Available online at www.sciencedirect.com
ScienceDirect
Molecular & Biochemical Parasitology 155 (2007) 103–112
MOLECULAR & BIOCHEMICAL PARASITOLOGY

Structural polymorphism and diversifying selection on the pregnancy malaria vaccine candidate VAR2CSA

Joseph Bockhorst[a,b], Fangli Lu[c,1], Joel H. Janes[d], Jon Keebler[e], Benoit Gamain[f], Philip Awadalla[e], Xin-zhuan Su[c], Ram Samudrala[g], Nebojsa Jojic[a], Joseph D. Smith[h,d,*]

[a] *Microsoft Research, Seattle, WA, USA*
[b] *University of Wisconsin—Milwaukee, WI, USA*
[c] *Laboratory of Malaria and Vector Research, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, MD, USA*
[d] *Department of Pathobiology, University of Washington, Seattle, WA, USA*
[e] *Department of Genetics, North Carolina State University, Raleigh, NC, USA*
[f] *Unité de Biologie des Interactions Hôte-Parasite, Institut Pasteur, Paris, France*
[g] *Department of Microbiology, University of Washington, Seattle, WA, USA*
[h] *Seattle Biomedical Research Institute, Seattle, WA, USA*

Received 7 April 2007; received in revised form 11 June 2007; accepted 12 June 2007
Available online 26 June 2007

Abstract

VAR2CSA is the main candidate for a pregnancy malaria vaccine, but vaccine development may be complicated by sequence polymorphism. Here, we obtained partial or full-length *var2CSA* sequences from 106 parasites and applied novel computational methods and three-dimensional modeling to investigate VAR2CSA geographic variation and selection pressure. Our analysis reveals structural patterns of VAR2CSA sequence variation in which polymorphic sites group into segments of limited diversity. Within these segments, two or three basic types characterize a substantial majority of the parasite samples. Comparison to the primate malaria *Plasmodium reichenowi* shows that these basic types have ancient origins. Globally, *var2CSA* genes are comprised of a mosaic of these ancestral polymorphic segments that have recombined extensively between *var2CSA* alleles. Three-dimensional modeling reveals that polymorphic segments concentrate in flexible loops at characteristic locations in the six VAR2CSA Duffy binding-like (DBL) adhesion domains. Individual DBL domain surfaces have distinct patterns of diversifying selection, suggesting that limited and differing portions of each DBL domain are targeted by host antibody. Since standard phylogenetic tree analysis is inadequate for highly recombining genes like *var2CSA*, we developed a novel phylogenetic approach that incorporates recombination and tracks new mutations in segment types. In the resulting tree, *P. reichenowi* is confirmed as an outlier and African and Asian *P. falciparum* isolates have slightly diverged. These findings validate a new approach to modeling protein evolution in the presence of frequent recombination and provide a clearer understanding of how *var* gene products function as immunoevasive binding ligands.

Keywords: Antigenic variation; Malaria; *Plasmodium falciparum*; *var* Genes

1. Introduction

Pregnancy associated malaria (PAM), a substantial cause of disease and death in pregnant women and newborns, is associated with a unique subset of *Plasmodium falciparum*-infected erythrocytes (IE) that bind chondroitin sulfate A (CSA) in the placenta [1–3]. Pregnant women acquire protective antibodies that cross-react with geographically diverse placental isolates [4–6] suggesting that surface molecule(s) expressed by PAM

Abbreviations: CSA, chondroitin sulfate A; DBL, Duffy binding-like; PfEMP1, *Plasmodium falciparum* erythrocyte membrane protein 1; PAM, pregnancy associated malaria

* Corresponding author at: Seattle Biomedical Research Institute, Seattle, WA, USA. Tel.: +1 206 256 7384; fax: +1 206 256 7229.
E-mail address: joe.smith@sbri.org (J.D. Smith).
[1] Present address: Sun Yat-sen University, Guangzhou, PR China.

0166-6851/$ – see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.molbiopara.2007.06.007

infected erythrocytes have conserved epitopes and, thus, that a PAM vaccine may be possible.

The specific targets of PAM immunity are currently being investigated, but recent evidence suggests that VAR2CSA, a member of the *P. falciparum* erythrocyte membrane 1 (PfEMP1) protein family, may have an important role in PAM disease and immunity [7]. PfEMP1 proteins are clonally variant parasite adhesion ligands expressed at the surface of infected erythrocytes [8]. Each parasite genome encodes approximately 60 different PfEMP1 proteins, or *var* genes, but only expresses one PfEMP1 protein at a time [9]. The *var2CSA* gene is unusual for the *var* gene family because it is found in all parasite isolates and has been shown to be transcriptionally upregulated in both placental isolates [10,11] and laboratory parasites selected to bind CSA [12]. VAR2CSA contains multiple CSA binding Duffy binding-like (DBL) domains [13] and appears to be one of the few or only PfEMP1 proteins mediating CSA binding because parasites in which the gene is experimentally disrupted lose the ability to adhere strongly to CSA [14,15]. Furthermore, VAR2CSA is the target of maternal antibodies [16–19] making it the leading candidate for a pregnancy malaria vaccine.

Because the VAR2CSA protein displays extensive sequence and antigenic polymorphism [12,16,17,20,21], designing an effective PAM vaccine will likely require detailed understanding of VAR2CSA sequence, structural and geographic variation. The three-dimensional structures for DBL domains contained in the binding regions of two different erythrocyte binding ligands involved in erythrocyte invasion were recently solved [22,23]. Despite limited sequence similarity, these domains were found to have highly related structures suggesting that all DBL domains, including those in VAR2CSA, may share a similar structure. This opens up new opportunities to investigate how VAR2CSA [16,17], and PfEMP1 proteins in general, have evolved as immunoevasive binding ligands [24].

A challenge of studying evolution of *var2CSA*, (and other highly recombinogenic gene families), is that recombination reduces the effectiveness of traditional phylogenetic approaches, which assume evolution via point mutation [25]. An additional practical complication is that *var2CSA* is large (∼10 kb), and consequently most gene comparisons have relied on small gene fragments. An important problem in computation biology is discovering correlations among nearby positions in amino acid or DNA alignments. The immune system, for example, recognizes short stretches of amino acids in pathogen proteins. Thus, understanding patterns of sequence diversity is important in vaccine design. In this study, we amplified full-length VAR2CSA sequences from a global collection of parasite isolates and applied novel "recombination aware" computational methods and molecular modeling to investigate the diversity of VAR2CSA.

2. Materials and methods

2.1. Parasite isolates

Genomic DNAs in this study were prepared from culture adapted parasite isolates that were previously published [26].

2.2. Amplification and sequencing of var2CSA sequences

Nearly complete or full-length *var2CSA* sequences were amplified from 10 different parasites isolates (Table S1) representing South East Asia (five isolates), East or West Africa (four isolates) or Central or South America (two isolates) using previously defined PCR conditions [20]. In brief, *var2CSA* sequences were amplified in a process of trial and error in two to five overlapping parts using published degenerate primers to the unique UpsE-type 5′ gene flanking sequence [20] and gene-specific primers to highly conserved coding regions. In addition, *var2CSA* sequences were collected from the HB3 sequencing project [Broad Institute of Harvard and MIT (http://www.broad.mit.edu)] and from the *P. falciparum* Ghana isolate at The Wellcome Trust Sanger Institute website at http://www.sanger.ac.uk/Projects/P_falciparum/.

2.3. Entropy calculations

Protein multiple sequence alignments are widely used to infer amino acid conservation within evolutionary related families. The entropy score, a common approach for measuring sequence variation [27], calculates amino acid variation at a single multiple alignment position using the formula:

$$S_{\text{entropy}} = -\sum_{i=1}^{20} p_i \log_2(p_i)$$

where p_i represents the observed frequency of residue type i in the aligned column. The minimum positional entropy (0) occurs at perfectly conserved positions and the maximum positional entropy ($4.32 = \log_2 20$) occurs at positions where all amino acids are observed with equal frequency. Following segmentation (see below), we compute the post-segmentation entropy of each position (see supplemental methods). A site's post-segmentation entropy, a number between zero and the site's positional entropy, is reduced if its variation strongly correlates with variations at other polymorphic sites in the same segment.

2.4. Segmentation

Segmentation analysis investigates correlations among multiple nearby polymorphic sites [28]. We begin by computing the optimal segmentation of our VAR2CSA multiple alignment using a maximum segment length of 15 amino acids and a maximum of three types per segment. To create longer segments, we subsequently tried merging adjacent segments; however, this did not lead to better segmentations. We also tried different maximum numbers of segment types, but three appeared to capture most of the variation in the VAR2CSA dataset (data not shown). The output of the segmentation process consists of (i) the segment boundaries, (ii) an identification of the total number of types found at each segment (either two or three), (iii) a type assignment to each sequence in each segment and (iv) a probabilistic sequence model of each segment's sequences (see supplemental methods). We use these models to assign types. Qualitatively, the segmentation procedure places segment boundaries so that polymorphic sites in the same segment are

strongly correlated while nearby polymorphic sites in different segments are less (or not) correlated. To assess the significance of the correlations implicit in the segmentation, we used a 10-fold cross validation methodology [29] to compare sequence models produced via segmentation with sequence models that assume mutations happen independently among sites (independent model). The total cross-validated log likelihood is much greater under the segmentation model (−24132) than under the independent models (−31476) (p-value $< 10^{-6}$ in a two-tailed paired t-test) indicating that the segmentation model more accurately predicts unseen sequences than the independent model.

2.5. Phylogenetic analysis

For the phylogenetic analysis we used the dnaml program of the PHYLIP package (http://evolution.genetics.washington.edu/phylip.html). Trees were compared using standard nucleotide alignments and type expanded nucleotide alignments. The type expanded alignment takes the output from the segmentation analysis and specifically compares nucleotide variation occurring within each of the basic segment types at that segment position thereby avoiding comparisons between different segment types because these are most likely introduced through gene conversion/recombination between *var2CSA* sequences. Question marks were introduced as gaps in the type expanded nucleotide alignment for segment types not present in a specific *var2CSA* sequence. The resulting comparison leads to what we term 'population trees', which emphasize new mutation in ancestral recombination blocks. This approach treats each observed sequence at a leaf of the phylogenetic tree as a representative of a recombining parasite population that contains all types at each segment. The leaf sequences have one observed type per segment (i.e., from that parasite genotype) and the other types in the parasite population are hidden. The interior nodes of the tree also represent populations whose sequences (as in standard tools) are hidden and to be inferred through phylogeny.

2.6. Evolutionary analysis

To test for positive selection among amino acid encoding codons we calculated estimates of rates of non-synonymous

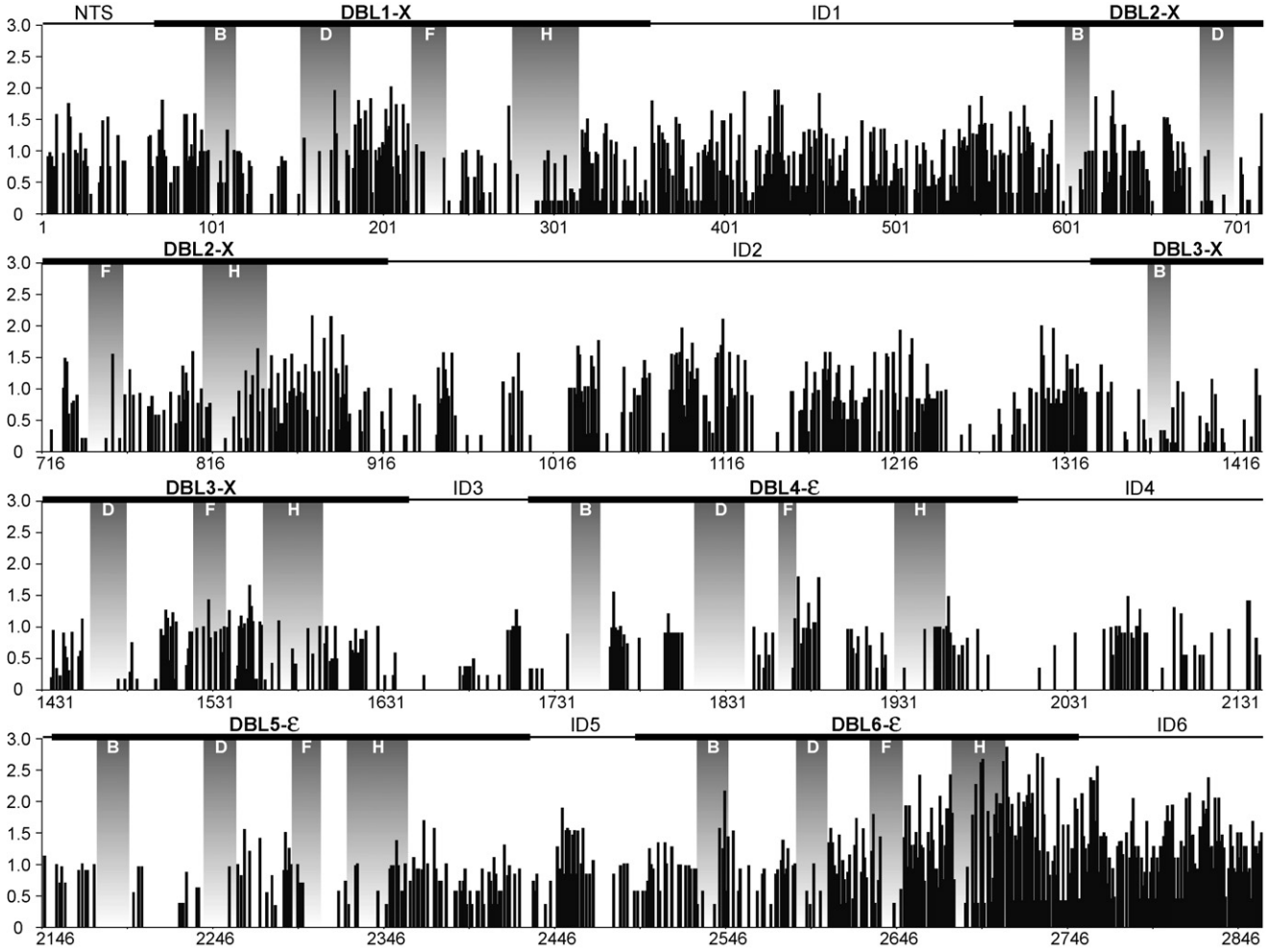

Fig. 1. Positional entropy values across a VAR2CSA amino acid alignment. Positional entropy values were calculated from a multiple alignment of 106 complete and partial VAR2CSA extracellular sequences. Entropy values were derived from all available sequences at each alignment position (range 11–55 sequences). The N-terminal segment (NTS), DBL and interdomain regions (ID) are labeled. The DBL semi-conserved homology blocks B, D, F and H are shown to scale.

(d*N*) and synonymous substitution (d*S*) among codons. We calculated these ratios, henceforth called ω, for each codon of the whole gene to ask whether particular codons were evolving under positive selection in order to ask whether a codon was under positive selection ($\omega > 1$) using a model of natural selection that allows for variable ω among codons. We used a Bayesian approach [30] to test for departures from neutrality among codons that allows for recombination to occur among lineages throughout the sequence (and hence, independent multiple genealogies).

2.7. Molecular modeling

The three-dimensional conformations of the IT4 VAR2CSA DBL domains were modeled on the EBA-175 region II [23] using the PROTINFO structure prediction server (http://protinfo.compbio.washington.edu). Modeling was performed using the comparative modeling protocol, which has been shown to work well in the CASP protein structure prediction experiments [31,32]. Initial models were constructed using a minimum perturbation approach that aims to preserve as much information as possible from the template structure solved by X-ray diffraction (the template with the Protein Data Bank identifier 1zro was used). Variable side chains and main chains were then built using a graph-theory clique-finding approach that explores a variety of possible conformations for the respective side chains and main chains and finds the optimal combination using an all-atom scoring function [33]. These approaches are described in further detail in [34,35]. The positional entropies and polymorphic segments determined from the alignment of 18 fuller-length *var2CSA* sequences were mapped onto the IT4var DBL models after correcting for indels.

3. Results and discussion

3.1. VAR2CSA DBL domains are structured into variable and semi-conserved blocks

Whereas VAR2CSA is the primary candidate for a PAM vaccine, relatively limited sequence information exists and most of this is partial gene fragments concentrated in only a small part of the protein. To acquire information about how VAR2CSA has diversified, we amplified full or nearly full-length *var2CSA* sequences from 10 parasite isolates from around the world, and collected all of the partial or complete *var2CSA* sequences present in GenBank or at *P. falciparum* genome sequencing projects (Table S1). In total, we compared 106 *var2CSA* sequences including the complete or nearly complete extracellular binding region from 18 *P. falciparum* isolates, 87 partial *var2CSA* sequences, and a partial VAR2CSA sequence from the chimpanzee malaria *Plasmodium reichenowi*.

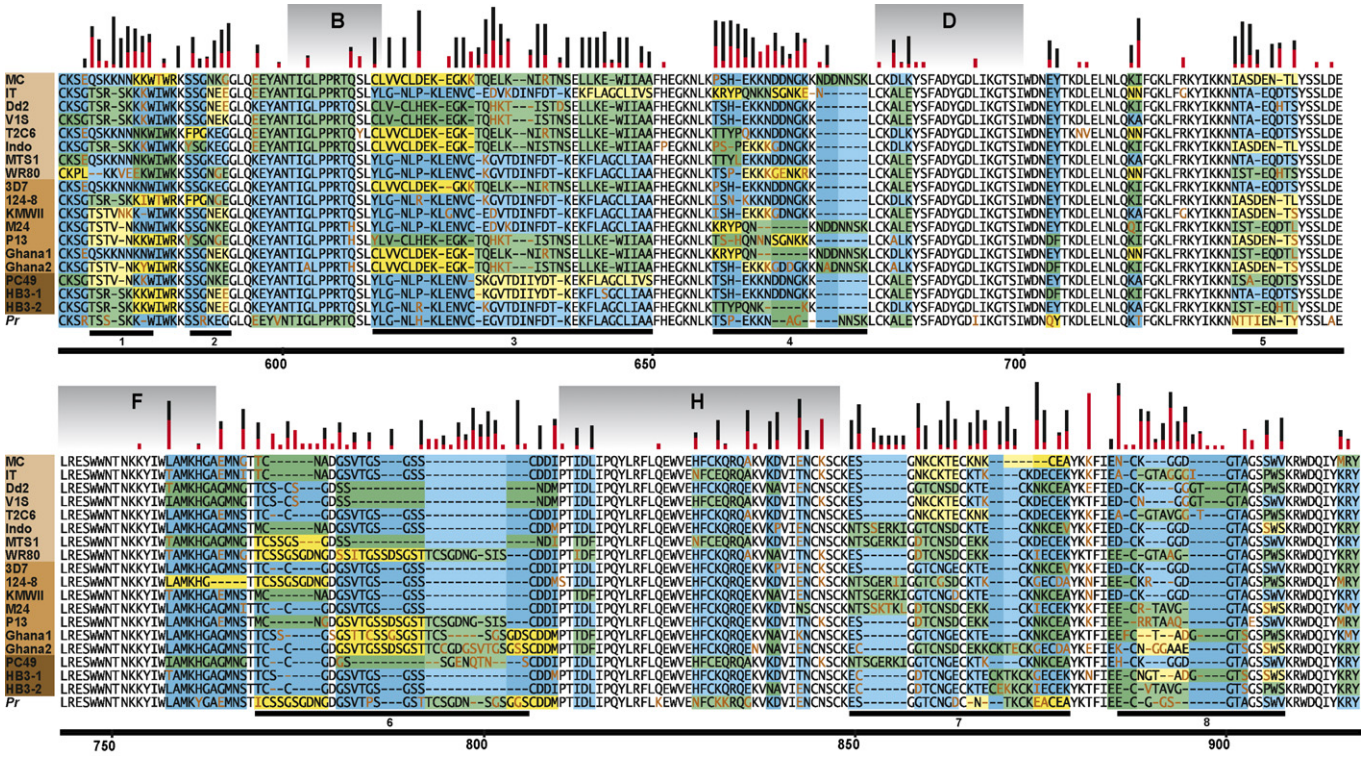

Fig. 2. Segmentation structure of the VAR2CSA DBL2X domain. A segmentation analysis was performed to identify local correlations in amino acid polymorphism. Regions under diversifying selection (d*N*/d*S* values > 1) are underlined. Within each segment, different types are shaded blue, yellow or green. Segment boundaries are indicated by unshaded positions or shading differences. Brown amino acids denote mutations from the consensus segment type. The B, D, F and H semi-conserved blocks are labeled. Bar heights indicate the amount of positional entropy prior to (black) or post-segmentation (red). Entropy calculations are based on a total of 106 VAR2CSA sequences and sequence tags. Parasite isolates are shaded by geographic origin; Asian (top), African (middle) and Central or South American (bottom). *Pr* denotes *P. reichenowi*.

Overall, the 11 most complete VAR2CSA sequences, representing parasite isolates from Asia, Africa and Central America, average 78% amino acid identity (range 75–83%). The resulting VAR2CSA alignment length was 2859 amino acids in which the maximum coverage per site in the alignment is 55 sequences (Table S1).

VAR2CSA is made up of six receptor-like DBL domains interspersed by variably sized interdomain (ID) regions. Previous sequence analysis revealed that DBL domains could be organized into 10 variable and 10 semi-conserved blocks (named A-J) [36]. The semi-conserved blocks correspond to structural scaffolding in solved DBL structures [22,23] and can be used as a frame of reference between DBL sequences. As the simplest measure of VAR2CSA diversification we assessed the positional entropy or amino acid conservation at each position in the alignment (Fig. 1). Ignoring gaps, 1321 (46.2%) of the alignment positions are polymorphic, and 1174 (41.1%) have a positional entropy > 0.32, (the entropy for a position with one mutation in 17 sequences). However, there exist substantial local correlations that decrease this complexity (discussed below). As expected, variable blocks have a greater concentration of both high entropy positions and gaps in the alignment (Figs. 1 and 2). Although the six DBL domains in VAR2CSA differ in amino acid conservation between 61 and 88% (Table S2), they all tend to have variable blocks adjacent to the B, D, F and H semi-conserved blocks. However, the distribution of variability differs between domains. For instance, there were variable blocks on both sides of semi-conserved block B in the DBL1, DBL2 and DBL6 domains, but only one side in the DBL4 and DBL5 domains, and the DBL3 domain was not highly variable at these locations (Fig. S1). These locations are also highly variable in alignments of all DBL sequences, suggesting that the DBL fold is relatively insensitive to sequence and length variability in these regions [24].

Unlike DBL domains, which have a variable/conserved block structure, no global patterns of variation apply to interdomain regions. Interdomain regions can be classified into larger regions (ID1, ID2 and ID4), which contain more than 100 amino acids, or smaller regions (ID3, ID5, ID6), which contain less than 100 amino acids (Fig. 1). Similar to DBL domains, sequence variability in the ID2 and ID4 regions concentrates in variable blocks (Fig. 1), suggesting these interdomain regions may also be preserving a specific three-dimensional fold. By comparison, ID1 has relatively few invariant residues, and these are not concentrated but rather are distributed throughout the ~200 amino acid region. Of the smaller interdomain regions, ID5 and ID6 are highly polymorphic while ID3 is extremely conserved, and may fold as part of the DBL domain [17]. Curiously, although the DBL6 and ID6 are membrane proximal and may have been expected to be less exposed, they are the most polymorphic (Table S2, Fig. S1).

3.2. *VAR2CSA mutations are not independent and gene diversification is associated with a high rate of segmental gene recombination/gene conversion*

To investigate multiple-site patterns of VAR2CSA polymorphism, we performed a segmentation analysis. Unlike positional

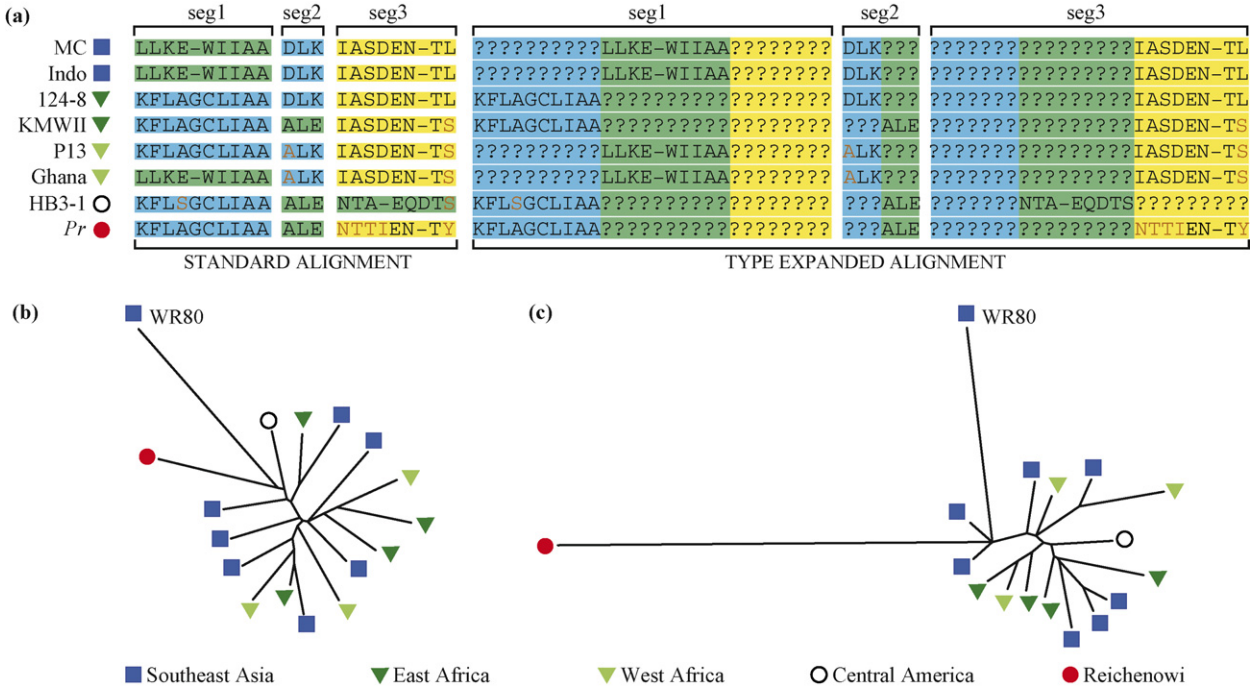

Fig. 3. Phylogenetic relationship between VAR2CSA sequences. (a) Example comparing a standard alignment to a type expanded alignment for three DBL2 segments. In the type expanded alignment, only segments of the same type are directly compared. Segment types absent from a specific VAR2CSA sequence are represented by question marks. Population trees are constructed by submitting the type expanded alignment to standard phylogenetic tools. A standard tree (b) and type expanded population tree (c) created from nucleotides of the 17 most complete VAR2CSA sequences between the beginning of the protein and the interdomain 2 (alignment length = 3405 nucleotides).

entropy, which treats sites as independent, segmentation analysis uses a statistical model that assumes the protein contains a number of 'correlated segments' (or simply 'segments') in which mutations are linked. Each segment may assume a number of basic types, where for simplicity a type can be thought of as a (possibly mutated) consensus sequence (see supporting materials). Under the segmentation model the total VAR2CSA entropy is decreased because many of the polymorphic sites are correlated (compare black and red bars, Figs. 2 and S1). The VAR2CSA alignment has 263 segments (average segment has 6.1 amino acids) that in total cover 1613 (56.4%) of the total positions and the majority (90.3%) of the strongly polymorphic (entropy > 0.32) positions, indicating the presence of substantial local correlations in amino acid polymorphism (Figs. 2 and S1). Correlated segments were identified in both variable and semi-conserved blocks.

In general, we observe two broad classes of differences among types in the same segment. Some types are relatively closely related to each other, suggesting they likely arose by point mutation. For example, the yellow (consensus KFLAG-CLIVS) and blue (consensus KFLAGCLIAA) types in the final segment in variable block 3 of DBL2 (Fig. 2). This is, however, the minority case as more typically types are quite different from each other (compare to the green type (consensus LLKE-WIIAA) in this same segment), suggesting that some mechanism other than point mutation brought them into alignment. Although within a segment, there is generally substantial divergence between types (21% identity across types at polymorphic sites compared to an overall 63% identity at polymorphic sites), the observed types are remarkably conserved between isolates. For instance, between the 105 *var2CSA* sequences and *P. reichenowi*, 70% of the polymorphic segments (4370/6216) exactly match one of the consensus types and an additional 22% differ by only one amino acid from consensus (Table 1). Furthermore, many of these basic types are present both across

Table 1
Variation at polymorphic segments

No. of mismatches from consensus types	Frequency among polymorphic segments[a]
0	4370 (70.3%)
1	1367 (22.0%)
2	308 (5.0%)
3	86 (1.4%)
4	50 (0.8%)
5	10 (0.1%)
6	13 (0.0%)
7	6
8	3
9	3

[a] Between 106 *var2CSA* sequences, a total of 6216 polymorphic segments compared.

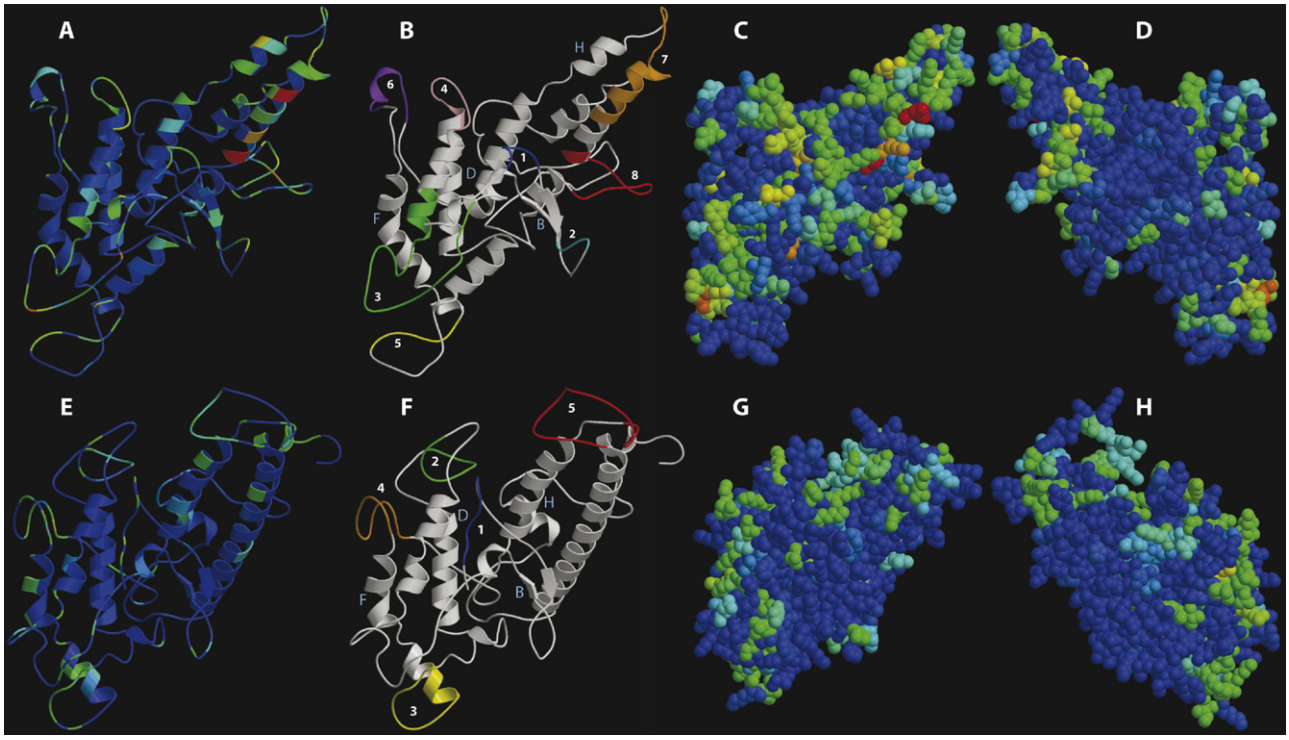

Fig. 4. Three-dimensional models of polymorphism in the VAR2CSA DBL2 and DBL3 domains. Domains were modeled based upon the EBA-175 F1 and F2 DBL domains [23]. Panels A–D correspond to VAR2CSA DBL2 and panels E–H to VAR2CSA DBL3. (A, E) The amino acid diversity or positional entropies (range: 0–2.1) determined from a multiple alignment of 18 full-length VAR2CSA sequences are shown using a temperature scale from blue to red (low to high entropy). (B, F) Mapping of variable sequence blocks determined by segmentation analysis. Polymorphic segments are colored and those under strong diversifying selection are numbered. Semi-conserved homology blocks B, D, F and H are indicated. Conserved residues are shown in grey. (C, D, G and H) Space-fill representation of positional entropy, panels (D) and (H) are rotated 180° relative to (C) and (G), respectively.

broad geographical regions and in *P. reichenowi* (Fig. S1) even though the primate malaria is estimated to have diverged from *P. falciparum* ~6–10 Mya [37,38]. Thus, we conclude that gene recombination/gene conversion of ancient and slowly mutating basic segment types likely creates the observed *var2CSA* combinatorial diversity and mosaicism (Figs. 3 and S1). In addition, evidence suggests that recombination sites typically fall between (and not within) correlated segments.

Using this data set, we investigated geographic relationships among VAR2CSA sequences to discover any geographically linked patterns of variation that may impact vaccine design. Standard phylogenetic tree approaches, however, assume sequences evolve through mutation and are not appropriate for studying highly recombinogenic gene families [25]. Indeed, standard VAR2CSA trees suggest some *P. falciparum* sequences are more distant from each other than from *P. reichenowi* (Fig. 3b). Alternative approaches that construct separate trees in each non-recombinant block are also not helpful [39]. First the extensive recombination of VAR2CSA would require hundreds of trees that would be challenging to comprehend. Second, and more importantly, the tree topologies would be determined primarily by large scale differences (i.e., type differences) between sequences even though such differences likely arose prior to *P. falciparum* continental separation and, thus, are not part of patterns of variation we wish to discover. Instead, we propose a novel phylogenetic analysis approach that tracks gene relationships based upon new mutations that have occurred since the basic segment types evolved. This approach differs from the multiple tree approach both because it produces a single tree and because recent evolution events determine the tree topology. Our 'population tree' approach assumes that entire populations of genes, each with all types for all segments present, evolve and spread independently. Fortunately, population trees can be obtained by submitting a 'type expanded' alignment as input to standard phylogenetic programs (Fig. 3). By avoiding alignments of different segment types, expanded

Table 2
Diversifying selection on VAR2CSA DBL domains

VAR2CSA region	Codon boundaries	Mean omega	Mean prob.[a]	VAR2CSA region	Codon boundaries	Mean omega	Mean prob.[a]
NTS	1–66	1.00	0.49	DBL4	1717–2002	0.46	0.18
DBL1	67–357	1.01	0.37	B BLOCK	1742–1757	0.04	0.00
VB1[b]	68–74	1.54	0.85	**VB1**	**1764–1774**	**1.48**	**0.84**
VB2	84–98	2.01	0.89	VB2	1796–1813	1.01	0.50
B BLOCK	96–114	0.96	0.59	D BLOCK	1814–1839	0.03	0.00
VB3	115–123	0.71	0.38	**F BLOCK**	**1863–1873**	**0.55**	**0.20**
VB4	139–143	0.41	0.04	**VB3**	**1872–1886**	**2.58**	**0.99**
D BLOCK	152–180	1.20	0.29	**VB4**	**1903–1915**	**0.86**	**0.28**
VB5[c]	**170–215**	**3.15**	**0.92**	H BLOCK	1930–1960	0.27	0.12
F BLOCK	217–237	0.34	0.17	VB5	1954–1965	0.93	0.46
VB6	**247–262**	**0.80**	**0.38**	DBL5	2152–2431	0.58	0.30
H BLOCK	276–315	0.20	0.03	**VB1**	**2167–2176**	**0.39**	**0.04**
VB7	**313–331**	**1.04**	**0.50**	B BLOCK	2178–2196	0.06	0.00
DBL2	570–917	1.05	0.54	VB2	2226–2241	0.87	0.38
VB1	573–581	1.38	0.85	D BLOCK	2241–2259	0.13	0.04
VB2	**587–592**	**1.85**	**0.88**	VB3	2260–2267	0.70	0.24
B BLOCK	600–612	0.46	0.29	**VB4**	**2286–2291**	**1.70**	**0.76**
VB3	**612–649**	**1.80**	**0.96**	F BLOCK	2292–2309	0.37	0.23
VB4	**658–678**	**1.81**	**0.92**	H BLOCK	2325–2359	0.57	0.31
D BLOCK	680–699	0.22	0.11	VB5	2349–2371	1.16	0.73
VB5	728–736	1.27	0.67	VB6	2387–2394	1.09	0.68
F BLOCK	743–763	0.21	0.11	VB7	2407–2419	1.16	0.73
VB6	769–805	1.16	0.74	DBL6	2493–2752	4.46	0.65
H BLOCK	811–846	0.59	0.29	VB1	2499–2529	1.11	0.65
VB7	**849–878**	**2.13**	**0.99**	**B BLOCK**	**2529–2546**	**1.34**	**0.46**
VB8	**885–907**	**1.42**	**0.50**	**VB2**	**2542–2550**	**2.52**	**0.74**
DBL3	1333–1647	0.52	0.25	**VB3**	**2578–2586**	**0.96**	**0.44**
VB1	**1335–1343**	**0.74**	**0.35**	D BLOCK	2587–2605	0.37	0.08
B BLOCK	**1365–1378**	**0.07**	**0.00**	**VB4**	**2606–2636**	**2.42**	**0.98**
VB2	1428–1455	1.19	0.62	F BLOCK	2630–2649	1.08	0.41
D BLOCK	1463–1480	0.07	0.01	**VB5**	**2650–2680**	**6.17**	**0.97**
VB3	1501–1517	1.21	0.74	**H BLOCK**	**2678–2709**	**9.07**	**0.76**
F BLOCK	1520–1539	1.00	0.60	**VB6**	**2690–2752**	**12.46**	**0.98**
VB4	**1537–1555**	**1.22**	**0.72**				
H BLOCK	1562–1596	0.26	0.03				
VB5	1597–1621	0.85	0.35				

[a] The posterior probability that a codon is subject to diversifying selection averaged across codons for the defined region.
[b] VB denotes variable block.
[c] Bolded regions have at least one codon inferred to be subject to significant positive selection.

alignments highlight new mutation. The tree built with the type expanded alignment correctly identifies *P. reichenowi* as an outlier and has more branch-like structure than the tree built with standard alignments (Fig. 3). Moreover, bootstrap analysis indicates that two sequences from the same geographical region are more likely to group in the same sub-tree than would be expected were there no geographical effect (p-value $< 10^{-5}$). Greater depth of VAR2CSA sequence coverage in multiple geographic regions will be important to determine if geographic considerations should be used in VAR2CSA vaccine selection.

3.3. Diversifying selection is highly biased on the DBL surface and differs in extent between the six VAR2CSA DBL domains

Although PfEMP1 proteins have a critical role in parasite immune evasion and pathogenesis [8], there has been limited investigation of diversity selection and this has primarily been focused on small protein regions [17,21]. Here, we analyzed selection pressures over all six VAR2CSA DBL domains and modeled the three-dimensional distribution of polymorphic sites in the VAR2CSA DBL2 and DBL3 domains. This analysis supports the hypothesis that the semi-conserved B, D, F and H blocks correspond to structural or scaffolding elements in DBL domains, and that variable blocks are either polymorphic loops that connect the scaffolding elements or surface exposed residues on the DBL scaffolding itself (Fig. 4). Notably, most polymorphic segments appear to be under strong diversifying or balancing selection (d*N*/d*S* ratios, or ω values, significantly >1.0, Table 2), consistent with B cell epitope mapping showing that polymorphic segments are subject to antibody pressure [16,17]. Conversely, semi-conserved blocks are more constrained, although individual residues exposed on the DBL scaffold can be polymorphic and have large d*N*/d*S* ratios.

Multi-domain comparisons show that VAR2CSA polymorphism is highly biased to one surface of the DBL2 fold and differs in total surface distribution between the DBL2 and DBL3 domains (Fig. 4) [17]. For instance, polymorphic segments 7 and 8 are highly diverse in the DBL2 domain, but the equivalent residues are nearly invariant in the DBL3 domain. Moreover, the pattern and extent of diversifying selection differs between individual domains (Table 2). These findings imply either that invariant surfaces are poorly targeted by antibody or may be less accessible in the native protein.

DBL domains have been found to engage different host receptors using distinct mechanisms and different binding sites on opposite sides of the DBL fold [22,23]. Although the precise CSA binding contact residues have not been defined in the VAR2CSA DBL domains, it may be significant that variable blocks 1, 2 and 3 in the DBL2 domain are located in the equivalent EBA-175::sialic acid binding region and variable blocks 4 and 6 are located in the equivalent Pkα-DBL::Duffy antigen binding region (Fig. 4). Both locations were also predicted to have variable loops in the DBL3 domain, although there was significantly less variability adjacent to the equivalent sialic acid binding site. Therefore, potential host interaction sites in VAR2CSA DBL domains are surrounded by amino acids that are under strong selection for amino acid diversification, presumably for immune evasion.

4. Concluding remarks

Our findings support a model of VAR2CSA evolution in which protein diversification is concentrated at flexible loops in the DBL fold and other surface exposed residues. Unexpectedly, there are extensive local correlations in VAR2CSA polymorphism and the same variable loop types were found in geographically diverse parasite isolates. Comparison to *P. reichenowi* suggests the polymorphic segments have ancient origins highlighting the importance of gene recombination/gene conversion in *var2CSA* diversification. While VAR2CSA does not appear to recombine extensively with other *var* genes [20], it is tempting to speculate that a similar mechanism may exist for the broader family of recombining *var* genes because polymorphic blocks are sometimes shared between otherwise distinct *var* sequences [40]. This mechanism may provide significant flexibility in DBL domains to bind and sequester infected erythrocytes from blood circulation and potentially introduces variability near previously defined DBL-host interaction sites [22,23].

A question raised by these findings is how do pregnant women develop a broad antibody response to an antigen as polymorphic as VAR2CSA? While the specific targets of maternal antibodies are only beginning to be defined [16,17], immune investigations suggest that CSA binding parasite lines display both common and diverse epitopes [5,41–44]. Although it has been postulated that antibody cross-recognition of placental isolates may be due to highly conserved epitopes, a distinct possibility suggested by this analysis is that antibody cross-reactivity may be caused by overlapping polymorphism between geographically diverse parasite isolates. During typical infections, parasites switch between PfEMP1 proteins to evade immunity [8]. For placental adherent isolates, the options for switching appear limited [14,15]. One possibility is that exposure to multiple different parasite genotypes during pregnancy may broaden the maternal antibody response to diverse VAR2CSA alleles. Curiously, although the VAR2CSA protein is highly polymorphic, a surprising amount of predicted DBL surfaces are invariant and diversifying selection differs between the six DBL domains in VAR2CSA. These results suggest that only a limited portion of each DBL domain is actively seen by the host immune system and may indicate the presence of previously unrecognized domain interactions within or between VAR2CSA and other proteins at the IE surface. If invariant residues or surfaces are exposed in the native protein, then vaccine efforts will need to redirect antibody responses to these less polymorphic sites. Conversely, if PAM immunity targets polymorphic segments, then information from segmentation analysis can now be applied in vaccine strategies to broaden antibody reactivity. The segmentation methodology may have wider application for uncovering sequence patterns in rapidly evolving and recombining gene families or highly polymorphic vaccine targets.

Acknowledgements

This research was supported by the Bill & Melinda Gates Foundation (J.D.S.), European Malaria Vaccine Initiative (B.G.) (grant no. 01/2005), and by the Intramural Research Program of the NIH, National Institute of Allergy and Infectious Diseases. VAR2CSA sequence data for the HB3 and Ghana isolates were obtained from the *Plasmodium falciparum* HB3 Sequencing Project [Broad Institute of Harvard and MIT (http://www.broad.mit.edu)] and from the *P. falciparum* Ghana isolate at The Wellcome Trust Sanger Institute website at http://www.sanger.ac.uk/Projects/P_falciparum/. Computational support provided by Technical Computing at Microsoft Research.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.molbiopara.2007.06.007.

References

[1] Beeson JG, Brown GV, Molyneux ME, et al. *Plasmodium falciparum* isolates from infected pregnant women and children are associated with distinct adhesive and antigenic properties. J Infect Dis 1999;180:464–72.
[2] Brabin BJ, Romagosa C, Abdelgalil S, et al. The sick placenta—the role of malaria. Placenta 2004;25:359–78.
[3] Fried M, Duffy PE. Adherence of *Plasmodium falciparum* to chondroitin sulfate A in the human placenta. Science 1996;272:1502–4.
[4] Duffy PE, Fried M. Antibodies that inhibit *Plasmodium falciparum* adhesion to chondroitin sulfate A are associated with increased birth weight and the gestational age of newborns. Infect Immun 2003;71:6620–3.
[5] Fried M, Nosten F, Brockman A, et al. Maternal antibodies block malaria. Nature 1998;395:851–2.
[6] Staalsoe T, Shulman CE, Bulmer JN, et al. Variant surface antigen-specific IgG and protection against clinical consequences of pregnancy-associated *Plasmodium falciparum* malaria. Lancet 2004;363:283–9.
[7] Smith JD, Deitsch KW. Pregnancy-associated malaria and the prospects for syndrome-specific antimalaria vaccines. J Exp Med 2004;200:1093–7.
[8] Miller LH, Baruch DI, Marsh K, et al. The pathogenic basis of malaria. Nature 2002;415:673–9.
[9] Frank M, Deitsch K. Activation, silencing and mutually exclusive expression within the *var* gene family of *Plasmodium falciparum*. Int J Parasitol 2006;36:975–85.
[10] Duffy MF, Caragounis A, Noviyanti R, et al. Transcribed *var* genes associated with placental malaria in Malawian women. Infect Immun 2006;74:4875–83.
[11] Tuikue Ndam NG, Salanti A, Bertin G, et al. High level of *var2csa* transcription by *Plasmodium falciparum* isolated from the placenta. J Infect Dis 2005;192:331–5.
[12] Salanti A, Staalsoe T, Lavstsen T, et al. Selective upregulation of a single distinctly structured *var* gene in chondroitin sulphate A-adhering *Plasmodium falciparum* involved in pregnancy-associated malaria. Mol Microbiol 2003;49:179–91.
[13] Gamain B, Trimnell AR, Scheidig C, et al. Identification of multiple chondroitin sulfate A (CSA)-binding domains in the var2CSA gene transcribed in CSA-binding parasites. J Infect Dis 2005;191:1010–3.
[14] Duffy MF, Maier AG, Byrne TJ, et al. VAR2CSA is the principal ligand for chondroitin sulfate A in two allogeneic isolates of *Plasmodium falciparum*. Mol Biochem Parasitol 2006;148:117–24.
[15] Viebig NK, Gamain B, Scheidig C, et al. A single member of the *Plasmodium falciparum var* multigene family determines cytoadhesion to the placental receptor chondroitin sulphate A. EMBO Rep 2005;6:775–81.
[16] Barfod L, Bernasconi NL, Dahlback M, et al. Human pregnancy-associated malaria-specific B cells target polymorphic, conformational epitopes in VAR2CSA. Mol Microbiol 2006.
[17] Dahlback M, Rask TS, Andersen PH, et al. Epitope mapping and topographic analysis of VAR2CSA DBL3X involved in *P. falciparum* placental sequestration. PLoS Pathog 2006;2:e124.
[18] Salanti A, Dahlback M, Turner L, et al. Evidence for the involvement of VAR2CSA in pregnancy-associated malaria. J Exp Med 2004;200:1197–203.
[19] Tuikue Ndam NG, Salanti A, Le-Hesran JY, et al. Dynamics of anti-VAR2CSA immunoglobulin G response in a cohort of senegalese pregnant women. J Infect Dis 2006;193:713–20.
[20] Kraemer SM, Smith JD. Evidence for the importance of genetic structuring to the structural and functional specialization of the *Plasmodium falciparum var* gene family. Mol Microbiol 2003;50:1527–38.
[21] Trimnell AR, Kraemer SM, Mukherjee S, et al. Global genetic diversity and evolution of *var* genes associated with placental and severe childhood malaria. Mol Biochem Parasitol 2006;148:169–80.
[22] Singh SK, Hora R, Belrhali H, et al. Structural basis for Duffy recognition by the malaria parasite Duffy-binding-like domain. Nature 2006;439:741–4.
[23] Tolia NH, Enemark EJ, Sim BK, et al. Structural basis for the EBA-175 erythrocyte invasion pathway of the malaria parasite *Plasmodium falciparum*. Cell 2005;122:183–93.
[24] Howell DP, Samudrala R, Smith JD. Disguising itself–insights into *Plasmodium falciparum* binding and immune evasion from the DBL crystal structure. Mol Biochem Parasitol 2006;148:1–9.
[25] Awadalla P. The evolutionary genomics of pathogen recombination. Nat Rev Genet 2003;4:50–60.
[26] Wootton JC, Feng X, Ferdig MT, et al. Genetic diversity and chloroquine selective sweeps in *Plasmodium falciparum*. Nature 2002;418:320–3.
[27] Valdar WS. Scoring residue conservation. Proteins 2002;48:227–41.
[28] Bockhorst J, Jojic N. Discovering patterns in biological sequences by optimal segmentation. Proceedings of the 23rd International Conference on Uncertainty in Artificial Intelligence, in press.
[29] Mitchell T. Machine learning. 1st ed. McGraw-Hill; 1997.
[30] Wilson DJ, McVean G. Estimating diversifying selection and functional constraint in the presence of recombination. Genetics 2006;172:1411–25.
[31] Hung LH, Samudrala R. PROTINFO: secondary and tertiary protein structure prediction. Nucleic Acids Res 2003;31:3296–9.
[32] Hung LH, Ngan SC, Liu T, et al. PROTINFO: new algorithms for enhanced protein structure predictions. Nucleic Acids Res 2005;33:W77–80.
[33] Samudrala R, Moult J. An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. J Mol Biol 1998;275:895–916.
[34] Samudrala R, Moult J. A graph-theoretic algorithm for comparative modeling of protein structure. J Mol Biol 1998;279:287–302.
[35] Samudrala R, Levitt M. A comprehensive analysis of 40 blind protein structure predictions. BMC Struct Biol 2002;2:3.
[36] Smith JD, Subramanian G, Gamain B, et al. Classification of adhesive domains in the *Plasmodium falciparum* erythrocyte membrane protein 1 family. Mol Biochem Parasitol 2000;110:293–310.
[37] Escalante AA, Barrio E, Ayala FJ. Evolutionary origin of human and primate malarias: evidence from the circumsporozoite protein gene. Mol Biol Evol 1995;12:616–26.
[38] Escalante AA, Ayala FJ. Evolutionary origin of Plasmodium and other Apicomplexa based on rRNA genes. Proc Natl Acad Sci USA 1995;92:5793–7.
[39] Kosakovsky Pond SL, Posada D, Gravenor MB, et al. Automated phylogenetic detection of recombination using a genetic algorithm. Mol Biol Evol 2006;23:1891–901.
[40] Ward CP, Clottey GT, Dorris M, et al. Analysis of *Plasmodium falciparum* PfEMP-1/*var* genes suggests that recombination rearranges constrained sequences. Mol Biochem Parasitol 1999;102:167–77.

[41] Beeson JG, Mann EJ, Elliott SR, et al. Antibodies to variant surface antigens of *Plasmodium falciparum*-infected erythrocytes and adhesion inhibitory antibodies are associated with placental malaria and have overlapping and distinct targets. J Infect Dis 2004;189:540–51.

[42] Beeson JG, Mann EJ, Byrne TJ, et al. Antigenic differences and conservation among placental *Plasmodium falciparum*-infected erythrocytes and acquisition of variant-specific and cross-reactive antibodies. J Infect Dis 2006;193:721–30.

[43] Elliott SR, Duffy MF, Byrne TJ, et al. Cross-reactive surface epitopes on chondroitin sulfate A-adherent *Plasmodium falciparum*-infected erythrocytes are associated with transcription of var2csa. Infect Immun 2005;73:2848–56.

[44] Tuikue Ndam NG, Fievet N, Bertin G, et al. Variable adhesion abilities and overlapping antigenic properties in placental *Plasmodium falciparum* isolates. J Infect Dis 2004;190:2001–9.