Genetic Diversity and Selection in Three *Plasmodium vivax* Merozoite Surface Protein 7 (*Pvmsp-7*) Genes in a Colombian Population

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

1 Fundación Instituto de Inmunología de Colombia – FIDIC, Bogotá DC, Colombia, **2** Microbiology postgraduate program, Universidad Nacional de Colombia, Bogotá DC, Colombia, **3** School of Medicine and Health Sciences, Universidad del Rosario, Bogotá DC, Colombia

Abstract

A completely effective vaccine for malaria (one of the major infectious diseases worldwide) is not yet available; different membrane proteins involved in parasite-host interactions have been proposed as candidates for designing it. It has been found that proteins encoded by the merozoite surface protein (*msp*)-7 multigene family are antibody targets in natural infection; the nucleotide diversity of three *Pvmsp-7* genes was thus analyzed in a Colombian parasite population. By contrast with *P. falciparum msp-7* loci and ancestral *P. vivax msp-7* genes, specie-specific duplicates of the latter specie display high genetic variability, generated by single nucleotide polymorphisms, repeat regions, and recombination. At least three major allele types are present in *Pvmsp-7C*, *Pvmsp-7H* and *Pvmsp-7I* and positive selection seems to be operating on the central region of these *msp-7* genes. Although this region has high genetic polymorphism, the C-terminus (Pfam domain ID: PF12948) is conserved and could be an important candidate when designing a subunit-based antimalarial vaccine.

Citation: Garzón-Ospina D, López C, Forero-Rodríguez J, Patarroyo MA (2012) Genetic Diversity and Selection in Three *Plasmodium vivax* Merozoite Surface Protein 7 (*Pvmsp-7*) Genes in a Colombian Population. PLoS ONE 7(9): e45962. doi:10.1371/journal.pone.0045962

Editor: Erika Martins Braga, Universidade Federal de Minas Gerais, Brazil

Received February 27, 2012; **Accepted** August 28, 2012; **Published** September 25, 2012

Funding: The authors have no support or funding to report.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: mapatarr.fidic@gmail.com

Introduction

Plasmodium vivax malaria should not be considered a benign disease anymore due to its wide geographical distribution [1,2] and ability to cause severe clinical syndromes [3,4,5,6]. It was estimated that 2.85 billion people were at risk of being infected by this parasite in 2009 [1]. Although several research groups have focused on developing a vaccine against *P. vivax*, relatively few vaccine candidate molecules have been thoroughly characterized [7,8]; some of them are membrane proteins making attractive targets for vaccine development since they are essential for the recognition and invasion of erythrocytes, hence immune responses directed against them could inhibit parasite entry [9]. The merozoite surface protein (MSP) family is a group of surface proteins which are involved in the initial interaction between the parasite and the host cell. However, the genetic diversity of these proteins has been the main problem for developing a vaccine [10,11,12,13,14]. Proteins displaying high antigenic diversity elicit strain-specific immune responses leading to low protective efficacy upon vaccination, while proteins displaying limited variability are attractive targets for testing as vaccine candidates. Research involving proteins displaying high overall polymorphism should be focused on the conserved functional domains [15] so that strain-specific responses may be avoided.

MSP-1, MSP-6 and MSP-7 form the most abundant protein complex on the *P. falciparum* merozoite surface [16,17]. *msp-1* and *msp-6* are both single copy genes while *msp-7* has a large number of genes which seem not to have diverged functionally. Furthermore, the *Plasmodium* genus has dissimilar copy numbers of this gene [18,19]. The *msp-7* genes have a single exon and have been named in alphabetical order according to their position regarding the PVX_082640 flanking gene [18]. Different *msp-7* genes are expressed simultaneously in the schizont stage of *Plasmodium* species [20,21,22] and several of them have been localized on the membrane surface [19,21,22,23], not only forming part of the main merozoite protein complex but also binding to erythrocytes [24]. Knockout and invasion inhibition assays have also shown that MSP-7 is involved in the invasion of erythrocytes [19,24,25]. Moreover, Wang's group has shown that a recombinant MSP-7 is recognized by sera from malaria-infected individuals where IgG3 subclass antibodies prevail [26]. Immunization with members of this family has been shown to protect vaccinated mice against experimental challenge [21].

P. falciparum has eight *msp-7* genes and *P. vivax* eleven, [18,27]; the members of this family have low genetic variability in the former specie, [28,29] as do the *P. vivax msp-7A* (GenBank ID: XM_001614080.1) and *msp-7K* (GenBank ID: XM_001614090.1) genes [30]. A recent study has shown that *P. vivax* specie-specific duplicates MSP-7C (GenBank ID: XM_001614082.1) and MSP-7H (GenBank ID: XM_001614087.1) are recognized by IgG antibodies from *P. vivax*-infected patient sera [31]. These proteins are phylogenetically related to PvMSP-7I (GenBank ID: XM_001614088.1) [18,27] and fragments from different MSP-7 proteins may have to be included to block this invasion route due to the biological implication that functional redundancy might have on vaccine development. Taking into account that some of these proteins are recognized by the immune system, their genetic diversity should be evaluated to determine their potential use as potent anti-malarial vaccine candidates.

Results

Forty-eight parasite samples were collected from symptomatic patients living in representative regions of Colombia for this study: Amazonian (Calamar, Guaviare department), Andean (Apartadó and El Bagre, Antioquia department), Caribbean (Tierra Alta and Puerto Libertador, Córdoba department), Orinoquia (Mapiripan, Meta department and Tauramena, Casanare department) and Pacific (Istmina, Chocó department and Tumaco, Nariño department) (Fig. S1). Forty-two samples corresponding to single *P. vivax msp-1* allele infections were considered for PCR amplification even though amplicons were not detected in a few samples (*Pvmsp-7C* n = 37, *Pvmsp-7H* n = 37 and *Pvmsp-7I* n = 42).

Polymorphism in *Pvmsp-7* Loci

1,167-bp *Pvmsp-7C*, 1,232-bp *Pvmsp-7H* and 1,182-bp *Pvmsp-7I* gene fragments were amplified and direct sequenced. Nucleotide sequence data here reported are available in GenBank: accession numbers JQ423957-JQ424037. Twenty-three haplotypes were found for *Pvmsp-7C* while twenty-eight haplotypes were found for both *Pvmsp-7H* and *Pvmsp-7I.* Haplotype diversity (Hd) was lower in the *Pvmsp-7C* gene than in *Pvmsp-7H* and *Pvmsp-7I* (Table 1). 148 sites from the total nucleotide sequence length analyzed were polymorphic in *Pvmsp-7C* and 142 sites in *Pvmsp-7H,* while 121 polymorphic sites were found in *Pvmsp-7I* (Table 1).

Nucleotide diversity for these three genes (*Pvmsp-7C* $\pi = 0.0548$, *Pvmsp-7H* $\pi = 0.0357$ and *Pvmsp-7I* $\pi = 0.0430$) was higher than that previously reported for other *P. vivax msp-7* family members (*Pvmsp-7A* $\pi = 0$ and *Pvmsp-7K* $\pi = 0.0022$) [30]. Furthermore, the *msp-7* genes from this study were among the most polymorphic *P. vivax* genes reported to date (Table S1). Regarding *Pvmsp-7* diversity (π value average) in the different Colombian regions, the Pacific area was the most diverse followed by the Andean, Caribbean and Orinoco regions while the Amazonian was the least polymorphic (Table S2).

Three major allele types were found in deduced PvMSP-7C amino acid sequences (Fig. 1 and Fig. S2) between residues 134 and 238 (numbered according to the alignment in Fig. S2). Allele types had similar frequency in the Colombian parasite population.

An AEAFG insertion-deletion from residue 146 to 150 and a $GTG_{(n)}GT[V/E]$ repeat region from residue 195 to 209 (numbered according to the alignment in Fig. S3) were revealed by protein sequence alignment in PvMSP-7H. Three regions throughout the protein sequences led to discriminating several allele types (Fig. 2 and Fig. S3). The first region from residue 153 to 170 (numbered according to the alignment given in the Fig. S3) had four different peptide sequences. Five different sequences were found in the second region (from residue 172 to 194) and three different ones were found from residue 229 to 247. These three regions were randomly associated (Fig. 2 and Fig. S3).

EEAVEGD and EA repeats could be observed in the deduced PvMSP-7I amino acid sequences. Similar to the genes mentioned above, several major allele types were observed between residues 131 and 219 (numbered according to the alignment in Fig. S4) characterized by two regions. The first region (between amino acids 131 and 140) had three different peptide sequences (Fig. 3 and Fig. S4). The second region ran from residue 163 to residue 219, having an extra four different peptide sequences. These regions were not randomly associated (Fig. 3 and Fig. S4). A further four peptide sequences were found from residue 221 to residue 234 and two more from residue 236 to residue 264 and, contrary to those mentioned above, these fragments were found to be associated with either of the alleles described above (Fig. 3 and Fig. S4).

Neutral Evolutionary Test and Selection in the *Pvmsp-7* Loci

Applying neutral evolution tests (Tajima, Fu & Li) to the Colombian *P. vivax* population gave significant values above 0 for *Pvmsp-7C* and *Pvmsp-7I* ($p<0.05$ Tajima D, $p<0.02$ Fu & Li D* and F*) (Table 1); these values indicated an excess of intermediate frequency alleles. The neutral evolution tests for *Pvmsp-7H* revealed no statistically significant differences; this gene therefore seems to be evolving under neutral expectations. However, a sliding window analysis for D, D* and F* statistics showed that different selective forces could be acting throughout the *Pvmsp-7* genes. It was found that balancing selection acted on the central gene region while negative selection could be acting at the 5′-ends (significant values for *Pvmsp-7C* and *Pvmsp-7H*) and 3′-ends (no significant values) (Fig. S5).

The average number of synonymous substitutions per synonymous site (d_S) and non-synonymous substitutions per non-synonymous site (d_N) was estimated to determine whether natural selection was affecting the *msp-7* loci. Although d_N was higher than d_S in *Pvmsp-7C* and *Pvmsp-7I,* d_N was lower than d_S in the *Pvmsp-7H* gene even though these values were not statistically significant (Table 2). Taking into account that a sliding window for ω (d_N/d_S) (Fig. S6) showed that the ω rate was higher than 1 (signal of positive selection) in the central region of the *msp-7* genes analyzed

Table 1. Estimates of DNA diversity and neutrality test for *Pvmsp-7* genes from a Colombian population.

n	Gene	Sites	Ss	S	Ps	H	Hd (se)	θ^{w} (se)	π (se)	Tajima	Fu & Li		Fu	Z_{ns}	ZZ	RM
										D	D*	F*	Fs			
37	*msp-7C*	1,098	148	1	147	23	0.93 (0.03)	0.0323 (0.001)	0.0548 (0.003)	2.094**	1.810*	2.287*	7.64	0.339	0.173*	22
37	*msp-7H*	1,131	142	19	123	28	0.96 (0.02)	0.0301 (0.002)	0.0357 (0.003)	0.388	0.952	0.987	-0.57	0.126	0.343*	27
42	*msp-7I*	1,058	127	6	121	28	0.97 (0.01)	0.0280 (0.002)	0.0430 (0.003)	1.420	1.584**	1.818*	2.12	0.232	0.334*	13

Ss: Number of segregating sites, S: number of singleton sites, Ps: number of parsimony-informative sites, H: number of haplotypes, Hd: haplotype diversity, θ^{W}: Watterson estimator, π: nucleotide diversity. (se): Standard deviation. Sites excluded from the analysis: *Pvmsp-7C*: nucleotides 1 to 17 (amino acids 1 to 6), 544 to 546 (amino acid 182), 616 to 618 (amino acid 206), 625 to 627 (amino acid 209), 631 to 648 (amino acids 211 to 216), 655 to 666 (amino acids 219 to 222), 676 to 684 (amino acids 226 to 228), 706 to 708 (amino acid 236), 1,105 to 1,107 (amino acid 369) and nucleotides 1,168 to 1,191 (amino acids 390 to 397); *Pvmsp-7H*: nucleotides 436 to 450 (amino acids 146 to 150), 448 to 486 (amino acid 162), 568 to 627 (amino acids 190 to 200) and nucleotides 772 to 774 (amino acid 258); *Pvmsp-7I*: nucleotides 1 to 18 (amino acids 1 to 6), 421 to 522 (amino acids 141 to 174) and nucleotides 526 to 537 (amino acids 176 to 179).
*: $p<0.02$, **: $p<0.05$.
doi:10.1371/journal.pone.0045962.t001

```
      129                                                                                                                244
      |                                                                                                                    |
Sal-I KGQADTAPSV KGDVSPPPNL PAAAASSPKE TVPAGTSNGL VEADYVVLNT PDGNPRPVGP GGGSRPSASG PDAASNL-QN -Q--ATAEAG G---STN--T QGSQTGGVST TPGANQ
AND5  .......... .......... .......... .......... .......... .....G.... .......... .......-.. -P--...... .---...--. .......... ......
PAC2  .....AGQT. ESN...R.AS S..DN.L..K .T.....S.V ..VR..NP.S ..S-.SDALS .......SQ. .GTSP.V-.. -.------S. ----NSQ--- AAAEA..STN ......
CAR9  .....ASQ.. ..ADT.GSK. ....D.P.RG .AAD.RNSHV ..IG.INR.S A.SS.LAA.S ..D.TL.... .GS..QIT.P SPSSPGGVP. NTLTNVQPQ. PVGAG..-TN ......
```

Figure 1. Alignment of deduced PvMSP-7C amino acid sequences, including the Sal-I sequence in which the three main allele types can be observed. Dots indicate conserved residues and dashes represent gaps introduced for alignment.
doi:10.1371/journal.pone.0045962.g001

here, the genes were thus split into 3 regions: 5′-end (*msp-7C*: between nucleotides 1 and 390, *msp-7H*: between nucleotides 1 and 471, *msp-7I*: between nucleotides 1 and 525), central (*msp-7C*: between nucleotides 391 and 717, *msp-7H*: between nucleotides 472 and 771, *msp-7I*: between nucleotides 526 and 789) and 3′-end (*msp-7C*: between nucleotide 718 and 1,191, *msp-7H*: between nucleotide 772 and 1,200, *msp-7I*: between nucleotide 790 and 1,188) (Table S3); d_N and d_S were then computed. d_S was higher than d_N for both 5′- and 3′-ends, just the 5′-end having significant statistical difference. On the other hand, the d_N substitutions in the central region were significantly higher than d_S substitutions for the three genes (Table 2).

Furthermore, several maximum likelihood-based methods were used for identifying which codon sites were under positive or negative selection. *Pvmsp-7C* had 19 sites under positive selection and 21 negatively selected sites according to SCAL, FEL, IFEL and REL methods. Positive selection signatures were found for *Pvmsp-7H* in 50 sites while another 32 sites were negatively selected and *Pvmsp-7I* had 10 sites under positive selection and 30 negatively selected sites (predicted by at least one method, Tables S4 and S5). Taking into account that recombination (see below) can affect the reliability of identifying sites under selection [32], the analysis was performed again, this time considering recombination. Two sites were thus positively selected and 43 were under negative selection for *Pvmsp-7C*. Ten sites were positively selected and 24 were negatively selected for *Pvmsp-7H*, and 4 sites were under positive selection while 13 sites were under negative selection for *Pvmsp-7I* (Tables S6 and S7).

Linkage Disequilibrium (LD) and Recombination

Random association was observed among *Pvmsp-7* haplotypes, therefore suggesting their independent segregation. LD analysis measured by Z_{nS} (average of $R^{\wedge 2}$) for whole data revealed no statistically significant values for the three *msp-7* genes (Table 1). The relationship of $R^{\wedge 2}$ (linkage disequilibrium) with physical distance in regression analysis showed that LD declined as nucleotide distance increased (Fig. S7). The ZZ statistic had significant values (Table 1). Therefore, both the regression analysis and ZZ statistic suggested that intragenic recombination was taking place in *Pvmsp-7C*, *Pvmsp-7H* and *Pvmsp-7I*. Recombination events were detected by using DnaSP which revealed several RM events in all genes (Table 1). Figure 4 shows the recombinant regions detected with RDP v3.4 software.

Clustering Analysis

Since recombination was detected for *msp-7* genes, phylogeny had to be inferred taking it into account. A phylogenetic tree was inferred for each recombinant region detected (Fig. S8, S9, S10). All topologies showed no geographical clustering among the Colombian isolates; moreover, Colombian and Salvadorian sequences clustered together (Fig. S8, S9, S10).

All *Pvmsp-7* sequences were aligned and trees were constructed by using ML with the TN93+*G* model. The tree showed three monophyletic groups (Fig. S11A); the first group clustered *Pvmsp-7H* sequences, close to the second group which clustered *Pvmsp-7I* sequences, while the third group clustered *Pvmsp-7C* sequences. This topology agreed with the phylogenetic relationship previously reported for *Pvmsp-7* [18,27]. However, *Pvmsp-7* sequences showed recombinant fragments possibly produced by gene conversion between paralogous genes at the 5′ and 3′-ends but not in the central region (Fig. S12 and Table S8). The *Pvmsp-7* genes became clustered into paraphyletic groups when trees were inferred taking gene conversion into account (Fig. S11B–E). Codons 38 and 112 in the three genes, as well as the 57, 70, 102 and 125 for *Pvmsp-7C* and *Pvmsp-7H*, and codons 48 and 86 for *Pvmsp-7H* and *Pvmsp-7I* (numbered according to the alignment in Fig. S13) were predicted as negatively selected sites but, taking into account that they lay within the gene conversion tracks detected, purifying selection at these sites could have been the result of gene conversion between paralogous gene loci.

The d_S and d_N rates between paralogous gene pairs were estimated to determine whether homogeneity at the *Pvmsp-7* ends

```
      151                                                                                              251
      |                                                                                                  |
Sal-I GGVPVTGNSASNSQSTGGSGSQNASPPQGSPSDSAQGSQVTNST--------GSTVTLNAPSSSHSTGQPQQSAGVSLPTGTAETVASNTAQTSPPAG
PAC2  ...T.....................Q.E.N.GGNP..T.....A--------...G..STS...Q....S...N.AEP.A..TQE.NT.AG.PP....
CAR2  ............................................--------..............................................
ORI3  ...T......T-.PPS..G..KSP.Q..DN........P-----GTGGTGGT......STS.....................................
AMA2  ...T......T-.PPS..G..ESP.Q..DN.GGG....P---..--GGTGGT.G....STS..GP...GS.P....APSV.N..A.VT..EHS.....
VCG-I ...T.......-.....DLN...P.Q.E.NNGGG....P-----GTGGTGGT.G....STS...P...GS.P..........................
AND3  ...........................E.NNGG.E..P...-----------EG.V...T...GP...GS.P..........................
AMA3  ..........................T....GGNP..T.....A--------...G..STS...Q....S.P....APSV.N..A.VT..EHS.....
AND11 ..A........-..PA.VW...SD.......GGD...T...G.A--------...G..STS...Q...GS.P....APSV.N..A.VT..EHS.....
```

Figure 2. Alignment of deduced PvMSP-7H amino acid sequences, including the Sal-I sequence. The alignment shows major allele types found in Colombian populations. Dots indicate conserved residues and dashes represent gaps introduced for alignment.
doi:10.1371/journal.pone.0045962.g002

```
     125                                                                           202
      |                                                                             |
Sal-I KIITKGAADK EEEEEDDEDE EAVEGD-EEAGERD EAEAEA VRGDTSHQSG SGDLTPSPPG SSERAPAV
AND11 .......... .......... ......-....... ...... .......... .......... ........
CAR4  .......... .......... ......-....... ...... .......... .......... ........
AND13 .......... .......... ......-....... ...... .......... .......... ........
CAR8  .......V.S N..DGG.... ......EA....EG --.E.. TNEVA..... .........A .G.L...T
AND3  ......VV.S N..DGG.... ......EA....EG --.E.. TNEVA....D .EASGE.D.A .G.L...T
PAC7  .......V.S N..DGG.... ......-.AEAGEE -G.E.. TNEVA....D .AVSPR.VSE .R.P.T.T
AMA4  ........SS ..G...---. .DD.D--G...QET ------ DKGST.D.PD .AVSPQRA.E .P.LTT.T
     201                                                                  270
      |                                                                     |
Sal-I APGETPRVTA TSQGPSQTHN PGSSGNAEPN SASARTPPAA AEAPLPQPAS PDVSTSGEES PNGRVPNV
AND11 .......D.. .......... .......... .......... .......... .......... ........
CAR4  .......... .......... .......... .......... .......... .......... ........
AND13 .......... .......... ....R.EGQS .T.SIN..P. PA.A...S.P QV...PEG.P ES......
CAR8  .R...ATDSP ...E.....T T...R.EGQS .T.SIN..P. PA.A...S.P QV...PEG.P ES......
AND3  .R..NATDSP ...E.....T T...R.EGQS .T.SIN..P. PA.A...S.P QV...PEG.P ES......
PAC7  V.A..A.DNP P..ETAP.NP ..P...V... .....N.... .A.A...... .V..APEG.P ES......
AMA4  .R.A.V.DSP ..P.L.P.NP S.Q...VGQS .T.SIN..P. PA.A...S.P QV...PEG.P ES......
```

Figure 3. Alignment of deduced PvMSP-7I amino acid sequences, including the Sal-I sequence in which the protein sequences differentiating major allele types can be observed. Dots indicate conserved residues and dashes represent gaps introduced for alignment.
doi:10.1371/journal.pone.0045962.g003

was caused by concerted evolution (Table 3); the d_S was higher than d_N at the 5′-end for the three genes. The d_S rate at the 3′-end between *Pvmsp-7C* and the other two genes was higher than d_N rate but the d_S was similar to d_N between *Pvmsp-7H* and *Pvmsp-7I* (Table 3).

Discussion

At least eleven MSPs have been reported in *P. falciparum* and nine *P. vivax* orthologous genes have been described so far. Two *P. vivax* MSP families (*msp-3* and *msp-7*) are particularly interesting since they have expanded differentially [18,33]. The *P. vivax msp-7* gene family has 11 functional genes [18] and at least 8 of them are transcribed during the *P. vivax* schizont stage [20]. MSP-7 forms part of the main protein complex interacting with host cells [16,22]. Taking into account that this protein is localized on the parasite membrane [19,21,22,23], it can be targeted by the host immune system and thus becomes an attractive vaccine candidate.

Besides being recognized by the immune system, vaccine candidates must have limited polymorphism. Results found in this study highlighted differences in allele diversity among *msp-7* gene family members, at least in the two main human malaria species. By contrast with the very little polymorphism displayed by *Pfmsp-7* genes [28,29], *Pvmsp-7* genes (*Pvmsp-7C*, *Pvmsp-7H* and *Pvmsp-7I*) have high genetic diversity. It is worth noting, however, that not all the *Pvmsp-7* genes display a similar pattern. *Pvmsp-7A* and *Pvmsp-7K*, like *Pfmsp-7A*, *Pfmsp-7B*, *Pfmsp-7C*, *Pfmsp-7D*, *Pfmsp-7E*, *Pfmsp-7F* and *Pfmsp-7H* (GenBank ID numbers: XM_001350038.1–44.1) display very low polymorphism [28,29,30] suggesting that *msp-7* genes might be exposed to different selective pressure (such as that exerted by the immune system) or have different biological constraints. Although low genetic polymorphism does not hold for *Pvmsp-7C*, *Pvmsp-7H* and *Pvmsp-7I*, vaccine development using these proteins could be focused on their conserved C-terminal domain (Pfam ID number PF12948) due to the high level of conservation it displays and, like in *P. falciparum*, the C-terminal domain could be involved in interaction with erythrocytes [24].

The genetic diversity found in these three genes places them among the most diverse *P. vivax* genes described so far (Table S1). Different allele types were found when Colombian samples were compared with Sal-I (GenBank ID numbers: XM_001614082.1, XM_001614087.1, XM_001614088.1) and unpublished Korean sequences, (GenBank ID numbers: GU476538.1, GU476534.1,

Table 2. Average number of synonymous substitutions per synonymous site (d_S) and non-synonymous substitutions per non-synonymous site (d_N) covering all sequence pairs at the 5′-end, central region, 3′-end and complete gene.

	5′-end		**Central region**		**3′-end**		**Full length gene**	
	d_S (se)	**d_N (se)**	**d_S (se)**	**d_N (se)**	**d_S (se)**	**d_N (se)**	**d_S (se)**	**d_N (se)**
msp-7C	0.070 (0.016)**	0.011 (0.004)	0.122 (0.023)	0.287 (0.031)•	0.004 (0.003)	0.000 (0.000)	0.053 (0.008)	0.059 (0.007)
msp-7H	0.059 (0.010)•	0.010 (0.003)	0.052 (0.016)	0.162 (0.019)•	0.010 (0.005)	0.005 (0.002)	0.038 (0.006)	0.037 (0.005)
msp-7I	0.041 (0.014)*	0.020 (0.007)	0.060 (0.019)	0.186 (0.021)•	0.017 (0.007)	0.007 (0.003)	0.036 (0.007)	0.048 (0.006)

se: standard error. 5′-end (*Pvmsp-7C*: nucleotides 1–390, *Pvmsp-7H*: nucleotides 1–471, *Pvmsp-7I*: nucleotides 1–525), central (*Pvmsp-7C*: nucleotides 391–717, *Pvmsp-7H*: nucleotides 472–771, *Pvmsp-7I*: nucleotides 526–789) and 3′-end (*Pvmsp-7C*: nucleotides 718–1,191, *Pvmsp-7H*: nucleotides 772–1,200, *Pvmsp-7I*: nucleotides 790–1,188).
*: $p<0.04$, **: $p<0.001$, •: $p<0.0001$.
doi:10.1371/journal.pone.0045962.t002

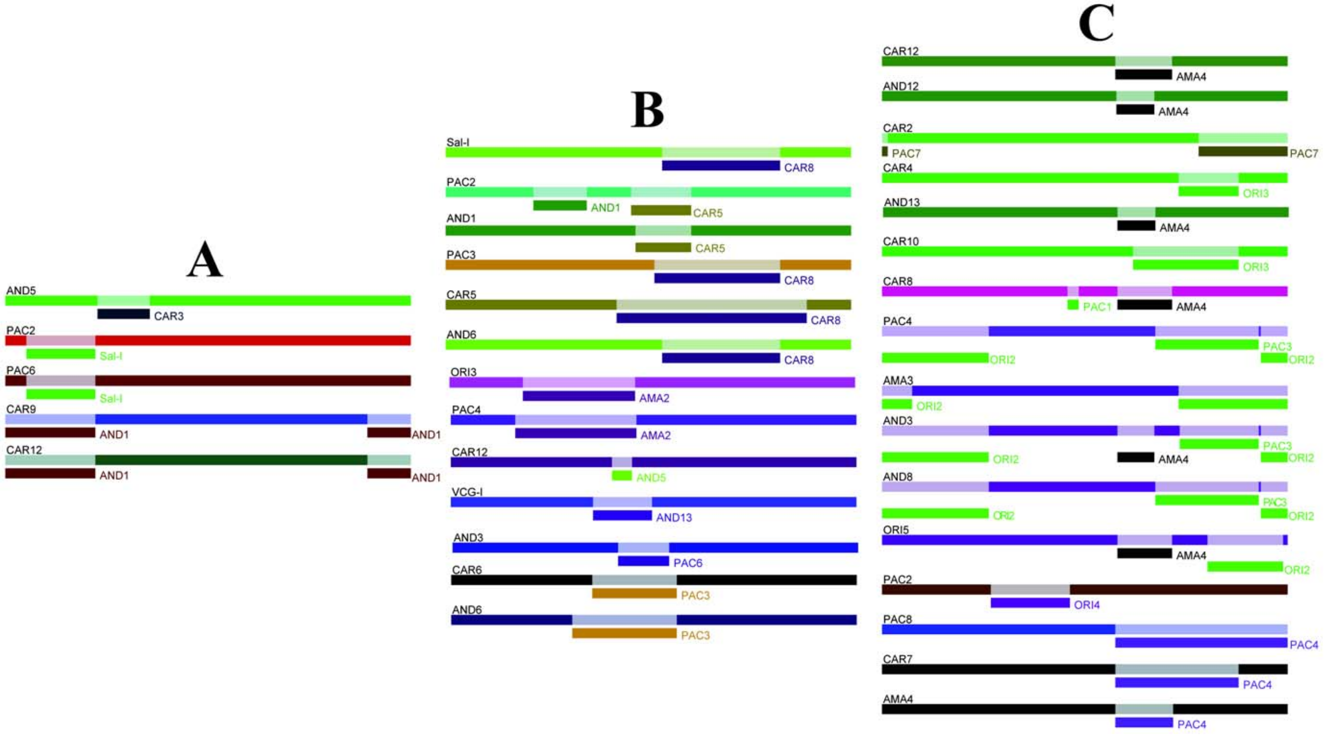

Figure 4. Schematic representation of recombination fragments identified by RDP3 v.3.4 for *Pvmsp-7C* (A), *Pvmsp-7H* (B) and *Pvmsp-7I* (C). The sequence names in black above the rectangles indicate the name of recombinant sequence. The rectangle with name to the right (name of the close relative minor parent) shown in different colors is a graphical representation of a sequence fragment that has potentially been derived through recombination. Only recombination events having $p<0.03$ were taken into account. SAL-I: Salvador strain, AMA: Amazonian, AND: Andean, CAR: Caribbean, ORI: Orinoco, PAC: Pacific.
doi:10.1371/journal.pone.0045962.g004

GU476518.1 and ACY66906-26) (data not shown), suggesting that such genetic diversity is distributed worldwide. Moreover, no correlation between nucleotide diversity (π) and *P. vivax* incidence in the geographical regions in question was observed (data not shown).

The *Pvmsp-7C* and *Pvmsp-7I* genes appear to be under balancing selection. It has been suggested that this type of selection increases

Table 3. Synonymous nucleotide substitution per synonymous site and nonsynonymous nucleotide substitution per nonsynonymous site and the standard deviations between three *Pvmsp-7* genes.

	Synonymous nucleotide substitution per synonymous site (d_S)								
	5′-end			**Central region**			**3′-end**		
	Pvmsp-7C	***Pvmsp-7H***	***Pvmsp-7I***	***Pvmsp-7C***	***Pvmsp-7H***	***Pvmsp-7I***	***Pvmsp-7C***	***Pvmsp-7H***	***Pvmsp-7I***
Pvmsp-7C									
Pvmsp-7H	0.241 (0.050)			1.016 (0.199)			0.879 (0.162)		
Pvmsp-7I	0.468 (0.085)	0.518 (0.092)		0.956 (0.167)	0.790 (0.140)		0.788 (0.138)	0.136 (0.031)	
	Non-synonymous nucleotide substitution per non-synonymous site (d_N)								
	5′-end			**Central region**			**3′-end**		
	Pvmsp-7C	***Pvmsp-7H***	***Pvmsp-7I***	***Pvmsp-7C***	***Pvmsp-7H***	***Pvmsp-7I***	***Pvmsp-7C***	***Pvmsp-7H***	***Pvmsp-7I***
Pvmsp-7C									
Pvmsp-7H	0.072 (0.017)			1.430 (0.238)			0.234 (0.035)		
Pvmsp-7I	0.125 (0.028)	0.152 (0.031)		1.729 (0.244)	1.841 (0.306)		0.323 (0.047)	0.122 (0.028)	

5′-end from nucleotides 1 to 384 (amino acids 1 to 128), central region from nucleotides 385 to 807 (amino acids 129 to 269) and 3′-end from nucleotides 808 to 1,248 (amino acids 270 to 416), numbered according to the alignment in Fig S13.
doi:10.1371/journal.pone.0045962.t003

variability within a population [34] maintaining alleles at intermediate frequencies, as can be observed in the alignment. Nevertheless, Tajima and Fu & Li tests are influenced not only by selection but also by the population history that can alter neutral allele frequency expectations. Therefore, the positive values in the tests could have been the result of a decrease in the population. However, a population affected by genetic drift (a mechanism that decreases the population) was not found for the Colombian population (Fs not statically significant). Moreover, the Hd and π values for both genes suggested a stable population having a long-term effective population size; allele frequency distribution would not therefore be influenced by a demographic process.

By contrast with the *Pvmsp-7C* and *Pvmsp-7I* results discussed above, *Pvmsp-7H* appeared to be under a standard neutral model of molecular evolution since there were no significant values in neutrality tests and no difference between d_N and d_S was found; consequently, high polymorphism would be expected in regions lacking functional constraints [35]. However, a deviation from neutral expectation would only be detected if the average from the whole gene was significantly greater or smaller than 0. The Nei-Gojobori method behaves similarly; positive selection can be detected only if the d_N average from the whole gene is significantly greater than d_S (the opposite occurs for negative selection). The sliding window test (for neutral statistics and ω) for different selective pressures appeared to be acting throughout *Pvmsp-7* gene sequences, as has been described for others proteins [36,37,38,39]. The tests showed that the 5′- and 3′-ends had purifying selection signals (negative values in the neutrality test and $\omega < 1$) while the central region of the three genes had balancing or positive selection signals (positive values in the neutrality test and $\omega > 1$). So, despite the loci displayed being under balancing selection or under neutrality, natural selection may have varied across codons. Two different approaches were thus followed to investigate this hypothesis. One estimated the d_N and d_S rates at the 5′- and 3′-ends and in the central region of the three genes. The central region was under positive selection while the 5′- and 3′-ends seemed to be under negative selection (d_S higher than d_N, only significant for the 5′-end). Similar results were observed when using a second approach in which several maximum likelihood methods were performed to determine positively and negatively selected codons. The central region of the *Pvmsp-7C*, *Pvmsp-7H* and *Pvmsp-7I* genes had codons under positive selection while sites under negative selection were preferentially located at the 5′- and 3′-ends which were relatively fully conserved in *Pvmsp-7* genes.

Besides selective pressure and functional constraints, evolution of malarial antigens might be affected by recombination [13,14,39,40,41]. Several statistics and algorithms showed that intragenic recombination played an important role in generating new allele variants in individual *msp-7* genes. These events affect the accuracy of detecting selected sites, increasing type I errors [32]. When sequences were screened for recombination, the positively selected sites decreased, suggesting that several of the positively selected sites were false positive. However, some were true positively selected sites (Table S6). These sites did not seem to be originated by the stochastic nature of the mutation process, since it would have been expected that d_N substitutions (and/or positive selection) would have been randomly found across the genes; instead, these sites were just observed in the *msp-7* central region and the non-synonymous substitutions in the positively selected sites were found as parsimonious sites and not as singleton sites. Recombination and selection thus seem to drive *Pvmsp-7* genes' antigenic variation, similar to what occurs in other pathogen antigens [42,43,44]. Likewise, it has been shown that both diversification by recombination and purifying selection take place in different regions of the *E. coli fimA* gene [45], and this also occurs in the *Pvmsp-7* genes (Fig. 4 and Table S7).

On the other hand, multigene families might be evolving by recombination among paralogous genes, thereby contributing to allele diversity or the homogenization of the multigene family in the event of unequal crossover or gene conversion [46]. The gene conversion and subsequent phylogenetic analysis revealed that these genes had not evolved independently; several conversion tracks at the 5′- and 3′-ends (but not in the central region) were detected among *msp-7* genes, suggesting that at least these three genes could be evolving by concerted evolution mediated by a biased gene conversion, homogenizing only the 5′ and 3′-ends. Thus, recombination would seem to affect the evolution of the *msp-7* family as a whole, and individual *msp-7* genes. However, protein homogeneity can also be attained by purifying selection. Under the assumption of gene conversion, it would be expected that d_S between duplicated genes would be similar to d_N, whereas if purifying selection is the major evolutionary force, then d_S would be much higher than d_N [47], According to d_S and d_N rate comparison, the homogeneity at the *Pvmsp-7* genes' 5′-end was apparently caused by functional constraints rather than by concerted evolution, since *Pvmsp-7* genes have diverged extensively by silent nucleotide substitution at these ends. The conservation of the 3′-end in *Pvmsp-7H* and *Pvmsp-7I* but not in *Pvmsp-7C* seems to be maintained by gene conversion. This behavior might be a consequence of closely spaced gene duplicates being more likely to undergo gene conversion [48,49]. *Pvmsp-7H* and *Pvmsp-7I* genes are neighbors separated by 1,086 base pairs while the *Pvmsp-7C* gene is separated from the others by 9,619 base pairs. Accordingly, the high *Pvmsp-7*s' C-terminal conservation may have been the result of functional constraints and purifying selection, or the result of gene conversion (between *Pvmsp-7H* and *Pvmsp-7I*) possibly due to the presence of a functional domain (Pfam ID number PF12948) within this region. Therefore, the central region (the most diverse) could have been under selective pressure exerted by the immune system; consequently, the intragenic recombination and, to a lesser extent, the positively selected sites increased genetic diversity, generating different allele variants as an evasion mechanism.

Phylogenetic analysis showed that, regardless of an isolate's origin, sequences tended to cluster without having clear geographical distribution. This pattern might indicate a constant *P. vivax* gene flow in Colombian regions. This result agreed with previous reports of other highly polymorphic *P. vivax* genes not involved in geographical clustering [13,14,39].

Previous studies have shown the potential role of MSP-7 in parasite invasion of erythrocytes [19,24,25,50] which, added to its immunogenicity [26,31], and following the rules for subunit-based vaccine development [51,52], make the MSP-7 conserved domain an attractive candidate to be evaluated when designing an antimalarial vaccine.

Materials and Methods

Ethics Statement

All *P. vivax*-infected patients who provided us with the blood samples signed an informed consent and the purpose of the study was carefully explained to them. All procedures carried out in this study were approved by our institute's ethics committee.

Source of Parasite DNA and Field Isolate Genotyping

Peripheral blood samples from patients proving positive for *P. vivax* malaria by microscope examination were collected from geographical regions of Colombia from 2007 to 2010. DNA was obtained using a Wizard Genomic DNA Purification kit (Promega) following manufacturer's instructions.

All parasite samples were genotyped by PCR-RFLP of the *Pvmsp-1* gene's blocks 6, 7 and 8 as previously described [53] for selecting only samples having a single *P. vivax msp-1* allele infection.

Amplification and Sequencing

Primers were designed to amplify *Pvmsp-7C*, *Pvmsp-7H* and *Pvmsp-7I* DNA fragments based on the Sal-I reference sequences (GenBank IDs: XM_001614082.1, XM_001614087.1 and XM_001614088.1, respectively). The DNA fragment from *Pvmsp-7C* was amplified with 7Cdto 5′ ACCACAAAGATGAATAAAACG 3′ and 7Crev 5′ CACCTCAATCGTGTTCAGC 3′ primers. *Pvmsp-7H* was amplified by using 7Hdto 5′ GTGTGCATCAGTATAGCGAC 3′ and 7Hrev 5′ AAGAAGGTTAGCCATAAATGC 3′ primers and *Pvmsp-7I* was amplified with 7Idto 5′ ACAATGAGGGGCAAGTACG 3′ and 7Irev 5′ TTCATTCGTTGCTCACTTCG 3′ primers. All PCR reactions were performed using KAPA HiFi HotStart Readymix containing 0.3 μM of each primer in a final 25 μL volume. Thermal conditions were set as follows: one cycle of 5 min at 95°C, 25 cycles of 20 sec at 98°C, 15 sec at 62°C for *Pvmsp-7C* and *Pvmsp-7*I and 60°C for *Pvmsp-7H,* 30 sec at 72°C, followed by a 5 min final extension at 72°C. PCR products were purified using the Wizard PCR preps kit (Promega), and then sequenced with a BigDye Terminator kit (MACROGEN, Seoul, South Korea) in both directions. Internal primers were used for sequencing (intdc 5′ CTGTTGGACCCGGTGGAG3′, intrc 5′ CTTGTTGATTCGCTCCTGG 3′; intdh 5′ TCAAATACAGCACAGACTTCC, intrh 5′ CCTCAGGACAACCCGAAAG 3′; intdi 5′ TCACAAACGCACAACCCAGG 3′ and 5′ GCTCCATTACCACAACCGG 3′). Two PCR products obtained from independent PCR amplifications were sequenced per isolate to discard errors.

Statistical Analysis for the *msp-7* Sequences

Electropherograms were assembled using CLC DNA workbench 6 (CLC bio, Cambridge, MA, USA) and Clustal W [54] was used for aligning the deduced amino acid sequences, followed by manual editing. Additionally, repeats with 90% similarity in the deduced *msp-7* amino acid sequences were detected by using the T-REKS algorithm [55]. The PAL2NAL program [56] was used for constructing codon alignments from the corresponding aligned amino acid sequences. Colombian *P. vivax msp-7C, msp-7H* and *msp-7I* sequences were compared against the previously described Sal-I strain sequences available in GenBank (ID numbers: XM_001614082.1, XM_001614087.1, and XM_001614088.1).

DNA polymorphism was calculated with DnaSP v.5 software [57]. Tests to assess departure from the neutral model were applied using Tajima's D and Fu & Li's D* and F* statistics. The former statistic compares the differences between the total number of segregating sites and the average number of nucleotide differences between sequence pairs. The Fu & Li test calculates the D* statistic which is based on the difference between the number of singletons and the total number of mutations, as well as the F* statistic which is based on the difference between the number of singletons and the average number of nucleotide differences between sequence pairs. Positive and negative values from both tests correspond to departures from neutrality. The Fs statistic [58] was used; it is based on gene frequency distribution. All tests were applied using DnaSP v.5 software, considering coalescent simulations for obtaining confidence intervals [57]. Natural selection was estimated using the modified Nei-Gojobori method [59] to calculate the average number of non-synonymous (d_N) and synonymous (d_S) substitutions. Differences between d_N and d_S were assessed by applying the Z-test using MEGA software v.5 [60]. Additionally, codon sites under positive or negative selection at the population level were assessed by using Datamonkey web server [61] with IFEL, a codon-based maximum likelihood method [62]. This method infers selective pressure at population level; positive or negatively selected sites were assessed by FEL, SLAC and REL methods [63]. All algorithms estimated the ω (d_N/d_S ratio) at every codon in the alignment. A $\leq$0.1 p-value was considered significant for IFEL, FEL and SLAC methods and $\geq$50 Bayes factor for REL. Only Colombian sequences were considered for all analyses performed; positions containing gaps or repeats in the alignment were not taken into account (Table S9).

Linkage disequilibrium (LD) was evaluated by calculating the Z_{nS} statistic [64] which is the average of $R\hat{}^2$ over all pairwise comparisons. A lineal regression between LD ($R\hat{}^2$) and nucleotide distances was performed to evaluate whether recombination was taking place in *Pvmsp-7* genes. Recombination events were assessed using DnaSP v.5 software [57] applying the ZZ statistic [65] and RM parameter [66]. The latter statistic incorporates the effective population size and probability of recombination between adjacent nucleotides per generation. RDP3 v3.4 software [67] was used for detecting recombination regions in *msp-7* genes. This tool looks for evidence of recombination among aligned sequences by examining all possible triplet combinations following a scanning approach with a range of different recombination detection algorithms. Additionally, the algorithm developed by Betrán *et al.* (1997) [68] incorporated in DnaSP [57] and the GENECONV program [69] incorporated in RDP3 v3.4 software [67] were used to detect gene conversion among paralogous genes.

Geographical Clustering

Maximum Likelihood (ML) trees describing the phylogenetic consequences of the recombination events were constructed using RDP3 v3.4 software [67] with the HKY model (selected by the ModelTest algorithm [70]) to evaluate relationships between polymorphism and the geographical distribution of the isolates. Additionally, *Pvmsp-7C*, *Pvmsp-7H* and *Pvmsp-7I* sequences were aligned and trees were constructed by using ML with the TN93+*G* model selected by the ModelTest algorithm [70]. Bootstrap analysis (with 1,000 replicates each) was used for assigning confidence levels to branch nodes. Positions containing gaps as well as regions in the alignment that contained repeats were not taken into account in the phylogenetic analysis (Table S9).

Supporting Information

Figure S1 Geographical location of the study regions within Colombia. Amazonian region (samples from Calamar in the Guaviare department), Andean region (samples from Apartadó and El Bagre in the Antioquia department), Caribbean region (samples from Tierra Alta and Puerto Libertador in the Córdoba department), Orinoco region (samples from Tauramena in the Casanare department and Mapiripán in the Meta department), and Pacific region (samples from Istmina in the Chocó department and Tumaco in the Nariño department). Black dots on the map represent the areas from which patients came who donated the infected blood samples. 1: Puerto Libertador, 2: Tierra Alta, 3: Istmina, 4: Tumaco, 5: Apartado, 6: El Bagre, 7: Tauramena, 8: Mapiripan, 9: Calamar.
(TIF)

Figure S2 Sliding window analysis of Tajima D and Fu & Li D* and F* statistics along the *Pvmsp-7C* (A), *Pvmsp-7H* (B) and *Pvmsp-7I* (C) genes. Bars [Tajima's D (blue), Fu & Li's D* (red) and F* (green)] below each figure represent the regions where the tests showed a significant deviation from the neutral expectation. 5′-end (*Pvmsp-7C*: nucleotides 1–390, *Pvmsp-7H*: nucleotides 1–471, *Pvmsp-7I*: nucleotides 1–525), central (*Pvmsp-7C*: nucleotides 391–717, *Pvmsp-7H*: nucleotides 472–771, *Pvmsp-7I*: nucleotides 526–789) and 3′-end (*Pvmsp-7C*: nucleotides 718–1,191, *Pvmsp-7H*: nucleotides 772–1,200, *Pvmsp-7I*: nucleotides 790–1,188).
(PDF)

Figure S3 Sliding window analysis for ω rates (d_N/d_S) throughout the *Pvmsp-7C* (Blue), *Pvmsp-7H* (Red) and *Pvmsp-7I* (Green) genes. Discontinuity of the curves is due to gaps within the alignments which were not considered for the analysis. 5′-end (*Pvmsp-7C*: nucleotides 1–390, *Pvmsp-7H*: nucleotides 1–471, *Pvmsp-7I*: nucleotides 1–525), central (*Pvmsp-7C*: nucleotides 391–717, *Pvmsp-7H*: nucleotides 472–771, *Pvmsp-7I*: nucleotides 526–789) and 3′-end (*Pvmsp-7C*: nucleotides 718–1,191, *Pvmsp-7H*: nucleotides 772–1,200, *Pvmsp-7I*: nucleotides 790–1,188).
(PDF)

Figure S4 The linkage disequilibrium (LD) plot for *P. vivax Pvmsp-7C* (A), *Pvmsp-7H* (B) and *Pvmsp-7I* (C). Trace line represents the regression line which declined as nucleotide distance increased suggesting that intragenic recombination was taking place in *msp-7* genes.
(PDF)

Figure S5 ML trees describing the phylogenetic consequences of the intragenic recombination in *Pvmsp-7C*. A topology is inferred for each recombinant fragment, (A) from nucleotides 685 to 855, (B) from nucleotides 1,060 to 1,167 (excluding nucleotides 1,105 to 1,107) and from 18 to 266, and (C) from nucleotides 62 to 266. Isolates clustered without a clear geographical distribution. Sal-I: Salvador strain, AMA: Amazon, AND: Andean, CAR: Caribbean, ORI: Orinoco, PAC: Pacific.
(TIF)

Figure S6 ML trees describing the phylogenetic consequences of the intragenic recombination in *Pvmsp-7H*. A topology is inferred for each recombinant fragment, (A) nucleotides 189 to 543 (excluding nucleotides 436 to 450 and nucleotides 484 to 486), (B) nucleotides 353 to 661 (excluding nucleotides 436 to 450 and 568 to 627), (C) nucleotides 425 to 655 (excluding nucleotides 436 to 450, 484 to 486 and 568 to 627), (D) nucleotides 500 to 1,057 (excluding nucleotides 568 to 627 and 772 to 774), (E) nucleotides 628 to 977 (excluding nucleotides 772 to 774), (F) nucleotides 630 to 997, (G) nucleotides 416 to 590 (excluding nucleotides 436 to 450 and 484 to 486), (H) nucleotides 472 to 531 (excluding nucleotides 484 to 486), and (I) nucleotides 482 to 646 (excluding nucleotides 568 to 627). Isolates clustered without a clear geographical distribution. Sal-I: Salvador strain, AMA: Amazon, AND: Andean, CAR: Caribbean, ORI: Orinoco, PAC: Pacific.
(TIF)

Figure S7 ML trees describing the phylogenetic consequences of the intragenic recombination in *Pvmsp-7I*. A topology is inferred for each recombinant fragment identified, (A) nucleotides 19 to 187 and nucleotides 683 to 1,188, (B) nucleotides 683 to 849, (C) nucleotides 684 to 797, (D) nucleotides 19 to 358 and nucleotides 796 to 1,188, (E) nucleotides 798 to 1,105, (F) nucleotides 868 to 1,188, (G) nucleotides 920 to 1,188, (H) nucleotides 733 to 1,044, (I) nucleotides 688 to 797, and (J) nucleotides 683 to 849. Isolates clustered without a clear geographical distribution. SAL-I: Salvador strain, AMA: Amazon, AND: Andean, CAR: Caribbean, ORI: Orinoco, PAC: Pacific.
(TIF)

Figure S8 (A) Phylogenetic tree obtained by ML for *Pvmsp-7* sequences based on the TN93+*G* model, ignoring recombination. Three monophyletic groups are shown; the first groups clustered *Pvmsp-7H* (red) sequences, the second group clustered *Pvmsp-7I* (green) sequences and the third group clustered *Pvmsp-7C* (blue) sequences. (B-E) Trees describing phylogenetic consequences of some gene conversion tracks identified. (B) nucleotides 71 to 265 (amino acids 24 to 89), (C) nucleotides 913 to 1,090 (amino acids 305 to 390), (D) nucleotides 913 to 1,090 (amino acids 305 to 354), (E) nucleotides 913 to 1,175 (amino acids 305 to 392). Positions are numbered according to the alignment in Fig. S13. These topologies suggest that at least *Pvmsp-7C*, *Pvmsp-7H* and *Pvmsp-7I* genes did not evolve independently. Numbers represent bootstrap values with 1,000 replicates.
(PDF)

Figure S9 Schematic representation of gene conversion tracks identified by DnaSP and GENECONV for *Pvmsp-7C* (blue), *Pvmsp-7H* (red) and *Pvmsp-7I* (green) as a combined data set. Rectangles in a different color are graphical representations of sequence fragments that have potentially been originated by gene conversion and are localized at the conserved 5′- and 3′-ends. The black bars delimit the 5′-end, the central region and the 3′-end, [5′-end (*Pvmsp-7C*: nucleotides 1–390, *Pvmsp-7H*: nucleotides 1–471, *Pvmsp-7I*: nucleotides 1–525), central (*Pvmsp-7C*: nucleotides 391–717, *Pvmsp-7H*: nucleotides 472–771, *Pvmsp-7I*: nucleotides 526–789) and 3′-end (*Pvmsp-7C*: nucleotides 718–1,191, *Pvmsp-7H*: nucleotides 772–1,200, *Pvmsp-7I*: nucleotides 790–1,188)].
(PDF)

Figure S10 Alignment of deduced PvMSP-7C amino acid sequences.
(PDF)

Figure S11 Alignment of deduced PvMSP-7H amino acid sequences.
(TIF)

Figure S12 Alignment of deduced PvMSP-7C amino acid sequences.
(TIF)

Figure S13 Alignment of the paralogous PvMSP-7deduced amino acid sequences.
(PDF)

Table S1 Nucleotide diversity for *P. vivax* antigens. n: number of isolates, π: nucleotide diversity.
(PDF)

Table S2 Nucleotide diversity (π) values for subpopulations within Colombia.
(PDF)

Table S3 Nucleotide and amino acid positions within the 5′- end, central region and 3′-end.
(PDF)

Table S4 Positively selected sites detected for *Pvmsp-7* genes without taking recombination into account. Numbers according to the reference Sal-I protein sequence *Pvmsp-7C*:

XP_001614132.1, *Pvmsp-7H*: XP_001614137.1 and *Pvmsp-7I*: XP_001614138.1.
(PDF)

Table S5 Negatively selected sites detected for *Pvmsp-7* genes without taking recombination into account. Numbers according to the reference Sal-I protein sequence *Pvmsp-7C*: XP_001614132.1, *Pvmsp-7H*: XP_001614137.1 and *Pvmsp-7I*: XP_001614138.1.
(PDF)

Table S6 Positively selected sites detected for *Pvmsp-7* genes taking recombination into account. Numbers according to the reference Sal-I protein sequence *Pvmsp-7C*: XP_001614132.1, *Pvmsp-7H*: XP_001614137.1 and *Pvmsp-7I*: XP_001614138.1.
(PDF)

Table S7 Negatively selected sites detected for *Pvmsp-7* genes taking recombination into account. Numbers according to the reference Sal-I protein sequence *Pvmsp-7C*: XP_001614132.1, *Pvmsp-7H*: XP_001614137.1 and *Pvmsp-7I*: XP_001614138.1.
(PDF)

Table S8 Conversion tracks identified by DnaSP and GENECONV, between *Pvmsp-7* genes. Nucleotides and amino acids based in alignment of the Fig. S13.
(PDF)

Table S9 Nucleotides and amino acid positions excluded from the analysis.
(PDF)

Acknowledgments

We would like to thank the Instituto Nacional de Salud (INS), Bogotá, Colombia, for supplying the samples. We would like to thank Jason Garry for translating the manuscript. Our special gratitude goes to Francy Homez Homez and Pamela Jory Rios for their technical assistance. We would also like to thank Professor Manuel Elkin Patarroyo for his invaluable comments and suggestions. In loving memory of Doctor Carlos Fernando Cortés, who passed away recently.

Author Contributions

Conceived and designed the experiments: DG-O MAP. Performed the experiments: DG-O JF-R CL. Analyzed the data: DG-O JF-R CL. Wrote the paper: DG-O MAP.

References

1. Guerra CA, Howes RE, Patil AP, Gething PW, Van Boeckel TP, et al. (2010) The international limits and population at risk of Plasmodium vivax transmission in 2009. PLoS Negl Trop Dis 4: e774.
2. Lysenko AJ, Semashko IN (1968) Medical geography: a medical-geographical study of an ancient disease. WHO Publisher's Code: 911.3: 616.936(100).
3. Poespoprodjo JR, Fobia W, Kenangalem E, Lampah DA, Hasanuddin A, et al. (2009) Vivax malaria: a major cause of morbidity in early infancy. Clin Infect Dis 48: 1704–1712.
4. Tjitra E, Anstey NM, Sugiarto P, Warikar N, Kenangalem E, et al. (2008) Multidrug-resistant Plasmodium vivax associated with severe and fatal malaria: a prospective study in Papua, Indonesia. PLoS Med 5: e128.
5. Price RN, Tjitra E, Guerra CA, Yeung S, White NJ, et al. (2007) Vivax malaria: neglected and not benign. Am J Trop Med Hyg 77: 79–87.
6. Kochar DK, Das A, Kochar SK, Saxena V, Sirohi P, et al. (2009) Severe Plasmodium vivax malaria: a report on serial cases from Bikaner in northwestern India. Am J Trop Med Hyg 80: 194–198.
7. Carvalho LJ, Daniel-Ribeiro CT, Goto H (2002) Malaria vaccine: candidate antigens, mechanisms, constraints and prospects. Scand J Immunol 56: 327–343.
8. Galinski MR, Barnwell JW (2008) Plasmodium vivax: who cares? Malar J 7 Suppl 1: S9.
9. O'Donnell RA, de Koning-Ward TF, Burt RA, Bockarie M, Reeder JC, et al. (2001) Antibodies against merozoite surface protein (MSP)-1(19) are a major component of the invasion-inhibitory response in individuals immune to malaria. J Exp Med 193: 1403–1412.
10. Genton B, Reed ZH (2007) Asexual blood-stage malaria vaccine development: facing the challenges. Curr Opin Infect Dis 20: 467–475.
11. Takala SL, Plowe CV (2009) Genetic diversity and malaria vaccine design, testing and efficacy: preventing and overcoming 'vaccine resistant malaria'. Parasite Immunol 31: 560–573.
12. Figtree M, Pasay CJ, Slade R, Cheng Q, Cloonan N, et al. (2000) Plasmodium vivax synonymous substitution frequencies, evolution and population structure deduced from diversity in AMA 1 and MSP 1 genes. Mol Biochem Parasitol 108: 53–66.
13. Gomez A, Suarez CF, Martinez P, Saravia C, Patarroyo MA (2006) High polymorphism in Plasmodium vivax merozoite surface protein-5 (MSP5). Parasitology 133: 661–672.
14. Putaporntip C, Udomsangpetch R, Pattanawong U, Cui L, Jongwutiwes S (2010) Genetic diversity of the Plasmodium vivax merozoite surface protein-5 locus from diverse geographic origins. Gene 456: 24–35.
15. Richie TL, Saul A (2002) Progress and challenges for malaria vaccines. Nature 415: 694–701.
16. Pachebat JA, Ling IT, Grainger M, Trucco C, Howell S, et al. (2001) The 22 kDa component of the protein complex on the surface of Plasmodium falciparum merozoites is derived from a larger precursor, merozoite surface protein 7. Mol Biochem Parasitol 117: 83–89.
17. Trucco C, Fernandez-Reyes D, Howell S, Stafford WH, Scott-Finnigan TJ, et al. (2001) The merozoite surface protein 6 gene codes for a 36 kDa protein associated with the Plasmodium falciparum merozoite surface protein-1 complex. Mol Biochem Parasitol 112: 91–101.
18. Garzon-Ospina D, Cadavid LF, Patarroyo MA (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.
19. Tewari R, Ogun SA, Gunaratne RS, Crisanti A, Holder AA (2005) Disruption of Plasmodium berghei merozoite surface protein 7 gene modulates parasite growth in vivo. Blood 105: 394–396.
20. Bozdech Z, Mok S, Hu G, Imwong M, Jaidee A, et al. (2008) The transcriptome of Plasmodium vivax reveals divergence and diversity of transcriptional regulation in malaria parasites. Proc Natl Acad Sci U S A 105: 16290–16295.
21. Mello K, Daly TM, Long CA, Burns JM, Bergman LW (2004) Members of the merozoite surface protein 7 family with similar expression patterns differ in ability to protect against Plasmodium yoelii malaria. Infect Immun 72: 1010–1018.
22. Mello K, Daly TM, Morrisey J, Vaidya AB, Long CA, et al. (2002) A multigene family that interacts with the amino terminus of plasmodium MSP-1 identified using the yeast two-hybrid system. Eukaryot Cell 1: 915–925.
23. Mongui A, Perez-Leal O, Soto SC, Cortes J, Patarroyo MA (2006) Cloning, expression, and characterisation of a Plasmodium vivax MSP7 family merozoite surface protein. Biochem Biophys Res Commun 351: 639–644.
24. Garcia Y, Puentes A, Curtidor H, Cifuentes G, Reyes C, et al. (2007) Identifying merozoite surface protein 4 and merozoite surface protein 7 Plasmodium falciparum protein family members specifically binding to human erythrocytes suggests a new malarial parasite-redundant survival mechanism. J Med Chem 50: 5665–5675.
25. Kadekoppala M, O'Donnell RA, Grainger M, Crabb BS, Holder AA (2008) Deletion of the Plasmodium falciparum merozoite surface protein 7 gene impairs parasite invasion of erythrocytes. Eukaryot Cell 7: 2123–2132.
26. Wang L, Crouch L, Richie TL, Nhan DH, Coppel RL (2003) Naturally acquired antibody responses to the components of the Plasmodium falciparum merozoite surface protein 1 complex. Parasite Immunol 25: 403–412.
27. Kadekoppala M, Holder AA (2010) Merozoite surface proteins of the malaria parasite: the MSP1 complex and the MSP7 family. Int J Parasitol 40: 1155–1161.
28. Tetteh KK, Stewart LB, Ochola LI, Amambua-Ngwa A, Thomas AW, et al. (2009) Prospective identification of malaria parasite genes under balancing selection. PLoS One 4: e5568.
29. Roy SW, Weedall GD, da Silva RL, Polley SD, Ferreira MU (2009) Sequence diversity and evolutionary dynamics of the dimorphic antigen merozoite surface protein-6 and other Msp genes of Plasmodium falciparum. Gene 443: 12–21.
30. Garzon-Ospina D, Romero-Murillo L, Tobon LF, Patarroyo MA (2011) Low genetic polymorphism of merozoite surface proteins 7 and 10 in Colombian Plasmodium vivax isolates. Infect Genet Evol 11: 528–531.
31. Chen JH, Jung JW, Wang Y, Ha KS, Lu F, et al. (2010) Immunoproteomics profiling of blood stage Plasmodium vivax infection by high-throughput screening assays. J Proteome Res 9: 6479–6489.
32. Anisimova M, Nielsen R, Yang Z (2003) Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. Genetics 164: 1229–1236.
33. Carlton JM, Adams JH, Silva JC, Bidwell SL, Lorenzi H, et al. (2008) Comparative genomics of the neglected human malaria parasite Plasmodium vivax. Nature 455: 757–763.

34. Nielsen R (2005) Molecular signatures of natural selection. Annu Rev Genet 39: 197–218.
35. Kimura M (1983) The neutral theory of molecular evolution. Cambridge Cambridgeshire; New York: Cambridge University Press. xv, 367 p.
36. Suarez CF, Patarroyo ME, Trujillo E, Estupinan M, Baquero JE, et al. (2006) Owl monkey MHC-DRB exon 2 reveals high similarity with several HLA-DRB lineages. Immunogenetics 58: 542–558.
37. Martinez P, Suarez CF, Gomez A, Cardenas PP, Guerrero JE, et al. (2005) High level of conservation in Plasmodium vivax merozoite surface protein 4 (PvMSP4). Infect Genet Evol 5: 354–361.
38. Putaporntip C, Jongwutiwes S, Ferreira MU, Kanbara H, Udomsangpetch R, et al. (2009) Limited global diversity of the Plasmodium vivax merozoite surface protein 4 gene. Infect Genet Evol 9: 821–826.
39. Martinez P, Suarez CF, Cardenas PP, Patarroyo MA (2004) Plasmodium vivax Duffy binding protein: a modular evolutionary proposal. Parasitology 128: 353–366.
40. Mascorro CN, Zhao K, Khuntirat B, Sattabongkot J, Yan G, et al. (2005) Molecular evolution and intragenic recombination of the merozoite surface protein MSP-3alpha from the malaria parasite Plasmodium vivax in Thailand. Parasitology 131: 25–35.
41. Putaporntip C, Jongwutiwes S, Seethamchai S, Kanbara H, Tanabe K (2000) Intragenic recombination in the 3′ portion of the merozoite surface protein 1 gene of Plasmodium vivax. Mol Biochem Parasitol 109: 111–119.
42. Orsi RH, Ripoll DR, Yeung M, Nightingale KK, Wiedmann M (2007) Recombination and positive selection contribute to evolution of Listeria monocytogenes inlA. Microbiology 153: 2666–2678.
43. Andrews TD, Gojobori T (2004) Strong positive selection and recombination drive the antigenic variation of the PilE protein of the human pathogen Neisseria meningitidis. Genetics 166: 25–32.
44. Polley SD, Conway DJ (2001) Strong diversifying selection on domains of the Plasmodium falciparum apical membrane antigen 1 gene. Genetics 158: 1505–1512.
45. Peek AS, Souza V, Eguiarte LE, Gaut BS (2001) The interaction of protein structure, selection, and recombination on the evolution of the type-1 fimbrial major subunit (fimA) from Escherichia coli. J Mol Evol 52: 193–204.
46. Nei M, Rooney AP (2005) Concerted and birth-and-death evolution of multigene families. Annu Rev Genet 39: 121–152.
47. Nei M, Rogozin IB, Piontkivska H (2000) Purifying selection and birth-and-death evolution in the ubiquitin gene family. Proc Natl Acad Sci U S A 97: 10866–10871.
48. Semple C, Wolfe KH (1999) Gene duplication and gene conversion in the Caenorhabditis elegans genome. J Mol Evol 48: 555–564.
49. Katju V, Lynch M (2003) The structure and early evolution of recently arisen gene duplicates in the Caenorhabditis elegans genome. Genetics 165: 1793–1803.
50. Woehlbier U, Epp C, Hackett F, Blackman MJ, Bujard H (2010) Antibodies against multiple merozoite surface antigens of the human malaria parasite Plasmodium falciparum inhibit parasite maturation and red blood cell invasion. Malar J 9: 77.
51. Cifuentes G, Bermudez A, Rodriguez R, Patarroyo MA, Patarroyo ME (2008) Shifting the polarity of some critical residues in malarial peptides' binding to host cells is a key factor in breaking conserved antigens' code of silence. Med Chem 4: 278–292.
52. Patarroyo ME, Patarroyo MA (2008) Emerging rules for subunit-based, multiantigenic, multistage chemically synthesized vaccines. Acc Chem Res 41: 377–386.
53. Imwong M, Pukrittayakamee S, Gruner AC, Renia L, Letourneur F, et al. (2005) Practical PCR genotyping protocols for Plasmodium vivax using Pvcs and Pvmsp1. Malar J 4: 20.
54. Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22: 4673–4680.
55. Jorda J, Kajava AV (2009) T-REKS: identification of Tandem REpeats in sequences with a K-meanS based algorithm. Bioinformatics 25: 2632–2638.
56. Suyama M, Torrents D, Bork P (2006) PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res 34: W609–612.
57. Librado P, Rozas J (2009) DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics 25: 1451–1452.
58. Fu YX (1997) Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. Genetics 147: 915–925.
59. Zhang J, Rosenberg HF, Nei M (1998) Positive Darwinian selection after gene duplication in primate ribonuclease genes. Proc Natl Acad Sci U S A 95: 3708–3713.
60. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, et al. (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol 28: 2731–2739.
61. Delport W, Poon AF, Frost SD, Kosakovsky Pond SL (2010) Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology. Bioinformatics 26: 2455–2457.
62. Pond SL, Frost SD, Grossman Z, Gravenor MB, Richman DD, et al. (2006) Adaptation to different human populations by HIV-1 revealed by codon-based analyses. PLoS Comput Biol 2: e62.
63. Kosakovsky Pond SL, Frost SD (2005) Not so different after all: a comparison of methods for detecting amino acid sites under selection. Mol Biol Evol 22: 1208–1222.
64. Kelly JK (1997) A test of neutrality based on interlocus associations. Genetics 146: 1197–1206.
65. Rozas J, Gullaud M, Blandin G, Aguade M (2001) DNA variation at the rp49 gene region of Drosophila simulans: evolutionary inferences from an unusual haplotype structure. Genetics 158: 1147–1155.
66. Hudson RR, Kaplan NL (1985) Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics 111: 147–164.
67. Martin DP, Lemey P, Lott M, Moulton V, Posada D, et al. (2010) RDP3: a flexible and fast computer program for analyzing recombination. Bioinformatics 26: 2462–2463.
68. Betran E, Rozas J, Navarro A, Barbadilla A (1997) The estimation of the number and the length distribution of gene conversion tracts from population DNA sequence data. Genetics 146: 89–99.
69. Sawyer S (1989) Statistical tests for detecting gene conversion. Mol Biol Evol 6: 526–538.
70. Posada D, Crandall KA (1998) MODELTEST: testing the model of DNA substitution. Bioinformatics 14: 817–818.