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

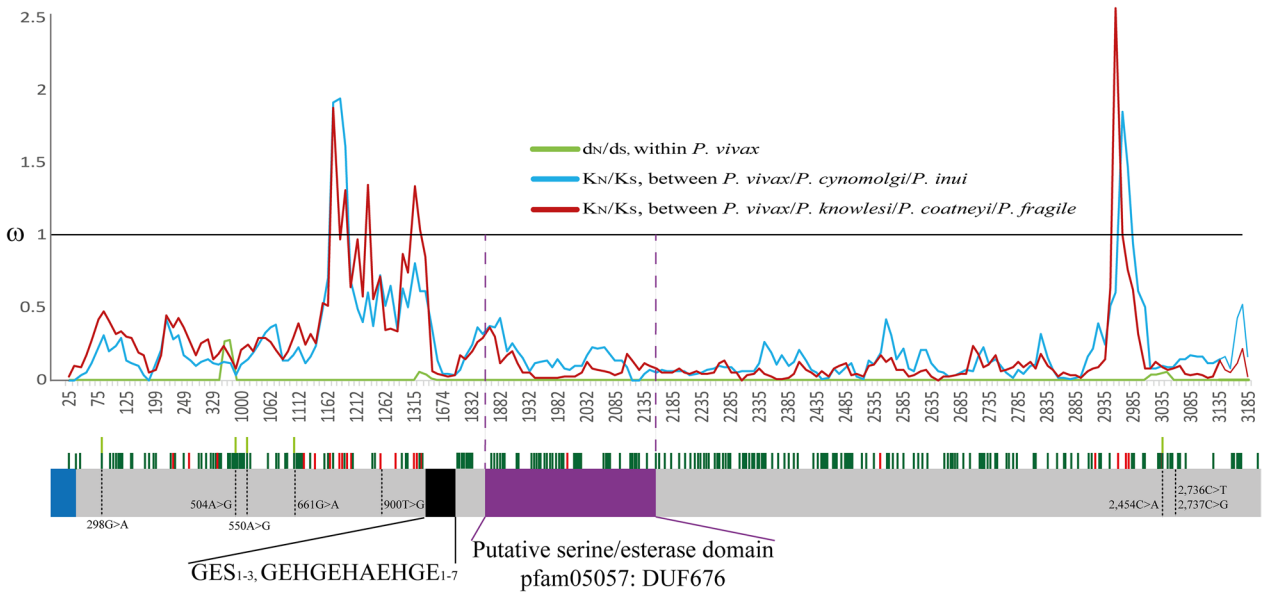

Fig. 1 Representation of *pvron4* encoding DNA sequence and sliding window for ω rate. *ron4* ω values (d_N/d_S) within *P. vivax* are shown in *green* throughout. Likewise, ω divergence rates (K_N/K_S) between the species infecting primates and *P. vivax* are shown in *blue* (for the species evolutionarily 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.

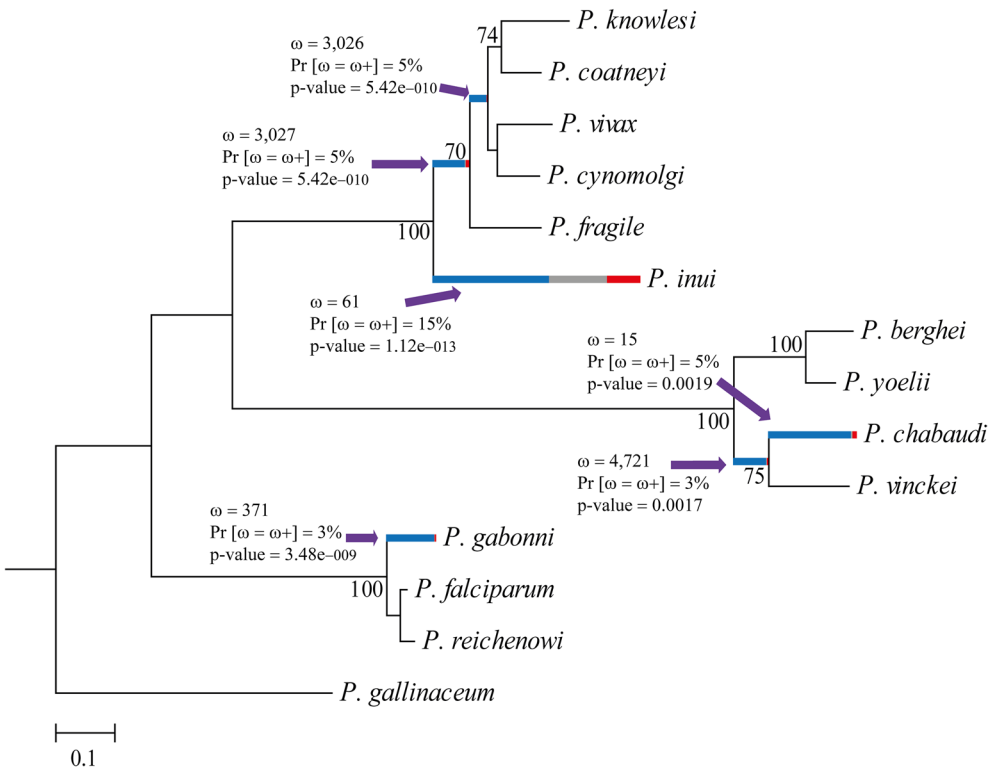

Fig. 2 Lineages subject to episodic diversifying selection in *ron4*. The branches (lineages) under episodic positive selection were identified by the Branch-site REL method. The figure shows the ω^+ values, the percentage of sites (Pr [ω = ω+]) and p values. The shade of each *colour* on branches indicates 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

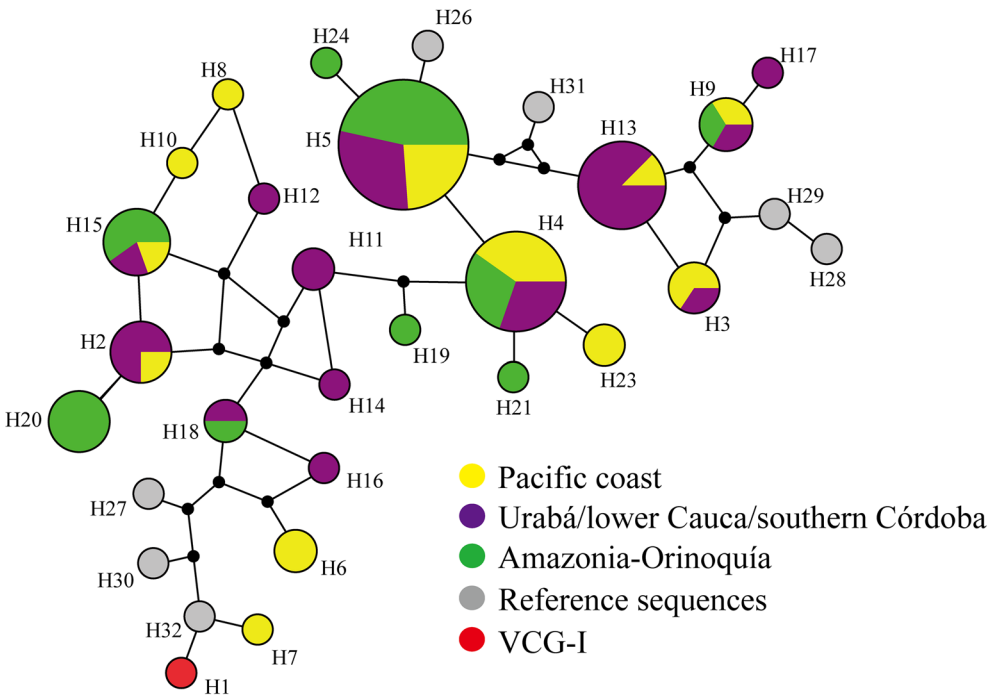

Fig. 3 Median-joining network for Colombian regions. The *Figure* shows the *pvron4* haplotypes identified from the isolates from the three regions of Colombia. Haplotypes 22 and 28 were included within haplotype 15 using the contraction star algorithm [86] for simplifying interpretation of the network. 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.