Replication and Review of Githinji & Bull 2017

Dieu My T. Nguyen[1, *] and Daniel B. Larremore[2, 3, †]

[1]*Interdisciplinary Quantitative Biology Program, University of Colorado Boulder*
[2]*Department of Computer Science, University of Colorado Boulder*
[3]*BioFrontiers Institute, University of Colorado Boulder*

The malaria-causing parasite *Plasmodium falciparum*'s virulence factor (*P. falciparum* erythrocyte membrane protein 1, PfEMP1) adheres to the infected surface of human erythrocytes. Natural malaria immunity is facilitated by the immune system's response to PfEMP1, which can also be the target of malaria vaccines. However, this protein is encoded by a diverse family of about 60 *var* genes. This diversity gives rise to antigenically diverse binding properties of the protein and therefore varying severity of malaria. There have been extensive sequence analyses of PfEMP1 sequences from the currently available seven parasite genomes. Due to the limited number of full-length *var* gene sequences available, many studies have also classified a specific conserved subdomain ("tag"') of PfEMP1 called Duffy binding-like alpha (DBLα) to draw information about cytoadhesion properties of the parasite's virulence factor and severity of malaria. Githinji & Bull 2017 compared DBLα tag classifications with sequence features of full-length *var* genes to show that the tags may provide insight into the functional specializations of *var* genes. In this review, we attempted to reproduce the results presented in Githinji & Bull 2017 and found that they were almost completely reproducible. As part of this replication, we provide open Python code, allowing the authors and others to see in detail how we used the datasets and implemented the methods in Python. This project was a 2-month rotation project in the Interdisciplinary Quantitative Biology program at the University of Colorado Boulder.

I. INTRODUCTION

The *Plasmodium falciparum* parasite is the most lethal of the five *Plasmodium* parasites responsible for malaria in humans. Once transmitted by the *Anopheles* mosquito to humans, the parasite exports to the surface of an infected red blood cell (RBC) a virulence factor called *Plasmodium falciparum* erythrocyte membrane protein 1 (PfEMP1), which is the target of the host immune system in malaria infections [6]. The type of PfEMP1 present on the infected RBC plays a key role in the clinical severity of the infection. Therefore, vaccines based on PfEMP1 immunity is a hope for malaria treatments. However, the parasite can produce antigenically different proteins by switching on and off about 60 different PfEMP1-encoding genes, called *var* [7]. The hyper-variant types of PfEMP1 have different binding properties to human endothelial receptors, and cause the immune system's antibodies to not always recognize the protein to kill the infected cell [7]. Thus, dissecting PfEMP1 diversity is a problem with possible clinical significance.

PfEMP1 molecules are made up of two to nine domains: N-terminal segment, Duffy binding-like (DBL), cysteine inter-domain region (CIDR), and acidic terminal segment domains [12]. The head structure of the protein containing DBL and CIDR domains that are known to mediate important binding properties of the parasite. Based on sequence similarity, the DBL and CIDR domains have been divided into subclasses ($\alpha, \beta, \gamma, \delta, \epsilon, \zeta$ and $\alpha, \beta, \gamma, \delta$ respectively) [12] . CIDRα domains encode PfEMP1 binding to the host receptors called cluster determinant 36 (CD36) and endothelial protein C receptor (EPCR), which in turn are linked to particular clinical symptoms such as cerebral malaria [9]. A subset of DBLα domains are linked to rosetting, a process that causes infected RBCs to bind to uninfected RBCs and has been clinically linked to respiratory distress [10]. Understanding the structure and composition of PfEMP1 proteins by analyzing the diverse makeup of the *var* genes that encode them is therefore critical to understanding malaria's abilities to evade the immune system and cause severe disease.

Due to their modular domain structure and their diversity, many different approaches have been taken to categorize *var* genes, in an effort to understand how *var* categories might represent functional or evolutionarily important groups. Based on full-length sequences from seven laboratory *P. falciparum* parasites, PfEMP1 *var* genes have been classified by broad structural methods. The upstream promoter region (ups) classification divides the sequences into groups A-E [12], [14]. Domain alignment of full-length sequences yields 23 *var* "domain cassettes" (DCs), some of which are linked to clearly defined functions, such as DC8 *var* gene proteins binding to brain endothelial cells.

Some studies have also explored the classification of the DBLα domains ("tags"), notably in two different approaches. The first approach involves grouping the tag sequences based on the number of cysteines and the mutually exclusive motifs MFK and REY [5]. These groups are called cys/positions of limited variability (Cys/PoLV) groups. The second approach uses network analysis to group together the sequences that share blocks of sequence with each other, with the two prominent groups being block-sharing groups 1 and 2 (BS1 and BS2) citebull2008.

[DM: True stats about number of methods?] In total, there are four different ways of classifying *var* genes, and two ways of categorizing their DBLα tags. In an effort to map the similarities and differences among these various classifications,

* dieu.nguyen@colorado.edu
† daniel.larremore@colorado.edu

Githinji & Bull 2017 aimed to assess the relationship between several DBLα tag classifications and the features of full-length *var* gene sequences, to show that the tags can provide information on some important features of full-length sequences. Here, we aim to reproduce the results of this paper to learn about the biology of the malaria parasite and the computational methods used in assessing the DBLα domain as a functional predictor of *var* gene sequences. We also refer to [4], [5] and [12] for more details on methods for DBLα tag and full sequence classifications used in Githinji & Bull 2017. Specifically, we replicate and provide open Python code [1] for the two approaches the authors have used to classify the DBLα tags (Cys/PoLV and block-sharing group classifications), and all figures presented in the paper.

This replication effort is the result of a two-month rotation project for the Interdisciplinary Quantitative Biology program at the University of Colorado Boulder. Replication code is written in Python 3.6.2 [1] and network visualizations were done in webweb [2].

II. METHODS & RESULTS

A. DBLα tag sequence classifications

We first explored the two different approaches that the authors have used to classify DBLα tags in previous papers, referred to as Cys/PoLV [5] and block-sharing groups [4]. For both approaches, we obtained the 1548 DBLα sequences from the file "1548_tags.fa" from the authors' Open Science Framework (OSF) storage at https://osf.io/uwcn2/ under "datasets."

1. *Cys/PoLV classification*

The Cys/PoLV approach, described in detail in [5], involves extracting two features from the sequences: 1) the number of cysteines and 2) motifs located at positions of limited variability (PoLV) – in particular, the presence or absence of mutually exclusive motifs MFK at PoLV1 and REY at PoLV2. As seen in Figure 1A, cys2 and cys4 groups have the most DBLα sequences, explaining the rationale behind the use of the two cys groups as the main groups for the Cys/PoLV classifications. The sequences are further grouped into six Cys/PoLV groups based on the [5]'s definitions:

Group 1: cys2, MFK* motif present at PoLV1
Group 2: cys2, *REY motif present at PoLV2
Group 3: cys2, not in groups 1, 2
Group 4: cys4, not group 5
Group 5: cys4, *REY motif present at PoLV2
Group 6: cys1, 3, 5, or >5

[5] hypothesized that genetically isolated sequences that do not recombine with one another maintain a distribution in sequence length. If the Cys/PoLV grouping based on some sequence similarities is accurate, the sequences in each group should have similar lengths (measured in terms of number of amino acids). Like the authors, we find that the lengths of the sequences are similar within groups and the groups follow a similar distribution of lengths (Figure 1B).

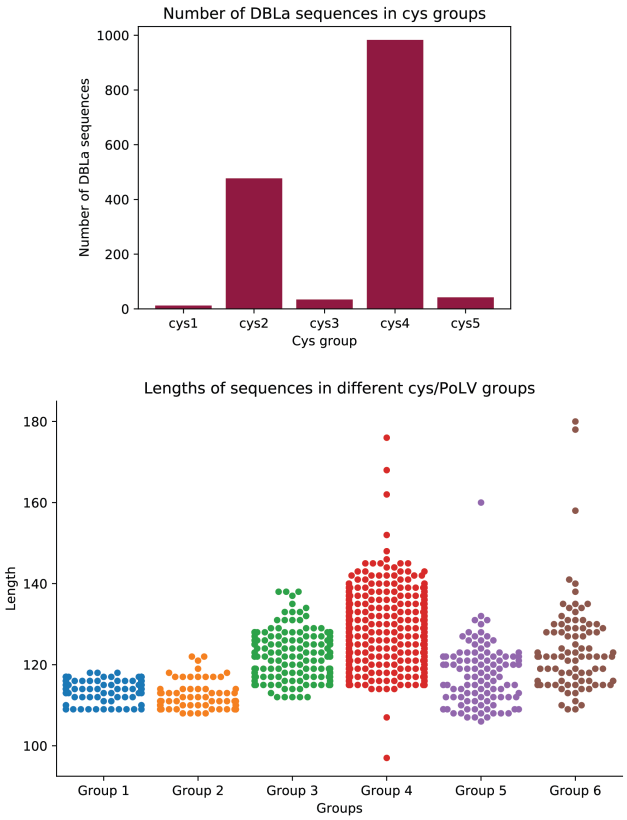

FIG. 1. **(A)** Number of DBLα sequences in each Cys group. **(B)** Comparisons of sequence lengths (based on number of amino acids) of DBLα sequences in different Cys/PoLV groups.

2. *Block-sharing network& classification*

Another method of classifying the DBLα tags comes from a network analysis approach, described in detail in [4]. For each sequence, we identify four polymorphic blocks at fixed locations based on three conserved anchor points which are annotated in Figure 2 (similar to [4] Figure 1B): D at the beginning, WW (or W followed by another amino acid) in the middle, and R at the end of the sequence. Each 10-amino acid (aa) block is a "position specific polymorphic block" (PSPB). These PSPBs are then used to construct a "block-sharing network" structure, in which sequences are represented by nodes that are linked by edges (lines in network) if they match at one or more PSPBs. Edges are not weighted in regard to number of shared PSPBs. The network structure is shown in Fig-

FIG. 2. **Position specific polymorphic block**. Four polymorphic blocks (purple) for four example tag sequences at fixed locations based on three conserved anchor points shaded in orange.

ure 3, in which we can also observe what [4] and Githinji & Bull 2017 saw: The network has two prominent lobes: a large one in the center and a smaller one on the right of large lobe. After construction of the network structure, we observe the sharing of 14-aa PSPBs within the DBLα tags from *P. falciparum* parasites from Kenyan children to find the two largest connected components: Block-sharing group 1 (BS1) and 2 (BS2), shown in Figure 7.

From the perspective of reproducibility, we note that partitioning the sequences into BS1 and BS2 was slightly challenging. [4] Figure 4C shows the network components obtained by using 14-aa long PSPBs, giving seven block-sharing groups, of which the two most prominent ones (Bs1 and Bs2) are used for sequence classifications in Githinji & Bull 2017. We could not find more details on how to identify the seven BS groups, so we followed [4] Perl script ("mmi0068-1519-SD3.pl") to assign sequences to BS1, BS2, or neither. In this way, the block-sharing groups are hard-coded within the Perl script, but cannot be reproduced *de novo*.

B. Full-length *var* gene sequence classifications

In this paper, the authors obtained full-length *var* genes classifications from the literature, notably [12]. These classifications include: 33 DBLα subdomains (DBLa 0.1-0.24, DBLα 1.1-1.8 and DBLa2), 5 ups groups (A-E), 628 homology blocks (HB), and 23 domain cassettes (DC). Some of these classes have been associated with severe malaria and are further discussed below.

C. Figures: Relationships between DBLα tags and full-length *var* sequences

Below is our reproduction of the figures in Githinji & Bull 2017 in the same order as the paper. Because we've confirmed above that the Cys/PoLV and block-sharing classifications were successfully reproduced, for the visualizations below, we use the data from Githinji & Bull 2017 file "curated_data_set.csv" (also on OSF Storage under "datasets") because it also includes full-length *var* gene classifications from [12] and other sources that we otherwise do not have access to.

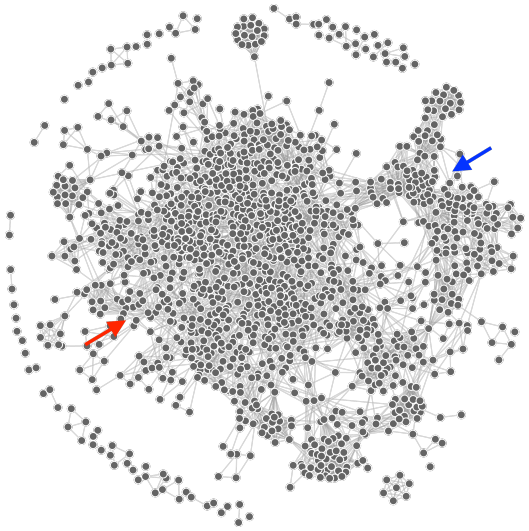

FIG. 3. **Block-sharing network structure.** The network is constructed based on DBLα tag sequences matching at one or more 10 amino acid PSPBs. Each sequence is represented by a node. Nodes that share PSPBs are linked by edges and are in the same region of the network. Two main lobes are observed: a large one in the center (pointed to by red arrow) and a smaller one right of large lobe (blue arrow). Nodes with few or no connections are placed at the perimeter of the network.

1. *Bar graphs*

The bar graphs provide a straightforward visualization of the relationship between different *var* gene classifications (upstream promoter sequence (ups), Cys/PoLV, BS, and HB and the specific DBLα domains, CIDR1 domains, and domain cassettes contained in the sequences. We use the same color scheme and arrangement of information (in decreasing upsA order) as the authors did, for easy comparison. Overall, the bar graphs below (Figure 4, Figure 5, Figure 6) are identical to those in Githinji & Bull 2017 Figure 1-3. As seen across the 3 figures, BS1 sequences are closely associated with upsA, while BS2 sequences with upsB and upsC. Most cys2 sequences (CP groups 1-3) are found in upsA sequences, but some are also found in upsB and upsC. Furthermore, DC8 sequences are associated with severe malaria [11]

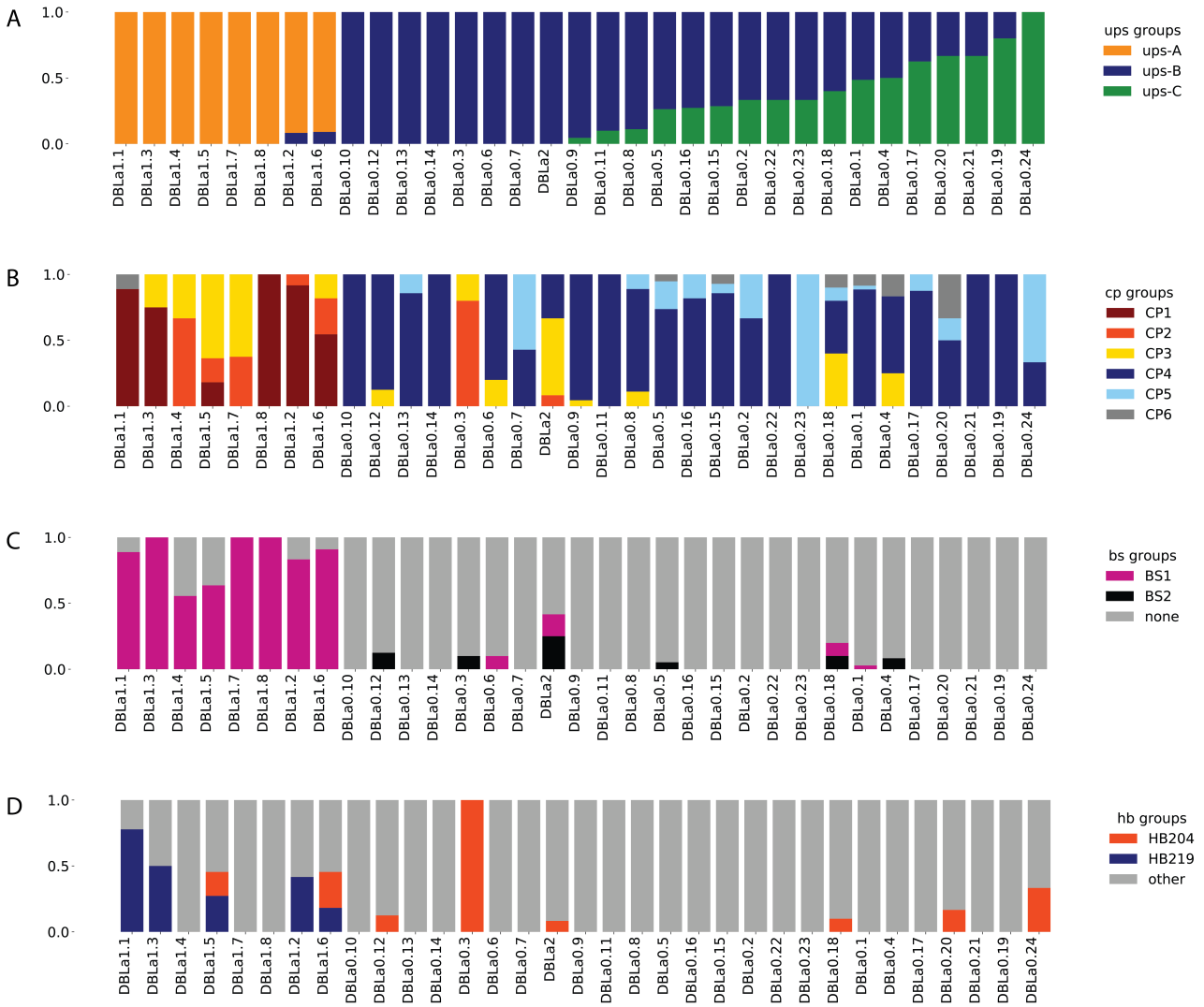

FIG. 4. **Correspondence between *var* sequence classifications and possession of specific DBLα domains**. *var* sequences are classified based on DBLα domains (horizontal axis) they contain. The proportion of the genes carrying other sequence features (ups, Cys/PoLV, block-sharing groups, select homology blocks) is shown on the vertical axis. Like in Githinji & Bull 2017, the DBLα domains are, from left to right, in order of decreasing upsA sequences.

[12]. In this analysis, DC8 sequences contain CP groups 2, 3, and 4 as well as most of the BS2 tags. This is consistent with the clinical finding of DC8-like sequences in two severe cases of malaria in Kenya [3]. Although this is based on limited information, as Githinji & Bull 2017 suggests, these findings may imply that *var* genes sampled from Africa may commonly share BS2 sequences.

2. *Network visualizations*

Built on the analysis in [4], the network visualizations in Githinji & Bull 2017 Figure 4, 5 provide information on how specific subsets of full-length *var* sequences are mapped onto the network based on the sharing of PSPBs by the DBLα tags. In Figure 7, we show our network analyses of several classifications: Cys/PoLV, block-sharing groups, UPS, DC (4,5,8,13), predicted EPCR binding, and CD36-binding. The clustering tendencies of the tag sequences in our networks are similar to those in Githinji & Bull 2017 Figure 4. Consistent with the bar graph analysis, DC8 sequences show to occupy

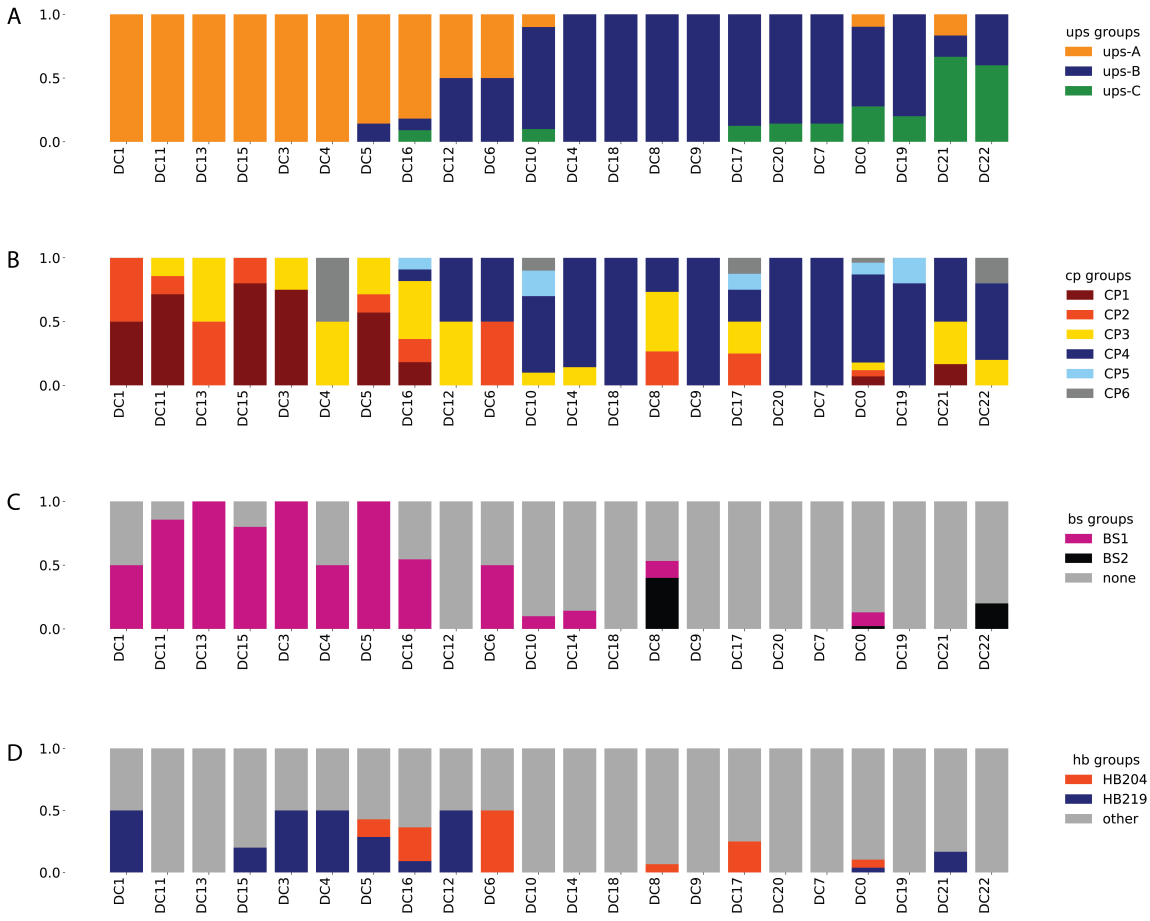

FIG. 5. **Correspondence between *var* sequence classifications and possession of specific domain cassettes (DCs).** ***var*** **sequences are classified based on DCs (horizontal axis) they contain**. The proportion of the genes carrying other sequence features (ups, Cys/PoLV, block-sharing groups, select homology blocks) is shown on the vertical axis. Like in Githinji & Bull 2017, the DCs are, from left to right, in order of decreasing upsA sequences.

the same region of the network as upsB and upsC sequences.

Furthermore, Figure 8 shows the analysis of the DBLα tags from known DC8 *var* genes, with the visualization created with Gephi 0.9.2. We are able to reproduce the largest connected component (star-shaped), the three groups outside of this largest component, and the isolated nodes/sequences shown in Githinji & Bull 2017 Figure 5. We also show the same results for the block-sharing group classification of each sequence, color-coded in the figure.

3. *Receiver operator characteristic curves*

When evaluating the quality of a parameterized prediction scheme, a common approach is to plot the relationship between sensitivity (false positives) and specificity (true positives). A set of these curves, called receiver operator curves (or ROC curves [sic]) was used in Githinji & Bull 2017 Figure 6 to how three DBLα tag classifications (cys2, cys2bs1, cys2bs_CP1) predict four *var* gene features (upsA [15], DC8 [11] [12], DC13 [15], CIDRα 1 [13]) which have been associated with malaria severity in previous papers. Our ROC curves in Figure 9A, C, D are similar to those in Githinji & Bull 2017 Figure 6A, C, D. As the authors noted, particularly, CIDRα 1 domain, which is associated with severe malaria due to binding to EPCR [13], is associated with "group A-like sequences" (cys2bs1). Previous reports have also shown associations between subsets of cys2 sequence tags and DC8 and DC13 *var* genes with severe disease phenotypes [15]. This is reproduced in Figure 9C showing prediction of DC13. With prediction for DC8 when compared to the authors? results, however, our ROC curves show higher sensitivity for predicting DC8 from cys2, and both lower sensitivity and lower specificity for predicting DC8 from cys2bs1

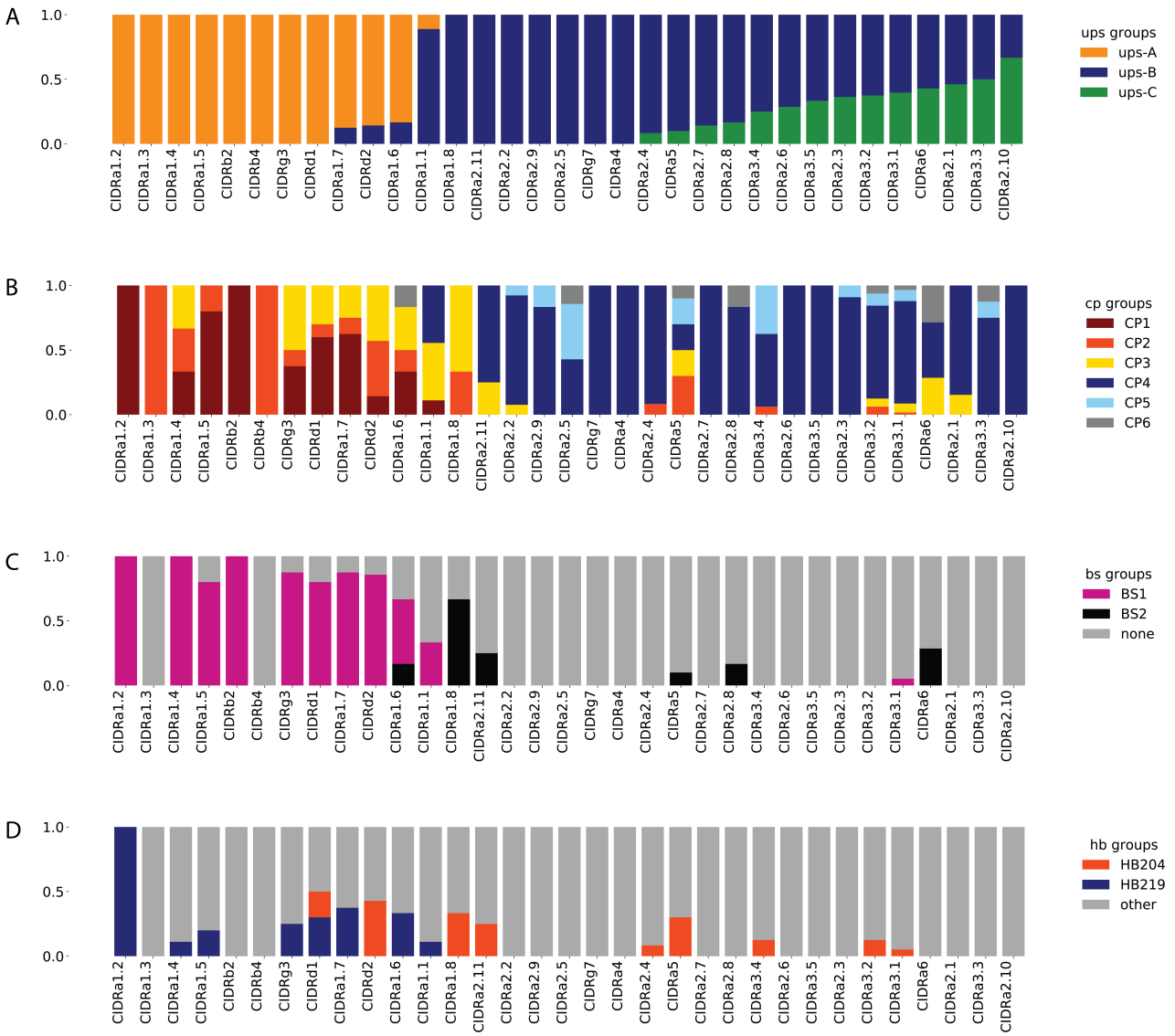

FIG. 6. **Correspondence between *var* sequence classifications and possession of specific CIDR1 domains. *var* sequences are classified based on CIDR1 domains (horizontal axis) they contain**. The proportion of the genes carrying other sequence features (ups, Cys/PoLV, block- sharing groups, select homology blocks) is shown on the vertical axis. Like in Githinji & Bull 2017, the CIDR1 domains are, from left to right, in order of decreasing upsA sequences.

and cys2bs1_CP1 such that these two curves are below the 45 diagonal of the ROC space. Githinji & Bull 2017 Figure 9B shows the ROC curves for the prediction of DC8 from cys2bs1 and cys2bs1_CP1 as roughly lying on the 45 diagonal. Together with our results, it seems that these two tag classifications are not highly accurate in providing prediction of the DC8 feature of *var* genes.

III. CONCLUSION

In summary, we have studied and reproduced the methods and results in Githinji & Bull 2017, which brings together previous papers to present an analysis of the correspondence between the biologically complex full-length *var* genes' features and one of their domains, the DBLα tags. This analysis shows that despite their diversity, DBLα tag classification can help us determine the features of the full-length *var* genes. Being able to predict the features that are associated

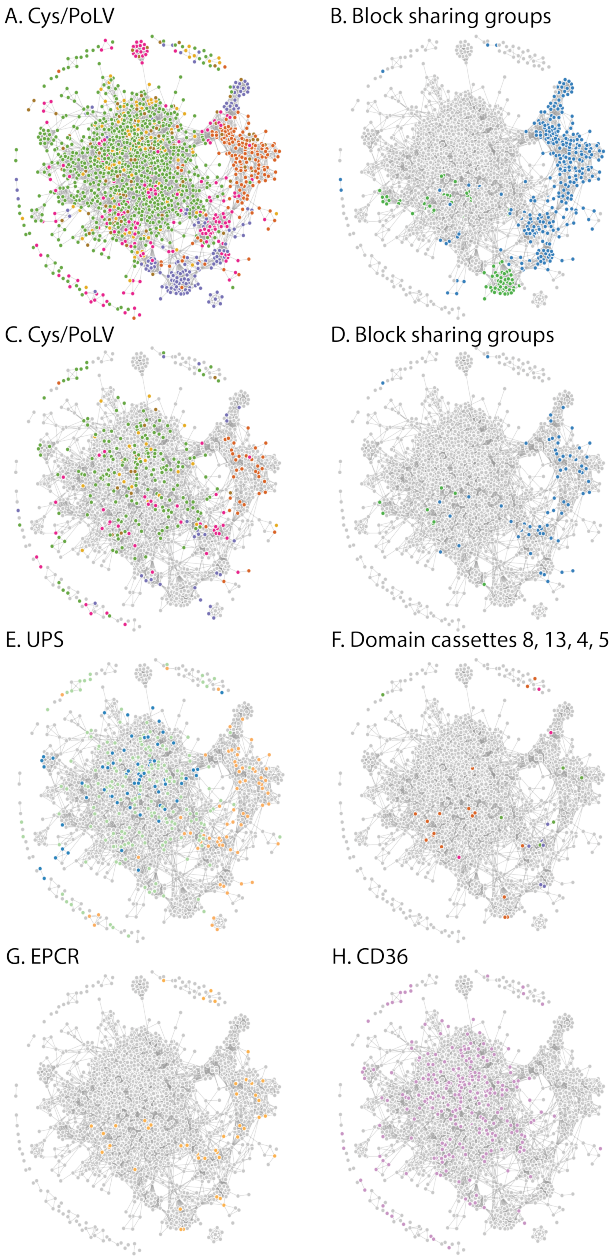

FIG. 7. **Various DBLα tag classifications mapped onto the block-sharing network**. (A) Cys/PoLV analysis for all sequences; (B) BS analysis for all sequences; (C) ups grouping; (D) Cys/PoLV analysis for full length *var* gene sequences from 6 laboratory isolates; (E) BS analysis for full length *var* gene sequences from 6 laboratory isolates; (F) domain cassette (DC) classification for DC4, DC5, DC8 and DC13; (G) predicted EPCR-binding phenotype due to CIDRα 1.1, CIDRα 1.4, CIDRα 1.5, CIDRα 1.6, CIDRα 1.7 or CIDRα 1.8 (Lau et al., 2015) for sequences with CIDRα information available; (H) predicted CD36-binding phenotype due to CIDRα 2, CIDRα 3, CIDRα 4, CIDRα 5 (Robinson et al., 2003) for sequences with CIDRα information available. Node colors: For all, unclassified = 0. (A&D) red = CP1, purple = CP2, pink = CP3, green = CP4, yellow = CP5, brown = CP6. (B& E) blue = BS1, green = BS2; (C) upsA = orange, green = upsB, blue = upsC; F) pink = DC8, purple = DC5, green = DC13, orange = DC4; G) orange = predicted EPCR binding; H) purple = predicted CD36 binding.

with severe malaria is clinically valuable, especially when sequencing the hyper-variable *var* genes is challenging but sequencing DBLα tags is more tractable.

The figures and methods described in the paper are clear and easily understood, making the paper almost completely reproducible, except for a minor difference in the ROC curves discussed in section C above. The open datasets and authors' code provide us a convenient way to access and use the same datasets in our replication and to compare our results. Reproducing this work has been a productive experience to learn the

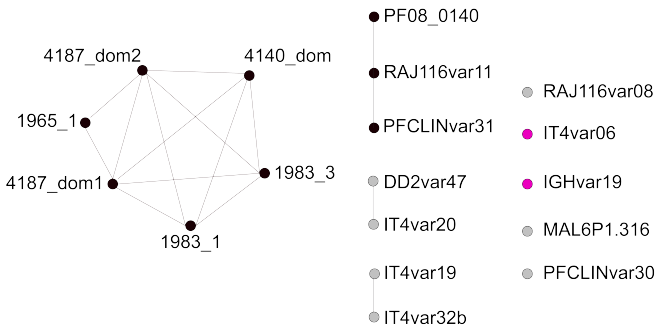

FIG. 8. **Network analysis of DBLα tag sequences from known DC8 *var* genes**. Node colors: grey = not in a BS group; pink = BS1; black = BS2.

biology of malaria as well as the analysis methods and findings this community of researchers. This paper also opens future directions for continuous exploration of the DBLα tags as a predictor of functional features of full-length *var* gene sequences, especially with the Sanger Institute releasing 1000 more *P. falciparum* whole genomes in the near future.

[1] http://github.com/dieumynguyen/githinji_vargenes.

[2] http://danlarremore.com/webweb/.

[3] BULL, P., BERRIMAN, M., KYES, S., AND ET AL. Plasmodium falciparum variant surface antigen expression patterns during malaria. *PLoS Pathog 1*, 3 (2005), e26.

[4] BULL, P., BUCKEE, C., KYES, S., KORTOK, M., THATHY, V., GUYAH, B., MCVEAN, S. J., NEWBOLD, C., AND MARSH, K. Plasmodium falciparum antigenic variation. mapping mosaic var gene sequences onto a network of shared, highly polymorphic sequence blocks. *Mol Microbiol 68*, 6 (2008), 1519–1534.

[5] BULL, P., KYES, S., BUCKEE, C., MONTGOMERY, J., KORTOK, M., NEWBOLD, C., AND MARSH, K. An approach to classifying sequence tags sampled from plasmodium falciparum var genes. *Mol Biochem Parasito 154*, 1 (2007), 98–102.

[6] CHAN, J., HOWELL, K., REILING, L., AND ET AL. Targets of antibodies against plasmodium falciparum-infected erythrocytes in malaria immunity. *J Clin Invest 122*, 9 (2017), 3227–3238.

[7] GARDNER, M., HALL, N., FUNG, E., AND ET AL. Genome sequence of the human malaria parasite plasmodium falciparum. *Nature 419*, 6906 (2002), 498–511.

[8] GITHINJI, G., AND BULL, P. A reassessment of gene-tag classification approaches for describing var gene expression patterns during human plasmodium falciparum malaria parasite infections. *Wellcome Open Res 2*, 86 (2017).

[9] HSIEH, F., TURNER, L., BOLLA, J., AND ET AL. The structural basis for cd36 binding by the malaria parasite. *Nat Commun 7*, 12837 (2016).

[10] LAU, C., TURNER, L., JESPERSEN, J., AND ET AL. Structural conservation despite huge sequence diversity allows epcr binding by the pfemp1 family implicated in severe childhood malaria. *Cell Host Microbe 17*, 1 (2015), 118–129.

[11] LAVSTSEN, T., TURNER, L., SAGUTI, F., AND 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 109*, 26 (2012), E1791–800.

[12] RASK, T., HANSEN, D., THEANDER, T., PEDERSEN, A., AND LAVSTEN, T. Plasmodium falciparum erythrocyte membrane protein 1 diversity in seven genomes–divide and conquer. *PLoS Comput Biol 6*, 9 (2010).

[13] TURNER, L., LAVSTSEN, T., BERGER, S., AND ET AL. Severe malaria is associated with parasite binding to endothelial protein c receptor. *Nature 498*, 7455 (2013), 502–505.

[14] VZQUEZ-MACAS, A., MARTNEZ-CRUZ, P., CASTAEDA-PATLN, M., AND ET AL. A distinct 5' flanking var gene region regulates plasmodium falciparum variant erythrocyte surface antigen expression in placental malaria. *Mol Microbiol 45*, 1 (2002), 155–167.

[15] WARIMWE, G., FEGAN, G., MUSYOKI, J., AND ET AL. Prognostic indicators of life-threatening malaria are associated with distinct parasite variant antigen profiles. *Sci Transl Med 4*, 129 (2012), 129ra45.

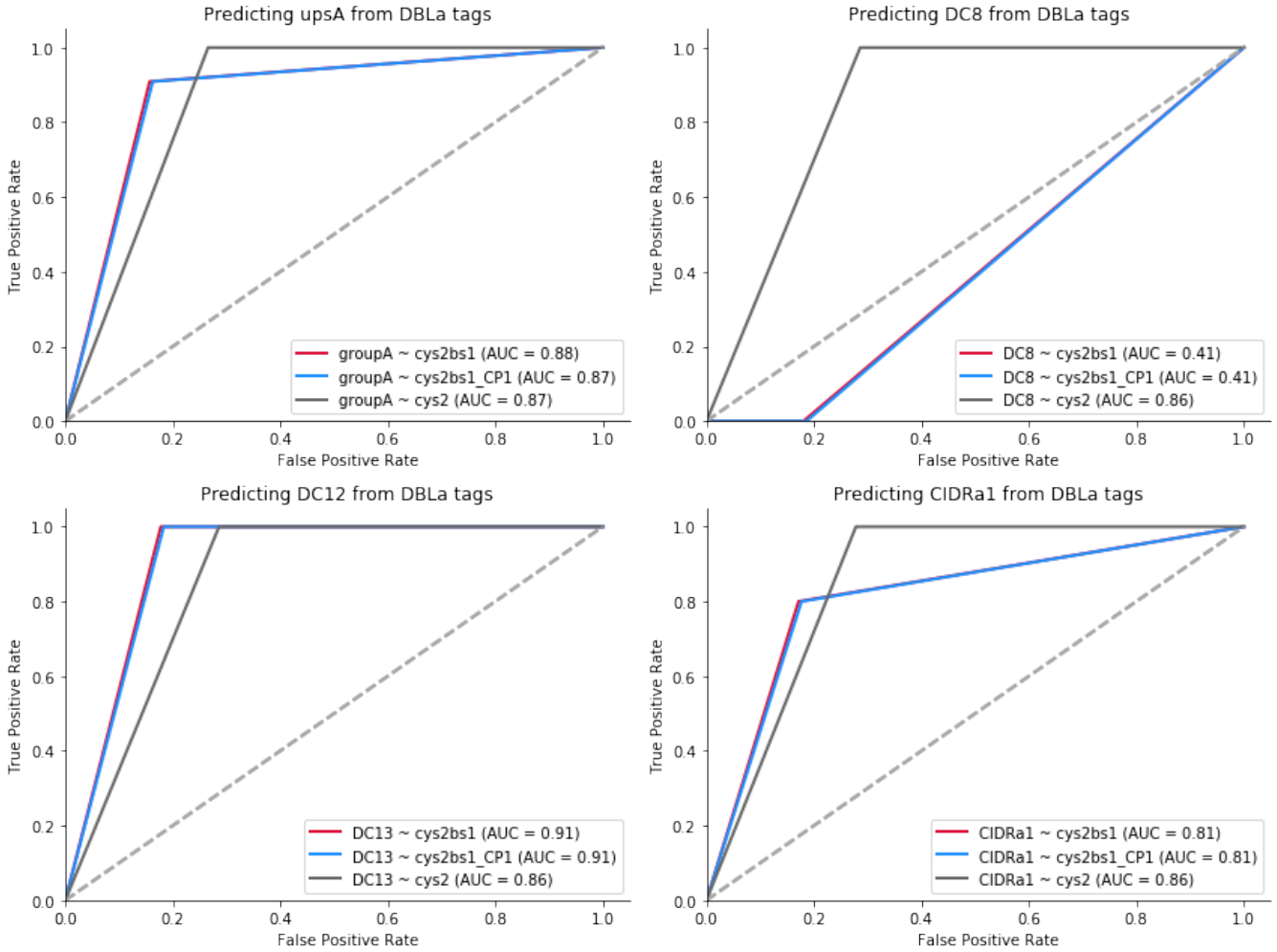

FIG. 9. **Receiver operator curves showing the sensitivity (true positive rate) and specificity (false positive rate) of three DBLα tag classifications (cys2, cys2bs1, cys2bs1_CP1) in predicting four *var* gene features associated with malaria severity: upsA , DC8, DC13 , CIDRα 1.** Sequences from the genomes 3D7 and IT4 were excluded because they were used in developing the BS classification. cys2 = two cysteines within the tag region; cys2bs1 = tag sequences in block-sharing group1 AND have two cysteines, defined as “group A-like”; cys2bs1_CP1 = cys2bs1 OR in Cys/PoLV group 1.