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.

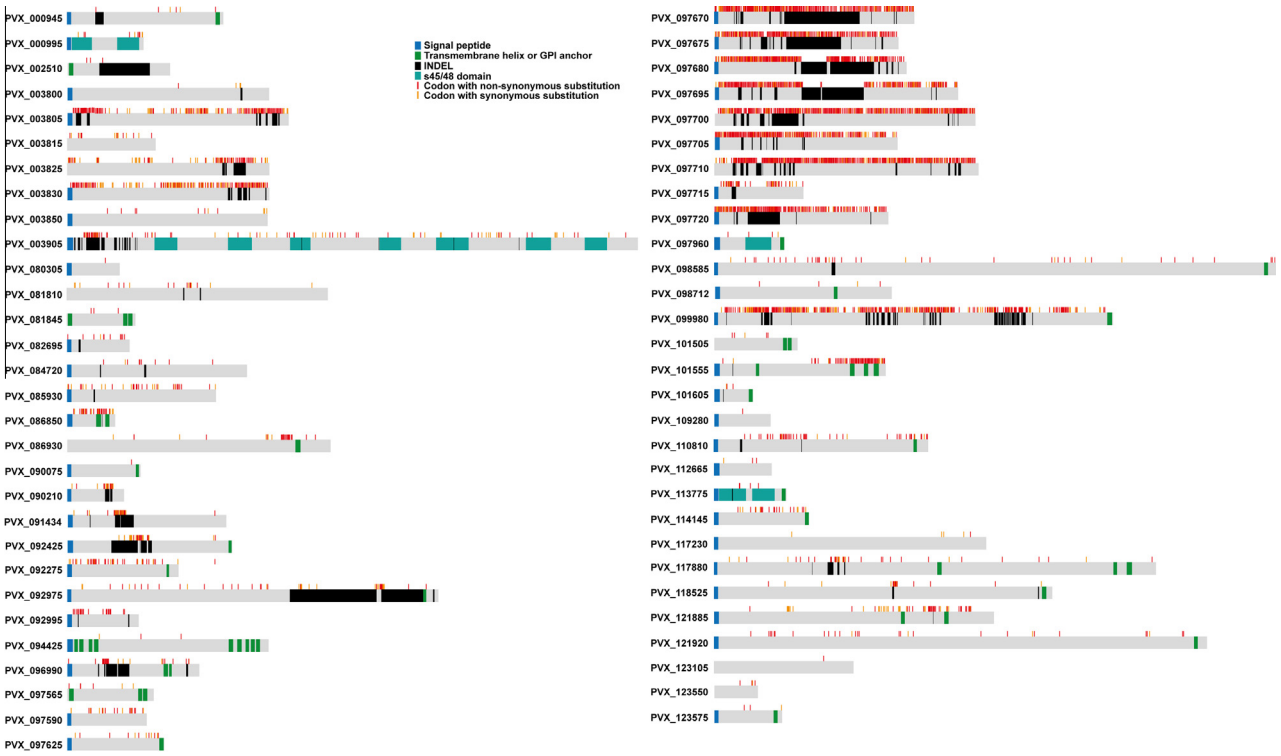

Fig. 1. Polymorphism distribution within the 59 *Plasmodium vivax* genes studied here. Codons having non-synonymous (red) and synonymous (orange) mutations are shown with vertical lines above each gene. Signal peptide-encoding regions (blue), predicted transmembrane helices or GPI anchors (green), s48/45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