

# SCIENTIFIC REPORTS



OPEN

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

Katrina J. Spensley<sup>1</sup>, Paul S. Wikramaratna<sup>2</sup>, Bridget S. Penman<sup>5</sup>, Andrew Walker<sup>3</sup>, Adrian L. Smith<sup>3</sup>, Oliver G. Pybus<sup>3</sup>, Létitia Jean<sup>4</sup>, Sunetra Gupta<sup>3</sup> & José Lourenço<sup>3</sup> 

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. Epidemiological models predict that strong immune selection can cause antigenic variants to structure into genetically discordant sets of antigenic types (e.g. serotypes). A corollary of this theory is that targets of immunity may be identified by searching for non-overlapping associations of amino acids among co-circulating antigenic variants. We propose a novel population genetics methodology that combines such predictions with phylogenetic analyses to identify genetic loci (epitopes) under strong immune selection. We apply this concept to the AMA-1 protein of the malaria parasite *Plasmodium falciparum* and find evidence of epitopes among certain regions of low variability which could render them ideal vaccine candidates. The proposed method can be applied to a myriad of multi-strain pathogens for which vast amounts of genetic data has been collected in recent years.

Among pathogens for which protective immune responses are directed against conserved targets (epitopes), a single infection ensures lifelong immunity, and the goal of vaccination is to elicit natural-like immunity against such conserved targets. Most infectious diseases against which we have effective vaccines fall into this category (e.g. measles). By contrast, the primary targets of protective immunity in several important pathogens such as *Plasmodium falciparum*, the human immunodeficiency virus (HIV) or the influenza A virus, 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<sup>1,2</sup>. Another solution is to target naturally immunogenic epitopes of very high variability (e.g. influenza A), with the drawback that vaccines may need recurrent updating. As opposed to targeting conserved epitopes of low immunogenicity or highly variable epitopes of high immunogenicity, an alternative and less researched strategy is to target epitopes of limited variability which are under immune selection. This strategy may allow to develop multi-allelic vaccines in which immunogenic epitopes that confer protection against a sub-population of strains are combined to provide broad coverage (e.g. Thompson *et al.*<sup>3</sup>).

In this study, we attempt to identify low variability epitopes under strong immune selection by using a novel approach based on the predictions of a multi-locus mathematical framework of pathogen evolution commonly referred to as *strain theory*<sup>4-6</sup>. The central premise we focus on, is that regions (or epitopes) under strong immune pressure are selected to self-organise into genetically discordant sets of antigenic types (e.g. dengue virus, meningococcal or pneumococcal serotypes).

This prediction is so far supported by empirical observations within a variety of host-pathogen systems<sup>7-13</sup>. In our proposed approach, we invert this prediction to search for signatures of immune selection by identifying genetic regions which exhibit a high degree of non-overlap, but critically excluding those that may have arisen as a consequence of shared ancestry or due to structural interactions. We term this novel approach the *reverse immunodynamic* method and present proof of concept results from a vast number of genetic sequences of the apical membrane antigen 1 protein (AMA-1).

<sup>1</sup>Imperial College London, London, W2 1PG, UK. <sup>2</sup>Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, EH9 3JT, UK. <sup>3</sup>Department of Zoology, University of Oxford, Oxford, OX1 3PS, UK. <sup>4</sup>The Sir William Dunn School of Pathology, University of Oxford, Oxford, OX1 3RE, UK. <sup>5</sup>School of Life Sciences, University of Warwick, Coventry, CV4 7AL, UK. Katrina J. Spensley and Paul S. Wikramaratna contributed equally. Correspondence and requests for materials should be addressed to J.L. (email: [jose.lourenco@zoo.ox.ac.uk](mailto:jose.lourenco@zoo.ox.ac.uk))

Received: 2 March 2018

Accepted: 20 November 2018

Published online: 15 February 2019

Association	Amino Acid Combinations				Scaled MI
	IN	IR	MN	MR	
496 vs 503	602	3	14	561	0.89
503 vs 512	565	52	40	524	0.61
451 vs 485	604	103	15	476	0.57
496 vs 512	552	68	54	523	0.53
283 vs 285	242	71	1	883	0.52
439 vs 451	660	72	46	390	0.5
206 vs 225	96	806	290	6	0.5
439 vs 485	588	144	30	406	0.41
448 vs 451	705	196	2	295	0.39
404 vs 405	108	635	367	84	0.34

**Table 1.** Counts of amino acid associations between pairs of sites with high non-overlap (scaled MI > 0.312).

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*<sup>14–20</sup>, *P. knowlesi*<sup>21</sup> and *P. vivax*<sup>22</sup>. Cellular responses to AMA-1 may provide immunity against both erythrocytic<sup>23</sup> and pre-erythrocytic forms of *P. falciparum*<sup>24</sup>. 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<sup>25,26</sup> as well as by difficulties in inducing lasting immune responses through vaccination<sup>27–30</sup>. 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 DIII surrounding a conserved hydrophobic trough<sup>31–36</sup> which plays a vital role in the invasion process<sup>37–39</sup>. 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<sup>26,40,41</sup>. Identifying low variability 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.

Previous studies on AMA-1 genetic data have demonstrated that selection pressure as measured by standard population genetics measures is strongest within DI and DIII<sup>34–36</sup>, and that while these are also highly variable, they contain a number of less variable residues<sup>35,36</sup>. Using our proposed *reverse immunodynamics* method we expand on these studies by detecting strong signatures of non-overlap among a small number of dimorphic residues in DIII, thus identifying epitopes within that wide region which are targets of protective immunity.

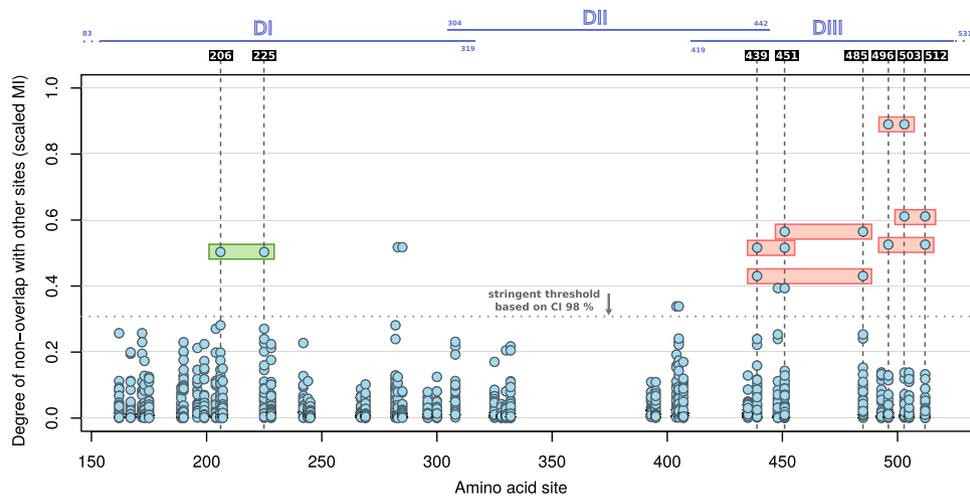
## Results

We implemented the reverse immunodynamic method by performing pairwise comparisons of all sites among 1198 unique PfAMA-1 sequences obtained from GenBank, to determine whether any of these pairs exhibited non-overlapping associations between amino acid residues (see Methods section for details). We used an Information Theory statistic known as Mutual Information (MI), scaled to its maximum value (scaled MI), to measure the degree of non-overlap in amino acid combinations between pairs of sites of PfAMA-1 sequences (see Supplementary Text 1 for mathematical formulations and theoretical background). The main results are focused on dimorphic sites as we aim at the discovery of low variability epitopes, but output of our approach for all polymorphic pairs of sites is made available in Table S1.

**A limited number of site-pairs exhibit high non-overlap.** Most pairs of sites showed low non-overlap, with only 10 pairs exhibiting a scaled MI in excess of 0.312 (Table 1), a threshold based on the top 1% of the distribution of all observed scaled MI values (upper limit of a stringent 98% confidence interval, CI) (Table S2). Of these pairs, 404/405, 448/451 and 283/285 were eliminated on grounds of sequence proximity based on the lower limit ( $d = 3$ ) 95% CI of all observed distances between dimorphic pairs. Figure 1A highlights the remaining 7 pairs among all the background scaled MI measures.

**For most pairs, high non-overlap is unlikely to have arisen by neutral evolutionary processes.**

To assess whether high non-overlap between sites could have arisen by neutral evolutionary processes (e.g. ancestry), we started by estimating a maximum likelihood (ML) phylogenetic tree from the sequences (see Methods for details). A major difference became evident upon inspection of the phylogenetic allelic structure of site-pair



**Figure 1.** Degree of non-overlap between amino acid residues at pairs of dimorphic sites. Each site on the x-axis (positions 149–534) that is dimorphic is compared with all other dimorphic sites. Those with high scores above a stringent scaled MI threshold are shown by bars spanning both sites (red and green). Scaled MI threshold is based on the top 1% of observed scaled MI values (CI 98% of the observed distribution, upper limit of 0.312, horizontal dotted grey line). Red bars mark pairs of sites of interest, and the green bar marks the pair that is not of interest once phylogenetic relationships are taken into account.

206/225 compared to the other pairs of interest (Fig. 2A). As is visually shown, the dominant non-overlapping pairs of amino acids at 206/225 are highly segregated and present an ancient dimorphic relationship, which has been observed in other *P. falciparum* genes such as DBLMSP/DBLMSP<sup>33</sup>, MSP1<sup>42</sup>, or MSP2<sup>43</sup>. The other pairs of sites exhibited a more interdigitated pattern, indicating multiple independent changes.

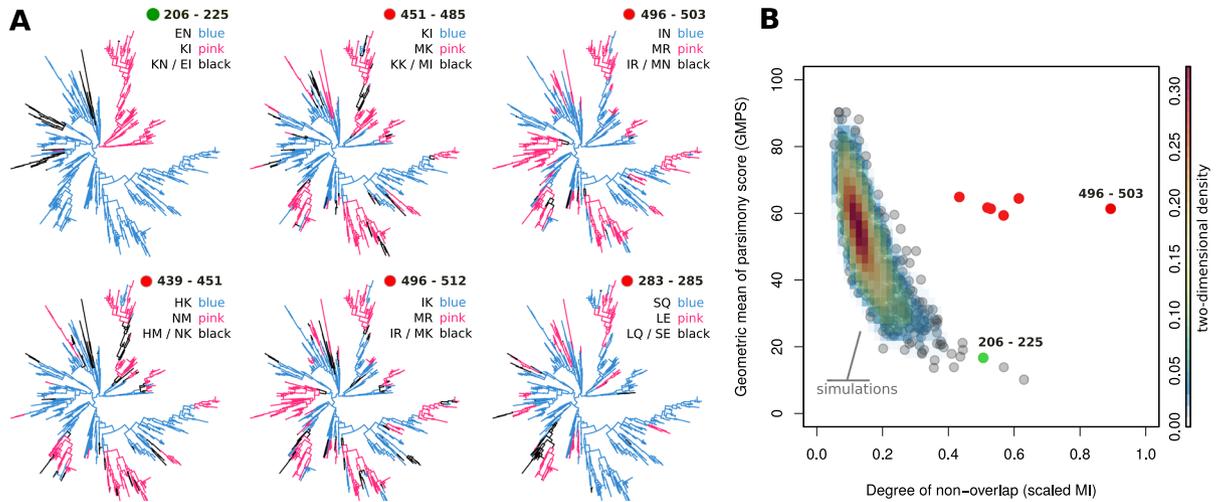
To try to quantify these differences, we calculated the parsimony score (PS) for each site (see Supplementary Text 1 for details), which is the minimum number of genetic changes required to explain the observed ancestral relationships (in this case, back and forth between the two possible amino acids at each site). We then calculated the product of scaled MI and the geometric mean of PS (GMPS) for each site, which we termed the S-score. Importantly, this new measure was found to preserve the same order as the scaled MI scores of the 7 pairs of interest (Table 1), except in the case of the 206/225 (values in Table S2). This meant that six out of seven pairs of sites had both intermediate-to-high scaled MI and GMPS values.

To obtain a null distribution of the S-score for hypothesis testing, we simulated 1000 sets of sequences which were (a) consistent with the ancestry of the empirical phylogenetic tree, and (b) had the same PS score at each dimorphic site as the empirical data (see Methods). Given that six out of seven pairs of sites of interest had the highest empirical S-scores, we selected, from each of the 1000 simulations, the highest S-score pair of sites (termed rank 1 pairs, values in Table S4). The two-dimensional distribution of rank 1 pairs of sites (Fig. 2B, grey points) showed that neutral evolution can generate sites with high PS or high scaled MI values, but is unlikely to generate pairs of sites with both measures being high. When the site pairs of interest (Table 1) were overlaid on this null distribution, they could be seen to present both high scaled MI and PS values (red points) as expected from their high S-scores, except for pair 206/225 (green point).

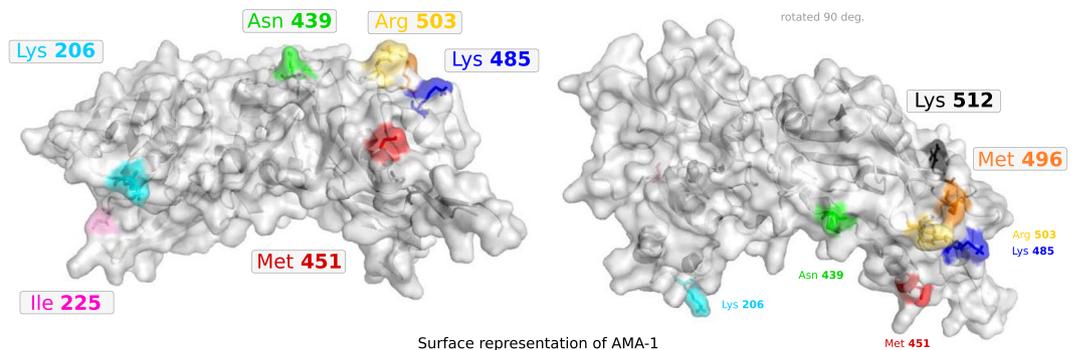
From both the phylogenetic allelic structures (Fig. 2A) and the S-score rank 1 distribution (Fig. 2B), we concluded that six out of the seven site-pairs of interest were unlikely to have arisen by neutral evolutionary processes (496/503, 503/512, 451/485, 496/512, 439/451, 439/485). Even when looking at the distributions of rank 1 to 6 (Table S4), these pairs were found significantly outside their own simulated distributions (Fig. S8). In contrast, we were unable to reach the same conclusion for site-pair 206/225.

**Physical proximity may explain some of the non-overlap.** Some of the pairs of sites which scored most highly on S-score and scaled MI comprised 496, 503 and 512. However, mapping these onto the tertiary structure of AMA-1 (Fig. 3) 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.

Other highly scored pairs included sites 439, 451, and 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 into discordant sets. 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. There should be no fundamental structural impediment to the existence



**Figure 2.** Empirical phylogenetic allelic structure, simulated parsimony score and scaled MI. **(A)** Example, ML phylogenetic trees of AMA-1 with branches coloured on the basis of observed combination of amino acid variants found at tips for different pairs of sites. Pink and cyan are used for the two most common amino acid variants (which are discordant for these sites, see Table 1) and black for the least common. **(B)** Empirical pairs of sites described in Fig. 1 and Table 1 are shown in relation to a null distribution of site-pairs from 1000 simulations that scored most highly (rank 1) on S-score (the product of scaled MI measuring non-overlap, and geometric mean of parsimony score indicating the minimum number of genetic changes required to explain the observed ancestral relationships, GMPS). Color key (scale) on the right is the two-dimensional density of GMPS and scaled MI of the simulated site-pairs. **(A,B)** Pairs of sites identified in red and green in the two panels match the pairs in Fig. 1.



**Figure 3.** Surface representation of the crystal structure of AMA-1 with specific sites highlighted. The highlighted amino acids are the same in both panels and the ones relevant and described in Table 1 and Figs 1 and 2. The right panel is a 90 degree rotation of the left panel, in which Lys 512 and Met 496 are revealed. Amino acids are the consensus of all sequences.

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.

**Non-overlapping associations are also observed at the level of local populations.** We also investigated the associations between sites 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 studied, both of the major sequence variant associations were present, with the secondary variants present at a lower frequency (Table 2). In particular, the dataset of 660 unique sequences from Mali showed high non-overlap for each pair. A smaller dataset comprising 79 sequences from the Gambia also yielded scaled 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 were seen to improve 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.

Association	Country and Amino Acid Combinations					Scaled MI
		HK	HM	NK	NM	
439 vs 451	<i>Gambia</i>	46	3	0	31	0.701
	<i>Kenya</i>	46	3	9	27	0.402
	<i>Mali</i>	395	25	11	211	0.637
		HI	HK	NI	NK	
439 vs 485	<i>Gambia</i>	39	7	1	30	0.549
	<i>Kenya</i>	45	4	1	35	0.682
	<i>Mali</i>	359	61	17	205	0.453
451 vs 485		KI	KK	MI	MK	
	<i>Gambia</i>	39	4	1	35	0.673
	<i>Kenya</i>	46	9	0	31	0.515
	<i>Mali</i>	366	40	10	244	0.608

**Table 2.** Country-specific counts of amino acid associations between positions 439, 451 and 485 within unique sequences obtained from The Gambia, Kenya and Mali.

Association	Amino Acid Combinations				Scaled MI
	HK	HM	NK	NM	
439 vs 451	61	3	9	55	0.564
439 vs 485	60	4	1	63	0.772
451 vs 485	61	9	0	59	0.644

**Table 3.** Counts of amino acid associations between positions 439, 451 and 485 in a population study (duplicate sequences not excluded) in Kilifi, Kenya.

It is nonetheless reassuring that these patterns are also evident under the more stringent conditions where only unique sequences are included.

## Discussion

In this study, we have set out to demonstrate reverse immunodynamics, a new population genetics approach that is able to identify targets of immune selection under the theoretical expectations of a well established, multi-locus framework of pathogen evolution. Our results suggest that DIII of AMA-1 contains several epitopes of lower variability that are under strong immune selection. These epitopes may play an important role in protective immunity alongside more polymorphic epitopes in the C1L region of DI which have so far compromised the development of an effective vaccine.

The identification of possible epitopes in DIII follows the conclusions of several studies based on standard population genetics (sPG) techniques that strongly indicate that DIII is under selection<sup>34–36,44–47</sup>. However, sPG such as Tajima's D are generally applied to windows of amino acids, and have critically not been devised for, and do not have the theoretical ground to identify immunity as the source of selection<sup>48,49</sup>. Reverse immunodynamics complements sPG techniques by finely mapping loci, which are tentatively under immune pressure within regions known to be under selection (through sPG analyses, as is the case of DIII).

Much work has also been done previously on AMA-1 site-pairs using measures of linkage disequilibrium (LD) such as D or R-squared (e.g.<sup>34–36</sup>), demonstrating that the alleles of many pairs are less mixed than expected by random chance (even in the presence of high recombination rates). Typically, such signatures can be used to infer that selection pressure has structured alleles into non-random associations. While it is tempting to assume that site-pairs high on LD are also high on MI, this is not necessarily true. The key difference is that MI is both a measure of linkage and non-overlap, while LD ignores the latter (for empirical and theoretical demonstration see Supplementary Text 1). We set on to look for signatures of non-overlap as these are critically predicted to be driven by strong immune selection, and therefore LD was not suited for the analyses. Finally, AMA-1 is known to present high recombination potential<sup>34–36</sup>, giving rise to a wide range of haplotypes<sup>35</sup>. According to complex transmission models based on *strain theory*, non-overlap is still a self-emergent signature among antigenic loci even in the presence of high recombination rates<sup>50</sup>.

So far, the evidence that DIII elicits protective antibody responses is limited, but much research supports this possibility. For instance, at a functional level, DIII has been shown to bind to the erythrocyte membrane protein Kx<sup>51</sup>, and immune responses that interrupt this process may, therefore, have a significant inhibitory effect on parasite growth. Nair *et al.*<sup>52</sup> have also shown that antibodies from a single blood donor affinity purified on DIII are inhibitory. In a study by Polley *et al.*<sup>15</sup>, it was found that antibody responses against recombinant full-length ectodomains were more 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 also been shown to elicit growth-inhibitory antibodies<sup>53</sup> and volunteers within a vaccine trial containing this AMA-1 peptide exhibited both reduced blood stage growth rates and a higher rate of morphological parasite abnormalities<sup>54</sup>. A recent analysis of allele-specific efficacy of a vaccine against AMA-1<sup>25</sup> 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. Interestingly, in the same study, the likelihood of a shift from a vaccine- 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.

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 CIL region of Domain I and the 439–451–485 triad (although the former exhibit strong non-overlap, Table S1). If our analyses are 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<sup>55</sup>. 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<sup>15</sup> but should be easier to induce by vaccination than immune responses to completely conserved epitopes<sup>56</sup>.

Following theoretical expectations, our analyses emphasize that targets of natural immunity are not necessarily polarised between the completely invariant and the highly polymorphic, but also critically include a range of epitopes of intermediate variability that have so far been largely overlooked<sup>3</sup>. The *reverse immunodynamics* method introduced in this study has the potential to identify and direct research towards such targets. While the method is likely to evolve with its application to other datasets, we foresee that its future use will help solve the impasse reached in vaccine discovery and development of many multi-strain pathogens.

## Methods

**Genetic sequence selection.** PfAMA-1 sequences were obtained by searching the NCBI Entrez nucleotide and protein databases. This yielded 2835 protein sequences and 2926 nucleotide sequences (up to and including year 2014). In cases when the GenBank nucleotide record did not include the corresponding protein sequence, these were translated using Biopython<sup>57</sup> (v1.69).

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 pseudoduplication, all identical sequences were removed, leaving 1198 unique sequences. These were then aligned using ClustalW<sup>58</sup> (Clustal2, v2.1).

**Indels and other undetermined states.** We first truncated our protein alignment to the region where we had >99% coverage. The resulting sequence alignment (from residue 149 to residue 534) had only 0.0025% of gaps (state ‘-’) and 0.0138% of undetermined amino acids (state ‘X’). In our analysis, we ignored all variants that had gap or undetermined alleles.

**Measuring non-overlap among dimorphic sites.** A frequency analysis of all residues within the three domains 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-analysis.** 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.** The alignment was used to construct a maximum likelihood (ML) tree using Mega6<sup>59</sup> 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). The resulting ML tree was used to calculate the maximum parsimony score (PS) for each dimorphic site (Fig. S2; Table S1). PS was calculated using the Phangorn R-package<sup>60</sup> (v2.3.1) function *parsimony* with default options. We defined a new measure herein termed S-score, as the product of the MI and the geometric mean of the PS for each site (values available in Table S2). We used Seq-Gen<sup>61</sup> (v1.3.4) to simulate 1000 sets of protein sequences with the same phylogenetic characteristics, restricting variation at each of the sites to two amino acids.

**Tertiary structures.** From the protein data bank (PDB) the sequences and tertiary structures of Plasmodium vivax AMA-1 (PvAMA-1, PDB 1W81) and PfAMA-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 PvAMA-1 (1W81) substituted by those from PfAMA-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.

## References

1. von Delft, A. *et al.* The generation of a simian adenoviral vectored HCV vaccine encoding genetically conserved gene segments to target multiple HCV genotypes. *Vaccine*, <https://doi.org/10.1016/j.vaccine.2017.10.079> (2017).
2. Coughlan, L. *et al.* Heterologous Two-Dose Vaccination with Simian Adenovirus and Poxvirus Vectors Elicits Long-Lasting Cellular Immunity to Influenza Virus A in Healthy Adults. *EBioMedicine* **29**, 146–154 (2018).
3. Thompson, C. P. *et al.* A naturally protective epitope of limited variability as an influenza vaccine target. *Nat. Commun.* **9**, 3859 (2018).
4. Gupta, S. *et al.* The maintenance of strain structure in populations of recombining infectious agents. *Nat. Med.* **2**, 437–42 (1996).
5. Gupta, S., Ferguson, N. M. & Anderson, R. M. Chaos, Persistence, and Evolution of Strain Structure in Antigenically Diverse Infectious Agents. *Science* (80-). **280**, 912–915 (1998).
6. Lourenço, J., Wikramaratna, P. S. P. S. & Gupta, S. MANTIS: an R package that simulates multilocus models of pathogen evolution. *BMC Bioinformatics* **16**, 176 (2015).
7. Lourenço, J. *et al.* Lineage structure of *Streptococcus pneumoniae* may be driven by immune selection on the groEL heat-shock protein. *Sci. Rep.* **7**, (2017).
8. Hinnebusch, J., Barbour, A. G., Restrepo, B. I. & Schwan, T. G. 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**, 432–440 (1998).
9. Johnson, D. R., Kaplan, E. L., VanGheem, A., Facklam, R. R. & Beall, B. 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**, 157–64 (2006).
10. Areschoug, T., Ståhlhammar-carlemalm, M., Larsson, C. & Lindahl, G. Group B Streptococcal Surface Proteins as Targets for Protective Antibodies: Identification of Two Novel Proteins in Strains of Serotype V Group B Streptococcal Surface Proteins as Targets for Protective Antibodies: Identification of Two Novel. *Proteins*. **67**, 6350–6357 (1999).
11. Buckee, C. O., Gupta, S., Kriz, P., Maiden, M. C. J. J. & Jolley, K. A. Long-term evolution of antigen repertoires among carried meningococci. *Proc. Biol. Sci.* **277**, 1635–1641 (2010).
12. Callaghan, M. J. *et al.* The effect of immune selection on the structure of the meningococcal opa protein repertoire. *PLoS Pathog.* **4**, e1000020 (2008).
13. Lourenço, J. & Recker, M. Dengue serotype immune-interactions and their consequences for vaccine impact predictions. *Epidemics* **16**, 40–8 (2016).
14. Remarque, E. J., Faber, B. W., Kocken, C. H. M. & Thomas, A. W. Apical membrane antigen 1: a malaria vaccine candidate in review. *Trends Parasitol.* **24**, 74–84 (2008).
15. Polley, S. D. *et al.* Human antibodies to recombinant protein constructs of *Plasmodium falciparum* Apical Membrane Antigen 1 (AMA1) and their associations with protection from malaria. *Vaccine* **23**, 718–728 (2004).
16. Fowkes, F. J. I., Richards, J. S., Simpson, J. A. & Beeson, J. G. The relationship between anti-merozoite antibodies and incidence of *Plasmodium falciparum* malaria: A systematic review and meta-analysis. *PLoS Med.* **7** (2010).
17. Richards, J. S. *et al.* Identification and Prioritization of Merozoite Antigens as Targets of Protective Human Immunity to *Plasmodium falciparum* Malaria for Vaccine and Biomarker Development. *J. Immunol.* **191**, 795–809 (2013).
18. Mugenyi, C. K. *et al.* Antibodies to Polymorphic Invasion-Inhibitory and Non-Inhibitory Epitopes of *Plasmodium falciparum* Apical Membrane Antigen 1 in Human Malaria. *PLoS One* **8** (2013).
19. Stanisic, D. I. *et al.* 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**, 646–660 (2015).
20. Osier, F. H. *et al.* Malaria: New antigens for a multicomponent blood-stage malaria vaccine. *Sci. Transl. Med.* **6** (2014).
21. Muh, F. *et al.* *In vitro* invasion inhibition assay using antibodies against *Plasmodium knowlesi* Duffy binding protein alpha and apical membrane antigen protein 1 in human erythrocyte-adapted P. knowlesi A1-H.1 strain. *Malar. J.* **17**, 1–11 (2018).
22. Vicentin, E. C. *et al.* Invasion-inhibitory antibodies elicited by immunization with *Plasmodium vivax* apical membrane antigen-1 expressed in *Pichia pastoris* yeast. *Infect. Immun.* **82**, 1296–1307 (2014).
23. Biswas, S. *et al.* Recombinant Viral-Vectored Vaccines Expressing *Plasmodium chabaudi* AS Apical Membrane Antigen 1: Mechanisms of Vaccine-Induced Blood-Stage Protection. *J. Immunol.* **188**, 5041–5053 (2012).
24. Schussek, S. *et al.* Immunization with apical membrane antigen 1 confers sterile infection-blocking immunity against *Plasmodium sporozooite* challenge in a rodent model. *Infect. Immun.* **81**, 3586–3599 (2013).
25. Ouattara, A. *et al.* Molecular basis of allele-specific efficacy of a blood-stage malaria vaccine: Vaccine development implications. *J. Infect. Dis.* **207**, 511–519 (2013).
26. Drew, D. R. *et al.* Defining the Antigenic Diversity of *Plasmodium falciparum* Apical Membrane Antigen 1 and the Requirements for a Multi-Allele Vaccine against Malaria. *PLoS One* **7** (2012).
27. Dicko, A. *et al.* Impact of a *Plasmodium falciparum* AMA1 vaccine on antibody responses in adult Malians. *PLoS One* **2**, 1–10 (2007).
28. Thera, M. A. *et al.* Safety and immunogenicity of an AMA-1 malaria vaccine in Malian adults: Results of a phase 1 randomized controlled trial. *PLoS One* **3** (2008).
29. Laurens, M. B. *et al.* 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** (2013).
30. Biswas, S. *et al.* Assessment of humoral immune responses to blood-stage malaria antigens following ChAd63-MVA immunization, controlled human malaria infection and natural exposure. *PLoS One* **9** (2014).
31. Bai, T. *et al.* Structure of AMA1 from *Plasmodium falciparum* reveals a clustering of polymorphisms that surround a conserved hydrophobic pocket. *Proc. Natl. Acad. Sci.* **102**, 12736–12741 (2005).
32. Lim, S. S. *et al.* Structure and dynamics of apical membrane antigen 1 from *Plasmodium falciparum* FVO. *Biochemistry* **53**, 7310–7320 (2014).
33. Crosnier, C. *et al.* Binding of *Plasmodium falciparum* merozoite surface proteins DBLMSP and DBLMSP2 to human immunoglobulin M Is conserved among broadly diverged sequence variants. *J. Biol. Chem.* **291**, 14285–14299 (2016).
34. Polley, S. D. & Conway, D. J. Strong diversifying selection on domains of the *Plasmodium falciparum* apical membrane antigen 1 gene. *Genetics* **158**, 1505–12 (2001).
35. Lumkul, L., Sawaswong, V., Simpallipan, P. & Kaewthamasorn, M. Unraveling Haplotype Diversity of the Apical Membrane Antigen-1 Gene in *Plasmodium falciparum* Populations in Thailand. **56**, 153–165 (2018).
36. Kang, J.-M. *et al.* Population genetic structure and natural selection of *Plasmodium falciparum* apical membrane antigen-1 in Myanmar isolates. *Malar. J.* **17**, 71 (2018).
37. Richard, D. *et al.* 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**, 14815–14822 (2010).
38. Lamarque, M. *et al.* The RON2-AMA1 interaction is a critical step in moving junction-dependent invasion by apicomplexan parasites. *PLoS Pathog.* **7**, (2011).
39. Srinivasan, P. *et al.* Binding of *Plasmodium* merozoite proteins RON2 and AMA1 triggers commitment to invasion. *Proc. Natl. Acad. Sci.* **108**, 13275–13280 (2011).

40. Healer, J. *et al.* Allelic polymorphisms in apical membrane antigen-1 are responsible for evasion of antibody-mediated inhibition in *Plasmodium falciparum*. *Mol. Microbiol.* **52**, 159–168 (2004).
41. Cortés, A. *et al.* Allele Specificity of Naturally Acquired Antibody Responses against *Plasmodium falciparum* Apical Membrane Antigen 1. *Society* **73**, 422–430 (2005).
42. Tanabe, K., Mackay, M., Goman, M. & Scaife, J. G. Allelic dimorphism in a surface antigen gene of the malaria parasite *Plasmodium falciparum*. *J. Mol. Biol.* **195**, 273–287 (1987).
43. Marshall, V. M. *et al.* A *Plasmodium falciparum* MSA-2 gene apparently generated by intragenic recombination between the two allelic families. *Mol. Biochem. Parasitol.* **45**, 349–351 (1991).
44. Arnott, A. *et al.* 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. *Malar J* **13**, 233 (2014).
45. Amambua-Ngwa, A. *et al.* Population Genomic Scan for Candidate Signatures of Balancing Selection to Guide Antigen Characterization in Malaria Parasites. *PLoS Genet.* **8** (2012).
46. Osier, F. H. A. *et al.* Allelic diversity and naturally acquired allele-specific antibody responses to *Plasmodium falciparum* apical membrane antigen 1 in Kenya. *Infect. Immun.* **78**, 4625–4633 (2010).
47. Tetteh, K. K. A. *et al.* Prospective identification of malaria parasite genes under balancing selection. *PLoS One* **4** (2009).
48. Pavlidis, P. & Alachiotis, N. A survey of methods and tools to detect recent and strong positive selection. *J. Biol. Res.* **24**, 7 (2017).
49. Bhatt, S., Katzourakis, A. & Pybus, O. G. Detecting natural selection in RNA virus populations using sequence summary statistics. *Infect. Genet. Evol.* **10**, 421–430 (2010).
50. Buckee, C. O., Recker, M., Watkins, E. R. & Gupta, S. Role of stochastic processes in maintaining discrete strain structure in antigenically diverse pathogen populations. *Proc. Natl. Acad. Sci. USA* **108**, 15504–15509 (2011).
51. Kato, K., Mayer, D. C. G., Singh, S., Reid, M. & Miller, L. H. Domain III of *Plasmodium falciparum* apical membrane antigen 1 binds to the erythrocyte membrane protein Kx. *Proc. Natl. Acad. Sci. USA* **102**, 5552–5557 (2005).
52. Nair, M. *et al.* Structure of domain III of the blood-stage malaria vaccine candidate, *Plasmodium falciparum* apical membrane antigen 1 (AMA1). *J. Mol. Biol.* **322**, 741–753 (2002).
53. Mueller, M. S. *et al.* 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**, 4749–4758 (2003).
54. Thompson, F. M. *et al.* Evidence of blood stage efficacy with a virosomal malaria vaccine in a phase IIa clinical trial. *PLoS One* **3** (2008).
55. Duan, J. *et al.* Population structure of the genes encoding the polymorphic *Plasmodium falciparum* apical membrane antigen 1: implications for vaccine design. *Proc. Natl. Acad. Sci. USA* **105**, 7857–62 (2008).
56. Dutta, S. *et al.* Overcoming Antigenic Diversity by Enhancing the Immunogenicity of Conserved Epitopes on the Malaria Vaccine Candidate Apical Membrane Antigen-1. *PLoS Pathog.* **9**, 1–17 (2013).
57. Cock, P. J. A. *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).
58. Larkin, M. A. *et al.* Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947–2948 (2007).
59. Tamura, K., Stecher, G., Peterson, D., Filipiński, A. & Kumar, S. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Mol. Biol. Evol.* **30**, 2725–9 (2013).
60. Schliep, K. P. phangorn: Phylogenetic analysis in R. *Bioinformatics* **27**, 592–593 (2011).
61. Rambaut, A. & Grassly, N. C. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Mol. Biol.* **13**, 235–238 (1997).

## Acknowledgements

We thank Dr. CRC Spensley for assistance with data manipulation and programming, and Dr. Pietro Roversi for assistance with generating the structure. JL and SG received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013, ERC grant number 268904 - DIVERSITY).

## Author Contributions

S.G., A.L.S., O.G.P., and J.L. developed the theory. K.J.S., J.L., P.S.W., S.G. and L.J. designed the study. K.J.S., J.L., P.S.W. and A.W. performed the experiments. K.J.S. and P.S.W. curated the data. K.J.S., P.S.W., S.G. and J.L. wrote the first draft of the manuscript. J. L. rewrote the manuscript after revision, including text, supplementary material and main figures. All authors revised the final manuscript.

## Additional Information

**Supplementary information** accompanies this paper at <https://doi.org/10.1038/s41598-018-37288-x>.

**Competing Interests:** The authors declare no competing interests.

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2019