bioRxiv preprint doi: https://doi.org/10.1101/159483; this version posted July 20, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Reverse immunodynamics: a new method for identifying targets of protective immunity

Katrina J Spensley[1*], José Lourenço[1*], Paul S Wikramaratna[2*], Bridget S Penman[1], Andrew Walker[1], Adrian L Smith[1], Oliver G Pybus[1], Létitia Jean[3] & Sunetra Gupta[1**]

1. Department of Zoology, University of Oxford, Oxford OX1 3PS
2. Institute of Evolutionary Biology, University of Edinburgh, Edinburgh EH9 3JT
3. The Sir William Dunn School of Pathology, South Parks Road, Oxford OX1 3RE

*these authors contributed equally to the manuscript

**Correspondence to: sunetra.gupta@zoo.ox.ac.uk (SG)

Abstract

Despite a dramatic increase in our ability to catalogue variation among pathogen genomes, we have made far fewer advances in using this information to identify targets of protective immunity. We propose a novel methodology that combines predictions from epidemiological models with phylogenetic and structural analyses to identify such targets. Mathematical models predict that strong immune selection can cause antigenic variants to exist in non-overlapping combinations. A corollary of this theory is that targets of immunity may be identified by searching for non-overlapping associations among antigenic variants. We applied this concept to the AMA-1 protein of the malaria parasite *Plasmodium falciparum* and found strong signatures of immune selection among certain regions of low variability which could render them ideal vaccine candidates.

bioRxiv preprint doi: https://doi.org/10.1101/159483; this version posted July 20, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Apical membrane antigen 1 (AMA-1) is a trans-membrane protein common to all *Plasmodium* species with a putative role in invasion at erythrocytic and pre-erythrocytic stages of the parasite life-cycle. Antibodies to AMA-1 appear to inhibit erythrocyte invasion and are thus thought to play a role in clinical protection against both *P. falciparum* [1-7] and *P. vivax* [8]. Cellular responses to AMA-1 may provide immunity against both erythrocytic [9] and pre-erythrocytic forms of *P. falciparum* [10]. The potential of AMA-1 as a component of a multi-stage malaria vaccine is compromised by the polymorphism of many of the targets of natural immunity [11,12] as well as by difficulties in inducing lasting immune responses through vaccination [13-16]. Identifying variable regions of AMA-1 that are under strong immune selection would afford the possibility of focusing vaccine-induced immune responses on these regions and ensure full coverage of variability at these critical epitopes.

The extracellular region of AMA-1 can be structurally resolved into three domains: DI, DII and DIII. The highest level of polymorphism is seen among residues in DI and DII surrounding a conserved hydrophobic trough [17,18] which plays a vital role in the invasion process [19-21]. Antibodies against this region have been shown to block invasion in a strain-specific manner; however, it is also clear that these antibodies constitute only a small proportion of the total inhibitory humoral response [12,22,23]. *Pf* AMA1 also contains a number of less variable residues across the three domains, and our aim was to identify which of these were also located within epitopes eliciting protective immune responses.

bioRxiv preprint doi: https://doi.org/10.1101/159483; this version posted July 20, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

We attempted to identify these additional epitopes under strong immune selection by using a novel approach based on the predictions of a multilocus mathematical model of pathogen evolution [24-26]. The central premise is that regions or epitopes under strong immune selection will self-organise into non-overlapping combinations of variants [24-26] and supported by empirical observations within a variety of host-pathogen systems [27-32]. We inverted this result to search for signatures of immune selection among polymorphic loci by identifying those pairs which exhibit a high degree of 'non-overlap' but excluding those that may have arisen as a consequence of shared ancestry or due to structural interactions. Using this "reverse immunodynamic" method, we found strong signatures of immune selection among a small number of dimorphic residues in Domain III of *Pf*AMA-1, suggesting that epitopes containing these residues are targets of protective immunity.

We implemented the reverse immunodynamic method by performing pairwise comparisons of all dimorphic sites among 1198 unique *Pf*AMA-1 sequences obtained from GenBank, to determine whether any of these pairs exhibited non-overlapping associations between amino acid residues. We used an Information Theory statistic known as Mutual Information (MI), scaled to its maximum value, to measure the degree of non-overlap in amino acid combinations between pairs of sites. Most pairs of sites showed low non-overlap, with only 10 pairs exhibiting a scaled MI in excess of 0.3 (Tables 1 & S1). Of these, pairs 404/405, 448/451 and 283/285 were eliminated on grounds of structural proximity, leaving 7 pairs as shown by the bars in Fig 1A.

bioRxiv preprint doi: https://doi.org/10.1101/159483; this version posted July 20, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

To assess whether high non-overlap between sites could have arisen by neutral evolutionary processes, we estimated a phylogenetic tree from the sequences and calculated the parsimony score (PS) for each site, which indicates the minimum number of mutations (in this case, back and forth between the two possible amino acids at each site) required to explain the observed ancestral relationships. A new statistic S, the product of MI and the geometric mean of PS for each site, was found to preserve the same order as the scaled MI scores of the 7 sites under consideration, except in the case of the 206/225 (Table S2). To obtain an appropriate null distribution of S for hypothesis testing, we simulated 1000 sets of sequences which were (a) consistent with the empirical phylogenetic tree and (b) had the same PS score at each dimorphic site as the empirical data. The null distribution of the top-scoring measures of S (Fig 1C) shows that neutral evolution can generate sites with high MI or high PS values, but not both. In contrast, the pairs of sites indicated in red in Fig 1A lie well outside the null distribution of high-scoring site-pairs (Fig 1C). Only site-pair 206/225 (shown in green) did not meet this criterion. Reasons for this pattern become evident upon inspection of the distribution of the amino acid variation on the phylogenetic tree (Fig 2). The dominant non-overlapping pairs of amino acids at 206/225 are highly segregated on the phylogeny and not improbable under the null hypothesis, while those of 451/485 exhibit a more interdigitated pattern, indicating multiple independent and convergent paired amino acid changes that are exceptionally unlikely in the absence of immune selection.

bioRxiv preprint doi: https://doi.org/10.1101/159483; this version posted July 20, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

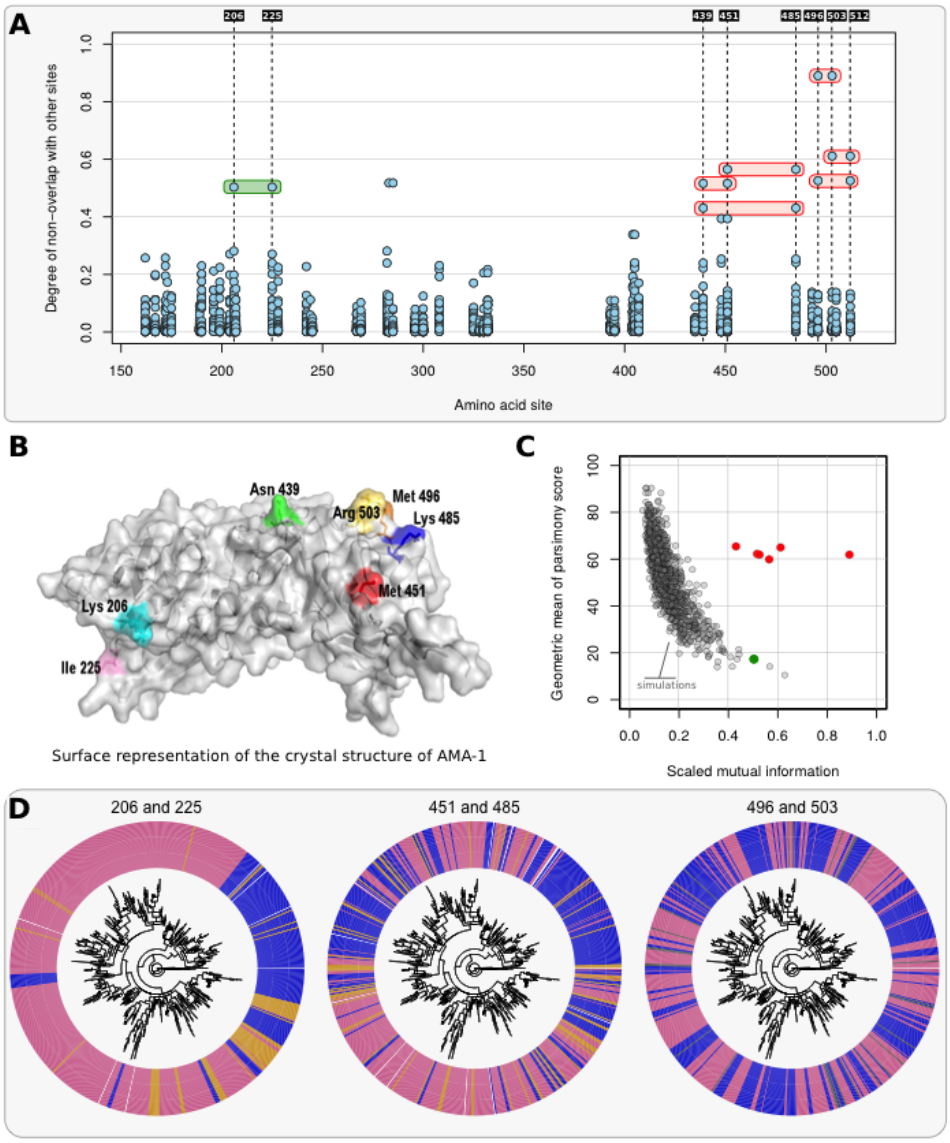

Figure 1: (A) Degree of non-overlap in associations between amino acid residues at pairs of dimorphic sites. Each site on the x-axis is compared with all other dimorphic sites: those with reciprocally high scores are shown by bars spanning both sites. Red bars show significant associations. Green bars represent associations that do not achieve significance when phylogenetic relationships are taken into account (B) Surface representation of the crystal structure of AMA-1 with specific amino acids highlighted [Lysine (Lys) 206 = Cyan; Isoleucine (Ile) 225 = Magenta; Asparagine (Asn) 439 = Green; Methionine (Met) 451 = Red, Lys 485 = Blue, Met 496 = Orange, Arginine (Arg) 503 = Yellow]. (C) Pairs of sites shown by red and green bars in Fig1A are shown in relation to a null distribution of site-pairs from 1000 simulations that scored most highly on the product of scaled Mutual Information (measuring non-overlap) and Geometric Mean of Parsimony Scores (indicating the minimum number of mutations required to explain the observed ancestral relationships). (D) Phylogenetic tree of AMA-1 with outside ring coloured on the basis of observed combination of amino acid variants found at tree tips for (A) sites 206 and 225 (pink = NE, blue = IK, orange = IE), (B) sites 451 and 485 (pink = IK, blue = KM, orange = KK), and (C) sites 496 and 503 (pink = NI, blue = RM, grey = rare variants).

bioRxiv preprint doi: https://doi.org/10.1101/159483; this version posted July 20, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

The group of sites which scored most highly on the conservative test (described above) comprised 496, 503 and 512. However, mapping these onto the tertiary structure of AMA-1 (Fig 1B) shows that they are in very close proximity of each other and therefore we were unable to exclude biochemical interactions as the primary cause of this striking association. It is still possible that variation at these sites is held in strong non-overlap through immune selection. Antibodies specific to epitopes containing amino acids I/N/K may cross-react with all other epitopic variants except those containing M/R/R, thereby promoting the co-circulation of these non-overlapping combinations of residues at these sites; however, because of their structural adjacency, it is impossible to privilege this explanation over intrinsic steric advantages to these combinations

High non-overlap was also observed among sites 439/451, 439/485 and 451/485 which were sufficiently far enough apart in the tertiary structure to represent potential examples of immune selection forcing non-structurally related residues to co-segregate. We cannot, of course, completely eliminate long-range physical interactions as a cause of these associations, and whether it is indeed immune selection that is responsible for these patterns can only be established by functional studies. It is clear that there is no fundamental structural impediment to the existence of intermediate forms since they do occur at low frequencies, but whether they are disadvantaged due to lower intrinsic fitness or immune-mediated selection remains to be resolved by further experimental work.

We next investigated the associations between positions 439, 451 and 485 on a smaller geographical scale to confirm that these patterns did not arise due to certain combinations of alleles dominating by chance in particular areas. In all countries

bioRxiv preprint doi: https://doi.org/10.1101/159483; this version posted July 20, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

studied, both of the major sequence variant associations were present, with the secondary variants present at a lower frequency (Table 2).

A dataset of 660 unique sequences was compiled for Mali and showed high non-overlap for each pair of associations. A smaller dataset comprising 79 sequences from the Gambia also yielded MI values in excess of those obtained from the global data set. Slightly weaker associations were found in 86 unique sequences from Kenya; however these improved significantly when the data were analysed as a collection of isolates rather than sequences: i.e. without excluding identical sequences present in different individuals (Table 3).

As the model predictions pertain to population level frequencies of allele combinations, rather than their presence in a collection of unique sequences, we believe that Table 3 presents a more valid test (than when only unique sequences are included) of the model's capabilities to pick out regions under strong immune selection; it is reassuring that these patterns are also evident under the more stringent conditions where only unique sequences are included.

Our results suggest that Domain III (DIII) of the AMA-1 protein is under strong immune selection, and may play an important role in protective immunity alongside the more polymorphic C1L region of Domain I. At a functional level, Domain III has been shown to bind to the erythrocyte membrane protein Kx [33]. Immune responses that interrupt this process may, therefore, have a significant inhibitory effect on parasite growth but, so far, the evidence that Domain III elicits protective antibody responses is limited. Nair *et al.* [34] have shown that antibodies from a single blood donor affinity purified on DIII are inhibitory. In a study by Polley *et al.* [2], it was found that antibody responses against recombinant full-length ectodomains were more

bioRxiv preprint doi: https://doi.org/10.1101/159483; this version posted July 20, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

strongly associated with protection than responses to those which did not express DIII, even though naturally acquired antibodies to recombinant DIII alone were rare. A virosomal formulation of loop I of DIII, comprising residues 446-490, has been shown to elicit growth-inhibitory antibodies [35] and volunteers within a vaccine trial containing this AMA1 peptide exhibited both reduced blood stage growth rates and a higher rate of morphological parasite abnormalities [36]. A recent analysis of allele-specific efficacy of a vaccine against AMA1 [11] showed that the relative risk ratio (RRR) of re-infection with strains containing vaccine-type amino acids among vaccinees vs controls was high (although not significant) at residues 439, 451 and 485. Residues associated with the polymorphic sites surrounding the hydrophobic trough displayed significant RRR, as might be expected; sites 308 and 332 also had high scores but no evidence of sustained immune selection at these sites was available in our analysis. In the same study, the likelihood of a shift from a vaccine-type to non-vaccine-type residue among sequences obtained following vaccination was found to be significantly increased at position 485 and non-significantly increased at position 451. Significant changes in residue composition were also found at highly polymorphic site 197 but not at any of the other polymorpic sites in the C1L region. Significant change was also observed at dimorphic site 435, but we found no non-overlapping associations in any of the pairwise comparisons between 435 and other sites. Several population genetic studies [37-41] based on site frequency spectrum analysis techniques (such as Tajima's D) strongly indicate that DIII is under balancing selection; our "reverse immunodynamic" method complements and extends these analyses by showing that we can use a measure of non-overlap between site-pairs in addition to standard population genetic measures to finely map regions under immune selection within DIII.

bioRxiv preprint doi: https://doi.org/10.1101/159483; this version posted July 20, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

An interesting feature of our analyses is that while particular groups of sites exhibit strong non-overlap, the associations between these groups appear to be random. For example, the triad 439-451-485 is in complete linkage equilibrium with triad 496-503-512. Furthermore, there are no non-overlapping associations present between the highly polymorphic residues in the C1L region of Domain I and the 439-451-485 triad (although the former exhibit strong non-overlap: Table S1). If the analysis we present here is correct, then it would be sufficient to present the dominant "non-overlapping" variants of 439-451-485 (i.e. HKI and NMK) in a manner that induced strong immune responses to the epitopes containing these residues, rather than to try and cover the entire haplotypic diversity of AMA-1 [42]. Immune responses to the epitopes containing residues 439, 451 and 485 appear to develop more slowly under natural conditions than antibodies to the highly polymorphic regions shielding the hydrophobic groove [2], but should be easier to induce by vaccination than immune responses to completely conserved epitopes [43].

Our analysis of AMA-1 highlights the role of less polymorphic epitopes in population-level immune evasion. At a more general level, our analyses emphasize that targets of natural immunity are not polarised between the completely invariant and the highly polymorphic but also include a range of epitopes of intermediate variability that have, so far, been largely overlooked. Among pathogens where protective immune responses are directed against conserved targets or epitopes, a single infection ensures lifelong immunity and the goal of vaccination is to elicit natural immunity against such conserved targets. Most infectious diseases (such as measles) against which we have effective vaccines fall into this category. By contrast,

bioRxiv preprint doi: https://doi.org/10.1101/159483; this version posted July 20, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

the primary targets of protective immunity in several important pathogens such as *P. falciparum*, are highly diverse and present a challenge to the development of vaccines. One solution to this problem that is currently being explored by several research groups is the artificial boosting of immunity towards conserved regions that are not sufficiently naturally immunogenic to prevent further infection upon single exposure but do eventually confer protection after multiple exposures. Another strategy is to develop multi-allelic vaccines in which immunogenic epitopes that confer protection against a sub-population of strains are combined to provide broad coverage. However, many of the known targets of protective immunity are too diverse for this strategy to be successfully implemented. Our work in this paper highlights that between these two extremes, there exist epitopes of intermediate variability which are nonetheless potentially under strong immune selection and thus offer alternative targets of vaccination.

Methods

SEQUENCE SELECTION: *Pf*AMA-1 sequences were obtained by searching the NCBI Entrez nucleotide and protein databases. On 04/07/2014 this yielded 2835 protein sequences and 2926 nucleotide sequences. In cases when the GenBank nucleotide record did not include the corresponding protein sequence, these were translated using Biopython, an open source set of tools (biopython.org).

The analysis was restricted to sequences between 400 and 622 amino acids in length. The lower limit aims to ensure that all the polymorphic regions across the three domains are included; the upper limit corresponds to the entire processed protein, and therefore removes any entries which contain the non-coding sequence. To avoid

bioRxiv preprint doi: https://doi.org/10.1101/159483; this version posted July 20, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

pseudoduplication, all identical sequences were removed, leaving 1198 unique sequences. These were then aligned using Clustal W [40].

MEASURING NON-OVERLAP AMONG DIMORPHIC SITES: A frequency analysis of all residues within the three domains (q.v.) was used to identify any polymorphic sites. Only variants with a frequency exceeding 2.5% were included in the subsequent analysis. This identified 44 dimorphic sites (Fig S1). We searched for signatures of immune selection by analysing the mutual information (MI) scores of all pairs of sites (Table S1).

GEOGRAPHIC SUB-ANALYSES: We repeated the calculation of MI for all isolates whose country of origin could be identified (either from the Entrez database or with reference to published papers).

PHYLOGENETIC ANALYSIS: We first truncated our protein alignment to the region where we had >99% coverage [namely, from residue 149 to residue 534]. This alignment was then used to construct a Maximum Likelihood Tree in Mega6 under the JTT model (with also: 4 discrete categories of gamma-distributed rates among sites; the nearest-neighbour-interchange heuristic; and a very strong branch swap filter) and this tree was used to calculate the maximum parsimony score (PS) for each dimorphic site (Fig S2; Table S1). We defined a new statistic S as the product of the MI and the geometric mean of the PS for each site (Table S2). We used Seq-Gen to simulate 1000 sets of protein sequences with the same phylogenetic characteristics, restricting variation at each of the sites to two amino acids. We were able to establish

bioRxiv preprint doi: https://doi.org/10.1101/159483; this version posted July 20, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

that the pairs of interest (see Table 1) lay significantly outside the distribution of the identically ranked (by S-score) sites from each simulation (Fig S3).

IMAGES: From the protein data bank (PDB) the sequences and tertiary structures of *Plasmodium vivax* AMA-1 (*Pv*AMA-1, PDB 1W81) and *Pf*AMA-1 domain III (PDB 1HN6) were downloaded. As the tertiary structure of 1HN6 was mostly unstructured, it could not be used on its own. Therefore a Fold and Function Assignment System (FFAS) was used to align the amino acid sequences of 1W81 and 1HN6. From this alignment a 'hybrid' structural model with 94% confidence was generated using SCWRL. This resulting model gave a single tertiary structure with the residues within domain III from *Pv*AMA-1 (1W81) substituted by those from *Pf*AMA-1 (1HN6) using the *P. vivax* structural coordinates. Pymol (http://www.pymol.org/) was used to visualise the 'hybrid' tertiary structure and to highlight the relevant residues.

Supplementary files

Table S1: parsimony score and MI for pairs of sites
Table S2: statistic S (product of MI and the geometric mean of the PS for each site)

Acknowledgements

We thank Dr. CRC Spensley for assistance with data manipulation and programming, and Dr. Pietro Roversi for assistance with generating the structure.

bioRxiv preprint doi: https://doi.org/10.1101/159483; this version posted July 20, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

References

1. Remarque EJ, Faber BW, Kocken CHM & Thomas AW (2008) Apical membrane antigen 1: A malaria vaccine candidate in review. *Trends Parasitol* 24(2): 74-84.
2. Polley SD, et al (2004) Human antibodies to recombinant protein constructs of plasmodium falciparum apical membrane antigen 1 (AMA1) and their associations with protection from malaria. *Vaccine* 23(5): 718-728.
3. Fowkes FJI, Richards JS, Simpson JA & Beeson JG (2010) The relationship between anti-merozoite antibodies and incidence of plasmodium falciparum malaria: A systematic review and meta-analysis. *Plos Medicine* 7(1): e1000218.
4. Richards JS, et al (2013) Identification and prioritization of merozoite antigens as targets of protective human immunity to plasmodium falciparum malaria for vaccine and biomarker development. *Journal of Immunology* 191(2): 795-809.
5. Mugyenyi CK, et al (2013) Antibodies to polymorphic invasion-inhibitory and non-inhibitory epitopes of plasmodium falciparum apical membrane antigen 1 in human malaria. *Plos One* 8(7): e68304.
6. Stanisic DI, et al (2015) Acquisition of antibodies against plasmodium falciparum merozoites and malaria immunity in young children and the influence of age, force of infection, and magnitude of response. *Infect Immun* 83(2): 646-660.
7. Osier FH, et al (2014) New antigens for a multicomponent blood-stage malaria vaccine. *Science Translational Medicine* 6(247)
8. Vicentin EC, et al (2014) Invasion-inhibitory antibodies elicited by immunization with plasmodium vivax apical membrane antigen-1 expressed in pichia pastoris yeast. *Infect Immun* 82(3): 1296-1307.
9. Biswas S, et al (2012) Recombinant viral-vectored vaccines expressing plasmodium chabaudi AS apical membrane antigen 1: Mechanisms of vaccine-induced blood-stage protection. *Journal of Immunology* 188(10): 5041-5053.
10. Schussek S, et al (2013) Immunization with apical membrane antigen 1 confers sterile infection-blocking immunity against plasmodium sporozoite challenge in a rodent model. *Infect Immun* 81(10): 3586-3599.
11. Ouattara A, et al (2013) Molecular basis of allele-specific efficacy of a blood-stage malaria vaccine: Vaccine development implications. *J Infect Dis* 207(3): 511-519.
12. Drew DR, et al (2012) Defining the antigenic diversity of plasmodium falciparum apical membrane antigen 1 and the requirements for a multi-allele vaccine against malaria. *Plos One* 7(12): e51023.
13. Dicko A, et al (2007) Impact of a plasmodium falciparum AMA1 vaccine on antibody responses in adult malians. *Plos One* 2(10): e1045.
14. Thera MA, et al (2008) Safety and immunogenicity of an AMA-1 malaria vaccine in malian adults: Results of a phase 1 randomized controlled trial. *Plos One* 3(1): e1465.
15. Laurens MB, et al (2013) Extended safety, immunogenicity and efficacy of a blood-stage malaria vaccine in malian children: 24-month follow-up of a randomized, double-blinded phase 2 trial. *Plos One* 8(11): e79323
16. Biswas S, et al (2014) Assessment of humoral immune responses to blood-stage malaria antigens following ChAd63-MVA immunization, controlled human malaria infection and natural exposure. *Plos One* 9(9): e107903.
17. Bai T, et al (2005) Structure of AMA1 from plasmodium falciparum reveals a clustering of polymorphisms that surround a conserved hydrophobic pocket. *Proc Natl Acad Sci U S A* 102(36): 12736-12741.
18. Lim SS, et al (2014) Structure and dynamics of apical membrane antigen 1 from plasmodium falciparum FVO. Biochemistry (N Y) 53(46): 7310-7320.
19. Richard D, et al (2010) Interaction between plasmodium falciparum apical membrane antigen 1 and the rhoptry neck protein complex defines a key step in the erythrocyte invasion process of malaria parasites. *J Biol Chem* 285(19): 14815-14822.

bioRxiv preprint doi: https://doi.org/10.1101/159483; this version posted July 20, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

20. Lamarque M, et al (2011) The RON2-AMA1 interaction is a critical step in moving junction-dependent invasion by apicomplexan parasites. *Plos Pathogens* 7(2): e1001276.
21. Srinivasan P, et al (2011) Binding of plasmodium merozoite proteins RON2 and AMA1 triggers commitment to invasion. *Proc Natl Acad Sci U S A* 108(32): 13275-13280.
22. Healer J, et al (2004) Allelic polymorphisms in apical membrane antigen-1 are responsible for evasion of antibody-mediated inhibition in plasmodium falciparum. *Mol Microbiol* 52(1): 159-168.
23. Cortes A, et al (2005) Allele specificity of naturally acquired antibody responses against plasmodium falciparum apical membrane antigen 1. *Infect Immun* 73(1): 422-430.
24. Gupta S, et al (1996) The maintenance of strain structure in populations of recombining infectious agents. *Nat Med* 2(4): 437-442.
25. Gupta S, Ferguson N & Anderson R (1998) Chaos, persistence, and evolution of strain structure in antigenically diverse infectious agents. *Science* 280(5365): 912-915.
26. Lourenço J, Wikramaratna SP, Gupta S (2015) MANTIS: an R package that simulates multilocus models of pathogen evolution. *BMC Bioinformatics*. 16:176.
27. Hinnebusch BJ, Barbour AG, Restrepo BI, Schwan TG. (1998) Population structure of the relapsing fever spirochete *Borrelia hermsii* as indicated by polymorphism of two multigene families that encode immunogenic outer surface lipoproteins. *Infect Immun* 66:432–40.
28. Johnson DR, Kaplan EL, VanGheem A, Facklam RR, Beall B (2006) Characterization of group A streptococci (Streptococcus pyogenes): Correlation of M-protein and emm-gene type with T-protein agglutination pattern and serum opacity factor. J Med Microbiol 55:157e164.
29. Areschoug, T., Stålhammar-Carlemalm, M., Larsson, C., & Lindahl, G. (1999). Group B Streptococcal Surface Proteins as Targets for Protective Antibodies: Identification of Two Novel Proteins in Strains of Serotype V. *Infection and Immunity*, *67*(12), 6350–6357.
30. Buckee CO, Gupta S, Kriz P, Maiden MC & Jolley KA (2010) Long-term evolution of antigen repertoires among carried meningococci. *Proc Biol Sci* 277:1635-41
31. Callaghan MJ, Buckee C, Jolley, KA, Kriz, P, Maiden MCJ , Gupta S. (2008) The Effect of Immune Selection on the Structure of the Meningococcal Opa Protein Repertoire. *PLoS Pathogens* 14;4(3):e1000020.
32. Polley SD, et al. (2007) Plasmodium falciparum merozoite surface protein 3 is a target of allele-specific immunity and alleles are maintained by natural selection. J Infect Dis 195:279e287.
33. Kato K, Mayer DCG, Singh S, Reid M & Miller LH (2005) Domain III of plasmodium falciparum apical membrane antigen 1 binds to the erythrocyte membrane protein kx. *Proc Natl Acad Sci U S A* 102(15): 5552-5557.
34. Nair M, et al (2002) Structure of domain III of the blood-stage malaria vaccine candidate, plasmodium falciparum apical membrane antigen 1 (AMA1). *J Mol Biol* 322(4): 741-753.
35. Mueller MS, et al (2003) Induction of parasite growth-inhibitory antibodies by a virosomal formulation of a peptidomimetic of loop I from domain III of plasmodium falciparum apical membrane antigen 1. *Infect Immun* 71(8): 4749-4758.
36. Thompson FM, et al (2008) Evidence of blood stage efficacy with a virosomal malaria vaccine in a phase IIa clinical trial. *Plos One* 3(1): e1493.
37. Polley SD, Conway DJ (2001) Strong diversifying selection on domains of the *Plasmodium falciparum* apical membrane antigen 1 gene. Genetics 158: 1505–1512.
38. Tetteh KK, Stewart LB, Ochola LI, Amambua-Ngwa A, Thomas AW, et al. (2009) Prospective identification of malaria parasite genes under balancing selection. PLoS ONE 4: e5568 doi:10.1371/journal.pone.0005568. doi: 10.1371/journal.pone.0005568
39. Osier FH, Weedall GD, Verra F, Murungi L, Tetteh KK, et al. (2010) Allelic diversity and naturally acquired allele-specific antibody responses to *Plasmodium falciparum* apical membrane antigen 1 in Kenya. Infect Immun 78: 4625–4633. doi: 10.1128/iai.00576-10
40. Amambua-Ngwa A, Tetteh KKA, Manske M, Gomez-Escobar N, Stewart LB, Deerhake ME, et al. (2012) Population Genomic Scan for Candidate Signatures of Balancing

bioRxiv preprint doi: https://doi.org/10.1101/159483; this version posted July 20, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Selection to Guide Antigen Characterization in Malaria Parasites. PLoS Genet 8(11): e1002992. doi:10.1371/journal.pgen.1002992

41. Arnott A, et al (2014) Distinct patterns of diversity, population structure and evolution in the AMA1 genes of sympatric plasmodium falciparum and plasmodium vivax populations of Papua New Guinea from an area of similarly high transmission. *Malaria Journal* 13: 233.
42. Duan J, et al (2008) Population structure of the genes encoding the polymorphic plasmodium falciparum apical membrane antigen 1: Implications for vaccine design. *Proc Natl Acad Sci U S A* 105(22): 7857-7862.
43. Dutta S, et al (2013) Overcoming antigenic diversity by enhancing the immunogenicity of conserved epitopes on the malaria vaccine candidate apical membrane antigen-1. *Plos Pathogens* 9(12): e1003840.
44. Larkin MA, et al (2007) Clustal W and clustal X version 2.0. *Bioinformatics* 23(21): 2947-2948.