RESEARCH ARTICLE

A re-assessment of gene-tag classification approaches for describing *var* gene expression patterns during human *Plasmodium falciparum* malaria parasite infections [version 1; referees: 2 approved]

George Githinji [1], Peter C. Bull[2]

[1]Kenya Medical Research Institute (KEMRI)-Wellcome Trust Research Programme, Centre for Geographic Medicine Research-Coast, Kilifi, Kenya
[2]Department of Pathology, University of Cambridge, Cambridge, UK

v1 **First published:** 19 Sep 2017, **2**:86 (doi: 10.12688/wellcomeopenres.12053.1)
Latest published: 19 Sep 2017, **2**:86 (doi: 10.12688/wellcomeopenres.12053.1)

Abstract

PfEMP1 are variant parasite antigens that are inserted on the surface of *Plasmodium falciparum* infected erythrocytes (IE). Through interactions with various host molecules, PfEMP1 mediate IE sequestration in tissues and play a key role in the pathology of severe malaria. PfEMP1 is encoded by a diverse multi-gene family called *var*. Previous studies have shown that that expression of specific subsets of *var* genes are associated with low levels of host immunity and severe malaria. However, in most clinical studies to date, full-length *var* gene sequences were unavailable and various approaches have been used to make comparisons between *var* gene expression profiles in different parasite isolates using limited information. Several studies have relied on the classification of a 300 – 500 base-pair “DBLα tag” region in the DBLα domain located at the 5’ end of most *var* genes.

We assessed the relationship between various DBLα tag classification methods, and sequence features that are only fully assessable through full-length *var* gene sequences. We compared these different sequence features in full-length *var* gene from six fully sequenced laboratory isolates. These comparisons show that despite a long history of recombination, DBLα sequence tag classification can provide functional information on important features of full-length *var* genes. Notably, a specific subset of DBLα tags previously defined as “group A-like” is associated with CIDRα1 domains proposed to bind to endothelial protein C receptor.

This analysis helps to bring together different sources of data that have been used to assess var gene expression in clinical parasite isolates.

KEMRI|Wellcome Trust This article is included in the KEMRI | Wellcome Trust gateway.

Open Peer Review

Referee Status: ✓ ✓

	Invited Referees 1	2
version 1 published 19 Sep 2017	✓ report	✓ report

1 **Mary M. Rorick**, University of Chicago, USA

2 **Thomas Lavstsen**, University of Copenhagen, Denmark

Discuss this article

Comments (0)

Corresponding author: George Githinji (GGithinji@kemri-wellcome.org)

Author roles: Githinji G: Conceptualization, Data Curation, Formal Analysis, Methodology, Project Administration, Software, Writing – Original Draft Preparation, Writing – Review & Editing; **Bull PC**: Conceptualization, Formal Analysis, Funding Acquisition, Investigation, Methodology, Project Administration, Resources, Supervision, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

How to cite this article: Githinji G and Bull PC. **A re-assessment of gene-tag classification approaches for describing *var* gene expression patterns during human *Plasmodium falciparum* malaria parasite infections [version 1; referees: 2 approved]** Wellcome Open Research 2017, **2**:86 (doi: 10.12688/wellcomeopenres.12053.1)

Grant information: This work was funded by the Wellcome Trust [084538], Strategic Award PhD studentship to GG; [084535], to PCB.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

First published: 19 Sep 2017, **2**:86 (doi: 10.12688/wellcomeopenres.12053.1)

Introduction

PfEMP1 is an important target of naturally acquired immunity to malaria (Chan *et al.,* 2012) and plays a central role in malaria pathology through interaction with host endothelial receptors such as ICAM-1 (Berendt *et al.,* 1989), CD36 (Barnwell *et al.,* 1989), CR1 (Rowe *et al.,* 1997) and endothelial protein-C receptor (EPCR) (Turner *et al.,* 2013). PfEMP1 undergo antigenic variation through epigenetically controlled, mutually exclusive expression of members of a diverse multi-gene family of around 60 *var* genes in every parasite genome (Gardner *et al.,* 2002).

Various cytoadhesive functions are encoded by specific PfEMP1 domain subsets. PfEMP1 molecules contain a combination of two to nine domains (Rask *et al.,* 2010; Smith *et al.,* 2000a) organized in a modular architecture comprising an N-terminal segment, Duffy binding-like (DBL), cysteine inter-domain region (CIDR) and acidic terminal segment domains. DBL domains have been classified into 5 broad groups (α, β, ϒ, δ, ε, and ζ) (Smith *et al.,* 2000b) and CIDR domains classified into four broad sub-groups (α, β, ϒ and δ) (Rask *et al.,* 2010; Smith *et al.,* 2000b) based on sequence similarity. ICAM1 binding is encoded by a subset of DBLβ domains (Brown *et al.,* 2013), CD36 and EPCR by distinct subsets of CIDRα domains (Hsieh *et al.,* 2016; Lau *et al.,* 2015) and rosetting by a subset of DBLα domains (Rowe *et al.,* 1997). Understanding the relationships between specific PfEMP1 variants and clinical malaria is not straightforward, since 1) due to recombination between *var* genes on non-homologous chromosomes, the overall architecture of PfEMP1 encoded by different parasites genotypes is extremely diverse and sequences are mosaics of many semi-conserved sequence blocks, and 2) multiple *var* genes are expressed simultaneously within the infecting parasite population. The range of *var* genes expressed at any one time in the infecting parasite population varies according to the antibodies and other *in vivo* selection pressures. 3) Analysis is further complicated by the high diversity of each domain subclass and lack of clear associations between specific adhesion phenotypes and classes of domains.

Based on full-length sequences from seven laboratory isolates, each domain class has been classified through global sequences alignment into further sub-classes (Rask *et al.,* 2010). For example, the DBLα domain, which has been reclassified into 33 sub-domains (DBLα 0.1 - 0.24, DBLα 1.1 - 1.8 and DBLα2).

Various broad classification methods have been employed to simplify this complex picture in the hope that a limited set of broad functional specializations may exist within *var* that may clarify the disease process. PfEMP1 genes can be classified in relation to their upstream promoter regions (*ups*). The *ups* classification partitions the sequences into groups A–E based on the sequence similarity of the 500 base-pair 5' flanking region and the *var* chromosomal location (Gardner *et al.,* 2002; Vázquez-Macías *et al.,* 2002; Voss *et al.,* 2003; Voss *et al.,* 2000). *Ups* E is associated exclusively with *var2CSA*, which plays a central role in placental malaria (Lavstsen *et al.,* 2003). UpsA *var* genes expression has been reported in several studies to be associated with severe disease (Kyriacou *et al.,* 2006; Lavstsen *et al.,* 2012; Rottmann *et al.,* 2006; Warimwe *et al.,* 2009; Warimwe *et al.,* 2012) and rosetting (Bull *et al.,* 2005a; Rowe *et al.,* 2002; Warimwe *et al.,* 2012). However, an increased transcription of upsB sequences has also been reported to be associated with severe malaria (Rottmann *et al.,* 2006). UpsC sequences have been shown to be expressed at higher levels in asymptomatic cases (Falk *et al.,* 2009; Kaestli *et al.,* 2006); however, expression of upsC sequences in severe malaria cases has also been reported (Kalmbach *et al.,* 2010).

PfEMP1 can be further described in terms of common configurations of different subclasses of domains. These common configurations have been labelled as "domain cassettes" (DCs) (Rask *et al.,* 2010). Twenty-three *var* DCs have been defined from full-length domain alignments of sequences from seven laboratory parasites. It was initially proposed that DCs may act as functional units. However, clearly defined functions have only been assigned at the level of individual domain sub-classes. Therefore, though common combinations of domains exist, it is unclear whether they represent functional units. For example: 1) specific CIDRα1 domains often found in the context of domain cassette 8 (DC8) and 13 (DC13) have been found to bind to EPCR (Turner *et al.,* 2013). *Var* genes containing DC8 cassettes from the IT4 line are suggested to bind to human endothelial cells from various organs and notably from the brain endothelial cells (Avril *et al.,* 2012; Claessens *et al.,* 2012); 2) DBLβ domains found within DC4 genes were reported to adhere to ICAM-1 and may be targets of broadly cross-reactive and adhesion-inhibitory IgG antibodies (Bengtsson *et al.,* 2013).

Clinical and laboratory studies have reported associations between DCs and disease severity. Using PCR primers designed to selectively amplify sequence features found within DC8 and DC13, expression of these DCs were found to be associated with severe malaria in a study conducted in Tanzania (Jespersen *et al.,* 2016; Lavstsen *et al.,* 2012), while a proteomic study in Benin linked the expression of DC8 with cerebral malaria (Bertin *et al.,* 2013).

Several clinical studies have relied on the classification of DBLα tags (Kirchgatter & Portillo, 2002; Kyriacou *et al.,* 2006; Warimwe *et al.,* 2009). We have previously classified these tags using two different approaches. In the first approach, we classified tags using the number of cysteine residues they contained and the existence of two mutually exclusive motifs MFK and REY (Bull *et al.,* 2005b; Bull *et al.,* 2007). Our second approach to classification relied on the fact that recombination between *var* genes appears to be non-random (Kraemer & Smith, 2003; Kraemer *et al.,* 2007). We used network analysis to define sequence groups that tend to share blocks of sequence with each other. We called the most prominent groups block sharing group 1 and block sharing group 2 (BS1 and BS2), respectively. Block sharing group 1 was found enriched in group-A *var* sequences carrying the *upsA* motif (Bull *et al.,* 2008). Based on sensitivity and specificity comparisons with known full length sequence data we defined sequences with 2 cysteines (CP1-3) that fell in

block sharing group 1 as "group A-like" sequences (Warimwe *et al.,* 2009). Clinical studies on *var* expression have shown that group A-like sequences are associated with severe malaria (Warimwe *et al.,* 2009; Warimwe *et al.,* 2012), while two other studies obtained similar results by simply partitioning tags to those with and those without two cysteines (Kirchgatter & Portillo, 2002; Kyriacou *et al.,* 2006). It is currently unclear whether DBLα tags provide information on specific cytoadhesive phenotypes. Furthermore, Lavstsen *et al.,* 2012 have suggested that information on EPCR binding by CIDRα1 within DC8 and DC13 may be unavailable within the DBLα tag due to a recombination hotspot situated between the DBLα tag region and the CIDRα domain.

In an attempt to bring together information from the DBLα tag with information available from the full length *var* gene, we examined associations between full length *var* gene classifications available from a recent study (Rask *et al.,* 2010) and *var* tag classifications used in previous studies of clinical parasite isolates (Bull *et al.,* 2005b; Bull *et al.,* 2007; Kirchgatter & Portillo, 2002; Kyriacou *et al.,* 2006; Warimwe *et al.,* 2009).

Methods

Data collection and sequence classification

DBLα sequence tags were extracted from a total of 403 full-length *var* genes that were sequenced from seven laboratory isolates in a study that explored sequence diversity and classification of PfEMP1 sequences (Rask *et al.,* 2010). The dataset comprised sequences from 3D7, IT4, HB3, DD2 from Indochina, RAJ116 and IGH-CR14 from India, and the Ghanaian isolate PFCLIN. The sequence tags from these genes were classified based on the Cys/PoLV approach (Bull *et al.,* 2007) and the block sharing group approach (Bull *et al.,* 2008), and information on the upstream promoter region and DCs was derived from (Rask *et al.,* 2010).

Var2CSA and sequences without 5' upstream promoter regions classification (ups) information were removed, leaving 313 sequences.

Mapping of *var* genes onto a network of shared polymorphic sequence blocks

A total of 1,548 published DBLα sequences was obtained from Kilifi (Bull *et al.,* 2008, n=1226) and from published parasite genomes (Rask *et al.,* 2010, n=313), together with three DC8 sequences from a study conducted in Tanzania (Lavstsen *et al.,* 2012) and six sequences from "sig2" sequences from (Bull *et al.*, 2005b). Sequences that shared 10 amino acid blocks were identified and used to draw a network of shared common sequences herein referred to as a block-sharing network. The block-sharing networks were generated using a described method (Bull *et al.,* 2008) and were visualized using Pajek 5.01 (Batagelj & Mrvar, 2004). A Perl script (Supplementary File 2) was used to build the sequence networks. For the network of 1,548 tag sequences, *var* tag sequences in fasta format (Dataset: 1548_tags.fa; Githinji, 2017) was used as the input and the output file saved with a .net extension for import into Pajek. The Pajek project used for network analysis is included as Supplementary File 3.

Definition of block sharing groups

The block sharing group (BS) classification of DBLα tags came from a sequence network analysis approach that aimed to visualize how different sequences share blocks of polymorphic sequence. Analysis of fully connected components of a sequence network constructed from observing the sharing of 14 amino acid blocks within DBLα tag sequences from parasites from Kenyan children showed that the largest component, called "block sharing group 1" (BS1) contained predominantly known upsA *var* genes. The second largest component was called block sharing group 2 (BS2) (Bull *et al.,* 2008). We subsequently allocated the newly sequenced DBLα tags to BS1 or BS2 if they contained one or more sequence blocks from the originally defined block sharing groups 1 or 2. We further defined sequences with two cysteines that were classified as BS1 (cys2BS1) as "group A like" (Warimwe *et al.,* 2009) and found that their expression was associated with cerebral malaria (Warimwe *et al.,* 2012).

Functional predictions from DBLα tag information

Receiver operator curves (ROC) were used to visualise the sensitivity and specificity of using specific subsets of DBLα sequence tags in the prediction of upsA, DC8, DC13 and CIDR1α, as outlined in Supplementary File 4.

The block sharing groups were originally defined using a global collection of sequences that included sequences from 3D7 and IT4 laboratory isolates (Bull *et al.,* 2008); therefore, sequences from 3D7 and IT4 isolates were excluded in the block-sharing group analysis presented here. Statistical analysis was done using R version 3.4.0 as outlined in Supplementary File 1.

Results and discussion

Our aim was to summarize the relationships between sequence features within DBLα tag sequences, and sequence features available from fully sequenced *var,* genes from seven fully sequenced genomes (Rask *et al.,* 2010). The relationships between these two levels of information were visualized using bar graphs (Figure 1, Figure 2 and Figure 3; Figure S1 and Figure S2) a network visualization approach (Figure 4 and Figure 5) and through a sensitivity, specificity analysis (Figure 6).

Figure 1 focuses on 313 DBL domains classified by (Rask *et al.,* 2010) into 33 DBLα sub-groups. The DBLα tag region within were classified by both the block-sharing (Bull *et al.,* 2008) and the cys/polv (Bull *et al.,* 2005b) classifications. The ups region of each corresponding gene is also shown. BS1 sequences were closely associated with upsA, and BS2 sequences were associated largely with upsB or upsC. While most cys2 sequences (CP1-3) were found within sequences containing the upsA promoter, some of them were also found in sequences containing upsB and upsC promoters. For example, sequences with DBLα-0.3

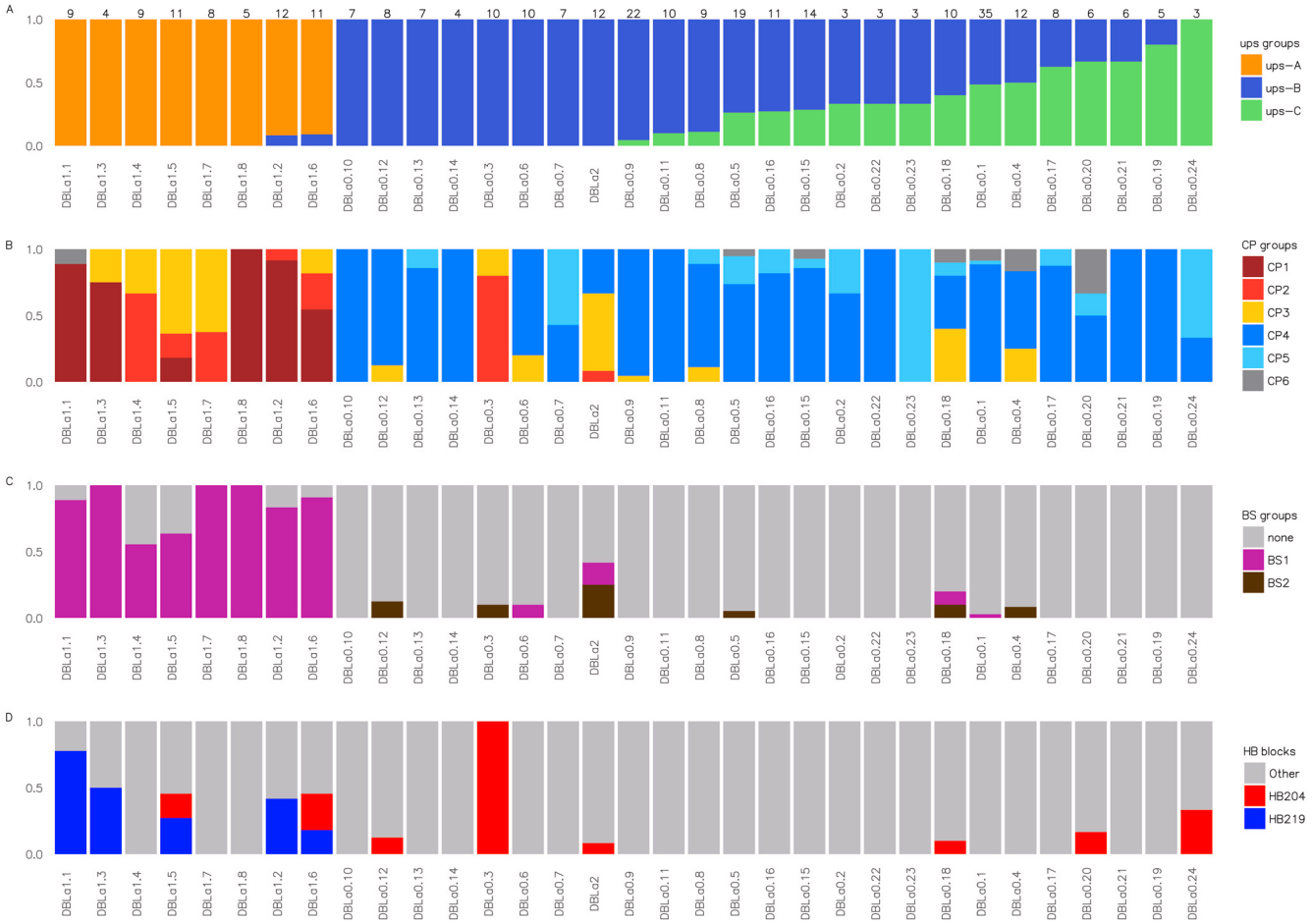

Figure 1. Correspondence between various *var* sequence classifications and possession of specific DBLα domains classified by (Rask *et al.*, 2010), for *var* genes sequenced from 6 laboratory isolates. Each *var* gene contains only one DBLα domain. For each subset of *var* genes, classified according to their DBLα domains (x axis), the proportion of genes carrying other sequence features is shown (y axis). (**A**) ups classification; (**B**) cys/polv classification (Bull *et al.*, 2005b); (**C**) block sharing group classification (Bull *et al.*, 2008); (**D**) selected homology block classifications (Rorick *et al.*, 2013). The domains are arranged from left to right in order of decreasing proportion of upsA to upsC-containing *var* gene sequences. The total number of sequences from each domain is shown at the top of the figure.

or DBLα-2 subdomains were largely upsB. However, they contained relatively high proportions of *var* sequences with two cysteines, specifically those from CP2 and CP3 Cys/PoLV groups.

DBLα *sub*-domains are not all homogeneous groups

Domain classification that was suggested by (Rask *et al.*, 2010) were partly based on global sequence alignments. Applying sequence alignment to a large collection of recombining *var* sequences is challenging because the alignment process does not consider the recombination history and potentially defines sequences as distinct when they are part of a network of recombining sequences.

Examination of DBLα tags suggests that MFK and REY motifs (highly enriched within subsequently defined homology blocks 219 and 204 (Rask *et al.*, 2010; Rorick *et al.*, 2013)) are never found on the same sequence (Bull *et al.*, 2005b). However, DBLα1.5, DBLα1.2 and DBLα1.6 groups defined by Rask and colleagues each comprise a mixture of MFK-containing and REY-containing sequences (Figure 1). The domain classification used in (Rask *et al.*, 2010) has therefore brought together distinct sequences within the same sequence classification. This suggests that the newly defined sub-domains do not always classify sequences into wholly genetically distinct groups. This discordance between methods of classification, employing global and local sequence comparisons reflects a mode of diversification of *var* sequences by *P. falciparum* that we might speculate leads to impaired recognition and clearance of PfEMP1 antigens by the immune system.

Existing DBLα tag classification cannot predict DC8 sequences from a global sequence collection

Similar to group A-like sequences, DC8 sequences are associated with severe malaria (Bengtsson *et al.*, 2013; Bertin *et al.*, 2013; Lavstsen *et al.*, 2012; Rask *et al.*, 2010) and contain a

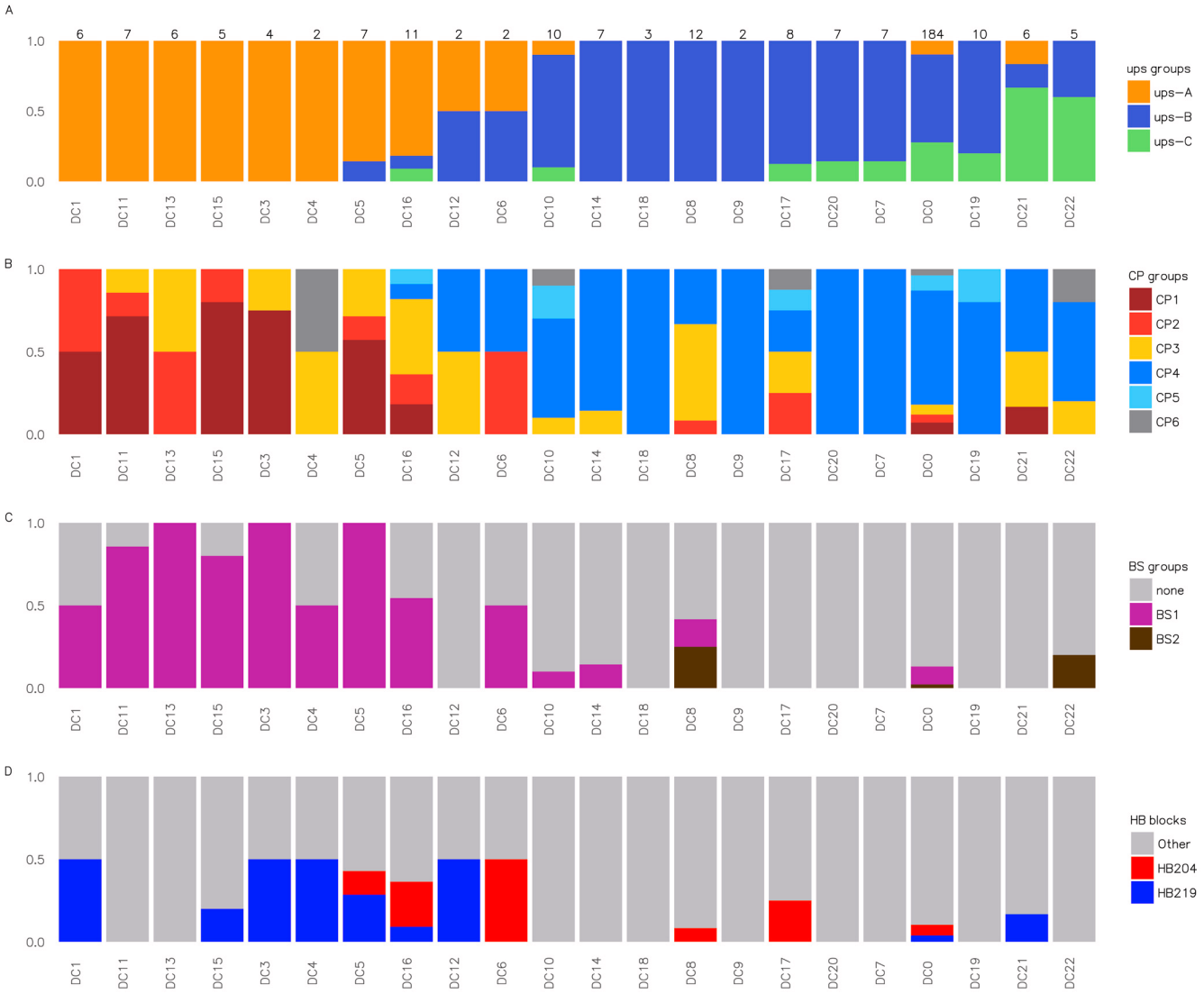

Figure 2. Correspondence between various *var* sequence classifications and possession of specific domain cassettes (DCs) for *var* genes sequenced from 6 laboratory isolates (Rask *et al.*, 2010). For each subset of *var* genes, classified according to their DC (x axis), the proportion of genes carrying other sequence features is shown (y axis). (**A**) ups classification; (**B**) cys/polv classification (Bull *et al.*, 2005b); (**C**) block sharing group classification (Bull *et al.*, 2008); (**D**) selected homology block classifications (Rorick *et al.*, 2013). The cassettes sorted from left to right such that the leftmost sequences contain the largest proportion of upsA *var* genes, while sequences to the right contain the largest proportion of upsC *var* genes. The number of sequences from each DC is shown at the top of the figure. Sequences that were not assigned to a domain are denoted as DC0.

specific class of DBLα2 sequences that appear to result from recombination events at a recombination hotspot proposed to be situated 3' of the DBLα tag region (Lavstsen *et al.*, 2012). Low levels of linkage disequilibrium between the DBLα tag region and parts of the genes encoding important cytoadhesive regions potentially limits the predictive information available within DBLα tag sequence. This is consistent with the observation that DC8 sequences contain multiple cys/PoLV groups CP2, CP3 and CP4 (Figure 2). However, none of the identified DC8 sequences contain CP1 tags, perhaps suggesting some level of linkage disequilibrium with the tag region. In support of this possibility, DC8 sequences contained the highest proportion of observed BS2 sequences of any DC. Furthermore, an additional set of DC8-like sequences identified in Tanzania (Lavstsen *et al.*, 2012) were similar to previously defined "sig2" sequences found in two severe malaria cases sampled from Kenyan children (Bull *et al.*, 2005b). Both sets of sequences are defined as BS2, CP2. We have previously suggested that BS2 sequences may be characteristic of *var* genes sampled from Africa (Bull *et al.*, 2008). It is possible that DC8 sequences sampled from limited

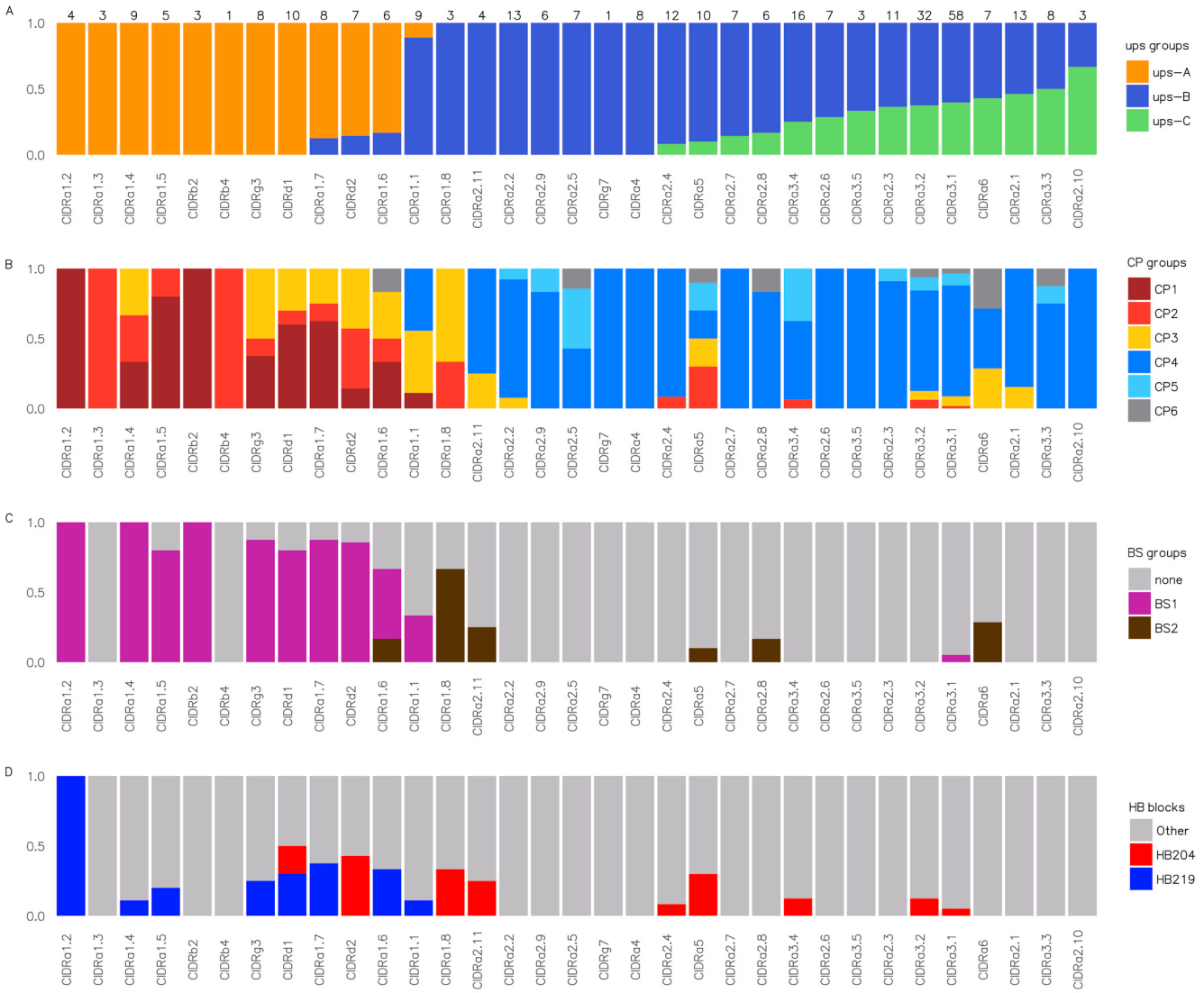

Figure 3. Correspondence between various *var* sequence classifications and possession of specific CIDR1 domains for *var* genes sequenced from 6 laboratory isolates (Rask *et al.*, 2010). For each subset of *var* genes, classified according to their CIDR1 domains (x-axis), the proportion of genes carrying other sequence features is shown (y-axis). (**A**) ups classification; (**B**) cys/polv classification (Bull *et al.*, 2005b); (**C**) block sharing group classification (Bull *et al.*, 2008); (**d**) selected homology block classifications (Rorick *et al.*, 2013). The CIDR domains are sorted from left to right, such that the left-most sequences contain the largest proportion of upsA, while sequences to the right contain the largest proportion of upsC *var* genes. The total number of *var* genes containing each of the CIDR1 domains is shown at the top of the figure.

geographical regions may show significant levels of linkage disequilibrium with DBLα tag sequence features (see Figure 5 below).

Mapping tag regions from full length *var* genes onto a network of DBLα tag sequences from Kenyan children

Patterns of diversification in sequences may give an indication of how these sequences evolve in the face of *in vivo* selection pressure. In Figure 4 and Figure 5, we used our previously described approach of visualizing the sharing of polymorphic blocks within DBLα to explore specific subsets of full length *var* genes.

To understand how various sequences with known DCs mapped to this network, we re-drew the network from (Bull *et al.*, 2008) whilst including the sequences from the 7 genomes. We also supplemented the figure with additional sequences including, the "sig 2" sequences identified in a previous analysis of isolates causing severe and non-severe malaria and DC8 sequences identified in Tanzania (Lavstsen *et al.*, 2012). As shown in Figure 4F, DC8 sequences were restricted mainly to the region of the network containing mainly upsB and upsC sequences, while DC13 were associated with the region of the network enriched in upsA sequences. Figure 5 further illustrates the relationships between DBLα tags from known DC8 genes.

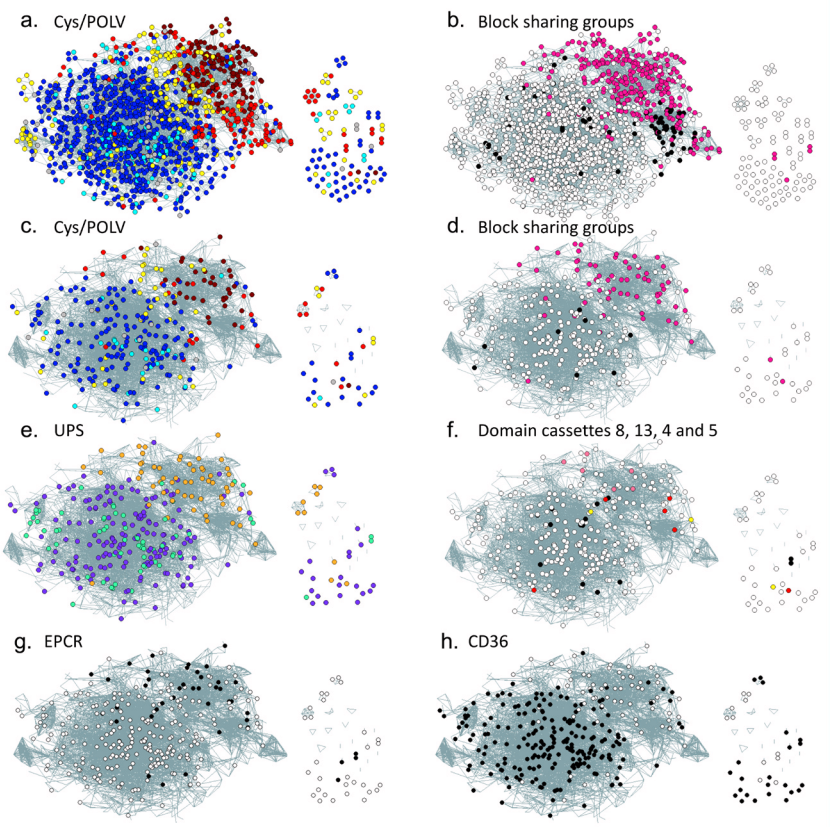

Figure 4. Network analysis of DBLα tag sequences collected from Kilifi (Bull *et al.*, 2008), 6 laboratory isolates (Rask *et al.*, 2010) and Tanzanian (Lavstsen *et al.*, 2012). The analysis builds on that described in (Bull *et al.*, 2008). (**a**) Cys/polv analysis for all sequences; (**b**) block sharing groups analysis for all sequences; (**c**) Cys/polv analysis for full length *var* gene sequences from 6 laboratory isolates; (**d**) block sharing groups analysis for full length *var* gene sequences from 6 laboratory isolates; (**e**) ups grouping for full length *var* gene sequences from 6 laboratory isolates; (**f**) domain cassette (DC) classification for DC4, DC5, DC8 and DC13 for full length *var* gene sequences from 6 laboratory isolates; (**g**) predicted EPCR-binding phenotype due to CIDRα1.1, CIDRα1.4, CIDRα1.5, CIDRα1.6, CIDRα1.7 or CIDRα1.8 (Lau *et al.*, 2015) for sequences with CIDRα information available; (**h**) predicted CD36-binding phenotype due to CIDRα2, CIDRα3, CIDRα4, CIDRα5 (Robinson *et al.*, 2003) for sequences with CIDRα information available. Colours of vertices match those defined in Figure 1: **a** and **c**) brown = cys/polv group 1 (CP1), red= CP2, yellow = CP3, blue = CP4, light-blue = CP5, grey = CP6; **b** and **d**) pink = block sharing group 1 (BS1), black = BS2, white = not a member of a block sharing group; **e**) orange = upsA, purple = upsB, light green = upsC; **f**) black = domain cassette 8 (DC8), red = DC5, pink = DC13, yellow = DC4; **g**) black = predicted EPCR binding; **h**) black

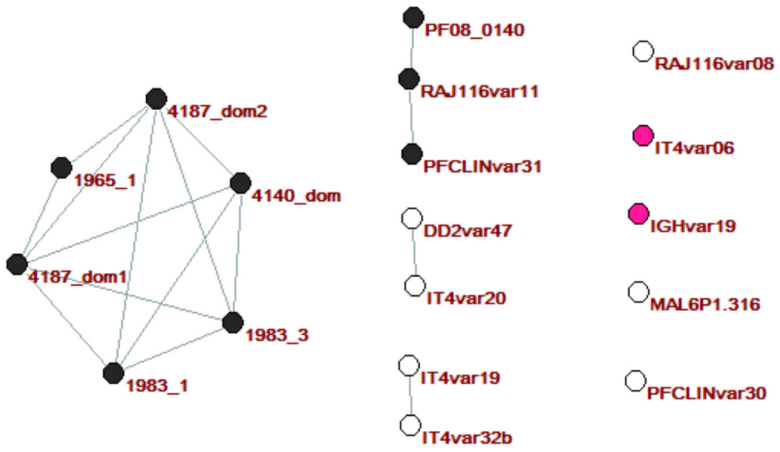

Figure 5. Network analysis of DBLα tag sequences from known DC8 *var* genes. sequences are from 6 genomes, DC8 Sequences 1983_3, 1983_1 and 1965_1 from a study in Tanzania (Lavstsen *et al.*, 2012) and "sig-2" sequences from Kenya, 4140_dom 4187_dom1 and 4187_dom2 (Bull *et al.*, 2005b). Colours of vertices match those defined in Figure 1: pink = block sharing group 1 (BS1); black = BS2; white = not a member of a BS.

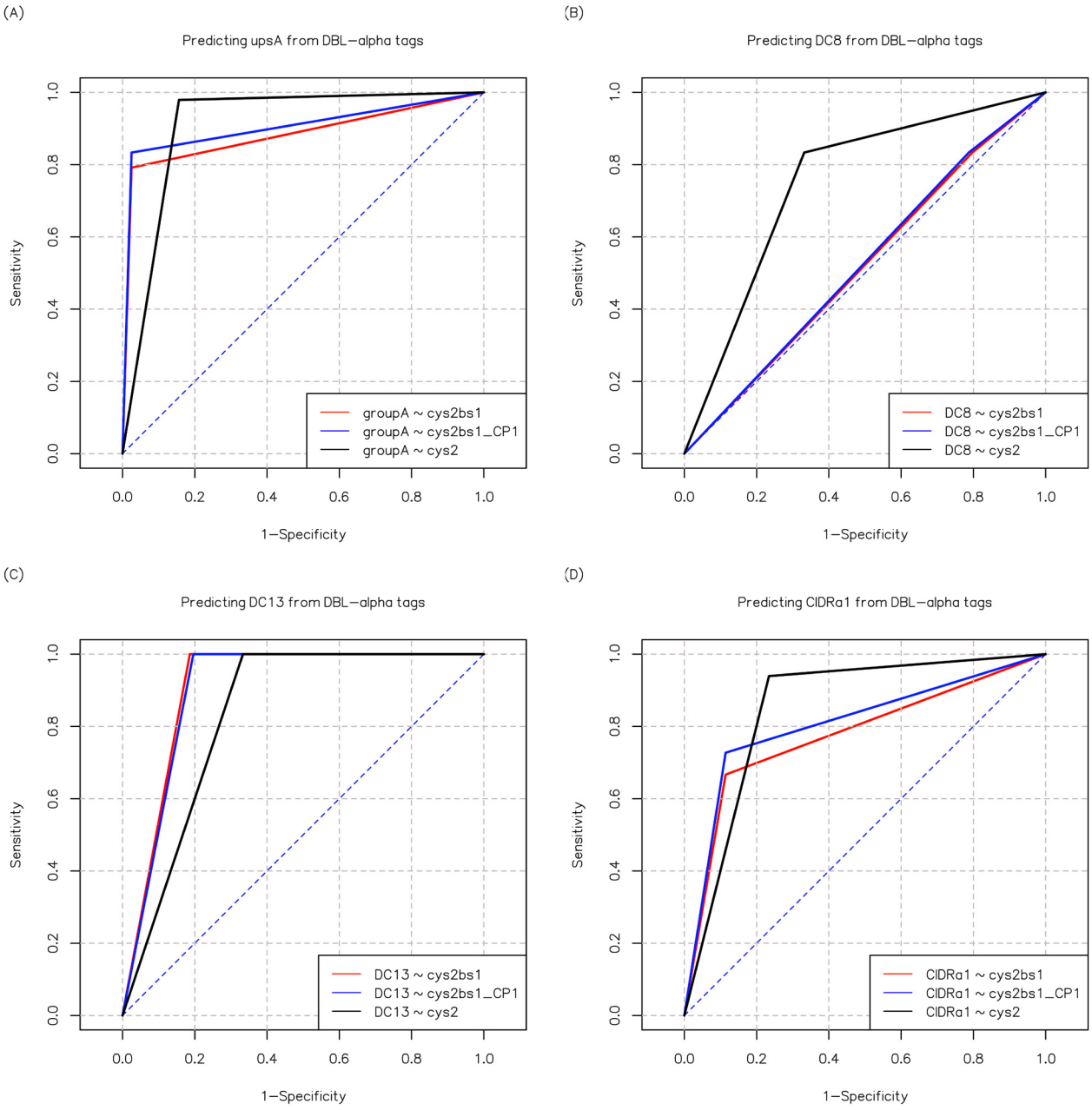

Figure 6. Receiver operator curves showing the sensitivity and specificity of three DBLα tag classifications in predicting *var* gene features associated with disease severity. (**A**) Sensitivity and specificity in predicting upsA sequences. (**B**, **C**) The prediction of DC8 and DC13 sequences. (**D**) The prediction of CIDR1α domains from tag information. Sequences from 3D7 and IT4 were excluded from the analysis because they were used for developing these these classifications (Bull *et al.*, 2008). cys2 = two cysteines within the tag region; cys2bs1 = tag sequences in block sharing group1 AND have two cysteines, defined as "group A-like" (Warimwe *et al.*, 2009); cys2bs1_CP1 = cys2bs1 OR in cys/PoLV group 1.

Sequences with DC4 cassettes are reported to be associated with binding to ICAM1 (Bengtsson *et al.*, 2013). In this data set, there were only 2 sequences with DC4 cassettes; one sequence has a CP3 DBLα tag region and the other a CP6 DBLα tag region (Figure 4F). These sequences map to distinct locations within the network. Sequences with DC5 cassettes were from different Cys/PoLV groups all of which belonged to BS1, three of which mapped to a similar region of the network (Figure 4F).

To map predicted cytoadhesive properties of the PfEMP1 antigens encoded by these genes, we made predictions based on existing information and mapped these cytoadhesive properties

onto the network (Figure 4). Endothelial protein C receptor binding and CD36 binding were predicted based on the binding properties of recombinant CIDR domains from (Lau *et al.,* 2015) and (Robinson *et al.,* 2003) respectively (Figures 4G and H). Though the number of sequences is very limited, this mapping of predicted cytoadhesive properties is consistent with the idea that functional specialization of *var* genes is associated with broad sequence differences that are detectable within DBLα tag sequences.

A recent study (Rorick *et al.,* 2013) has further explored this possibility by classifying DBLα tags using homology blocks defined in (Rask *et al.,* 2010). They found in datasets from Kenya and Mali that homology block 204 (closely related to CP2) was associated with impaired consciousness and homology block 219 (closely related to CP1) was associated with rosetting. Figure 1 and Figure 2 also summarizes how these two homology blocks relate to other DBLα tag classifications.

Sensitivity and specificity analysis

In summary, this analysis shows that some information about functionally relevant *var* gene sequence features from existing DBLα tag sequence classification methods. Most notably, the presence of a CIDRα1 domain, predicted to bind to endothelial protein C receptor (Lau *et al.,* 2015) and associated with severe malaria (Jespersen *et al.*, 2016) is associated with "group A-like" sequences (bs1cys2), which potentially explains previously reported associations between both the expression of related subsets of cys2 sequence tags and DC8 and DC13 *var* genes, with severe malaria (Kirchgatter & Portillo, 2002; Kyriacou *et al.,* 2006; Warimwe *et al.,* 2012). Figure 6 summarizes sensitivity and specificity analyses for the associations described. Supplementary File 5 (Tables 1–12) shows the corresponding statistical significance. Figure 6 also illustrates the slightly increased sensitivity of prediction of presence of a CIDRα1 domain through expanding the definition of group A-like to include all CP1 sequences (cys2bs1_CP1). Associations between DBLα tag classifications and full length *var,* sequences are useful for bringing together and explaining findings from previous studies. However, such analyses will soon be replaced by methods such as RNAseq (Otto *et al.,* 2010) or mass spectrometry (Bertin *et al.,* 2013) that allow access to information from full length *var* genes and PfEMP1 sequences from clinical isolates.

Data availability

The data and analysis scripts used in this analysis are available from OSF: http://doi.org/10.17605/OSF.IO/UWCN2 (Githinji, 2017).

Competing interests

No competing interests were disclosed.

Grant information

This work was funded by the Wellcome Trust [084538], Strategic Award PhD studentship to GG; [084535], to PCB.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Supplementary material

Supplementary File 1: R script - *var* expression analysis.

Click here to access the data.

Supplementary File 2: Perl script - network constructor.

Click here to access the data.

Supplementary File 3: Pajek project – network of 1548 *var* tag sequences.

Click here to access the data.

Supplementary File 4: R script – sensitivity specificity analysis.

Click here to access the data.

Supplementary File 5: Tables 1–12 show Fisher's exact tests.

Click here to access the data.

Figure S1: Bar plots showing the distribution and proportion of block sharing groups (BS) across 23 domain cassettes. All the full-length sequences that contained DC13 cassette architectures contained DBLα sequence tags that belonged to BS1. Full-length sequences with DC8 cassette architecture contained DBL sequence tags that belonged to either BS1 or BS2, and several the DBL tags did not belong either BS1 or BS2.

Click here to access the data.

Figure S2: The relationship between HB204, HB219 and cys/PoLV groups.

Click here to access the data.

References

Avril M, Tripathi AK, Brazier AJ, *et al.*: **A restricted subset of *var* genes mediates adherence of *Plasmodium falciparum*-infected erythrocytes to brain endothelial cells.** *Proc Natl Acad Sci U S A.* 2012; **109**(26): E1782–90.
PubMed Abstract | Publisher Full Text | Free Full Text

Barnwell JW, Asch AS, Nachman RL, *et al.*: **A human 88-kD membrane glycoprotein (CD36) functions *in vitro* as a receptor for a cytoadherence ligand on *Plasmodium falciparum*-infected erythrocytes.** *J Clin Invest.* 1989; **84**(3): 765–772.
PubMed Abstract | Publisher Full Text | Free Full Text

Batagelj V, Mrvar A: **Pajek—analysis and visualization of large networks**. *Graph drawing software.* 2004; 77–103.
Publisher Full Text

Bengtsson A, Joergensen L, Rask TS, *et al.*: **A novel domain cassette identifies *Plasmodium falciparum* PfEMP1 proteins binding ICAM-1 and is a target of cross-reactive, adhesion-inhibitory antibodies.** *J Immunol.* 2013; **190**(1): 240–249.
PubMed Abstract | Publisher Full Text | Free Full Text

Berendt AR, Simmons DL, Tansey J, *et al.*: **Intercellular adhesion molecule-1 is an endothelial cell adhesion receptor for *Plasmodium falciparum*.** *Nature.* 1989; **341**(6237): 57–59.
PubMed Abstract | Publisher Full Text

Bertin GI, Lavstsen T, Guillonneau F, *et al.*: **Expression of the domain cassette 8 *Plasmodium falciparum* erythrocyte membrane protein 1 is associated with cerebral malaria in Benin.** *PLoS One.* 2013; **8**(7): e68368.
PubMed Abstract | Publisher Full Text | Free Full Text

Brown A, Turner L, Christoffersen S, *et al.*: **Molecular architecture of a complex between an adhesion protein from the malaria parasite and intracellular adhesion molecule 1.** *J Biol Chem.* 2013; **288**(8): 5992–6003.
PubMed Abstract | Publisher Full Text | Free Full Text

Bull PC, Buckee CO, Kyes S, *et al.*: ***Plasmodium falciparum* antigenic variation. Mapping mosaic *var* gene sequences onto a network of shared, highly polymorphic sequence blocks.** *Mol Microbiol.* 2008; **68**(6): 1519–1534.
PubMed Abstract | Publisher Full Text | Free Full Text

Bull PC, Berriman M, Kyes S, *et al.*: ***Plasmodium falciparum* variant surface antigen expression patterns during malaria.** *PLoS Pathog.* 2005b; **1**(3): e26.
PubMed Abstract | Publisher Full Text | Free Full Text

Bull PC, Kyes S, Buckee CO, *et al.*: **An approach to classifying sequence tags sampled from *Plasmodium falciparum var* genes.** *Mol Biochem Parasitol.* 2007; **154**(1): 98–102.
PubMed Abstract | Publisher Full Text | Free Full Text

Bull PC, Pain A, Ndungu FM, *et al.*: ***Plasmodium falciparum* antigenic variation: relationships between *in vivo* selection, acquired antibody response, and disease severity.** *J Infect Dis.* 2005a; **192**(6): 1119–1126.
PubMed Abstract | Publisher Full Text

Chan JA, Howell KB, Reiling L, *et al.*: **Targets of antibodies against *Plasmodium falciparum*-infected erythrocytes in malaria immunity.** *J Clin Invest.* 2012; **122**(9): 3227–3238.
PubMed Abstract | Publisher Full Text | Free Full Text

Claessens A, Adams Y, Ghumra A, *et al.*: **A subset of group A-like *var* genes encodes the malaria parasite ligands for binding to human brain endothelial cells.** *Proc Natl Acad Sci U S A.* 2012; **109**(26): E1772–81.
PubMed Abstract | Publisher Full Text | Free Full Text

Falk N, Kaestli M, Qi W, *et al.*: **Analysis of *Plasmodium falciparum var* genes expressed in children from Papua New Guinea.** *J Infect Dis.* 2009; **200**(3): 347–356.
PubMed Abstract | Publisher Full Text

Gardner MJ, Hall N, Fung E, *et al.*: **Genome sequence of the human malaria parasite *Plasmodium falciparum*.** *Nature.* 2002; **419**(6906): 498–511.
PubMed Abstract | Publisher Full Text | Free Full Text

Githinji G: **A Reassessment of Gene-Tag Classification Approaches for Describing *Var* Gene Expression Patterns during Human *Plasmodium Falciparum* Malaria Parasite Infections.** *Open Science Framework.* 2017.
Data Source

Hsieh FL, Turner L, Bolla JR, *et al.*: **The structural basis for CD36 binding by the malaria parasite.** *Nat Commun.* 2016; **7**: 12837.
PubMed Abstract | Publisher Full Text | Free Full Text

Jespersen JS, Wang CW, Mkumbaye SI, *et al.*: ***Plasmodium falciparum var* genes expressed in children with severe malaria encode CIDRα1 domains.** *EMBO Mol Med.* 2016; **8**(8): 839–850.
PubMed Abstract | Publisher Full Text | Free Full Text

Kaestli M, Cockburn IA, Cortés A, *et al.*: **Virulence of malaria is associated with differential expression of *Plasmodium falciparum var* gene subgroups in a case-control study.** *J Infect Dis.* 2006; **193**(11): 1567–1574.
PubMed Abstract | Publisher Full Text | Free Full Text

Kalmbach Y, Rottmann M, Kombila M, *et al.*: **Differential *var* gene expression in children with malaria and antidromic effects on host gene expression.** *J Infect Dis.* 2010; **202**(2): 313–317.
PubMed Abstract | Publisher Full Text

Kirchgatter K, Portillo Hdel A: **Association of severe noncerebral *Plasmodium falciparum* malaria in Brazil with expressed PfEMP1 DBL1 alpha sequences lacking cysteine residues.** *Mol Med.* 2002; **8**(1): 16–23.
PubMed Abstract | Free Full Text

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**(5): 1527–1538.
PubMed Abstract | Publisher Full Text

Kraemer SM, Kyes SA, Aggarwal G, *et al.*: **Patterns of gene recombination shape *var* gene repertoires in *Plasmodium falciparum*: comparisons of geographically diverse isolates.** *BMC Genomics.* 2007; **8**: 45.
PubMed Abstract | Publisher Full Text | Free Full Text

Kyriacou HM, Stone GN, Challis RJ, *et al.*: **Differential *var* gene transcription in *Plasmodium falciparum* isolates from patients with cerebral malaria compared to hyperparasitaemia.** *Mol Biochem Parasitol.* 2006; **150**(2): 211–218.
PubMed Abstract | Publisher Full Text | Free Full Text

Lau CK, Turner L, Jespersen JS, *et al.*: **Structural conservation despite huge sequence diversity allows EPCR binding by the PfEMP1 family implicated in severe childhood malaria.** *Cell Host Microbe.* 2015; **17**(1): 118–129.
PubMed Abstract | Publisher Full Text | Free Full Text

Lavstsen T, Salanti A, Jensen AT, *et al.*: **Sub-grouping of *Plasmodium falciparum* 3D7 *var* genes based on sequence analysis of coding and non-coding regions.** *Malar J.* 2003; **2**: 27.
PubMed Abstract | Publisher Full Text | Free Full Text

Lavstsen T, Turner L, Saguti F, *et al.*: ***Plasmodium falciparum* erythrocyte membrane protein 1 domain cassettes 8 and 13 are associated with severe malaria in children.** *Proc Natl Acad Sci U S A.* 2012; **109**(26): E1791–800.
PubMed Abstract | Publisher Full Text | Free Full Text

Otto TD, Wilinski D, Assefa S, *et al.*: **New insights into the blood-stage transcriptome of *Plasmodium falciparum* using RNA-Seq.** *Mol Microbiol.* 2010; **76**(1): 12–24.
PubMed Abstract | Publisher Full Text | Free Full Text

Rask TS, Hansen DA, Theander TG, *et al.*: ***Plasmodium falciparum* erythrocyte membrane protein 1 diversity in seven genomes--divide and conquer.** *PLoS Comput Biol.* 2010; **6**(9): pii: e1000933.
PubMed Abstract | Publisher Full Text | Free Full Text

Robinson BA, Welch TL, Smith JD: **Widespread functional specialization of *Plasmodium falciparum* erythrocyte membrane protein 1 family members to bind CD36 analysed across a parasite genome.** *Mol Microbiol.* 2003; **47**(5): 1265–1278.
PubMed Abstract | Publisher Full Text

Rorick MM, Rask TS, Baskervill EB, *et al.*: **Homology blocks of *Plasmodium falciparum var* genes and clinically distinct forms of severe malaria in a local population.** *BMC Microbiol.* 2013; **13**(1): 244.
PubMed Abstract | Publisher Full Text | Free Full Text

Rottmann M, Lavstsen T, Mugasa JP, *et al.*: **Differential expression of *var* gene groups is associated with morbidity caused by *Plasmodium falciparum* infection in Tanzanian children.** *Infect Immun.* 2006; **74**(7): 3904–3911.
PubMed Abstract | Publisher Full Text | Free Full Text

Rowe JA, Moulds JM, Newbold CI, *et al.*: ***P. falciparum* rosetting mediated by a parasite-variant erythrocyte membrane protein and complement-receptor 1.** *Nature.* 1997; **388**(6639): 292–295.
PubMed Abstract | Publisher Full Text

Rowe JA, Shafi J, Kai OK, *et al.*: **Nonimmune IgM, but not IgG binds to the surface of *Plasmodium falciparum*-infected erythrocytes and correlates with rosetting and severe malaria.** *Am J Trop Med Hyg.* 2002; **66**(6): 692–699.
PubMed Abstract | Publisher Full Text

Smith JD, Craig AG, Kriek N, *et al.*: **Identification of a *Plasmodium falciparum* intercellular adhesion molecule-1 binding domain: a parasite adhesion trait implicated in cerebral malaria.** *Proc Natl Acad Sci U S A.* 2000a; **97**(4): 1766–1771.
PubMed Abstract | Publisher Full Text | Free Full Text

Smith JD, Subramanian G, Gamain B, *et al.*: **Classification of adhesive domains in the *Plasmodium falciparum* erythrocyte membrane protein 1 family.** *Mol Biochem Parasitol.* 2000b; **110**(2): 293–310.
PubMed Abstract | Publisher Full Text

Turner L, Lavstsen T, Berger SS, *et al.*: **Severe malaria is associated with parasite binding to endothelial protein C receptor.** *Nature.* 2013; **498**(7455): 502–505.
PubMed Abstract | Publisher Full Text | Free Full Text

Vázquez-Macías A, Martínez-Cruz P, Castañeda-Patlán MC, *et al.*: **A distinct 5' flanking *var* gene region regulates *Plasmodium falciparum* variant erythrocyte**

surface antigen expression in placental malaria. *Mol Microbiol.* 2002; **45**(1): 155–167.
PubMed Abstract | Publisher Full Text

Voss TS, Kaestli M, Vogel D, *et al.*: **Identification of nuclear proteins that interact differentially with *Plasmodium falciparum var* gene promoters.** *Mol Microbiol.* 2003; **48**(6): 1593–1607.
PubMed Abstract | Publisher Full Text

Voss TS, Thompson JK, Waterkeyn J, *et al.*: **Genomic distribution and functional characterisation of two distinct and conserved *Plasmodium falciparum var* gene 5' flanking sequences.** *Mol Biochem Parasitol.* 2000; **107**(1): 103–115.
PubMed Abstract | Publisher Full Text

Warimwe GM, Fegan G, Musyoki JN, *et al.*: **Prognostic indicators of life-threatening malaria are associated with distinct parasite variant antigen profiles.** *Sci Transl Med.* 2012; **4**(129): 129ra45.
PubMed Abstract | Publisher Full Text | Free Full Text

Warimwe GM, Keane TM, Fegan G, *et al.*: ***Plasmodium falciparum var* gene expression is modified by host immunity.** *Proc Natl Acad Sci U S A.* 2009; **106**(51): 21801–21806.
PubMed Abstract | Publisher Full Text | Free Full Text

Open Peer Review

Current Referee Status: ✔ ✔

Version 1

Referee Report 09 October 2017

doi:10.21956/wellcomeopenres.13039.r26166

✔ **Thomas Lavstsen**
Centre for Medical Parasitology, Department of Immunology and Microbiology (ISIM), Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark

Githinji and Bull present an analysis of the associations between previously developed annotation tools whole PfEMP1 sequences and for the “DBLa-tag”, a short PCR-amplifiable sequence found in all PfEMP1 encoding genes allowing unique identification of specific var genes. Reconciling sequence traits of known PfEMP1 receptor binding phenotype, defined PfEMP1 domain types and DBL-tag annotation methods is important as despite advances in high throughput sequencing, analysis of DBLa-tag along with qPCR analysis still represent most efficient and precise detection of the polymorphic var genes expressed by parasites in patients.

The paper presents the analyses in a set of intuitively easily understood graphs, - however PfMEP1 domain composition and nomenclature can easily become confusing. Most of my comments and suggestions relate to improvement and corrections of explanations towards a simpler and hopefully clearer presentation of the current knowledge and relevance of this study. One additional analysis regarding adding predicted ICAM1 binding PfEMP1 should be added to one of the figures.

Specific comments:

Although shown many times, it will be useful to have a very simple diagram showing PfEMP1 domain structure indicating the position of the DBLa-tag, the known hotspots of recombination at the DBLa-tag end, and mid var region. This will highlight the purpose and challenge of this whole exercise.
Quote: “3) Analysis is further complicated by the high diversity of each domain subclass and lack of clear associations between specific adhesion phenotypes and classes of domains.”
I would phrase this differently. Although many binding phenotypes have been proposed for iRBCs, well characterized interactions for PfEMP1 are more limited. In fact only few interactions are studied to the extent that these can be used to predict PfEMP1 function: CSA, CD36, EPCR and ICAM1. For these there is only a small uncertainly for determining ICAM1 binding domains.
Binding to HABP1, PECAM1, IgM, etc is, as stated not clearly linked to specific domains or sequence traits. As parasite adhesion phenotypes cannot be investigated in vivo, "specific adhesion phenotypes" is defined from a combination of clear association of PfMEP1 domain type with binding to a specific receptor, as well as validation of this by iRBC binding assays; and thus observed iRBC binding to various receptors cannot be taken as a gold standard on its own.

I suggest Correcting next paragraph to: Based on full-length sequences from seven laboratory isolates, each domain class was divided through global sequence alignment into further sub-classes (Rask *et al.,*

2010). For example, the DBLa domains were reclassified into 33 sub-domains (DBLa0.1 - 0.24, DBLa1.1 - 1.8 and DBLa2).

I suggest changing references as: "*Ups* E is associated exclusively with *var2CSA*(Lavstsen *et al.,* 2003), which plays a central role in placental malaria (Reference Salanti et al J ex Med).

Quote: "It was initially proposed that DCs may act as functional units. However, clearly defined functions have only been assigned at the level of individual domain sub-classes. Therefore, though common combinations of domains exist, it is unclear whether they represent functional units."
I agree with this. I would even suggest that it is clear that most domain cassettes, although useful to define molecular tools, does not appear to reflect functional units. However, I do not think the examples given does not elude clearly to this. In line with my comment above, I think it should be made clear that the binding phenotypes we understand well today, (EPCR; CD36 and ICAM1) all are associated with and fully contained within single domains. However, some subsets of these domains appear to have co-evolved - like ICAM1 binding (DBLb) in group A is always found in CIDRa1 (EPCR binding) variants, (co-evolution clearly seen within the DC13 context); whereas ICAM1-binding DBLb are rarely found in CIDRa1 (EPCR binding) containing DC8; and DBLb are not specifically associated with any Domain subclass when found in group B. The comment of DBLb domains being targets of cross reactive antibodies seems irrelevant in this context and confusing here.

Paragraph. "Clinical and laboratory studies have reported…..".
The DC8 and DC13 are useful for understanding how these variable genes can be probed and detected iin vivo. But given the limited usefulness of the DCs to describe known binding phenotypes, and their relation clinical outcome;
And the recent studies of var expression (Those referenced, and the Mkumbaye et al IAI 2017, not included but should be) - which are in line with previous work as described in two paragraphs before;
And to simplify for those new to the field –
I think it would be best to use this paragraph to describe the consensus from these studies, that CIDRa1 is the only common trait of var genes whose expression is associated with malaria pathology; regardless of symptomology. And EPCR+IMCA1 but not CD36+icam1 is found more frequently in CM (lennartz et al 2017). These findings does not mean that future studies relying on DBLa tag is not useful or needed.

M&M section: "Definition of block sharing groups"
This is a more detailed (and required) description of the statement above on SB1, correct? Perhaps just refer back to this. It is unclear which part of the text refers to previous work or and which to the re-analysis performed here. Leave out the sentence: "and found that their expression was associated with cerebral malaria (Warimwe *et al.,* 2012).". This is not relevant here.

I think following concluding statement is not needed ,as the whole premise for the study is to compare classification methods, - it is given or should already be stated that the sequences has evolved to under pressure to diversify in response to immune recognition, and to maintain structural fold to retain function:
"This discordance between methods of classification, employing global and local sequence comparisons reflects a mode of diversification of *var* sequences by *P. falciparum* that we might speculate leads to impaired recognition and clearance of PfEMP1 antigens by the immune system.

"Similar to group A-like sequences, DC8 sequences are associated with severe malaria (Bengtsson *et al.,* 2013; Bertin *et al.,* 2013; Lavstsen *et al.,* 2012; Rask *et al.,* 2010)……"
The main point to iterate here should be that DC8s have CIDRa1 and a recombined B/A DBLa domain named DBLa2. The link to SM is part of this fact; . ie the presence of the CIDRa1 domain.

I am not sure which part the Bengtsson et al 2013 reference plays here.
Also, specify that DC8 is (the only known) a B/A recombination - thus B from UPS to DBLa tag end, and A like in its DBLaS3 and downstream from there. This is important to understand, as the A vs B/C grouping is tied to the chromosomal localization of these genes which ensures that A does not normally recombine with B (otherwise lethal chromosomes will be formed). This is probably why the DC8 is the best conserved domain cassette, and why DBLa tag analysis is particularly difficult (and important) to apply for these genes.

“Sequences with DC4 cassettes are reported to be associated with…..”.
The ICAM1 prediction has been refined considerable in Lennartz et al 2017. The authors should color in sequences which contain group B DBLb5s show/predicted to bind ICAM1 and the DBLb3/1 domaains found and predicted to bind ICAM1 (Lennartz et al; + the IT4var07 shown in PMID: 26119044 & PMID: 27406562 ; which I believe was not included in Lennarz et al).

“Sequences with DC5 cassettes were from different…”
..As expected from this C terminal group A domain cassette - also previously described as not associated with N terminal seq features (Rask et al 2010).

“However, such analyses will soon be replaced by methods such …”
I disagree with this prediction. RNAseq may indeed prove useful if costs are reduced further to allow enough depth perform the required assembly of var genes, which are few and rare in RNA exptracted from blood in vivo; RNA seq may prove useful if the upcoming analysis of Sangers 1000 Pf genomes suggest so. Using MS analysis it is extraordinarily difficult to do de novo assembly of multiple rare polymorphic sequences in a patient sample, which is what is required to elude further to current knowledge. On the contrary I think this work lays the ground for a developing a much more cost effective var type prediction tool using DBLa tag expression analysis, once such a tool could be developed and validated from the ~1000 pf genomes to be released from Sanger.

Is the work clearly and accurately presented and does it cite the current literature?
Partly

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Referee Report 06 October 2017

doi:10.21956/wellcomeopenres.13039.r26465

Mary M. Rorick
Department of Ecology and Evolution, University of Chicago, Chicago, IL, USA

A nice computational study comparing multiple methods of categorization for the ultra-diverse, biologically complex, and clinically important family of *var* genes of the malaria parasite *P. falciparum*. The *var* genes encode *Plasmodium falciparum* Erythrocyte Membrane Protein 1 (*Pf*EMP1). Due to the diverse, recombining nature of the *var*, from non-laboratory stains it is typically only feasible to sample a very short tag region of about 125 amino acids from the relatively conserved *var* DBLα domain. Therefore, most of the *var/Pf*EMP1 sequence variation currently available from the clinical setting consists of only this tag region. This region does not include most of the variable host endothelial binding sites that have been proposed in the literature, and which are potentially relevant to severe malaria disease. The careful mapping of the relationships between different *Pf*EMP1 classification schemes is therefore important for deciphering this protein's multiple and variable binding functions, and the various disease manifestations that likely result as a consequence. Until it is possible to obtain the complete *var*/*Pf*EMP1 sequences from large numbers of clinical isolates, this type of study is of useful for the progression of the field.

The research methods appear to be of high quality, and the paper is clear and well-written. However, there are some typos and areas where the writing could be improved. Due to the complex nature of the topic I have included many detailed suggested edits below. The majority of the suggested changes are simply to improve the clarity of the manuscript - something that is important for readers from outside the community of scientists who study *var*/*Pf*EMP1.

1. This paper relies heavily on the results of Rask et al. 2010, which lay an extensive groundwork for *var*/*Pf*EMP1 categorization. In that paper, the authors discuss the association between HB36 and cys4/cp4-6 sequences. The authors could maybe include a discussion of HB36 when they talk about predicting *ups*A/CIDRa1.

2. Throughout the manuscript, I believe PfEMP1 should be formatted with Pf in italics: *Pf*EMP1.

3. The authors explain the diversity and mosaicism of *var* genes as follows: "due to recombination between *var* genes on non-homologous chromosomes, the overall architecture of PfEMP1 encoded by different parasite genotypes is extremely diverse and sequences are mosaics of many semi-conserved sequence blocks". I think this is slightly overstating our current knowledge of the genetic mechanisms and evolutionary and ecological dynamics shaping and maintaining var diversity. Why the variants have a mosaic nature is likely party due to ectopic recombination (recombination between *vars* on non-homologous chromosomes), but also likely due to recombination between *vars* at the *same* genetic location within *homologous* chromosomes. *Var* genes surely recombine in a homologous manner at least as frequently as they recombine ectopically, and due to the diversity among different parasites even at a single *var* locus, this more normal type of recombination is also likely to generate mosaicism. Another reason it is an overstatement/misstatement: if we were going to give a reason for why there is all the *var* diversity, balancing selection is a much more direct explanation as opposed to non-homologous

recombination. Balancing selection must be invoked to explain why so much *var* diversity is maintained within the population. In my view at least, the question of the “extreme diversity” of the *var* genes is not really addressed at all by just invoking the immediate genetic mechanisms generating the variants. Given the above, I recommend simply removing the following clause from the sentence: “due to recombination between *var* genes on non-homologous chromosomes”, and replacing it with: “due to rapid recombination among *var* genes, and likely balancing selection” .

4. The authors use the phrase “the infecting parasite population” twice near the end of the second paragraph of the introduction, and in both cases I believe they specifically mean the parasites within an single, individual host. I think this phrasing is more confusing than it needs to be. For any system with population-level dynamics occurring within organisms there is the possibility for confusion about the level of hierarchy the dynamics are operating on. Specifically, in this case, I think some readers may think the authors mean the population of infective parasites rather than only those parasites that exist simultaneously within a given host individual. Another possible source of confusion is that some readers may be familiar with the fact that the *var* genes are expressed in a strictly mutually exclusive manner, and it may not be obvious that this does not translate to strict mutually exclusive expression at the level of the host individual (also I believe some early clinical results have contributed to some of this confusion). I suggest rephrasing as follows: “multiple *var* genes are expressed simultaneously within the infecting parasite population” could be changed to “while *var* genes are expressed in a strictly mutually exclusive manner at the level of the individual cell, multiple *var* genes are expressed simultaneously at the level of the infected host”; and “The range of *var* genes expressed at any one time in the infecting parasite population” could be changed to “The range of *var* genes expressed at any one time within a given host”.

5. Starting with the third paragraph of the introduction I felt the structure of the manuscript begins to get a bit confusing. I had to work too hard as a reader to follow where they were going with the introduction, and why they were presenting this information in this order. I therefore suggest simplifying the writing in the third, fourth and fifth paragraphs of the introduction, and giving the reader a bit more of an explicit “road map” indicating how the paragraphs are connected and where they are taking us. Specific suggestions follow.

6. A topic sentence could be added to the very beginning of the third paragraph: “The hope has been that it may be possible to identify a limited set of *Pf*EMP1 functional specializations, which may in turn clarify the disease process; however, it remains unclear which aspects of *var* diversity are the most relevant for achieving this goal.”

7. For simplicity, I suggest deleting “Based on full length sequences from seven laboratory isolates …”, adding the topic sentence suggested above, and changing the original first sentence of the paragraph (which would now be the second sentence of the paragraph) to the following: “All *Pf* EMP1 domain classes have been classified through global sequence alignment into a large number of highly refined and specific domain sub-classes (Rask et al., 2010).”

8. The second sentence of the third paragraph of the Introduction is not a grammatically well-formed sentence. Also it mentions 33 sub-domains, which implies regions smaller than a domain, but I believe that is not what the authors mean. I believe they mean domain sub-classes (i.e., smaller, more refined categories of domains). I suggest changing the first part of the sentence to the following: “For example, the DBLα domain can be classified into 33 domain sub-classes….”

9. I suggest rewriting the fourth paragraph as follows: “In addition to the refined domain classification schemes—which were based on the handful of sequenced laboratory strains for which we have complete *var* sequences—various *broader* classification methods have also been employed for the *var* genes that use a sparser set of their sequence features. *Pf*EMP1 genes can be classified into just five broad functional and recombination groups based on sequence similarity of their upstream promoter regions (*ups*) and chromosomal location. This classification partitions the sequences into groups A-E (citations). *UpsE* is associated exclusively with *var2CSA*, which plays a central role in placental malaria (citation). *UpsA var* gene [note typo correction here] expression is associated with severe disease (citations) and rosetting (citations). Increased transcription of *upsB* sequences have also been reported in cases of severe malaria (citation). And, while some research indicates that *upsC* sequences are expressed at higher levels in asymptomatic cases (citations), it has also been reported that *upsC* expression is associated with severe malaria (citation).”

10. Choose a consistent notation for *ups* groups (i.e. with or without italics, and with or without a space).

11. Choose whether to say “subclasses/subclass” or “sub-classes/sub-class”. Both are used in the manuscript, and consistency is the important thing. I prefer the word without the hyphen, but its just personal preference.

12. Again, mostly just for clarity, I suggest rewriting the fifth paragraph of the introduction as indicated below. I removed the “for example” and the numbering because, while the information is relevant, it is not clear to me that they are really *examples* of the initial statement.

    “*Pf*EMP1 can also be described in terms of common configurations of different subclasses of domains. These common configurations are called “domain cassettes” (DCs) (Rask et al, 2010). Twenty-three *var* DCs have been defined from full-length domain alignments of sequences from seven laboratory strains. It was initially proposed that DCs may act as functional units. However clearly defined functions have still only been assigned at the level of individual domain subclasses. Therefore, it remains unclear whether DCs represent functional units under natural selection, or whether they are just neutral artifacts of the recombinatoric diversification process. [Paragraph break here.] Research pertaining to DCs has revealed the following: Specific CIDRα1 domains, often found in the context of domain cassette 8 (DC8) and 13 (DC13), appear to bind EPCR (citation). Var genes containing DC8 cassettes seem to bind human endothelial cells from various organs, including—notably—those from brain endothelial cells (citations). DBLβ domains found within DC4 genes reportedly adhere to ICAM-1 and may be targets of broadly cross-reactive and adhesion-inhibitory IgG antibodies (citations). [Remove paragraph break that is currently here.] Clinical and laboratory studies have reported associations between DCs....”

13. Page 3, second column, fifth line from the bottom: I would remove the word “respectively”.

14. Page 4, first column, first paragraph: “by simply partitioning tags to those with and those without two cysteins” does not make grammatical sense, and it implies 2 versus 0 cysteins, which I believe is not what the authors mean. This could be changed to: “by simply partitioning tags by whether they contain two cysteins or some other number of cysteins”.

15. Page 4, first column, line 9: I would remove “Furthermore”.

16. Page 4, first column: For clarity, I recommend changing the sentence “Furthermore, Lavstsen et al., 2012 have suggested that information on EPCR binding by CIDRα1 within DC8 and DC13 may be unavailable within the DBLα tag due to a recombination hotspot situated between the DBLα tag region and the CIDRα domain” to the following: “Lavstsen et al., 2012 have suggested that the DBLα domain tag may not be informative about whether its flanking CIDRα1 domain binds EPCR because there is a recombination hotspot situated between the DBLα tag region and the CIDRα1 domain” [also, note that I changed CIDRα to CIDRα1 since I thought that was likely a typo].

17. When referring to a paper within a sentence (as opposed to the parenthetical manner at the end of a sentence) the authors sometimes use no parentheses: “Lavstsen et al., 2012 have suggested…” and other times use parentheses: “classification that was suggested by (Rask et al., 2010)”. The style should at least be consistent, and ideally also consistent with the journal’s formatting recommendations for this type of citation.

18. At least in two places there is a weird comma after “*var”* that does not appear to belong: Within the first paragraph of “Results and discussion”, and within the 6th line up from the very end of the manuscript.

19. Final line of the first paragraph of “Results and discussion”: I recommend changing “sensitivity, specificity analysis” to “sensitivity-specificity analysis”.

20. For consistency and clarity, use italics for “ups” or spell it out and don’t use the abbreviation. For example, page 4, second column, 7th line up from the bottom.

21. I believe there is an error in the title of the section “DBLα sub-domains are not all homogeneous groups”. I believe it should read “DBLα domain subclasses are not all homogeneous groups”. A “sub-domain” and a domain “sub-class” are completely different things. The first is physically smaller than a domain, the second is a smaller category of a complete domain. As far as I can tell, everywhere the authors use the term “sub-domain” they actually mean “domain sub-class” (or equivalently “domain subclass”).

22. First sentence of the section “DBLα sub-domains are not all homogeneous groups”: Grammatical errors. Insert “The” before “Domain classification that was suggested…”, and replace “were” with “was”.

23. Page 5, second column: These two sentences feel pretty meaningless to me, plus it seems the meaning the authors are trying to convey is redundant with the sentences the precede and follow. Therefore I would just delete both of the sentences: “The domain classification used in (Rask et al., 2010) has therefore brought together distinct sequences within the same sequence classification. This suggests that the newly defined sub-domains do not always classify sequences into wholly genetically distinct groups.”

24. I think “This discordance between methods of classification, employing global and local sequence comparisons…” would read more clearly as follows: “This discordance between methods of classification when employing global versus local sequence comparisons…”

25. The above sentence continues by describing “a mode of diversification… that we might speculate leads to impaired recognition and clearance of PfEMP1 antigens by the immune system.” I think this sentence presents an interesting biological hypothesis that could be elaborated on more. To

me it is not immediately obvious that the discordance should be interpreted in this biological manner. An alternative interpretation might be that some categorization methods are less informative about recombination patterns, or more noisy, for example. I simply find the hypothesis interesting and warranting of further discussion.

26. Page 6, first column: I recommend changing: “recombination hotspot proposed to be situated 3’ of the DBLα tag region” to “recombination hotspot purportedly situated 3’ of the DBLα tag region”.

27. Page 6, first column: I recommend changing: “DBLα tag region and parts of the genes encoding important cytoadhesive regions potentially limits the predictive information available with DBLα tag sequence” to “DBLα tag region and **the** parts of the **gene** encoding important cytoadhesive regions potentially limits the predictive information available with **the** DBLα tag sequence”.

28. Page 6, first column: I recommend adding a colon to “multiple cys/PoLV groups CP2, CP3 and CP4” so it reads “multiple cys/PoLV groups: CP2, CP3 and CP4”.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
No source data required

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Referee Expertise: Evolutionary genetics, ecology and evolution of infectious disease

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.