Molecular Phylogenetics and Evolution 55 (2010) 399–408

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

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

# Differential expansion of the merozoite surface protein (*msp*)-7 gene family in *Plasmodium* species under a birth-and-death model of evolution

Diego Garzón-Ospina [a,b], Luis F. Cadavid [c], Manuel A. Patarroyo [a,b,*]

[a] Molecular Biology Department, Fundación Instituto de Inmunología de Colombia (FIDIC), Colombia
[b] School of Medicine and Health Sciences, Universidad del Rosario, Colombia
[c] Institute of Genetics and Department of Biology, Universidad Nacional de Colombia, Colombia

ARTICLE INFO

*Article history:*
Received 1 April 2009
Revised 3 February 2010
Accepted 12 February 2010
Available online 19 February 2010

*Keywords:*
*Plasmodium*
*msp-7*
Multigene family
Gene duplication
Functional redundancy

ABSTRACT

MSP-7 is a surface protein expressed by the *Plasmodium* merozoite as part of a protein-complex involved in initial interaction between merozoite and erythrocyte. Contigs of seven *Plasmodium* species were analyzed in order to identify all *msp-7* family genes. The search identified annotated and unannotated open reading frames (ORFs) and showed an uneven number of *msp-7* genes among the different species of the *Plasmodium* lineage. A phylogenetic analysis established the presence of at least two ancestral genes and identified various lineage- and species-specific duplication events. An estimation of synonymous ($d_S$) and non-synonymous substitutions ($d_N$) showed higher $d_S$ values compared to $d_N$ values, suggesting the action of purifying selection on these genes, moreover no changes in $\omega$ (evolutive rates) were found in codon models test. These data together with the data obtained from the Gu's type-I functional divergence test and comparisons between evolutionary rates among orthologous and paralogous genes suggest functional redundancy. Finally, an analysis of recombination events suggests that several sequences are undergoing such process and that this mechanism could therefore be playing an important role in the emergence of new sequences. We conclude that evolution of the *msp-7* family is in agreement with a birth-and-death model of evolution, as *msp-7* genes have expanded until reaching an optimal gene copy number in each *Plasmodium* species in order to adapt to different niches.



## 1. Introduction

Malaria is an infectious disease caused by protozoa of the genus *Plasmodium* that parasitize reptiles, birds and mammals (Escalante and Ayala, 1994; Rich and Ayala, 2003). Only *P. ovale*, *P. malariae*, *P. falciparum* and *P. vivax* are known to infect humans. Among these, *P. falciparum* causes the most lethal form of the disease and *P. vivax* is the most widely distributed (Mendis et al., 2001). Evolutive history studies show a correlation between the phylogenies of the genus *Plasmodium* and their hosts, such that *P. falciparum* (one of the species causing human malaria) is the species most closely related to *P. reichenowi* (parasite infecting chimpanzees) and both form a monophyletic group that diverged approximately at the same time as the chimpanzees and human lineages (Carter and Mendis, 2002). However, *P. vivax*, another parasite species infecting humans appears to form a monophyletic group with parasite species that infect Old world monkeys, suggesting a host switch. Other monophyletic lineages distinguished in the genus *Plasmodium* include the lineage of rodent parasites and the bird/reptiles lineage (Rich and Ayala, 2003; Hayakawa et al., 2008).*Plasmodium* spp. have a complex life cycle involving both an invertebrate and a vertebrate host (Rich and Ayala, 2003). Infection begins when sporozoites are inoculated into the vertebrate host's bloodstream by the bite of an infected female *Anopheles* mosquito. Sporozoites rapidly migrate and invade liver cells, inside which they differentiate into schizonts in a process that takes just a few days, until the infected cell bursts releasing thousands of merozoites that proceed to invade erythrocytes, thereby initiating the intraerythrocytic cycle. Merozoite recognition and invasion of erythrocytes is a complex process that involves multiple interactions between host-cell receptors and merozoite surface molecules. Many of these merozoite surface molecules are members of the merozoite surface protein (MSP) family (Cowman et al., 2002), and have been considered promising vaccine candidates due to their exposure to the host's immune system. However, their function and structure remain to be fully understood. To date, 11 MSPs have been identified in *P. falciparum*, some of which (i.e., MSP-1, -4, -5, -8 and -10) are bound to the cell membrane by a glycosylphosphatidylinositol (GPI) anchor and have one or two epidermal growth factor (EGF)-like domains. Others, such as MSP-3, -6 and -7, are non-covalently associated to the merozoite surface, forming complexes

* Corresponding author. Address: Molecular Biology Department, Fundación Instituto de Inmunología de Colombia (FIDIC), Carrera 50 No. 26-20, Bogotá, Colombia. Fax: +57 1 4815269.
*E-mail address:* mapatarr@fidic.org.co (M.A. Patarroyo).

1055-7903/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.ympev.2010.02.017

with other MSPs (Cowman et al., 2002; Mello et al., 2004; Kauth et al., 2006). Indeed, MSP-1 forms a protein-complex with MSP-6 and MSP-7 and is involved in the initial parasite–erythrocyte interaction (Pachebat et al., 2001; Kauth et al., 2006). The *P. falciparum msp-7* gene (*Pfmsp-7*) contains a single exon coding for a 42-kDa protein that is expressed during the schizont stage. This protein undergoes several proteolytic cleavages in its N-terminal region resulting in a 22-kDa fragment (Pachebat et al., 2001; Kauth et al., 2006) that remains bound to the MSP-1 complex, possibly playing a stabilization role.

We have previously characterized a *Pfmsp-7* homologue in *P. vivax* named *Pvmsp-7* (Pv082695), which is encoded in a single exon 937-bp gene (Mongui et al., 2006) and shows high identity and similarity values with upstream and downstream genes. Studies have shown that *msp-7* belongs to a gene family with variable gene copy number within *Plasmodium* species. In *P. falciparum*, expression assays with members of this family (MSP-7 and MSP-7-related proteins (MSRPs)) suggest that these genes are expressed simultaneously but regulated independently (Mello et al., 2002). Furthermore, it has been shown that some MSRPs have greater affinity to MSP-1 than MSP-7 and that immunization with MSRPs induces protection in mice (Mello et al., 2004). Moreover, disruption of the *msp*-7 gene in *P. berghei* (*Pbmsp-7*) reduces the parasite's ability to invade erythrocytes during early infection stages compared to wild-type lines, yet invasion efficacy is fully restored a few days later (Tewari et al., 2005). In *P. falciparum*, MSRP mRNA has been detected in parasite blood stages and, although the encoded proteins (PfMSRP-1 and -2) co-localize with MSP-1, they do not seem to be part of the MSP-1 complex (Mello et al., 2002).

To date, the available data suggest that MSP-7 and MSRPs play an important role in merozoite invasion to erythrocytes, yet little is known about their gene organization, structure and evolution in *Plasmodium* species. Such analysis could provide important clues to understanding the function of this gene family and could provide insights to current efforts in vaccine development. In this study, we have analyzed the *msp-7* chromosomal region in seven *Plasmodium* species and have evaluated phylogenetic relationships of *msp-7* genes within and between species. In addition, we have assessed the possible mechanisms that have driven the evolution of this gene family.

## 2. Materials and methods

### 2.1. Sequence data

DNA and amino acid sequences of *msp-7* from 7 *Plasmodium* species (*P. vivax*, *P. knowlesi*, *P. yoelii*, *P. berghei*, *P. chabaudi*, *P. reichenowi* and *P. falciparum*) were obtained from the PlasmoDB (http://plasmodb.org/plasmo/), Sanger (http://www.sanger.ac.uk/) and OrthoMCL (http://orthomcl.cbil.upenn.edu/cgi-bin/OrthoMclWeb.cgi) databases. Since several partial sequences are reported in these databases, missing regions were obtained using tBLASTn (http://www.sanger.ac.uk/DataSearch/blast.shtml, http://tigrblast.tigr.org/er-blast/index.cgi?project=pya1 and http://plasmodb.org/plasmo/). Framework genes Pf13_0190 and MAL13P1.177 (and their homologous) flanked the genomic region containing the *msp-7* family. Since most *msp-7* genes have a single exon, contigs were analyzed using ORF Finder (http://www.ncbi.nlm.nih.gov/projects/gorf/) and Gene Runner tools to identify open reading frames (ORFs) encoding proteins of more than 500 amino acids (Table 1). Deduced amino acid sequences obtained with Gene Runner were aligned with the sequences obtained from the databases mentioned above in order to confirm that they indeed corresponded to annotated *msp-7* family genes.

### 2.2. Sequence alignment and phylogenetic tree reconstruction

Amino acid and DNA sequence alignments were performed with MUSCLE (multiple sequence comparison by log-expectation) (Edgar, 2004) available at http://www.ebi.ac.uk/Tools/muscle/index.html and manually edited using GeneDoc software (Nicholas and Nicholas, 1997). The best models for both amino acid and nucleotide substitutions were selected by Akaike's information criterion using the ProtTest program (Abascal et al., 2005) and the ModelTest methodology (Posada and Crandall, 1998), respectively. The phylogenetic trees from the amino acid alignments were inferred through distance and Bayesian methods. The distance method used in this study was neighbor joining (NJ) with the JTT model of amino acid substitution, applying different rates among the alignment sites according to a Gamma distribution. Bayesian phylogenetic analysis (BY) was conducted with a Metropolis-coupled Markov chain Monte Carlo (mcmc) algorithm (Altekar et al., 2004). Analyses were run for 1 million generations with the mixed model option and the JTT + G model. For the latter model, the frequencies of all amino acids were changed by the observed amino acid frequencies in our data set. Trees for DNA alignments were constructed using the NJ method and the Bayesian phylogenetic analysis. For this latter analysis, the mcmc algorithm was run for 1 million generations with the GTR + G + I model. The reliability of both protein and DNA tree topologies was evaluated by bootstrap using 1000 iterations. In Bayesian analysis, the *sump* and *sumt* commands were used to tabulate posterior probabilities and build consensus trees. Distance method analyses were performed with MEGA v.4 software (Tamura et al., 2007) and Bayesian inference with MrBayes v.3.1 software (Ronquist and Huelsenbeck, 2003).

### 2.3. Estimation of divergence times

Two different likelihood ratio tests were performed in order to evaluate the molecular clock hypothesis. The first analysis was used to assess a global clock for every sub-tree of the entire phylogeny, whereas the second one evaluated a local clock by constraining only a rooted sub-tree of the phylogeny. Divergence dates were obtained for Bayesian mcmc analyses by using BEATS software (Drummond and Rambaut, 2007). For the obtained dates, both options were used: relaxed clock (Drummond et al., 2006) and strict clock using the following data as calibration dates: Pviv/Pkno 30.5 ± 0.4 MYA, Pber/Pyoe 17.9 ± 4.5 MYA and Pfal/Prei 8.9 ± 0.4 MYA (Escalante et al., 1995).

### 2.4. Evolutionary analysis to test functional divergence

We estimated the rate between the mean number of non-synonymous substitutions per non-synonymous sites ($d_N$) and the mean number of synonymous substitutions per synonymous sites ($d_S$) to determine the type of selective pressure affecting the evolution of the *msp-7* multigene family and establish whether they have evolved by functional divergence or functional redundancy. The Nei–Gojobori's method was used to estimate $d_N$ and $d_S$ (Nei and Gojobori, 1986), and the differences between both values were assessed by applying a *Z*-test with MEGA v.4 software (Tamura et al., 2007). Evolutionary rates ($\omega$) were estimated with the Jukes–Cantor correction and differences between clades were identified by implementing a pairwise Student's *t*-test. A likelihood-based analysis was carried out by using HyPhy software (Pond et al., 2005) to asses if the evolutionary rates in the sequences varied from site to site. For this purpose the MG94xHKY85 matrix had to be modified (Pond et al., 2005) to allow each site to have its own synonymous and non-synonymous rates. Finally, we used an approach proposed by Yang et al. (2000) who employed the empirical Bayes technique to compute the posterior probability for testing