The genetic diversity of four of these seven genes has already been evaluated [21,22]; this study was therefore aimed at evaluating the genetic variability of the three remaining members (*pvmsp-7E*, *pvmsp-7F* and *pvmsp-7L*) which are expressed during the intra-erythrocyte stage. *pvmsp-7E* displayed high polymorphism and its central region had undergone rapid evolution whilst *pvmsp-7F* and *pvmsp-7L* were seen to be highly conserved. The genes' 3′-ends tended to be conserved by negative selection, suggesting that they encode the functional region for these proteins. Similar to what happened with the *msp-1* gene [26,27], *msp-7* genes seem to have diverged due to positive selection, which could have resulted from malaria parasites adaptation to different hosts.

Methods

Ethics statement

All *P. vivax*-infected patients who provided us with the blood samples were informed about the purpose of the study and all gave their written consent. All procedures carried out in this study were approved by the ethics committee of the Fundación Instituto de Inmunología de Colombia.

Parasite DNA and genotyping

Thirty-six peripheral blood samples from patients proving positive for *P. vivax* malaria by microscope examination were collected from some of Colombia's departments (Chocó and Nariño in the south-west, Guainía, Guaviare and Meta in the south-east, Tolima in the Midwest, and Atlántico, Antioquia and Córdoba in the north-west) between 2007 and 2010 (nine isolates in 2007, seven isolates in 2008, 8 isolates in 2009 and twelve isolates in 2010). DNA was obtained using a Wizard Genomic DNA Purification kit (Promega), following the manufacturer's instructions, and stored at −20°C until use. The parasite samples were genotyped by PCR-RFLP of the *pvmsp-1* gene as previously described [28]. Samples having single *P. vivax msp-1* allele infection were used for PCR amplification.

PCR amplification and sequencing

Primers were designed for amplifying *pvmsp-7E*, *pvmsp-7F* and *pvmsp-7L* DNA fragments, based on Sal-I sequences (PlasmoDB IDs: PVX_082665, PVX_082670 and PVX_082700, respectively). The *pvmsp-7E* gene fragment was amplified with 7Edto 5′ GCCGATCTGTTGTCTT TTCC 3′ and 7Erev 5′ CCTTACGACACGTCAAATGG 3′ primers. *pvmsp-7F* was amplified by using 7Fdto 5′ TCCTCTCCTTGCTGATACTCC 3′ and 7Frev 5′ CAGC CGCTTAAATCACTTC 3′ primers whilst *pvmsp-7L* was amplified with 7Ldto 5′ AGTACTATTCTTCTTGCCG TCC 3′ and 7Lrev 5′ TCCCCTCAGTAGTAAAACATCG 3′ primers. All PCR reactions were performed using KAPA HiFi HotStart Readymix containing 0.3 μM of each primer in a final 25 μL volume. Thermal conditions were set as follows: one cycle of 5 min at 95°C, 30 cycles of 20 sec at 98°C, 15 sec at 63°C, 30 sec at 72°C, followed by a 5 min final extension at 72°C. PCR products were purified using the UltraClean PCR Clean-up (MO BIO) kit, and then sequenced with a BigDye Terminator kit (Macrogen, Seoul, South Korea) in both directions. Three PCR products obtained from independent PCR amplifications were sequenced per isolate to discard errors. Sequences having a different haplotype to the previously reported ones were deposited in the GenBank database (accession numbers KM212276 - KM212302).

Phylogenetic analysis for *Plasmodium cynomolgi msp-7* orthologous identification

A similar approach used for *msp-7* identification in other *Plasmodium* species [12] was adopted for identifying *msp-7* genes in *Plasmodium cynomolgi* (*pc*) and establishing their orthologous relationships. The genomic region (obtained from GenBank, accession number: NC_020405) encoded by the PCYB_122860 and PCYB_122720 genes (homologues to PVX_082640 and PVX_082715 which circumscribed the *msp-7* region in *P. vivax*) was analysed

using ORF Finder [29] and Gene Runner software for identifying open reading frames (ORFs) encoding proteins larger than 300 amino acids. Deduced amino acid sequences obtained with Gene Runner were aligned with *P. vivax* (12 proteins) and *Plasmodium knowlesi* (5 proteins) MSP-7 sequences using the MUSCLE algorithm [30]. The best model for amino acid substitutions was selected by Akaike's information criterion using the ProtTest program [31]. Phylogenetic trees were inferred through Maximum Likelihood (ML) and Bayesian (BY) methods using the JTT+G model. The observed amino acid frequencies (JTT+G+F) were also considered in Bayesian phylogenetic analysis and the analysis was run for one million generations. ML topology reliability was evaluated by bootstrap, using 1,000 iterations, whilst the sump and sumt commands in Bayesian analysis were used for tabulating posterior probability and building consensus trees. MEGA v.5 software was used for ML analysis and MrBayes v.3.2 software for assessing Bayesian inference. The *P. falciparum* MSP-7H (*Pf*MSP-7H) sequence was used as outgroup in both methods.

DNA diversity and evolutionary analysis in *pvmsp-7* genes

CLC Main workbench (CLC bio, Cambridge, MA, USA) software was used to assemble forward and reverse sequences from three independent PCR fragments per isolate. Deduced amino acids from Colombian isolates' *pvmsp-7* sequences and those obtained from several sequencing projects (Sal-I, Brazil-I, Mauritania-I, India-VII and North Korean reference sequences) [4,32] were aligned using the MUSCLE algorithm [30], followed by manual editing. PAL2NAL software [33] was then used for inferring codon alignments from the aligned amino acid sequences. The T-REKS algorithm [34] was used for searching for repeats having 90% similarity with the deduced *msp-7* amino acid sequences.

DnaSP v.5 software [35] was used for calculating the number of polymorphic segregating sites (Ss), the number of singleton sites (s), the number of parsimony-informative sites (Ps), the number of haplotypes (H), the Watterson estimator (θ^w) and nucleotide diversity per site (π) for all available sequences (reference sequences and Colombian ones), as well as for the Colombian population alone. Departure from the neutral model was assessed in the Colombian population by frequency spectrum-based tests (Tajima's D, Fu and Li's D* and F* statistics Fay and Wu's H) and tests based on the distribution of haplotypes (Fu's Fs and K-tests and H-test (for the latter test haplotype diversity obtained from DnaSP software was multiplied by (n-1)/n according to Depaulis and Veuille [35,36])). DnaSP v.5 and/or ALLELIX software were used for these tests, coalescent simulations being used for obtaining confidence intervals [35]. Positions containing gaps or repeats in the alignment were not taken into account.

Natural selection was assessed by using the modified Nei-Gojobori method [37] which calculated non-synonymous (d_N) and synonymous (d_S) rate substitution. Differences between d_N and d_S were assessed by applying Fisher's exact test (suitable for d_N and $d_S < 10$ [38]) and the Z-test available in MEGA software v.5 [39]. The Datamonkey web server [40] was used for assessing codon sites under positive or negative selection at population level, along with the IFEL codon-based maximum likelihood method [41]. Positive or negatively selected sites were also assessed by FEL, SLAC, REL [42], MEME [43] and FUBAR [44] methods. A <0.1 p-value was considered significant for IFEL, FEL, SLAC and MEME methods, a >50 Bayes factor for REL and a >0.9 posterior probability for FUBAR. Recombination was considered before running these tests. The branch-site REL method was used for identifying branches (lineages) when a percentage of sites have evolved under positive selection for exploring the long-term selection effect. Non-synonymous divergence (K_N) and synonymous divergence (K_S) rate substitutions were also calculated using the modified Nei-Gojobori method [37] with Jukes-Cantor correction [45]. Positive and negative selection at every codon for *P. vivax*/*P. cynomolgi msp-7* alignments were also evaluated by FEL, SLAC, MEME, REL and FUBAR methods.

Linkage disequilibrium (LD) was evaluated by calculating the Z_{nS} statistic [46]. Linear regression between LD and nucleotide distances was evaluated to ascertain whether recombination was taking place in *pvmsp-7* genes. Recombination was also assessed by the GARD method [47] and by ZZ [48] and RM tests [49]. RDP3 v3.4 software was used for detecting recombinant fragments in *pvmsp-7* genes [50].

Results and discussion

Genotyping natural isolates

The thirty-six samples used in this study were genotyped by PCR-RFLP of the *pvmsp-1* marker. All samples were infected by a single strain (a single *P. vivax msp-1* allele was detected) and considered for PCR amplification of *pvmsp-7* genes. The RFLP pattern confirmed the presence of different genotypes in the isolates so obtained. In spite of all samples amplifying the *pvmsp-1* fragment, no amplimers were detected in some samples for some of the *msp-7* genes (*pvmsp-7E* n = 31, *pvmsp-7F* n = 36 and *pvmsp-7L* n = 31).

The *msp-7* family structure in *Plasmodium cynomolgi* and phylogenetic analysis

Prior analysis has suggested that there are several *msp-7* species-specific duplications in *P. vivax* [12]. The recent sequencing of the *P. cynomolgi* genome [51], a species phylogenetically close to *P. vivax* [3], has meant that

new sequences from this multigene family are now available. The *P. cynomolgi* genomic region flanked by the PCYB_122860 and PCYB_122720 genes contained eleven 0.9 to 1.4 Kb length ORFs having the same transcription orientation. A shorter 0.5 Kb fragment having 30% similarity with the identified ORFs was also observed. A 314 bp region having 75.8% identity with the 285 bp fragment in *P. vivax* between the *pvmsp-7I* and *pvmsp-7K* genes was also found (Figure 1). The *P. cynomolgi msp-7* genes (and/or fragments) were named in alphabetical order, according to their location regarding the PCYB_122860 gene (Figure 1). Contrasting with PlasmoDB annotation, our group found that PcMSP-7C, −7F, −7H, −7I, −7K proteins might be encoded by a single exon like *P. vivax* MSP-7K [52] (Additional file 1). The deduced amino acid sequences from these ORFs had a signal peptide, but no membrane-anchoring regions. The domain characteristic of the MSP-7 family (MSP_7C, Pfam domain ID: PF12948) was absent in the deduced PcMSP-7L protein sequences due to a premature stop codon.

Orthologous relationships were established for *P. cynomolgi* MSP-7 (PcMSP-7) sequences by inferring phylogenies, using these sequences together with *P. vivax* (Pv) and *P. knowlesi* (Pk) MSP-7 sequences (Figure 2) with the *P. falciparum* MSP-7H (PfMSP-7H) as outgroup. The topologies revealed eleven clades having good statistical support (Figure 2); each *P. cynomolgi* MSP-7 had a counterpart in *P. vivax* or *P. knowlesi*. However, clustering for PcMSP-7B and PcMSP-7E differed regarding the other MSP-7s. PcMSP-7E formed a group with PvMSP-7E (Figure 2A) in ML topology, even though not having very high statistical support (72%). The BY topology method (Figure 2B) gave a subgroup formed by PcMSP-7B and PcMSP-7E which appeared as an external PvMSP-7B/PvMSP-7E group suggesting that these genes are inparalogous and, therefore, the duplication events occurred after the divergence of *P. cynomolgi* and *P. vivax*. Even though *pcmsp-7B* and *pcmsp-7E* did not form a group with *pvmsp-7B* and *pvmsp-7E*, respectively, their location regarding PVX_082640 and PCYB_122860 genes was the same (Figure 1). Moreover, genetic distances between *pcmsp-7B* and *pcmsp-7E* (and/or *pvmsp-7B* and *pvmsp-7E*) were similar to *pcmsp-7B* and *pvmsp-7B* (or *pcmsp-7E* and *pvmsp-7E*). Furthermore, genetic distance between *pcmsp-7B* and *pvmsp-7B* was smaller than that between *pcmsp-7B* and *pvmsp-7E* and the distance between *pcmsp-7E* and *pvmsp-7E* was less than the distance between *pcmsp-7E* and *pvmsp-7B* (Additional file 2). Consequently, there is little probability that duplication events following the divergence of these two species independently led to the same order of *msp-7B* and *msp-7E* genes. Evidence of gene conversion (mechanism conducting paralogous homogenization) between *P. vivax msp-7* members has been reported previously [21]; if such mechanism occurs between *msp-7B* and *msp-7E* genes it would be expected that they would become clustered in the phylogeny and not with their counterparts from a sister taxon; however, further analysis is needed for confirming such hypothesis.

The aforementioned results have shown that *P. vivax* and *P. cynomolgi* share the whole *msp-7* repertoire described to date, suggesting that the duplications which gave rise to the *msp-7B*, *−7C*, *−7E*, *−7F*, *−7I*, *−7L* and *-7M* genes occurred before the divergence between *P. vivax* and *P. cynomolgi* and after their divergence from *P. knowlesi*, and thus *msp-7B*, *−7C*, *−7E*, *−7F*, *−7I*, *−7L* and *-7M* are not exclusive to *P. vivax*.

pvmsp-7E genetic diversity

168 segregant sites were found in the *pvmsp-7E* gene (Table 1), showing that this gene is highly polymorphic. The nucleotide diversity (π) estimated for this gene (Table 1) was comparable to that found in other genes encoding surface proteins (*pvmsp-1* [53], *pvmsp-3* [54] and *pvmsp-5* [55,56]) as well as other members of *pvmsp-7* family [21]. Even though genes such as *pvmsp-1*, *pvmsp-3* and *pvmsp-5* are highly polymorphic, their diversity at protein level is usually located in determined regions. These regions are usually immune response targets and have thus tended to evolve more rapidly, accumulating mutations which alter the protein sequence and thereby evading the host's immune system. Around 60% of the polymorphism found in *pvmsp-7E* was located at the gene's central region whilst the gene's ends were relatively high conserved (Additional files 3 and 4).

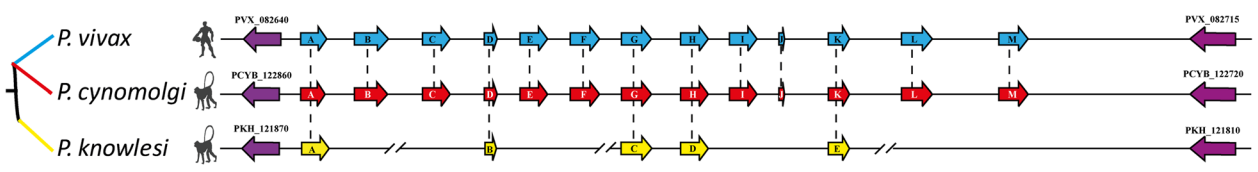


Figure 1 Schematic representation of the *msp-7* family in *Plasmodium vivax*, *P. cynomolgi* and *P. knowlesi*. The genes flanking the *msp-7* chromosome region in these three species are represented by purple boxes. The blue boxes represent *P. vivax msp-7* genes, the red ones represent *P. cynomolgi msp-7* genes and the yellow ones represent *P. knowlesi msp-7* genes. The genes are given in alphabetical order from left to right. The dashed lines connect orthologous genes. All genes are represented to scale, but in *P. cynomolgi* and *P. knowlesi* the distance between them is not representative.

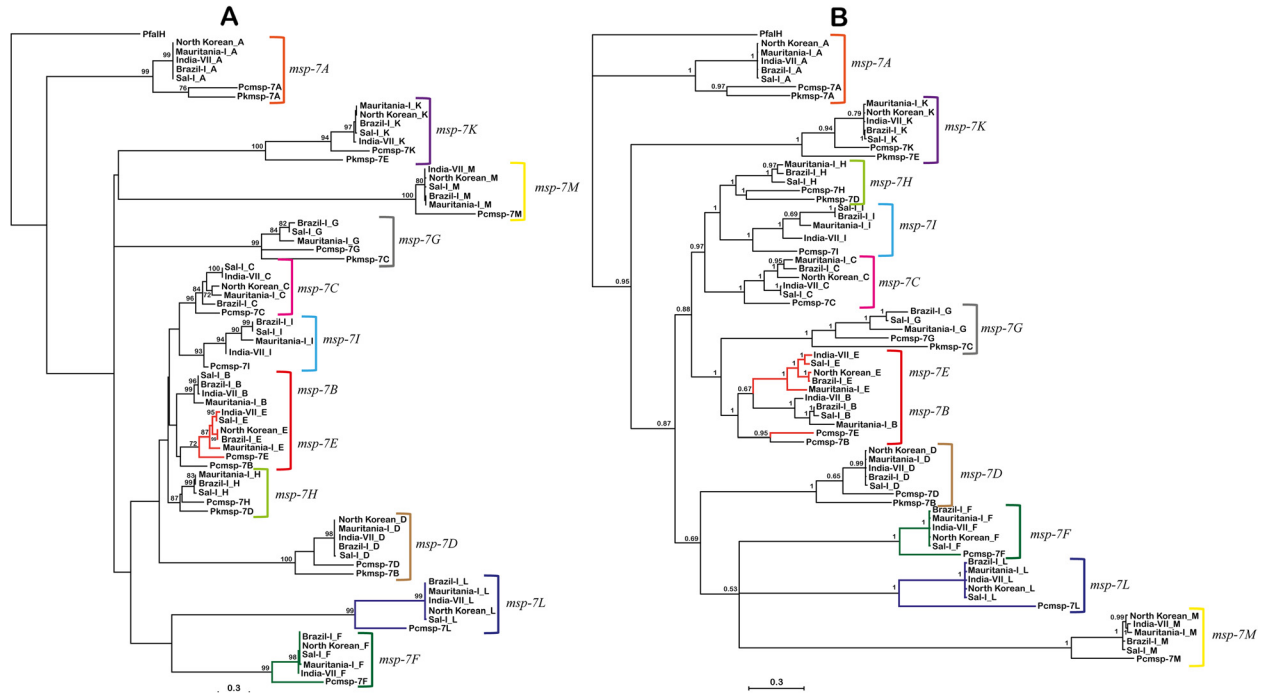


Figure 2 *msp-7* gene family phylogeny inferred by Maximum Likelihood (A) and Bayesian (B) methods for *P. vivax* and two closely related species. The coloured branches show the MSP-7E (red), MSP-7F (green) and MSP-7L (blue) sequences. Sequence clustering reflects orthologous relationships between the *msp-7* genes from these three species. The *P. falciparum* MSP-7F (PfalH) sequence was used as outgroup. Numbers over the branches are bootstrap and probability values.

This pattern has been previously observed in other *pvmsp*-7 genes [21].

Regarding DNA, 23 haplotypes have been found worldwide in *pvmsp-7E* (Table 1 and Additional file 4); fourteen different haplotypes have been found in this gene's 3′-end, whilst eleven haplotypes have been found in the central region and five in the 5′-end (Additional file 3). Nineteen of these 23 haplotypes were found in the Colombian population (Table 1 and Additional file 4; haplotype 10: 26%; haplotypes 5, 9: 15%; haplotypes 7, 11, 13, 18: 10%; and haplotypes 6, 8, 12, 14–17, 19–23: 5%) over the course of a 3-year period (2007–2010) without any longitudinal or spatial trends. Thirteen haplotypes were found in Colombia at amino acid level (Additional file 5); ten haplotypes having similar frequencies were observed in the *pvmsp-7E* central region in the Colombian population whilst three and four haplotypes were distinguished towards the N- and C-terminals, respectively (Additional file 5).

Table 1 DNA polymorphism measurements for *pvmsp-7* genes

n	Gene	Sites	Ss	S	Ps	H	θ^W (SD)	π (SD)
Worldwide genetic diversity								
35	***msp-7E***	1,044	168	8	160	23	0.0390 (0.0030)	0.0573 (0.0040)
41	***msp-7F***	1,164	3	1	2	5	0.0006 (0.0003)	0.0008 (0.0001)
36	***msp-7L***	1,212	5	1	4	7	0.0010 (0.0004)	0.0006 (0.0001)
Local genetic diversity								
31	***msp-7E***	1,044	164	6	158	19	0.0393 (0.0031)	0.0558 (0.0045)
36	***msp-7F***	1,176	2	0	2	4	0.0004 (0.0003)	0.0007 (0.0004)
31	***msp-7L***	1,212	4	0	4	6	0.0008 (0.0004)	0.0006 (0.0001)

Ss: number of segregating sites, S: number of singleton sites, Ps: number of parsimony-informative sites, H: number of haplotypes, θ^W: Watterson estimator, π: nucleotide diversity. (SD): standard deviation. Worldwide genetic diversity: the analysis involved the reference sequences together with the Colombian sequences. Local genetic diversity: Analysis for the Colombian sequences.

The T-REKS algorithm did not find repeats in the deduced protein sequences. Of the thirty-six sequences analysed for this gene, the North Korean sequence had a premature stop codon, suggesting that *pvmsp-7E* is a pseudogene in this strain; however, the annotation in the Broad institute for the North Korean *pvmsp-7E* gene (accession: PVNG_00513.1) suggests that this could have an intron. Further cDNA analysis is needed to confirm such issue regarding this strain.

pvmsp-7E neutrality and selection tests

Several tests based on the neutral model of molecular evolution were used with *pvmsp-7E* sequences from the Colombian population for evaluating whether this gene deviated from neutral expectations (Table 2 and Additional file 6). Tests based on the polymorphism frequency spectrum had significant values (Table 2). Fu and Li's D* and F* estimators

Table 2 Neutrality, linkage disequilibrium and recombination tests for *pvmsp-7* genes for the Colombian population

n	Gene	Tajima	Fu and Li		Fay and Wu's H	Fu's Fs	K-test	H-test (SD)	Z_{nS}	ZZ	RM
		D	D*	F*							
31	***msp-7E***	1.452	1.603**	1.839*	−51.002*	7.501**	18*	0.922 (0.02)*	0.246*	0.453*	12
36	***msp-7F***	1.401	0.783	1.112	−0.365	0.114	4	0.560 (0.07)	0.327	0.000	1
31	***msp-7L***	−0.610	1.054	0.658	0.656	−2.453	6	0.578 (0.08)	0.050	−0.031	1

msp-7E haplotype number (K-test) and diversity (H-test) were lower than expected under neutrality. (SD): standard deviation. *: p <0.05, **: p <0.02.

had values greater than zero whilst Fay and Wu's H estimator had statistically significant negative values (Table 2), indicating deviation from the neutral model of evolution. In addition, the haplotype distribution-based tests also gave statistically significant values; Fu's Fs test gave values greater than zero and the haplotype number (18) and haplotype diversity (0.922) were lower than that expected under neutrality (Table 2).

A sliding window for D, D*, F* and H statistics gave values greater than zero (D, D* and F*) in the gene's central region and lower than zero in the 3′-end, this being the region where the most negative value for H was located (Additional file 7). The gene was divided into three fragments: 5′-end (nucleotide 1 to 390), central (nucleotide 391 to 747) and 3′-end (nucleotide 748 to 1158); the aforementioned neutrality estimators were calculated for each of them (Additional file 6). Fu and Li's D* and Fu's Fs tests gave values greater than zero at the 5′-end whilst Tajima's D, Fu and Li's D*, F* and Fu's Fs test scores were greater than zero in the central region. The haplotype number and haplotype diversity were lower than expected under neutrality in the 5′-end and central region. Only Fay and Wu's H tests gave statistically significant values at the 3′-end (Additional file 6).

Natural selection seemed to act in different ways within the *pvmsp-7E* gene. A sliding window for the non-synonymous substitutions per non-synonymous site and synonymous substitutions per synonymous site rate ($d_N/d_S = \omega$) gave values greater than 1 ($\omega > 1$) in the central region and a peak at the 3′-end (Figure 3). The d_S rate at the 5′ and 3′-ends was significantly greater than d_N, whilst d_N was significantly greater than d_S in this gene's central region (Table 3). Eight positively selected sites were identified in the central region and one in the 3′-end. Twenty-six negatively selected sites were identified, mainly in the 5′-end (9 sites) and 3′-end (14 sites) (Additional files 4 and 8).

These results suggested that the central region was under natural positive selection. According to Tajima and Fu and Li tests (which had significant values higher than zero) *pvmsp-7E* seemed to be under balancing selection favouring the existence of different alleles in the population. This type of selection frequently occurs in antigens exposed to the immune system; immune responses therefore seemed to be directed towards the *pvmsp-7E* central region in which the mutations were accumulated at a greater rate by positive selection (Additional files 4 and 8).

Interestingly, as Fay and Wu's H tests gave significant negative values and as K-test and H-test were lower than expected under neutrality (Table 2), then a selective sweep would be probable and strong LD and low genetic diversity would thus be expected. Z_{nS} values suggested that *pvmsp-7E* had non-random polymorphism association, as expected in a selective sweep (Table 2). However, differently to what was expected, *pvmsp-7* had high genetic diversity (Table 1). When recombination is present in a locus under selective sweep, then it would be expected that genetic diversity would only become reduced close to the selection site [57]. There was evidence of recombination throughout the *pvmsp-7E* gene (Table 1 and Figure 4, see below) and since the deepest H "valley" was located at the 3′-end (Additional file 7), the selective sweep may not have affected the gene completely but just this region. The 5′-end and central regions showed no evidence of selective sweep (Additional file 6) and π was high in the central region and low at the 5′ extreme (Additional file 3). The 3′-end had significant values in H and Z_{nS} tests (Additional file 6), suggesting that the selection site should have been located in this region. The π value in 3′-end was considerably reduced regarding that for the central region (Additional file 3). However, the number of SNPs seems to be higher than that expected in a selective sweep. Our results suggested that a selective sweep affected the *pvmsp-7E* gene, the selection site was located in the gene's 3′-end and this seemed to be an incomplete selective sweep due to the presence of recombination since not all variability had been lost. However, if the selective sweep has not been recent, new mutations could have become fixed following such sweep [57].

Fu's Fs test gave values greater than zero (Table 2 and Additional file 6) which may have resulted from a reduction in haplotypes due to a recent bottleneck. Consequently, a low genetic diversity throughout the gene pool is expected in *P. vivax* Colombia population. Prior studies have shown high genetic diversity in parasitic antigens in the Colombian population [21,55], meaning that such demographic event is highly unlikely. This test's result may have been due to a reduction of *pvmsp-7E* haplotypes by the selective sweep, causing the number of haplotypes to be lower than that expected.

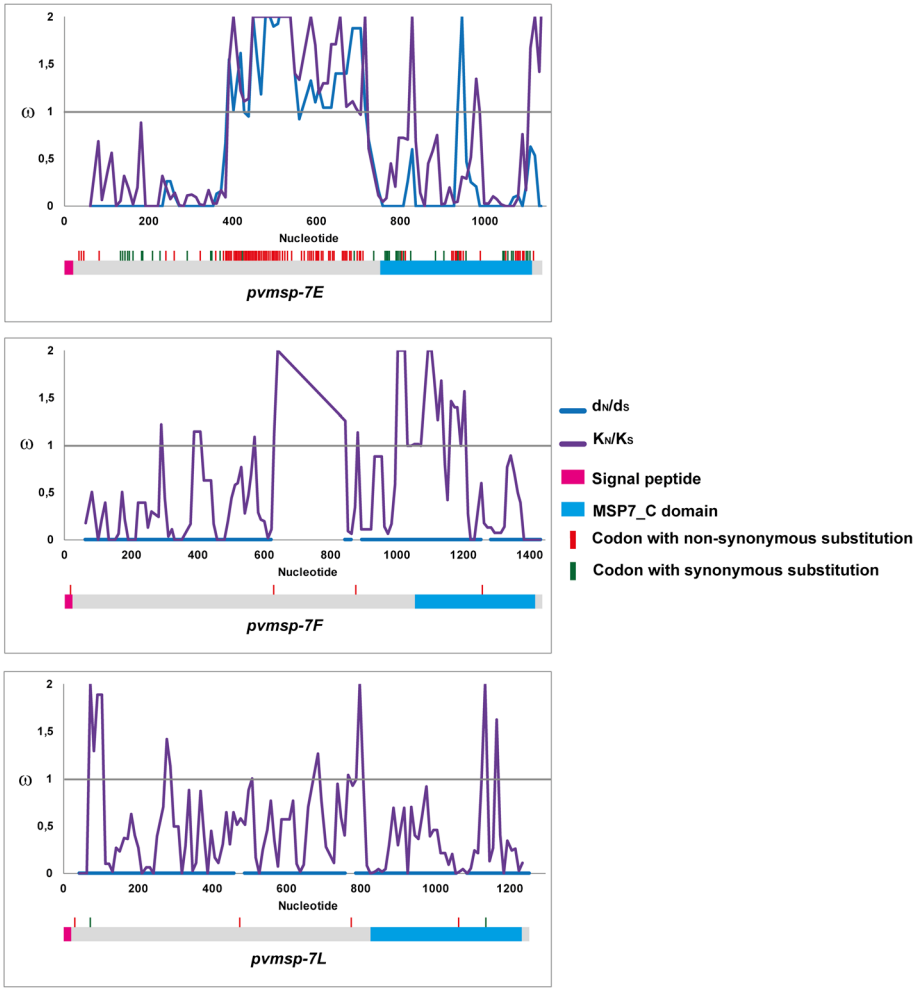


Figure 3 Sliding window for ω rate. *Plasmodium vivax msp-7E*, *msp-7F* and *msp-7L* genes' ω (d_N/d_S) values represented in blue, whilst the divergence omega (ω: K_N/K_S) between *P. vivax* and *P. cynomolgi* is shown in purple. A diagram of each gene is given below the sliding window. Intra-species non-synonymous substitutions (red) and synonymous substitutions (green) are shown by vertical lines above each gene.

Table 3 Average number of *pvmsp-7* gene synonymous substitutions per synonymous site (d_S) and non-synonymous substitutions per non-synonymous site (d_N)

n	Gene	5′-end		Central		3′-end		Full-length gene	
Worldwide isolates		**d_S (SE)**	**d_N (SE)**	**d_S (SE)**	**d_N (SE)**	**d_S (SE)**	**d_N (SE)**	**d_S (SE)**	**d_N (SE)**
35	***msp-7E***	0.0488 (0.0118)•	0.0082 (0.0032)	0.0717 (0.0163)	0.1573 (0.0134)•	0.0665 (0.0130)•	0.0171 (0.0045)	0.0626 (0.0078)	0.0551 (0.0055)
41	***msp-7F***	0.0000 (0.0000)	0.0000 (0.0000)	0.0000 (0.0000)	0.0013 (0.0012)	0.0000 (0.0000)	0.0018 (0.0018)	0.0000 (0.0000)	0.0011 (0.0007)^
36	***msp-7L***	0.0006 (0.0005)	0.0000 (0.0000)	0.0000 (0.0000)	0.0012 (0.0008)	0.0013 (0.0012)	0.0008 (0.0008)	0.0006 (0.0005)	0.0007 (0.0004)
Colombian isolates									
31	***msp-7E***	0.0473 (0.0115)•	0.0079 (0.0032)	0.0706 (0.0168)	0.1549 (0.0135)•	0.0633 (0.0124)•	0.0161 (0.0042)	0.0606 (0.0077)	0.0539 (0.0056)
36	***msp-7F***	0.0000 (0.0000)	0.0000 (0.0000)	0.0000 (0.0000)	0.0012 (0.0011)	0.0000 (0.0000)	0.0017 (0.0016)	0.0000 (0.0000)	0.0010 (0.0007)^
31	***msp-7L***	0.0000 (0.0000)	0.0000 (0.0000)	0.0000 (0.0000)	0.0012 (0.0009)	0.0010 (0.0010)	0.0009 (0.0008)	0.0004 (0.0004)	0.0007 (0.0004)

SE: standard error. 5′-end (*pvmsp-7E*: nucleotide 1–390, *pvmsp-7F*: nucleotide 1–432, *pvmsp-7L*: nucleotide 1–381), central (*pvmsp-7E*: nucleotide 391–747, *pvmsp-7F*: nucleotide 433–1,053, *pvmsp-7L*: nucleotide 382–816) and 3′-end (*pvmsp-7E*: nucleotide 748–1,158, *pvmsp-7F*: nucleotide 1,054–1,449, *pvmsp-7L*: nucleotide 817–1,275). Numbering based on Additional files 4, 9 and 10. ^: $p = 0.06$, •: $p < 0.001$.

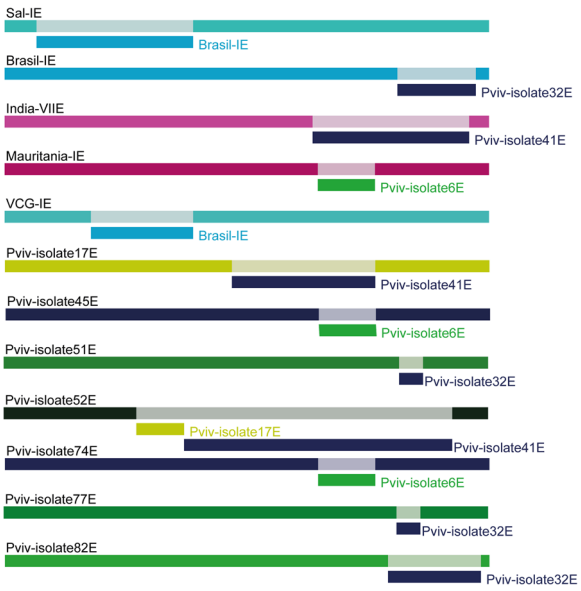


Figure 4 Schematic representation of recombination segments identified in *pvmsp-7E* by RDP3. Recombination fragments were only found in *pvmsp-7E*. Each variant is represented by a colour bar; a segment having a different colour below the bar represents the recombination event so identified.

pvmsp-7F and *pvmsp-7L* genetic diversity

In contrast to *pvmsp-7E*, *pvmsp-7F* and *pvmsp-7L* had low genetic diversity (Table 1). This genetic diversity was similar to that observed in *pvmsp-4* [58,59], *pvmsp-8* [60], *pvmsp-10* [22,60], *pv12*, *pv38* [61], *pv41*[62], as well as in *pvmsp-7A* and *pvmsp-7K* [22]. Aligning the Colombian sequences with those obtained from the databases (reference sequences, Additional files 9 and 10) revealed that these genes only had four and six segregant sites, respectively. The π values and the number of haplotypes for these genes were low (Table 1). The most frequently occurring *pvmsp-7F* allele in Colombia was haplotype 2 (61%), followed by haplotype 1 (19%), haplotype 3 (17%) and haplotype 4 (3%), whilst haplotype 1 (61%) was the most frequent for *pvmsp-7L*, followed by haplotype 3 (16%), haplotype 2 (14%) and haplotypes 4, 5 and 6 (3%).

pvmsp-7F and *pvmsp-7L* neutrality and selection tests

Neutrality for *pvmsp-7F* and *-7L* genes in the Colombian population could not be ruled out as no statistically significant values were found for the tests based on the neutral model of molecular evolution (Table 2 and Additional file 6). Likewise, no natural selection signals were found to be acting on these genes when d_N and d_S rates were calculated (Table 3). However, when the effect of selection on each codon was evaluated, it was seen that codon 424 regarding *pvmsp-7F*, was under positive selection. Concerning *pvmsp-7L*, codons 159, 260 and 357 showed positive selection signals (Additional files 8, 9 and 10).

Intragene linkage disequilibrium (LD) and recombination in *pvmsp-7* genes

As mentioned above, there were non-random associations regarding polymorphism for *pvmsp-7E* according to the Z_{nS} test (Table 2 and Additional file 6). No evidence of LD was found in *pvmsp-7F* or *pvmsp-7L* (Table 2 and Additional file 6), indicating that polymorphism within these genes was not associated. A linear regression between LD and nucleotide distance for *pvmsp-7s* gave a line sloping downwards as nucleotide distance increased in *pvmsp-7E*, suggesting intragene recombination. Twelve minimal recombination (RM) events were found for *pvmsp-7E* whilst only one RM was found in *pvmsp-7F* and *pvmsp-7L* (Table 2). The ZZ test and GARD method suggested recombination in *pvmsp-7E* (ZZ = $p < 0.05$ and GARD 2 breakpoints, $p < 0.0004$) but not in *pvmsp-7F* or *pvmsp-7L*. Figure 4 shows the fragments produced by recombination in *pvmsp-7E*.

pcmsp-7 and *pvmsp-7* genes appear to have diverged by positive selection

Natural selection's long-term effect on evolutionary history can be evaluated by comparing the orthologous genes from phylogenetically-related species [60-64]. *msp-7E* was highly divergent when compared to the *pvmsp-7E* and *pcmsp-7E* genes (Figure 3). In spite of *msp-7F* and *msp-7L* being highly conserved in *P. vivax*, they also have been shown to be highly divergent when compared to *P. cynomolgi* orthologous genes (Figure 3). The random effects branch-site model (Branch-site REL) was performed for determining how natural selection had acted during *P. vivax* and *P. cynomolgi* evolutionary history. This test displayed lineage-specific diversifying selection signals in *msp-7E* and *msp-7L* ($\omega > 1$, Figure 5). Moreover, the sliding window for the non-synonymous divergence per non-synonymous site rate (K_N) and the synonymous divergence per synonymous site rate (Ks) (divergent omega, $K_N/Ks = \omega$) gave highly divergent areas in the central region and 3′-end of these three genes (Figure 3). A statistically significant $K_N > K_S$ was found in the central region ($p < 0.001$) in *pvmsp-7E* (Table 4). No significant values were found in *msp-7F*, but in *msp-7L*, K_S was significantly higher than K_N (Table 4). However, when intraspecific polymorphism was compared to interspecific divergence using the McDonald-Kreitman test (MKT) no statistically significant values were found for these genes. The methods for estimating ω values for each codon (SLAC, FEL, REL, MEME and FUBAR) identified twenty-five (for *msp-7E*), four (for *msp-7F*) and seven (for *msp-7L*) codons under

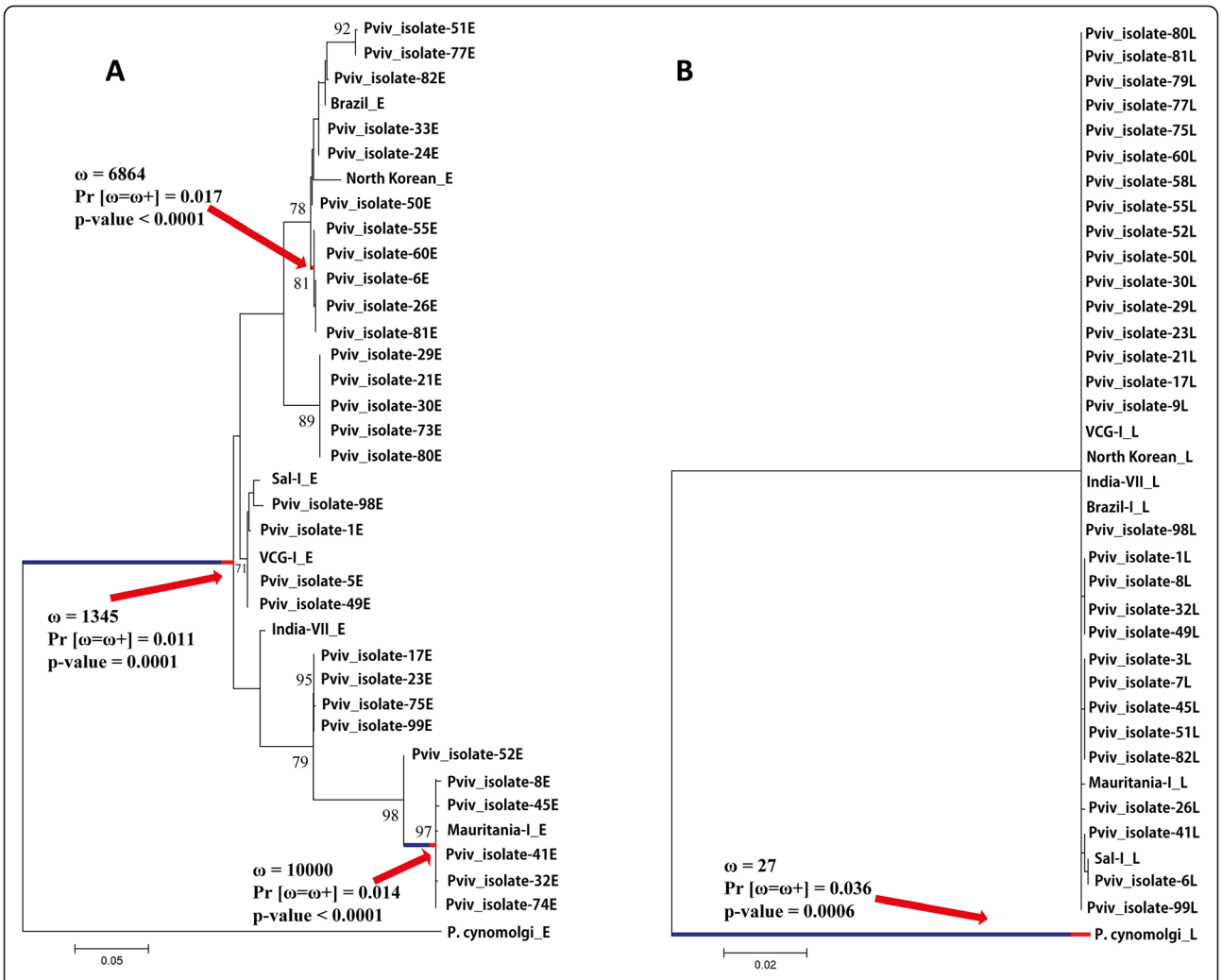


Figure 5 Positive lineage-specific selection in *msp-7E* (A) and *msp-7L* (B) genes. Branches under diversifying selection were identified by Branch-site REL method. ω values, the percentage of selected sites (Pr [ω = ω+]) and p-values are shown.

Table 4 Average number of *msp-7* gene synonymous divergence per synonymous site (K_S) and non-synonymous divergence per non-synonymous site (K_N)

n	Gene	5′ end		Central		3′ end		Full-length gene	
Worldwide isolates		K_S (SE)	K_N (SE)	K_S (SE)	K_N (SE)	K_S (SE)	K_N (SE)	K_S (SE)	K_N (SE)
36	***msp-7E***	0.0746 (0.0156)•	0.0126 (0.0034)	0.0813 (0.0185)	0.1976 (0.0191)•	0.0980 (0.0163)•	0.0299 (0.0056)	0.0833 (0.0092)	0.0676 (0.0066)
42	***msp-7F***	0.0098 (0.0023)†	0.0047 (0.0009)	0.0146 (0.0029)	0.0141 (0.0022)	0.0086 (0.0022)	0.0100 (0.0022)	0.0110 (0.0014)	0.0096 (0.0010)
37	***msp-7L***	0.0140 (0.0031)‡	0.0092 (0.0017)	0.0172 (0.0035)	0.0149 (0.0020)	0.0111 (0.0025)‡	0.0070 (0.0013)	0.0139 (0.0017)*	0.0102 (0.0010)
Colombian isolates									
32	***msp-7E***	0.0755 (0.0151)•	0.0128 (0.0032)	0.0808 (0.0183)	0.1963 (0.0191)•	0.0982 (0.0156)•	0.0305 (0.0053)	0.0834 (0.0089)	0.0676 (0.0065)
37	***msp-7F***	0.0111 (0.0027)†	0.0053 (0.0011)	0.0165 (0.0033)	0.0156 (0.0023)	0.0098 (0.0025)	0.0110 (0.0022)	0.0125 (0.0016)	0.0107 (0.0011)
32	***msp-7L***	0.0156 (0.0037)‡	0.0107 (0.0020)	0.0199 (0.0041)	0.0170 (0.0023)	0.0123 (0.0027)‡	0.0081 (0.0015)	0.0157 (0.0019)*	0.0118 (0.0011)

SE: standard error. 5′-end (*pvmsp-7E*: nucleotide 1–390, *pvmsp-7F*: nucleotide 1–432, *pvmsp-7L*: nucleotide 1–381), central (*pvmsp-7E*: nucleotide 391–747, *pvmsp-7F*: nucleotide 433–1,053, *pvmsp-7L*: nucleotide 382–816) and 3′-end (*pvmsp-7E*: nucleotide 748–1,158, *pvmsp-7F*: nucleotide 1,054–1,449, *pvmsp-7L*: nucleotide 817–1,275). Numbering based on Additional files 4, 9 and 10. ‡: p <0.09, *: p <0.04, †: p <0.02, •: p <0.001.

positive selection between *pvmsp-7* and *pcmsp-7* sequences (Additional files 4, 9, 10 and 11).

These results suggested that these genes have become diversified by positive selection; a similar pattern which have been reported for the *pvmsp-1* gene [26,27]. Divergence due to positive selection in *msp-1* coinciding with Asian macaque radiation [26,65] 3 to 6 million years ago means that divergence by positive selection in *msp-1* appears to be the result of adaptations to available new hosts [26,65]. *P. falciparum* MSP-1 and MSP-7 form a protein complex involved in invasion [9,10]. Assuming the formation of a protein complex between MSP-1 and MSP-7 in *P. vivax*, MSP-7s would be under the same selective pressures and may thus have evolved in a similar way. Theoretically [66,67], it has been suggested that a strong selective sweep may result in population differentiation at the hitchhiking locus, provided that the gene flow between these populations is low. Since malarial parasites could become diversified by sympatric events [68,69], *msp-7* (similar to *msp-1*) may have become diversified by positive selection (Figure 5) as a mechanism for adapting the ancestral *P. vivax* population to a new host during the switch to humans [70] and thus the selective sweep detected in *msp-7E* might have been an effect of such adaptation.

Negative selection within and between species supports the idea that the 3′-end encodes the functional region in MSP-7 proteins

In spite of divergence by positive selection, *msp-7* functional regions could have evolved more slowly due to their role during invasion and thus the accumulation of substitutions would have been mainly synonymous. $K_S > K_N$ was revealed in *msp-7E* and *msp-7L* when comparing *P. vivax* and *P. cynomolgi* sequences (Table 4). Fifty-seven sites were revealed to be under negative selection in *msp-7E*, twenty-four in *msp-7F* and thirty-six in *msp-7L* (Additional files 4, 9, 10 and 11). A large percentage of negatively selected sites were located in the gene's 3′-end encoding the *msp-7* family's characteristic domain (MSP7_C, Pfam domain ID: PF12948). The protein's C-terminal region encoded by these genes was highly conserved in *pvmsp-7A*, *−7C*, *−7H*, *−7I*, *−7K* [21,22], *−7E*, *−7F* and *-7L*; furthermore, this region has been conserved for a long period of time (2.6 to 5.2 million years ago [3]), at least in *msp-7E* (84.8% similarity between *P. vivax* and *P. cynomolgi*), *−7F* (86.8%) and *-7L* (95.4%). The negative selection signals identified at the 3'-end of these three genes (Additional files 8 and 11) suggested that the biological structure encoded by this region has been stable slowly evolving since divergence between *P. vivax* and *P. cynomolgi* due to its functional importance. These results support the idea that this region encodes this family's functional domain [21].

The *pvmsp-7* and *pcmsp-7* sequences have different gene structures

Marked differences were observed between *P. vivax* and *P. cynomolgi msp-7* genes. *pcmsp-7F* had a long insertion (one hundred ninety-two nucleotides) compared to *pvmsp-7F* (Additional file 9); however, the ORF remained open. *pcmsp-7L* had a premature stop codon caused by the deletion of one or two nucleotides from the sequence (Additional file 10). The protein encoded by this gene thus had no domain characteristic of this family (MSP7_C, Pfam domain ID: PF12948); however, many synonymous substitutions between species were observed in the region encoding this domain (the gene's 3`-end) when *P. vivax* and *P. cynomolgi* sequences were compared. Thirteen sites in this region were under negative selection in *msp-7L* (Additional files 10 and 11). The GeneScan algorithm [71] was then used for searching for exon/intron splice sites in *pcmsp-7F* and *pcmsp-7L* sequences. GeneScan analysis revealed regions which could act as donor and acceptor sequences in *pcmsp-7L* but not in *pcmsp-7F*. There was a thymine in *pcmsp-7L* nucleotide 609, whilst there was a cytosine in the homologous position in its orthologue in *P. vivax* (nucleotide 615 in the Sal-I sequence). Such change may have produced a putative donor (GT) site in *pcmsp-7L* whilst a putative acceptor site was located in position 1,030/1,031 (Additional file 12); an intron region was thus located in *pcmsp-7L* between nucleotides 608 and 1,031. Such exon-intron-exon structure in *pcmsp-7L* can be observed in the annotation of the *P. cynomolgi* genome available from PlasmoDB; however, the intron predicted in PlasmoDB was shorter than that predicted by GeneScan. This exon-intron-exon structure allowed *pcmsp-7L* to encode a protein having the MSP7_C domain.

Conclusions

Our results confirmed that the *P. vivax msp-7* family has a heterogeneous genetic diversity pattern. Some members were seen to be highly conserved whilst other had high genetic diversity. Consequently, *P. vivax msp-7* genes must have evolved differently from those in *P. falciparum* which have low polymorphism [23,24]. The PvMSP-7s C-terminal region (the gene's 3′-end) tended to be conserved within and between genes [21]. This region's conservation tended to be maintained by negative selection in *msp-7E*, *msp-7F* and *msp-7L*, suggesting that this is the functional region for this group of proteins. On the other hand, PvMSP-7 highly diverse members (*pvmsp-7C*, *−7H*, *−7I* [21] and *-7E*) were seen to have undergone rapid evolution at the protein's central region; immune responses would thus been directed towards this portion of the protein. New alleles have consequently arisen in the population and been maintained by balancing selection as a mechanism for evading an immune response. In addition to this type of evasion, the *P. vivax msp-7* family (similar to that suggested for the *pvmsp-3*

family [72]) would follow a model of multi-allele diversifying selection where functionally redundant paralogues [12] would increase evasion of the immune responses by antigenic diversity.

Our results have shown that *P. vivax* and *P. cynomolgi* share the whole *msp-7* repertoire described to date and have revealed lineage-specific positive selection signals which are similar to those reported for *pvmsp-1*. Mutations occurring in *msp-7s* genes during host switch may thus have succeeded in adapting the ancestral *P. vivax* parasite to humans.

Additional files

Additional file 1: Putative *P. cynomolgi msp-7* gene sequences obtained from chromosome 12, whole genome shotgun sequence GenBank accession number: NC_020405. ORF Finder and Gene Runner software were used to identify open reading frames encoding *P. cynomolgi* MSP-7 proteins.

Additional file 2: Genetic distance between *pcmsp-7B*, *pcmsp-7E* and *pvmsp-7B*, *pvmsp-7E* sequences from 5 *P. vivax isolates*. The number of nucleotide differences per site was estimated as well as the standard error regarding the *pvmsp-7E* and *pvmsp-7B* reference sequences and the *pcmsp-7E* and *pcmsp-7B* sequences.

Additional file 3: DNA polymorphism measurements at the 5′-end, central region and 3′-end for *pvmsp-7* genes in the Colombian population. Ss: number of segregating sites, S: number of singleton sites, Ps: number of parsimony-informative sites, H: number of haplotypes, θ^W: Watterson estimator, π: nucleotide diversity. (SD): standard deviation. 5′-end (*pvmsp-7E*: nucleotide 1–390, *pvmsp-7F*: nucleotide 1–432, *pvmsp-7L*: nucleotide 1–381), central (*pvmsp-7E*: nucleotide 391–747, *pvmsp-7F*: nucleotide 433–1,053, *pvmsp-7L*: nucleotide 382–816) and 3′-end (*pvmsp-7E*: nucleotide 748–1,158, *pvmsp-7F*: nucleotide 1,054–1,449, *pvmsp-7L*: nucleotide 817–1,275). Numbers based on Additional files 4, 9 and 10.

Additional file 4: *pvmsp-7E* gene alignment. The alignment shows the 23 haplotypes found in *pvmsp-7E* together with the *pcmsp-7E* haplotype. Haplotype 1, Sal-I; haplotype 2, Brazil-I; haplotype 3, India-VII; haplotype 4, Mauritania-I; haplotype 5–23, Colombian isolates. Dots represent nucleotide identity. Codons under positive selection are shown in green (intra-species) and turquoise (inter-species) and those under negative selection are shown in yellow (intra-species) and fuchsia (inter-species).

Additional file 5: Haplotype alignment at PvMSP-7E protein level in Colombia. The alignment shows the 13 haplotypes at protein level found in PvMSP-7E. Dots represent amino acid identity.

Additional file 6: Neutrality, linkage disequilibrium and recombination tests at the 5′-end, central region and 3′-end for *pvmsp-7* genes in the Colombian population. 5′-end (*pvmsp-7E*: nucleotide 1–390, *pvmsp-7F*: nucleotide 1–432, *pvmsp-7L*: nucleotide 1–381), central (*pvmsp-7E*: nucleotide 391–747, *pvmsp-7F*: nucleotide 433–1,053, *pvmsp-7L*: nucleotide 382–816) and 3′-end (*pvmsp-7E*: nucleotide 748–1,158, *pvmsp-7F*: nucleotide 1,054–1,449, *pvmsp-7L*: nucleotide 817–1,275). Numbers based on Additional files 4, 9 and 10. •: $p < 0.02$, *: $p < 0.05$.

Additional file 7: Neutrality test sliding window for the *pvmsp-7E* gene. Tajima's D (blue), Fu and Li's D* (red), F* (green) and Fay and Wu's H (purple). The gene was divided into 3 regions: the 5′-end (nucleotide 1 to 390), central (nucleotide 391 to 747) and 3′-end region (nucleotide 748 to 1,158). The bars at the bottom indicate that a test gave significant values in each region. Numbering based on the alignment shown in Additional file 4.

Additional file 8: Intra-species positively and negatively selected sites detected for *pvmsp-7* genes. 5′-end (*pvmsp-7E*: nucleotide 1–390, *pvmsp-7F*: nucleotide 1–432, *pvmsp-7L*: nucleotide 1–381), central (*pvmsp-7E*: nucleotide 391–747, *pvmsp-7F*: nucleotide 433–1,053, *pvmsp-7L*: nucleotide 382–816) and 3′-end (*pvmsp-7E*: nucleotide 748–1,158, *pvmsp-7F*: nucleotide 1,054–1,449, *pvmsp-7L*: nucleotide 817–1,275). Numbers based on Additional files 4, 9 and 10.

Additional file 9 *pvmsp-7F* gene alignment. The alignment shows the 8 haplotypes found in *pvmsp-7F* together with *pcmsp-7F* haplotype. Haplotype 1, Sal-I; haplotype 2, Brazil-I and North Korea; haplotype 3, India-VII; haplotype 4, Mauritania-I; haplotypes 5–8, Colombian isolates. Dots represent nucleotide identity. Codons under positive selection are shown in green (intra-species) and turquoise (inter-species) and those under negative selection are shown in fuchsia (inter-species).

Additional file 10: *pvmsp-7L* gene alignment. The alignment shows the 7 haplotypes found in *pvmsp-7L* together with *pcmsp-7L* haplotype. Haplotype 1, Sal-I; haplotype 2, Brazil-I, India-VII and North Korea; haplotype 3, Mauritania-I; haplotypes 4–7, Colombian isolates. The dots represent nucleotide identity. Codons under positive selection are shown in green (intra-species) and in turquoise (inter-species) and those under negative selection are shown in fuchsia (inter-species).

Additional file 11: Inter-species positively and negatively selected sites detected for *msp-7* genes. 5′-end (*msp-7E*: nucleotide 1–390, *msp-7F*: nucleotide 1–432, *msp-7L*: nucleotide 1–381), central (*msp-7E*: nucleotide 391–747, *msp-7F*: nucleotide 433–1,053, *msp-7L*: nucleotide 382–816) and 3′-end (*msp-7E*: nucleotide 748–1,158, *msp-7F*: nucleotide 1,054–1,449, *msp-7L*: nucleotide 817–1,275). Numbers based on Additional files 4, 9 and 10.

Additional file 12: *pcmsp-7L* putative donor and acceptor sites. An alignment was made between the Sal-I strain *pvmsp-7L* sequences, *pcmsp-7L* and the sequence resulting from GeneScan analysis (*pcmsp-7L*_mRNA). The red arrows indicate the putative donor and acceptor sites in *pcmsp-7L*.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

DG-O and JF-R devised and designed the study, performed the experiments, made the population genetics analysis and wrote the manuscript. MAP coordinated the study, and helped to write the manuscript. All the authors have read and approved the final manuscript.

Acknowledgements

We would like to thank Jason Garry for translating the manuscript. This work was financed by the "Departamento Administrativo de Ciencia, Tecnología e Innovación (COLCIENCIAS)" through contracts RC # 0309–2013 and # 0709–2013. JF-R received financing through COLCIENCIAS cooperation agreement # 0719–13.

Received: 21 October 2014 Accepted: 10 December 2014
Published: 13 December 2014

References

1. Singh J, Purohit B, Desai A, Savardekar L, Shanbag P, Kshirsagar N: **Clinical Manifestations, treatment, and outcome of hospitalized patients with *Plasmodium vivax* malaria in two Indian States: A Retrospective Study.** *Malar Res Treat* 2013, **2013:**341862.
2. Jain V, Agrawal A, Singh N: **Malaria in a tertiary health care facility of Central India with special reference to severe vivax: implications for malaria control.** *Pathog Glob Health* 2013, **107:**299–304.
3. Pacheco MA, Battistuzzi FU, Junge RE, Cornejo OE, Williams CV, Landau I, Rabetafika L, Snounou G, Jones-Engel L, Escalante AA: **Timing the origin of human malarias: the lemur puzzle.** *BMC Evol Biol* 2011, **11:**299.
4. 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, *et al*: **Comparative genomics of the neglected human malaria parasite *Plasmodium vivax*.** *Nature* 2008, **455:**757–763.

5. Iyer J, Gruner AC, Renia L, Snounou G, Preiser PR: **Invasion of host cells by malaria parasites: a tale of two protein families.** *Mol Microbiol* 2007, **65:**231–249.
6. Chitnis CE, Blackman MJ: **Host cell invasion by malaria parasites.** *Parasitol Today* 2000, **16:**411–415.
7. Rodriguez LE, Urquiza M, Ocampo M, Curtidor H, Suarez J, Garcia J, Vera R, Puentes A, Lopez R, Pinto M, Rivera Z, Patarroyo ME: ***Plasmodium vivax* MSP-1 peptides have high specific binding activity to human reticulocytes.** *Vaccine* 2002, **20:**1331–1339.
8. Urquiza M, Rodriguez LE, Suarez JE, Guzman F, Ocampo M, Curtidor H, Segura C, Trujillo E, Patarroyo ME: **Identification of *Plasmodium falciparum* MSP-1 peptides able to bind to human red blood cells.** *Parasite Immunol* 1996, **18:**515–526.
9. Kauth CW, Woehlbier U, Kern M, Mekonnen Z, Lutz R, Mucke N, Langowski J, Bujard H: **Interactions between merozoite surface proteins 1, 6, and 7 of the malaria parasite *Plasmodium falciparum*.** *J Biol Chem* 2006, **281:**31517–31527.
10. Pachebat JA, Ling IT, Grainger M, Trucco C, Howell S, Fernandez-Reyes D, Gunaratne R, Holder AA: **The 22 kDa component of the protein complex on the surface of *Plasmodium falciparum* merozoites is derived from a larger precursor, merozoite surface protein 7.** *Mol Biochem Parasitol* 2001, **117:**83–89.
11. Trucco C, Fernandez-Reyes D, Howell S, Stafford WH, Scott-Finnigan TJ, Grainger M, Ogun SA, Taylor WR, Holder AA: **The merozoite surface protein 6 gene codes for a 36 kDa protein associated with the *Plasmodium falciparum* merozoite surface protein-1 complex.** *Mol Biochem Parasitol* 2001, **112:**91–101.
12. Garzon-Ospina D, Cadavid LF, Patarroyo MA: **Differential expansion of the merozoite surface protein (msp)-7 gene family in Plasmodium species under a birth-and-death model of evolution.** *Mol Phylogenet Evol* 2010, **55:**399–408.
13. Garcia Y, Puentes A, Curtidor H, Cifuentes G, Reyes C, Barreto J, Moreno A, Patarroyo ME: **Identifying merozoite surface protein 4 and merozoite surface protein 7 *Plasmodium falciparum* protein family members specifically binding to human erythrocytes suggests a new malarial parasite-redundant survival mechanism.** *J Med Chem* 2007, **50:**5665–5675.
14. Kadekoppala M, O'Donnell RA, Grainger M, Crabb BS, Holder AA: **Deletion of the *Plasmodium falciparum* merozoite surface protein 7 gene impairs parasite invasion of erythrocytes.** *Eukaryot Cell* 2008, **7:**2123–2132.
15. Tewari R, Ogun SA, Gunaratne RS, Crisanti A, Holder AA: **Disruption of *Plasmodium berghei* merozoite surface protein 7 gene modulates parasite growth in vivo.** *Blood* 2005, **105:**394–396.
16. Mello K, Daly TM, Morrisey J, Vaidya AB, Long CA, Bergman LW: **A multigene family that interacts with the amino terminus of Plasmodium MSP-1 identified using the yeast two-hybrid system.** *Eukaryot Cell* 2002, **1:**915–925.
17. Chen JH, Jung JW, Wang Y, Ha KS, Lu F, Lim CS, Takeo S, Tsuboi T, Han ET: **Immunoproteomics profiling of blood stage *Plasmodium vivax* infection by high-throughput screening assays.** *J Proteome Res* 2010, **9:**6479–6489.
18. Wang L, Crouch L, Richie TL, Nhan DH, Coppel RL: **Naturally acquired antibody responses to the components of the *Plasmodium falciparum* merozoite surface protein 1 complex.** *Parasite Immunol* 2003, **25:**403–412.
19. Woehlbier U, Epp C, Hackett F, Blackman MJ, Bujard H: **Antibodies against multiple merozoite surface antigens of the human malaria parasite *Plasmodium falciparum* inhibit parasite maturation and red blood cell invasion.** *Malar J* 2010, **9:**77.
20. Mello K, Daly TM, Long CA, Burns JM, Bergman LW: **Members of the merozoite surface protein 7 family with similar expression patterns differ in ability to protect against *Plasmodium yoelii* malaria.** *Infect Immun* 2004, **72:**1010–1018.
21. Garzon-Ospina D, Lopez C, Forero-Rodriguez J, Patarroyo MA: **Genetic diversity and selection in three *Plasmodium vivax* merozoite surface protein 7 (Pvmsp-7) genes in a Colombian population.** *PLoS One* 2012, **7:**e45962.
22. Garzon-Ospina D, Romero-Murillo L, Tobon LF, Patarroyo MA: **Low genetic polymorphism of merozoite surface proteins 7 and 10 in Colombian *Plasmodium vivax* isolates.** *Infect Genet Evol* 2011, **11:**528–531.
23. Roy SW, Weedall GD, da Silva RL, Polley SD, Ferreira MU: **Sequence diversity and evolutionary dynamics of the dimorphic antigen merozoite surface protein-6 and other Msp genes of *Plasmodium falciparum*.** *Gene* 2009, **443:**12–21.
24. Tetteh KK, Stewart LB, Ochola LI, Amambua-Ngwa A, Thomas AW, Marsh K, Weedall GD, Conway DJ: **Prospective identification of malaria parasite genes under balancing selection.** *PLoS One* 2009, **4:**e5568.
25. Bozdech Z, Mok S, Hu G, Imwong M, Jaidee A, Russell B, Ginsburg H, Nosten F, Day NP, White NJ, Carlton JM, Preiser PR: **The transcriptome of *Plasmodium vivax* reveals divergence and diversity of transcriptional regulation in malaria parasites.** *Proc Natl Acad Sci U S A* 2008, **105:**16290–16295.
26. Sawai H, Otani H, Arisue N, Palacpac N, de Oliveira ML, Pathirana S, Handunnetti S, Kawai S, Kishino H, Horii T, Tanabe K: **Lineage-specific positive selection at the merozoite surface protein 1 (msp1) locus of *Plasmodium vivax* and related simian malaria parasites.** *BMC Evol Biol* 2010, **10:**52.
27. Tanabe K, Escalante A, Sakihama N, Honda M, Arisue N, Horii T, Culleton R, Hayakawa T, Hashimoto T, Longacre S, Pathirana S, Handunnetti S, Kishino H: **Recent independent evolution of msp1 polymorphism in *Plasmodium vivax* and related simian malaria parasites.** *Mol Biochem Parasitol* 2007, **156:**74–79.
28. Imwong M, Pukrittayakamee S, Gruner AC, Renia L, Letourneur F, Looareesuwan S, White NJ, Snounou G: **Practical PCR genotyping protocols for *Plasmodium vivax* using *Pvcs* and *Pvmsp1*.** *Malar J* 2005, **4:**20.
29. *ORF Finder (Open Reading Frame Finder).* [http://www.ncbi.nlm.nih.gov/projects/gorf/]
30. Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput.** *Nucleic Acids Res* 2004, **32:**1792–1797.
31. Abascal F, Zardoya R, Posada D: **ProtTest: selection of best-fit models of protein evolution.** *Bioinformatics* 2005, **21:**2104–2105.
32. Neafsey DE, Galinsky K, Jiang RH, Young L, Sykes SM, Saif S, Gujja S, Goldberg JM, Young S, Zeng Q, Chapman SB, Dash AP, Anvikar AR, Sutton PL, Birren BW, Escalante AA, Barnwell JW, Carlton JM: **The malaria parasite *Plasmodium vivax* exhibits greater genetic diversity than *Plasmodium falciparum*.** *Nat Genet* 2012, **44:**1046–1050.
33. Suyama M, Torrents D, Bork P: **PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments.** *Nucleic Acids Res* 2006, **34:**W609–612.
34. Jorda J, Kajava AV: **T-REKS: identification of Tandem REpeats in sequences with a K-meanS based algorithm.** *Bioinformatics* 2009, **25:**2632–2638.
35. Librado P, Rozas J: **DnaSP v5: a software for comprehensive analysis of DNA polymorphism data.** *Bioinformatics* 2009, **25:**1451–1452.
36. Depaulis F, Veuille M: **Neutrality tests based on the distribution of haplotypes under an infinite-site model.** *Mol Biol Evol* 1998, **15:**1788–1790.
37. Zhang J, Rosenberg HF, Nei M: **Positive Darwinian selection after gene duplication in primate ribonuclease genes.** *Proc Natl Acad Sci U S A* 1998, **95:**3708–3713.
38. Nei M, Kumar S: *Molecular evolution and phylogenetics.* Oxford New York: Oxford University Press; 2000.
39. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S: **MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods.** *Mol Biol Evol* 2011, **28:**2731–2739.
40. Delport W, Poon AF, Frost SD, Kosakovsky Pond SL: **Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology.** *Bioinformatics* 2010, **26:**2455–2457.
41. Pond SL, Frost SD, Grossman Z, Gravenor MB, Richman DD, Brown AJ: **Adaptation to different human populations by HIV-1 revealed by codon-based analyses.** *PLoS Comput Biol* 2006, **2:**e62.
42. Kosakovsky Pond SL, Frost SD: **Not so different after all: a comparison of methods for detecting amino acid sites under selection.** *Mol Biol Evol* 2005, **22:**1208–1222.
43. Murrell B, Wertheim JO, Moola S, Weighill T, Scheffler K, Kosakovsky Pond SL: **Detecting individual sites subject to episodic diversifying selection.** *PLoS Genet* 2012, **8:**e1002764.
44. Murrell B, Moola S, Mabona A, Weighill T, Sheward D, Kosakovsky Pond SL, Scheffler K: **FUBAR: a fast, unconstrained bayesian approximation for inferring selection.** *Mol Biol Evol* 2013, **30:**1196–1205.
45. Jukes TH, Cantor CR: **Evolution of protein molecules.** In *Mammalian Protein Metabolism.* Edited by Munro HN. New York: Academic Press; 1969.
46. Kelly JK: **A test of neutrality based on interlocus associations.** *Genetics* 1997, **146:**1197–1206.
47. Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SD: **Automated phylogenetic detection of recombination using a genetic algorithm.** *Mol Biol Evol* 2006, **23:**1891–1901.
48. Rozas J, Gullaud M, Blandin G, Aguade M: **DNA variation at the rp49 gene region of *Drosophila simulans*: evolutionary inferences from an unusual haplotype structure.** *Genetics* 2001, **158:**1147–1155.

49. Hudson RR, Kaplan NL: **Statistical properties of the number of recombination events in the history of a sample of DNA sequences.** *Genetics* 1985, **111:**147–164.
50. Martin DP, Lemey P, Lott M, Moulton V, Posada D, Lefeuvre P: **RDP3: a flexible and fast computer program for analyzing recombination.** *Bioinformatics* 2010, **26:**2462–2463.
51. Tachibana S, Sullivan SA, Kawai S, Nakamura S, Kim HR, Goto N, Arisue N, Palacpac NM, Honma H, Yagi M, Tougan T, Katakai Y, Kaneko O, Mita T, Kita K, Yasutomi Y, Sutton PL, Shakhbatyan R, Horii T, Yasunaga T, Barnwell JW, Escalante AA, Carlton JM, Tanabe K: ***Plasmodium cynomolgi* genome sequences provide insight into *Plasmodium vivax* and the monkey malaria clade.** *Nat Genet* 2012, **44:**1051–1055.
52. Mongui A, Perez-Leal O, Soto SC, Cortes J, Patarroyo MA: **Cloning, expression, and characterisation of a *Plasmodium vivax* MSP7 family merozoite surface protein.** *Biochem Biophys Res Commun* 2006, **351:**639–644.
53. Figtree M, Pasay CJ, Slade R, Cheng Q, Cloonan N, Walker J, Saul A: ***Plasmodium vivax* synonymous substitution frequencies, evolution and population structure deduced from diversity in AMA 1 and MSP 1 genes.** *Mol Biochem Parasitol* 2000, **108:**53–66.
54. Mascorro CN, Zhao K, Khuntirat B, Sattabongkot J, Yan G, Escalante AA, Cui L: **Molecular evolution and intragenic recombination of the merozoite surface protein MSP-3alpha from the malaria parasite *Plasmodium vivax* in Thailand.** *Parasitology* 2005, **131:**25–35.
55. Gomez A, Suarez CF, Martinez P, Saravia C, Patarroyo MA: **High polymorphism in *Plasmodium vivax* merozoite surface protein-5 (MSP5).** *Parasitology* 2006, **133:**661–672.
56. Putaporntip C, Udomsangpetch R, Pattanawong U, Cui L, Jongwutiwes S: **Genetic diversity of the *Plasmodium vivax* merozoite surface protein-5 locus from diverse geographic origins.** *Gene* 2010, **456:**24–35.
57. Nurminsky D: *Selective sweep.* Georgetown, Tex. New York, N.Y: Landes Bioscience/Eurekah.com; Kluwer Academic/Plenum Publishers; 2005.
58. Putaporntip C, Jongwutiwes S, Ferreira MU, Kanbara H, Udomsangpetch R, Cui L: **Limited global diversity of the *Plasmodium vivax* merozoite surface protein 4 gene.** *Infect Genet Evol* 2009, **9:**821–826.
59. Martinez P, Suarez CF, Gomez A, Cardenas PP, Guerrero JE, Patarroyo MA: **High level of conservation in Plasmodium vivax merozoite surface protein 4 (PvMSP4).** *Infect Genet Evol* 2005, **5:**354–361.
60. Pacheco MA, Elango AP, Rahman AA, Fisher D, Collins WE, Barnwell JW, Escalante AA: **Evidence of purifying selection on merozoite surface protein 8 (MSP8) and 10 (MSP10) in Plasmodium spp.** *Infect Genet Evol* 2012, **12:**978–986.
61. Forero-Rodriguez J, Garzon-Ospina D, Patarroyo MA: **Low genetic diversity and functional constraint in loci encoding *Plasmodium vivax* P12 and P38 proteins in the Colombian population.** *Malar J* 2014, **13:**58.
62. Forero-Rodriguez J, Garzon-Ospina D, Patarroyo MA: **Low genetic diversity in the locus encoding the *Plasmodium vivax* P41 protein in Colombia's parasite population.** *Malar J* 2014, **13:**388.
63. Chenet SM, Pacheco MA, Bacon DJ, Collins WE, Barnwell JW, Escalante AA: **The evolution and diversity of a low complexity vaccine candidate, merozoite surface protein 9 (MSP-9), in *Plasmodium vivax* and closely related species.** *Infect Genet Evol* 2013, **20:**239–248.
64. Pacheco MA, Ryan EM, Poe AC, Basco L, Udhayakumar V, Collins WE, Escalante AA: **Evidence for negative selection on the gene encoding rhoptry-associated protein 1 (RAP-1) in Plasmodium spp.** *Infect Genet Evol* 2010, **10:**655–661.
65. Carlton JM, Das A, Escalante AA: **Genomics, population genetics and evolutionary history of *Plasmodium vivax*.** *Adv Parasitol* 2013, **81:**203–222.
66. Slatkin M, Wiehe T: **Genetic hitch-hiking in a subdivided population.** *Genet Res* 1998, **71:**155–160.
67. Nurminsky DI: **Genes in sweeping competition.** *Cell Mol Life Sci* 2001, **58:**125–134.
68. Perez-Tris J, Hellgren O, Krizanauskiene A, Waldenstrom J, Secondi J, Bonneaud C, Fjeldsa J, Hasselquist D, Bensch S: **Within-host speciation of malaria parasites.** *PLoS One* 2007, **2:**e235.
69. Sutherland CJ, Tanomsing N, Nolder D, Oguike M, Jennison C, Pukrittayakamee S, Dolecek C, Hien TT, do Rosario VE, Arez AP, Pinto J, Michon P, Escalante AA, Nosten F, Burke M, Lee R, Blaze M, Otto TD, Barnwell JW, Pain A, Williams J, White NJ, Day NP, Snounou G, Lockhart PJ, Chiodini PL, Imwong M, Polley SD: **Two nonrecombining sympatric forms of the human malaria parasite *Plasmodium ovale* occur globally.** *J Infect Dis* 2010, **201:**1544–1550.
70. Mu J, Joy DA, Duan J, Huang Y, Carlton J, Walker J, Barnwell J, Beerli P, Charleston MA, Pybus OG, Su XZ: **Host switch leads to emergence of *Plasmodium vivax* malaria in humans.** *Mol Biol Evol* 2005, **22:**1686–1693.
71. Burge C, Karlin S: **Prediction of complete gene structures in human genomic DNA.** *J Mol Biol* 1997, **268:**78–94.
72. Rice BL, Acosta MM, Pacheco MA, Carlton JM, Barnwell JW, Escalante AA: **The origin and diversification of the merozoite surface protein 3 (msp3) multi-gene family in *Plasmodium vivax* and related parasites.** *Mol Phylogenet Evol* 2014, **78C:**172–184.

doi:10.1186/1475-2875-13-495
Cite this article as: Garzón-Ospina *et al.*: Heterogeneous genetic diversity pattern in *Plasmodium vivax* genes encoding merozoite surface proteins (MSP) -7E, –7F and -7L. *Malaria Journal* 2014 **13**:495.

Buitrago *et al. Malar J (2016) 15:501*
DOI 10.1186/s12936-016-1563-4

Malaria Journal

RESEARCH Open Access

Size polymorphism and low sequence diversity in the locus encoding the *Plasmodium vivax* rhoptry neck protein 4 (PvRON4) in Colombian isolates

Sindy P. Buitrago[1,2], Diego Garzón-Ospina[1,3] and Manuel A. Patarroyo[1,3*]

Abstract

Background: Designing a vaccine against *Plasmodium vivax* has focused on selecting antigens involved in invasion mechanisms that must have domains with low polymorphism for avoiding allele-specific immune responses. The rhoptry neck protein 4 (RON4) forms part of the tight junction, which is essential in the invasion of hepatocytes and/or erythrocytes; however, little is known about this locus' genetic diversity.

Methods: DNA sequences from 73 Colombian clinical isolates from *pvron4* gene were analysed for characterizing their genetic diversity; *pvron4* haplotype number and distribution, as well as the evolutionary forces determining diversity pattern, were assessed by population genetics and molecular evolutionary approaches.

Results: *ron4* has low genetic diversity in *P. vivax* at sequence level; however, a variable amount of tandem repeats at the N-terminal region leads to extensive size polymorphism. This region seems to be exposed to the immune system. The central region has a putative esterase/lipase domain which, like the protein's C-terminal fragment, is highly conserved at intra- and inter-species level. Both regions are under purifying selection.

Conclusions: *pvron4* is the locus having the lowest genetic diversity described to date for *P. vivax*. The repeat regions in the N-terminal region could be associated with immune evasion mechanisms while the central region and the C-terminal region seem to be under functional or structural constraint. Bearing such results in mind, the PvRON4 central and/or C-terminal portions represent promising candidates when designing a subunit-based vaccine as they are aimed at avoiding an allele-specific immune response, which might limit vaccine efficacy.

Keywords: *Plasmodium vivax*, Rhoptry, Genetic diversity, Tandem repeat, *pvron4*, Natural selection, Functional restriction

Background

Malaria is the parasitic disease having the greatest impact on public health [1]. It is caused by different species from the *Plasmodium* genus, these being widely distributed throughout the world's tropical and sub-tropical regions [2]. These parasites cause 140–300 million clinical cases and more than half a million deaths annually [3, 4]. *Plasmodium falciparum* is considered the most lethal species, mainly affecting vulnerable populations in sub-Saharan Africa [4]. Even though efforts were initially concentrated on controlling this species, reports of ever-increasingly severe cases caused by *Plasmodium vivax* [5] and the appearance of drug-resistant strains during the last few years [6, 7] has made this species a growing public health problem, affecting more than a third of the world's population, having high prevalence in Asia and South and Central America [3, 7, 8].

Designing an anti-malarial vaccine against *P. vivax* (as for *P. falciparum*) has been focused on blocking

*Correspondence: mapatarr.fidic@gmail.com
[1] Fundación Instituto de Inmunología de Colombia (FIDIC), Carrera 50 No. 26-20, Bogotá D.C., Colombia
Full list of author information is available at the end of the article

parasite-host interactions during different parasitic stages, especially during the blood phase responsible for the disease's clinical manifestations [9, 10]. A large amount of *P. vivax* antigens have been characterized to date [10, 11], however, their genetic diversity should be assessed for selecting the best antigens for vaccine development [10, 12]. Highly polymorphic antigens can provoke allele-specific immune responses leading to protection having low efficacy after vaccination. On the contrary, those having limited diversity are attractive targets for being evaluated as candidates as they avoid an allele-specific immune response [13].

Most antigens characterized to date have been merozoite proteins [10, 11], including the microneme AMA1 protein and rhoptry neck (RONs) proteins. The interaction between these proteins (specifically AMA1–RON2) has been well described in *Toxoplasma gondii* and *P. falciparum,* these being the structural basis for the tight junction (TJ), a connective ring through which a parasite enters a host cell [14–18].

The RON protein complex (characterized in *P. falciparum*) consists of RON2, RON4 and RON5 proteins [14, 17, 19]. Even though the mechanisms regarding function and interaction between the complex's proteins are not clear, they are considered important targets for blocking invasion. Various studies have highlighted the potential of AMA1 and RON2 as vaccine candidates, however, current knowledge concerning the other RONs is deficient. Co-localization studies and invasion models described for *Plasmodium* spp and *T. gondii* have led to establishing RON4's convincing participation in the TJ [15, 17, 20, 21]. Likewise, its expression in the parasite's invasive forms [21–23], the ability to provoke an immune response in natural malarial infections [23] and the protein's conserved nature, specifically towards the C-terminal (inferred by comparative analysis between *Pf*RON4 and *Tg*RON4 amino acid sequences) [24] suggest that this protein plays an important role for the parasite and could thus be evaluated as vaccine candidate.

The *P. falciparum* RON4 orthologue has recently been characterized in the *P. vivax* VCG-I strain (Vivax Colombia Guaviare-I) [22]. PvRON4 (*P. vivax* RON4) is encoded by a gene having around 2657 bp in this species, expressed during the last hours of the intra-erythrocyte cycle and secreted from the rhoptry neck. This consists of signal peptide sequence, a low complexity domain formed by two types of tandem repeats, a double spiral alpha helix domain and five conserved cysteines towards the C-terminal [22]; the latter region seems to be highly conserved among *P. vivax* and parasite species infecting monkeys [25].

Bearing RON4's potential participation in invasion in mind and given that parasite antigen genetic diversity is an obstacle for designing a completely effective vaccine against *P. vivax*, this study was thus aimed at using Colombian clinical isolates for evaluating *pvron4* locus genetic diversity and the evolutionary mechanisms determining its variation pattern.

Methods

Sample collection

Plasmodium vivax genomic DNA was obtained from 73 clinical isolates collected from 2007 to 2015 (2007: 10, 2008: 12, 2010: 18, and 2015: 33 samples). These came from Colombia's Pacific coast region (Chocó and Nariño departments), Urabá/lower Cauca/southern Córdoba (Córdoba and Antioquia departments) and the Orinoquia-Amazonia region (Amazonas and Guainía departments), representing the three regions having the greatest transmissibility in Colombia [26]. More than 360,000 cases of *P. vivax* infection were recorded between 2007 and 2015, more than 14 % of them regarding Colombia's Pacific coastal region, Urabá/lower Cauca/southern Córdoba 62 % while 7.5 % of *P. vivax* cases were recorded in the Orinoquia-Amazonia region. Malaria symptomatic patients (living in the regions described above) were diagnosed with *P. vivax* infection by microscopy and then invited to donate 5 mL of venous blood. Some Amazonia samples were collected and diagnosed, as has been reported elsewhere [27]. Male and female patients aged 16–64 years were invited to participate in the study. DNA was extracted and stored at −20 °C before being processed and were genotyped by PCR-RFLP from the *pvmsp-3α* gene.

Amplifying, cloning and sequencing the *pvron4* locus

A set of primers was designed to amplify and clone *pvron4* based on Sal-I genomic sequence (GenBank accession number AAKM01000005.1), sequences being as follows: *pvron4* dir 5′ CACAGTGCAACCATGTCTCG 3′ (20 bp) and *pvron4* rev 5′ GCAAGCTAATTTCACAA GTCTTC 3′ (23 bp) primers. Touchdown-PCR was used for amplification using the KAPA-HiFi HotStart Readymix enzyme (Kapa Biosystems) in 25 µL reactions using VCG-I strain genomic DNA as positive control. Thermal conditions were as follows: a 5 min denaturing step at 95 °C, 10 cycles consisting of 20 s at 98 °C, 15 s at 68 °C (temperature was reduced by one degree per cycle) and 1 min at 72 °C, followed by 35 cycles of 20 s at 98 °C, 15 s at 60 °C, 1 min at 72 °C and a final 5-min extension at 72 °C. PCR products were purified using an UltraClean PCR Clean-up purification kit (MOBIO).

The amplicons were ligated in pGEM T-easy cloning vector and then used for transforming *Escherichia coli* JM109 strain cells. Recombinant bacteria were selected by the alpha complementation method and their growth

ability in the presence of ampicillin. These were confirmed by PCR using MangoTaq DNA polymerase and internal primers for the gene (intdir: 5′ TGTGGGTGG CGAGAGTG 3′ (17 bp), and intrev: 5′ ATATTTCC ATTGCTGTACTAGG 3′ (22 bp), designed on Sal-I genomic sequence) using the following thermal conditions a 5 min denaturing step at 95 °C, 35 cycles of 20 s at 98 °C, 15 s at 60 °C, 1 min at 72 °C and a final 5-min extension at 72 °C. The plasmid from the two recombinant colonies per sample was extracted using an UltraClean 6 Minute Mini Plasmid Prep kit (MOBIO) and sent to be sequenced using a BigDye Terminator kit (MACROGEN, Seoul, South Korea), with universal primers SP6 Promoter Primer (Cat.# Q5011), T7 Promoter Primer (Cat.# Q5021) [28] and a set of internal primers (*pvron4dseq*: 5′ CACTAGAAAAGCTAAACATA AACC 3′ (24 bp), and *pvron4rseq*: 5′ ACTCCAATGGT CCTCAAC 3′ (18 bp) designed on the Sal-I genomic sequence) for sequencing.

Statistical analysis of *pvron4* sequences

The electropherograms obtained by sequencing were assembled using CLC DNA workbench v.3 software (CLC bio, Cambridge, MA, USA). The 73 consensus sequences obtained in this study (Additional file 1), 7 reference sequences (from the Salvador-I (Sal-I, GenBank: XM_001615228.1), Mauritania-I (GenBank: AFNI01000694.1), India-VII (GenBank: AFBK01001155.1), Brazil-I (GenBank: AFMK01001544.1/AFMK01001546.1), North Korea (GenBank: AFNJ01001110.1), ctg (GenBank: AAKM01000005) and P01 (GeneDB: PVP01_0916600.1) strains) and 13 orthologous sequences (from *Plasmodium cynomolgi* (GeneBank: BAEJ01000746.1), *Plasmodium knowlesi* (GeneBank: NC_011910.1), *Plasmodium inui* (GeneBank: AMYR01000790.1/AMYR01000791.1), *Plasmodium fragile* (GeneBank: NW_012192637.1), *Plasmodium coatneyi* (GeneBank: CM002860.1), *Plasmodium reichenowi* (GeneBank: LVLA01000012.1), *P. falciparum* (GeneBank: XM_001347803.2), *Plasmodium bergei* (GeneBank: CAAI01002287.1), *Plasmodium yoelii* (GeneBank: AABL01000590.1), *Plasmodium chabaudi* (GeneBank: CAAJ01004050.1), *Plasmodium vinckei* (GeneBank: AMYS01000107.1), *Plasmodium gaboni* (GeneBank: LVLB01000012.1), and *Plasmodium gallinaceum* (Sanger Institute: gal28a.d000001405.Contig1/gal28a.d000000110.Contig1) were used for obtaining the deduced amino acid sequence used for multiple alignment with the MUSCLE algorithm [29]. Such alignment was manually edited to ensure correct repeat region alignment and then used for inferring DNA alignment by Pal2Nal software [30]. The T-REKS algorithm was used for identifying repeat regions [31].

DnaSP v.5 software [32] was used for calculating genetic diversity regarding Colombian sequences and *P. vivax* reference sequences alignment using estimators based on single nucleotide polymorphism (SNP) and sequence length (InDels). Tajima [33], Fu and Li [34], Fu [35], Fay and Wu [36] tests were used for evaluating deviations from the neutral model of molecular evolution, bearing coalescence simulations in mind for obtaining confidence intervals and their statistical significance. The repeat regions or those having gaps were not taken into account for analysis.

The Nei-Gojobori modified method [37] with MEGA v.6 software [38] was used for calculating the average number of synonymous substitutions per synonymous site (d_S) and the average number of non-synonymous substitutions per non-synonymous site (d_N) at intra-species level. The average amount of synonymous divergences per synonymous site (K_S) and the average amount of non-synonymous divergences per non-synonymous site (K_N) were calculated by modified Nei-Gojobori method with Jukes-Cantor correction [39] for determining natural selection signals throughout *Plasmodium* spp evolutionary history (using the *P. vivax* sequences, together with phylogenetically close parasites' orthologous sequences). The differences between intra- and inter-species substitution rates were determined by Fisher's exact test (recommended when the amount of synonymous and/or no-synonymous substitutions is fewer than ten) or the Z-test incorporated in MEGA software v6. Additionally, the McDonald–Kreitman (MK) test [40] with Jukes-Cantor correction was used for evaluating neutrality deviations using the Standard & Generalized McDonald–Kreitman Test web server [41, 42].

A sliding window was used for analysing evolutionary rate ($\omega = d_N/d_S$ and/or K_N/K_S) by evaluating the effect of selection throughout the gene. Individual sites under selection were identified by calculating synonymous and non-synonymous substitution rates per codon using SLAC, FEL, REL, IFEL [43], MEME [44], and FUBAR methods [45] in the Datamonkey online server [46]. Repeat regions or those having gaps were not taken into account for this analysis.

The random effects likelihood (REL)-branch-site method [47] was used for evaluating the existence of lineages under episodic diversifying selection in *Plasmodium* for the *ron4* locus. The MUSCLE algorithm was used for aligning 14 orthologous protein sequences from different species from the genus; this was then used for inferring the best evolutionary model using MEGA software. Phylogeny was then inferred by using the maximum likelihood method with the JFF + G + F model. This is used as reference for analysing lineage-specific positive selection with the REL-branch-site method in the HyPhy package

[48], using a DNA alignment inferred by Pal2Nal from aligning amino acids.

Effective number of codons

ENCprime [49] and DnaSP software were used for estimating the effective number of codons (ENC). This is a measurement of selective pressure at translational level, leading to protein function loss or gain [49]. This test compares the use of each codon versus a null distribution (uniform use of synonymous codons). ENC values close to 61 indicate that all synonymous codons for each amino acid are used equitably, while values close to 0 suggest bias or preferential codon use [50]. Statistical significance could be affected by gene length or recombination [50]. The codon bias index (CBI) was thus used, which takes values ranging from 0 (uniform use of synonymous codons) to 1 (maximum codon bias) [51].

Linkage disequilibrium and recombination

Linkage disequilibrium (LD) was evaluated by calculating the Z_{nS} estimator [52]. A linear regression between this and the nucleotide distance was performed to see whether intragenic recombination occurred in *pvron4*. Recombination was also evaluated by ZZ estimator [53], the minimum number of recombination events (Rm) [54], the GARD algorithm [55] and the RDP v.3 software [56].

pvron4 locus differentiation and population genetic structure

The degree of genetic differentiation (or inter-population heterogeneity) in the *pvron4* locus among Colombian *P. vivax* malaria-endemic regions was estimated by analysis of molecular variance (AMOVA) and Wright's fixating index (F_{ST}), using the Arlequin v.3.1 software [57]. NETWORK v.5 software was used for constructing a median joining network for evaluating possible mutational pathways giving rise to *pvron4* haplotypes, their distribution and frequencies. This method expresses multiple plausible evolutionary pathways as cycles, bound by vectors interpreted as extinct ancestral sequences [58].

Predicting *pvron4* putative domains and antigenic potential

The Blastp algorithm from the NCBI web server was used for identifying putative domains in PvRON4 using the Sal-I sequence as reference. The B-cell epitope prediction tool available at the immune epitope database (IEDB) server was used for evaluating PvRON4s antigenic potential regarding its antigenicity [59], its hydrophobicity [60], protein solvent availability [61] and its potential linear B-cell epitopes [62]. These tests were done with two PvRON4 haplotypes differentiated by the amount of repeats: haplotype 6 (one copy of GEHGEHAEHGE) and haplotype 17 (seven copies of the repeat), to evaluate the effect of repeat number on protein antigenic behavior.

Results

pvron4 locus genetic diversity

Seventy-three sequences from the *pvron4* locus obtained from the *P. vivax* Colombian population (24 from Orinoquia-Amazonia, 21 from the Pacific coast and 28 from Urabá/lower-Cauca/southern Córdoba) were analysed. *pvron4* was initially annotated from the VCG-I strain as being a 2657 bp gene [22], however, the locus analysed from Colombian samples had a variation in length due to two tandem repeats located towards the gene's 5′-end (Fig. 1). These repeats consisted of copies of the GTGG CGAGA nucleotide sequence encoding GES amino acids (repeated one to three times) and a longer sequence CGGAGAGCACGGTGAACACGCTGAACATGGGGA GCA encoding the GEHGEHAEHGE peptide (repeated one to seven times).

Few SNPs were found. Regarding the 2542 sites analysed concerning the Colombian sequences and the reference ones, the number of SNPs varied from five to eight (Table 1). The genetic diversity estimators gave low values ($\theta_w = 4.7 \times 10^{-4}$ and $\pi = 4.1 \times 10^{-4}$) in the *P. vivax* Colombian population. Such values remained constant when comparing the Colombian sequences to the reference ones (Table 1). Likewise, the total amount of *pvron4* haplotypes identified and haplotype diversity were low (Table 1), however, 15 haplotypes in the Colombian population (21 when Colombian and reference sequences are analysed together) and high diversity estimator values (Table 1) were found when analysing insertions/deletions (InDels) in *pvron4*. Bearing the SNPs and InDels between the *pvron4* reference sequence and Colombian sequences in mind, 32 haplotypes were identified (Additional file 2).

Evaluating the effect of selection on the *pvron4* locus

No statistically significant values were found for *pvron4* when using the Tajima, Fu and Li, Fay and Wu and Fu estimators (Table 2). Likewise, the MK test did not reveal any deviations regarding neutrality (Table 3), however, the $d_N - d_S$ difference (Table 3) showed that the synonymous substitution rate was higher than the non-synonymous substitution rate ($p = 0.000$, Fisher's exact test), suggesting that *pvron4* was under purifying selection. The sliding windows led to $\omega < 1$ values being observed throughout the gene (Fig. 1).

When comparing phylogenetically related species, the sliding window gave $\omega < 1$ values, with few regions having $\omega \geq 1$ (positions 1172–1325, 2955–2975 and in position 2955; Fig. 1). The $K_N - K_S$ difference (Table 3) gave negative values suggesting that purifying selection has

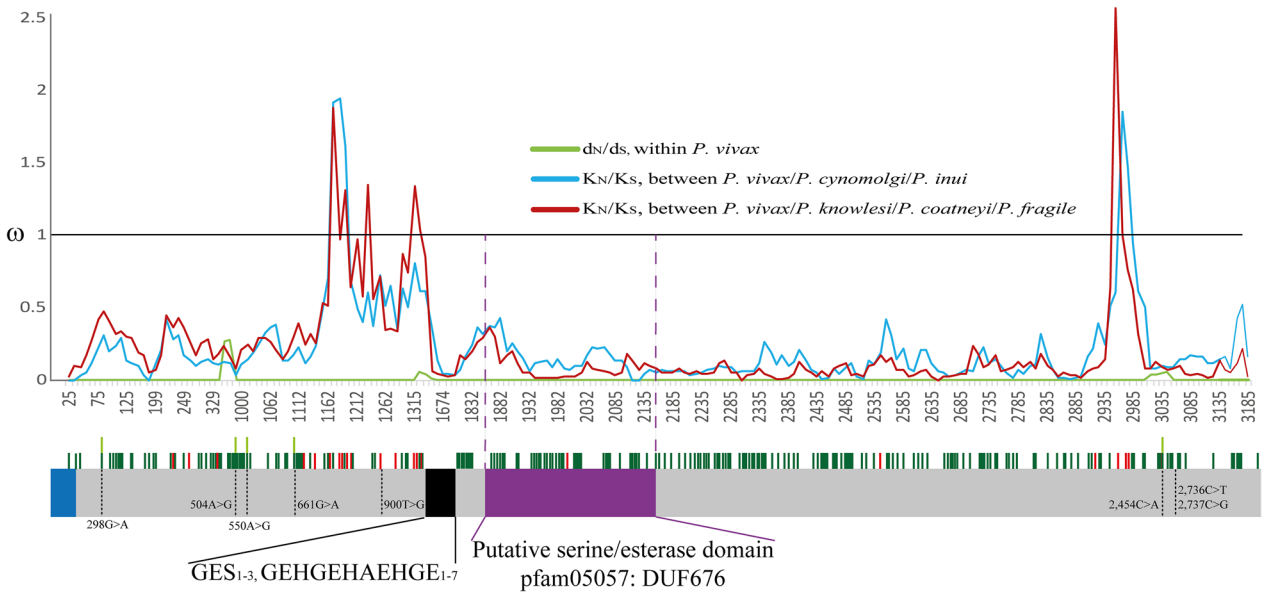


Fig. 1 Representation of *pvron4* encoding DNA sequence and sliding window for ω rate. *ron4* ω values (d_N/d_S) within *P. vivax* are shown in *green* throughout. Likewise, ω divergence rates (K_N/K_S) between the species infecting primates and *P. vivax* are shown in *blue* (for the species evolutionarily closest to *P. vivax*) and *red* (most distant). A representation of the *ron4* gene is given below the sliding window indicating signal peptide (*blue*), the repeat region (*black*) and the putative esterase/lipase domain (*purple*). The *P. vivax* SNP positions are shown as *discontinue lines*, their numbering is based on Additional file 2 alignment. The sites under purifying selection within species are represented by *light green lines* and *dark green* between species while positively selected sites between species are shown by *red lines*

Table 1 *pvron4* genetic diversity estimators calculated from single nucleotide polymorphism and sequence length polymorphism

Single nucleotide polymorphism									Sequence length polymorphism				
n	Sites	Ss	S	Ps	H	Hd	θ_w	π	Sites	No InDels	H	Hd	π
Colombian and reference isolates													
80	2542	8	3	5	11	0.67	7.6×10^{-4}	4.3×10^{-4}	408	21	21	0.82	9.9×10^{-4}
Colombian isolates													
73	2464	5	0	5	8	0.65	4.7×10^{-4}	4.1×10^{-4}	289	15	15	0.78	8.0×10^{-4}

Genetic diversity estimators were calculated using the reference sequences obtained from databases together with Colombian isolates as well as for just Colombian isolates' sequences

n number of isolates analysed, *sites* total of sites analysed excluding gaps, *Ss* number of segregating sites, *S* number of singleton sites, *Ps* number of informative-parsimonious sites, *H* number of haplotypes, *Hd* haplotype diversity, θ_w Watterson estimator, *π* nucleotide diversity per site

played an important role in this locus' evolutionary history in the genus *Plasmodium*. On the other hand, four negatively selected sites (28, 112, 149, 643; Fig. 1) were found throughout the gene in *P. vivax* when calculating selection per codon based on maximum probability methods (SCAL, FEL, IFEL, REL, and FUBAR) and the Bayesian method (MEME), while no sites were found to be under positive selection (Fig. 1). These methods revealed 162 sites under negative selection and 21 under positive selection between species (Fig. 1).

Phylogeny was inferred from orthologous sequences from 14 *Plasmodium* species. This was used as reference framework for the REL-branch-site test. This method led to finding six branches (lineages) having evidence of episodic selection (Fig. 2). Three were ancestral branches (internal) while the other three (external) had given rise to *Plasmodium inui*, *Plasmodium chabaudi* and *Plasmodium gaboni*. The MEME method revealed codons under diversifying episodic selection.

Effective number of codons

Given the high conservation of the *pvron4* locus, deviation regarding the effective use of codons was evaluated as a means of selection at translational level. The ENC

Table 2 Neutrality, linkage disequilibrium and recombination tests for the *pvron4* gene in the *Plasmodium vivax* Colombian population

N	Gene	Tajima	Fu and Li		Fay and Wu's H	Fu´s Fs	Z_{nS}	ZZ	RM
		D	D*	F*					
Colombian and reference isolates									
80	2578	NP	NP	NP	NP	NP	0.154$^{\tau}$	−0.002	0
Colombian isolates									
73	2578	−0.037	−0.035	−0.020	−0.039	−0.028	0.163$^{\tau}$	−0.002	0

No statistically significant values were found in neutrality tests

Z_{nS} average of R^2 for all comparison pairs, *ZZ:* $Z_{nS} - Za$ difference, *Rm* minimum amount of recombination events, *NP* not performed, since not all sequences belonged to the same population

$^{\tau}$ $p < 0.05$

Table 3 Difference between $d_N - d_S$, $K_N - K_S$ and the neutral index from MK test

P. vivax	*P. vivax/Plasmodium* ssp						
	P. knowlesi	*P. inui*	*P. coatneyi*	*P. cynomolgi*	*P. fragile*	*P.cyn/P.inu*	*P.kno/P.coa/P.fra*
$d_N - d_S$	$K_N - K_S$						
−0.002*	−0.011$^{\beta}$	−0.006$^{\beta}$	−0.006$^{\beta}$	−0.009$^{\beta}$	−0.010$^{\beta}$	−0.015$^{\beta}$	−0.025$^{\beta}$
	NI						
	0.696	0.597	0.857	1.133	0.919		

Non-synonymous substitution rate (d_N) and synonymous substitution rate (d_S) within *P. vivax*. Non-synonymous (K_N) and synonymous (K_S) divergence between *P. vivax* and phylogenetically close species. Neutrality index (NI) estimated by McDonald–Kreitman test using Jukes Cantor correction

P.cyn P. cynomolgi, *P.inu P. inui*, *P.kno P. knowlesi*, *P.coa P. coatneyi*, *P.fra P. fragile*

* $p < 0.01$

$^{\beta}$ $p < 0.001$

for *pvron4* estimated by ENCprime (N'c = 53, scaled $X^2 = 0.103$) and DnaSP (ENC = 55.4, scaled $X^2 = 0.159$) gave values close to 61, with a CBI value of 0.162, suggesting that there was no bias regarding the effective use of codons, thereby ruling out translational selection.

Linkage disequilibrium and recombination

The Z_{nS} estimator was used for evaluating LD between *pvron4* polymorphisms, giving 0.15. Linear regression between the LD and nucleotide distance showed a reduction in LD as distance increased, thereby suggesting recombination. However, the ZZ estimator gave −0.0019 and no minimum recombination sites were detected. Likewise, GARD methods found no breakpoints nor did RDP software reveal recombination tracks.

The *Plasmodium vivax* Colombian population's genetic structure regarding the *pvron4* locus

An AMOVA between the three Colombian regions was calculated for evaluating *pvron4* geospatial genetic diversity in Colombia, as well as Wright's fixation index (F_{ST}) between the different populations (regions). AMOVA revealed statistically significant differences between the sub-populations within the regions ($F_{SC} = 0.06$, $p = 0.04$). The Nariño sub-population (from the Pacific region) could be responsible for genetic differences regarding each of the other subpopulations (Additional files 3, 4). Calculating the F_{ST} index between populations (regions) gave values close to 0. There was a statistically significant difference between the Urabá/lower Cauca/southern Córdoba and Orinoquia-Amazonia regions (Table 4).

A median joining network was used for better understanding the evolutionary relationship between *pvron4* haplotypes for describing the set of potential mutational pathways giving rise to the 32 haplotypes available for the locus (Fig. 3 and Additional file 3). The network showed that the parasite's populations shared haplotypes, regardless of geographical region, which were related by different mutational pathways and ancestral sequences (median vectors) (Fig. 3 and Additional file 3). The most frequently occurring haplotypes were H5 (53.1 %), followed by H4 (31.2 %), H13 (25 %), H3 and H20 (12.5 % each). The presence of unique haplotypes in the Nariño (H7, H8 and H10) and Amazonas populations (H21, H22, H2) should be noted as they could be considered rare or specific regional alleles.

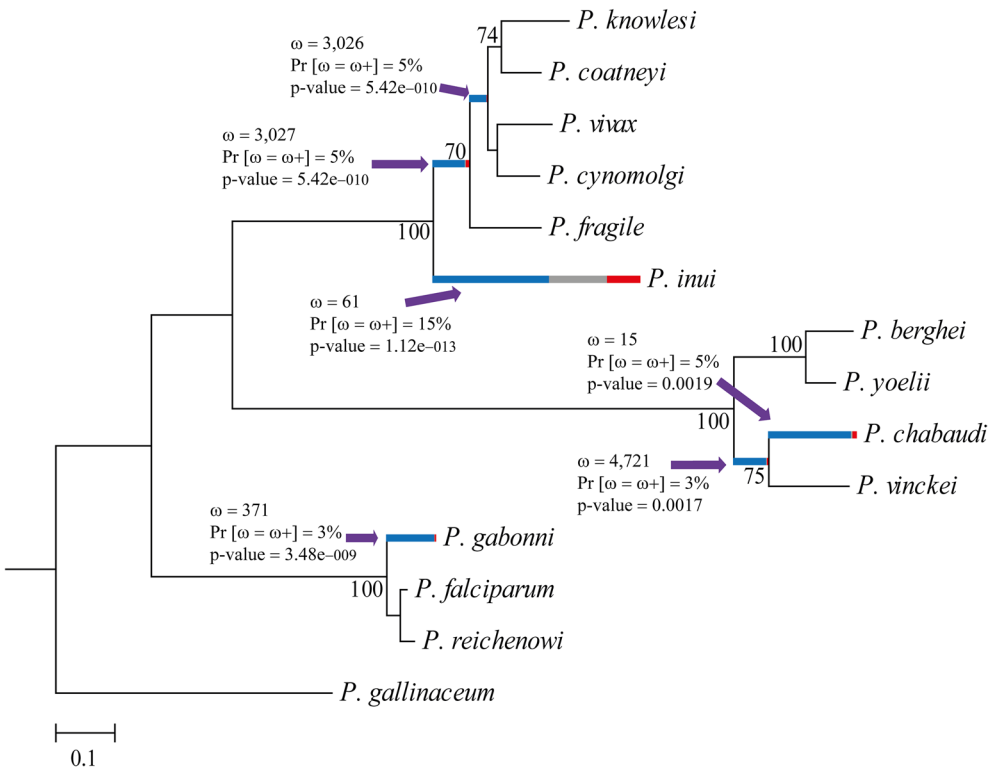


Fig. 2 Lineages subject to episodic diversifying selection in *ron4*. The branches (lineages) under episodic positive selection were identified by the Branch-site REL method. The figure shows the ω^+ values, the percentage of sites (Pr [ω = ω+]) and p values. The shade of each *colour* on branches indicates strength of selection (*red* shows $\omega^+ > 15$, *blue* $\omega^- \leq 1$ and *grey* $\omega = 1$). The size of each *colour* represents the percentage of sites in the corresponding class. Branches have been classified as undergoing episodic diversifying selection by the p value corrected for multiple testing using the Holm-Bonferroni method at $p < 0.05$

Table 4 Inter-population F_{ST} statistic for *pvron4*

F_{ST}	Pacific coast	Urabá/lower Cauca/southern Córdoba	Orinoquia-Amazonia
Pacific coast		0.44043	0.31836
Urabá/lower Cauca/southern Córdoba	−0.00581		0.01465
Orinoquia-Amazonia	0.00542	*0.08128*	

F_{ST} was calculated for parasite populations in three Colombian regions. Values close to 0 indicate low genetic differentiation while values close to 1 indicate high genetic differentiation. Values below the diagonal indicate the F_{ST} value and those above the diagonal represent the p values. Values in italics indicate significant differences having $p < 0.02$

Predicting *pvron4* putative domains and antigenic potential

Analysing the PvRON4 sequence from the Sal-I strain revealed the presence of a putative domain for the esterase/lipase (pfam05057) protein superfamily between amino acids 311–425 (nucleotides 931 to 1275, numbers based on the Sal-I reference sequence, GenBank access number: XM_001615228.1) e-value 1.10e-03. This domain was located after the repeat region and was highly conserved (Fig. 1).

According to antigenicity and B-cell linear epitope prediction, there was a potentially antigenic region between positions 41–171 and 178 up to 256 for haplotype 6 and up to position 340 for haplotype 17 (Additional file 5). This agreed with hydrophobicity and solvent accessibility predictions so that the PvRON4 N-terminal region seems to be a potential immune target. By contrast, the central region (following the repeat region) and the C-terminal seemed to be less antigenic, being less solvent-exposed (Additional file 5).

Discussion

Various proteins contained within the parasite's apical organelles seem to be crucial for host cell invasion and thus represent promising vaccine targets. RON complex proteins are among the proteins localized in the apical organelles, forming part of the TJ [17, 21, 63, 64]. This TJ plays a decisive role in parasite entry to a host cell and closure of the parasitophorous vacuole [17]. The RON4

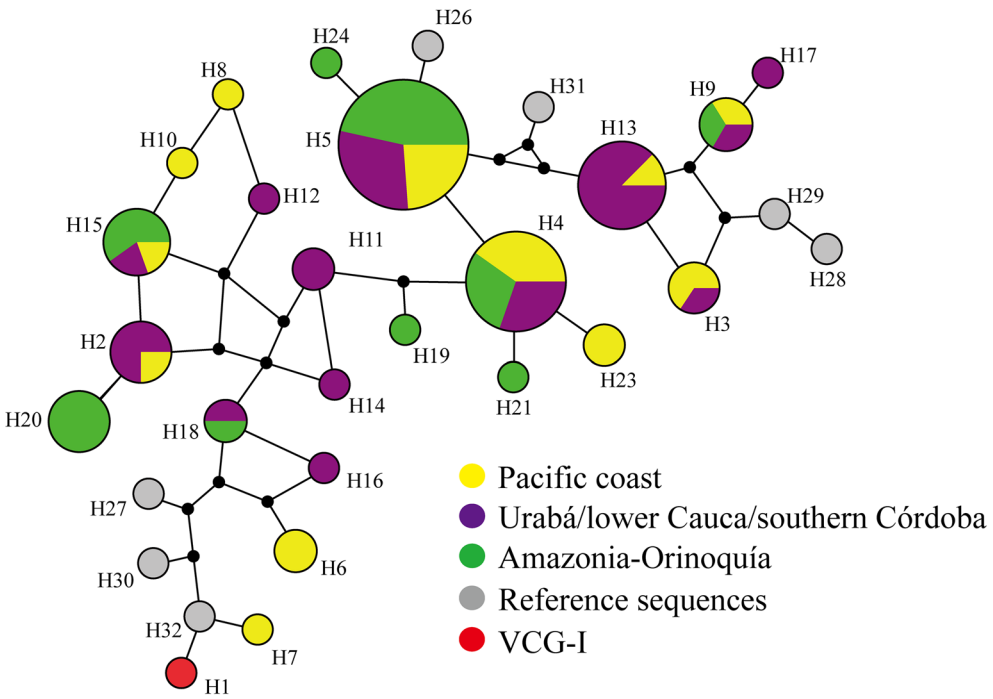


Fig. 3 Median-joining network for Colombian regions. The *Figure* shows the *pvron4* haplotypes identified from the isolates from the three regions of Colombia. Haplotypes 22 and 28 were included within haplotype 15 using the contraction star algorithm [86] for simplifying interpretation of the network. Each node is a haplotype and its size indicates its frequency. The *lines* connecting the haplotypes represent the different mutational paths and the median vectors are the ancestral sequences explaining the relationship and evolutionary origin

protein located in the invasion complex is present in Phylum Apicomplexa members [15, 23, 24], suggesting that it forms part of a conserved invasion pathway. This protein has thus been described as a potential vaccine candidate.

A vaccine candidate must have several characteristics [10, 65]; one of them is to have low genetic diversity to avoid allele-specific immune responses, which could reduce vaccine efficacy. The analysis of *P. falciparum* laboratory strains from different geographically origins showed the *pfron4* locus as being a highly conserved locus, having just one amino acid substitution [23]. *Plasmodium vivax ron4* seems to have the same pattern, as analysing five reference sequences revealed low genetic diversity [25]. This study thus analysed 73 clinical isolates from the Colombian population for confirming *pvron4* as a highly conserved gene in *P. vivax*. In spite of increasing the number of sequences analysed, *pvron4* diversity remained low, the present study's results showing that *pvron4* had lower genetic diversity than in a previous report [25]. Only eight SNPs were identified in the 80 available sequences compared to 14 previously identified ones [25]. Such high number of previously reported SNPs (i.e., 14) was due to erroneous repeat region alignment. *pvron4* had a similar pattern to that of other apical organelle proteins [66, 67]. *pvrap1* and *pvrap2* had 0.0009 to 0.001 π [67] while *pvron4* had a much lower value than *pvrap1*, suggesting it had low genetic diversity, the locus being more conserved to date for *P. vivax*. As it has been suggested for *pvrap1* and *pvrap*2 [67] the low diversity in *pvron4* could be the consequence of functional/structural constraint (see below) due to the key role of this protein in parasite invasion.

Even though *pvron4* was a highly conserved sequence regarding SNP occurrence, it had high polymorphism regarding size. Previous studies have identified two types of repeats towards the N-terminal of the encoded protein [22]. These repeats were reported as being imperfect copies of amino acids GGEH/SGEH/S and G/AEH. However, the analysis here performed showed that the *pvron4* repeat region consisted of two types of repeats having 100 % identity; the first encoded three GES amino acids (one to three copies) and the second one GEHGEHAEHGE amino acids (one to seven copies). These repeats gave a high number of different haplotypes (alleles) in *P. vivax*.

Previous studies have suggested that tandem repeats could play an important role as host immune response evasion mechanism [68–71]. In this study, 21 haplotypes were identified in PvRON4 when the InDels were

analysed. PvRON4 N-terminal region seems to be the most exposed protein according to solvent availability and hydrophobicity results. This region (between signal peptide and repeat region) seems to be a potential antigenic target due to this being where the largest amount of potential B-cell linear epitopes was predicted. The repeat sequences identified broadened the solvent-exposed region and the protein's antigenic potential. The PvRON4 N-terminal region could thus be the region exposed to a host's immune system and repeats could be acting as an immunological smokescreen. Further antigenic and immunogenic studies are needed to confirm such hypothesis. As the repeat region was highly conserved regarding sequence, it could play an important structural or functional role, as has been suggested recently for the CSP [72, 73].

While the N-terminal region might to be exposed to the immune system, the central and C-terminal regions seem to be under functional constraint. Neutrality tests (e.g., Tajima, Fu and Li) gave no statistically significant values and neutrality was thus not ruled out. If *pvron4* is under neutrality then it should show high polymorphism unless there is a functional or structural constraint [74]. Given that this locus was highly conserved regarding sequence, functional/structural constraint is probable. However, selection at translational level could also be responsible for high conservation in the *pvron4* sequence. Analysing regarding preferential codon use did not reveal bias regarding codon use (ENC = 53–55). In fact this value was similar to that reported for the complete genome (ENC = 52.18) [75], suggesting that high *pvron4* locus conservation was not due to selection at translational level and could have been a result of strong purifying selection.

The d_S rate was significantly greater than the d_N rate according to Fisher's exact test, suggesting that this locus has evolved under purifying selection. However, it is not easy to evaluate how natural selection acts in highly conserved antigens [66, 76, 77]. Previous studies have compared *P. vivax* sequences to phylogenetically related species to evaluate the effect of selection on parasite antigens [66, 76–78]. Sliding window analysis of ω rate gave values less than 1 towards the protein's central region as well as towards the C-terminal. The K_S was statistically greater than K_N and various sites under purifying selection between species were detected in these regions, suggesting that purifying selection plays an important role during the locus' evolution in the genus. Bearing in mind that functional regions tend to have slower evolution and are usually conserved between species [79], these results suggest that the PvRON4 C-terminal and central regions could be functionally important. The presence of conserved cysteines in the C-terminal portion (usually associated with protein–protein interaction) could be mediating the interaction between RON4 and AMA-1 and/or other RONs [16, 80], while the presence of a putative esterase/lipase domain in the protein's central region could be involved in RON4 entry to the host cell.

Plasmodium falciparum and *T. gondii* studies have shown that the RON4 C-terminal region seems to play an important role in invasion [16, 24]. RON4 is located inside red blood cells (RBC), anchoring the AMA1/RON2 complex [17, 18, 24]; RON4 must thus be secreted and enter RBC during initial invasion stages by a yet-unknown mechanism. The presence of an esterase/lipase domain in the PvRON4 central region could provide a clue regarding the action mechanism. This is one of the protein's most structured regions, being highly conserved among species and containing several sites under purifying selection, suggesting a functional/structural role. Therefore, while the PvRON4 N-terminal region seems to be associated with evasion of the immune response, the central region (containing the esterase/lipase domain) could be associated with the rupture of ester bonds in the phospholipids constituting host cell membrane. Such rupture would enable RON4 entry to RBC or hepatocyte cytoplasm. Once inside, the RON4 C-terminal region anchors RON proteins, which, in turn, enable AMA1-mediated interaction between the parasite and host cells. It can thus be hypothesized that such putative esterase/lipase domain could play a role regarding RON4 entry to a host cell, however, further functional assays are needed to confirm this.

In spite of *ron4* being highly conserved between species and that purifying selection seems to be important during this locus' evolution in *Plasmodium*, some sites under positive selection were identified, coinciding with the regions where $\omega > 1$ was observed. Previous studies have shown that some antigens (regardless of their genetic diversity) have regions/codons under episodic positive selection, which could have enabled adaptation to different hosts [76, 81, 82]. The topology obtained for *ron4* was similar to that obtained when analysing mitochondrial DNA [83]. The phylogenetic relationships of species infecting rodents and hominids can be seen in *ron4* phylogeny, however, such relationships have not been seen for species infecting monkeys. These species have a complex evolutionary history, which includes biogeographic aspects, adaptation to new macaque hosts and even a change from monkeys to humans [81, 84]. The episodic selection observed here might thus have been a consequence of this group of parasites' rapid diversification (in the N-terminal region and a small portion of the C-terminal region) thereby enabling RON4 to adapt from an ancestral population to new available hosts, as previously suggested [76, 81, 82, 84].

A relatively high number of haplotypes has been found in the *pvron4* locus in Colombia, resulting from a combination of SNPs and tandem repeats. AMOVA analysis and median joining showed that Colombian regions shared most haplotypes and seemed to be genetically similar. However, it was observed that a 6 % of estimated variation between these regions was due to differences between the subpopulations constituting them. The F_{ST} value showed that some subpopulations might not be genetically similar; this could be associated with the presence of unique haplotypes. This agreed with studies in Colombia involving other parasite antigens [77], as well as mitochondrial DNA studies in America [85], suggesting that the parasite population in America is structured and has limited gene flow. However, since some subpopulations analysed here had limited sample size, the number of sequences must be increased for such results to be confirmed.

Conclusions

Designing a vaccine which is completely effective against the parasites causing malaria requires antigens having limited genetic diversity to avoid allele-specific immune responses. The *pvron4* locus was seen to have low genetic diversity regarding SNPs but had a large amount of haplotypes due to tandem repeats located in the proteins' N-terminal, which could be involved in evading the immune response. On the other hand, the central and C-terminal regions are highly conserved, even between species. Such regions are under purifying selection, suggesting that they are under functionally or structurally constraint. The central region has a putative esterase/lipase domain, leading to the hypothesis that this domain enables RON4 entry to host cells while the C-terminal region anchors the AMA1/RON complex. Bearing the aforementioned results in mind, PvRON4 central/C-terminal region would seem to be a promising candidate for inclusion when designing a subunit-based vaccine against *P. vivax*.

Additional files

Additional file 1. The sequences of the 73 Colombian isolates obtained in this study, 7 reference sequences and 13 orthologous sequences from the *ron4* locus.

Additional file 2. Aligning the 32 haplotypes identified for *pvron4*.

Additional file 3. Median–joining network for Colombian departments. The Figure shows the evaluative relationship between the 32 *pvron4* haplotypes identified from isolates from five Colombian departments, together with the reference sequence. Haplotypes 22 and 28 were included within haplotype 15 when using the star contraction algorithm [86] for simplifying interpretation of the network. Each node is a haplotype and its size indicates its frequency. The lines connecting the haplotypes represent mutational paths and the median vectors are the ancestral sequences explaining the evolutionary relationship and origin.

Additional file 4. Inter-population F_{ST} statistic for *pvron4* per department. F_{ST} was calculated for parasite subpopulations in Colombia. Values close to 0 indicate low genetic differentiation while values close to 1 indicate high genetic differentiation. Values below the diagonal indicate the F_{ST} value and those above the diagonal represent the p-values. Values in bold indicate significant differences having $p < 0.03$.

Additional file 5. Predicting potentially antigenic regions in PvRON4. Predictive tests for A. Linear B epitopes (threshold=0.35); B. Kolaskar and Tongaonkar antigenicity (threshold=1.00); C. Parker hydrophobicity (threshold=2.31) and D. Accessibility to solvent (threshold=1,000) using haplotype 6 and 17 sequences. The dotted line shows the regions having the greatest probability of being recognised by the immune system according to score on each prediction. The inverted antigenicity values arose from calculating antigenic propensity, regardless of the possible occurrence of amino acids in epitopes and on protein surface.

Abbreviations

AMA1: apical membrane antigen 1; AMOVA: analysis of molecular variance; BepiPred: B cell epitope prediction; CBI: codon bias index; CSP: circumsporozoite protein; d_N: non-synonymous substitutions per non-synonym site; d_S: synonyms substitutions per synonym site; ENC: effective number of codons; FEL: fixed effects likelihood; F_{SC}: F-statistic, used to estimate the proportion of genetic variability found among populations within groups; F_{ST}: Wright's fixation index, used to estimate the proportion of genetic variability found among populations; FUBAR: fast, unconstrained Bayesian approximation for inferring selection; GARD: genetic algorithm recombination detection; IEDB: immune epitope database and analysis resource; IFEL: internal fixed effects likelihood; InDels: insertions/deletions; K_N: average number of non-synonyms divergences per non-synonym site; K_S: average number of synonyms divergences per synonym site; LD: linkage disequilibrium; MEME: mixed effects model of evolution; MK: McDonald–Kretiman test; N'c: ENC prime; PCR–RFLP: polymerase chain reaction-restriction fragment length polymorphism; *pfron4*: gene encoding *Plasmodium falciparum* rhoptry neck protein 4; *pvmsp-3α*: alpha subunit of gene encoding *Plasmodium vivax* merozoite surface protein-3; *pvrap1*: gene encoding *Plasmodium vivax* rhoptry-associated protein 1; *pvrap2*: gene encoding *Plasmodium vivax* rhoptry-associated protein 2; *pvron4*: gene encoding *Plasmodium vivax* rhoptry neck protein 4; RDP3: recombination detection program 3; REL: random effects likelihood; Rm: minimum number of recombination events; RON2: *Plasmodium vivax* rhoptry neck protein 2; RONs: Rhoptry neck proteins; SLAC: single likelihood ancestor counting; SNP: single nucleotide polymorphisms; *Tg*RON4: *Toxoplasma gondii* rhoptry neck protein 4; TJ: tight junction; VCG-I: Vivax Colombia Guainía-I.

Authors' contributions

SPB performed the experiments as well as the population genetics and molecular evolutionary analysis and wrote the manuscript. DG-O devised and designed the study participated in the experiments as well as in the population genetics and molecular evolutionary analysis and writing the manuscript. MAP coordinated the study and helped to write the manuscript. All the authors read and approved the final manuscript.

Author details

[1] Fundación Instituto de Inmunología de Colombia (FIDIC), Carrera 50 No. 26-20, Bogotá D.C., Colombia. [2] Microbiology Postgraduate Program, Universidad Nacional de Colombia, Bogotá D.C., Colombia. [3] School of Medicine and Health Sciences, Universidad del Rosario, Bogotá D.C., Colombia.

Acknowledgements

We would like to thank Jason Garry for translating and reviewing the manuscript. We would also like to thank Johana Barreto Badillo by her technical assistance and Johanna Forero-Rodríguez and Carlos Fernando Suárez for their valuable comments and suggestions. MAP would like to especially thank Liliana Andrea Córdoba for all her love and support during the last couple of years.

Competing interests

The authors declare that they have no competing interests.

Availability of data and material
The sequences obtained here and which showed different haplotypes to already-reported ones were stored in the GenBank database, accession numbers KX513800- KX513824. These sequences, together with the others analysed in this study are available in Additional file 1.

Ethics approval and consent to participate
The *P. vivax*-infected patients from whom parasite genomic DNA was obtained, gave their informed consent after having been notified about the research's purpose. All procedures were approved by the *Fundación Instituto de Inmunología de Colombia* and the *Universidad del Rosario's* ethics committees.

Funding
This work was financed by the *Departamento Administrativo de Ciencia, Tecnología e Innovación* (COLCIENCIAS) through grant RC # 0309-2013. SPB received financing through COLCIENCIAS cooperation agreement # 0555-2015.

Received: 24 August 2016 Accepted: 7 October 2016
Published online: 18 October 2016

References

1. Naghavi M, Wang H, Lozano R, Davis A, Liang X, Zhou M. GBD 2013 mortality and causes of death collaborators. Global, regional, and national age-sex specific all-cause and cause-specific mortality for 240 causes of death, 1990–2013: a systematic analysis for the Global Burden of Disease Study 2013. Lancet. 2015;385:117–71.
2. Hay SI, Guerra CA, Tatem AJ, Noor AM, Snow RW. The global distribution and population at risk of malaria: past, present, and future. Lancet Infect Dis. 2004;4:327–36.
3. WHO. World malaria report 2015. Geneva: World Health Organization; 2015. http://www.who.int/malaria/publications/world-malaria-report-2015/wmr2015-without-profiles.pdf?ua=1.
4. UNICEF. Achieving the malaria MDG target. Reversing the incidence of malaria 2000–2015 http://www.unicef.org/publications/files/Achieving_the_Malaria_MDG_Target.pdf.
5. Price RN, Douglas NM, Anstey NM. New developments in *Plasmodium vivax* malaria: severe disease and the rise of chloroquine resistance. Curr Opin Infect Dis. 2009;22:430–5.
6. Tjitra E, Anstey NM, Sugiarto P, Warikar N, Kenangalem E, Karyana M, et al. Multidrug-resistant *Plasmodium vivax* associated with severe and fatal malaria: a prospective study in Papua Indonesia. PLoS Med. 2008;5:e128.
7. Winter DJ, Pacheco MA, Vallejo AF, Schwartz RS, Arevalo-Herrera M, Herrera S, et al. Whole genome sequencing of field isolates reveals extensive genetic diversity in *Plasmodium vivax* from Colombia. PLoS Negl Trop Dis. 2015;9:e0004252.
8. Guerra CA, Howes RE, Patil AP, Gething PW, Van Boeckel TP, Temperley WH, et al. The international limits and population at risk of *Plasmodium vivax* transmission in 2009. PLoS Negl Trop Dis. 2010;4:e774.
9. Birkett AJ, Moorthy VS, Loucq C, Chitnis CE, Kaslow DC. Malaria vaccine R&D in the decade of vaccines: breakthroughs, challenges and opportunities. Vaccine. 2013;31(Suppl 2):B233–43.
10. Barry AE, Arnott A. Strategies for designing and monitoring malaria vaccines targeting diverse antigens. Front Immunol. 2014;5:359.
11. Patarroyo MA, Calderon D, Moreno-Perez DA. Vaccines against *Plasmodium vivax*: a research challenge. Expert Rev Vaccines. 2012;11:1249–60.
12. Arnott A, Barry AE, Reeder JC. Understanding the population genetics of *Plasmodium vivax* is essential for malaria control and elimination. Malar J. 2012;11:14.
13. Takala SL, Plowe CV. Genetic diversity and malaria vaccine design, testing and efficacy: preventing and overcoming 'vaccine resistant malaria'. Parasite Immunol. 2009;31:560–73.
14. Harvey KL, Gilson PR, Crabb BS. A model for the progression of receptor-ligand interactions during erythrocyte invasion by *Plasmodium falciparum*. Int J Parasitol. 2012;42:567–73.
15. Lebrun M, Michelin A, El Hajj H, Poncet J, Bradley PJ, Vial H, Dubremetz JF. The rhoptry neck protein RON4 re-localizes at the moving junction during *Toxoplasma gondii* invasion. Cell Microbiol. 2005;7:1823–33.
16. Takemae H, Sugi T, Kobayashi K, Gong H, Ishiwa A, Recuenco FC, et al. Characterization of the interaction between *Toxoplasma gondii* rhoptry neck protein 4 and host cellular beta-tubulin. Sci Rep. 2013;3:3199.
17. Weiss GE, Gilson PR, Taechalertpaisarn T, Tham WH, de Jong NW, Harvey KL, et al. Revealing the sequence and resulting cellular morphology of receptor-ligand interactions during *Plasmodium falciparum* invasion of erythrocytes. PLoS Pathog. 2015;11:e1004670.
18. Cao J, Kaneko O, Thongkukiatkul A, Tachibana M, Otsuki H, Gao Q, et al. Rhoptry neck protein RON2 forms a complex with microneme protein AMA1 in *Plasmodium falciparum* merozoites. Parasitol Int. 2009;58:29–35.
19. Paul AS, Egan ES, Duraisingh MT. Host-parasite interactions that guide red blood cell invasion by malaria parasites. Curr Opin Hematol. 2015;22:220–6.
20. Giovannini D, Spath S, Lacroix C, Perazzi A, Bargieri D, Lagal V, et al. Independent roles of apical membrane antigen 1 and rhoptry neck proteins during host cell invasion by apicomplexa. Cell Host Microbe. 2011;10:591–602.
21. Boucher LE, Bosch J. The apicomplexan glideosome and adhesins-structures and function. J Struct Biol. 2015;190:93–114.
22. Arevalo-Pinzon G, Curtidor H, Abril J, Patarroyo MA. Annotation and characterization of the *Plasmodium vivax* rhoptry neck protein 4 (PvRON4). Malar J. 2013;12:356.
23. Morahan BJ, Sallmann GB, Huestis R, Dubljevic V, Waller KL. *Plasmodium falciparum*: genetic and immunogenic characterisation of the rhoptry neck protein PfRON4. Exp Parasitol. 2009;122:280–8.
24. Alexander DL, Arastu-Kapur S, Dubremetz JF, Boothroyd JC. *Plasmodium falciparum* AMA1 binds a rhoptry neck protein homologous to TgRON4, a component of the moving junction in *Toxoplasma gondii*. Eukaryot Cell. 2006;5:1169–73.
25. Garzon-Ospina D, Forero-Rodriguez J, Patarroyo MA. Inferring natural selection signals in *Plasmodium vivax*-encoded proteins having a potential role in merozoite invasion. Infect Genet Evol. 2015;33:182–8.
26. Gestión para la vigilancia entomológica y control de la transmisión de malaria. Guía de Vigilancia Entomológica y Control de Malaria. http://www.ins.gov.co/temas-de-interes/Documentacin%20Malaria/03%20Vigilancia%20entomo%20malaria%20.pdf.
27. Camargo-Ayala PA, Cubides JR, Nino CH, Camargo M, Rodriguez-Celis CA, Quinones T, et al. High *Plasmodium malariae* prevalence in an endemic area of the Colombian Amazon region. PLoS ONE. 2016;11:e0159968.
28. pGEM®-T and pGEM®-T Easy Vector Systems, Instructions for Use of Products. https://www.promega.com/-/media/files/resources/protocols/technical-manuals/0/pgem-t-and-pgem-t-easy-vector-systems-protocol.pdf.
29. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32:1792–7.
30. Suyama M, Torrents D, Bork P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. 2006;34:W609–12.
31. Jorda J, Kajava AV. T-REKS: identification of Tandem REpeats in sequences with a K-meanS based algorithm. Bioinformatics. 2009;25:2632–8.
32. Librado P, Rozas J. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics. 2009;25:1451–2.
33. Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics. 1989;123:585–95.
34. Fu YX, Li WH. Statistical tests of neutrality of mutations. Genetics. 1993;133:693–709.
35. Fu YX. Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. Genetics. 1997;147:915–25.
36. Fay JC, Wu CI. Hitchhiking under positive Darwinian selection. Genetics. 2000;155:1405–13.
37. Zhang J, Rosenberg HF, Nei M. Positive Darwinian selection after gene duplication in primate ribonuclease genes. Proc Natl Acad Sci USA. 1998;95:3708–13.
38. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. MEGA6: molecular evolutionary genetics analysis version 6.0. Mol Biol Evol. 2013;30:2725–9.
39. Jukes TH, Cantor CR. Evolution of protein molecules. In: Munro HN, editor. Mammalian protein metabolism. New York: Academic Press; 1969.
40. McDonald JH, Kreitman M. Adaptive protein evolution at the Adh locus in *Drosophila*. Nature. 1991;351:652–4.
41. Standard & generalized McDonald–Kreitman Test. http://mkt.uab.es/mkt/MKT.asp.

42. Egea R, Casillas S, Barbadilla A. Standard and generalized McDonald–Kreitman test: a website to detect selection by comparing different classes of DNA sites. Nucleic Acids Res. 2008;36:W157–62.
43. Kosakovsky Pond SL, Frost SD. Not so different after all: a comparison of methods for detecting amino acid sites under selection. Mol Biol Evol. 2005;22:1208–22.
44. Murrell B, Wertheim JO, Moola S, Weighill T, Scheffler K, Kosakovsky Pond SL. Detecting individual sites subject to episodic diversifying selection. PLoS Genet. 2012;8:e1002764.
45. Murrell B, Moola S, Mabona A, Weighill T, Sheward D, Kosakovsky Pond SL, et al. FUBAR: a fast, unconstrained bayesian approximation for inferring selection. Mol Biol Evol. 2013;30:1196–205.
46. Delport W, Poon AF, Frost SD, Kosakovsky Pond SL. Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology. Bioinformatics. 2010;26:2455–7.
47. Kosakovsky Pond SL, Murrell B, Fourment M, Frost SD, Delport W, Scheffler K. A random effects branch-site model for detecting episodic diversifying selection. Mol Biol Evol. 2011;28:3033–43.
48. Pond SL, Frost SD, Muse SV. HyPhy: hypothesis testing using phylogenies. Bioinformatics. 2005;21:676–9.
49. Novembre JA. Accounting for background nucleotide composition when measuring codon usage bias. Mol Biol Evol. 2002;19:1390–4.
50. Shields DC, Sharp PM, Higgins DG, Wright F. "Silent" sites in *Drosophila* genes are not neutral: evidence of selection among synonymous codons. Mol Biol Evol. 1988;5:704–16.
51. Morton BR. Chloroplast DNA codon use: evidence for selection at the psb A locus based on tRNA availability. J Mol Evol. 1993;37:273–80.
52. Kelly JK. A test of neutrality based on interlocus associations. Genetics. 1997;146:1197–206.
53. Rozas J, Gullaud M, Blandin G, Aguade M. DNA variation at the rp49 gene region of *Drosophila simulans*: evolutionary inferences from an unusual haplotype structure. Genetics. 2001;158:1147–55.
54. Hudson RR, Kaplan NL. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics. 1985;111:147–64.
55. Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SD. Automated phylogenetic detection of recombination using a genetic algorithm. Mol Biol Evol. 2006;23:1891–901.
56. Martin D, Rybicki E. RDP: detection of recombination amongst aligned sequences. Bioinformatics. 2000;16:562–3.
57. Excoffier L, Laval G, Schneider S. Arlequin (version 3.0): an integrated software package for population genetics data analysis. Evol Bioinform Online. 2005;1:47–50.
58. Bandelt HJ, Forster P, Rohl A. Median-joining networks for inferring intraspecific phylogenies. Mol Biol Evol. 1999;16:37–48.
59. Kolaskar AS, Tongaonkar PC. A semi-empirical method for prediction of antigenic determinants on protein antigens. FEBS Lett. 1990;276:172–4.
60. Parker JM, Guo D, Hodges RS. New hydrophilicity scale derived from high-performance liquid chromatography peptide retention data: correlation of predicted surface residues with antigenicity and X-ray-derived accessible sites. Biochemistry. 1986;25:5425–32.
61. Emini EA, Hughes JV, Perlow DS, Boger J. Induction of hepatitis A virus-neutralizing antibody by a virus-specific synthetic peptide. J Virol. 1985;55:836–9.
62. Larsen JE, Lund O, Nielsen M. Improved method for predicting linear B-cell epitopes. Immunome Res. 2006;2:2.
63. Tonkin ML, Roques M, Lamarque MH, Pugniere M, Douguet D, Crawford J, et al. Host cell invasion by apicomplexan parasites: insights from the co-structure of AMA1 with a RON2 peptide. Science. 2011;333:463–7.
64. Cowman AF, Crabb BS. Invasion of red blood cells by malaria parasites. Cell. 2006;124:755–66.
65. Richie TL, Saul A. Progress and challenges for malaria vaccines. Nature. 2002;415:694–701.
66. Pacheco MA, Ryan EM, Poe AC, Basco L, Udhayakumar V, Collins WE, et al. Evidence for negative selection on the gene encoding rhoptry-associated protein 1 (RAP-1) in *Plasmodium* spp. Infect Genet Evol. 2010;10:655–61.
67. Garzon-Ospina D, Romero-Murillo L, Patarroyo MA. Limited genetic polymorphism of the *Plasmodium vivax* low molecular weight rhoptry protein complex in the Colombian population. Infect Genet Evol. 2010;10:261–7.
68. Verra F, Hughes AL. Biased amino acid composition in repeat regions of *Plasmodium* antigens. Mol Biol Evol. 1999;16:627–33.
69. Hisaeda H, Yasutomo K, Himeno K. Malaria: immune evasion by parasites. Int J Biochem Cell Biol. 2005;37:700–6.
70. Ferreira MU, da Silva Nunes M, Wunderlich G. Antigenic diversity and immune evasion by malaria parasites. Clin Diagn Lab Immunol. 2004;11:987–95.
71. Ramasamy R. Molecular basis for evasion of host immunity and pathogenesis in malaria. Biochim Biophys Acta. 1998;1406:10–27.
72. Ferguson DJ, Balaban AE, Patzewitz EM, Wall RJ, Hopp CS, Poulin B, et al. The repeat region of the circumsporozoite protein is critical for sporozoite formation and maturation in *Plasmodium*. PLoS ONE. 2014;9:e113923.
73. Aldrich C, Magini A, Emiliani C, Dottorini T, Bistoni F, Crisanti A, et al. Roles of the amino terminal region and repeat region of the *Plasmodium berghei* circumsporozoite protein in parasite infectivity. PLoS ONE. 2012;7:e32524.
74. Kimura M. The neutral theory of molecular evolution. Cambridge: Cambridge University Press; 1983.
75. Cornejo OE, Fisher D, Escalante AA. Genome-wide patterns of genetic polymorphism and signatures of selection in *Plasmodium vivax*. Genome Biol Evol. 2015;7:106–19.
76. Forero-Rodriguez J, Garzon-Ospina D, Patarroyo MA. Low genetic diversity in the locus encoding the *Plasmodium vivax* P41 protein in Colombia's parasite population. Malar J. 2014;13:388.
77. Forero-Rodriguez J, Garzon-Ospina D, Patarroyo MA. Low genetic diversity and functional constraint in loci encoding *Plasmodium vivax* P12 and P38 proteins in the Colombian population. Malar J. 2014;13:58.
78. Pacheco MA, Elango AP, Rahman AA, Fisher D, Collins WE, Barnwell JW, et al. Evidence of purifying selection on merozoite surface protein 8 (MSP8) and 10 (MSP10) in *Plasmodium* spp. Infect Genet Evol. 2012;12:978–86.
79. Graur D, Zheng Y, Price N, Azevedo RB, Zufall RA, Elhaik E. On the immortality of television sets: "function" in the human genome according to the evolution-free gospel of ENCODE. Genome Biol Evol. 2013;5:578–90.
80. Narum DL, Nguyen V, Zhang Y, Glen J, Shimp RL, Lambert L, et al. Identification and characterization of the *Plasmodium yoelii* PyP140/RON4 protein, an orthologue of *Toxoplasma gondii* RON4, whose cysteine-rich domain does not protect against lethal parasite challenge infection. Infect Immun. 2008;76:4876–82.
81. Muehlenbein MP, Pacheco MA, Taylor JE, Prall SP, Ambu L, Nathan S, et al. Accelerated diversification of nonhuman primate malarias in Southeast Asia: adaptive radiation or geographic speciation? Mol Biol Evol. 2015;32:422–39.
82. Sawai H, Otani H, Arisue N, Palacpac N, de Oliveira Martins L, Pathirana S, et al. Lineage-specific positive selection at the merozoite surface protein 1 (msp1) locus of *Plasmodium vivax* and related simian malaria parasites. BMC Evol Biol. 2010;10:52.
83. Pacheco MA, Cranfield M, Cameron K, Escalante AA. Malarial parasite diversity in chimpanzees: the value of comparative approaches to ascertain the evolution of *Plasmodium falciparum* antigens. Malar J. 2013;12:328.
84. Mu J, Joy DA, Duan J, Huang Y, Carlton J, Walker J, et al. Host switch leads to emergence of *Plasmodium vivax* malaria in humans. Mol Biol Evol. 2005;22:1686–93.
85. Taylor JE, Pacheco MA, Bacon DJ, Beg MA, Machado RL, Fairhurst RM, et al. The evolutionary history of *Plasmodium vivax* as inferred from mitochondrial genomes: parasite genetic diversity in the Americas. Mol Biol Evol. 2013;30:2050–64.
86. Forster P, Torroni A, Renfrew C, Rohl A. Phylogenetic star contraction applied to Asian and Papuan mtDNA evolution. Mol Biol Evol. 2001;18:1864–81.

Baquero et al. Parasites & Vectors (2017) 10:251
DOI 10.1186/s13071-017-2183-8

Parasites & Vectors

RESEARCH Open Access

PvGAMA reticulocyte binding activity: predicting conserved functional regions by natural selection analysis

Luis A. Baquero[1†], Darwin A. Moreno-Pérez[1,2†], Diego Garzón-Ospina[1,2], Johanna Forero-Rodríguez[1], Heidy D. Ortiz-Suárez[1] and Manuel A. Patarroyo[1,3*]

Abstract

Background: Adhesin proteins are used by *Plasmodium* parasites to bind and invade target cells. Hence, characterising molecules that participate in reticulocyte interaction is key to understanding the molecular basis of *Plasmodium vivax* invasion. This study focused on predicting functionally restricted regions of the *P. vivax* GPI-anchored micronemal antigen (*Pv*GAMA) and characterising their reticulocyte binding activity.

Results: The *pvgama* gene was initially found in *P. vivax* VCG-I strain schizonts. According to the genetic diversity analysis, *Pv*GAMA displayed a size polymorphism very common for antigenic *P. vivax* proteins. Two regions along the antigen sequence were highly conserved among species, having a negative natural selection signal. Interestingly, these regions revealed a functional role regarding preferential target cell adhesion.

Conclusions: To our knowledge, this study describes *Pv*GAMA reticulocyte binding properties for the first time. Conserved functional regions were predicted according to natural selection analysis and their binding ability was confirmed. These findings support the notion that *Pv*GAMA may have an important role in *P. vivax* merozoite adhesion to its target cells.

Keywords: Adhesin protein, *Plasmodium vivax*, Genetic diversity, Conserved functional region, Reticulocyte binding activity

Background

Plasmodium vivax is a human malaria-causing parasite whose eradication is a priority on the international health agenda [1]. As a strategy for eradicating this species, several research groups have focused their efforts on developing a vaccine, as vaccination has been successful at controlling and eradicating other infectious diseases [2].

It has been suggested that vaccines should consist of key proteins or their fragments used by infectious agents to bind to the target cells [3, 4]. Hence, knowledge of proteins expressed by the parasite at the end of its intra-erythrocyte life-cycle, especially those interacting with red blood cells (RBC), should prove most suitable as candidate vaccine components.

Current efforts to develop an anti-malarial vaccine have mainly focused on *P. falciparum*, given the availability of robust in vitro culturing techniques for this parasite (currently unavailable for *P. vivax*) which has led to a large-scale identification of genes [5], transcripts [6] and proteins [7]. This information has led to an improved understanding of the molecules involved in *P. falciparum* merozoite invasion of erythrocytes. For example, several adhesin molecules have been described in the apical organelles (rhoptries and micronemes), that facilitate interaction with cell receptors and promote parasite internalisation within the target cell [8]. Several of these proteins are immunogenic and are being evaluated as vaccine candidates in clinical studies [9]. The GPI-anchored micronemal antigen (GAMA) represents one apical protein that has an adhesive role in *Plasmodium* and *Toxoplasma*. *Plasmodium falciparum* GAMA (*Pf*GAMA) binds to human erythrocytes, an interaction

* Correspondence: mapatarr.fidic@gmail.com
†Equal contributors
[1]Molecular Biology and Immunology Department, Fundación Instituto de Inmunología de Colombia (FIDIC), Carrera 50 No. 26-20, Bogotá DC, Colombia
[3]Basic Sciences Department, School of Medicine and Health Sciences, Universidad del Rosario, Carrera 24 No. 63C-69, Bogotá DC, Colombia
Full list of author information is available at the end of the article

mediated by its binding region which is located in the amino terminal sequence, and is involved in the sialic acid-independent invasion pathway [10]. On the other hand, GAMA knockouts of *T. gondii* (*Tg*GAMA) show a reduction in the ability of tachyzoites to attach to the host cell during invasion as well as a delay in the time to death in an in vivo model, suggesting a function during parasite adhesion and invasion [11].

Unfortunately, basic *P. vivax* research has been delayed mainly due to the parasite's preference for invading reticulocytes which are difficult to obtain in the high percentages needed for propagating *P. vivax* in vitro [12, 13]. However, it has been possible to characterise several molecules forming part of the parasite's selective human reticulocyte invasion route, such as reticulocyte binding proteins (RBPs) [14, 15], merozoite surface protein 1 (MSP-1) [16], some proteins from the tryptophan-rich antigen (TRAg) family [17] and the recently described rhoptry neck protein 5 (RON5) [18]. Some of these contain specific binding regions that have been identified using several strategies, such as mapping using peptides labelled with radioactive iodine, ELISA, flow cytometry or rosetting assays. However, these methodologies are laborious when large molecules must be analysed. Furthermore, sometimes it is not known whether these regions are polymorphic between isolates, which would be counterproductive for the development of a broadly protective vaccine.

A new strategy has recently been proposed for identifying selection signals and that enables the determination of conserved antigens or those having potential functional regions [19]. Cornejo et al. [20] and Garzón-Ospina et al. [19] identified natural selection signals in *P. vivax* genes when analysing the sequences of five genomes from different locations [21]. These results were supported by earlier studies, increasing the number of sequences analysed [22–24]. This type of analysis could therefore provide a viable approach for selecting conserved antigens that are subject to functional restrictions. However, no experimental evidence has been produced to support such approach.

Given the importance of conserved functional region prediction and the role of adhesin proteins during host-parasite interaction, and considering the interesting features displayed by GAMA in other apicomplexa, the present study aimed at characterising *P. vivax* VCG-I strain GAMA functional regions by selection signal prediction and then determine the role of such regions in binding to reticulocytes.

Methods

An approach to GAMA genetic diversity and evolutionary forces

Evolutionary methods compare the non-synonymous mutations rate (d_N, mutations altering protein sequences) to the synonymous mutations rate (d_S, those encoding the same amino acid) in the search for natural selection signals. Deleterious mutations are usually removed from populations by negative natural selection ($d_N < d_S$ or $\omega < 1$). Regions displaying this kind of selection might have functional/structural importance, maintaining high sequence conservation between species [25]. On the other hand, mutations having an adaptive advantage (or a beneficial role) are fixed in a population by positive natural selection ($dN > dS$ or $\omega > 1$). Taking the above into account, functional regions could be predicted by evolutionary approaches [19]. *pvgama* gene DNA sequences from 6 *P. vivax* strains (VCG-I, Sal-I, Brazil-I, India-VII, Mauritania-I and North Korea [21]) and 5 phylogenetically-related species (*P. cynomolgi*, *P. inui*, *P. fragile*, *P. knowlesi* and *P. coatneyi*) [26] were obtained by tblastn (except for VCG-I) from the whole-genome shotgun contigs (wgs) NCBI database for assessing genetic diversity and evolutionary forces regarding GAMA. The MUSCLE algorithm [27] was used to align the sequences and the alignment was manually corrected. Nucleotide diversity per site (π) was estimated from the *P. vivax* sequences and the modified Nei-Gojobori method [28] was used to assess natural selection signals by calculating the difference between synonymous and non-synonymous substitution rates (d_N-d_S). Natural selection was also assessed by estimating the difference between synonymous and non-synonymous divergence rates (K_N-K_S) using sequences from *P. vivax* and related species through the modified Nei-Gojobori method and Jukes-Cantor correction [29]. Specific codons under natural selection amongst species were identified using codon-based Bayesian or maximum likelihood approaches (SLAC, FEL, REL [30], MEME [31] and FUBAR [32]), following recombination by the GARD method [33]. Codon-based methods estimate the evolutionary rate (ω) at each codon using a statistical test to determine whether ω is significantly different to 1 (neutral evolution). The Branch-site REL algorithm [34] was used to identify lineages under episodic positive selection (selection occasionally having transient periods of adaptive evolution masked by negative selection or neutral evolution). The Datamonkey web server was used to perform these analyses [35].

Primer design, cloning and sequencing

The *Plasmodium vivax gama* (*pvgama*) gene sequence was taken from the PlasmoDB database [36] and scanned for PCR priming sites (Table 1) using Generunner software (version 3.05). Primers were designed to amplify either the entire *pvgama* gene or several smaller-sized fragments according to the natural selection analysis (Fig. 1). The gDNA (extracted using a Wizard Genomic purification kit; Promega, Madison, USA) and cDNA (synthesised with SuperScript III enzyme (RT+) (Invitrogen,

Table 1 Primer designed for *pvgama* gene amplification

Target	Primer sequence (5′ – 3′)[a]	MT (°C)	Product size (bp)	aa position
pvgama	Fwd: ATGAAGTGCAACGCCTCC Rev: AAAAATGAATAGGAGCAACG	58	2313	1 to 771
pvgama -Nt	Fwd: ATACGGAATGGAAACAACC Rev: AGTCGGTTCGTTATTCTCG		1284	22 to 449
pvgama -Ct	Fwd: CTGCTCAAGAACACGAAC Rev: GCTTCCACTCTGCAATTC		948	434 to 749
pvgama -CR1	Fwd: GACGATCATCTGTGTTCAAAAA Rev: GACCTCATTTTTGGACTTCTC	60	666	87 to 308
pvgama -VR1	Fwd: GGCGCCTTCCTGCAGTC Rev: CATTAACATGGTGTTGTCGCT		438	330 to 475
pvgama -CR2	Fwd: CAGGCGGCCATCTTACTAA Rev: GCTCCCGTTGACGCCCTT		321	482 to 588
pvgama -VR2	Fwd: GCCGCAAACGCAGACGCC Rev: GTTTGCCGAGAAGCTTCCAC		384	626 to 753

Abbreviations: *Nt and Ct* amino and carboxyl terminal; *CR* conserved region, *VR* variable region; *Fwd* forward, *Rev* reverse, *MT* melting temperature, *bp* base pair, *aa* amino acid

[a]Protein's expression start codon was included in forward primer's 5′ end

Carlsbad, USA) samples from *P. vivax* VCG-I strain schizont-stage enriched parasites (propagated and obtained as previously described [37, 38]) were used as template in 25 μl PCR reactions containing 1× KAPA HiFi HotStart ReadyMix (KAPA Biosystems, Woburn, MA, USA), 0.3 μM primers and DNAse-free water. Temperature cycling for PCR involved a denaturing step of 95 °C for 5 min, followed by 35 cycles of 98 °C for 20 s, Tm °C (Table 1) for 15 s and 72 °C for 30 s or 1 min and 30 s depending on product size. A Wizard PCR preps kit (Promega) was used for purifying amplicons obtained from PCR with the RT+ and gDNA samples, once quality had been evaluated on agarose gel. Purified products were ligated to the pEXP5 CT/TOPO expression vector or pGEM (Promega) (for the gene obtained from gDNA) and transformed in TOP10 *E. coli* cells (Invitrogen). Several clones obtained from independent PCR reactions were grown for purifying the plasmid using an UltraClean mini plasmid prep purification kit (MO BIO Laboratories, California, USA). Insert integrity and correct orientation were then confirmed by sequencing, using an ABI-3730 XL sequencer (MACROGEN, Seoul, South Korea). ClustalW (NPS@) software was used for comparing gene sequences from Sal-I reference strain and the primate-adapted VCG-I strain [39]. The *pvgama* gene sequence from *P. vivax* VCG-I strain was deposited in NCBI under accession number KT248546.

Recombinant protein expression

The pEXP-*pvgama* recombinant plasmids were transformed in *E. coli* BL21-DE3 (Invitrogen), according to the manufacturer's recommendations. Cells were grown

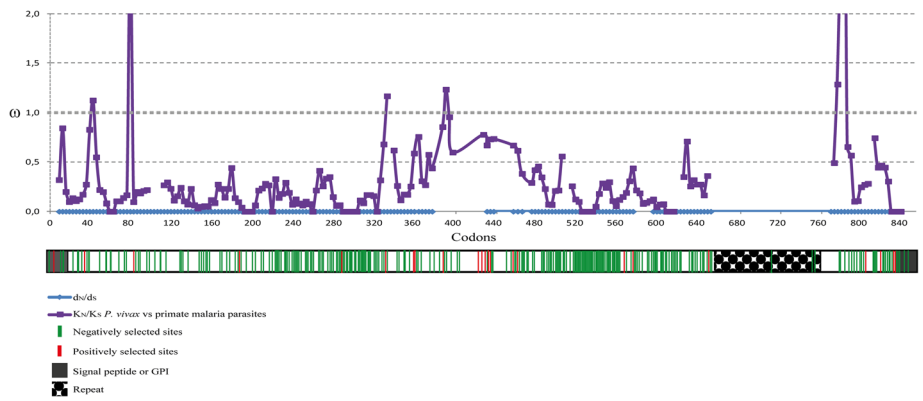


Fig. 1 Evolutionary rate (ω) sliding window. Intra-species ω values (d_N/d_S) are represented in *blue* whilst inter-species ω values (K_N/K_S between *P. vivax* and malarial parasites infecting primates) are shown in *purple*. A ω value equal 1 means neutral evolution, ω < 1 negative selection whilst ω > 1 means positive selection. A diagram of the gene can be observed *below* the sliding window. Negatively selected inter-species codons are shown in *green* whilst positively selected sites are shown in *red*. Numbering is based on the alignment in Additional file 1: Figure S1

overnight at 37 °C in 50 ml Luria Bertani (LB) medium containing 100 μg/ml ampicillin using a Lab-line Incubator Shaker. The initial inoculum was then seeded in 1 l of LB with ampicillin (100 μg/ml) and left to grow at 37 °C with shaking at ~300× *rpm* until reaching 0.5 OD_{600}. The culture was incubated on ice for 30 min and then IPTG 1 mM was used to induce expression by incubation for 16 h at room temperature (RT) with shaking at ~200× *rpm*. The culture was then spun at 2400× *g* for 20 min and the pellet was collected for extraction of the recombinant protein.

Denaturing extraction

The cell pellet obtained from *E. coli* expressing *Pv*GAMA-Nt and *Pv*GAMA-Ct fragments was homogenised in denaturing extraction buffer (DEB) (6 M urea, 10 mM Tris, 100 mM NaH_2PO_4 and 20 mM imidazole) containing the SIGMAFAST protease inhibitor cocktail (Sigma-Aldrich, St. Louis, USA) and then lysed by incubating with 0.1 mg/ml lysozyme overnight at 4 °C at 10× *rpm* using a tube rotator (Fisher Scientific, Waltham, USA). The supernatant was collected by spinning at 16,000× *g* for 1 h.

Native extraction

*Pv*GAMA-CR1, *Pv*GAMA-VR1, *Pv*GAMA-CR2 and *Pv*GAMA-VR2 were extracted using a method for obtaining the molecules in native conditions with the respective positive and negative controls (region II and III/IV from the Duffy binding protein, DBP) (unpublished data). Briefly, the pellet was frozen/thawed for 3 cycles and then homogenised in native extraction buffer (NEB) (50 mM Tris, 300 mM NaCl, 25 mM imidazole, 0.1 mM EGTA and 0.25% Tween-20, pH 8.0). The mixture was incubated for 1 h at 4 °C at 10× *rpm* and the supernatant was collected by spinning at 16,000× *g* for 1 h.

Protein purification

Total lysate supernatant was incubated with Ni^{+2}-NTA resin (Qiagen, Valencia, CA, USA) for purifying the proteins by solid-phase affinity chromatography, once protein expression had been verified by western blot. Briefly, the resin was pre-equilibrated with the respective buffer used for extracting proteins and then incubated with the *E. coli* lysate overnight at 4 °C. The protein-resin mixture was placed on a column and then weakly bound proteins were eluted by washing with 20 ml buffer containing 0.1% Triton X-114 followed by 50 ml of the same buffer without detergent. The proteins extracted in denaturing conditions were dialysed on the column by passing 20 ml DEB with urea in descending concentrations (6 M, 3 M, 1.5 M, 0.75 M and PBS). Bound proteins were then eluted with PBS containing imidazole at increasing concentrations (50 mM to 500 mM) in 3 ml fractions; those having a single band (confirmed on 12% SDS-PAGE by Coomassie blue staining and by western blot using anti-polyhistidine antibodies) were pooled and dialysed extensively in PBS, pH 7.2. A micro BCA protein assay kit (Thermo Scientific, Rockford, USA) was used for quantifying each protein, using the bovine serum albumin (BSA) curve as reference.

Peptide synthesis

One 6 histidine peptide was synthesised according to a previously-established methodology [40], polymerised, lyophilised and characterised by RP-HPLC and MALDI-TOF MS. The peptide was homogenised in PBS and then stored at -20 °C until use.

Blood sample collection and processing

Individuals with a clinical history of *P. vivax* (37 subjects) or *P. falciparum* (30 subjects) malaria, aged 18 to 50 year-old and living in malaria-endemic areas of Colombia (Chocó, Nariño, Córdoba, Vichada and Guaviare) were selected for this study. Sera from healthy individuals (16 adult subjects) who had never been affected by the disease and who were living in non-endemic areas were used as negative controls. The blood samples were collected in BD Vacutainer tubes without anticoagulant by personnel from the Fundación Instituto de Inmunología de Colombia (FIDIC) from October 2006 to March 2011 (for *P. vivax*) and June to October 1993 (for *P. falciparum*) and stored at 4 °C until transport. Samples were then transported to Bogotá for processing. Total blood was spun at 5000× *g* for 5 min and the serum was then recovered and stored at -80 °C in FIDIC serum bank (to date).

Enzyme-linked immunosorbent assay (ELISA)

*Pv*GAMA antigenicity was evaluated in triplicate using serum from patients who had suffered episodes of *P. vivax* or *P. falciparum* infection. Briefly, 96-well polysorb plates were covered with 1 μg r*Pv*GAMA-Nt, or r*Pv*GAMA-Ct, overnight at 4 °C and then incubated at 37 °C for 1 h. The dishes were blocked with 200 μl 5% skimmed milk - PBS-0.05% Tween for 1 h at 37 °C. Antibody reactivity against the recombinant protein was evaluated by incubating the plates with 1:100 dilution of each human serum in 5% skimmed milk - PBS-0.05% Tween for 1 h at 37 °C. The dishes were incubated with peroxidase-coupled goat anti-human IgG monoclonal secondary antibody (1:10,000) (Catalogue 1222H, ICN) diluted in 5% skimmed milk - PBS-0.05% Tween for 1 h at 37 °C and then a peroxidase substrate solution (KPL Laboratories, Gaithersburg, MD, USA) was added to reveal the reaction, according to the manufacturer's recommendations. Optical density (OD) at 620 nm (detected by MJ ELISA Multiskan Reader) was

calculated by subtracting the OD value obtained from the control well value (no antigen). The cut-off value for evaluating the positivity threshold was determined by taking the average of the OD plus twice the standard deviation (± 2SD) of healthy individuals' sera reactivity.

Cord blood sample processing

The newborn umbilical cord blood samples used in this research were collected by personnel from the Hemocentro Distrital (Bogotá) and then processed by SEPAX Cell Processing System (Biosafe, Eysins, Switzerland) to reduce nucleated cells, according to the manufacturer's recommendations. The samples were stored at 4 °C and Duffy antigen receptor for chemokines (DARC) presence was determined by agglutination assay using antibodies directed against the molecule's Fya or Fyb fraction. The percentage of nucleated cells was scored in 20 fields at 100× magnification using Wright's stain before carrying out the binding assay.

Cell binding assay

Reticulocyte binding was tested in triplicate by flow cytometry and using the total cells from cord blood sample (Fya^- Fyb^+ phenotype). Briefly, 5 μl samples were incubated with 25 μg of each recombinant protein (*Pv*GAMA-CR1, *Pv*GAMA-VR1, *Pv*GAMA-CR2 and *Pv*GAMA-VR2) for 16 h at 4 °C at 4× *rpm*. Twenty-five μg of DBP region II and III/IV were used as positive and negative controls, respectively. The 6 histidine peptide was also used as control once the recombinant proteins contained a 6-histidine tag. A binding inhibition assay was also performed by incubating *Pv*GAMA conserved recombinant proteins (CR1 and CR2) with a mixture of human sera (1:10 dilution) for 1 h at 4 °C before putting them in contact with cells. The samples were then incubated with mouse anti-His-PE monoclonal antibody (1:40 dilution) (MACSmolecular-Miltenyi Biotec, San Diego, CA, USA) for 30 min in the dark after washing with 1% BSA-PBS solution (v/v). White cells and reticulocytes were stained by incubating with anti-CD45 APC clone 2D1 (1:80 dilution) (Becton Dickinson, Franklin Lakes, NJ, USA) and anti-CD71 APC-H7 clone M-A712 (1:80 dilution) (Becton Dickinson) monoclonal antibodies for 20 min at RT. Subsequently, reticulocyte (CD71 + CD45-PE+) and mature erythrocyte (CD71-CD45-PE+) binding was quantified by analysing 1 million events using a FACSCanto II cytometer (BD, San Diego, CA, USA) and Flowjo V10 software. PE signal intensity in the reticulocyte population was evaluated regarding CD71 signal to determine CD71 low ($CD71^{lo}$) and high ($CD71^{hi}$) cells.

Statistical analysis

Mean values and standard deviations (SD) were calculated from the measurements of three independent experiments. Statistical significance was assessed by comparing means using a 0.05 significance level for testing a stated hypothesis. Student's *t*-test and analysis of variance (ANOVA) were used for comparing the means of each experimental group to those for control. Tukey's multiple comparison test was used for multiple comparison of experimental group means to those for control. GradhPad Software (San Diego, CA) was used for all statistical analysis.

Results

*Pv*GAMA genetic diversity and selection signals

Pvgama sequences were obtained from genomes of 5 different strains from different geographical regions (North Korea, Brazil, Mauritania and India). These were aligned with the VCG-I strain sequence and orthologous sequences from 5 phylogenetically-related species. The alignment revealed a size polymorphism in *pvgama* due to the [C/T]C[G/C]C[A/T]AA[C/T][C/G][A/G/C][G/A]AC[G/C/A] repeat which was not present in *P. cynomolgi*, *P. inui*, *P. fragile*, *P. knowlesi* or *P. coatneyi* (Additional file 1: Figure S1). Regarding *P. vivax*, 5 segregating sites and $\pi = 0.0008$ were observed.

No significant values were found when evaluating synonymous and non-synonymous substitution rates (d_N-d_S = -0.001 (0.001), $P > 0.1$). However, synonymous divergence was greater than non-synonymous divergence ($P < 0.0001$) when comparing *pvgama* sequences to each related species: K_N-K_S *P. vivax/P. cynomolgi* = -0.041 (0.006); K_N-K_S *P. vivax/P. inui* = -0.062 (0.008); K_N-K_S *P. vivax/P. fragile* = -0.030 (0.006); K_N-K_S *P. vivax/P. knowlesi* = -0.072 (0.009); K_N-K_S *P. vivax/P. coatneyi* = -0.049 (0.007). The evolutionary rate ω (dN/dS and KN/KS) sliding window showed that two highly conserved regions amongst species (codons 80–320 and 514–624) might be under negative selection ($\omega < 0.5$). Furthermore, 308 negatively-selected codons were observed amongst species (Fig. 1); a lot of them were in the conserved regions. The Branch-site REL algorithm identified episodic positive selection signals in the lineages giving rise to *P. knowlesi* and *P. coatneyi* as well as the lineage formed by *P. cynomolgi* and *P. fragile* (Additional file 2: Figure S2). 22 sites showed evidence of positive selection amongst species (Fig. 1).

Antigenic response was directed against the GAMA carboxyl fragment

Based on the polymorphism analysis results, it was hypothesised that the carboxyl region was more antigenic than the amino one by the presence of the repetitive region. Hence, r*Pv*GAMA-Nt and r*Pv*GAMA-Ct antigenicity (obtained recombinantly; Additional file 3: Figure S3a, b) was evaluated using sera from 37 patients suffering of *P. vivax* malaria and sera from people who had never suffered the disease. r*Pv*GAMA-Nt reacted

positively with 64.8% of the sera in screening (0.26 cut-off point) whilst 67.5% of them recognised r*Pv*GAMA-Ct (0.47 cut-off point). These data agreed with a study of the profile of the humoral immune response for *P. vivax* in which r*Pv*GAMA was recognised by 54.5% of the sera used in the array [41]. The statistical test for the assay with r*Pv*GAMA-Nt gave a significant difference between the means (m) of the groups (ANOVA: $F_{(1,41)} = 4.73$, $P = 0.035$; $m = 0.38$ for the group of infected patients and $m = 0.12$ for the control group). Likewise, there was a significant difference between the means of the groups (ANOVA: $F_{(1,41)} = 14.75$, $P = 0.0001$; $m = 0.67$ for the group of infected patients and $m = 0.14$ for the control group) when r*Pv*GAMA-Ct was detected by human sera (Fig. 2a). There was also a statistically significant difference when analysing the means of recognition for r*Pv*GAMA-Nt and r*Pv*GAMA-Ct (ANOVA: $F_{(1,72)} = 16.01$, $P = 0.0002$). Taking into account that the response was higher against *Pv*GAMA-Ct, it was decided to confirm whether the antibodies generated during *P. falciparum* natural infection were able to detect this fragment. No significant difference (ANOVA: $F_{(1,38)} = 0.036$, $P = 0.850$) was seen for *Pv*GAMA-Ct recognition by these sera (Fig. 2b). The significant reactivity against the recombinants by *P. vivax*-infected individuals' sera indicated that the protein could trigger an antigenic response during natural infection, this being higher and species-specific against the *Pv*GAMA carboxyl region.

*Pv*GAMA bound to human reticulocytes

Red blood cell samples having the $Fya^{-}Fyb^{+}$ phenotype (Duffy +) taken from umbilical cord blood were incubated with conserved (CR1 and CR2) and variable (VR1 and VR2) regions extracted and purified in their soluble form (Additional file 3: Figure S3c), predicted by natural selection analysis and then evaluated by flow cytometry to quantify the protein-cell interaction. The percentage of each recombinant binding to erythrocytes was calculated using the gating strategy described in Additional file 4: Figure S4, which enabled selecting the mature (CD71-CD45-) or immature (CD71 + CD45-) cell population to which a target protein was bound (labelled with anti-His PE antibody). All recombinant proteins had a curve shift when the PE signal was compared to control (cells not incubated with recombinant proteins) in the histogram (Fig. 3). Interestingly, the GAMA fragments bound to reticulocytes to a much higher percentage compared to mature erythrocytes (CR1: *t*-test: $t_{(4)} = 24.9$, $P = 0.0001$; VR1: *t*-test: $t_{(4)} = 9.02$, $P = 0.001$; CR2: *t*-test: $t_{(4)} = 12.4$, $P = 0.0001$; VR2: *t*-test: $t_{(4)} = 24.8$, $P = 0.0001$) (Fig. 4a). The conserved regions showed highest interaction with the reticulocytes compared to negative binding controls (ANOVA-Tukey: $F_{(6,\ 12)} = 72.64$, $P < 0.0001$). CR2 recombinant protein bound to 10.11% (SD = 1.33) of target cells, which was very similar to the positive control ($m \pm SD = 11.8 \pm 1.15$) ($P > 0.189$), whilst CR1 were able to bind to 6.36% (SD = 0.30) of the cells (Fig. 4a). Regarding *Pv*GAMA variable regions, VR1 was able to bind to 3.08% (SD = 0.54) of the reticulocytes whilst VR2 bound 5.64% (SD = 0.37). CR1, CR2 and VR2 fragments had the highest interaction with $CD71^{hi}$ reticulocytes when binding percentages were analysed as a function of CD71 APC-H7 signal (CR1: *t*-test: $t_{(4)} = 7.32$, $P = 0.002$; CR2: *t*-test: $t_{(4)} = 16.04$, $P = 0.0001$; VR2: *t*-test: $t_{(4)} = 3.71$, $P = 0.021$), unlike VR1 and DBP-RII (VR1: *t*-test: $t_{(4)} = 1.52$, $P = 0.202$; DBP-RII: *t*-test: $t_{(4)} = 0.19$, $P = 0.853$) (as previously found [42])

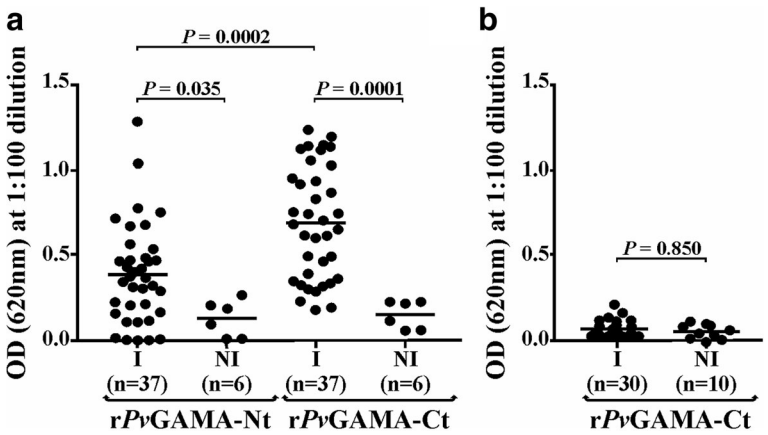


Fig. 2 *Pv*GAMA antigenicity during natural malaria infection. The *dot plot* shows OD distribution (Y-axis) for detecting r*Pv*GAMA-Nt or r*Pv*GAMA-Ct by *P. vivax* (**a**) or r*Pv*GAMA-CT by *P. falciparum* (**b**) infected (I) and non-infected (NI) patients' sera (X-axis). r*Pv*GAMA-Nt: infected individuals $n = 37$, $m \pm SD = 0.38 \pm 0.29$; control individuals $n = 6$, $m \pm SD = 0.12 \pm 0.1$. r*Pv*GAMA-Ct: infected individuals $n = 37$, $m \pm SD = 0.67 \pm 0.32$; control individuals $n = 6$, $m \pm SD = 0.14 \pm 0.08$. r*Pv*GAMA-Ct recognised by *P. falciparum* infected patients' sera: infected individuals $n = 30$, $m \pm SD = 0.06 \pm 0.04$; control individuals $n = 10$, $m \pm SD = 0.06 \pm 0.03$

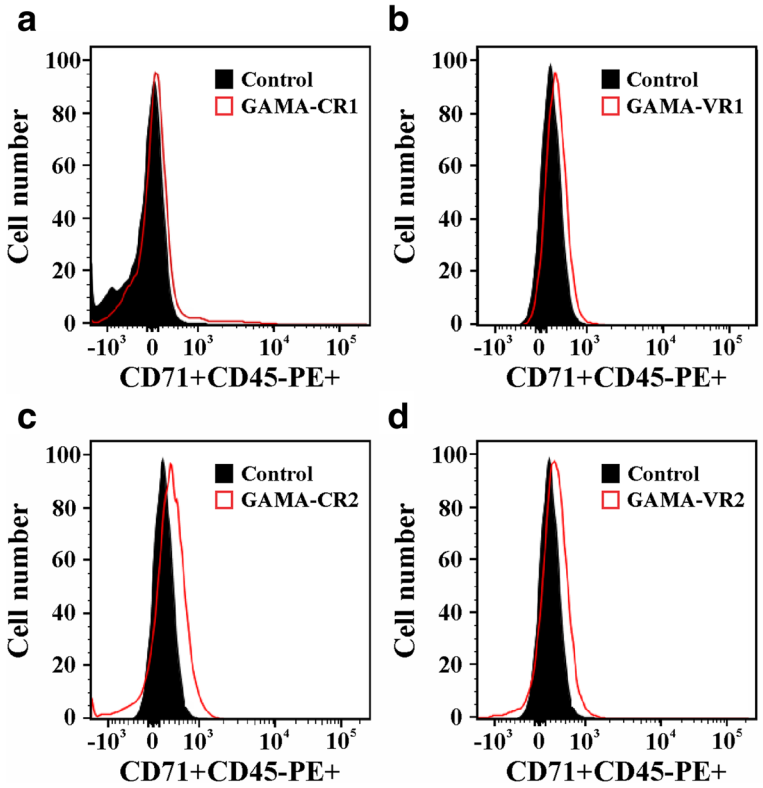


Fig. 3 Flow cytometry analysis. Histograms of conserved (**a** and **c**) and variable (**b** and **d**) GAMA fragments compared to control (cells not incubated with the protein). Each figure is representative from three independent experiments

(Fig. 4b). These findings suggested that GAMA in *P. vivax* has a functional role in preferential interaction with human reticulocytes.

Natural antibodies did not affect *Pv*GAMA binding activity

A cytometry adhesion inhibition assay was performed with sera from individuals suffering *P. vivax* malaria to determine whether the antibodies produced during natural infection could inhibit functional conserved regions (CR1 and CR2) interaction with reticulocytes. Figure 4c shows that conserved recombinant proteins pre-incubated with human sera were able to bind to target cells (CR1: $m \pm SD = 6.21 \pm 0.27$; CR2: $m \pm SD = 9.83 \pm 0.09$), giving a similar percentage to that for controls (CR1: $m \pm SD = 6.5 \pm 0.08$; CR2: $m \pm SD = 10.01 \pm 0.95$) (CR1: *t*-test: $t_{(3)} = 0.55$, $P = 0.617$; CR2: *t*-test: $t_{(4)} = 0.37$, $P = 0.730$), suggesting that the immune response was directed against regions which are not implicated in cell binding.

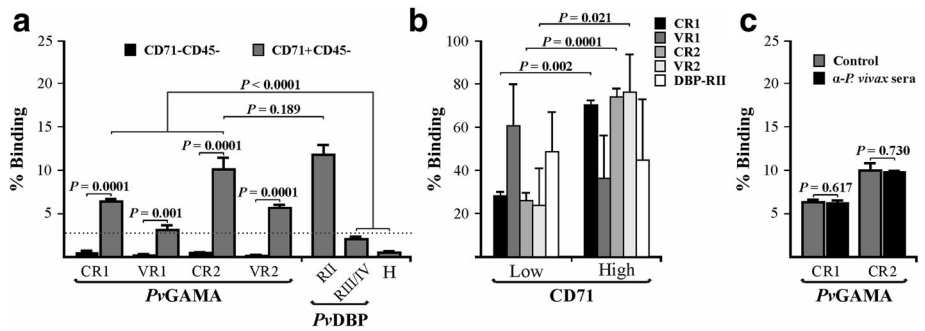


Fig. 4 *Pv*GAMA human reticulocyte binding activity. Flow cytometry analysis showing the recombinant binding percentage to CD71-CD45- and CD71 + CD45- cells (**a**) and regarding CD71-APCH7 signal (only for CD71 + CD45- cells) (**b**). Positive (DBP-RII) and negative (DBP-RIII/IV and H (6 histidine peptide)) binding controls are also shown. **c** CR1 and CR2 reticulocyte binding inhibition assay using human sera (α-*P. vivax* sera). Binding percentage in both analyses were expressed as mean ± SD of three independent experiments

Discussion

Merozoite invasion of erythrocytes involves the participation of several parasite molecules expressed at the end of the intra-erythrocyte lifecycle, mainly those contained in the apical organelles, such as the rhoptries and micronemes [8]. Only a few of these molecules possessing a reticulocyte binding role in *P. vivax* have been identified and their binding domains mapped, suggesting an urgent need for performing further studies to supplement current knowledge on *P. vivax* adhesins. This will improve our understanding of the molecular basis of parasite invasion of reticulocytes. This study aimed at using natural selection analysis for identifying GAMA functional regions playing a potential role in reticulocyte binding.

According to the phylogenetic analysis, a repeat region (RR) localised between amino acids 591 and 695 consisting of residues [A/L]AN[A/G][N/D] was predicted. This RR was common in different *P. vivax* strains but not in phylogenetically-related species (Additional file 1: Figure S1). This characteristic has been found in several *P. vivax* antigens described in the *P. vivax* VCG-I strain located on the parasite surface (Pv12 [12], ARP [43]) or in the apical pole (Pv34 [44], RON1 [45], RON2 [46] and RON4 [47, 48]). DNA sequences from different *P. vivax* strains and phylogenetically-related species were thus compared to ascertain whether *gama* gene diversity has been modulated by immune pressure. Evidence of episodic positive selection was found in some parasite lineages (Additional file 2: Figure S2). As shown for other antigens [49–51], the episodic selection found in GAMA could be the outcome of adaptation to different hosts during malaria-primate evolution [50, 51]. Therefore, the insertions found in *P. vivax* could be an adaptation of the species to humans since the RR in malaria are associated with evasion of the host's immune response, making such response become directed against functionally unimportant regions [52, 53]. This hypothesis was supported by the fact that r*Pv*GAMA-Ct (where the RR is located) can trigger a species-specific immune response (Fig. 2) which did not inhibit CR2 binding activity to reticulocytes (Fig. 4c).

Polymorphic regions induce high levels of strain-specific antibodies (allele specific) whilst conserved regions (directly implicated in interaction with cell receptors) are usually non-antigenic [54]. Therefore, the immune response must be directed against conserved regions to avoid different parasite strains evading immunity, thereby reducing vaccine efficacy. According to the selection signal identification strategy, low genetic diversity was found in the GAMA-encoding gene, comparable to that observed in *msp4* [55, 56], *msp7A/7 K/7 F/7 L* [57, 58], *msp8* [59], *msp10* [57, 59], *pv12*, *pv38* [22, 24], *pv41* [23, 24], *rap1/2* [60] and *ron4* [48] which seem involved in host cell invasion. Despite the lack of statistically significant values for d_N-d_S difference, K_S divergence amongst species was greater than K_N, suggesting negative selection. Many codons were found to be experiencing negative selection which probably plays an important role in GAMA evolution. Two regions along the antigen were highly conserved amongst species, giving a < 0.5 evolutionary rate (ω) (Fig. 1).

Given the polymorphism and selection analysis, it was decided to determine *Pv*GAMA conserved and variable region interaction with reticulocytes to validate the *in silico* prediction of functional regions (Figs. 3 and 4) and elucidate the protein's function. A reticulocyte sample having a Duffy positive phenotype was used, given that *Pv*GAMA reportedly has a binding role regardless of such antigen's expression [61]. Unlike Cheng and his group, the anti-CD71 monoclonal antibody was included for identifying GAMA regions' preference for immature reticulocyte binding as *P. vivax* merozoites have tropism for this cell type (characterised by the expression of the CD71 receptor [62]). Given that the CD71 marker is also present in activated lymphocytes, a nucleated cell depleted umbilical cord blood sample was used. The anti-CD45 was also included to totally exclude the lymphocytes from the analysis once the Wright staining revealed 0.4% of such cells (also confirmed by cytometry analysis) (Additional file 4: Figure S4). It was also confirmed that there was no difference in reticulocyte percentage by incubating the samples for 4 and 16 h at 4 °C (4 h: $m \pm SD = 1.24 \pm 0.27$; 16 h: $m \pm SD = 1.31 \pm 0.07$) (*t*-test: $t_{(2)} = 0.32$, $P > 0.777$). However, it was decided to use a prolonged incubation time to enable complete protein-cell interaction.

It was found that all *Pv*GAMA fragments bound to mature erythrocytes (CD71-CD45-) though to a lesser extent compared to reticulocytes (CD71 + CD45-) (Fig. 4a), thereby supporting the fact that the protein preferentially interacts with the latter cell type. The conserved fragment located in the carboxyl region (CR2) had higher reticulocyte binding than the amino one (CR1) (Fig. 4a) coinciding with that shown recently for *Pv*GAMA where this fragment [F2 (aa 345 to 589) or F7 (408 to 589) regions in that study] showed higher rosetting activity, unlike the F1 region (aa 22 to 344) (amino fragment) [61]. Interestingly, CR1 and CR2 had higher $CD71^{hi}$ reticulocyte binding percentages than to $CD71^{lo}$ (Fig. 4b), suggesting that GAMA mainly binds to such cell type's most immature stage. It has been reported that some reticulocytes' integral membrane components decrease as cells mature [63]. Therefore, the findings found here suggest that *Pv*GAMA receptor is less abundant in $CD71^{lo}$ cells unlike $CD71^{hi}$, as a consequence of cell maturation. The fact that more than 69% of the CD71 + CD45- cells were $CD71^{lo}$ ($m \pm$

SD = 69.3 ± 3.3) can be the explanation of why *Pv*GAMA fragment binding to 100% of the CD71+ reticulocytes was not found (Fig. 4a). It has been observed that several *P. vivax* proteins, such as DBP [64], MSP-1 [16], RBP1 [14], the erythrocyte binding protein (EBP) [42], RBP1a, RBP1b [65] and RBP2 [15], have preferential reticulocyte binding activity, being the RBPs particularly important in parasite cell selection. Taking the results obtained here into account, it can be suggested that *P. vivax* target cell selection is not only governed by the RBPs but other ligands are also taking place in this process, such as DBP, MSP-1, EBP and now *Pv*GAMA.

Immunoreactive proteins are considered potential candidates for developing a vaccine as it has been seen that an immune response induced during infection is related to naturally-acquired immunity [66]. Antigenicity is thus one of the classical parameters for selecting molecules when developing a vaccine. Although there was an immune response against *Pv*GAMA (Fig. 2), this was not sufficient to inhibit the conserved regions binding to reticulocytes (Fig. 4c). It has been observed that *P. falciparum* proteins' conserved regions (implicated in target cell binding) cannot trigger an immune response when used as vaccine candidates in the *Aotus* model whilst non-conserved ones trigger protective responses upon parasite challenge but those are strain-specific [54]. Accordingly, the *Pv*GAMA antibodies produced/induced during natural *P. vivax* infection were directed against immunodominant epitopes which are unimportant in binding activity. Bearing in mind that functional regions usually evolve more slowly and that natural negative selection tends to keep these regions conserved amongst species [25], our experimental findings suggested that CR1 and CR2 located between residues 80–320 (40% of negatively selected sites) and 514–624 (64.5% of negatively selected sites) are functionally/structurally restricted and that vaccine design should thus be focused on them.

Conclusions

To our knowledge, this study described *Pv*GAMA reticulocyte binding properties for the first time. The *Pv*GAMA antigenic response was principally directed against its carboxyl fragment which comprises by a repetitive region. On the other hand, it was shown that *Pv*GAMA consists of two conserved binding fragments that bind preferentially to most immature human reticulocytes, which is consistent with the *P. vivax* invasion phenotype and highlights the fact that functional regions can be predicted by analysing natural selection. Further studies aimed at discerning the function of conserved regions as vaccine components are required.

Additional files

Additional file 1: Figure S1. GAMA antigen alignment. *pvgama* sequences from 6 *P. vivax* strains were aligned with orthologous sequences from *P. cynomolgi*, *P. inui*, *P. fragile*, *P. coatneyi* and *P. knowlesi*. a DNA sequence alignment. b Deduced amino acid alignment. The sequences were obtained from GenBank: access numbers being India-VII AFBK01000586-AFBK01000587, North Korean AFNJ01000531, Brazil-I AFMK01000508-AFMK01000509, Mauritania-I AFNI01000333-AFNI01000334, *P. inui* NW_0084818881, *P. fragile* NW_012192586, *P. cynomolgi* BAEJ01000249, *P. coatneyi* CM0028561 and *P. knowlesi* NC_0119061. (PDF 373 kb)

Additional file 2: Figure S2. Lineage-specific positive selection. Branches under positive episodic selection were identified by using the REL-site branch method. Episodic selection acts very quickly and involves a switch from negative to positive natural selection and back to negative and might enable adaptation to a new host. Phylogeny was inferred in MEGA v6 by the maximum likelihood method using the GTR + G evolutionary model. ω^+ model: ω rate values. Pr [ω = ω +]: percentage of sites evolving under positive selection. *P*-value corrected for multiple tests using the Holm-Bonferroni method. (TIF 470 kb)

Additional file 3: Figure S3. Obtaining recombinant proteins. a, b Recombinant GAMA protein expression and purification. Lanes 2–3 show non-induced and induced cell lysate, respectively. Lanes 4–5 show purified r*Pv*GAMA-Nt and -Ct stained with Coomassie blue or analysed by western blot using anti-polyhistidine antibodies, respectively. **c** Purifying conserved (CR1 and CR2) and variable (VR1 and VR2) *Pv*GAMA regions. Lanes 2, 4, 6 and 8 show purified recombinant proteins and lanes 3, 5, 7 and 9 show western blot detection. The proteins' molecular markers are indicated in Lane 1 on all figures. (TIF 5327 kb)

Additional file 4: Figure S4. Selection strategy for immature and mature erythrocyte populations. The doublets were excluded by plotting FSC-H against FSC-A. Cells were then selected by their granularity, using an SSC-A vs FSC-A cytogram. The CD45 vs CD71 signal was plotted for selecting reticulocyte (CD71 + CD45-) and mature erythrocyte (CD71-CD45-) populations and omitting activated lymphocytes (CD71 + CD45+). The percentage of cells having bound protein was calculated using the PE signal (CD71 + CD45-PE+). A representative histogram from three independent experiments analysing the PE signal for the CR2 binding assay compared to control is also shown. (TIF 10448 kb)

Abbreviations

ANOVA: Analysis of variance; $CD71^{hi}$: CD71 high; $CD71^{lo}$: CD71 low; CR: Conserved region; DARC: Duffy antigen receptor for chemokines; DBP: Duffy binding protein; DEB: Denaturing extraction buffer; EBP: Erythrocyte binding protein; ELISA: Enzyme-linked immunosorbent assay; LB: Luria bertani; MALDI-TOF: Matrix-assisted laser desorption/ionization-time of flight; MS: Mass spectrometry; MSP-1: Merozoite surface protein 1; NEB: Native extraction buffer; OD: Optical density; PBS: Phosphate buffered saline; *Pv*GAMA: *P. vivax* GPI-anchored micronemal antigen; *Pvgama*: *Plasmodium vivax gama*; RBP: Reticulocyte binding protein; RON5: Rhoptry neck protein 5; RP-HPLC: Reverse phase high-performance liquid chromatography; RR: Repeat region; RT: Room temperature; SD: Standard deviation; TRAg: Tryptophan-rich antigen; VCG-I: Vivax Colombia Guaviare 1; VR: Variable region

Acknowledgements

We would like to thank Ana María Perdomo and Bernardo Camacho for supplying the umbilical cord blood, Diana Díaz for technical support in cytometry and Jason Garry for translating this manuscript.

Funding

This research was financed by the Colombian Science, Technology and Innovation Department (COLCIENCIAS), contract RC#0309-2013. During the course of this research JFR and HDOS were financed via COLCIENCIAS cooperation agreement # 0719–2013 and # 0555–2015, respectively. The sponsors had no role in study design, collection and analysis or data interpretation.

Availability of data and materials
All data generated or analysed during this study are included within this article and its additional files. The *pvgama* sequence from *P. vivax* VCG-I strain was deposited in the GenBank database under accession number KT248546.

Authors' contributions
LAB and DAMP devised and designed the study; LAB, DAMP, DGO, JFR and HDOS performed the experiments; LAB, DAMP, DGO and MAP analysed the results and wrote the manuscript. All authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Consent for publication
Not applicable.

Ethics approval and consent to participate
All individuals who participated in this research (including progenitors regarding umbilical cord samples) signed an informed consent form after receiving detailed information regarding the study's goals. All procedures were approved by FIDIC's ethics committee.

Publisher's Note

Author details
[1]Molecular Biology and Immunology Department, Fundación Instituto de Inmunología de Colombia (FIDIC), Carrera 50 No. 26-20, Bogotá DC, Colombia. [2]PhD Programme in Biomedical and Biological Sciences, Universidad del Rosario, Carrera 24 No. 63C-69, Bogotá DC, Colombia. [3]Basic Sciences Department, School of Medicine and Health Sciences, Universidad del Rosario, Carrera 24 No. 63C-69, Bogotá DC, Colombia.

Received: 27 October 2016 Accepted: 10 May 2017
Published online: 19 May 2017

References
1. WHO. World malaria report. 2014.
2. WHO. State of the art of new vaccine research and development. 2006.
3. Patarroyo ME, Bermudez A, Patarroyo MA. Structural and immunological principles leading to chemically synthesized, multiantigenic, multistage, minimal subunit-based vaccine development. Chem Rev. 2011;111:3459–507.
4. Lanzavecchia A, Fruhwirth A, Perez L, Corti D. Antibody-guided vaccine design: identification of protective epitopes. Curr Opin Immunol. 2016;41:62–7.
5. Gardner MJ, Hall N, Fung E, White O, Berriman M, Hyman RW, et al. Genome sequence of the human malaria parasite *Plasmodium falciparum*. Nature. 2002;419:498–511.
6. Bozdech Z, Llinas M, Pulliam BL, Wong ED, Zhu J, DeRisi JL. The transcriptome of the intraerythrocytic developmental cycle of *Plasmodium falciparum*. PLoS Biol. 2003;1:E5.
7. Lasonder E, Ishihama Y, Andersen JS, Vermunt AM, Pain A, Sauerwein RW, et al. Analysis of the *Plasmodium falciparum* proteome by high-accuracy mass spectrometry. Nature. 2002;419:537–42.
8. Cowman AF, Berry D, Baum J. The cellular and molecular basis for malaria parasite invasion of the human red blood cell. J Cell Biol. 2012;198:961–71.
9. Richards JS, Beeson JG. The future for blood-stage vaccines against malaria. Immunol Cell Biol. 2009;87:377–90.
10. Arumugam TU, Takeo S, Yamasaki T, Thonkukiatkul A, Miura K, Otsuki H, et al. Discovery of GAMA, a *Plasmodium falciparum* merozoite micronemal protein, as a novel blood-stage vaccine candidate antigen. Infect Immun. 2011;79:4523–32.
11. Huynh MH, Carruthers VB. A *Toxoplasma gondii* Ortholog of *Plasmodium* GAMA contributes to parasite attachment and cell invasion. mSphere. 2016;1: doi:10.1128/mSphere.00012-16.
12. Moreno-Perez DA, Areiza-Rojas R, Florez-Buitrago X, Silva Y, Patarroyo ME, Patarroyo MA. The GPI-anchored 6-Cys protein *Pv*12 is present in detergent-resistant microdomains of *Plasmodium vivax* blood stage schizonts. Protist. 2013;164:37–48.
13. Patarroyo MA, Calderón D, Moreno-Pérez DA. Vaccines against *Plasmodium vivax*: a research challenge. Expert Rev Vaccines. 2012;11:1249–60.
14. Urquiza M, Patarroyo MA, Mari V, Ocampo M, Suarez J, Lopez R, et al. Identification and polymorphism of *Plasmodium vivax* RBP-1 peptides which bind specifically to reticulocytes. Peptides. 2002;23:2265–77.
15. Franca CT, He WQ, Gruszczyk J, Lim NT, Lin E, Kiniboro B, et al. *Plasmodium vivax* reticulocyte binding proteins are key targets of naturally acquired immunity in young Papua New Guinean children. PLoS Negl Trop Dis. 2016;10:e0005014.
16. Rodriguez LE, Urquiza M, Ocampo M, Curtidor H, Suarez J, Garcia J, et al. *Plasmodium vivax* MSP-1 peptides have high specific binding activity to human reticulocytes. Vaccine. 2002;20:1331–9.
17. Zeeshan M, Tyagi RK, Tyagi K, Alam MS, Sharma YD. Host-parasite interaction: selective Pv-fam-a family proteins of *Plasmodium vivax* bind to a restricted number of human erythrocyte receptors. J Infect Dis. 2015;211:1111–20.
18. Arevalo-Pinzon G, Bermudez M, Curtidor H, Patarroyo MA. The *Plasmodium vivax* rhoptry neck protein 5 is expressed in the apical pole of *Plasmodium vivax* VCG-1 strain schizonts and binds to human reticulocytes. Malaria J. 2015;14:106.
19. Garzon-Ospina D, Forero-Rodriguez J, Patarroyo MA. Inferring natural selection signals in *Plasmodium vivax*-encoded proteins having a potential role in merozoite invasion. Infect Genet Evol. 2015;33:182–8.
20. Cornejo OE, Fisher D, Escalante AA. Genome-wide patterns of genetic polymorphism and signatures of selection in *Plasmodium vivax*. Genome Biol Evol. 2014;7:106–19.
21. Neafsey DE, Galinsky K, Jiang RH, Young L, Sykes SM, Saif S, et al. The malaria parasite *Plasmodium vivax* exhibits greater genetic diversity than *Plasmodium falciparum*. Nat Genet. 2012;44:1046–50.
22. Forero-Rodriguez J, Garzon-Ospina D, Patarroyo MA. Low genetic diversity and functional constraint in loci encoding *Plasmodium vivax* P12 and P38 proteins in the Colombian population. Malaria J. 2014;13:58.
23. Forero-Rodriguez J, Garzon-Ospina D, Patarroyo MA. Low genetic diversity in the locus encoding the *Plasmodium vivax* P41 protein in Colombia's parasite population. Malaria J. 2014;13:388.
24. Wang Y, Ma A, Chen SB, Yang YC, Chen JH, Yin MB. Genetic diversity and natural selection of three blood-stage 6-Cys proteins in *Plasmodium vivax* populations from the China-Myanmar endemic border. Infect Genet Evol. 2014;28:167–74.
25. Graur D, Zheng Y, Price N, Azevedo RB, Zufall RA, Elhaik E. On the immortality of television sets: "function" in the human genome according to the evolution-free gospel of ENCODE. Genome Biol Evol. 2013;5:578–90.
26. Escalante AA, Cornejo OE, Freeland DE, Poe AC, Durrego E, Collins WE, et al. A monkey's tale: the origin of *Plasmodium vivax* as a human malaria parasite. Proc Natl Acad Sci USA. 2005;102:1980–5.
27. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32:1792–7.
28. Zhang J, Rosenberg HF, Nei M. Positive Darwinian selection after gene duplication in primate ribonuclease genes. Proc Natl Acad Sci USA. 1998;95:3708–13.
29. Jukes THaCRC. Evolution of protein molecules. In: Munro HN, editor. Mammalian protein metabolism. New York: Academic Press; 1969.
30. Kosakovsky Pond SL, Frost SD. Not so different after all: a comparison of methods for detecting amino acid sites under selection. Mol Biol Evol. 2005;22:1208–22.
31. Murrell B, Wertheim JO, Moola S, Weighill T, Scheffler K, Kosakovsky Pond SL. Detecting individual sites subject to episodic diversifying selection. PLoS Genet. 2012;8:e1002764.
32. Murrell B, Moola S, Mabona A, Weighill T, Sheward D, Kosakovsky Pond SL, et al. FUBAR: a fast, unconstrained bayesian approximation for inferring selection. Mol Biol Evol. 2013;30:1196–205.
33. Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SD. Automated phylogenetic detection of recombination using a genetic algorithm. Mol Biol Evol. 2006;23:1891–901.
34. Kosakovsky Pond SL, Murrell B, Fourment M, Frost SD, Delport W, Scheffler K. A random effects branch-site model for detecting episodic diversifying selection. Mol Biol Evol. 2011;28:3033–43.
35. Delport W, Poon AF, Frost SD, Kosakovsky Pond SL. Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology. Bioinformatics. 2010;26:2455–7.
36. Aurrecoechea C, Brestelli J, Brunk BP, Dommer J, Fischer S, Gajria B, et al. PlasmoDB: a functional genomic database for malaria parasites. Nucleic Acids Res. 2009;37:D539–543.

37. Moreno-Perez DA, Degano R, Ibarrola N, Muro A, Patarroyo MA. Determining the *Plasmodium vivax* VCG-1 strain blood stage proteome. J Proteomics. 2014;113C:268–80.
38. Pico de Coana Y, Rodriguez J, Guerrero E, Barrero C, Rodriguez R, Mendoza M, et al. A highly infective *Plasmodium vivax* strain adapted to *Aotus* monkeys: quantitative haematological and molecular determinations useful for *P. vivax* malaria vaccine development. Vaccine. 2003;21:3930–7.
39. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 1994;22:4673–80.
40. Houghten RA. General method for the rapid solid-phase synthesis of large numbers of peptides: specificity of antigen-antibody interaction at the level of individual amino acids. Proc Natl Acad Sci USA. 1985;82:5131–5.
41. Lu F, Li J, Wang B, Cheng Y, Kong DH, Cui L, et al. Profiling the humoral immune responses to *Plasmodium vivax* infection and identification of candidate immunogenic rhoptry-associated membrane antigen (RAMA). J Proteomics. 2014;102C:66–82.
42. Ntumngia FB, Thomson-Luque R, Torres Lde M, Gunalan K, Carvalho LH, Adams JH. A novel erythrocyte binding protein of *Plasmodium vivax* suggests an alternate invasion pathway into Duffy-positive reticulocytes. mBio. 2016;7: doi: 10.1128/mBio.01261-16.
43. Moreno-Perez DA, Saldarriaga A, Patarroyo MA. Characterizing *Pv*ARP, a novel *Plasmodium vivax* antigen. Malaria J. 2013;12:165.
44. Mongui A, Angel DI, Gallego G, Reyes C, Martinez P, Guhl F, et al. Characterization and antigenicity of the promising vaccine candidate *Plasmodium vivax* 34 kDa rhoptry antigen (*Pv*34). Vaccine. 2009;28:415–21.
45. Moreno-Perez DA, Montenegro M, Patarroyo ME, Patarroyo MA. Identification, characterization and antigenicity of the *Plasmodium vivax* rhoptry neck protein 1 (*Pv*RON1). Malaria J. 2011;10:314.
46. Arevalo-Pinzon G, Curtidor H, Patino LC, Patarroyo MA. *Pv*RON2, a new *Plasmodium vivax* rhoptry neck antigen. Malaria J. 2011;10:60.
47. Arevalo-Pinzon G, Curtidor H, Abril J, Patarroyo MA. Annotation and characterization of the *Plasmodium vivax* rhoptry neck protein 4 (*Pv*RON4). Malaria J. 2013;12:356.
48. Buitrago SP, Garzon-Ospina D, Patarroyo MA. Size polymorphism and low sequence diversity in the locus encoding the *Plasmodium vivax* rhoptry neck protein 4 (*Pv*RON4) in Colombian isolates. Malaria J. 2016;15:501.
49. Garzon-Ospina D, Forero-Rodriguez J, Patarroyo MA. Evidence of functional divergence in MSP7 paralogous proteins: a molecular-evolutionary and phylogenetic analysis. BMC Evol Biol. 2016;16:256.
50. Muehlenbein MP, Pacheco MA, Taylor JE, Prall SP, Ambu L, Nathan S, et al. Accelerated diversification of nonhuman primate malarias in Southeast Asia: adaptive radiation or geographic speciation? Mol Biol Evol. 2015;32:422–39.
51. Sawai H, Otani H, Arisue N, Palacpac N, de Oliveira ML, Pathirana S, et al. Lineage-specific positive selection at the merozoite surface protein 1 (msp1) locus of *Plasmodium vivax* and related simian malaria parasites. BMC Evol Biol. 2010;10:52.
52. Schofield L. On the function of repetitive domains in protein antigens of *Plasmodium* and other eukaryotic parasites. Parasitol Today. 1991;7:99–105.
53. Ferreira MU, da Silva NM, Wunderlich G. Antigenic diversity and immune evasion by malaria parasites. Clin Diagn Lab Immunol. 2004;11:987–95.
54. Patarroyo ME, Patarroyo MA. Emerging rules for subunit-based, multiantigenic, multistage chemically synthesized vaccines. Acc Chem Res. 2008;41:377–86.
55. Martinez P, Suarez CF, Gomez A, Cardenas PP, Guerrero JE, Patarroyo MA. High level of conservation in *Plasmodium vivax* merozoite surface protein 4 (*Pv*MSP4). Infect Genet Evol. 2005;5:354–61.
56. Putaporntip C, Jongwutiwes S, Ferreira MU, Kanbara H, Udomsangpetch R, Cui L. Limited global diversity of the *Plasmodium vivax* merozoite surface protein 4 gene. Infect Genet Evol. 2009;9:821–6.
57. Garzon-Ospina D, Romero-Murillo L, Tobon LF, Patarroyo MA. Low genetic polymorphism of merozoite surface proteins 7 and 10 in Colombian *Plasmodium vivax* isolates. Infect Genet Evol. 2011;11:528–31.
58. Garzon-Ospina D, Forero-Rodriguez J, Patarroyo MA. Heterogeneous genetic diversity pattern in *Plasmodium vivax* genes encoding merozoite surface proteins (MSP) -7E, -7 F and -7 L. Malaria J. 2014;13:495.
59. Pacheco MA, Elango AP, Rahman AA, Fisher D, Collins WE, Barnwell JW, et al. Evidence of purifying selection on merozoite surface protein 8 (MSP8) and 10 (MSP10) in *Plasmodium* spp. Infect Genet Evol. 2012;12:978–86.
60. Garzon-Ospina D, Romero-Murillo L, Patarroyo MA. Limited genetic polymorphism of the *Plasmodium vivax* low molecular weight rhoptry protein complex in the Colombian population. Infect Genet Evol. 2010;10:261–7.
61. Cheng Y, Lu F, Wang B, Li J, Han JH, Ito D, et al. *Plasmodium vivax* GPI-anchored micronemal antigen (*Pv*GAMA) binds human erythrocytes independent of Duffy antigen status. Sci Reports. 2016;6:35581.
62. Malleret B, Li A, Zhang R, Tan KS, Suwanarusk R, Claser C, et al. *Plasmodium vivax*: restricted tropism and rapid remodeling of CD71-positive reticulocytes. Blood. 2015;125:1314–24.
63. Wilson MC, Trakarnsanga K, Heesom KJ, Cogan N, Green C, Toye AM, et al. Comparison of the proteome of adult and cord erythroid cells, and changes in the proteome following reticulocyte maturation. Mol Cel Proteomics MCP. 2016;15:1938–46.
64. Ocampo M, Vera R, Eduardo R, Curtidor H, Urquiza M, Suarez J, et al. *Plasmodium vivax* Duffy binding protein peptides specifically bind to reticulocytes. Peptides. 2002;23:13–22.
65. Han JH, Lee SK, Wang B, Muh F, Nyunt MH, Na S, et al. Identification of a reticulocyte-specific binding domain of *Plasmodium vivax* reticulocyte-binding protein 1 that is homologous to the *Pf*Rh4 erythrocyte-binding domain. Sci Reports. 2016;6:26993.
66. Dent AE, Nakajima R, Liang L, Baum E, Moormann AM, Sumba PO, et al. *Plasmodium falciparum* protein microarray antibody profiles correlate with protection from symptomatic malaria in Kenya. J Infect Dis. 2015;212:1429–38.

frontiers in Genetics

ORIGINAL RESEARCH
published: 25 January 2018
doi: 10.3389/fgene.2018.00010

Identifying Potential *Plasmodium vivax* Sporozoite Stage Vaccine Candidates: An Analysis of Genetic Diversity and Natural Selection

Diego Garzón-Ospina[1,2], *Sindy P. Buitrago*[1], *Andrea E. Ramos*[1] and *Manuel A. Patarroyo*[1,3]*

[1] Molecular Biology and Immunology Laboratory, Fundación Instituto de Inmunología de Colombia, Bogotá, Colombia, [2] PhD Programme in Biomedical and Biological Sciences, School of Medicine and Health Sciences, Universidad del Rosario, Bogotá, Colombia, [3] Basic Sciences Department, School of Medicine and Health Sciences, Universidad del Rosario, Bogotá, Colombia

Parasite antigen genetic diversity represents a great obstacle when designing a vaccine against malaria caused by *Plasmodium vivax*. Selecting vaccine candidate antigens has been focused on those fulfilling a role in invasion and which are conserved, thus avoiding specific-allele immune responses. Most antigens described to date belong to the blood stage, thereby blocking parasite development within red blood cells, whilst studying antigens from other stages has been quite restricted. Antigens from different parasite stages are required for developing a completely effective vaccine; thus, pre-erythrocyte stage antigens able to block the first line of infection becoming established should also be taken into account. However, few antigens from this stage have been studied to date. Several *P. falciparum* sporozoite antigens are involved in invasion. Since 77% of genes are orthologous amongst *Plasmodium* parasites, *P. vivax* sporozoite antigen orthologs to those of *P. falciparum* might be present in its genome. Although these genes might have high genetic diversity, conserved functionally-relevant regions (ideal for vaccine development) could be predicted by comparing genetic diversity patterns and evolutionary rates. This study was thus aimed at searching for putative *P. vivax* sporozoite genes so as to analyse their genetic diversity for determining their potential as vaccine candidates. Several DNA sequence polymorphism estimators were computed at each locus. The evolutionary force (drift, selection and recombination) drawing the genetic diversity pattern observed was also determined by using tests based on polymorphism frequency spectrum as well as the type of intra- and inter-species substitutions. Likewise, recombination was assessed both indirectly and directly. The results showed that sporozoite genes were more conserved than merozoite genes evaluated to date. Putative domains implied in cell traversal, gliding motility and hepatocyte interaction had a negative selection signal, being conserved amongst different species in the genus. PvP52, PvP36, PvSPATR, PvPLP1, PvMCP1, PvTLP, PvCelTOS, and PvMB2 antigens or functionally restricted regions within them would thus seem promising vaccine candidates and could be used when designing a pre-erythrocyte and/or multi-stage vaccine against *P. vivax* to avoid allele-specific immune responses that could reduce vaccine efficacy.

Keywords: ***Plasmodium vivax*, genetic diversity, sporozoite, vaccine, natural selection, hepatocyte invasion**

OPEN ACCESS

Edited by:
José M. Álvarez-Castro, Universidade de Santiago de Compostela, Spain

Reviewed by:
Marcelo R. S. Briones, Federal University of São Paulo, Brazil
Gonzalo Gajardo, University of Los Lagos, Chile

***Correspondence:**
Manuel A. Patarroyo
mapatarr.fidic@gmail.com

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *24 October 2017*
Accepted: *09 January 2018*
Published: *25 January 2018*

Citation:
Garzón-Ospina D, Buitrago SP, Ramos AE and Patarroyo MA (2018) Identifying Potential Plasmodium vivax Sporozoite Stage Vaccine Candidates: An Analysis of Genetic Diversity and Natural Selection. Front. Genet. 9:10. doi: 10.3389/fgene.2018.00010

INTRODUCTION

Plasmodium vivax (Pv) is one of the five *Plasmodium* species causing malaria in human beings [Coatney and National Institute of Allergy and Infectious Diseases (U.S.), 1971; Rich and Ayala, 2003]. Outdoor biting of less-anthropophilic mosquitos (than the main *Plasmodium falciparum* vectors) transmitting it, and endemic regions' social-economic conditions make *P. vivax* an emergent public health problem (Mueller et al., 2015). This parasite exclusively invades reticulocytes and is characterized by relapses from dormant liver stages; it produces early and continuous gametocytes (Price et al., 2009; Patarroyo et al., 2012; Adams and Mueller, 2017) and has great genetic diversity throughout its genome (Neafsey et al., 2012; Winter et al., 2015). All these features make *P. vivax* control and elimination a great challenge.

Vaccine development has been considered as one of the most cost-effective interventions for controlling malaria. Designing a vaccine against this disease has focused on selecting antigens able to induce an effective immune response that block invasion of target cells. The *Plasmodium* life-cycle should be considered when designing an anti-malarial vaccine. Malarial infection begins with an infected female mosquito's bite. *Plasmodium* sporozoites (Spz or pre-erythrocyte stage) in the vertebrate bloodstream must migrate to the host's liver, traversing the endothelial and Kupffer cells that form the sinusoidal barrier. They then migrate through some hepatocytes before infecting one of them (Menard, 2001; Frevert, 2004). Inside hepatocytes, Spz differentiate into thousands of merozoites (Mrz) which after their release proceed to invade red blood cells (RBC), initiating the erythrocyte or blood stage. Within RBC the Mrz could differentiate in new Mrz which will infect new RBC or into gametocytes which can be taken by the mosquito vector to start the sexual stage.

As mentioned above, proteins involved in parasite-host cell interactions are the main targets for vaccine development. However, the genetic diversity found in the parasite has become a challenge for designing a fully-effective vaccine (Patarroyo et al., 2012; Barry and Arnott, 2014). Such polymorphisms are typically found within functionally irrelevant gene/protein regions enabling the evasion of host immune responses, whilst functionally important regions remain conserved due to functional/structural constraints (Garzón-Ospina et al., 2012; Baquero et al., 2017); these regions could therefore be taken into account for vaccine design in order to avoid allele-specific immune responses. Several studies have measured potential vaccine candidates' genetic diversity (Putaporntip et al., 1997, 2009, 2010; Gomez et al., 2006; Garzón-Ospina et al., 2010, 2011, 2012, 2014; Dias et al., 2011; Premaratne et al., 2011; Chenet et al., 2013; Barry and Arnott, 2014; Forero-Rodriguez et al., 2014a,b; Buitrago et al., 2016; Chaurio et al., 2016; Mehrizi et al., 2017). Likewise, the evolutionary forces (mutation, natural selection, genetic drift, recombination, and migration) modulating polymorphism (Casillas and Barbadilla, 2017) have also been determined. This has been used for monitoring anti-malarial vaccine targets (Barry and Arnott, 2014) but might also be used for predicting functional regions which are usually conserved amongst species (Kimura, 1983; Graur et al., 2013). Accordingly, promising vaccine candidates must likely be parasite proteins playing an important role during target cell invasion but displaying limited genetic diversity or, at least, a domain having such pattern. These genes or domains must thus have a negative selection signal ($\omega < 1$ evolutionary rate). Moreover, vaccine candidates should be able to induce an immune response in natural or experimental infection (Patarroyo et al., 2012; Barry and Arnott, 2014; Weiss et al., 2015).

Despite there being three intervention points [pre-erythrocyte, blood and gametocyte stages (Barry and Arnott, 2014)], potential *P. vivax* candidates characterized to date have mainly been described for the blood stage to avoid Mrz entry to RBC, preventing the disease's typical symptomatology (Patarroyo et al., 2012). On the contrary, few pre-erythrocyte phase antigens have been studied, in spite of the fact that blocking Spz interaction with hepatocytes would greatly reduce the possibility of developing the disease. This strategy would also avoid *P. vivax* dormant form formation in the liver and hence infected patients' constant relapses (Price et al., 2009; Hulden, 2011). Several Spz antigens involved in invasion have been characterized in *P. falciparum* and other parasite species (Sultan et al., 1997; Wengelnik et al., 1999; Menard, 2001; Matuschewski et al., 2002; Romero et al., 2004; Kariu et al., 2006; Labaied et al., 2007a,b; Moreira et al., 2008; Rosado et al., 2008; Engelmann et al., 2009; Alba et al., 2011; Curtidor et al., 2011; Aldrich et al., 2012; Annoura et al., 2014; Ferguson et al., 2014; Jimah et al., 2016; Kublin et al., 2017; Manzoni et al., 2017; Yang et al., 2017a). Since around 77% of *Plasmodium* genes are orthologous (Carlton et al., 2008), many being essential for parasite survival (Bushell et al., 2017), orthologous Spz antigens might also be present in the *P. vivax* genome. Given the *P. vivax* genome's high genetic diversity (Neafsey et al., 2012), *P. vivax* Spz genes might also be highly polymorphic; however, as functionally important regions tend to have low polymorphism (Kimura, 1983; Graur et al., 2013), *P. vivax* and related *Plasmodium* parasites' highly conserved regions could

Abbreviations: BepiPred, B-cell epitope prediction; CBI, codon bias index; ENC, effective number of codons; FEL, fixed effects likelihood; FUBAR, fast, unconstrained Bayesian approximation for inferring selection; GARD, genetic algorithm recombination detection; GPI, glycosylphosphatidylinositol membrane anchor; GTP, guanosine triphosphate; IEDB, immune epitope database and analysis resource; IFEL, internal fixed effects likelihood; MEME, mixed effects model of evolution; Mrz, merozoite; *pvspect1,* gene encoding *Plasmodium vivax* sporozoite protein essential for cell traversal; *pvsiap1,* gene encoding *Plasmodium vivax* sporozoite invasion-associated protein-1; *pvp52,* gene encoding *Plasmodium vivax* 6-cysteine protein P52; *pvp36*, gene encoding *Plasmodium vivax* 6-cysteine protein P36; *pvspatr*, gene encoding *Plasmodium vivax* secreted protein with altered thrombospondin repeat domain; *pvtrsp,* gene encoding *Plasmodium vivax* thrombospondin-related sporozoite protein; *pvtrap*, gene encoding *Plasmodium vivax* thrombospondin-related anonymous protein; *pvsiap2*, gene encoding *Plasmodium vivax* sporozoite invasion-associated protein-2; *pvmaebl,* gene encoding *Plasmodium vivax* merozoite adhesive erythrocytic binding protein; *pvplp1*, gene encoding *Plasmodium vivax* perforin-like protein 1; *pvmcp1*, gene encoding *Plasmodium vivax* merozoite capping protein 1; *pvtlp*, gene encoding *Plasmodium vivax* thioredoxin-like protein; *pvceltos*, gene encoding *Plasmodium vivax* cell traversal protein for ookinetes and sporozoites; *pvmb2,* gene encoding *Plasmodium vivax* MB2 protein; REL, random effects likelihood; SLAC, single likelihood ancestor counting; SNP, single nucleotide polymorphisms; Spz, sporozoite.

be functionally important and should therefore be taken into account for vaccine development (Patarroyo et al., 2012). As few *P. vivax* pre-erythrocyte stage antigens have been described to date (Menard, 2001; Castellanos et al., 2007; Barry and Arnott, 2014), and because characterizing antigens from this stage is laborious, this study involved a search for orthologous genes in *P. vivax* to those previously studied in *P. falciparum* for *in silico* characterization so as to assess their genetic diversity. The evolutionary forces modulating the variation pattern observed were analyzed for identifying conserved and functional regions; then proteins' antigenic potential was predicted for identifying those potentially able to induce an immune response and thus determine which of these antigens should be prioritized and taken into account when designing a pre-erythrocyte and/or multi-stage vaccine against malaria caused by *P. vivax*.

METHODOLOGY

Sequences Data Set and *in Silico* Characterization of *P. vivax* Loci

Few genes encoding putative *Plasmodium vivax* Spz stage vaccine antigens have been assessed regarding their genetic diversity. Fifteen *P. falciparum* proteins have been suggested as promising vaccine candidates (Curtidor et al., 2011). Orthologous sequences to these *P. falciparum* vaccine candidates were sought in the *P. vivax* Salvador I (Sal-I) strain from the PlasmoDB database (Release 32). One hundred and seventy-one *P. vivax* natural strain DNA sequences from regions worldwide, analyzed by whole genome sequencing (Chan et al., 2012; Neafsey et al., 2012; Hester et al., 2013; Hupalo et al., 2016) available in the PlasmoDB database, were also obtained for these putative antigens. Sal-I sequences from these genes were then used as query for searching orthologous sequences in closely-related species [*Plasmodium cynomolgi* (GCA_000321355.1), *Plasmodium inui* (GCA_000524495.1), *Plasmodium fragile* (GCA_000956335.1), *Plasmodium knowlesi* (GCA_000006355.1) and *Plasmodium coatneyi* (GCA_000725905.1)], using the whole genome data available in GenBank. Orthologous sequences from less related *Plasmodium* species (*Plasmodium berghei, Plasmodium yoelii, Plasmodium chabaudi, Plasmodium vinckei, Plasmodium falciparum, Plasmodium reichenowi, Plasmodium gaboni*, and *Plasmodium gallinaceum*) were obtained from the PlasmoDB database (Release 32).

Potential vaccine candidates described in Mrz are characterized by having a signal peptide and some have membrane anchoring structures [i.e., transmembrane helices and/or glycosylphosphatidylinositol (GPI) anchor], whilst some others have binding and/or protein-protein interaction domains (Patarroyo et al., 2012), therefore, the Sal-I Spz protein sequences were used for *in silico* characterization using several bioinformatics tools. SignalP (Nielsen, 2017) and Phobius (Kall et al., 2007) predictors were used for ascertaining signal peptide presence; the BaCelLo algorithm (Pierleoni et al., 2006) was used for predicting antigen location. Transmembrane and/or GPI domains were evaluated with Phobius, TMHMM (Sonnhammer et al., 1998) and GPI-SOM (Fankhauser and Maser, 2005) and the Pfam database was searched for putative domains.

Alignment and Sequence Analysis

The 171 sequences from different locations worldwide obtained for each gene were screened to rule out gene sequences having missing data or ambiguous nucleotides; introns were removed from those genes having them. Multiple DNA alignment for each gene was performed based on amino acid (aa) information using the TranslatorX (Abascal et al., 2010) server with the Muscle algorithm (Edgar, 2004).

DnaSP v5 software (Librado and Rozas, 2009) was then used for calculating several genetic diversity estimators for 14 Spz loci. On the other hand, the effective number of codons (ENC) and codon bias index (CBI) were obtained as a measure of selective pressure at translational level (Novembre, 2002). Tajima (1989), Fu and Li (1993), and Fay and Wu (2000) tests were used for evaluating a neutral model of molecular evolution. Repeat regions or those having insertions/deletions were not taken into account for analysis. Tests based on the type of nucleotide substitution were used to infer natural selection signals within genes. The Nei-Gojobori modified method (Zhang et al., 1998) was used for calculating the difference between non-synonymous and synonymous substitution rates at intra-species level (d_N-d_S). The difference between non-synonymous and synonymous divergences rates (K_N-K_S) was calculated by modified Nei-Gojobori method with Jukes-Cantor correction (Jukes and Cantor, 1969), using *P. vivax* sequences together with orthologous sequences from phylogenetically-closely related species for determining natural selection signals at inter-species level. MEGA v6 software (Tamura et al., 2013) was used for all such analysis. A sliding window for omega rates ($\omega = d_N/d_S$ and/or K_N/K_S) was used for evaluating the effect of natural selection throughout the gene. Likewise, individual sites (codons) under selection were identified by calculating synonymous and non-synonymous substitution rates per codon using SLAC, FEL, REL (Kosakovsky Pond and Frost, 2005), MEME (Murrell et al., 2012) and FUBAR methods (Murrell et al., 2013) in the Datamonkey online server (Delport et al., 2010). The McDonald-Kreitman (MK) test (McDonald and Kreitman, 1991) with Jukes-Cantor correction was also used for evaluating neutrality deviations by using the http://mkt.uab.es/mkt/MKT.asp web server (Egea et al., 2008).

Lineage-Specific Positive Selection

Lineages under episodic diversifying selection were assessed for each gene using the random effects likelihood (REL)-branch-site method (Kosakovsky Pond et al., 2011). Orthologous sequences from 13 *Plasmodium* species were aligned using the MUSCLE algorithm; this was then used for inferring the best evolutionary model using JModelTest (Posada, 2008). Phylogeny was then inferred using the Bayesian method (Ronquist et al., 2012) with a corresponding evolutionary model; these were used as reference when analyzing lineage-specific positive selection using the HyPhy package (Pond et al., 2005). CIPRES Science Gateway (Miller et al., 2010) web application was used for choosing the evolutionary model and Bayesian analysis.

Linkage Disequilibrium and Recombination

Linkage disequilibrium (LD) was assessed by using the Z_{nS} estimator (Kelly, 1997), followed by linear regression between LD and nucleotide distance to ascertain whether intra-gene recombination could have taken place regarding any gene. Recombination was also evaluated using the ZZ estimator (Rozas et al., 2001), the minimum number of recombination events (Rm) (Hudson and Kaplan, 1985) and the GARD algorithm (Kosakovsky Pond et al., 2006).

Predicting Proteins' Antigenic Potential

A previous study has shown a correlation between predicting potential B-cell epitopes and antigenic protein regions in a *P. vivax* antigen (Rodrigues-da-Silva et al., 2017). Potential linear B-cell epitopes were therefore predicted by using the immune epitope database (IEDB) server (Kolaskar and Tongaonkar, 1990) for each protein, using Sal-I reference sequences as query.

RESULTS

P. vivax Spz Loci *in Silico* Characterization

The currently available information for the Sal-I strain published in PlasmoDB database was used as the source for obtaining sequences from Spz genes orthologous to those identified as vaccine candidates in *P. falciparum* (**Table 1**). An ortholog search in other *Plasmodium* species showed that all genes (except for *siap2*, which was only present in species infecting primates) had an ortholog in the *Plasmodium* species analyzed here. The Sal-I protein sequence for each gene was then inferred from searching protein features within them (**Figure 1**). A positive secretion signal sequence was predicted for all antigens, except PvP36 (6-Cys protein family member), PvSIAP2 (sporozoite invasion-associated protein 2) and PvMCP1 (merozoite capping protein) proteins. In addition to signal peptide, a post-translational modification (consisting of a C-terminal GPI anchor sequence) was predicted for PvP52 (6-Cys protein family member) and PvTRSP (thrombospondin-related sporozoite protein), whilst several proteins seemed to have a transmembrane helix (**Figure 1** and **Table 1**); transmembrane helices were found at the N-terminal end in PvP36, PvPLP1 (perforin-like protein 1) and PvSIAP2. A signal peptide and transmembrane helix were predicted at the same position in PvPLP1. The transmembrane helices in PvP36, PvPLP1 and PvSIAP2 could thus have resulted from misidentification and they could actually have been signal peptide sequences.

PvTRSP, PvTRAP (thrombospondin-related anonymous protein) and PvTLP (thioredoxin-like protein) proteins, having a transmembrane region, also had a thrombospondin type 1 putative domain, whilst PvTRAP and PvTLP had a von Willebrand factor type A putative domain. A membrane attack complex/perforin (MACPF) putative domain was found in PvPLP1 and the AhpC/TSA domain in PvMCP1. A single sexual stage antigen s48/45 domain was predicted in PvP52 and PvP36. Both elongation factor Tu GTP (guanosine triphosphate)-binding domain and translation-initiation factor 2 were predicted for PvMB2 (**Figure 1** and **Table 1**).

Genetic Diversity at *P. vivax*-Sporozoite Loci

Sequences from different regions worldwide were analyzed for quantifying Spz antigen genetic diversity (**Table 2**); these parasite antigens have limited genetic diversity ($\pi < 0.003$). According to the diversity parameters estimated here, *pvtrap*, *pvsiap2* and *pvceltos* (cell traversal protein for ookinetes and sporozoites) genes had the highest nucleotide and protein diversity values ($\pi > 0.0015$; $\rho > 0.0039$); *pvspect1* (sporozoite protein essential for cell traversal), *pvplp1*, *pvspatr* (secreted protein with altered thrombospondin repeat domain), *pvtlp*, *pvsiap1*, and *pvmaebl* (merozoite adhesive erythrocytic binding protein) genes formed part of the most conserved genes/proteins ($\pi < 0.0009$; $\rho < 0.0020$) (**Table 2**). The haplotype number was low in most loci. However, *pvtrap*, *pvsiap2*, *pvplp1*, *pvmcp1*, *pvtlp*, and *pvmb2* were the Spz loci with the highest haplotype number.

Assessing Neutral Evolution in *P. vivax*-Sporozoite Loci

The ENC and CBI parameters were evaluated to assess whether codon bias had taken place in Spz genes. These parameters gave values higher than 47 for ENC and lower than 0.47 for CBI (**Table 3**). Tests based on polymorphism frequency spectrum were used with the 14 Spz *P. vivax* genes to assess any departure from neutral expectations. Six of these 14 genes had an overall negative statistically significant value for at least one test (**Table 3**), suggesting that natural selection could have been acting regarding these genes; *pvp36* and *pvmb2* genes had specific regions within each gene having a statistical significant negative value (Supplementary Material 1).

Regarding d_N-d_S rates, negative values were found for *pvmaebl*, *pvplp1*, and *pvmb2* whilst positive values was observed for the *pvsiap2* gene (**Table 3**). When species divergence was assessed, there was evidence of inter-specific negative selection for all genes, except *trsp*, *siap2*, and *celtos*; the sliding window for ω rate (Supplementary Material 2) gave values lower than 1 for these genes. Likewise, codon-based methods identified several codons under negative selection with few codons under positive selection. Many codons under negative selection were located in regions encoding putative domains (e.g., thrombospondin type 1 domain, Supplementary Material 2). It was found that *siap1* and *trap* loci had statistically significant values higher than 1 when polymorphism and divergence were compared by MK test, whilst a value lower than 1 was observed for *maebl*.

Lineage-Specific Positive Selection

Plasmodium species orthologous sequences were used for constructing phylogenies to assess whether positive natural selection had taken place during antigen evolutionary history (Supplementary Material 3). Topologies gave three monophyletic clusters (*siap2* did not, since it is only present in primate-infecting parasites); the first involved monkey-malaria parasites, the second clustered *Plasmodium* species infected rodents and the third was formed by hominid-malaria parasites. The latter cluster and the rodent clade represent *Plasmodium* phylogenetic relationships. However, the monkey-malaria

TABLE 1 | *In silico* characterization of 14 *P. vivax* sporozoite proteins.

Gene	Signal peptide (position)	Transm helix (position)	GPI	Domain (position)	Location
siap1	Yes (1–22)	No	No	–	Secretory
p52	Yes (1–18)	No	C-ter	Sexual stage antigen s48/45 domain (159–281)	Secretory/membrane
p36	No	Yes (10–32 and 37–59)	No	Sexual stage antigen s48/45 domain (204–332)	Secretory/membrane
spatr	Yes (1–20)	No	No	–	Secretory
trsp	Yes (1–18)	Yes (131–154)	C-ter	Thrombospondin type 1 domain (59–105)	Secretory/membrane
trap	Yes (1–24)	Yes (494–515)	No	von Willebrand factor type A domain (44–225) and Thrombospondin type 1 domain (241–284)	Secretory/membrane
spect	Yes (1–19)	No	No	–	Secretory
siap2	No	Yes (7–28)	No	–	Secretory
maebl	Yes (1–19)	Yes (1,799–1,816)	No	–	Secretory/membrane
plp1	Yes (1–23)	Yes (7–26)	No	Membrane attack complex/perforin (MACPF) (232–581)	Secretory
mcp1	No	No	No	AhpC/TSA family (9–134)	Non cytoplasmic
tlp	Yes (1–23)	Yes (1,426–1,445)	No	Thrombospondin type 1 domain (264–311) and von Willebrand factor type A domain (334–508)	Secretory/membrane
celtos	Yes (1–35)	No	No	–	Non cytoplasmic
mb2	Yes (1–22)	No	No	Elongation factor Tu GTP binding domain (766–931) and Translation-initiation factor 2 (1,151–1,263)	Secretory

The key structures for a potential vaccine candidate (such as a signal peptide, a transmembrane helix or glycosylphosphatidylinositol (GPI) anchor and functional domains) were described from the P. vivax Sal–I strain aa sequence. A protein's type of cell location is also indicated. The numbers in brackets indicate the position where a determined structure was predicted.

parasite topology had different branch patterns regarding malaria-species relationships which could have resulted from positive selection (Sawai et al., 2010). Nine of these topologies had lineages (branches) where some sites were under positive selection. Most branches having evidence of positive selection lead to a particular species, whilst few of them were ancestral lineages (Supplementary Material 3).

Linkage Disequilibrium and Recombination

Linkage disequilibrium (LD) at intra-gene level was assessed by Zns estimator (**Table 4**). Non-random associations between SNPs were found for *pvmaebl* but not for the remaining genes. However, there was a decreasing linear regression tendency between LD and nucleotide distance in all genes, suggesting recombination action. Likewise, most of these genes had at least one recombination event (RM), while the ZZ estimator just gave statistically significant values for *pvp52* and *pvspatr*. The GARD algorithm gave one recombination breakpoint for *pv52*, *pvtrap*, *pvplp1*, *pvtlp*, and *pvceltos* and two recombination breakpoints for *pvmb2* (**Table 4**).

P. vivax-Spz Proteins' Antigenic Potential

Previous studies have shown a B-epitope and solvent accessibility prediction correlation with antigenic regions in natural infections (Rodrigues-da-Silva et al., 2017). The Sal-I sequences for each Spz protein studied here were thus analyzed by the BepiPred server for determining their antigenic potential (Supplementary Material 4). Large-sized proteins, such as MAEBL, TLP, TRAP, P52, and MCP1, had regions toward the C-terminal which could be recognized as linear B-epitopes; these aa regions were the regions having greater solvent accessibility (Supplementary Material 4). It was seen that the most exposed PLP1 region was located toward the N-terminal. Smaller proteins, such as SPECT1, SPART, and SIAP1/2, did not seem to have clearly-defined recognition regions all along their sequences, whilst 3 regions having antigen potential were observed for CelTOS (Supplementary Material 4). Many P36, TRSP, PLP1, MB2 antigen sequences seemed to be exposed and several regions having potential B-epitopes were found.

DISCUSSION

A prospective *Plasmodium vivax*-malarial vaccine has been delayed regarding *P. falciparum*; however, knowledge acquired concerning the latter species could be useful for designing a fully-effective anti-*P. vivax* vaccine. Vaccine development involves several challenges; for instance, high antigen diversity has made vaccines not fully-protective since polymorphism provokes allele-specific immune responses; genetic diversity is therefore an immune avoidance mechanism. Consequently, conserved antigens (or regions within them) should be used as vaccines candidates to avoid this kind of response (Richie and Saul, 2002; Patarroyo et al., 2012). Even more, just one antigen might not be enough to produce full protection regarding a particular vaccine, so several antigens would be necessary. Since malaria parasites have multiple stages, a fully-effective vaccine must have several conserved antigens (or regions containing them) from different parasite stages.

Antigen identification is not an easy task due to *P. vivax* having a complex biology. Most antigens described to date

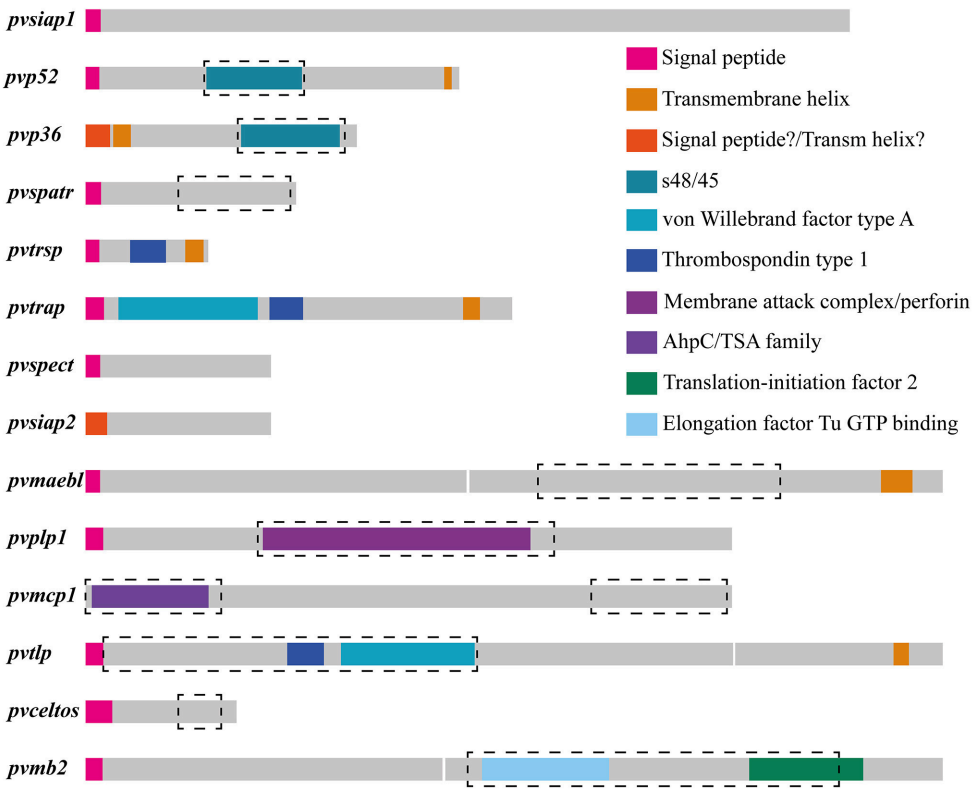


FIGURE 1 | Schematic models for *Plasmodium vivax* sporozoite antigen encoding genes and their predicted functionally constrained regions. The 14 putative *P. vivax* sporozoite antigens were characterized *in silico*. Each antigen's predicted structure is shown in different colors and the predicted regions under functional constraint (highly conserved between species, having a $\omega < 1$ and several negatively selected codons) are enclosed by a black dashed rectangle. In most cases, these regions correspond with regions having a putative domain. A region encoding a transmembrane helix was predicted for *pvp36* and *pvsiap2* at the 5′-end, but this could have been a misleading prediction (i.e., it could actually have been a signal peptide).

TABLE 2 | Estimating 14 *P. vivax* sporozoite genes' genetic diversity.

n	Gene	Sites	Ss	S	Ps	H	$\hat{k}^{DNA}$	θ_W (SD)	π^{DNA} (SD)	$\hat{d}^{AA}$	ρ^{AA}
15	*siap1*	2,991	12	8	4	8	2.55	0.0012 (0.0005)	0.0008 (0.0002)	2.00	0.0020
86	*p52*	1,461	9	2	7	13	1.98	0.0012 (0.0012)	0.0014 (0.0001)	1.49	0.0031
102	*p36*	1,059	6	3	3	9	1.13	0.0011 (0.0005)	0.0011 (0.0001)	0.67	0.0019
91	*spatr*	822	4	2	2	5	0.21	0.0009 (0.0005)	0.0003 (0.0001)	0.09	0.0003
63	*trsp*	480	2	0	2	3	0.59	0.0009 (0.0006)	0.0012 (0.0002)	0.59	0.0037
81	*trap*	1,668	28	9	19	20	5.11	0.0036 (0.0011)	0.0031 (0.0002)	3.91	0.0070
86	*spect1*	723	1	0	1	2	0.05	0.0003 (0.0003)	0.0001 (0.0000)	0.05	0.0002
79	*siap2*	1,239	25	8	17	25	2.09	0.0041 (0.0013)	0.0017 (0.0002)	1.99	0.0048
12	*maebl*	5,598	25	22	3	6	5.15	0.0015 (0.0006)	0.0009 (0.0005)	2.49	0.0013
73	*plp1*	2,529	13	5	8	14	0.77	0.0011 (0.0004)	0.0003 (0.0001)	0.40	0.0004
92	*mcp1*	1,455	13	4	9	19	1.44	0.0020 (0.0007)	0.0009 (0.0001)	0.84	0.0017
57	*tlp*	4,308	25	11	14	16	2.45	0.0013 (0.0004)	0.0006 (0.0001)	1.68	0.0012
101	*celtos*	588	6	3	3	10	0.89	0.0020 (0.0009)	0.0015 (0.0002)	0.76	0.0039
45	*mb2*	4,044	35	16	19	30	5.08	0.0020 (0.0006)	0.0012 (0.0001)	2.28	0.0017

The genetic diversity estimators were calculated from a variable amount of sequences for each antigen. n: amount of sequences analyzed. Sites, total of sites analyzed, excluding gaps; Ss, amount of segregating sites; S, amount of singleton sites; Ps, amount of informative-parsimonious sites; H, amount of haplotypes; $\hat{k}^{DNA}$, The average amount of nucleotide differences; θ_w, Watterson estimator; π^{DNA}, The average amount of nucleotide differences per site between two sequences; $\hat{d}^{AA}$, The average amount of aa differences; ρ^{AA}, The average amount of aa differences per site between two sequences; SD, standard deviation.

TABLE 3 | Neutrality and codon usage bias tests.

Gene	Tajima	Fu & Li		Fay & Wu	d_N - d_S (ES)	K_N - K_S (ES)	MKT	ENC	CBI
	D	D	F	H			NI		
siap1	−1.20	−1.71	−1.84	−0.44	0.000 (0.000)	−0.153 (0.010)††	13.6††	52.0	0.27
p52	0.27	1.29	0.94	−2.93*	0.000 (0.000)	−0.023 (0.003)**	2.87	53.6	0.22
p36	−0.36	−1.80	−1.79	0.44	−0.001 (0.002)	−0.023 (0.003)††	3.82	54.2	0.23
spatr	−1.42	−1.47	−1.71	−0.22	−0.001 (0.001)	−0.034 (0.005)††	1,56	49.6	0.39
trsp	0.64	0.72	0.81	0.46	0.002 (0.001)	−0.007 (0.005)	Null	50.7	0.45
trap	−0.48	−0.24	−0.36	−11.61**	0.002 (0.001)	−0.014 (0.004)††	6.46††	55.8	0.20
spect1	−0.91	0.50	0.09	0.04	0.000 (0.000)	−0.014 (0.003)††	Null	47.5	0.29
siap2	−1.80*	−0.75	−1.35	−2.07	0.002 (0.001)***	−0.003 (0.003)	Null†	47.5	0.36
maebl	−1.68	−1.20	−1.64	−10.21**	−0.001 (0.000)**	−0.066 (0.007)††	0.39†	49.1	0.30
plp1	−2.11*	−0.78	−1.46	−3.16*	−0.002 (0.001)*	−0.033 (0.003)*	2.09	52.3	0.28
mcp1	−1.19	−1.44	−1.64	0.57	−0.001 (0.001)	−0.045 (0.005)††	1.64	47.8	0.28
tlp	−1.76	−1.39	−1.95	−5.22**	−0.0001 (0.000)	−0.051 (0.004)††	1.81	52.7	0.20
celtos	−0.80	−0.82	−0.87	−1.49	0.001 (0.001)	−0.036 (0.0027)	6.36	53.8	0.35
mb2	−1.31	−1.67	−1.84	−3.41	−0.001 (0.001)*	−0.074 (0.004)††	1.26	53.4	0.25

*Selective pressure on each gene was inferred from neutrality tests and estimating effective codon usage. Non-synonymous substitution rate (d_N) and synonymous substitution rate (d_S) in P. vivax. Non-synonymous (K_N) and synonymous (K_S) divergence between P. vivax and the phylogenetically closest specie. Neutrality index (NI) estimated by McDonald-Kreitman test, using Jukes-Cantor correction. The preferential use of the synonymous codons was evaluated by estimating the effective amount of codons (ENC) and the codon bias index (CBI). Genes under selection in the tests had *$p < 0.05$, **$p < 0.01$, †$p < 0.03$, ***$p < 0.006$, and ††$p < 0.0001$.*

regarding this parasite have been from the blood stage; few genes/proteins from others stages have been characterized. Although *P. vivax* biology cannot be assessed directly, some new technologies enable making inferences about it. Whole genome sequences could be used to infer the genes in a particular parasite and thus gene ontology could provide clues about its biology. At least 15 *P. falciparum* Spz stage proteins are involved in parasite invasion and could thus become vaccine candidates. Since many genes are shared amongst *Plasmodium* species (Carlton et al., 2008), and several of them seem to be essential for parasite survival (Bushell et al., 2017), orthologs to the 15 *P. falciparum* Spz antigens might be present in the *P. vivax* genome and could thus be taken into account for vaccine development. Fourteen of the 15 *P. falciparum* Spz genes found in *P. vivax* had similar gene/protein structures, suggesting that they could be involved in a conserved *Plasmodium* Spz invasion pathway. However, gene/protein identification is the first step in vaccine design. As *P. vivax* genetic diversity represents an even greater challenge than in *P. falciparum* concerning vaccine design, the next step in this work was to assess the genetic diversity of these Spz loci using available whole genome sequences to find out which of them could be promising vaccine candidates against *P. vivax*.

P. vivax Spz Loci Genetic Diversity

It has been suggested that the *P. vivax* genome's large genetic diversity would hinder *P. vivax* control and elimination. Several *P. vivax* blood-stage antigens have been considered as vaccine candidates; however, they have high genetic diversity (Putaporntip et al., 1997, 2010; Gomez et al., 2006; Dias et al., 2011; Premaratne et al., 2011; Garzón-Ospina et al., 2012, 2014, 2015), representing one of the challenges to be overcome when designing a completely-effective anti-malarial vaccine. Contrasting with the aforementioned antigens, the Spz loci analyzed here had low genetic diversity ($\pi < 0.003$) which is a desirable feature when designing a fully-effective antimalarial vaccine. Few segregating sites were found at each locus; some were singleton sites. This could have resulted from genetic differentiation amongst *P. vivax* populations worldwide, as has been shown previously (Taylor et al., 2013; Hupalo et al., 2016). Nevertheless, fully-conserved regions were also identifiable in all loci. The observed polymorphism in Spz genes was comparable to that found in the most of the conserved Mrz genes described to date (Putaporntip et al., 2009; Garzón-Ospina et al., 2010, 2011, 2015; Pacheco et al., 2012; Chenet et al., 2013; Forero-Rodriguez et al., 2014a,b; Buitrago et al., 2016) being *pvspect*, *pvspart*, *pvsiap1*, *pvplp1*, and *pvtlp* the loci with the lowest diversity. The most polymorphic gene was *pvtrap*; its diversity pattern has previously been reported for the *P. falciparum* ortholog (Ohashi et al., 2014).

Evolutionary Forces Modulating *P. vivax* Sporozoite Diversity

Contrasting with Mrz proteins (Putaporntip et al., 1997, 2010; Gomez et al., 2006; Dias et al., 2011; Garzón-Ospina et al., 2012, 2014, 2015), Spz antigens evaluated here displayed low diversity. This could have resulted from various evolutionary forces. Taking into account that a low amount of synonymous and non-synonymous polymorphism was found, low diversity could have been a consequence of codon usage bias. ENC and CBI parameters were thus estimated for evaluating such hypothesis. ENC values close to 61 indicated that all synonymous codons for each aa were used equitably (values close to 0 suggest bias or preferential codon use). CBI values range from 0, meaning

TABLE 4 | LD and intra-gene recombination estimators.

Gene	Zns	ZZ	RM	GARD
siap1	0.12	−0.00	1	–
p52	0.08	−0.06*	2	369†
p36	0.01	−0.00	1	–
spatr	0.54	0.18*	0	–
trsp	0.04	0.00	0	–
trap	0.10	0.06	22	1,200†
spect1	–	–	–	–
siap2	0.03	−0.02	2	–
maebl	0.73*	0.16	0	–
plp1	0.03	−0.01	2	1,275**
mcp1	0.02	−0.02	2	–
tlp	0.08	0.12	2	2,551†
celtos	0.02	0.01	0	355†
mb2	0.04	0.04	7	1,368† and 2,605†

*GARD, recombination breakpoint position. *$p < 0.04$, **$p < 0.005$, †$p < 0.001$.*

uniform synonymous codon usage, to 1 (i.e., maximum codon bias) (Morton, 1993). Spz genes' ENC and CBI values suggested that all of them had random synonymous codon usage. Such values were similar to those previously described for genes participating in Mrz invasion and the value reported for the complete genome (Cornejo et al., 2014). This meant continuous transcription related to these proteins' level of expression during invasion (Gajbhiye et al., 2017; Uddin, 2017) which was not affected by any type of selective pressure or preferential codon usage. The low genetic diversity found in Spz genes was therefore not a consequence of selection at translation level and thus other evolutionary forces must be causing low genetic diversity in theses antigens.

Test of neutral molecular evolution (e.i. Tajima, Fu & Li, Fay & Wu) gave negative values, just a few being statistically significant; the neutrality could not thus be ruled out in antigens lacking significant values. Since these genes were highly conserved, a functional/structural constraint was likely (Kimura, 1983; Graur et al., 2013). Although some genes had no statistically significant values, some of them had specific regions where neutrality could be ruled out, suggesting that negative selection was acting in such regions. Likewise, genes having an overall statistically significant negative value in these tests suggested that negative selection was responsible for low protein diversity and consequently functional/structural constrains would also be expected under this kind of selection. This was confirmed when non-synonymous and synonymous rates, as well as evolutionary rate (ω) sliding windows, were computed. Synonymous mutations seemed to fix at a higher rate than non-synonymous ones after speciation involving monkey-malaria parasites; this thus agreed with a hypothesis of functional/structural constraint. Furthermore, several negative selected sites (codons) were found; many were located in putative functional domains, suggesting that negative selection is an important force for maintaining protein domain integrity. These regions had statistically significant negative values in the tests based on polymorphism frequency spectrum (e.i. Tajima, Fu & Li, Fay & Wu). These patterns could thus be used to predict functionally important regions. A recent report regarding *P. vivax* showed that regions having low ω rates having several negatively selected sites are regions used by the parasite to recognize host cells (Baquero et al., 2017). Consequently, the regions from the 14 antigens considered here having low ω rates and several codons under negative selection could be those used to recognize hepatocytes. However, these regions might not necessarily be involved in host-parasite interaction and they could have other functions. However, they could be taken into account for *P. vivax* vaccine development since they had low diversity and were under functional/structural constraint. The aforementioned results could then also be used for elucidating aspects of *P. vivax* sporozoite invasion of target cells by assessing the domains having functional constraint in invasion assays.

Few positively selected sites were found; they could have been fixed to adapt to different selective pressures. Molecular adaptations could play an important role during parasite evolutionary history. Monkey-malaria clades diversify three to four times more rapidly than those infecting other mammalians (Muehlenbein et al., 2015). Taking host species' rapid diversification into account (Ziegler et al., 2007), adaptive radiation in monkey-malaria parasites could explain this accelerated cladogenesis and therefore several molecular adaptations could have arisen during such radiation. Phylogenies inferred for Spz antigens showed that monkey-malaria parasite topology did not agree with malaria-species relationships which could have resulted from positive selection (Sawai et al., 2010). There was evidence of selective sweep in five *P. vivax* genes; this pattern has already been observed for some of these genes in previous studies (Shen et al., 2017). Some mutations fixed in this parasite would thus allow it to adapt to a new host after host-switch decreasing genetic diversity.

Additionally, phylogenetically-based analysis could provide greater insight into the role of selection (at these loci) during a parasite's evolutionary history if this was the result of adaptations to new hosts and/or environments (Muehlenbein et al., 2015). There was evidence of episodic positive selection for nine of these 14 antigens. Throughout phylogenies, several lineages had some sites under positive selection; however, few of them were ancestral lineages. Since most branches under selection lead to particular species, episodic selection could have resulted from adaptation to different host, for instance, to avoid immune response to a particular host during sympatric speciation or to recognize a new host receptor. However, this behavior seems to be common in the *Plasmodium* genus and not just for monkey-malaria parasites.

Recombination is an evolutionary force which can increase genetic diversity. It could be acting on Mrz blood-stage antigens leading to new haplotypes to arise, being maintained in the parasite population to evade the host's immune responses (Garzón-Ospina et al., 2012). Even though low genetic diversity was observed, some Spz genes had a large amount of haplotypes. Nevertheless, larger-sized genes accumulated a greater amount

of single nucleotide polymorphisms in their sequences; these characteristics were not related to the amount of haplotypes. Some haplotypes could thus have arisen by recombination. Evidence of intra-gene recombination was found for *pvp52*, *pvplp1*, *pvtlp*, *pvceltos*, and *pvmb2* genes. Recombination is thus an evolutionary force in these loci increasing genetic diversity.

P. vivax Spz Antigens' Putative Roles

The mechanism by which these Spz proteins act in *P. vivax* is not certain; however, their possible role could be elucidated from them having putative domains. SPECT, PLP1, and TLP seem to be essential in different *Plasmodium* species for traversing host cells (Ishino et al., 2004; Yang et al., 2017a) since deleting them has reduced parasite capability to traverse the sinusoidal barrier and thus gain access to hepatocytes (Ishino et al., 2004; Kaiser et al., 2004). No described domains were identified for PvSPECT and PvSPART and their role thus requires further investigation. Nevertheless, orthologs to these proteins have been implicated in parasite interaction with epithelial or hepatic host cells in *P. falciparum* (Curtidor et al., 2011). Even though putative protein-protein interaction domains were not identified, the regions involved in such interaction can be predicted regarding their degree of conservation between species, as can their evolutionary rates (ω) (Graur et al., 2013; Baquero et al., 2017). According to the sliding window for the gene encoding to SPART in *P. vivax*, the C-terminal region seemed to be functionally restricted as low genetic diversity (intra- e inter-species) was observed, as well as lower than 1 ω rate and various sites under negative selection (at inter-specific level); this region could be implicated in interaction with hepatocytes. Additionally, PfSPART is antigenic in natural infections and antibodies against it have blocked interaction with hepatocytes *in vitro* (Palaeya et al., 2013). Similar to PfSPATR, PvSPATR has antigenic potential given the 4 regions along its sequence having high solvent accessibility values and potential linear B-epitopes, suggesting that PvSPART is a promising antigen when designing an anti-*P. vivax* vaccine.

Similar to the aforementioned antigens, particular domains were not found in the whole PvSIAP1 sequence. Orthologs for this protein in *Plasmodium* spp, are involved in Spz exit from oocysts, as well as their colonization in a mosquito's salivary glands (Engelmann et al., 2009). SIAP1 in a vertebrate host seems to be mediated by pathogen-host interaction (Curtidor et al., 2011). This protein's ortholog in *P. vivax* has been shown to have low diversity ($\omega < 1$) throughout its sequence, as well as several negatively-selected sites. However, putative functional regions have not been clearly defined due to a large amount of sites under positive selection all along the sequence of the gene encoding this protein. Nevertheless, the C-terminal region was the region where most codons were found to be under negative selection and could thus be the region used for *P. vivax* interaction with a particular host.

PLP1 is not required for entry to hepatocytes, but does play an important role in exit from transitory vacuoles during cell traversal (Risco-Castillo et al., 2015). A putative MAC/perforin domain was identified in *P. vivax* which had a high sequence conservation, several codons under negative selection and a $\omega < 1$; this functionally restricted region could thus be mediating membrane destabilization and pore formation in *P. vivax*, as has been suggested in other species (Rosado et al., 2008; Patarroyo et al., 2016; Yang et al., 2017a).

It is well known that the TLP (Moreira et al., 2008) and TRAP (Sultan et al., 1997) are essential for Spz gliding motility (Sultan et al., 1997). *P. falciparum* TLP is the most conserved member of the TRAP/MIC2 family and the first to be seen to play a role traversing cells (Moreira et al., 2008). *P. vivax* TLP was one of the most conserved proteins of those evaluated here; the N-terminal region was highly conserved within and amongst species. Thrombospondin type 1 (TSP1) and von Willebrand factor type A (vWa) domains were observed in this region. The *P. falciparum*, TSP1 domain mediates glycosaminoglycan binding whilst the vWa domain is involved in cell-cell, cell-matrix, matrix-matrix interactions and includes a metal-ion dependent adhesion site. Proteins containing such domains might be associated with parasite invasion ability (Wengelnik et al., 1999; Matuschewski et al., 2002; Mongui et al., 2010); the PvTLP N-terminal region could thus be mediating *P. vivax* Spz invasion of hepatocytes.

Unlike a PvTLP, PvTRAP (having the same domains toward the N-terminal region) was the protein having the greatest diversity of those evaluated here. This gene is highly polymorphic in *P. falciparum* and is under positive selection (Ohashi et al., 2014); the region encoding the vWA domain is where *pftrap* diversity is concentrated (Moreira et al., 2008). Such pattern seems to be similar in *P. vivax* and its related species; the sliding window for this gene showed that this region evolved rapidly, having several sites under positive selection amongst species. This could have resulted from lineage-specific adaptations for the recognition of particular receptors for each host or is the target region for immune responses, similar to that which occurs in PvDBP (VanBuskirk et al., 2004; Chootong et al., 2014). A prediction of B-epitopes and solvent accessibility showed that TRAP should have a potentially antigenic region toward the N-terminal where the vWa domain would be located. However, the C-terminal region seemed to have greater antigenic potential. According to previous studies PvTRAP can induce an IgG1 and IgG3 response following natural infection; producing this type of antibody is usually correlated with protection against disease in a hyper-endemic region (Nazeri et al., 2017). This protein has thus been proposed as being an important immune response target regarding pre-erythrocyte stages of malaria caused by both *P. falciparum* and *P. vivax* (Ohashi et al., 2014). Nevertheless, the diversity observed in this protein could produce allele-specific immune responses and would thus not be a good vaccine candidate.

Unlike TRAP and TLP, PvTRSP only has the TSP1 domain, this protein is involved in *P. berghei* invasion where the deletion of *pbtrsp* reduces mutant parasite capability to enter hepatocytes (Labaied et al., 2007a). This gene was highly conserved in *P. vivax*; however, when the sequences from *P. vivax* and related species were compared, the TSP1 domain was seen to have a $\omega > 1$ and few sites under negative selection. On the other hand, the C-terminal region seemed to be this protein's antigenic region.

MAEBL has been described as being a type I membrane protein in *P. falciparum*, having erythrocyte-binding activity

and seeming to fulfill a function similar to that of AMA1, given the high sequence similarity (Yang et al., 2017b). This protein is necessary for Spz invasion of a mosquito's salivary glands and for them to traverse vertebrate host cells (Yang et al., 2017b). MAEBL has 4 tandem repeats in *P. vivax*, located toward the gene's 3′ region, between positions 3,427 and 4,756. Each repeat has 90% similarity; 9 imperfect copies of GCTAGAAGGGCTGAGGAGT residues, 3 copies of AGAAAGGCGGAAGAGGCA, 17 copies of GCAAGGAAGGCAGAGGATGCTAGAAAGGCAG AGGCGGCTA and 6 copies of GCTAAAAAGGCTGAA GCAGCAAGGAAGGCAGAGGCA residues were found. In spite of an accumulation of repeats being responsible for size polymorphism, they did not show such polymorphism when sequences from different isolates were analyzed. Calculating the omega rate between *P. vivax* isolates and related species revealed that such repeats were highly conserved in *Plasmodium*, as has been suggested previously (Leite et al., 2015). The fact that this repeat region was found to be under negative selection suggested an important role, similar to that already reported for CSP (Aldrich et al., 2012; Ferguson et al., 2014). A prediction of linear epitopes suggested that the C-terminal region where these repeats were located could be antigenic. Even though the repeats have been suggested as targets distracting an immune response, this is still not clear because they are highly conserved, even between species. Conversely, the *maebl* 5′-end was highly divergent amongst species, having several sites under positive selection, possibly resulting from species-specific adaptations.

MB2 (just like MAEBL) is one of the largest proteins expressed in *Plasmodium* spp Spz and is involved in hepatocyte invasion (Nguyen et al., 2009). It contains a characteristic GTP-binding domain (Nguyen et al., 2001; Romero et al., 2004), that is also present in its *P. vivax* counterpart; in addition, a translation-initiation factor domain was predicted for PvMB2. Both were functionally restricted, being conserved in *P. vivax* as well as between phylogenetically-related species. Even though it is not clear whether these domains are mediated by pathogen-host interaction, they do seem to be important for this protein's function. This protein is antigenic in *P. falciparum* and has been recognized by patients showing protection against *Plasmodium* infection following experimental immunization (Nguyen et al., 2009). The region so recognized is the N-terminal region, which had antigenic potential in *P. vivax*, given its solvent accessibility and the prediction of B-linear epitopes. Given low PvMB2 diversity and its antigenic potential, it could also be taken into account when designing a vaccine against *P. vivax*.

Members of the 6-cys family are expressed during different *Plasmodium* spp. stages. Two members of this family (P36 and P52) are expressed during the pre-erythrocyte stage and seem to play an important role in invasion (Labaied et al., 2007b; Annoura et al., 2014; Kublin et al., 2017; Manzoni et al., 2017). Like 6-Cys members expressed in *P. vivax* Mrz (Forero-Rodriguez et al., 2014a,b), *pvp36* and *pvp52* displayed low genetic diversity, accompanied by low evolutionary rates. The s48/45 domains, characteristic of this family, seemed to be functionally restricted (limited diversity, $\omega < 1$ and various sites under negative selection) and thus might have been responsible for the interaction between *P. vivax* Spz and hepatocytes. The proteins encoded by these genes also have antigenic potential; the PvP52 C-terminal region seemed to be exposed, having potential linear B-epitopes, whilst the PvP38 central region and C-terminal could be antigenic regions. Thus, the same as other 6-Cys family members (Forero-Rodriguez et al., 2014a,b), PvP36 and PvP52 would seem to be promising antigens when designing a vaccine.

In addition to TRAP, SIAP2, and CELTOS were the antigens having the highest diversity values amongst those evaluated here. These proteins are predominantly expressed during the Spz stage in other species where they cover parasite surface (Siau et al., 2008). Unlike the other proteins evaluated here, SIAP2 only had orthologs in *Plasmodium* species infecting primates. This is a potential vaccine candidate as it seems to interact specifically with heparin sulfate and chondroitin sulfate-type membrane receptors (Siau et al., 2008; Alba et al., 2011). Potential functional regions could not be defined for SIAP2 (like TRAP) since it had many sites under positive selection amongst species throughout its sequence. Predicted results regarding PvSIAP2 antigenicity suggested that this could be exposed to the immune system, having various potential B-epitopes along its sequence. Its use as vaccine candidate is limited due to its diversity and the large amount of haplotypes found.

CelTOS has been considered a potential vaccine candidate given its association with clinical protection regarding naturally-acquired immunity (Kanoi et al., 2017). This protein has a sole specificity for phosphatidic acid (a lipid found predominantly on plasma membrane inner face) and breaks liposomes consisting of phosphatidic acid through pore formation. It has been shown that parasites lacking CelTOS can enter target cells but remain trapped inside. The foregoing supposes that CelTOS targets cell membrane inner leaflet which can burst due to pore formation inside infected cells and favors parasite exit (Kariu et al., 2006; Jimah et al., 2016). Even though its function seems to be intracellular, predictive analysis suggested that PvCelTOS could be exposed to the immune system, having potential B-epitopes toward the C-terminal region.

A recent study has shown a correlation between the prediction and this protein's antigenic regions. Naturally-infected patients generate immune responses against PvCelTOS, predominantly toward the protein's C-terminal region (Rodrigues-da-Silva et al., 2017), coinciding with the most exposed regions and having potential B-epitopes. This would suggest that PvCelTOS is a potential antigen for inclusion in a vaccine against *P. vivax*. The antigenic region coincided with the gene's region where various sites under negative selection were found, suggesting that the CelTOS functional region could be located in the C-terminal region; however, some sites under positive inter-species selection were found toward this region. If this region is functionally restricted, but is also the region toward which the immune response is directed in different species, then some sites could have been positively selected in a species-specific manner fixing specific mutations in each species, resulting in a positive selection signal. In addition to its antigenic potential, PvCelTOS has been shown to have limited diversity within *P. vivax*. The gene encoding this protein in Iran had, on average, less than one mutation in paired comparisons (Mehrizi et al., 2017); such results were similar to those reported here. PvCelTOS

might thus be considered a vaccine candidate, given its antigenic characteristics and limited diversity.

CONCLUSIONS

Designing a completely effective vaccine against *P. vivax* must include antigens or highly conserved regions containing them to abolish allele-specific immune responses; antigens from different stages must also be included. As most research has been focused on the blood stage, new strategies must be implemented for identifying potential vaccine candidates from other parasite stages. As genetic diversity is a characteristic to be born in mind when designing a vaccine, the present work could be taken as a basis for selecting Spz antigens which might become potential vaccine candidates.

P. vivax genetic diversity would be expected to be high as it has been shown in whole genome analysis (Neafsey et al., 2012) and in several Mrz antigens (Gomez et al., 2006; Putaporntip et al., 2010; Dias et al., 2011; Premaratne et al., 2011; Garzón-Ospina et al., 2012). However, this study's results showed that some Spz antigens had limited genetic diversity and, consequently, they could be good vaccine candidates, since low genetic diversity is ideal for avoiding an allele-specific immune response. The analysis described above enabled determining which regions in these proteins could be functionally important. Recent studies have shown that regions predicted to have functional restriction coincide with regions used by the parasite to bind to a host cell (Baquero et al., 2017). Therefore, regions in Spz antigens which were conserved between species having low ω values and codons under negative selection in the parts of a protein might be thus implicated in these antigens' function. This could help to elucidate aspects of *P. vivax* Spz invasion. Further assays regarding interaction with hepatocytes are thus required for determining whether such regions are involved in pathogen-host interaction.

Promising antigens to be taken into account for designing a fully effective vaccine are those having a limited genetic diversity or at least one domain with such pattern. These genes or domains must have a negative selection signal, as well as a $\omega < 1$ (Garzón-Ospina et al., 2015). Thus, the results reported here suggest that the PvP52, PvP36, PvSPATR, PvPLP1, PvMCP1, PvTLP, PvCelTOS, and PvMB2 antigens (or functionally restricted regions of them) are promising vaccine candidates and thus should be prioritized in future studies aimed at developing a completely effective vaccine against *P. vivax*.

AUTHOR CONTRIBUTIONS

DG-O, SB, and AR acquired the genomic data and carried out the genetic diversity analyses. DG-O performed the general interpretation of the data and together with SB and AR participated in writing the paper. MP coordinated the study and helped to write the manuscript. All the authors have read and approved the final version of the manuscript.

FUNDING

This work was financed by the Departamento Administrativo de Ciencia, Tecnología e Innovación (COLCIENCIAS), through grant RC # 0309-2013.

ACKNOWLEDGMENTS

We would like to thank Jason Garry for translating and reviewing the manuscript and also thank the team in charge of updating the PlasmoDB database for providing the genetic information from which this work was carried out.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2018.00010/full#supplementary-material

REFERENCES

Abascal, F., Zardoya, R., and Telford, M. J. (2010). TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. *Nucleic Acids Res.* 38, W7–W13. doi: 10.1093/nar/gkq291

Adams, J. H., and Mueller, I. (2017). The biology of *plasmodium vivax*. *Cold Spring Harb. Perspect. Med.* 7, 1–13. doi: 10.1101/cshperspect.a025585

Alba, M. P., Almonacid, H., Calderón, D., Chacón, E. A., Poloche, L. A., Patarroyo, M. A., et al. (2011). 3D structure and immunogenicity of *Plasmodium falciparum* sporozoite induced associated protein peptides as components of fully-protective anti-malarial vaccine. *Biochem. Biophys. Res. Commun.* 416, 349–355. doi: 10.1016/j.bbrc.2011.11.039

Aldrich, C., Magini, A., Emiliani, C., Dottorini, T., Bistoni, F., Crisanti, A., et al. (2012). Roles of the amino terminal region and repeat region of the *Plasmodium berghei* circumsporozoite protein in parasite infectivity. *PLoS ONE* 7:e32524. doi: 10.1371/journal.pone.0032524

Annoura, T., van Schaijk, B. C., Ploemen, I. H., Sajid, M., Lin, J. W., Vos, M. W., et al. (2014). Two Plasmodium 6-Cys family-related proteins have distinct and critical roles in liver-stage development. *FASEB J.* 28, 2158–2170. doi: 10.1096/fj.13-241570

Baquero, L. A., Moreno-Pérez, D. A., Garzón-Ospina, D., Forero-Rodríguez, J., Ortiz-Suárez, H. D., and Patarroyo, M. A. (2017). PvGAMA reticulocyte binding activity: predicting conserved functional regions by natural selection analysis. *Parasit. Vectors* 10:251. doi: 10.1186/s13071-017-2183-8

Barry, A. E., and Arnott, A. (2014). Strategies for designing and monitoring malaria vaccines targeting diverse antigens. *Front. Immunol.* 5:359. doi: 10.3389/fimmu.2014.00359

Buitrago, S. P., Garzón-Ospina, D., and Patarroyo, M. A. (2016). Size polymorphism and low sequence diversity in the locus encoding the *Plasmodium vivax* rhoptry neck protein 4 (PvRON4) in Colombian isolates. *Malar. J.* 15:501. doi: 10.1186/s12936-016-1563-4

Bushell, E., Gomes, A. R., Sanderson, T., Anar, B., Girling, G., Herd, C., et al. (2017). Functional profiling of a plasmodium genome reveals an abundance of essential genes. *Cell* 170, 260–272 e268. doi: 10.1016/j.cell.2017.06.030

Carlton, J. M., Adams, J. H., Silva, J. C., Bidwell, S. L., Lorenzi, H., Caler, E., et al. (2008). Comparative genomics of the neglected human malaria parasite *Plasmodium vivax*. *Nature* 455, 757–763. doi: 10.1038/nature07327

Casillas, S., and Barbadilla, A. (2017). Molecular population genetics. *Genetics* 205, 1003–1035. doi: 10.1534/genetics.116.196493

Castellanos, A., Arévalo-Herrera, M., Restrepo, N., Gulloso, L., Corradin, G., and Herrera, S. (2007). *Plasmodium vivax* thrombospondin related adhesion protein: immunogenicity and protective efficacy in rodents and *Aotus monkeys*. *Mem. Inst. Oswaldo Cruz* 102, 411–416. doi: 10.1590/S0074-02762007005000047

Chan, E. R., Menard, D., David, P. H., Ratsimbasoa, A., Kim, S., Chim, P., et al. (2012). Whole genome sequencing of field isolates provides robust characterization of genetic diversity in *Plasmodium vivax*. *PLoS Negl. Trop. Dis.* 6:e1811. doi: 10.1371/journal.pntd.0001811

Chaurio, R. A., Pacheco, M. A., Cornejo, O. E., Durrego, E., Stanley, C. E. Jr., Castillo, A. I., et al. (2016). Evolution of the transmission-blocking vaccine candidates Pvs28 and Pvs25 in *Plasmodium vivax*: geographic differentiation and evidence of positive selection. *PLoS Negl. Trop. Dis.* 10:e0004786. doi: 10.1371/journal.pntd.0004786

Chenet, S. M., Pacheco, M. A., Bacon, D. J., Collins, W. E., Barnwell, J. W., and Escalante, A. A. (2013). The evolution and diversity of a low complexity vaccine candidate, merozoite surface protein 9 (MSP-9), in *Plasmodium vivax* and closely related species. *Infect. Genet. Evol.* 20, 239–248. doi: 10.1016/j.meegid.2013.09.011

Chootong, P., McHenry, A. M., Ntumngia, F. B., Sattabongkot, J., and Adams, J. H. (2014). The association of Duffy binding protein region II polymorphisms and its antigenicity in *Plasmodium vivax* isolates from Thailand. *Parasitol. Int.* 63, 858–864. doi: 10.1016/j.parint.2014.07.014

Coatney, G.R., and National Institute of Allergy and Infectious Diseases (U.S.) (1971). *The Primate Malarias*. Bethesda, MD: U.S. National Institute of Allergy and Infectious Diseases; for sale by the Supt. of Docs., U S Government Printing off Washington.

Cornejo, O. E., Fisher, D., and Escalante, A. A. (2014). Genome-wide patterns of genetic polymorphism and signatures of selection in *Plasmodium vivax*. *Genome Biol. Evol.* 7, 106–119. doi: 10.1093/gbe/evu267

Curtidor, H., Vanegas, M., Alba, M. P., and Patarroyo, M. E. (2011). Functional, immunological and three-dimensional analysis of chemically synthesised sporozoite peptides as components of a fully-effective antimalarial vaccine. *Curr. Med. Chem.* 18, 4470–4502. doi: 10.2174/092986711797287575

Delport, W., Poon, A. F., Frost, S. D., and Kosakovsky Pond, S. L. (2010). Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology. *Bioinformatics* 26, 2455–2457. doi: 10.1093/bioinformatics/btq429

Dias, S., Longacre, S., Escalante, A. A., and Udagama-Randeniya, P. V. (2011). Genetic diversity and recombination at the C-terminal fragment of the merozoite surface protein-1 of *Plasmodium vivax* (PvMSP-1) in Sri Lanka. *Infect. Genet. Evol.* 11, 145–156. doi: 10.1016/j.meegid.2010.09.007

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32, 1792–1797. doi: 10.1093/nar/gkh340

Egea, R., Casillas, S., and Barbadilla, A. (2008). Standard and generalized McDonald-Kreitman test: a website to detect selection by comparing different classes of DNA sites. *Nucleic Acids Res.* 36, W157–W162. doi: 10.1093/nar/gkn337

Engelmann, S., Silvie, O., and Matuschewski, K. (2009). Disruption of Plasmodium sporozoite transmission by depletion of sporozoite invasion-associated protein 1. *Eukaryotic Cell* 8, 640–648. doi: 10.1128/EC.00347-08

Fankhauser, N., and Mäser, P. (2005). Identification of GPI anchor attachment signals by a Kohonen self-organizing map. *Bioinformatics* 21, 1846–1852. doi: 10.1093/bioinformatics/bti299

Fay, J. C., and Wu, C. I. (2000). Hitchhiking under positive Darwinian selection. *Genetics* 155, 1405–1413.

Ferguson, D. J., Balaban, A. E., Patzewitz, E. M., Wall, R. J., Hopp, C. S., Poulin, B., et al. (2014). The repeat region of the circumsporozoite protein is critical for sporozoite formation and maturation in Plasmodium. *PLoS ONE* 9:e113923. doi: 10.1371/journal.pone.0113923

Forero-Rodríguez, J., Garzón-Ospina, D., and Patarroyo, M. A. (2014a). Low genetic diversity and functional constraint in loci encoding *Plasmodium vivax* P12 and P38 proteins in the Colombian population. *Malar. J.* 13:58. doi: 10.1186/1475-2875-13-58

Forero-Rodríguez, J., Garzon-Ospina, D., and Patarroyo, M. A. (2014b). Low genetic diversity in the locus encoding the *Plasmodium vivax* P41 protein in Colombia's parasite population. *Malar. J.* 13:388. doi: 10.1186/1475-2875-13-388

Frevert, U. (2004). Sneaking in through the back entrance: the biology of malaria liver stages. *Trends Parasitol.* 20, 417–424. doi: 10.1016/j.pt.2004.07.007

Fu, Y. X., and Li, W. H. (1993). Statistical tests of neutrality of mutations. *Genetics* 133, 693–709.

Gajbhiye, S., Patra, P. K., and Yadav, M. K. (2017). New insights into the factors affecting synonymous codon usage in human infecting Plasmodium species. *Acta Trop.* 176, 29–33. doi: 10.1016/j.actatropica.2017.07.025

Garzón-Ospina, D., Forero-Rodríguez, J., and Patarroyo, M. A. (2014). Heterogeneous genetic diversity pattern in *Plasmodium vivax* genes encoding merozoite surface proteins (MSP)-7E, -7F and -7L.. *Malar. J.* 13:495. doi: 10.1186/1475-2875-13-495

Garzón-Ospina, D., Forero-Rodríguez, J., and Patarroyo, M. A. (2015). Inferring natural selection signals in *Plasmodium vivax*-encoded proteins having a potential role in merozoite invasion. *Infect. Genet. Evol.* 33, 182–188. doi: 10.1016/j.meegid.2015.05.001

Garzón-Ospina, D., López, C., Forero-Rodríguez, J., and Patarroyo, M. A. (2012). Genetic diversity and selection in three *Plasmodium vivax* merozoite surface protein 7 (Pvmsp-7) genes in a Colombian population. *PLoS ONE* 7:e45962. doi: 10.1371/journal.pone.0045962

Garzón-Ospina, D., Romero-Murillo, L., and Patarroyo, M. A. (2010). Limited genetic polymorphism of the *Plasmodium vivax* low molecular weight rhoptry protein complex in the Colombian population. *Infect. Genet. Evol.* 10, 261–267. doi: 10.1016/j.meegid.2009.12.004

Garzon-Ospina, D., Romero-Murillo, L., Tóbon, L. F., and Patarroyo, M. A. (2011). Low genetic polymorphism of merozoite surface proteins 7 and 10 in Colombian *Plasmodium vivax* isolates. *Infect. Genet. Evol.* 11, 528–531. doi: 10.1016/j.meegid.2010.12.002

Gomez, A., Suarez, C. F., Martinez, P., Saravia, C., and Patarroyo, M. A. (2006). High polymorphism in *Plasmodium vivax* merozoite surface protein-5 (MSP5). *Parasitology* 133(Pt 6), 661–672. doi: 10.1017/S0031182006001168

Graur, D., Zheng, Y., Price, N., Azevedo, R. B., Zufall, R. A., and Elhaik, E. (2013). On the immortality of television sets: "function" in the human genome according to the evolution-free gospel of ENCODE. *Genome Biol. Evol.* 5, 578–590. doi: 10.1093/gbe/evt028

Hester, J., Chan, E. R., Menard, D., Mercereau-Puijalon, O., Barnwell, J., Zimmerman, P. A., et al. (2013). De novo assembly of a field isolate genome reveals novel *Plasmodium vivax* erythrocyte invasion genes. *PLoS Negl. Trop. Dis.* 7:e2569. doi: 10.1371/journal.pntd.0002569

Hudson, R. R., and Kaplan, N. L. (1985). Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics* 111, 147–164.

Hulden, L. (2011). Activation of the hypnozoite: a part of *Plasmodium vivax* life cycle and survival. *Malar. J.* 10:90. doi: 10.1186/1475-2875-10-90

Hupalo, D. N., Luo, Z., Melnikov, A., Sutton, P. L., Rogov, P., Escalante, A., et al. (2016). Population genomics studies identify signatures of global dispersal and drug resistance in *Plasmodium vivax*. *Nat. Genet.* 48, 953–958. doi: 10.1038/ng.3588

Ishino, T., Yano, K., Chinzei, Y., and Yuda, M. (2004). Cell-passage activity is required for the malarial parasite to cross the liver sinusoidal cell layer. *PLoS Biol.* 2:e4. doi: 10.1371/journal.pbio.0020004

Jimah, J. R., Salinas, N. D., Sala-Rabanal, M., Jones, N. G., Sibley, L. D., Nichols, C. G., et al. (2016). Malaria parasite CelTOS targets the inner leaflet of cell membranes for pore-dependent disruption. *Elife* 5:e20621. doi: 10.7554/eLife.20621

Jukes, T. H., and Cantor, C. R. (1969). "Evolution of protein molecules," in *Mammalian Protein Metabolism*, ed H. N. Munro (New York, NY: Academic Press), 21–132.

Kall, L., Krogh, A., and Sonnhammer, E. L. (2007). Advantages of combined transmembrane topology and signal peptide prediction–the Phobius web server. *Nucleic Acids Res.* 35, W429–W432. doi: 10.1093/nar/gkm256

Kaiser, K., Camargo, N., Coppens, I., Morrisey, J. M., Vaidya, A. B., and Kappe, S. H. I. (2004). A member of a conserved *Plasmodium* protein family with membrane-attack complex/perforin (MACPF)-like domains localizes to the micronemes of sporozoites. *Mol. Biochem. Parasitol.* 133, 15–26. doi: 10.1016/j.molbiopara.2003.08.009

Kanoi, B. N., Takashima, E., Morita, M., White, M. T., Palacpac, N. M., Ntege, E. H., et al. (2017). Antibody profiles to wheat germ cell-free system synthesized

Plasmodium falciparum proteins correlate with protection from symptomatic malaria in Uganda. *Vaccine* 35, 873–881. doi: 10.1016/j.vaccine.2017.01.001

Kariu, T., Ishino, T., Yano, K., Chinzei, Y., and Yuda, M. (2006). CelTOS, a novel malarial protein that mediates transmission to mosquito and vertebrate hosts. *Mol. Microbiol.* 59, 1369–1379. doi: 10.1111/j.1365-2958.2005.05024.x

Kelly, J. K. (1997). A test of neutrality based on interlocus associations. *Genetics* 146, 1197–1206.

Kimura, M. (1983). *The Neutral Theory of Molecular Evolution.* Cambridge; New York: Cambridge University Press. doi: 10.1017/CBO9780511623486

Kolaskar, A. S., and Tongaonkar, P. C. (1990). A semi-empirical method for prediction of antigenic determinants on protein antigens. *FEBS Lett.* 276, 172–174. doi: 10.1016/0014-5793(90)80535-Q

Kosakovsky Pond, S. L., and Frost, S. D. (2005). Not so different after all: a comparison of methods for detecting amino acid sites under selection. *Mol. Biol. Evol.* 22, 1208–1222. doi: 10.1093/molbev/msi105

Kosakovsky Pond, S. L., Murrell, B., Fourment, M., Frost, S. D., Delport, W., and Scheffler, K. (2011). A random effects branch-site model for detecting episodic diversifying selection. *Mol. Biol. Evol.* 28, 3033–3043. doi: 10.1093/molbev/msr125

Kosakovsky Pond, S. L., Posada, D., Gravenor, M. B., Woelk, C. H., and Frost, S. D. (2006). Automated phylogenetic detection of recombination using a genetic algorithm. *Mol. Biol. Evol.* 23, 1891–1901. doi: 10.1093/molbev/msl051

Kublin, J. G., Mikolajczak, S. A., Sack, B. K., Fishbaugher, M. E., Seilie, A., Shelton, L., et al. (2017). Complete attenuation of genetically engineered *Plasmodium falciparum* sporozoites in human subjects. *Sci. Transl. Med.* 9:eaad9099. doi: 10.1126/scitranslmed.aad9099

Labaied, M., Camargo, N., and Kappe, S. H. (2007a). Depletion of the Plasmodium berghei thrombospondin-related sporozoite protein reveals a role in host cell entry by sporozoites. *Mol. Biochem. Parasitol.* 153, 158–166. doi: 10.1016/j.molbiopara.2007.03.001

Labaied, M., Harupa, A., Dumpit, R. F., Coppens, I., Mikolajczak, S. A., and Kappe, S. H. (2007b). Plasmodium yoelii sporozoites with simultaneous deletion of P52 and P36 are completely attenuated and confer sterile immunity against infection. *Infect. Immun.* 75, 3758–3768. doi: 10.1128/IAI.00225-07

Leite, J. A., Bargieri, D. Y., Carvalho, B. O., Albrecht, L., Lopes, S. C., Kayano, A. C., et al. (2015). Immunization with the MAEBL M2 domain protects against lethal plasmodium yoelii infection. *Infect. Immun.* 83, 3781–3792. doi: 10.1128/IAI.00262-15

Librado, P., and Rozas, J. (2009). DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* 25, 1451–1452. doi: 10.1093/bioinformatics/btp187

Manzoni, G., Marinach, C., Topçu, S., Briquet, S., Grand, M., Tolle, M., et al. (2017). Plasmodium P36 determines host cell receptor usage during sporozoite invasion. *Elife* 6. doi: 10.7554/eLife.25903

Matuschewski, K., Nunes, A. C., Nussenzweig, V., and Ménard, R. (2002). Plasmodium sporozoite invasion into insect and mammalian cells is directed by the same dual binding system. *EMBO J.* 21, 1597–1606. doi: 10.1093/emboj/21.7.1597

McDonald, J. H., and Kreitman, M. (1991). Adaptive protein evolution at the Adh locus in Drosophila. *Nature* 351, 652–654. doi: 10.1038/351652a0

Mehrizi, A. A., Torabi, F., Zakeri, S., and Djadid, N. D. (2017). Limited genetic diversity in the global *Plasmodium vivax* Cell traversal protein of Ookinetes and Sporozoites (CelTOS) sequences; implications for PvCelTOS-based vaccine development. *Infect. Genet. Evol.* 53, 239–247. doi: 10.1016/j.meegid.2017.06.005

Ménard, R. (2001). Gliding motility and cell invasion by Apicomplexa: insights from the Plasmodium sporozoite. *Cell. Microbiol.* 3, 63–73. doi: 10.1046/j.1462-5822.2001.00097.x

Miller, M. A., Pfeiffer, W., and Schwartz, T. (2010). "Creating the CIPRES science gateway for inference of large phylogenetic trees," in *Gateway Computing Environments Workshop (GCE)* (New Orleans, LA: IEEE), 1–8.

Mongui, A., Angel, D. I., Moreno-Perez, D. A., Villarreal-Gonzalez, S., Almonacid, H., Vanegas, M., et al. (2010). Identification and characterization of the *Plasmodium vivax* thrombospondin-related apical merozoite protein. *Malar. J.* 9:283. doi: 10.1186/1475-2875-9-283

Moreira, C. K., Templeton, T. J., Lavazec, C., Hayward, R. E., Hobbs, C. V., Kroeze, H., et al. (2008). The plasmodium TRAP/MIC2 family member, TRAP-Like Protein (TLP), is involved in tissue traversal by sporozoites. *Cell. Microbiol.* 10, 1505–1516. doi: 10.1111/j.1462-5822.2008.01143.x

Morton, B. R. (1993). Chloroplast DNA codon use: evidence for selection at the psb A locus based on tRNA availability. *J. Mol. Evol.* 37, 273–280. doi: 10.1007/BF00175504

Muehlenbein, M. P., Pacheco, M. A., Taylor, J. E., Prall, S. P., Ambu, L., Nathan, S., et al. (2015). Accelerated diversification of nonhuman primate malarias in Southeast Asia: adaptive radiation or geographic speciation? *Mol. Biol. Evol.* 32, 422–439. doi: 10.1093/molbev/msu310

Mueller, I., Shakri, A. R., and Chitnis, C. E. (2015). Development of vaccines for *Plasmodium vivax* malaria. *Vaccine* 33, 7489–7495. doi: 10.1016/j.vaccine.2015.09.060

Murrell, B., Moola, S., Mabona, A., Weighill, T., Sheward, D., Kosakovsky Pond, S. L., et al. (2013). FUBAR: a fast, unconstrained bayesian approximation for inferring selection. *Mol. Biol. Evol.* 30, 1196–1205. doi: 10.1093/molbev/mst030

Murrell, B., Wertheim, J. O., Moola, S., Weighill, T., Scheffler, K., and Kosakovsky Pond, S. L. (2012). Detecting individual sites subject to episodic diversifying selection. *PLoS Genet.* 8:e1002764. doi: 10.1371/journal.pgen.1002764

Nazeri, S., Zakeri, S., Mehrizi, A. A., and Djadid, N. D. (2017). Naturally acquired immune responses to thrombospondin-related adhesion protein (TRAP) of *Plasmodium vivax* in patients from areas of unstable malaria transmission. *Acta Trop.* 173, 45–54. doi: 10.1016/j.actatropica.2017.05.026

Neafsey, D. E., Galinsky, K., Jiang, R. H., Young, L., Sykes, S. M., Saif, S., et al. (2012). The malaria parasite *Plasmodium vivax* exhibits greater genetic diversity than *Plasmodium falciparum*. *Nat. Genet.* 44, 1046–1050. doi: 10.1038/ng.2373

Nguyen, T. V., Fujioka, H., Kang, A. S., Rogers, W. O., Fidock, D. A., and James, A. A. (2001). Stage-dependent localization of a novel gene product of the malaria parasite, *Plasmodium falciparum*. *J. Biol. Chem.* 276, 26724–26731. doi: 10.1074/jbc.M103375200

Nguyen, T. V., Sacci, J. B. Jr., de la Vega, P., John, C. C., James, A. A., and Kang, A. S. (2009). Characterization of immunoglobulin G antibodies to *Plasmodium falciparum* sporozoite surface antigen MB2 in malaria exposed individuals. *Malar. J.* 8:235. doi: 10.1186/1475-2875-8-235

Nielsen, H. (2017). Predicting secretory proteins with signalP. *Methods Mol. Biol.* 1611, 59–73. doi: 10.1007/978-1-4939-7015-5_6

Novembre, J. A. (2002). Accounting for background nucleotide composition when measuring codon usage bias. *Mol. Biol. Evol.* 19, 1390–1394. doi: 10.1093/oxfordjournals.molbev.a004201

Ohashi, J., Suzuki, Y., Naka, I., Hananantachai, H., and Patarapotikul, J. (2014). Diversifying selection on the thrombospondin-related adhesive protein (TRAP) gene of *Plasmodium falciparum* in Thailand. *PLoS ONE* 9:e90522. doi: 10.1371/journal.pone.0090522

Pacheco, M. A., Elango, A. P., Rahman, A. A., Fisher, D., Collins, W. E., Barnwell, J. W., et al. (2012). Evidence of purifying selection on merozoite surface protein 8 (MSP8) and 10 (MSP10) in Plasmodium spp. *Infect. Genet. Evol.* 12, 978–986. doi: 10.1016/j.meegid.2012.02.009

Palaeya, V., Lau, Y. L., Mahmud, R., Chen, Y., and Fong, M. Y. (2013). Cloning, expression, and immunocharacterization of surface protein containing an altered thrombospondin repeat domain (SPATR) from *Plasmodium knowlesi*. *Malar. J.* 12:182. doi: 10.1186/1475-2875-12-182

Patarroyo, M. A., Calderón, D., and Moreno-Pérez, D. A. (2012). Vaccines against *Plasmodium vivax*: a research challenge. *Expert Rev. Vaccines* 11, 1249–1260. doi: 10.1586/erv.12.91

Patarroyo, M. E., Arévalo-Pinzón, G., Reyes, C., Moreno-Vranich, A., and Patarroyo, M. A. (2016). Malaria parasite survival depends on conserved binding peptides' critical biological functions. *Curr. Issues Mol. Biol.* 18, 57–78. doi: 10.21775/cimb.018.057

Pierleoni, A., Martelli, P. L., Fariselli, P., and Casadio, R. (2006). BaCelLo: a balanced subcellular localization predictor. *Bioinformatics* 22, e408–e416. doi: 10.1093/bioinformatics/btl222

Pond, S. L., Frost, S. D., and Muse, S. V. (2005). HyPhy: hypothesis testing using phylogenies. *Bioinformatics* 21, 676–679. doi: 10.1093/bioinformatics/bti079

Posada, D. (2008). jModelTest: phylogenetic model averaging. *Mol. Biol. Evol.* 25, 1253–1256. doi: 10.1093/molbev/msn083

Premaratne, P. H., Aravinda, B. R., Escalante, A. A., and Udagama, P. V. (2011). Genetic diversity of *Plasmodium vivax* Duffy Binding Protein II (PvDBPII) under unstable transmission and low intensity malaria in Sri Lanka. *Infect. Genet. Evol.* 11, 1327–1339. doi: 10.1016/j.meegid.2011.04.023

Price, R. N., Douglas, N. M., and Anstey, N. M. (2009). New developments in *Plasmodium vivax* malaria: severe disease and the rise of chloroquine resistance. *Curr. Opin. Infect. Dis.* 22, 430–435. doi: 10.1097/QCO.0b013e32832f14c1

Putaporntip, C., Jongwutiwes, S., Ferreira, M. U., Kanbara, H., Udomsangpetch, R., and Cui, L. (2009). Limited global diversity of the *Plasmodium vivax* merozoite surface protein 4 gene. *Infect. Genet. Evol.* 9, 821–826. doi: 10.1016/j.meegid.2009.04.017

Putaporntip, C., Jongwutiwes, S., Tanabe, K., and Thaithong, S. (1997). Interallelic recombination in the merozoite surface protein 1 (MSP-1) gene of *Plasmodium vivax* from Thai isolates. *Mol. Biochem. Parasitol.* 84, 49–56. doi: 10.1016/S0166-6851(96)02786-7

Putaporntip, C., Udomsangpetch, R., Pattanawong, U., Cui, L., and Jongwutiwes, S. (2010). Genetic diversity of the *Plasmodium vivax* merozoite surface protein-5 locus from diverse geographic origins. *Gene* 456, 24–35. doi: 10.1016/j.gene.2010.02.007

Rich, S. M., and Ayala, F. J. (2003). Progress in malaria research: the case for phylogenetics. *Adv. Parasitol.* 54, 255–280. doi: 10.1016/S0065-308X(03)54005-2

Richie, T. L., and Saul, A. (2002). Progress and challenges for malaria vaccines. *Nature* 415, 694–701. doi: 10.1038/415694a

Risco-Castillo, V., Topçu, S., Marinach, C., Manzoni, G., Bigorgne, A. E., Briquet, S., et al. (2015). Malaria Sporozoites Traverse Host Cells within Transient Vacuoles. *Cell Host Microbe* 18, 593–603. doi: 10.1016/j.chom.2015.10.006

Rodrigues-da-Silva, R. N., Soares, I. F., Lopez-Camacho, C., Martins da Silva, J. H., Perce-da-Silva, D. S., Têva, A., et al. (2017). *Plasmodium vivax* cell-traversal protein for ookinetes and sporozoites: naturally acquired humoral immune response and B-cell epitope mapping in brazilian amazon inhabitants. *Front. Immunol.* 8:77. doi: 10.3389/fimmu.2017.00077

Romero, L. C., Nguyen, T. V., Deville, B., Ogunjumo, O., and James, A. A. (2004). The MB2 gene family of Plasmodium species has a unique combination of S1 and GTP-binding domains. *BMC Bioinformatics* 5:83. doi: 10.1186/1471-2105-5-83

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D. L., Darling, A., Höhna, S., et al. (2012). MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* 61, 539–542. doi: 10.1093/sysbio/sys029

Rosado, C. J., Kondos, S., Bull, T. E., Kuiper, M. J., Law, R. H., Buckle, A. M., et al. (2008). The MACPF/CDC family of pore-forming toxins. *Cell. Microbiol.* 10, 1765–1774. doi: 10.1111/j.1462-5822.2008.01191.x

Rozas, J., Gullaud, M., Blandin, G., and Aguadé, M. (2001). DNA variation at the rp49 gene region of *Drosophila simulans*: evolutionary inferences from an unusual haplotype structure. *Genetics* 158, 1147–1155.

Sawai, H., Otani, H., Arisue, N., Palacpac, N., de Oliveira Martins, L., Pathirana, S., et al. (2010). Lineage-specific positive selection at the merozoite surface protein 1 (msp1) locus of *Plasmodium vivax* and related simian malaria parasites. *BMC Evol. Biol.* 10:52. doi: 10.1186/1471-2148-10-52

Shen, H. M., Chen, S. B., Wang, Y., Xu, B., Abe, E. M., and Chen, J. H. (2017). Genome-wide scans for the identification of *Plasmodium vivax* genes under positive selection. *Malar. J.* 16:238. doi: 10.1186/s12936-017-1882-0

Siau, A., Silvie, O., Franetich, J. F., Yalaoui, S., Marinach, C., Hannoun, L., et al. (2008). Temperature shift and host cell contact up-regulate sporozoite expression of *Plasmodium falciparum* genes involved in hepatocyte infection. *PLoS Pathog.* 4:e1000121. doi: 10.1371/journal.ppat.1000121

Sonnhammer, E. L., von Heijne, G., and Krogh, A. (1998). A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 6, 175–182.

Sultan, A. A., Thathy, V., Frevert, U., Robson, K. J., Crisanti, A., Nussenzweig, V., et al. (1997). TRAP is necessary for gliding motility and infectivity of plasmodium sporozoites. *Cell* 90, 511–522. doi: 10.1016/S0092-8674(00)80511-5

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123, 585–595.

Tamura, K., Stecher, G., Peterson, D., Filipski, A., and Kumar, S. (2013). MEGA6: molecular evolutionary genetics analysis version 6.0. *Mol. Biol. Evol.* 30, 2725–2729. doi: 10.1093/molbev/mst197

Taylor, J. E., Pacheco, M. A., Bacon, D. J., Beg, M. A., Machado, R. L., Fairhurst, R. M., et al. (2013). The evolutionary history of *Plasmodium vivax* as inferred from mitochondrial genomes: parasite genetic diversity in the Americas. *Mol. Biol. Evol.* 30, 2050–2064. doi: 10.1093/molbev/mst104

Uddin, A. (2017). Codon usage bias: a tool for understanding molecular evolution. *J. Proteomics Bioinform.* 10:e32. doi: 10.4172/jpb.1000e32

VanBuskirk, K. M., Sevova, E., and Adams, J. H. (2004). Conserved residues in the *Plasmodium vivax* Duffy-binding protein ligand domain are critical for erythrocyte receptor recognition. *Proc. Natl. Acad. Sci. U.S.A.* 101, 15754–15759. doi: 10.1073/pnas.0405421101

Weiss, G. E., Gilson, P. R., Taechalertpaisarn, T., Tham, W. H., de Jong, N. W., Harvey, K. L., et al. (2015). Revealing the sequence and resulting cellular morphology of receptor-ligand interactions during *Plasmodium falciparum* invasion of erythrocytes. *PLoS Pathog.* 11:e1004670. doi: 10.1371/journal.ppat.1004670

Wengelnik, K., Spaccapelo, R., Naitza, S., Robson, K. J., Janse, C. J., Bistoni, F., et al. (1999). The A-domain and the thrombospondin-related motif of *Plasmodium falciparum* TRAP are implicated in the invasion process of mosquito salivary glands. *EMBO J.* 18, 5195–5204. doi: 10.1093/emboj/18.19.5195

Winter, D. J., Pacheco, M. A., Vallejo, A. F., Schwartz, R. S., Arevalo-Herrera, M., Herrera, S., et al. (2015). Whole genome sequencing of field isolates reveals extensive genetic diversity in *Plasmodium vivax* from Colombia. *PLoS Negl. Trop. Dis.* 9:e0004252. doi: 10.1371/journal.pntd.0004252

Yang, A. S. P., O'Neill, M. T., Jennison, C., Lopaticki, S., Allison, C. C., Armistead, J. S., et al. (2017a). Cell traversal activity is important for *Plasmodium falciparum* liver infection in humanized mice. *Cell Rep.* 18, 3105–3116. doi: 10.1016/j.celrep.2017.03.017

Yang, A. S. P., Lopaticki, S., O'Neill, M. T., Erickson, S. M., Douglas, D. N., Kneteman, N. M., et al. (2017b). AMA1 and MAEBL are important for *Plasmodium falciparum* sporozoite infection of the liver. *Cell. Microbiol.* 19:e12745. doi: 10.1111/cmi.12745

Zhang, J., Rosenberg, H. F., and Nei, M. (1998). Positive Darwinian selection after gene duplication in primate ribonuclease genes. *Proc. Natl. Acad. Sci. U.S.A.* 95, 3708–3713. doi: 10.1073/pnas.95.7.3708

Ziegler, T., Abegg, C., Meijaard, E., Perwitasari-Farajallah, D., Walter, L., Hodges, J. K., et al. (2007). Molecular phylogeny and evolutionary history of Southeast Asian macaques forming the *M. silenus* group. *Mol. Phylogenet. Evol.* 42, 807–816. doi: 10.1016/j.ympev.2006.11.015

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 10 September 2018
doi: 10.3389/fgene.2018.00372

On the Evolution and Function of *Plasmodium vivax* Reticulocyte Binding Surface Antigen (*pvrbsa*)

*Paola Andrea Camargo-Ayala[1,2], Diego Garzón-Ospina[1,3], Darwin Andrés Moreno-Pérez[1,4], Laura Alejandra Ricaurte-Contreras[1], Oscar Noya[5] and Manuel A. Patarroyo[1,6]**

[1] Department of Molecular Biology and Immunology, Fundación Instituto de Inmunología de Colombia (FIDIC), Bogotá, Colombia, [2] Microbiology Postgraduate Programme, Universidad Nacional de Colombia, Bogotá, Colombia, [3] PhD Programme in Biomedical and Biological Sciences, Universidad del Rosario, Bogotá, Colombia, [4] Livestock Sciences Faculty, Universidad de Ciencias Aplicadas y Ambientales, Bogotá, Colombia, [5] Instituto de Medicina Tropical, Facultad de Medicina, Universidad Central de Venezuela, Caracas, Venezuela, [6] School of Medicine and Health Sciences, Universidad del Rosario, Bogotá, Colombia

OPEN ACCESS

Edited by:
Miguel Arenas,
University of Vigo, Spain

Reviewed by:
Eduardo Castro-Nallar,
Universidad Andrés Bello, Chile
Eugenia Lo,
University of North Carolina at Charlotte, United States
S. Noushin Emami,
Stockholm University, Sweden

***Correspondence:**
Manuel A. Patarroyo
mapatarr.fidic@gmail.com

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: 25 May 2018
Accepted: 23 August 2018
Published: 10 September 2018

Citation:
Camargo-Ayala PA, Garzón-Ospina D, Moreno-Pérez DA, Ricaurte-Contreras LA, Noya O and Patarroyo MA (2018) On the Evolution and Function of Plasmodium vivax Reticulocyte Binding Surface Antigen (pvrbsa). Front. Genet. 9:372. doi: 10.3389/fgene.2018.00372

The RBSA protein is encoded by a gene described in *Plasmodium* species having tropism for reticulocytes. Since this protein is antigenic in natural infections and can bind to target cells, it has been proposed as a potential candidate for an anti-*Plasmodium vivax* vaccine. However, genetic diversity (a challenge which must be overcome for ensuring fully effective vaccine design) has not been described at this locus. Likewise, the minimum regions mediating specific parasite-host interaction have not been determined. This is why the *rbsa* gene's evolutionary history is being here described, as well as the *P. vivax rbsa* (*pvrbsa*) genetic diversity and the specific regions mediating parasite adhesion to reticulocytes. Unlike what has previously been reported, *rbsa* was also present in several parasite species belonging to the monkey-malaria clade; paralogs were also found in *Plasmodium* parasites invading reticulocytes. The *pvrbsa* locus had less diversity than other merozoite surface proteins where natural selection and recombination were the main evolutionary forces involved in causing the observed polymorphism. The N-terminal end (PvRBSA-A) was conserved and under functional constraint; consequently, it was expressed as recombinant protein for binding assays. This protein fragment bound to reticulocytes whilst the C-terminus, included in recombinant PvRBSA-B (which was not under functional constraint), did not. Interestingly, two PvRBSA-A-derived peptides were able to inhibit protein binding to reticulocytes. Specific conserved and functionally important peptides within PvRBSA-A could thus be considered when designing a fully-effective vaccine against *P. vivax*.

Keywords: *Plasmodium vivax*, *rbsa*, genetic diversity, evolutionary forces, protein biding, parasite–host interaction, antimalarial vaccine

Abbreviations: ω, omega rate (d_N/d_S and/or K_N/K_S); BY, Bayesian method; Chr, chromosome; LD, linkage disequilibrium; LRR, leucine-rich region; ML, maximum likelihood; Mrz, merozoite; RBSA, reticulocyte binding surface antigen; rRBSA, recombinant RBSA protein; WB, Western blot.

INTRODUCTION

Plasmodium vivax is a parasite which emerged in Asia (Escalante et al., 2005; Carlton et al., 2013) (although an African origin is also likely, Liu et al., 2014), involving a host-switch from monkeys to humans (Mu et al., 2005) around 1.3–2.9 million years ago (Pacheco et al., 2011), *Plasmodium cynomolgi* being the most phylogenetically related species (Mu et al., 2005; Pacheco et al., 2011; Tachibana et al., 2012). *P. vivax* then reached countries on the 5 continents through human migration (Rodrigues et al., 2018), currently predominating in Asia and America (Guerra et al., 2010; Gething et al., 2012). This species is considered the second most important human-malaria parasite worldwide, due to the morbidity it causes (WHO, 2017).

Plasmodium vivax incidence has decreased since 2010. An estimated 6–11 million cases were attributed to this parasite in 2016 (around 7 million less than in 2010) (WHO, 2017). Despite that, the social and economic burden of malaria in endemic countries (Suh et al., 2004) could still be huge. Even though control measures against this parasite have been shown to be useful, *P. vivax* elimination (and/or malarial elimination) is not an easy task. The most relevant challenges for eliminating malaria concern the social and economic conditions of the places most affected by this illness, the social and political conflicts in several endemic areas and the anomalous weather patterns (WHO, 2017). These, together with the spread of insecticide-resistant mosquitoes and drug-resistant parasites, could bring about a recurrence of this disease (Maxmen, 2012; Price et al., 2014; Huijben and Paaijmans, 2018). Consequently, new interventions must be designed which can reduce the parasite reservoir, limiting the time that a human (or mosquito) host is infectious (Barry and Arnott, 2014). This goal could be achieved by developing a fully effective antimalarial vaccine which, together with existing control measures, could contribute towards a malaria-free world (Barry and Arnott, 2014; White et al., 2014).

Malaria-related vaccine research and development efforts were exclusively focused on *Plasmodium falciparum* for a long time (Galinski and Barnwell, 2008). During the last few decades, *P. vivax* research has been increased and more and more groups are studying approaches to an anti-*P. vivax* vaccine (Galinski and Barnwell, 2008; Chen et al., 2010; Valencia et al., 2011; The malERA Consultative Group on Vaccines, 2011; Arnott et al., 2012; Patarroyo et al., 2012; Nanda Kumar et al., 2016). Initial research was based on the knowledge acquired regarding *P. falciparum* (Patarroyo et al., 2012); however, given *P. vivax*'s unique attributes (such as hypnozoite forms in the liver, invasion preference for reticulocytes or the rapid gametocyte formation) (Galinski et al., 2013) and since *P. falciparum* and *P. vivax* have had different evolutionary paths, research regarding *P. vivax*-exclusive antigens could be relevant. For instance, the Duffy antigen, which is not present in *P. falciparum*, is essential for *P. vivax* invasion (Chitnis and Miller, 1994). PvDBP-conserved epitopes can trigger strong neutralizing antibodies; it has therefore been suggested as one of the main *P. vivax* vaccine candidates (Chen et al., 2016; Ntumngia et al., 2017).

One antigen shared just by *P. vivax* and *P. cynomolgi* (both having reticulocyte predilection) has been characterized recently (Moreno-Perez et al., 2017). This antigen (named reticulocyte binding surface antigen – RBSA) is encoded by a two-exon gene (Moreno-Perez et al., 2017) located on Chr 3. The protein product has a signal peptide and two transmembrane helices and is located on mature Mrz membrane. This protein seems to be involved in Mrz invasion, since recombinant *P. vivax* RBSA (rPvRBSA) binds to a subpopulation of immature reticulocytes having a Duffy positive phenotype; it is also recognized by the human immune system in natural infections. It has thus been suggested as a putative vaccine candidate for anti-*P. vivax* malaria vaccine development (Moreno-Perez et al., 2017).

However, since parasite's high genetic diversity is one of the challenges to be overcome during *P. vivax* vaccine design (Neafsey et al., 2012); the antigens' genetic diversity must therefore be assessed (Arnott et al., 2012; Barry and Arnott, 2014) for selecting those having limited polymorphism or the conserved functional regions within them (Richie and Saul, 2002). *pvrbsa* genetic diversity has not been assessed to date and, although PvRBSA binds to reticulocytes (Moreno-Perez et al., 2017), the regions involved in Mrz-host specific interactions have not been established yet.

PvRBSA is an Mrz membrane protein and is recognized by the human immune system in natural infection; it could therefore be a highly polymorphic antigen. Nevertheless, functionally important parts of the protein (i.e., regions involved in host-parasite interaction) should be functionally constrained, being maintained by negative selection as highly conserved within and between species (Graur et al., 2013; Garzon-Ospina et al., 2015). Consequently, inferring this kind of selection on *pvrbsa* could be used to predict regions under functional constraint which then might be used as putative vaccine candidates. This study evaluated *pvrbsa* genetic diversity, assessed the evolutionary forces involved in causing the observed polymorphism and established minimum regions involved in protein–cell interaction.

MATERIALS AND METHODS

Ethics Approval and Consent to Participate

All *P. vivax*-infected patients who provided us with the blood samples were informed about the purpose of the study and all gave their written consent. Regarding newborn umbilical cord blood samples, the progenitors signed an informed consent form after having received detailed information regarding the study's goals. All procedures carried out in this study were approved by the ethics committees of the Fundación Instituto de Inmunología de Colombia (IRB number: ACTA N° 037-CEEPA), the Universidad del Rosario (IRB number: CEI-ABN026-0001061) and the Bioethics' committee from the Instituto de Medicina Tropical "Dr. Félix Pifano" at the Universidad Central de Venezuela (IRB number: CEC-IMT11/2018).

Parasite DNA

One hundred and sixty-seven peripheral blood samples and/or blood drops collected on FTA cards from patients proving

positive for *P. vivax* malaria by microscope examination were collected in different Colombian (Chocó, n = 37, Córdoba, n = 39, and Amazonas, n = 41) and Venezuelan (Bolivar, n = 29 and Venezuela's coastal area, n = 19) endemic areas between 2010 and 2016 (**Supplementary Figure S1**). A Wizard Genomic DNA Purification kit (Promega) was used for obtaining DNA from Chocó and Córdoba samples, following the manufacturer's instructions. A Pure Link Genomic DNA mini kit (Invitrogen) was used for extracting DNA from Amazon samples from each drop of blood collected on the FTA cards, according to the manufacturer's specifications whilst blood were extracted from Venezuelan samples by salting-out, following Welsh and Bunce modifications (Welsh and Bunce, 1999). The recovered DNAs were stored at −20°C until use. All samples had a single *P. vivax msp3* allele, suggesting that these samples came from single *P. vivax* infection.

Identifying the *rbsa* Gene in *Plasmodium* Monkey Lineage

A Blast search (using the 1,506 bp from Sal-I-*P. vivax rbsa* gene sequence as query) was carried out using available whole genome sequences from *P. cynomolgi* (GenBank accession number: GCA_000321355.1), *P. inui* (GenBank accession number: GCA_000524495.1), *P. knowlesi* (GenBank accession number: GCA_000006355.1), *P. coatneyi* (GenBank accession number: GCA_000725905.1) and *P. fragile* (GenBank accession number: GCA_000956335.1) to identify orthologous in these *Plasmodium* species. The MUSCLE method (Edgar, 2004) was used for aligning the recovering contigs having high identity sequences (>60). The best model for DNA substitutions was then selected using the JModelTest v.2.1.3 (Posada, 2008) with the Akaike's information criterion and MEGA v.6 software (Tamura et al., 2013). Phylogenetic trees were inferred through ML and Bayesian (BY) methods, using the GTR and/or the TVM models (selected as the best models). MEGA v.6 software was used for ML analysis and topology reliability was evaluated by bootstrap, using 1,000 iterations. Bayesian phylogenetic analysis was conducted with MrBayes v.3.2 software (Ronquist et al., 2012) and the analysis was run until reaching a lower than 0.01 standard deviation of split frequencies value; sump and sumt commands were then used for tabulating posterior probabilities and building a consensus tree.

PCR Amplification and Sequencing

One hundred and sixty-seven parasite DNAs from endemic Colombian (n = 117) and Venezuelan (n = 50) regions were used for amplifying the *pvrbsa* locus by nested-PCR, as follows. The first reaction involved 3.3 μL of ultrapure water, 5.4 μL of KAPA HiFi HotStart Readymix, 0.3 μM of each primer (*rbsafwd*: 5′-TTTATTTCATTTTGACGTTGTAACTTG-3′ and *rbsarev*: 5′-TTAAGAAATGATCCCAACTCG-3′) and 1 μL of DNA. The PCR was carried out as follows: one 5 min step at 95°C, a second 35 cycle step for 20 s at 98°C, 15 s at 57°C and 45 s at 72°C, followed by a final step of 10 min at 72°C. Two μL of the first PCR product were added to the second PCR reaction containing 7.5 μL of ultrapure water, 12.5 μL of KAPA HiFi HotStart Readymix and 0.3 μM of each primer (*rbsa2fwd*: 5′-GAAATACAAGATGAAAGGAATAATG-3′ and *rbsa2rev*: 5′-GATCCCAACTCGGTTTATC-3′). The thermal conditions were the same as those used in the first PCR reaction. PvRBSA fragments towards the amino and carboxyl (PvRBSA-A and PvRBSA-B, respectively) encoding regions were amplified using KAPA-HiFi HotStart Readymix (KAPA Biosystems) and the genomic DNA previously extracted from the Vivax Colombia Guaviare-I (VCG-I) strain (Moreno-Perez et al., 2014). The 25 μL PCR reaction contained 7.5 μL ultrapure water, 12.5 μL enzyme, 0.3 μM primer (designed taking into account the natural selection signatures observed for *pvrbsa*, PvRBSA-A: *rbsa-a-fwd*: 5′-GGGGTACCACAGCAAGTAGTGAGTCTCT-3′ and *rbsa-a-rev*: 5′-CCCTCGAGCTCACATTCTCCACCAC TTAA-3′; PvRBSA-B: *rbsa-b-fwd*: 5′-GGGGTACCCA TATAGAAGTAGGATCCGAA-3′ and *rbsa-b*-rev: 5′-CCCT CGAGCAATTGTTCTTCTCCGTATATAT-3′) and 50 ng DNA as template. Temperature cycling conditions involved 1 step of 3 min at 95°C, followed by 35 cycles of 20 sec at 98°C, 15 s at 60°C and 15 s at 72°C, and a final step of 30 s at 72°C. All amplicons were then purified by low-melt agarose gel using the Wizard SV Gel and PCR Clean-Up System (Promega) and then sequenced with a BigDye Terminator kit (Macrogen, Seoul, South Korea) in both directions using the second PCR primers and an internal primer (*rbsaintseq*: 5′-TTTATATTTACACTATTCCTTTGG-3′). Singleton SNPs were confirmed by an independent PCR amplification. The sequences obtained here were deposited in the GenBank database (accession numbers MH391806 - MH391972).

Obtaining Recombinant Plasmids With RBSA Fragments

pvrbsa-A and *pvrbsa-B* PCR products were digested with KpnI (New England Biolabs) and AvaI (NEB) enzymes and then ligated into the pET32b+ vector using T4 ligase (NEB). Briefly, 0.5 μg of each purified product, 1x of CutSmart buffer and 1 U/μL of each enzyme were used in a 25 μL reaction which was incubated for 1 h at 37°C and then inactivated at 80°C for 20 min. Ligation to pET32b+ vector was performed in a 20 μL volume containing 1x T4 buffer, 30 U/μL T4 ligase and vector/product in a 1:3 ratio. The reaction was incubated at 16°C for 16 h and then inactivated at 65°C for 20 min. Each recombinant plasmid was transformed into *Escherichia coli* JM109 cells (Invitrogen) according to the manufacturer's recommendations and recombinant colonies were then confirmed by PCR using the primers from each product. Three positive colonies were used for extracting plasmids with an UltraClean 6 Minute Mini Plasmid Prep kit (MOBIO), following the manufacturer's recommendations and then sequenced bidirectionally using *pet32b-fwd*: 5′-CGGTGAAGTGGCGGCAA-3′ and *pet32-Rev*: 5′-CCAAGGGGTTATGCTAGT-3′ primers.

pvrbsa Gene DNA Diversity and Evolutionary Analysis

Three chromatograms (forward, reverse, and internal primer) were obtained from Colombian and Venezuelan samples from

sequencing; they were assembled using CLC Main workbench v.3 software (CLC bio, Cambridge, MA, United States). Colombian and Venezuelan sequences were compared and aligned [using the MUSCLE method (Edgar, 2004)] with reference strain sequences (Sal-I GenBank accession number: AAKM01000020.1, Brazil-I GenBank accession number: AFMK01000195.1/AFMK01000194.1, India-VII GenBank accession number: AFBK01001271.1 and North Korean GenBank accession number: AFNJ01000313.1) (Carlton et al., 2008; Neafsey et al., 2012) and with sequences obtained from several sequencing projects (Chan et al., 2012; Hester et al., 2013; Hupalo et al., 2016) available in PlasmoDB database (sequences were screened to rule out those having missing data or ambiguous nucleotides); 232 sequences from different regions around the world were thus used. Afterwards the intron was ruled out from all sequences and codon alignments were inferred using the TranslatorX web server (Abascal et al., 2010). This alignment was manually edited (**Supplementary Data Sheet 1A**) to ensure correct repeat alignment for further analysis.

DnaSP v.5 software (Librado and Rozas, 2009) was used for calculating the amount of singleton sites (s), the amount of parsimony-informative sites (Ps), the amount of haplotypes (H), the haplotype diversity (Hd), the nucleotide polymorphism (or Watterson estimator, θ^{w}) and the nucleotide diversity per site (π) for all available sequences (worldwide samples), as well as for the Colombian and Venezuelan populations and for the subpopulations within both populations. Departure from the neutral model of molecular evolution was assessed by Tajima (1989); Fu and Li (1993), and Fay and Wu (2000) estimators. These are frequency spectrum-based tests for comparing two estimators of the population mutation parameter θ which characterizes mutation–drift equilibrium (neutral model). Under neutrality, several unbiased estimators of θ should be equal. Rejection of the neutral expectations suggests that selection or a demographic process could be taking place, the Fay and Wu test being suitable for detecting selective sweep (Fay and Wu, 2000). On the other hand, since each new polymorphic site has a high probability of delineating a new haplotype (Depaulis and Veuille, 1998), tests based on haplotype distribution have been developed [K- and Hd-test (Depaulis and Veuille, 1998) as well as Fu's Fs (Fu, 1997)]. Similar to frequency spectrum-based tests, departures from neutral expectation could be the consequence of selection or demographic history, Fu's Fs estimator being a more sensitive indicator of population expansion than Tajima's test (Fu, 1997). DnaSP v.5 and/or ALLELIX software were used for these tests, coalescent simulations being used for obtaining confidence intervals (Librado and Rozas, 2009). Sites containing gaps or repeats in the alignment were not taken into account.

The aforementioned methods do not consider the classes of mutations (non-synonymous and synonymous); natural selection signatures were therefore also assessed using different methods which classified mutations as non-synonymous or synonymous. The aBSREL method (Smith et al., 2015) was used to test whether positive selection had occurred on a percentage of branches regarding *rbsa* phylogeny. The modified Nei-Gojobori method (Zhang et al., 1998) with Jukes-Cantor correction (Jukes and Cantor, 1969) was then used for calculating the non-synonymous and synonymous substitution difference rate (d_N-d_S) within *P. vivax*, using a *Z*-test available in MEGA software v.6 (Tamura et al., 2013) to identify significant statistical values. Two tests were performed for assessing natural selection signals comparing different species; the McDonald–Kreitman test (McDonald and Kreitman, 1991) was calculated using a web server (Egea et al., 2008) using *P. cynomolgi* orthologous sequences. Likewise, non-synonymous divergence and synonymous divergence substitution difference rates (K_N-K_S) were also calculated, using the Z-test for identifying statistically significant values. Both tests were carried out taking Jukes-Cantor divergence correction into account.

A sliding window for omega (ω) rates (d_N/d_S and/or K_N/K_S) was then used for assessing how natural selection acts throughout the gene. The Datamonkey web server (Delport et al., 2010) was used for assessing codon sites under positive or negative selection by using codon-based ML and BYs [IFEL (Pond et al., 2006), FEL, SLAC, REL (Kosakovsky Pond and Frost, 2005), MEME (Murrell et al., 2012) and FUBAR (Murrell et al., 2013)]. A < 0.1 *p*-value was considered significant for IFEL, FEL, SLAC and MEME methods and a > 0.9 posterior probability for FUBAR. Since ignored recombination can bias d_N/d_S estimation (Anisimova et al., 2003; Arenas and Posada, 2010, 2014), the GARD method (Kosakovsky Pond et al., 2006) was considered regarding recombination before running these tests.

Linkage disequilibrium was evaluated by calculating Z_{NS} (Kelly, 1997) to assess whether recombination was/is taking place in *pvrbsa*; this was followed by linear regression between LD and nucleotide distances. Evidence of recombination was also assessed by the GARD method (Kosakovsky Pond et al., 2006), as well as by ZZ (Rozas et al., 2001) and RM (Hudson and Kaplan, 1985) tests. The degree of genetic differentiation amongst *P. vivax* malaria-endemic regions (subpopulations) regarding the *pvrbsa* locus was evaluated by analysis of molecular variance (AMOVA) and by computing Wright's fixating index (F_{ST}), using Arlequin population genetics data analysis v.3.1 software (Excoffier et al., 2007). The mutational pathways giving rise to *pvrbsa* haplotypes, their distribution and frequencies were inferred by median Joining method, using Network v.5 software (Bandelt et al., 1999).

Assessing Functional Regions in the PvRBSA Protein

Whole PvRBSA, as well as its A and B regions, were recombinantly expressed and purified, as previously described (Moreno-Perez et al., 2017) with minor modifications. For example, the protein was expressed for 4 h at 37°C (for complete PvRBSA) or 30°C (for PvRBSA-A and PvRBSA-B regions) adding 0.2 mM IPTG. The cell pellet was homogenized in B1 buffer (20 mM Tris-Cl, 500 mM NaCl and 1 mM EDTA, pH 8.5) during native protein extraction and then lysed by sonication on ice for 3 min 30 s, with 0.2 s pulses ON, 0.2 s OFF, at 40% amplitude. Protein purification and cell binding assays involved using the same protocols described for PvGAMA and PvRBSA proteins, including PvDBP-RII as positive control and PvDBP-RIII/IV as negative control (Baquero et al., 2017; Moreno-Perez et al., 2017). The competition assay was performed by pre-incubating

rPvRBSA-A-derived peptides 40893 (TASSESLAESNDAPS NSYES), 40894 (FPEIRENLTASEESLTSCEE), 40895 (SLTGSN ESLTGSNE), 40896 (SLTESRESLEASRESLRASR), 40897 (ESL AASRESLNDFCGSEESV) and 40898 (ACEGEPNEKTFMGDV LSGGE) [synthesized as described (Arevalo-Pinzon et al., 2017)] in a 1:20 (protein:peptide) molar ratio with 2×10^7 red blood cells from umbilical cord blood for 1 hour at 4°C at 4 rpm using a tube rotator (Fisher Scientific). Furthermore, a peptide (P7) from *Mycobacterium tuberculosis* (39266 – APSNETLVKTFSPGEQVTTY) (Carabali-Isajar et al., 2018), was also used as negative control. All cysteine residues were replaced by a threonine to avoid RBSA peptide polymerisation. RBSA reticulocyte binding activity inhibition percentage was quantified by analysing 100,000 events using a FACS Canto II (Biosciences) cytometer and FACSDiva software (BD).

Statistical Analysis

Differences between medians (m) were compared by Kruskal–Wallis test when comparing multiple groups. Statistical significance was assessed by comparing m, using a 0.05 significance level. Median values and standard deviations (SD) were calculated from the measurements of three independent experiments.

RESULTS

Identifying the *rbsa* Gene in *Plasmodium* Monkey Lineage

The Sal-I-*P. vivax rbsa* gene (*pvrbsa*) sequence was used for a Blast search of *P. vivax* phylogenetically related species to identify orthologs in the monkey-parasite lineage. A contig was found in *P. inui* and *P. fragile* having higher than 60% identity whilst two different contigs were found in *P. cynomolgi* having 80% identity with Sal-I *pvrbsa* (**Supplementary Data Sheet 1B**). Interestingly, both *P. cynomolgi* contigs had the same identity but they did not belong to the same chr; one of the two fragments occurred at the chr3 and the other one at chr10. There was 99.57% identity between these *P. cynomolgi* DNA fragments.

A similar blast search using Sal-I *pvrbsa* sequence was then performed against other *P. vivax* reference stains' whole genome sequences (Neafsey et al., 2012), revealing extra DNA fragments having 63% identity at the same chr, lacking the first exon and the intron sequences (**Supplementary Data Sheet 1C**), probably derived from duplication. The duplicated fragment was around 4,000 bp (**Figure 1**, **Supplementary Data Sheet 1C** and **Supplementary Figure S2**), taking upstream and downstream DNA regions into account (Sal-I Ch3 GenBank accession number: AAKM01000020.1) from *pvrbsa* and the *pvrbsa* paralog (*pvrbsap*). Counterparts to this 4,000 bp-DNA fragment were also identified in *P. cynomolgi* (at both chr), *P. inui* and *P. fragile*, which were aligned and used to infer a phylogenetic tree (**Supplementary Figure S2**). Three well-supported groups were identified in the tree; the first one placed *P. vivax* sequences together, the second group clustered *P. cynomolgi* sequences and in the third, *P. inui* and *P. fragile* sequences were found (**Supplementary Figure S2**).

Aligning *P. cynomolgi rbsa* (*pcrbsa,* Ch3 GenBank accession number: BAEJ01000151.1) and *pcrbsa* paralogue (*pcrbsap*, Ch10 GenBank accession number: BAEJ01000874.1) DNA fragments revealed a 10 Mb DNA region between chr3 and ch10 (**Figure 1**, **Supplementary Data Sheet 1D** and **Supplementary Figure S2**), having 98.59% identity.

pvrbsa Gene PCR Amplification and Sequencing

One hundred and sixty-seven parasite DNAs from endemic Colombian and Venezuelan regions were used for amplifying the *pvrbsa* locus. The amplicons had 1,454–1,580 bp sizes. These PCR fragments were then purified and sequenced. At least three chromatograms were assembled for each sample, thereby obtaining a consensus sequence.

pvrbsa Gene DNA Diversity and Evolutionary Analysis

Together with the 167 sequences obtained here, 65 additional sequences recovered from PlasmoDB (Colombia n = 18, Peru n = 16, Brazil n = 3, Mexico n = 5, Madagascar n = 2, China n = 2, Cambodia n = 2, Thailand n = 8, Papua New Guinea n = 6, India n = 1, North Korea n = 1 and the Sal-I sequences, **Supplementary Data Sheet 1A**) were used to assess genetic diversity in *pvrbsa*. The *pvrbsa* encoding region had 44 segregating sites whilst just 3 polymorphic sites were found at intron worldwide. *pvrbsa* had an intermediate diversity according to the nucleotide diversity estimator (π) (**Table 1**, $\pi < 0.01$). The Venezuelan parasite population had a higher π than the Colombian population; both subpopulations within Venezuela (Bolivar and Venezuelan's Coastal area) had similar π values. Regarding Colombian subpopulations, parasites from Meta (sequences available in PlasmoDB) had the lowest diversity, whilst Córdoba had the highest diversity value for Colombian subpopulations (**Table 1**). The 232 sequences could be clustered into 80 different haplotypes (70 regarding CDS, **Supplementary Data Sheet 1E**). The Colombian population had 37 haplotypes whilst 25 haplotypes were observed in Venezuela (**Table 1**).

Since different evolutionary forces (drift, selection, recombination and migration) could determine the genetic diversity pattern (Casillas and Barbadilla, 2017) observed in natural populations, several evolutionary tests were performed to infer which of them were modulating the *pvrbsa* gene diversity observed here. Tests based on a neutral molecular evolutionary model were performed for population and subpopulations to assess drift/selection force. The Colombia population had negative statistically significant values just for the Fay and Wu test (**Table 2**) indicating a possible selective sweep. When subpopulations were analyzed separately, Meta had positive values for the Fu and Li test, being statistically significant ($p < 0.05$) which could have resulted from balancing selection or a decrease in population. The Chocó subpopulation had a value lower than 0 ($p < 0.03$) for Fay and Wu's H test (**Table 2**). A sliding window for frequency spectrum-based tests showed regions inside *pvrbsa* having significant values (**Supplementary Figure S3**). On the other hand, the Fu and Li's F estimator gave

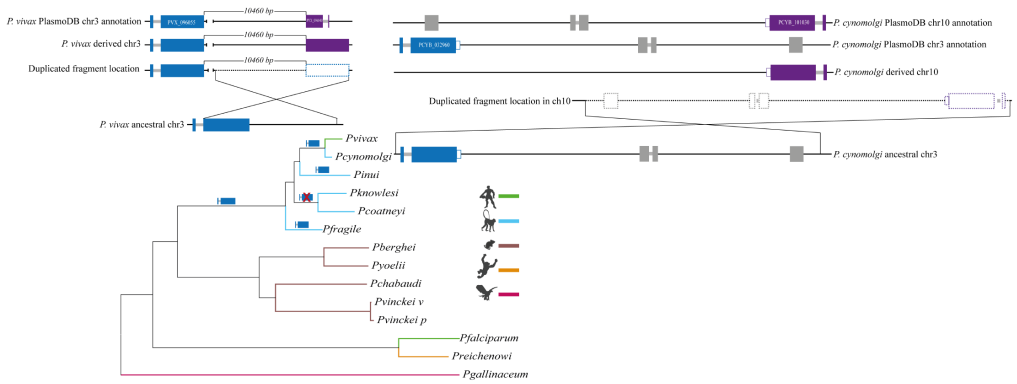


FIGURE 1 | Phylogenetic tree mapping the presence/absence of *rbsa* locus and duplication events. The phylogenetic tree represents *Plasmodium* evolutionary history. The branches are colored regarding their vertebrate host. The *rbsa* locus was found in parasites from monkey-malaria clade, but it was not clear whether the *rbsa* gene became lost from the other clades or arose exclusively in the monkey-malaria clade. *rbsa* became lost in *P. knowlesi* and *P. coatneyi* whilst duplication events occurred in *P. vivax* and *P. cynomolgi*. Models of putative ancestral chromosome 3 are displayed, indicating the arrangement after the duplication event (chr derived). PlasmoDB genome fragment annotation is also shown. *P. cynomolgi rbsa* (*pcrbsa*) and its paralog (*pcrbsap*) had a premature stop codon regarding *P. vivax rbsa* sequences, indicated by a white rectangle in the *pcrbsa* and *pcrbsap* gene models.

TABLE 1 | Genetic diversity estimators.

	n		Aligned length	S	Ps	H	θw (sd)	π (sd)	Hd (sd)
Worldwide isolates	232	Gene	1,354	15	32	80	0.0067 (0.0010)	0.0085 (0.0002)	0.970 (0.005)
	232	CDS	1,113	14	30	70	0.0066 (0.0010)	0.0080 (0.0002)	0.962 (0.006)
Colombia[a]	135	Gene	1,354	13	26	37	0.0061 (0.0010)	0.0079 (0.0002)	0.951 (0.009)
	135	CDS	1,113	12	24	32	0.0059 (0.0010)	0.0074 (0.0002)	0.937 (0.010)
Amazonas[b]	41	Gene	1,391	6	20	13	0.0050 (0.0010)	0.0072 (0.0004)	0.890 (0.028)
	41	CDS	1,170	6	18	13	0.0049 (0.0010)	0.0067 (0.0004)	0.890 (0.028)
Chocó[b]	37	Gene	1,354	14	20	18	0.0072 (0.0012)	0.0079 (0.0005)	0.954 (0.024)
	37	CDS	1,113	13	18	16	0.0069 (0.0012)	0.0074 (0.0005)	0.936 (0.024)
Córdoba[b]	39	Gene	1,405	3	24	19	0.0053 (0.0010)	0.0080 (0.0006)	0.956 (0.023)
	39	CDS	1,170	3	22	17	0.0051 (0.0010)	0.0076 (0.0006)	0.944 (0.024)
Meta[b]	18	Gene	1,506	0	16	8	0.0036 (0.0009)	0.0055 (0.0003)	0.961 (0.033)
	18	CDS	1,290	0	14	8	0.0034 (0.0008)	0.0050 (0.0003)	0.961 (0.033)
Venezuela[a]	50	Gene	1,363	2	28	25	0.0056 (0.0010)	0.0089 (0.0003)	0.958 (0.023)
	50	CDS	1,149	2	26	25	0.0055 (0.0010)	0.0087 (0.0003)	0.958 (0.023)
Bolívar[b]	29	Gene	1,363	0	27	17	0.0059 (0.0011)	0.0088 (0.0005)	0.975 (0.027)
	29	CDS	1,149	0	25	17	0.0057 (0.0011)	0.0085 (0.0005)	0.975 (0.027)
Coastal area[b]	19	Gene	1,363	6	22	13	0.0070 (0.0013)	0.0094 (0.0005)	1.000 (0.036)
	19	CDS	1,149	5	21	13	0.0069 (0.0013)	0.0093 (0.0005)	1.000 (0.036)

Genetic diversity estimators were calculated using all the pvrbsa sequences available (worldwide isolates) as well as for populations (Colombia and Venezuela) and subpopulations (Amazonas, Chocó, Córdoba, Meta, Bolivar, and Venezuela's coastal area). n, amount of isolates analyzed; aligned length, total sites analyzed, excluding gaps; Ss, amount of segregating sites; Ps, amount of informative parsimonious sites; H, amount of haplotypes; θw, nucleotide polymorphism; π, nucleotide diversity per site; Hd, haplotype diversity. sd, standard deviation. Diversity estimates were corrected for sample size by multiplying by n/(n-1). Although nucleotide diversity was not affected by sample size, it was also corrected. [a]Populations. [b]Subpopulations.

statistically significant values for the Venezuelan population; this pattern was also observed for the Bolivar subpopulation (**Table 2**). The sliding window analysis revealed regions inside *pvrbsa* having significant values for Venezuelan subpopulations. Tests based on haplotype distribution did not have statistically significant values (**Table 2**).

A test based on non-synonymous and synonymous mutations was also computed to search natural selection signals in *pvrbsa*. The aBSREL method found evidence of episodic diversifying selection on *rbsa* phylogeny (**Figure 2A**). The d_N-d_S difference had positive values within *P. vivax*; however, they were not statistically significant, except for the Meta subpopulation (**Table 3**). A sliding window was then inferred to assess how d_N and d_S have been accumulated throughout the *pvrbsa* gene (**Figure 2B**). The 3′-end showed omega (ω) values higher than 1 (**Figure 2B**) which is an indicator of positive selection. The region from nucleotide 200–1,100 had ω = 0 which is expected under negative selection. Although, the McDonald–Kreitman test did not have significant values, the K_N-K_S differences (comparing *P. vivax* sequences to *P. cynomolgi* ones) had

TABLE 2 | Neutrality tests for Colombia, Venezuela, and subpopulations.

	n		Tajima	Fu and Li		Fay and Wu's *H* (*p*-value)	Fu's Fs (*p*-value)	*K*-test (*p*-value)	*H*-test (*p*-value)
			D (*p*-value)	D (*p*-value)	F (*p*-value)				
CDS	135	Colombia[a]	0.637 ($p > 0.1$)	−0.779 ($p > 0.1$)	−0.304 ($p > 0.1$)	−12.819* ($p < 0.01$)	−3.447 ($p > 0.1$)	32 ($p > 0.05$)	0.930 ($p > 0.05$)
	41	Amazonas[b]	1.192 ($p > 0.1$)	0.262 ($p > 0.1$)	0.789 ($p > 0.1$)	−4.821 ($p > 0.1$)	1.192 ($p > 0.1$)	13 ($p > 0.05$)	0.868 ($p > 0.05$)
	37	Chocó[b]	0.174 ($p > 0.1$)	−0.460 ($p > 0.1$)	−0.285 ($p > 0.1$)	−11.339* ($p < 0.03$)	−0.755 ($p > 0.1$)	16 ($p > 0.05$)	0.911 ($p > 0.05$)
	39	Córdoba[b]	1.424 ($p > 0.1$)	1.046 ($p > 0.1$)	1.447 ($p > 0.1$)	−5.036 ($p > 0.1$)	−0.758 ($p > 0.1$)	17 ($p > 0.05$)	0.920 ($p > 0.05$)
	18	Meta[b]	1.856 ($p > 0.1$)	1.543* ($p < 0.05$)	2.118* ($p < 0.02$)	−1.673 ($p > 0.1$)	1.050 ($p > 0.1$)	8 ($p > 0.05$)	0.908 ($p > 0.05$)
	50	Venezuela[a]	1.865 ($p > 0.1$)	−5.036 ($p > 0.05$)	2.168* ($p < 0.02$)	−4.140 ($p > 0.1$)	−3.851 ($p > 0.1$)	25 ($p > 0.05$)	0.939 ($p > 0.05$)
	29	Bolívar[b]	1.735 ($p > 0.05$)	1.806* ($p < 0.02$)	2.236* ($p < 0.02$)	−5.197 ($p > 0.1$)	−1.969 ($p > 0.1$)	17 ($p > 0.05$)	0.941 ($p > 0.05$)
	19	Coastal area[b]	1.426 ($p > 0.1$)	0.910 ($p > 0.1$)	1.392 ($p > 0.1$)	−4.520 ($p > 0.1$)	−1.196 ($p > 0.1$)	13 ($p > 0.05$)	0.947 ($p > 0.05$)

*The tests were carried out with just the rbsa encoding region (CDS). *Statistically significant values. [a]Populations. [b]Subpopulations.*

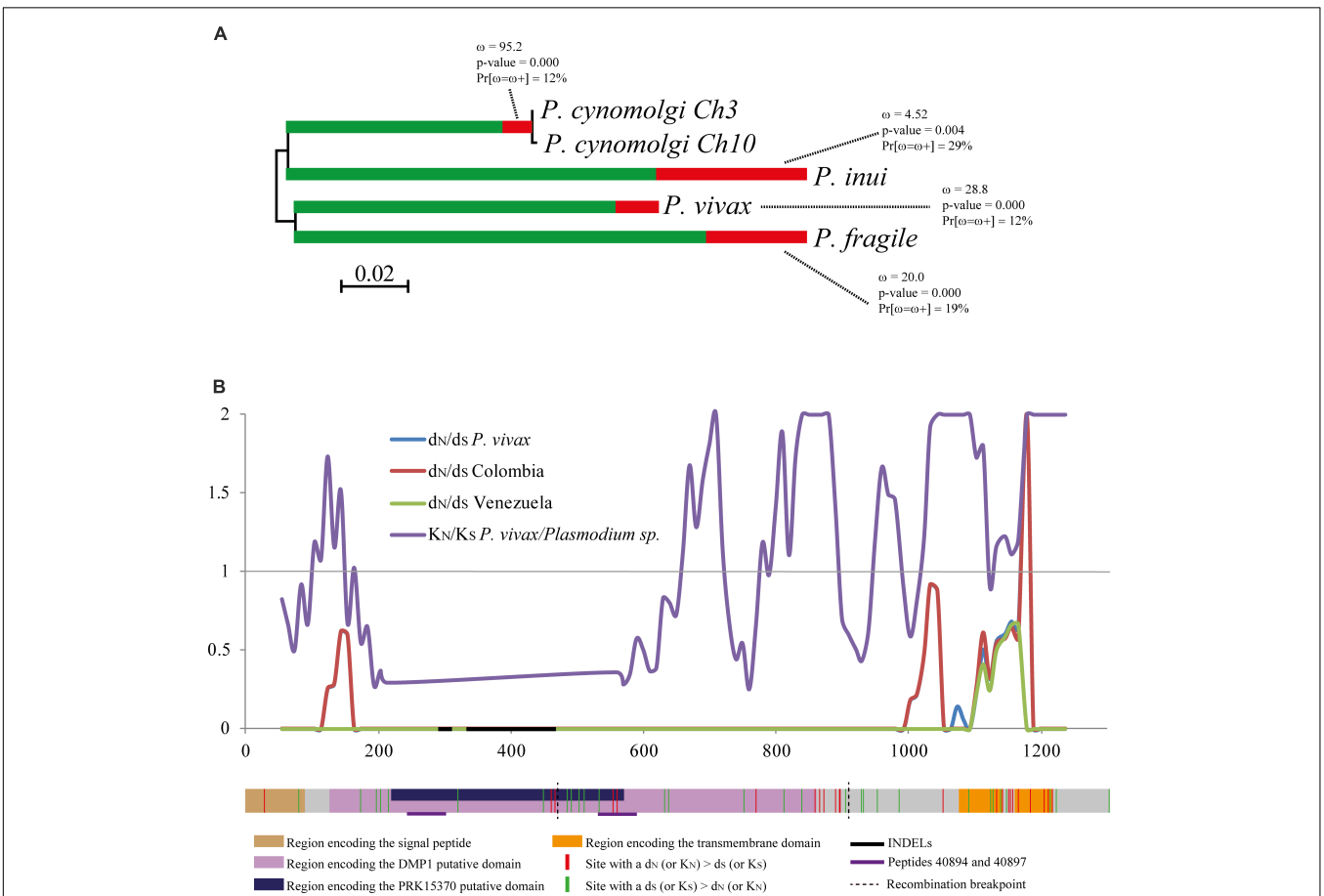


FIGURE 2 | Identifying natural selection signatures in the *rbsa* gene. **(A)** A phylogeny tree inferred from *pvrbsa* CDS and its orthologous CDS was analyzed by the aBSREL method. The shade of each color on branches indicates selection strength [red indicates positive selection ($\omega > 1$) and green negative selection ($\omega < 1$)]. The size of each colored segment represents the percentage of selected sites in the corresponding ω class. The Figure shows the ω values, the percentage of positive selected sites (Pr [ω = ω+]) and *p*-values. Branches have been classified as undergoing episodic diversifying selection, by the *p*-value corrected for multiple testing, using the Holm–Bonferroni method at $p < 0.05$. **(B)** Representation of the *rbsa* encoding DNA sequence and the sliding window for the ω rate. The ω rates within *P. vivax* (d_N/d_S) and between the species (K_N/K_S) are shown. A gene model of the *rbsa* gene is given below the sliding window, indicating the sites under purifying selection (green) and positive selection (red). Gene model and numbering were based on the **Supplementary Data Sheet 1E** alignment.

positive selection signals (positive significant values, **Table 4**). A similar result was obtained when *P. vivax* and all orthologous sequences from phylogenetically related species were analyzed (**Table 4**). The sliding ω (K_N/K_S) window for this dataset showed positive selection signatures ($\omega > 1$) towards the 5′-end, as well as in a region between nucleotide 600 – 1,572 (hereinafter called *Pv*RBSA-B). The region between nucleotide 200 – 600 (hereinafter called *Pv*RBSA-A) had a negative

selection signal ($\omega < 1$, **Figure 2B**). Consequently, the gene was split into two regions (PvRBSA-A and PvRBSA-B) and d_N-d_S (as well as K_N-K_S) were computed again for each one. No statistically significant values were observed for d_N-d_S in either of the two regions (except for the Meta subpopulation, **Table 3**). Conversely, positive values (indicator of positive selection) having a *p*-value ≤ 0.045 were found in PvRBSA-B when K_N-K_S was computed; by contrast, PvRBSA-A had a negative selection signature (negative values having $p > 0.05$, **Table 4**).

ω rate was then computed for each codon using codon-based methods. Twenty-one positive selected codons were found between species whilst another 30 were under negative selection. Several of the negative selected sites were found in PvRBSA-A (**Figure 2B**). A protein blast (using NCBI database with Sal-I PvRBSA haplotype as query) gave two putative domains in the

TABLE 3 | d_N-d_S difference within *P. vivax.*

Population	PvRBSA-A (196 - 600 bp)	PvRBSA-B (601 – 1,353 bp)	Full length gene
	d_N - d_S (se)	**d_N - d_S (se)**	**d_N - d_S (se)**
Worldwide isolates	0.0014 (0.0022)	0.0060 (0.0046)	0.0044 (0.0028)
	$p = 0.255$	$p = 0.102$	$p = 0.062$
Colombian[a]	0.0011 (0.002)	0.0050 (0.004)	0.0036 (0.003)
	$p = 0.318$	$p = 0.145$	$p = 0.099$
Amazonas[b]	0.0016 (0.001)	0.0033 (0.005)	0.0027 (0.003)
	$p = 0.083$	$p = 0.251$	$p = 0.177$
Chocó[b]	0.0023 (0.002)	0.0050 (0.005)	0.0039 (0.003)
	$p = 0.161$	$p = 0.133$	$p = 0.078$
Córdoba[b]	0.0006 (0.004)	0.0036 (0.005)	0.0027 (0.003)
	$p = 0.441$	$p = 0.811$	$p = 0.173$
Meta[b]	−0.0024 (0.004)	0.0082 (0.003)*	0.0045 (0.002)*
	$p = 0.248$	$p = 0.009$	$p = 0.023$
Venezuela[a]	0.0015 (0.002)	0.0064 (0.005)	0.0046 (0.003)
	$p = 0.269$	$p = 0.120$	0.070
Bolívar[b]	0.0023 (0.002)	0.0058 (0.005)	0.0045 (0.003)
	$p = 0.166$	$p = 0.124$	$p = 0.076$
Coastal area[b]	0.0002 (0.003)	0.0070 (0.006)	0.0047 (0.003)
	$p = 0.472$	$p = 0.099$	$p = 0.081$

*Non-synonymous substitution rate (d_N) and synonymous substitution rate (d_S) within P. vivax. se, standard error. *Statistically significant values. [a]Populations. [b]Subpopulations.*

TABLE 4 | K_N–K_S difference between *Plasmodium* species.

Population	RBSA-A (196 – 600 bp)	RBSA-B (601 – 1,353 bp)	Full length gene
	K_N – K_S (se)	**K_N – K_S (se)**	**K_N – K_S (se)**
P. vivax*/*P. cynomolgi			
Worldwide isolates	−0.0010 (0.003)	0.0080 (0.005)*	0.0053 (0.003)*
	$p = 0.372$	$p = 0.045$	$p = 0.033$
Colombian	−0.0030 (0.004)	0.0080 (0.005)*	0.0053 (0.003)*
	$p = 0.249$	$p = 0.033$	$p = 0.032$
Venezuela	−0.0092 (0.010)	0.0150 (0.006)*	0.0090 (0.004)*
	$p = 0.188$	$p = 0.005$	$p = 0.012$
***P. vivax*/*Plasmodium* sp.**			
Worldwide isolates	−0.0031 (0.004)	0.0101 (0.005)*	0.0068 (0.003)*
	$p = 0.234$	$p = 0.017$	$p = 0.012$
Colombian	−0.0064 (0.007)	0.0118 (0.005)*	0.0077 (0.003)*
	$p = 0.184$	$p = 0.008$	$p = 0.007$
Venezuela	−0.0174 (0.002)	0.0238 (0.006)*	0.0152 (0.005)*
	$p = 0.167$	$p = 0.0001$	$p = 0.001$

*Non-synonymous (K_N) and synonymous (K_S) divergence between P. vivax/P. cynomolgi and P. vivax/phylogenetically (P. cynomolgi/P. inui/P. fragile) related species. se, standard error. *Statistically significant values.*

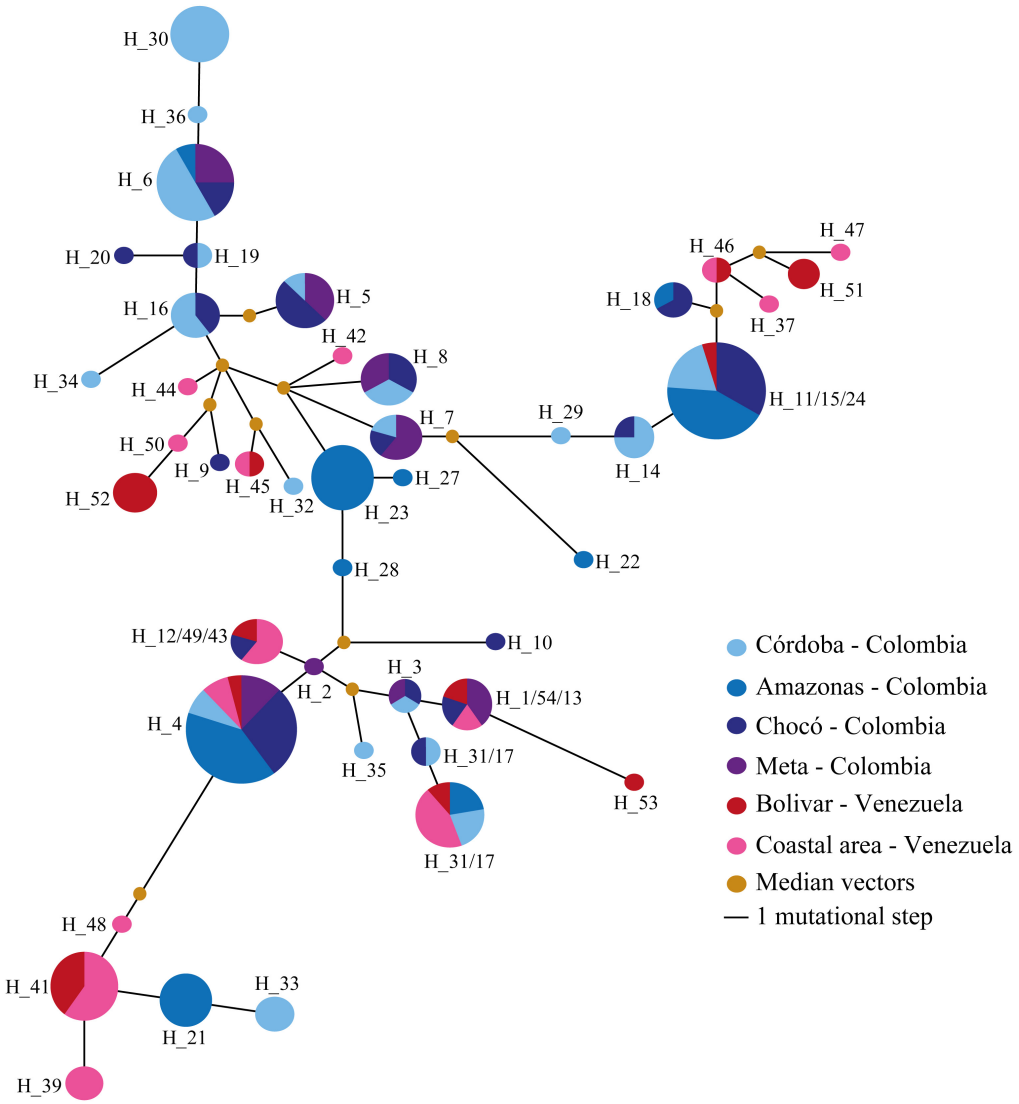


FIGURE 3 | Median-joining network for Colombian and Venezuelan subpopulations. The Figure shows the *pvrbsa* haplotypes identified from Colombian and Venezuelan isolates. Some haplotypes were included within another haplotype using the contraction star algorithm (Forster et al., 2001) for simplifying network interpretation. Each node is a haplotype and its size indicates its frequency. The lines connecting the haplotypes represent the different mutational paths and the median vectors are the ancestral sequences explaining the relationship and evolutionary origin.

regions having $\omega < 1$ (PvRBSA-A). The first belonged to the DMP1 domain and the other to a domain containing a LRR.

Recombination is an evolutionary force which can provide fresh opportunities for overcoming selective pressures to adapt to new environments and/or hosts by linking (within the same DNA region) independently arising variants (Perez-Losada et al., 2015). In fact, it has been observed that diversity levels increase with recombination rate (Hellmann et al., 2003; Kulathinal et al., 2008; Rao et al., 2011). Several tests were thus performed to assess whether this evolutionary force takes/has taken place in *pvrbsa*. A linear regression between LD and nucleotide distance showed that LD decreased regarding increased nucleotide distance, a pattern expected under recombination. Statistically significant ZZ values and several Rm were observed for all populations, thereby confirming recombination action (**Supplementary Data Sheet 2**). Two breakpoints were identified by the GARD method at nucleotides 618 and 1058 ($p = 0.0004$, nucleotide number base in the Sal-I sequence) (**Figure 2**). Recombination was thus taking place at this locus.

A haplotype network was inferred (**Figure 3**) and AMOVA used for assessing whether migration was also involved in shaping *pvrbsa* locus diversity (**Table 5**). Several haplotypes were shared between populations (Colombia and Venezuela, i.e. H_4) but also amongst subpopulations (Chocó, Córdoba, Amazonas, Meta, Bolivar, and the Venezuelan coastal area) lacking clear population structure (**Figure 3**). However, some haplotypes were restricted to particular subpopulations (i.e.

TABLE 5 | Analysis of molecular variance (AMOVA) analysis and inter-population F_{ST} values.

Source of variation	% of variation	*p*-value
Between populations (F_{CT})	2.81	0.064
Amongst subpopulations within populations (F_{SC})	4.39	0.000*
Amongst subpopulations (F_{ST})	92.79	0.001*

	F_{ST}					
	Meta	**Chocó**	**Amazonas**	**Córdoba**	**Bolívar**	**Coastal area**
Meta		0.286	0.004	0.007	0.000	0.000
Chocó	0.007		0.000	0.006	0.000	0.001
Amazonas	0.074*	0.055*		0.000	0.000	0.000
Córdoba	0.041*	0.027*	0.086*		0.000	0.000
Bolívar	0.057*	0.049*	0.083*	0.061*		0.146
Coastal area	0.076*	0.056*	0.094*	0.065*	0.014	

*F_{ST} was calculated for parasite subpopulations within Colombian and Venezuela. Values close to 0 indicate low genetic differentiation whilst values close to 1 indicate high genetic differentiation. Values below the diagonal correspond to the F_{ST} value and those above the diagonal represent the respective p-values. *Statistically significant values.*

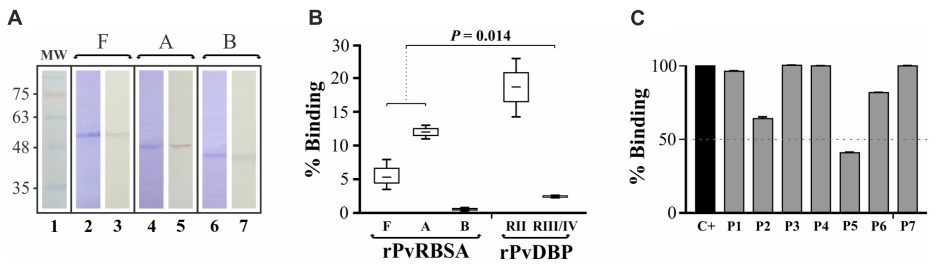


FIGURE 4 | PvRBSA reticulocyte binding activity. **(A)** Purification of full rPvRBSA (F) or its fragments (A and B). The Figure shows each purified molecule analysed by Coomassie blue staining (lines 2, 4, and 6) or Western blot (lines 3, 5, and 7). MW, molecular weight marker (line 1). **(B)** rPvRBSA target cell binding assay. DBP-RII and -RIII/IV indicates positive and negative binding controls, respectively. **(C)** Binding inhibition assay. The Figure shows the rPvRBSA reticulocyte binding inhibition percentage using different PvRBSA-A-derived peptides as well as a *M. tuberculosis* peptide (negative controls). Standard deviation is shown for each assay. C+, shows the assay just with the rPvRBSA protein (positive control). P1-P6 indicate peptides 40893, 40894, 40895, 40896, 40897, and 40898, respectively. P7 is a *M. tuberculosis* peptide (39266) used as negative control. All peptides were pre-incubated with cells before incubating with the complete protein.

H_30, H_23, H_52, H_39) or populations (i.e., H5 - H_8, H_41, H_46). AMOVA was then used to address population differentiation (**Table 5**); analysis showed that the source of variation was between populations (around 4%, $p = 0.00$), the greatest variation (93%) occurring amongst subpopulations ($p = 0.00$). The F_{ST} had statistically significant values (**Table 5**). However, comparing Meta/Chocó (as well as Bolivar/Venezuela's coastal area) revealed no statistically significant F_{ST} values (**Table 5**).

Assessing PvRBSA Functional Regions

It has been shown that immature reticulocytes are *P. vivax* target cells, expressing the CD71 receptor abundantly on their surface (Malleret et al., 2015). Therefore, a reticulocyte-binding assay was performed by flow cytometry to evaluate whether the PvRBSA region predicted under negative selection (PvRBSA-A having a $\omega < 1$) was able to bind to target cells (labeled with anti-CD71 antibody). First, the full PvRBSA protein and the PvRBSA-A and PvRBSA-B regions were recombinantly expressed and obtained in soluble form. Each protein was purified and recognized by WB using a monoclonal anti-polyhistidine antibody (**Figure 4A**). When proteins were incubated with umbilical cord blood (containing 6–7% reticulocytes), only the rPvRBSA ($m \pm SD = 5.5 \pm 1.8$) and rPvRBSA-A regions ($m \pm SD = 13 \pm 1.7$) bound to a CD71+CD45- cell population (reticulocytes), unlike the rPvRBSA-B region (**Figure 4B**). As can be observed, rPvRBSA and rPvRBSA-A binding activity had a statistically significant difference compared to negative control ($m \pm SD = 2.4 \pm 1.3$) (Kruskal–Wallis: $p = 0.014$) (**Figure 4B**). Once the PvRBSA functional region's reticulocyte interaction activity had been determined, a binding inhibition experiment was performed with PvRBSA-A-derived peptides to search for specific regions within this region involved in interaction with target cells (**Figure 4C**). Peptide 40898 was able to inhibit protein-cell interaction by 18% whilst peptide 40894 produced a 35% reduction. Only one peptide (40897) inhibited rPvRBSA-reticulocyte interaction by more than 50%. The *Mycobacterium* peptide was unable to inhibit rPvRBSA binding activity to reticulocytes. Interestingly, synthesized peptides 40894 and 40897 were conserved in several *P. vivax* isolates and were located in the PRK15370 putative domain (**Figure 2B**).

DISCUSSION

Malaria remains a public health problem in several tropical and subtropical regions worldwide (WHO, 2017). Although an antimalarial vaccine appears to be a good cost-effective intervention which would help in controlling malaria, a fully effective vaccine has not been developed yet. Since *P. vivax* has a complex biology (Galinski et al., 2013) (for instance, it has tropism for reticulocytes, making it difficult to obtain a continuous *in vitro* culture), antigen identification and characterization for an antimalarial vaccine is a slow task (Patarroyo et al., 2012). Several antigens suggested as potential *P. vivax* vaccine candidates have been characterized, taking into account the candidates proposed for *P. falciparum* or other *Plasmodium* species (Arevalo-Pinzon et al., 2011, 2013, 2015; Patarroyo et al., 2012; Moreno-Perez et al., 2013). However, *P. vivax* and *P. falciparum* have different features and this could be the consequence of their different evolutionary paths. Consequently, several species-specific antigens could be found in both species. These species-specific antigens could thus be taken into account when designing an antimalarial vaccine.

The RBSA has recently been identified in species invading reticulocytes (*P. vivax* and *P. cynomolgi*) suggesting that this antigen could be specific for invading this kind of host cell and could therefore be considered a vaccine candidate when designing an anti-*P. vivax* malaria vaccine (Moreno-Perez et al., 2017). However, this antigen appears not to be exclusive for *P. vivax* and *P. cynomolgi*. A Blast search in whole monkey-malaria lineage genomes showed that *pvrbsa* orthologs were also present in species invading normocytes, such as *P. inui* and *P. fragile* (**Supplementary Data Sheet 1B**). Since the *rbsa* gene was present in *P. fragile* (the most basal species in monkey-malaria lineage, **Figure 1**) (Carlton et al., 2013; Muehlenbein et al., 2015), this gene could be present in the ancestor of all monkey-malaria parasites and, because orthologs were not found in *P. knowlesi* and *P. coatneyi*, then it has been lost in some species within this clade.

The *rbsa* gene was found as a single copy gene in *P. inui* (*pirbsa*) and *P. fragile* (*pfrbsa*), but this gene appears to be duplicated in *P. vivax* (*pvrbsa*) and *P. cynomolgi* (*pcrbsa*). The phylogenetic tree (**Supplementary Figure S2**) showed that *pvrbsa* was closer to *pvrbsap* than *pcrbsa*. This could be the consequence of concerted evolution which can homogenize duplicate gene fragments (Nei and Rooney, 2005). However, duplication was incomplete or was found at different chromosomes; the duplication event must therefore have taken place after *P. vivax* and *P. cynomolgi* divergence. A 4,000 bp fragment was duplicated at chr3 in *P. vivax* (it is present in all *P. vivax* strain genomes available in the PlasmoDB database; duplication and pseudogenisation should therefore have been taking place during early *P. vivax* evolutionary history); this duplicate (*pvrbsap*) lacked the first exon and the intron (**Figure 1**) and could, consequently, be a pseudogene. By contrast, both *pcrbsa* copies were complete in *P. cynomolgi*; in fact, they had 99.91% identity. Furthermore, *pcrbsap* was not found at chr3; instead, it was located at chr10. At least a 10 Mbp were duplicated for this species; however, the mechanism involved in this large fragment's duplication at a different chr is not clear. Likewise, it is not yet clear whether the high identity found in *pcrbsa* and *pcrbsap* has been due to negative selection acting on both copies or has been due to a recent duplication event. According to the aforementioned results, two independent duplication events must have happened.

On the other hand, the *pvrbsa* gene was amplified in 167 samples from Colombian and Venezuelan populations. The derived sequences were analysed together with 65 sequences from different regions around the world. This gene had length polymorphism due to repeats located at the 5′-end. Worldwide, the *pvrbsa* encoding region had 44 segregating sites (47, taking encoding and non-encoding regions into account) giving 70 haplotypes (**Supplementary Data Sheet 1E**). The *pvrbsa* π value (0.0080 ± 0.0002) was lower than that observed for other Mrz surface proteins [*pvmsp1*, $\pi > 0.05200$ (Valderrama-Aguirre et al., 2011; Garzon-Ospina et al., 2015); *pvmsp3α*, $\pi > 0.0349$ (Mascorro et al., 2005; Garzon-Ospina et al., 2015); *pvmsp7C*, $\pi = 0.0548$; *pvmsp7H*, $\pi = 0.0357$; *pvmsp7I*, $\pi = 0.0430$ (Garzon-Ospina et al., 2012); *pvmsp7E*, $\pi = 0.0573$ (Garzon-Ospina et al., 2014)], but similar to that found in *pvama1* [$\pi = 0.0067$ (Gunasekera et al., 2007; Garzon-Ospina et al., 2015)] and *pvdbp* [$\pi = 0.0101$ (Nobrega de Sousa et al., 2011)] and higher than *pvmsp7* ($-A$ $\pi = 0.0002$, $-K$ $\pi = 0.0025$, $-F$ $\pi = 0.0008$, and $-L$ $\pi = 0.0006$) (Garzon-Ospina et al., 2011; Garzon-Ospina et al., 2014), *pvmsp8* ($\pi = 0.0022$) (Pacheco et al., 2012), *pvmsp10* ($\pi = 0.0002$) (Garzon-Ospina et al., 2011; Pacheco et al., 2012), *pv12* ($\pi = 0.0004$), *pv38* ($\pi = 0.0026$) and *pv41* ($\pi = 0.0037$) (Forero-Rodriguez et al., 2014a,b; Wang et al., 2014) or rhoptry proteins [*rap1*, $\pi = 0.00088$, *rap2*, $\pi = 0.00141$, and *ron4*, $\pi = 0.0004$ (Garzon-Ospina et al., 2010; Pacheco et al., 2010; Buitrago et al., 2016)].

The Venezuelan subpopulations had more genetic diversity than subpopulations within Colombia regarding this gene (**Table 1**); however, the forces causing such diversity seemed to be similar. The sliding windows for frequency spectrum-based tests (**Supplementary Figure S3**) had statistically significant values within *pvrbsa*, suggesting that natural selection seems to be determining the pattern of diversity. The negative values ($p < 0.02$) found between nucleotides 100 and 300 (**Supplementary Figure S3**, numbers based on **Supplementary Data Sheet 1E**) in populations and subpopulations, suggested directional (negative or positive) selection; this kind of selection decreased diversity and could have been a consequence of functional constraint. On the other hand, the 750–940 nucleotide region was under balancing selection, as suggested by positive values for the frequency spectrum-based tests for these positions (**Supplementary Figure S3**). Directional selection was also identified between nucleotides 1,000–1,090 in the Colombian subpopulations whilst balancing selection was also found between nucleotides 1,090–1,200 in the Venezuelan subpopulations (**Supplementary Figure S3**).

Tests based on non-synonymous and synonymous mutations confirmed natural selection. A percentage of sites under positive selection were identified in *P. vivax* and other species regarding the aBSREL method (**Figure 2A**). This pattern has also been observed in other genes, suggesting species-specific adaptation during parasite evolution (Muehlenbein et al., 2015; Buitrago et al., 2016; Garzon-Ospina et al., 2016). Evidence of positive selection was found in PvRBSA-B (**Figure 2** and **Tables 3**, **4**), agreeing with the frequency spectrum-based test results. All this data (**Figure 2**, **Tables 3**, **4**, and **Supplementary Figure S3**) suggested that non-synonymous mutations were maintained in populations by balancing (or diversifying) selection, providing the parasite with an advantage to avoid host immune responses, as has been proposed for other antigens (Garzon-Ospina et al., 2012).

Whilst PvRBSA-B could be involved in host immune evasion, PvRBSA-A might be involved in parasite-host interaction. Negative values were found in d_N-d_S and K_N-K_S tests; however, they were not statistically significant (**Tables 3**, **4**). Nevertheless, statistically significant negative frequency spectrum-based test values were observed. The ω rate was lower than 1 in this region, several codons being under negative selection (**Figure 2**) according codon-based methods; this suggested that negative selection is/has been operating in PvRBSA-A and this region therefore seems to be under functional constraint. This premise was also supported because two putative domains were inferred within this region (**Figure 2B**); one was the DMP1 superfamily domain which is found in dentin matrix protein 1 acting as transcriptional component in mammals (Narayanan et al., 2003). However, how this domain could act in the parasite is not clear yet.

More interesting was the finding of the PRK15370 domain. This domain belongs to the NEL superfamily where family members have an LRR. The LRR has been involved in host-pathogen interaction (Kedzierski et al., 2004). Just PvRBSA-A bound when PvRBSA-A and -B recombinant protein fragments were assessed regarding their ability to bind to human reticulocytes (**Figure 4**). The binding percentage for PvRBSA-A was higher than that for the complete protein, which has been observed for other invasion-related *Plasmodium* parasite antigens (Chitnis and Miller, 1994; Fraser et al., 2001; Kato et al., 2005; Arevalo-Pinzon et al., 2017). An inhibitory assay was then performed to identify minimum specific regions within PvRBSA-A able to inhibit rPvRBSA-reticulocyte interaction. Two peptides (FPEIRENLTASEESLTSCEE and ESLAASRESLNDFCGSEESV) were able to decrease rPvRBSA-reticulocyte interaction. Both peptides were located in the PRK15370 putative domain, the first located towards the N-terminal domain whilst the other one was at the end of this. Remarkably, this region was found in the PvRBSA repeat region; repeats are usually used by the parasite as an immune evasion mechanism. Nevertheless, repeats could also be functionally important, as has been observed in the CS protein (Aldrich et al., 2012; Ferguson et al., 2014). The PvRBSA region involved in parasite-host interaction was thus located between amino acids 76 to 176 (numbers based on Sal-I protein sequences), FPEIRENLTASEESLTSCEE and ESLAASRESLNDFCGSEESV peptides (which were fully conserved in all *P. vivax* sequences analyzed here) being critical for binding, meaning that these peptides could be considered in vaccine development.

The aforementioned results thus suggest that natural selection is an evolutionary force modulating *pvrbsa* genetic diversity whilst negative selection acts at the gene's 5′-end, the 3′-end is under diversifying or balancing selection. Nonetheless, other forces could also be involved. Recombination might increase diversity by interchanging DNA fragments during sexual reproduction. LD decreased in *pvrbsa* as nucleotide distance increased (**Supplementary Data Sheet 2**), suggesting that recombination was/has been taking place in this gene. Recombination was confirmed by using ZZ, Rm and GARD method (**Supplementary Data Sheet 2**), suggesting that recombination can increase genetic diversity by intra-gene recombination.

A previous study has suggested that American populations are structured (Taylor et al., 2013); this pattern has also been observed in other *P. vivax* antigens in Colombia (Forero-Rodriguez et al., 2014a,b; Buitrago et al., 2016), hence migration was the last force evaluated here. The haplotype network inferred by using all available sequences did not have clear structuring; *pvrbsa* haplotypes were shared by different countries, even around the world. Several haplotypes found on the American continent were also present in Asia; consequently, most haplotypes should have arisen before *P. vivax* spread worldwide. Therefore, due to different *P. vivax* introductions to America by human migration (Rodrigues et al., 2018), *pvrbsa* haplotypes have reached different American counties. A structured population was not found when a haplotype network was inferred. In fact, several haplotypes were shared between Colombia and Venezuela, as well as amongst all subpopulations (**Figure 4**); however, AMOVA suggested that subpopulations were genetically different (**Table 5**). Statistically significant values were found for most comparisons (**Table 5**) when the fixation index (F_{ST} based on haplotype diversity) was computed. This could have been the consequence of limited gene flow amongst Colombian subpopulations or due to local adaptation. Synonymous and non-synonymous mutations were analysed independently to try to determine which event(s) led to this structure (**Supplementary Data Sheet 3**). Since synonymous mutations are typically silent, they are considered to follow neutral expectations and, therefore, they could represent the demographic history (migration). On the other hand, non-synonymous mutations are subject to natural selection and could consequently represent local adaptations. The synonymous data set showed that Bolívar and Venezuela's coastal area in Venezuela as well as the Chocó, Córdoba, Amazonas and Meta departments in Colombia seem to be genetically similar populations since non statistically significant F_{ST} were found. However, the non-synonymous data set showed that all (except Bolívar and Venezuela's coastal area as well as Chocó and Meta) subpopulations were genetically different. This suggested that local adaptation is responsible for the

observed structure. However, limited gene flow (partly) might also have provoked such structuring in Colombia. Nevertheless, migration in Venezuela seems to be/have been an important force modulating *pvrbsa* diversity. Further analysis using neutral markers within Colombia and Venezuela populations could confirm this issue.

CONCLUSION

Although the RBSA protein has previously been identified as an antigen exclusive to *Plasmodium* species invading reticulocytes (Moreno-Perez et al., 2017), it is actually present in several monkey-malaria lineage species. The encoding gene should have been present in the last monkey-malaria parasite common ancestor and it then became lost in some species (i.e. *P. knowlesi* and *P. coatneyi*). An independent duplication event took place in *P. vivax* and *P. cynomolgi*. The *pvrbsa* paralog appears to be a pseudogene whilst the *pcrbsa* paralog is a functional gene having just one mutation between *pcrbsa* and *pcrbsap*.

The *pvrbsa* locus has lower genetic diversity ($\pi = 0.008$) than other Mrz surface proteins (Mascorro et al., 2005; Valderrama-Aguirre et al., 2011; Garzon-Ospina et al., 2012, 2014); this diversity is modulated by natural selection, recombination and migration (the latter for Venezuela but not for Colombia). According to Tajima, Fu and Li, K_N-K_S and codon-based tests the RBSA's C-terminal end (PvRBSA-B) is under balancing (or diversifying) selection, likely due to this region being involved in immune response evasion whilst PvRBSA-A is under directional selection due to a functional/structural constraint ($\omega < 1$). The latter region has the PRK15370 putative domain (characterized by an LRR) and is involved in host-parasite interaction according to binding assays. Inhibition assays showed that two PRK15370 domain-derived peptides which were conserved in *P. vivax* isolates have been particularly involved in the specific interaction between PvRBSA and reticulocytes. Thus, these minimum regions could be considered when designing a fully effective anti-*P. vivax* vaccine.

AUTHOR CONTRIBUTIONS

PC-A performed the molecular evolutionary, recombinant expression and binding assays, and wrote the manuscript. DG-O devised and designed the study, performed the molecular evolutionary analysis, and wrote the manuscript. DM-P devised and designed the study, performed recombinant expression and binding assays, and wrote the manuscript. LR-C performed recombinant expression and binding assays and helped in writing the manuscript. ON and MP coordinated the study and helped to write the manuscript. All the authors have read and approved the final version of the manuscript.

FUNDING

This work was financed by the Departamento Administrativo de Ciencia, Tecnología e Innovación (COLCIENCIAS) through grant RC # 0309-2013.

ACKNOWLEDGMENTS

We would like to thank the Instituto de Ciencia, Biotecnología e Innovación en Salud (IDCBIS) in Bogotá for supplying the umbilical cord blood and Jason Garry for translating and reviewing the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2018.00372/full#supplementary-material

REFERENCES

Abascal, F., Zardoya, R., and Telford, M. J. (2010). TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. *Nucleic Acids Res.* 38(Web Server issue), W7–W13. doi: 10.1093/nar/gkq291

Aldrich, C., Magini, A., Emiliani, C., Dottorini, T., Bistoni, F., Crisanti, A., et al. (2012). Roles of the amino terminal region and repeat region of the *Plasmodium berghei* circumsporozoite protein in parasite infectivity. *PLoS One* 7:e32524. doi: 10.1371/journal.pone.0032524

Anisimova, M., Nielsen, R., and Yang, Z. (2003). Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. *Genetics* 164, 1229–1236.

Arenas, M., and Posada, D. (2010). Coalescent simulation of intracodon recombination. *Genetics* 184, 429–437. doi: 10.1534/genetics.109.109736

Arenas, M., and Posada, D. (2014). "The influence of recombination on the estimation of selection from coding sequence alignments," in *Natural Selection: Methods and Applications*, ed. M. A. Fares (Boca Raton, FL: CRC Press/Taylor & Francis), 112–125.

Arevalo-Pinzon, G., Bermudez, M., Curtidor, H., and Patarroyo, M. A. (2015). The *Plasmodium vivax* rhoptry neck protein 5 is expressed in the apical pole of *Plasmodium vivax* VCG-1 strain schizonts and binds to human reticulocytes. *Malar. J.* 14:106. doi: 10.1186/s12936-015-0619-1

Arevalo-Pinzon, G., Bermudez, M., Hernandez, D., Curtidor, H., and Patarroyo, M. A. (2017). *Plasmodium vivax* ligand-receptor interaction: PvAMA-1 domain I contains the minimal regions for specific interaction with CD71+ reticulocytes. *Sci. Rep.* 7:9616. doi: 10.1038/s41598-017-10025-6

Arevalo-Pinzon, G., Curtidor, H., Abril, J., and Patarroyo, M. A. (2013). Annotation and characterization of the *Plasmodium vivax* rhoptry neck protein 4 (PvRON4). *Malar. J.* 12:356. doi: 10.1186/1475-2875-12-356

Arevalo-Pinzon, G., Curtidor, H., Patino, L. C., and Patarroyo, M. A. (2011). PvRON2, a new *Plasmodium vivax* rhoptry neck antigen. *Malar. J.* 10:60. doi: 10.1186/1475-2875-10-60

Arnott, A., Barry, A. E., and Reeder, J. C. (2012). Understanding the population genetics of *Plasmodium vivax* is essential for malaria control and elimination. *Malar. J.* 11, 14. doi: 10.1186/1475-2875-11-14

Bandelt, H. J., Forster, P., and Rohl, A. (1999). Median-joining networks for inferring intraspecific phylogenies. *Mol. Biol. Evol.* 16, 37–48. doi: 10.1093/oxfordjournals.molbev.a026036

Baquero, L. A., Moreno-Perez, D. A., Garzon-Ospina, D., Forero-Rodriguez, J., Ortiz-Suarez, H. D., and Patarroyo, M. A. (2017). PvGAMA reticulocyte

binding activity: predicting conserved functional regions by natural selection analysis. *Parasit Vectors* 10:251. doi: 10.1186/s13071-017-2183-8

Barry, A. E., and Arnott, A. (2014). Strategies for designing and monitoring malaria vaccines targeting diverse antigens. *Front. Immunol.* 5:359. doi: 10.3389/fimmu.2014.00359

Buitrago, S. P., Garzon-Ospina, D., and Patarroyo, M. A. (2016). Size polymorphism and low sequence diversity in the locus encoding the *Plasmodium vivax* rhoptry neck protein 4 (PvRON4) in Colombian isolates. *Malar. J.* 15:501. doi: 10.1186/s12936-016-1563-4

Carabali-Isajar, M. L., Ocampo, M., Rodriguez, D. C., Vanegas, M., Curtidor, H., Patarroyo, M. A., et al. (2018). Towards designing a synthetic antituberculosis vaccine: The Rv3587c peptide inhibits mycobacterial entry to host cells. *Bioorg. Med. Chem.* 26, 2401–2409. doi: 10.1016/j.bmc.2018.03.044

Carlton, J. M., Adams, J. H., Silva, J. C., Bidwell, S. L., Lorenzi, H., Caler, E., et al. (2008). Comparative genomics of the neglected human malaria parasite *Plasmodium vivax*. *Nature* 455, 757–763. doi: 10.1038/nature07327

Carlton, J. M., Das, A., and Escalante, A. A. (2013). Genomics, population genetics and evolutionary history of *Plasmodium vivax*. *Adv. Parasitol.* 81, 203–222. doi: 10.1016/B978-0-12-407826-0.00005-9

Casillas, S., and Barbadilla, A. (2017). Molecular population genetics. *Genetics* 205, 1003–1035. doi: 10.1534/genetics.116.196493

Chan, E. R., Menard, D., David, P. H., Ratsimbasoa, A., Kim, S., Chim, P., et al. (2012). Whole genome sequencing of field isolates provides robust characterization of genetic diversity in *Plasmodium vivax*. *PLoS Negl. Trop. Dis.* 6:e1811. doi: 10.1371/journal.pntd.0001811

Chen, E., Salinas, N. D., Huang, Y., Ntumngia, F., Plasencia, M. D., Gross, M. L., et al. (2016). Broadly neutralizing epitopes in the *Plasmodium vivax* vaccine candidate duffy binding protein. *Proc. Natl. Acad. Sci. U.S.A.* 113, 6277–6282. doi: 10.1073/pnas.1600488113

Chen, J. H., Jung, J. W., Wang, Y., Ha, K. S., Lu, F., Lim, C. S., et al. (2010). Immunoproteomics profiling of blood stage *Plasmodium vivax* infection by high-throughput screening assays. *J. Proteome Res.* 9, 6479–6489. doi: 10.1021/pr100705g

Chitnis, C. E., and Miller, L. H. (1994). Identification of the erythrocyte binding domains of *Plasmodium vivax* and *Plasmodium knowlesi* proteins involved in erythrocyte invasion. *J. Exp. Med.* 180, 497–506. doi: 10.1084/jem.180.2.497

Delport, W., Poon, A. F., Frost, S. D., and Kosakovsky Pond, S. L. (2010). Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology. *Bioinformatics* 26, 2455–2457. doi: 10.1093/bioinformatics/btq429

Depaulis, F., and Veuille, M. (1998). Neutrality tests based on the distribution of haplotypes under an infinite-site model. *Mol. Biol. Evol.* 15, 1788–1790. doi: 10.1093/oxfordjournals.molbev.a025905

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32, 1792–1797. doi: 10.1093/nar/gkh340

Egea, R., Casillas, S., and Barbadilla, A. (2008). Standard and generalized McDonald-Kreitman test: a website to detect selection by comparing different classes of DNA sites. *Nucleic Acids Res.* 36, W157–W162. doi: 10.1093/nar/gkn337

Escalante, A. A., Cornejo, O. E., Freeland, D. E., Poe, A. C., Durrego, E., Collins, W. E., et al. (2005). A monkey's tale: the origin of *Plasmodium vivax* as a human malaria parasite. *Proc. Natl. Acad. Sci. U.S.A.* 102, 1980–1985. doi: 10.1073/pnas.0409652102

Excoffier, L., Laval, G., and Schneider, S. (2007). Arlequin (version 3.0): an integrated software package for population genetics data analysis. *Evol. Bioinform. Online* 1, 47–50.

Fay, J. C., and Wu, C. I. (2000). Hitchhiking under positive Darwinian selection. *Genetics* 155, 1405–1413.

Ferguson, D. J., Balaban, A. E., Patzewitz, E. M., Wall, R. J., Hopp, C. S., Poulin, B., et al. (2014). The repeat region of the circumsporozoite protein is critical for sporozoite formation and maturation in *Plasmodium*. *PLoS One* 9:e113923. doi: 10.1371/journal.pone.0113923

Forero-Rodriguez, J., Garzon-Ospina, D., and Patarroyo, M. A. (2014a). Low genetic diversity and functional constraint in loci encoding *Plasmodium vivax* P12 and P38 proteins in the Colombian population. *Malar. J.* 13, 58. doi: 10.1186/1475-2875-13-58

Forero-Rodriguez, J., Garzon-Ospina, D., and Patarroyo, M. A. (2014b). Low genetic diversity in the locus encoding the *Plasmodium vivax* P41 protein in Colombia's parasite population. *Malar. J.* 13:388. doi: 10.1186/1475-2875-13-388

Forster, P., Torroni, A., Renfrew, C., and Rohl, A. (2001). Phylogenetic star contraction applied to Asian and Papuan mtDNA evolution. *Mol. Biol. Evol.* 18, 1864–1881. doi: 10.1093/oxfordjournals.molbev.a003728

Fraser, T. S., Kappe, S. H., Narum, D. L., VanBuskirk, K. M., and Adams, J. H. (2001). Erythrocyte-binding activity of *Plasmodium yoelii* apical membrane antigen-1 expressed on the surface of transfected COS-7 cells. *Mol. Biochem. Parasitol.* 117, 49–59. doi: 10.1016/S0166-6851(01)00326-7

Fu, Y. X. (1997). Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics* 147, 915–925.

Fu, Y. X., and Li, W. H. (1993). Statistical tests of neutrality of mutations. *Genetics* 133, 693–709.

Galinski, M. R., and Barnwell, J. W. (2008). *Plasmodium vivax*: who cares? *Malar J.* 7(Suppl. 1), S9. doi: 10.1186/1475-2875-7-S1-S9

Galinski, M. R., Meyer, E. V., and Barnwell, J. W. (2013). *Plasmodium vivax*: modern strategies to study a persistent parasite's life cycle. *Adv. Parasitol.* 81, 1–26. doi: 10.1016/B978-0-12-407826-0.00001-1

Garzon-Ospina, D., Forero-Rodriguez, J., and Patarroyo, M. A. (2014). Heterogeneous genetic diversity pattern in *Plasmodium vivax* genes encoding merozoite surface proteins (MSP) -7E, -7F and -7L. *Malar. J.* 13, 495. doi: 10.1186/1475-2875-13-495

Garzon-Ospina, D., Forero-Rodriguez, J., and Patarroyo, M. A. (2015). Inferring natural selection signals in *Plasmodium vivax*-encoded proteins having a potential role in merozoite invasion. *Infect. Genet. Evol.* 33, 182–188. doi: 10.1016/j.meegid.2015.05.001

Garzon-Ospina, D., Forero-Rodriguez, J., and Patarroyo, M. A. (2016). Evidence of functional divergence in MSP7 paralogous proteins: a molecular-evolutionary and phylogenetic analysis. *BMC Evol. Biol.* 16:256. doi: 10.1186/s12862-016-0830-x

Garzon-Ospina, D., Lopez, C., Forero-Rodriguez, J., and Patarroyo, M. A. (2012). Genetic diversity and selection in three *Plasmodium vivax* merozoite surface protein 7 (Pvmsp-7) genes in a Colombian population. *PLoS One* 7:e45962. doi: 10.1371/journal.pone.0045962

Garzon-Ospina, D., Romero-Murillo, L., and Patarroyo, M. A. (2010). Limited genetic polymorphism of the *Plasmodium vivax* low molecular weight rhoptry protein complex in the Colombian population. *Infect. Genet. Evol.* 10, 261–267. doi: 10.1016/j.meegid.2009.12.004

Garzon-Ospina, D., Romero-Murillo, L., Tobon, L. F., and Patarroyo, M. A. (2011). Low genetic polymorphism of merozoite surface proteins 7 and 10 in Colombian *Plasmodium vivax* isolates. *Infect. Genet. Evol.* 11, 528–531. doi: 10.1016/j.meegid.2010.12.002

Gething, P. W., Elyazar, I. R., Moyes, C. L., Smith, D. L., Battle, K. E., Guerra, C. A., et al. (2012). A long neglected world malaria map: *Plasmodium vivax* endemicity in 2010. *PLoS Negl. Trop Dis.* 6:e1814. doi: 10.1371/journal.pntd.0001814

Graur, D., Zheng, Y., Price, N., Azevedo, R. B., Zufall, R. A., and Elhaik, E. (2013). On the immortality of television sets: "function" in the human genome according to the evolution-free gospel of ENCODE. *Genome Biol. Evol.* 5, 578–590. doi: 10.1093/gbe/evt028

Guerra, C. A., Howes, R. E., Patil, A. P., Gething, P. W., Van Boeckel, T. P., Temperley, W. H., et al. (2010). The international limits and population at risk of *Plasmodium vivax* transmission in 2009. *PLoS Negl. Trop Dis.* 4:e774. doi: 10.1371/journal.pntd.0000774

Gunasekera, A. M., Wickramarachchi, T., Neafsey, D. E., Ganguli, I., Perera, L., Premaratne, P. H., et al. (2007). Genetic diversity and selection at the *Plasmodium vivax* apical membrane antigen-1 (PvAMA-1) locus in a Sri Lankan population. *Mol. Biol. Evol.* 24, 939–947. doi: 10.1093/molbev/msm013

Hellmann, I., Ebersberger, I., Ptak, S. E., Paabo, S., and Przeworski, M. (2003). A neutral explanation for the correlation of diversity with recombination rates in humans. *Am. J. Hum. Genet.* 72, 1527–1535. doi: 10.1086/375657

Hester, J., Chan, E. R., Menard, D., Mercereau-Puijalon, O., Barnwell, J., Zimmerman, P. A., et al. (2013). De novo assembly of a field isolate genome reveals novel *Plasmodium vivax* erythrocyte invasion genes. *PLoS Negl. Trop Dis.* 7:e2569. doi: 10.1371/journal.pntd.0002569

Hudson, R. R., and Kaplan, N. L. (1985). Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics* 111, 147–164.

Huijben, S., and Paaijmans, K. P. (2018). Putting evolution in elimination: winning our ongoing battle with evolving malaria mosquitoes and parasites. *Evol. Appl.* 11, 415–430. doi: 10.1111/eva.12530

Hupalo, D. N., Luo, Z., Melnikov, A., Sutton, P. L., Rogov, P., Escalante, A., et al. (2016). Population genomics studies identify signatures of global dispersal and drug resistance in *Plasmodium vivax*. *Nat. Genet.* 48, 953–958. doi: 10.1038/ng.3588

Jukes, T. H., and Cantor, C. R. (1969). "Evolution of protein molecules," in *Mammalian Protein Metabolism*, ed. H. N. Munro (New York, NY: Academic Press). doi: 10.1016/B978-1-4832-3211-9.50009-7

Kato, K., Mayer, D. C., Singh, S., Reid, M., and Miller, L. H. (2005). Domain III of *Plasmodium falciparum* apical membrane antigen 1 binds to the erythrocyte membrane protein Kx. *Proc. Natl. Acad. Sci. U.S.A.* 102, 5552–5557. doi: 10.1073/pnas.0501594102

Kedzierski, L., Montgomery, J., Curtis, J., and Handman, E. (2004). Leucine-rich repeats in host-pathogen interactions. *Arch. Immunol. Ther. Exp. (Warsz)* 52, 104–112.

Kelly, J. K. (1997). A test of neutrality based on interlocus associations. *Genetics* 146, 1197–1206.

Kosakovsky Pond, S. L., and Frost, S. D. (2005). Not so different after all: a comparison of methods for detecting amino acid sites under selection. *Mol. Biol. Evol.* 22, 1208–1222. doi: 10.1093/molbev/msi105

Kosakovsky Pond, S. L., Posada, D., Gravenor, M. B., Woelk, C. H., and Frost, S. D. (2006). Automated phylogenetic detection of recombination using a genetic algorithm. *Mol. Biol. Evol.* 23, 1891–1901. doi: 10.1093/molbev/msl051

Kulathinal, R. J., Bennett, S. M., Fitzpatrick, C. L., and Noor, M. A. (2008). Fine-scale mapping of recombination rate in *Drosophila* refines its correlation to diversity and divergence. *Proc. Natl. Acad. Sci. U.S.A.* 105, 10051–10056. doi: 10.1073/pnas.0801848105

Librado, P., and Rozas, J. (2009). DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* 25, 1451–1452. doi: 10.1093/bioinformatics/btp187

Liu, W., Li, Y., Shaw, K. S., Learn, G. H., Plenderleith, L. J., Malenke, J. A., et al. (2014). African origin of the malaria parasite *Plasmodium vivax*. *Nat. Commun.* 5:3346. doi: 10.1038/ncomms4346

Malleret, B., Li, A., Zhang, R., Tan, K. S., Suwanarusk, R., Claser, C., et al. (2015). *Plasmodium vivax*: restricted tropism and rapid remodeling of CD71-positive reticulocytes. *Blood* 125, 1314–1324. doi: 10.1182/blood-2014-08-596015

Mascorro, C. N., Zhao, K., Khuntirat, B., Sattabongkot, J., Yan, G., Escalante, A. A., et al. (2005). Molecular evolution and intragenic recombination of the merozoite surface protein MSP-3alpha from the malaria parasite *Plasmodium vivax* in Thailand. *Parasitology* 131(Pt 1), 25–35. doi: 10.1017/S0031182005007547

Maxmen, A. (2012). Malaria surge feared. *Nature* 485:293. doi: 10.1038/485293a

McDonald, J. H., and Kreitman, M. (1991). Adaptive protein evolution at the Adh locus in *Drosophila*. *Nature* 351, 652–654. doi: 10.1038/351652a0

Moreno-Perez, D. A., Baquero, L. A., Chitiva-Ardila, D. M., and Patarroyo, M. A. (2017). Characterising PvRBSA: an exclusive protein from *Plasmodium* species infecting reticulocytes. *Parasit Vect.* 10:243. doi: 10.1186/s13071-017-2185-6

Moreno-Perez, D. A., Degano, R., Ibarrola, N., Muro, A., and Patarroyo, M. A. (2014). Determining the *Plasmodium vivax* VCG-1 strain blood stage proteome. *J. Proteomics* 113C, 268–280. doi: 10.1016/j.jprot.2014.10.003

Moreno-Perez, D. A., Saldarriaga, A., and Patarroyo, M. A. (2013). Characterizing PvARP, a novel *Plasmodium vivax* antigen. *Malar. J.* 12:165. doi: 10.1186/1475-2875-12-165

Mu, J., Joy, D. A., Duan, J., Huang, Y., Carlton, J., Walker, J., et al. (2005). Host switch leads to emergence of *Plasmodium vivax* malaria in humans. *Mol. Biol. Evol.* 22, 1686–1693. doi: 10.1093/molbev/msi160

Muehlenbein, M. P., Pacheco, M. A., Taylor, J. E., Prall, S. P., Ambu, L., Nathan, S., et al. (2015). Accelerated diversification of nonhuman primate malarias in Southeast Asia: adaptive radiation or geographic speciation? *Mol. Biol. Evol.* 32, 422–439. doi: 10.1093/molbev/msu310

Murrell, B., Moola, S., Mabona, A., Weighill, T., Sheward, D., Kosakovsky Pond, S. L., et al. (2013). FUBAR: a fast, unconstrained bayesian approximation for inferring selection. *Mol. Biol. Evol.* 30, 1196–1205. doi: 10.1093/molbev/mst030

Murrell, B., Wertheim, J. O., Moola, S., Weighill, T., Scheffler, K., and Kosakovsky Pond, S. L. (2012). Detecting individual sites subject to episodic diversifying selection. *PLoS Genet* 8:e1002764. doi: 10.1371/journal.pgen.1002764

Nanda Kumar, Y., Jeyakodi, G., Gunasekaran, K., and Jambulingam, P. (2016). Computational screening and characterization of putative vaccine candidates of *Plasmodium vivax*. *J. Biomol. Struct. Dyn.* 34, 1736–1750. doi: 10.1080/07391102.2015.1090344

Narayanan, K., Ramachandran, A., Hao, J., He, G., Park, K. W., Cho, M., et al. (2003). Dual functional roles of dentin matrix protein 1, implications in biomineralization and gene transcription by activation of intracellular Ca2+ store. *J. Biol. Chem.* 278, 17500–17508. doi: 10.1074/jbc.M212700200

Neafsey, D. E., Galinsky, K., Jiang, R. H., Young, L., Sykes, S. M., Saif, S., et al. (2012). The malaria parasite *Plasmodium vivax* exhibits greater genetic diversity than *Plasmodium falciparum*. *Nat. Genet.* 44, 1046–1050. doi: 10.1038/ng.2373

Nei, M., and Rooney, A. P. (2005). Concerted and birth-and-death evolution of multigene families. *Annu. Rev. Genet.* 39, 121–152. doi: 10.1146/annurev.genet.39.073003.112240

Nobrega de Sousa, T., Carvalho, L. H., and Alves de Brito, C. F. (2011). Worldwide genetic variability of the Duffy binding protein: insights into *Plasmodium vivax* vaccine development. *PLoS One* 6:e22944. doi: 10.1371/journal.pone.0022944

Ntumngia, F. B., Pires, C. V., Barnes, S. J., George, M. T., Thomson-Luque, R., Kano, F. S., et al. (2017). An engineered vaccine of the *Plasmodium vivax* Duffy binding protein enhances induction of broadly neutralizing antibodies. *Sci. Rep.* 7:13779. doi: 10.1038/s41598-017-13891-2

Pacheco, M. A., Battistuzzi, F. U., Junge, R. E., Cornejo, O. E., Williams, C. V., Landau, I., et al. (2011). Timing the origin of human malarias: the lemur puzzle. *BMC Evol. Biol.* 11:299. doi: 10.1186/1471-2148-11-299

Pacheco, M. A., Elango, A. P., Rahman, A. A., Fisher, D., Collins, W. E., Barnwell, J. W., et al. (2012). Evidence of purifying selection on merozoite surface protein 8 (MSP8) and 10 (MSP10) in *Plasmodium* spp. *Infect. Genet. Evol.* 12, 978–986. doi: 10.1016/j.meegid.2012.02.009

Pacheco, M. A., Ryan, E. M., Poe, A. C., Basco, L., Udhayakumar, V., Collins, W. E., et al. (2010). Evidence for negative selection on the gene encoding rhoptry-associated protein 1 (RAP-1) in *Plasmodium* spp. *Infect. Genet. Evol.* 10, 655–661. doi: 10.1016/j.meegid.2010.03.013

Patarroyo, M. A., Calderon, D., and Moreno-Perez, D. A. (2012). Vaccines against *Plasmodium vivax*: a research challenge. *Exp. Rev. Vaccines* 11, 1249–1260. doi: 10.1586/erv.12.91

Perez-Losada, M., Arenas, M., Galan, J. C., Palero, F., and Gonzalez-Candelas, F. (2015). Recombination in viruses: mechanisms, methods of study, and evolutionary consequences. *Infect. Genet. Evol.* 30, 296–307. doi: 10.1016/j.meegid.2014.12.022

Pond, S. L., Frost, S. D., Grossman, Z., Gravenor, M. B., Richman, D. D., and Brown, A. J. (2006). Adaptation to different human populations by HIV-1 revealed by codon-based analyses. *PLoS Comput. Biol.* 2:e62. doi: 10.1371/journal.pcbi.0020062

Posada, D. (2008). jModelTest: phylogenetic model averaging. *Mol. Biol. Evol.* 25, 1253–1256. doi: 10.1093/molbev/msn083

Price, R. N., von Seidlein, L., Valecha, N., Nosten, F., Baird, J. K., and White, N. J. (2014). Global extent of chloroquine-resistant *Plasmodium vivax*: a systematic review and meta-analysis. *Lancet Infect. Dis.* 14, 982–991. doi: 10.1016/S1473-3099(14)70855-2

Rao, Y., Sun, L., Nie, Q., and Zhang, X. (2011). The influence of recombination on SNP diversity in chickens. *Hereditas* 148, 63–69. doi: 10.1111/j.1601-5223.2010.02210.x

Richie, T. L., and Saul, A. (2002). Progress and challenges for malaria vaccines. *Nature* 415, 694–701. doi: 10.1038/415694a

Rodrigues, P. T., Valdivia, H. O., de Oliveira, T. C., Alves, J. M. P., Duarte, A., Cerutti-Junior, C., et al. (2018). Human migration and the spread of malaria parasites to the New World. *Sci. Rep.* 8:1993. doi: 10.1038/s41598-018-19554-0

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D. L., Darling, A., Hohna, S., et al. (2012). MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* 61, 539–542. doi: 10.1093/sysbio/sys029

Rozas, J., Gullaud, M., Blandin, G., and Aguade, M. (2001). DNA variation at the rp49 gene region of *Drosophila* simulans: evolutionary inferences from an unusual haplotype structure. *Genetics* 158, 1147–1155.

Smith, M. D., Wertheim, J. O., Weaver, S., Murrell, B., Scheffler, K., and Kosakovsky Pond, S. L. (2015). Less is more: an adaptive branch-site random effects model for efficient detection of episodic diversifying selection. *Mol. Biol. Evol.* 32, 1342–1353. doi: 10.1093/molbev/msv022

Suh, K. N., Kain, K. C., and Keystone, J. S. (2004). Malaria. *CMAJ* 170, 1693–1702. doi: 10.1503/cmaj.1030418

Tachibana, S., Sullivan, S. A., Kawai, S., Nakamura, S., Kim, H. R., Goto, N., et al. (2012). *Plasmodium cynomolgi* genome sequences provide insight into Plasmodium vivax and the monkey malaria clade. *Nat. Genet.* 44, 1051–1055. doi: 10.1038/ng.2375

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123, 585–595.

Tamura, K., Stecher, G., Peterson, D., Filipski, A., and Kumar, S. (2013). MEGA6: molecular evolutionary genetics analysis version 6.0. *Mol. Biol. Evol.* 30, 2725–2729. doi: 10.1093/molbev/mst197

Taylor, J. E., Pacheco, M. A., Bacon, D. J., Beg, M. A., Machado, R. L., Fairhurst, R. M., et al. (2013). The evolutionary history of *Plasmodium vivax* as inferred from mitochondrial genomes: parasite genetic diversity in the Americas. *Mol. Biol. Evol.* 30, 2050–2064. doi: 10.1093/molbev/mst104

The malERA Consultative Group on Vaccines (2011). A research agenda for malaria eradication: vaccines. *PLoS Med* 8:e1000398. doi: 10.1371/journal.pmed.1000398

Valderrama-Aguirre, A., Zuniga-Soto, E., Marino-Ramirez, L., Moreno, L. A., Escalante, A. A., Arevalo-Herrera, M., et al. (2011). Polymorphism of the Pv200L fragment of merozoite surface protein-1 of *Plasmodium vivax* in clinical isolates from the Pacific coast of Colombia. *Am. J. Trop. Med. Hyg.* 84, 64–70. doi: 10.4269/ajtmh.2011.09-0517

Valencia, S. H., Rodriguez, D. C., Acero, D. L., Ocampo, V., and Arevalo-Herrera, M. (2011). Platform for *Plasmodium vivax* vaccine discovery and development. *Mem. Inst. Oswaldo Cruz* 106(Suppl 1), 179–192. doi: 10.1590/S0074-02762011000900023

Wang, Y., Ma, A., Chen, S. B., Yang, Y. C., Chen, J. H., and Yin, M. B. (2014). Genetic diversity and natural selection of three blood-stage 6-Cys proteins in *Plasmodium vivax* populations from the China-Myanmar endemic border. *Infect. Genet. Evol.* 28, 167–174. doi: 10.1016/j.meegid.2014.09.026

Welsh, K., and Bunce, M. (1999). Molecular typing for the MHC with PCR-SSP. *Rev. Immunogenet.* 1, 157–176.

White, N. J., Pukrittayakamee, S., Hien, T. T., Faiz, M. A., Mokuolu, O. A., and Dondorp, A. M. (2014). Malaria. *Lancet* 383, 723–735. doi: 10.1016/S0140-6736(13)60024-0

WHO (2017). *World Malaria Report 2017*. Geneva: World Health Organization.

Zhang, J., Rosenberg, H. F., and Nei, M. (1998). Positive darwinian selection after gene duplication in primate ribonuclease genes. *Proc. Natl. Acad. Sci. U.S.A.* 95, 3708–3713. doi: 10.1073/pnas.95.7.3708

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.