

# Immune Signature Against *Plasmodium falciparum* Antigens Predicts Clinical Immunity in Distinct Malaria Endemic Communities

## Authors

Carla Proietti, Lutz Krause, Angela Trieu, Daniel Dodoo, Ben Gyan, Kwadwo A. Koram, William O. Rogers, Thomas L. Richie, Peter D. Crompton, Philip L. Felgner, and Denise L. Doolan

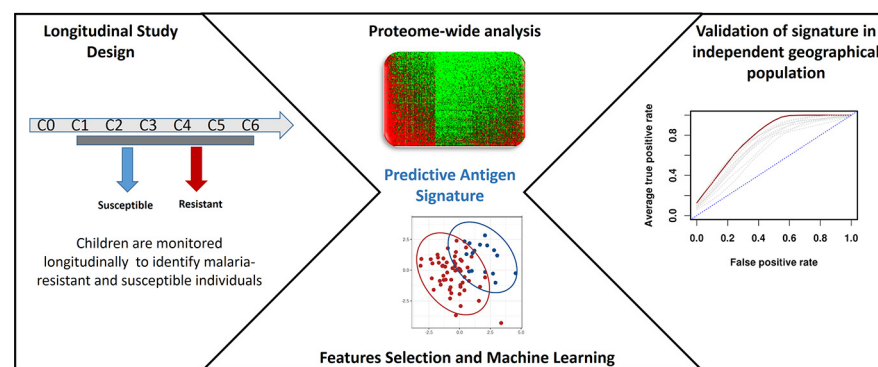
## Correspondence

denise.doolan@jcu.edu.au

## Graphical Abstract

## In Brief

We have established a predictive modelling framework to systematically analyze IgG antibody responses against a large panel of *P. falciparum*-specific antigens and identify a predictive signature of naturally acquired immunity to malaria. Our results show that an individual's immune status can be accurately predicted by measuring IgG antibody responses to a parsimonious set of 15 target antigens. The identified immune signature is highly versatile and capable of providing precise and accurate estimates of clinical protection from malaria in demographically distinct populations.



## Highlights

- A predictive modelling framework has been established to analyze IgG antibody responses against a large panel of *P. falciparum*-specific antigens to identify a specific antigen signature of NAI.
- An individual's immune status can be accurately predicted by measuring IgG responses against a small set of 15 defined parasite antigens.
- Proteins identified in the 15-antigen signature represent potential candidates for next-generation malaria vaccines or biomarkers for monitoring the impact of malaria interventions.
- The developed predictive framework can be adapted for developing novel surveillance and intervention tools for other infectious diseases.



# Immune Signature Against *Plasmodium falciparum* Antigens Predicts Clinical Immunity in Distinct Malaria Endemic Communities\*<sup>§</sup>

Carla Proietti<sup>‡§</sup>, Lutz Krause<sup>¶</sup>, Angela Trieu<sup>‡</sup>, Daniel Dodoo<sup>||</sup>, Ben Gyan<sup>||</sup>, Kwadwo A. Koram<sup>||</sup>, William O. Rogers<sup>\*\*</sup>, Thomas L. Richie<sup>\*\*</sup>, Peter D. Crompton<sup>‡‡</sup>, Philip L. Felgner<sup>§§</sup>, and Denise L. Doolan<sup>‡§¶¶</sup>

**A large body of evidence supports the role of antibodies directed against the *Plasmodium* spp. parasite in the development of naturally acquired immunity to malaria, however an antigen signature capable of predicting protective immunity against *Plasmodium* remains to be identified. Key challenges for the identification of a predictive immune signature include the high dimensionality of data produced by high-throughput technologies and the limitation of standard statistical tests in accounting for synergetic interactions between immune responses to multiple targets. In this study, using samples collected from young children in Ghana at multiple time points during a longitudinal study, we adapted a predictive modeling framework which combines feature selection and machine learning techniques to identify an antigen signature of clinical immunity to malaria. Our results show that an individual's immune status can be accurately predicted by measuring antibody responses to a small defined set of 15 target antigens. We further demonstrate that the identified immune signature is highly versatile and capable of providing precise and accurate estimates of clinical protection from malaria in an independent geographic community. Our findings pave the way for the development of a robust point-of-care test to identify individuals at high risk of disease and which could be applied to monitor the impact of vaccinations and other interventions. This approach could be also translated to biomarker discovery for other infectious diseases. *Molecular & Cellular Proteomics* 19: 101–113, 2020. DOI: 10.1074/mcp.RA118.001256.**

Epidemiological and experimental studies support the role of antibodies directed against *P. falciparum* antigens in pro-

TECTIVE immunity to malaria (1). However, despite decades of intensive efforts, little is known about the parasite antigens that function as targets of naturally acquired immunity (NAI)<sup>1</sup> and there are no defined correlates of protection. Several studies by us and others have demonstrated that immunity is associated with combinations of reactivity against multiple antigens, rather than the recognition of any single antigen (2–7). Identifying the key antigens targeted by NAI and understanding how NAI develops and is maintained within a population would be immensely beneficial for the development of a diagnostic tool to assess whether individuals or populations are at a high risk of disease and whether this risk changes after the implementation of malaria control measures. Moreover, the identification of an immune signature associated with clinical protection would facilitate the design and development of an effective malaria vaccine comprising the subset of antigens shown to be associated with protection. The small number of antigens under development as vaccine candidates reflects our current limited understanding of immunity against malaria (8).

To address this, we have pioneered studies using protein microarrays expressing the complete or partial proteome of *Plasmodium* parasites to profile the immune response on a proteome-wide scale in individuals naturally exposed to or experimentally infected with malaria (3, 9–12). Those studies have shown that ~30% of the proteome is reproducibly recognized and that some antigens are serodominant but other are not (12). Proteome-wide studies provide information on immune responses against a large fraction of the parasite proteome, but the overwhelming amount of generated data has been challenging to analyze and interpret with standard statistical approaches and has limited the success in identi-

From the <sup>‡</sup>Centre for Molecular Therapeutics, Australian Institute of Tropical Health and Medicine, James Cook University, Cairns, QLD, Australia; <sup>§</sup>QIMR Berghofer Medical Research Institute, Brisbane, QLD, Australia; <sup>¶</sup>The University of Queensland Diamantina Institute, Brisbane, QLD, Australia; <sup>||</sup>Noguchi Memorial Institute for Medical Research, University of Ghana, Ghana; <sup>\*\*</sup>Naval Medical Research Center, Silver Spring Maryland; <sup>‡‡</sup>Malaria Infection Biology and Immunity Section, Laboratory of Immunogenetics, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Rockville, Maryland; <sup>§§</sup>Department of Medicine, Division of Infectious Diseases, University of California Irvine, Irvine, California

\* Author's Choice—Final version open access under the terms of the Creative Commons CC-BY license.

Received December 3, 2018, and in revised form, October 23, 2019

Published, MCP Papers in Press, October 28, 2019, DOI 10.1074/mcp.RA118.001256

fyng a protective immune signature. For example, the number of variables (proteins) measured in protein microarray experiments usually far outnumbers the sample size, making data analysis challenging and limiting the applicability of standard statistical tests. Traditional statistical approaches that are based on individual antigens are not robust and accurate enough to predict the immune status at an individual level. To overcome these limitations, multivariate methods and machine learning techniques have been recently applied for the analysis of high dimensional “omics” datasets (transcriptomics, metabolomics, proteomics, metagenomics, etc.) to identify predictive biomarker signatures of vaccination, infection or exposure. Insights into the mechanisms of natural and vaccine-induced immunity have been reported for several diseases, including yellow fever, influenza, and tuberculosis (5, 13–16), but not malaria.

Herein, we have established a predictive modeling framework combining feature selection and machine learning to systematically analyze IgG antibody responses against a large panel of *P. falciparum*-specific antigens to identify a predictive signature of clinical immunity to malaria. In a longitudinal study in Ghana, blood samples were collected from well-defined cohorts of young children in the process of acquiring clinical immunity to malaria during multiple time points of the malaria season. Sera were probed against a protein microarray containing 1,080 *P. falciparum* antigens. By analyzing the antibody profiles before the malaria season, we were able to identify a parsimonious set of antibody responses that could predict an individual's immune status (clinically resistant or susceptible) with high accuracy (86%). We further validated this signature in a distinct epidemiological and demographical setting, among 2–10 year-old children and 18–25 year-old adults in Mali. The predictive modeling framework presented here proved to be a powerful approach to identify a sensitive and specific immune signature of NAI to malaria.

### EXPERIMENTAL PROCEDURES

**Population and Study Design**—Studied children were recruited from the Kassena-Nankana District (KND) of the Upper East region of northern Ghana. In this region malaria transmission occurs throughout the year with two main seasons: a dry season from approximately October to April, and a wet season from approximately May to October (supplemental Fig. S1). The characteristics of the area and study details have been published elsewhere (17–23). Briefly, three hundred children were passively followed up over one calendar year (from May 2004 to May 2005) and were visited seven times (every 2

months). Clinical, hematological and parasitological data were collected at the beginning of the study and during each of the seven visits. A blood sample was obtained by fingerpick (~0.5–1.0 ml) for thick and thin blood film, for serological analysis and rapid malaria diagnostic test (DiaMed Optimal Rapid Malaria test). Children suffering from uncomplicated malaria were treated with chloroquine and sulfadoxine-pyrimethamine (SP) (Fansidar®) as the first-line drugs according to Ghana Ministry of Health policy at the time. Asexual parasites were counted against 200 white blood cells and converted to parasites/ $\mu$ l assuming a density of 8,000 white blood cells/ $\mu$ l blood. Concurrent parasitemia was defined as any documented parasite density at the time of sample collection. Fever was defined as axillary temperature  $\geq 37.5$  °C. Hemoglobin level was measured by Hemocue photometer (Leo Diagnostics, Helsingborg, Sweden). Anemia was defined as hemoglobin level  $< 9$  g/dL. Serum samples from 80 subjects (39 young children aged 1–2 years and 41 old children aged 4–5 years) were randomly selected from this cohort to assay total IgG antibody reactivity against 1080 *P. falciparum* proteins or protein fragments by protein microarray.

**Ethics Statement**—The study protocol for clinical specimens complied with all applicable federal regulations governing the protection of human subjects. The protocol was approved by the Ghana Health Service, the Navrongo Health Research Centre, the Noguchi Memorial Institute for Medical Research, the Naval Medical Research Institutional Review Board, the Office of the Special Assistant for Human Subject Protections at the Naval Bureau of Medicine and Surgery, and the Human Subjects Research Review Board of the Army Surgeon General. All study subjects gave written informed consent. The protein microarray studies were approved by the Queensland Institute of Medical Research Human Research Ethics Committee, and the University of California Irvine Institutional Review Board.

**Protein Microarray Probing**—Total IgG antibody reactivity against *P. falciparum* was assayed using a protein array containing 1080 *P. falciparum* recombinant proteins. Proteins were selected for inclusion based on evidence for blood-stage expression by microarray, proteomics, or expressed sequence tags (ESTs), or predicted to be in the blood-stage secretome (PlasmoDB.org). The array was fabricated as previously described (9). Briefly, coding sequences were PCR-amplified from *P. falciparum* (clone 3D7, (MRA-102, MR4)) genomic DNA and cloned into the PXT7 plasmid with a T7 transcription terminator, and tagged with 5' polyhistidine (HIS) and 3' hemagglutinin (HA) epitopes. Recombinant proteins were expressed using *E. coli* cell-free *in vitro* transcription and translation reactions (RTS 100 HY kits from 5 PRIME, Gaithersburg, MD) according to the manufacturer's instructions. Protein arrays were printed as previously described, with each recombinant protein spotted once in each array (3, 9). Once printed, the expression of each recombinant protein printed on the array was verified using anti-polyhistidine (clone His-1, Sigma) and anti-hemagglutinin (clone 3F10, Roche) monoclonal antibodies, as previously described (9). Epstein-Barr Nuclear Antigen 1 (EBNA-1) was included on each microarray chip as a positive control, and an empty T7 vector rapid translation system (RTS) reaction (no DNA) as a negative control. Sera were pre-absorbed against *E. coli* lysate in protein array blocking buffer (1:100 dilution) and then 500  $\mu$ l of the pre-absorbed plasma was added to each protein array and incubated overnight at 4 °C. Serum antibodies were detected with biotin-conjugated goat anti-human IgG secondary antibody (1:1000 dilution) and visualized with a streptavidin P-3-conjugated antibody (1:200 dilution). Air-dried slides were scanned on an Axon GenePix 4300A array scanner (Molecular Devices, CA) and results read as raw fluorescence signal intensities (SI) were quantified using the Axon GenePix Pro 7 software after correction for spot-specific background. IgG antibody reactivity to the positive control (EBNA-1 fragment) was

<sup>1</sup> The abbreviations used are: NAI, naturally acquired immunity; AMA1, apical membrane antigen 1; AUC, area under the curve; CSP, circumsporozoite protein; EBA175, erythrocyte binding antigen-175; EBNA, Epstein-Barr nuclear antigen; LSA1–3, liver stage antigen1–3; MSP1–2, merozoite surface protein 1–2; PCoA, principal coordinates analysis; RDA, redundancy analysis; ROC, receiver operating characteristic; RTS, rapid translation system; SI, signal intensity; sPLS-DA, sparse partial least squares discriminant analysis; SVM, support vector machine; TRA, thrombospondin-related anonymous; vsn, variance stabilizing normalization.

detected in all individuals, whereas there was no detectable reactivity to the no-DNA (data not shown).

**Array Data and Statistical Analysis**—Protein array data were analyzed primarily in R (<http://www.r-project.org>) and the GMine data-mining server (6) (<http://cgenome.net/gmine>). Raw data were measured as the mean pixel signal intensity for each spot. After subtracting slide background, the mean intensity of negative-control spots (no DNA) was subtracted from each test spot to adjust for any cross-reaction effects from the *E. coli* vector used to express the proteins. Next, the corrected data were transformed using the variance stabilizing normalization (vsN) method in GMine using the VSN Bioconductor package. Proteins were defined as positive if the transformed signal intensity was larger than the mean plus 2 standard deviations (SDs) of the transformed negative-control spots (no DNA). For each child, the breadth of the antibody responses was calculated as the number of positive responses against the 1080 *P. falciparum* proteins. Antibody profiles against the 1080 *P. falciparum* proteins were associated with multiple explanatory variables using the multivariate statistical methods redundancy analysis (RDA). Age group (young/old children), time of the malaria season (dry/wet), parasite positive (yes/no), hemoglobin level (g/dL), fever (yes/no) and gender (M/F) were included as explanatory variables and the antibody responses matrix was included as dependent variable.

**Identification of Proteins Associated with Immunity to Malaria**—Associations between antibody responses against individual *P. falciparum* proteins at baseline and subsequent protection from symptomatic malaria were identified by logistic regression. Immune status (resistant/susceptible), age and parasite status (parasite positive/parasite negative) at baseline were included as explanatory variables (predictors) and baseline antibody intensity (SI) against individual *P. falciparum* proteins was modeled as dependent (response) variable. The regression model had the form: antibody signal intensity = protection status (resistant/susceptible) + age group (younger/older) + parasite status at baseline (positive/negative). Individuals were categorized as susceptible if they had a recorded episode of symptomatic malaria (clinical disease) within our one-year study period, and resistant otherwise. Symptomatic malaria was defined as febrile illness (axillary temperature  $\geq 37.5$  °C) with concurrent *P. falciparum* parasitemia  $>2500$  parasites/ $\mu$ l. To minimize the misclassification bias that arises because of difficulties in ascertaining exposure to the *P. falciparum* parasite (24), we restricted this analysis to 72 children (35 aged 1–2 years and 37 aged 4–5 years) with any documented parasitemia in at least one contact during the longitudinal study. *p* values were Bonferroni corrected to adjust for multiple statistical comparisons and the significance level  $\alpha$  was set to 0.05.

**Identification of a Predictive Immune Signature**—To identify predictive immune signatures underlying clinical protection, we employed sparse Partial Least Squares Discriminant Analysis (sPLS-DA) (25). The high dimensionality and sparsity of protein microarray data, where the number of measured variables (proteins) far outnumbers the sample size, represent challenges that affect applicability of standard statistical tests. sPLS-DA is a powerful method for identifying the key variables of complex and sparse omics datasets that are associated with a biological outcome of interest and can be used for feature selection. The sPLS-DA implemented in the MixOmics R package (26) was run via the GMine data-mining server (6). This procedure involves dimension reduction using Partial Least Squares regression (PLS) for discriminant analysis in combination with a Lasso penalization for feature selection. To reduce the dimensionality of the input data for the sPLS-DA to a manageable size, only the subset of *P. falciparum* proteins was included that had a *p* value  $< 0.1$  in the logistic regression (described above) and signal intensity higher in the resistant group. The immune status (susceptibility or resistance) of children during the one-year study period was included as a response

variable. Longitudinal changes in antibody responses against individual antigens were identified by mixed-effects linear regression (27), which can account for the repeated measures study design with multiple samples for each subject. Mixed-effects linear regression modeled antibody signal intensity as dependent variable and time point as fixed effect and individual as a random effect. Following sPLS-DA based feature selection, we then employed a Support Vector Machine (SVM) to test the power of the identified immune signature to correctly discriminate between resistant and susceptible individuals. SVMs are a robust and powerful classification technique that have achieved excellent classification abilities for a wide range of applications, but that are not well suited for feature selection. The SVM was trained on the baseline signal intensities of the selected antigens and evaluated by Monte Carlo cross-validation. Iteratively, 30 samples were randomly selected as a training set and the remaining 42 samples were used as test set. The model was then fit to the training data and evaluated on the test set and the predictive accuracy was defined as the number of correct predictions divided by the total number of predictions. This process was repeated 40 times. The predictive performance, of the model was then evaluated by averaging the receiver operating characteristic (ROC) curves across all 40 Monte Carlo cross-validation runs. The ROCR R package was used for plotting ROC curves and determining the Area Under the Curve (AUC).

**Effect of the Definition of Clinical Malaria on the Predictive Performance of the Signature**—We defined symptomatic malaria (clinical disease) as febrile illness (axillary temperature  $\geq 37.5$  °C) with concurrent *P. falciparum* parasitemia  $>2500$  parasites/ $\mu$ l. To assess the effect that different clinical case definitions have on the performance of the identified antigen signature in predicting the immune status of an individual, we performed a sensitivity and specificity analysis as described below. We considered a positive episode of symptomatic malaria to be one in which there was both fever and parasitemia greater than a specified threshold. For each parasitemia threshold, we counted the number of individuals who had at least one episode of symptomatic malaria during the study period, and we randomly selected an equal number of resistant individuals from the entire cohort. For each parasite threshold, we trained SVM on the baseline signal intensities of the selected antigens and evaluated the performance by a leave-one-out cross-validation. The process was repeated 5 times for each parasite threshold, each time with a different random subset of resistant children from the cohort as a control group. Each time, sensitivity and specificity were calculated using the following formula:

Sensitivity:  $TP/(TP+FN)$

Specificity:  $TN/(TN+FP)$

Where

TP: true positives (*i.e.* the number of resistant individuals correctly classified as resistant).

FP: false positives (*i.e.* the number of susceptible individuals incorrectly classified as resistant).

TN: true negatives (*i.e.* the number of susceptible individuals correctly classified as susceptible).

FN: false negatives (*i.e.* the number of resistant individuals incorrectly classified as susceptible).

The sensitivity and specificity of the classifier was then averaged across the 5 runs and plotted versus the parasite threshold.

**Validation of the Predictive Immune Signature in an Independent Cohort**—The predictive immune antigen signature discovered by sPLS-DA was next validated in an independent cohort of children exposed to malaria. As a validation dataset, we used our previously published protein microarray data generated in a prospective study investigating host immune response to the malaria parasite *P. falciparum* (10) in 225 individuals from Kambila, Mali. In that study, we used a protein microarray consisting of 2320 probes (representing ~23%

of the *P. falciparum* proteome and including all of the 1080 proteins on the array tested herein) to profile host immune response (IgG) against malaria parasite proteins. Plasma samples were collected from each individual before and after the 8-month malaria season. Individuals were classified as resistant (no malaria episode during the 8 months study) or susceptible ( $\geq 1$  malaria episodes). In the current study, we included the subset of 194 individuals between the ages of 2–10 years and 18–25 years as an independent validation set. Of these, 128 experienced clinical malaria during the follow up (susceptible) and 66 individuals remained asymptomatic (resistant). Signal intensities were transformed and normalized as previously described (6). Using a SVM we assessed if the predictive antigen signature identified in the Ghana cohort was also able to discriminate susceptible and resistant children from the Mali cohort. Using the baseline data of the Mali cohort, a SVM was trained on the signal intensities of our predictive signature discovered in Ghana and evaluated by Monte Carlo cross-validation as described above. Predictive accuracy, sensitivity and specificity were calculated as described above. A modified randomization test was then carried out to evaluate if the achieved classification accuracy was better than expected by chance. In 1000 iterations, a subset of the 1080 *P. falciparum* proteins was randomly selected (a subset with the same size as the predictive antigen signature discovered in Ghana) and a SVM was trained and validated by Monte Carlo cross-validation on this subset. Finally, we counted the number of times that our predictive antigen signature performed better than the 1000 random subsets.

**Development and Validation of a Simple Diagnostic Decision Rule to Distinguish Resistant from Susceptible Individuals**—We aimed to develop a simple decision rule using the 15-antigen signature discovered by sPLS-DA that could be used to adequately distinguish between resistant or susceptible individuals and aid in diagnostic classification. We developed a simple decision rule like the method developed by Pan *et al.* (28).

First, signal intensities of the n-antigen signature were converted to a binary output by assigning 1 (high) to signal intensities above a selected threshold and 0 (low) to signal intensities below the threshold. The level threshold for each antigen was defined as the mean signal intensity of that antigen in the resistant group minus 1 standard deviation for the Ghana dataset and 0.5 standard deviations for the Mali dataset; the cut-offs were different to account for differences in the variance of signal intensities in the two datasets. Next, an “immunity score” was defined as the sum of the binary outputs for the n-antigen signature. Samples with an immunity score greater than  $z$  (with  $1 \leq z \leq n$ ) were classified as resistant and as susceptible otherwise. The optimal  $z$  value was determined by calculating predictive accuracy, sensitivity and specificity for all possible  $z$  in the range 1 to  $n$ .

**Conservation of the Antigens in the Predictive Immune Signature**—The amino acid sequences from the proteins identified as predictive immune signature were blasted against the entire NCBI dataset for *Plasmodium* (*Plasmodium* taxid:5820) using protein-protein BLAST, v2.2.29+. P-P BLAST was run without the filter and an E-value cutoff of  $1e^{-15}$ .

The genomes included in the analysis of conservation are all publicly available *Plasmodium falciparum* strains ( $n = 15$ ) and 11 human-infective *Plasmodium* species. Species and strains included are listed as follows: *Plasmodium falciparum* 7G8, *Plasmodium falciparum* CAMP Malaysia, *Plasmodium falciparum* Dd2, *Plasmodium falciparum* FCH 4, *Plasmodium falciparum* HB3, *Plasmodium falciparum* IGH CR14, *Plasmodium falciparum* MaliPS096\_E11, *Plasmodium falciparum* NF135 5C10, *Plasmodium falciparum* NF54, *Plasmodium falciparum* Palo Alto Uganda, *Plasmodium falciparum* RAJ116, *Plasmodium falciparum* Santa Lucia, *Plasmodium falciparum* Tanzania 200070, *Plasmodium falciparum* UGT51, *Plasmodium falciparum* Viet0m Oak Knoll FVO, *Plasmodium knowlesi*, *Plasmodium knowlesi*

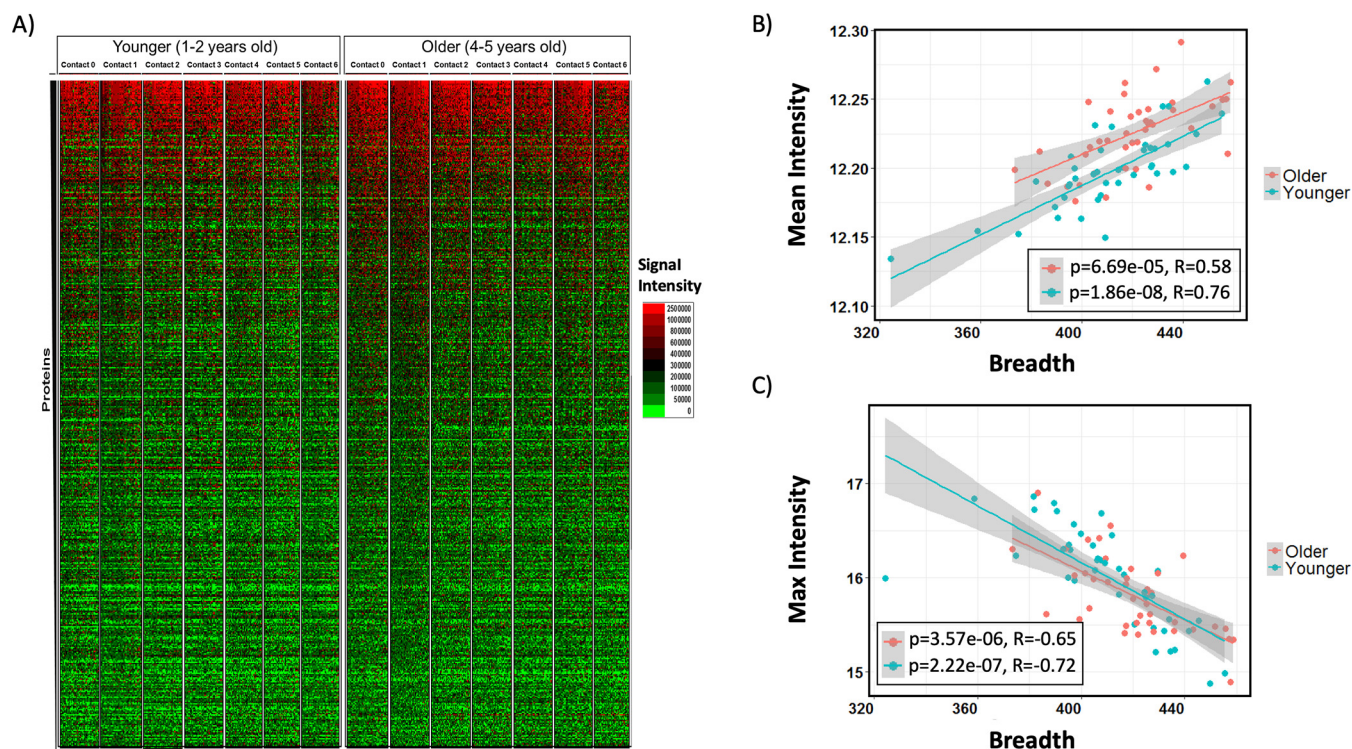
strain H, *Plasmodium malariae*, *Plasmodium ovale*, *Plasmodium ovale curtisi*, *Plasmodium ovale wallikeri*, *Plasmodium vivax*, *Plasmodium vivax* Brazil I, *Plasmodium vivax* India VII, *Plasmodium vivax* Mauritania I, *Plasmodium vivax* North Korean.

## RESULTS

**Characteristics of the Study Population and Clinical Outcomes in Relation to the Transmission Season and Age**—A total of 80 children were randomly selected from a previously described longitudinal study in Ghana (23): 39 young children aged 1 to 2 years and 41 children aged 4 to 5 years (supplemental Table S1). *P. falciparum* infection was prevalent during the entire year, with 58–69% of younger children and 69–95% of older children carrying parasites throughout the study period (supplemental Table S1, supplemental Fig. S2A). In general, older children showed concurrent parasitemia at the time of sampling more frequently than younger children (supplemental Fig. S2A) ( $p = 6.95E^{-5}$ , Wilcoxon rank-sum test), but with a similar level of parasite density ( $p = 0.934$ , Wilcoxon rank-sum test) (supplemental Table S1, supplemental Fig. S2B). As expected, parasite prevalence and density were significantly higher during the wet season ( $p = 0.032$ ,  $p = 0.0007$  respectively; Wilcoxon paired rank-test) (supplemental Table S1). Younger children were more frequently anemic than older children ( $p = 2.51E^{-7}$ , Wilcoxon rank-sum test), with a higher hemoglobin level during the dry season compared with the wet season ( $p = 4.06E^{-4}$ , Wilcoxon paired rank-test) (supplemental Table S1).

**Antibody Reactivity Associated with Age, Gender, Concurrent Parasitemia and Malaria Season**—Blood samples from each included child ( $n = 80$ ) were collected every two months during the malaria season (7 time points per subject; 539 samples in total) and examined for IgG reactivity against 1,080 individual *P. falciparum* proteins (Fig. 1A). As expected, a high degree of variability in antibody responses was observed across individuals, and within individuals longitudinally during the study period. The overall mean of antibody intensity was significantly higher in older children ( $p = 4.85E^{-05}$  Wilcoxon rank-sum test), and the mean intensities were higher during the wet season compared with the dry season ( $p = 6.93E^{-5}$  Wilcoxon paired rank-test) (supplemental Fig. S3).

Antibody responses to some of the malaria vaccine candidate antigens included in the protein array (erythrocyte binding antigen-175 (EBA175), thrombospondin-related anonymous protein (TRAP), and circumsporozoite protein (CSP)) were generally higher in older children ( $p < 0.05$ , Wilcoxon rank-sum test) consistent with other studies (2, 10, 21, 29) (supplemental Fig. S4). Conversely, antibody responses against the leading liver-stage vaccine candidates included in the array, LSA-1 and LSA-3, were surprisingly higher in younger children compared with older children (supplemental Fig. S4). Other blood-stage vaccine candidates included in the array showed similar levels of reactivity across the two study



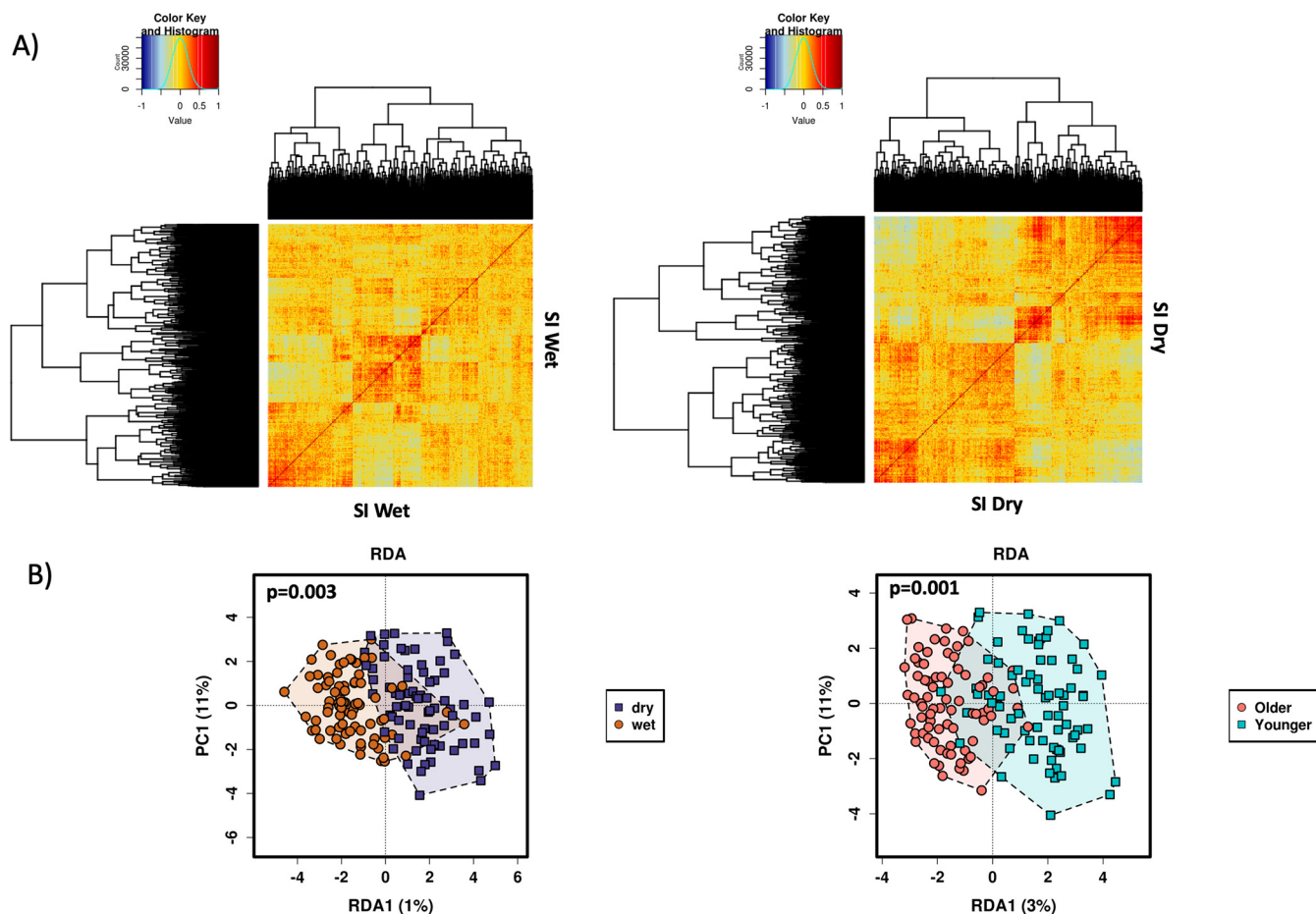
**FIG. 1. Longitudinal study of IgG response of 39 young children (1–2 years of age) and 41 older children (4–5 years of age) against *P. falciparum* antigens.** A, Heat map of the top 500 most reactive *P. falciparum* antigens by time point ordered by reactivity. Serum samples were collected from each participant every two months, starting at the beginning of the wet season (May 2002) and ending 12 months later at the end of the dry season (May 2003). Antibody responses were profiled using a protein microarray expressing 1080 *P. falciparum* antigens. Antibody responses were generally stronger in older versus younger children and for the same children during the wet versus dry season. B–C, Correlation between mean signal intensity and breadth of response (number of reactive proteins per child) (B) and between maximum intensity and breadth of response (C). Mean intensities showed a positive correlation with breadth of response whereas maximum intensity was negatively correlated. Breadth of response was defined as the number of reactive proteins per child (at least 2 S.D. above mean of negative controls) and averaged among the six contacts. Mean intensity was calculated as the mean intensities of the 1,080 proteins for each subject and averaged among the 6 contacts. Significance was tested by Pearson's correlation. The red line and the blue line represent the best linear fit for younger and older children respectively. The gray area represents the 95% CI of the best-fit line

groups or higher levels in younger children ( $p > 0.05$ , Wilcoxon rank-sum test) (supplemental Fig. S4).

Antibody responses showed a gradient of reactivity across the population with some individuals displaying stronger signals and broader recognition compared with others (Fig. 1A). Not surprisingly, we observed a positive correlation between the breadth of antibody responses (*i.e.* the number of proteins to which IgG reacted in each subject) and the mean intensity of all IgG reactivity on the array ( $p = 2.826E-13$ ,  $r = 0.70$ , Pearson's correlation) (Fig. 1B). However, unexpectedly, we observed a significant inverse correlation between the breadth of antibody responses and the maximum antibody reactivity ( $p = 3.21E-14$ ,  $r = -0.71$ , Pearson's correlation) (Fig. 1C). The inverse correlation was stronger for younger children ( $p = 2.22e-07$ ,  $r = -0.72$ , Pearson's correlation) compared with older children ( $p = 3.57e-06$ ,  $r = -0.65$ , Pearson's correlation). These results suggest that high antibody responses against individual dominant proteins may suppress the response to other protein, resulting in either a low number of highly reactive proteins or a broader number

of less reactive antigens. The inverse correlation between breadth and intensity may represent somatic hypermutation and affinity maturation occurring after long term exposure to the parasite. In early years, the antibody responses may be relatively low affinity against a large collection of antigens whereas as the immune response matures over years it may become more focused with higher affinity against a few antigens.

Antibody responses to the 1,080 *P. falciparum* proteins were positively correlated (Fig. 2A), probably as a result of a co-acquisition of antibodies to parasite antigens. However, antibody profiles were highly subject-specific and clustered by age group, and by the malaria season (wet/dry) (Fig. 2B). The multivariate statistical method redundancy analysis (RDA) was applied to test the significance of the observed associations. In both the dry and wet seasons antibody responses were significantly associated with age ( $p = 0.001$  wet,  $p = 0.002$  dry, RDA) and hemoglobin level (hemoglobin level  $<9$  g/dL) ( $p = 0.001$  both season, RDA). The presence or absence of fever, gender and parasite density were not associated with antibody reactivity in either of the two seasons (wet season:



**FIG. 2. Antibody responses clustered by age group and time point.** A, Pearson's correlation of the antibody responses to the 1,080 *P. falciparum* proteins at a single time point during the wet season (September) (LHS) and dry season (March) (RHS). Antibody responses to the *P. falciparum* proteins were positively correlated in both seasons. B, Multivariate statistical methods redundancy analysis (RDA) showed a significant clustering of the antibody profiles against the 1080 *P. falciparum* proteins by time of the malaria season (dry/wet) ( $p = 0.003$ ) (LHS) and by age group (younger/older) ( $p = 0.001$ ) (RHS).

$p = 0.500$ ,  $p = 0.704$ ,  $p = 0.526$  respectively; dry season:  $p = 0.183$ ,  $p = 0.124$ ,  $p = 0.397$  respectively, RDA).

**Individual Antigens Do Not Predict Protection Against Clinical Malaria**—To determine if antibody responses against individual *P. falciparum* proteins could prospectively discriminate resistant from susceptible children, we compared the baseline antibody profiles of children who did not experience any malaria episode during the following one-year study period ( $n = 56$ ) (resistant) compared with those who experienced at least one episode ( $n = 16$ ) (susceptible). Among the 16 susceptible individuals, 13 reported one episode of symptomatic malaria during the year whereas 3 individuals reported two episodes. No significant differences were observed in parasite density ( $p = 0.313$ , Wilcoxon rank-sum test), breadth ( $p = 0.329$ , Wilcoxon rank-sum test) or hemoglobin level ( $p = 0.635$ , Wilcoxon rank-sum test) between resistant and susceptible children at baseline (Table I). In our cohort, none of the 1,080 *P. falciparum* proteins individually was able to discriminate resistant and susceptible children after correct-

ing for multiple testing (Bonferroni  $p > 0.05$ , logistic regression adjusted for age and parasitemia at the baseline) (supplemental Table S2). Of note, among the antibody responses against the leading malaria vaccine candidates included in the protein array (apical membrane antigen 1 (AMA1), merozoite surface protein 1 (MSP1), merozoite surface protein 2 (MSP2), erythrocyte binding antigen-175 (EBA175)), thrombospondin-related anonymous protein (TRAP), liver-stage antigen 1 and 3 (LSA1-LSA3) and circumsporozoite protein (CSP), none were within the top 100 proteins in discriminating resistant and susceptible children (supplemental Table S2) consistent with recently published findings by us and others (6, 30, 31).

**Identification of a Predictive Antigen Signature for Protection Against Symptomatic Malaria**—Based on our previous observations (3, 6, 9, 10) and other published studies (2, 5–7), susceptibility to symptomatic malaria is more likely to be predicted by measuring antibody responses to a combination of antigens rather than individual antigens. We therefore sought to identify a signature of the most relevant proteins

TABLE I  
Characteristic of resistant and susceptible children at the baseline (contact 0)

Children were defined as “resistant” ( $n = 56$ ) if they did not experience a clinical malaria episode during the 12-month study period and “susceptible” ( $n = 16$  children) otherwise. Clinical malaria was defined as fever (temperature above 37.5 °C) and parasite density above 2500 parasite/ $\mu$ L.

Characteristic	Resistant	Susceptible
No. of individuals	56	16
Younger (1–2 years), $n$ (%)	26 (46.4)	9 (56.3)
Female sex, $n$ (%)	24 (42.9)	8 (50.0)
Parasitemic at time of sample collection, $n$ (%)	46 (82.1)	12 (75.0)
Median <i>P. falciparum</i> density at time of sample collection (IQR) (parasites/ $\mu$ L)	640 (110–2120)	680 (240–2520)
Median Haemoglobin level (g/dL) at time of sample collection	10.65 (9.56–11.13)	10.50 (9.28–12.03)
Median breadth <sup>a</sup> at time of sample collection (IQR)	419 (385–437)	424 (407–439)

Only participants who had at least one *P. falciparum* infection recorded in the year were included.

<sup>a</sup>Breadth was defined as the number of positive proteins per individual.

that could predict a child’s immune status to malaria. Sparse Partial Least Squares Discriminant Analysis (sPLS-DA) effectively identified a signature of 15 *P. falciparum* antigens capable of discriminating susceptible versus resistant children (Fig. 3A, supplemental Table S3). As expected, Pearson’s correlation of the antibody responses against the 15-antigen signature showed a null or very low positive correlation between antigens (Fig. 3B) and a significantly higher intensity in the “resistant” group compared with the “susceptible” group (Fig. 3C). Using linear mixed-effects regression, we observed no significant longitudinal changes in antibody reactivity to the antigens in the signature (supplemental Fig. S5), suggesting stability of the antibody signature over different time points of the malaria transmission season.

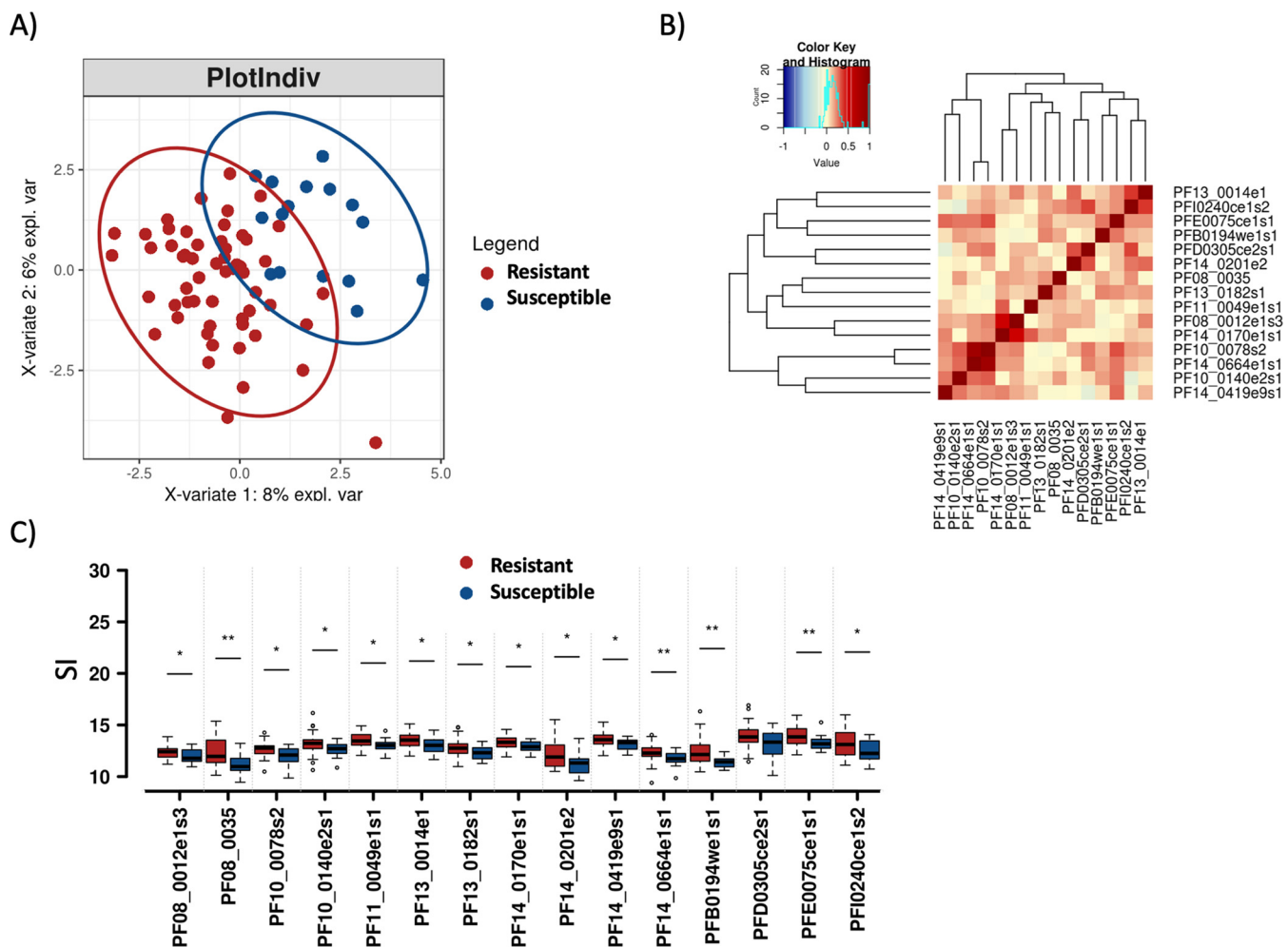
The discriminatory power of the 15-antigen signature to correctly predict immunity to malaria was examined using a Support Vector Machine (SVM) evaluated by Monte Carlo cross-validation. When all children (1–5 years) were included, the SVM correctly predicted immunity to malaria using the baseline antibody responses against the 15 selected antigens with 86% accuracy, 88% sensitivity, 82% specificity and an AUC of 0.94 (Fig. 4A). This predictive performance dropped substantially when the model was applied only to younger children (1–2 years) (accuracy 69%, sensitivity 85%, specificity 85%, AUC = 0.75) but remained significantly high when only older children (4–5 years) were included (accuracy 83%, sensitivity 97%, specificity 88%, AUC = 0.97) (Fig. 4B). Because the predictive performance may depend on the clinical case definition used in the study, we considered how an over- or under-estimation of clinical malaria cases might affect the performance of the 15-antigen signature in predicting immunity. As described under Experimental procedures, we considered a series of clinical case definitions of symptomatic malaria. Each definition required the presence of an axillary temperature  $\geq 37.5$  °C and a parasitemia above a specified threshold. We modeled the effect of changing parasitemia thresholds on the predictive performance of the 15-antigen signature. Supplemental Fig. S6 shows the sensitivity and specificity of the performance of the 15-antigen signature in

predicting symptomatic and resistant individuals as a function of the parasitemia threshold. For resistant individuals, varying the parasitemia threshold across a wide range had a relatively minor effect on sensitivity (supplemental Fig. S6A) whereas susceptible individuals were more sensitive to the threshold chosen, varying from 58% to 96% as the threshold varied from any detectable parasitemia to up to 4000 parasites/ $\mu$ L. The sensitivity decreased to 88% at 5000 parasites/ $\mu$ L. The specificity varied greatly for both resistant and susceptible individuals (supplemental Fig. S6B), increasing from 64% to 94% and from 71% to 85% from 1 to 4000 parasites/ $\mu$ L respectively, for resistant and susceptible individuals and then decreased to 88 and 79%, respectively, at 5000 parasites/ $\mu$ L.

*Validation of 15-antigen Signature in an Independent Cohort From Mali*—We used previously published protein microarray data from Malian children (10) aged 2–10 years old and adults aged 18–25 years to determine whether the predictive 15-antigen signature identified in Ghanaian children could also predict the immune status in individuals from Mali. Subjects were defined as either resistant if they did not experience a clinical malaria episode during the 8-month study period ( $n = 66$ ), or susceptible if they did ( $n = 128$ ). A SVM trained on the antibody responses at baseline against the 15-antigen signature and evaluated by Monte Carlo cross-validation was able to predict the immune status of Mali individuals with an accuracy of 87%, sensitivity 68%, specificity 82% and an AUC of 0.82. When the model training only included children, the average performance of the model maintained an accuracy 86% and specificity of 73% but with a lower AUC (0.74) and lower sensitivity (33%), as a result of misclassification of resistant individuals.

To assess whether these results were better than those expected by chance, we carried out a re-randomization analysis. From the 1080 *P. falciparum* proteins spotted onto the array, we randomly selected 15 proteins and tested the accuracy in classifying resistant individuals in the Mali dataset by SVM with Monte Carlo cross-validation as described above. The analysis was repeated 1000 times using 1000 random sets of 15 proteins. The 15-antigen signature identi-





**FIG. 3. sPLS-DA identified a 15-antigen signature discriminating susceptible and resistant children.** A, sPLS-DA of antibody responses at contact 0 (baseline) showed a separation between resistant ( $n = 56$ ) and susceptible children ( $n = 16$ ). Children aged 1–5 years ( $n = 72$ ) were defined as “resistant” ( $n = 56$ ) if they did not experience a clinical malaria episode during the 12-month study period and “susceptible” ( $n = 16$  children) if they experienced clinical malaria. Antibody responses were measured at the beginning of the wet season (contact 0). Ellipses highlight 95% confidence intervals. B, Pearson’s correlation of the antibody responses against the 15-antigen signature, presented as a heat map. The heat map presents correlations by color code, ranging from blue (negative association) to red (positive correlation). The heat map showed a null or low positive correlation between antigens. C, Boxplot of the 15-antigen signature showed a significant difference between resistant and symptomatic children ( $*p < 0.05$ ,  $**p < 0.01$   $t$  test).

fied in the Ghana cohort performed better than 984 of the 1000 random sets ( $p = 0.016$ ) (Fig. 5). When the analysis only included children, the number of times the signature performed better than 1000 random sets decreased to 925 ( $p = 0.075$ ), suggesting that the predictive power of the signature increases with the age and/or exposure, as reported in a previous study (2, 30). The ability of the 15-antigen signature to predict the immune status in Mali suggests that the signature may be generalizable across epidemiologically and demographically distinct populations.

**Development and Validation of a Simple Diagnostic Decision Rule to distinguish Resistant From Susceptible Individuals**—We developed a simple decision rule based on our 15-antigen signature that could be used to distinguish between resistant or susceptible individuals and aid in diagnostic clas-

sification. Signal intensities were transformed into binary values (0/1 = low/high) based on defined antigen-specific threshold and individuals were classified as resistant if more than  $z$  of the 15 antigens had high intensities (*i.e.* the sum of binary values was greater or equal  $z$ ). Our 15-antigen signature achieved the best performance for  $z = 11$  with 87% accuracy, 76% sensitivity, and 100% specificity, in our Ghana cohort (Table II; Fig. 6A). The performance of the proposed decision rule was evaluated on an independent cohort of individuals from Mali, achieving 72% accuracy, 71% sensitivity and 72% specificity (Fig. 6B). The high accuracy observed for the independent validation set demonstrates the robustness of our decision rule in identifying resistant individuals in a different geographical population.

FIG. 4. **IgG responses to a signature of 15 selected antigens predicts protection from clinical malaria.** A, support vector machine (SVM) was trained on the antibody responses against the 15 *P. falciparum* antigens selected by sPLS-DA to discriminate between resistant and susceptible children. The SVM achieved a classification accuracy of 86%. The performance was evaluated by Monte Carlo cross-validation which randomly selected 30 samples for the training set and 42 for the testing set. The model was fit to the training data, and the predictive accuracy of the model in classifying samples as protected or susceptible was assessed using the testing set. This process was repeated 40 times (dashed black ROC curves). The solid black line represents the mean ROC curve. The dashed blue line represents an AUC = 0.5. B, The predictive performance of the SVM was significantly reduced when only young children were included (accuracy 69%) whereas it remained significantly higher when only older children (accuracy 83%) were included.

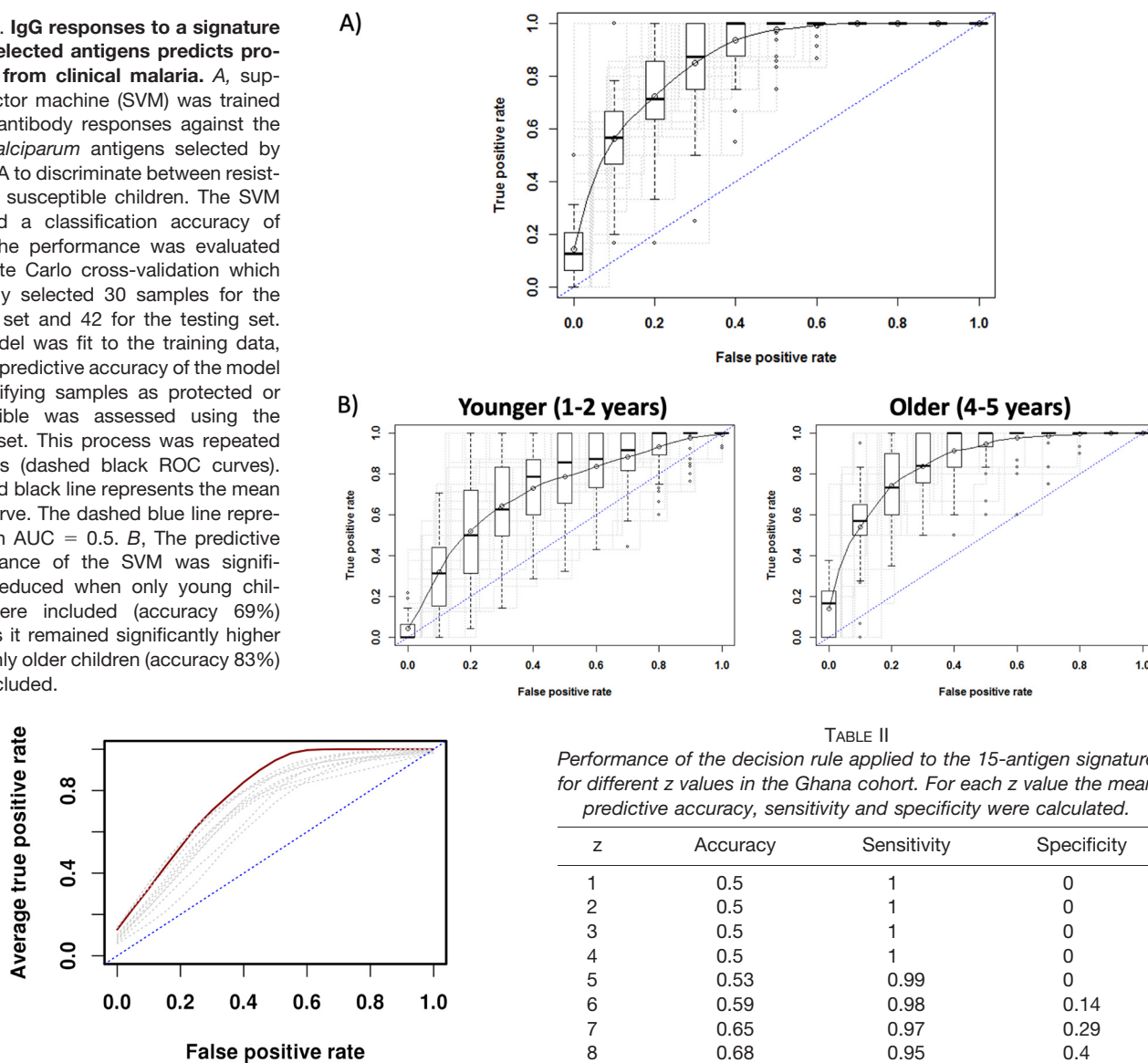


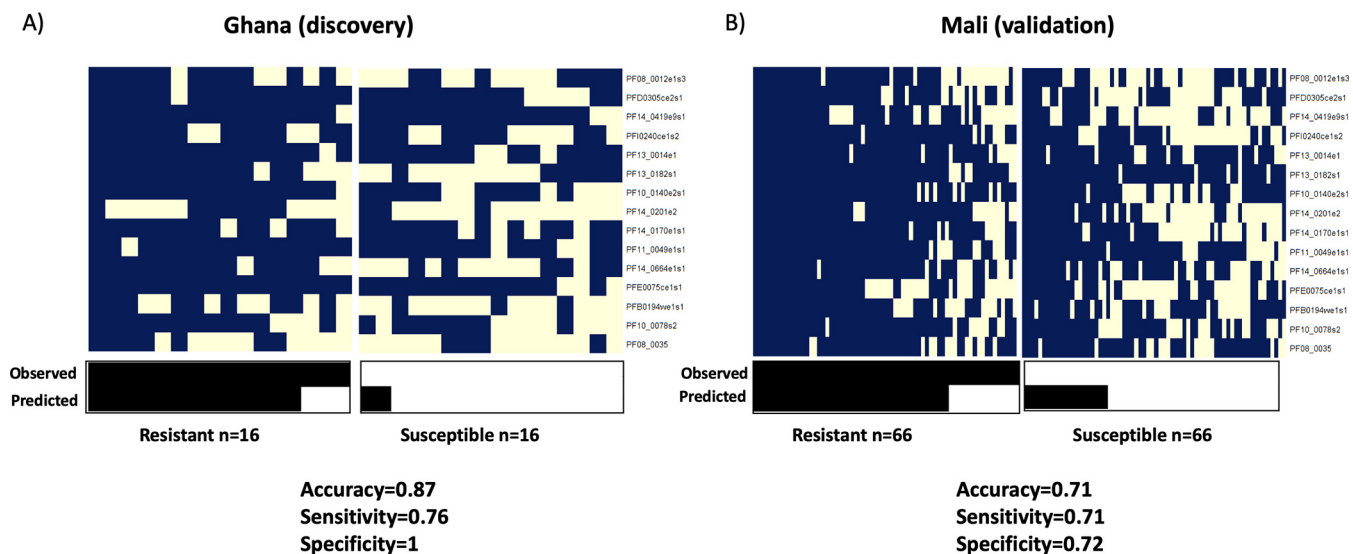
FIG. 5. **Validation of the 15-antigen signature in an independent cohort.** The 15-antigen signature discovered by sPLS-DA in Ghanaian children was validated in an independent cohort of 194 children aged 2–10 years old and adults 18–25 years old from Mali. First, an SVM was trained on the 15-antigen signature in the Mali cohort and validated by Monte Carlo cross-validation (red ROC curve). Next, 1000 sets of 15 antigens (out of 1080 antigens) were randomly selected and each time trained by SVM in the Mali cohort and validated by Monte Carlo cross-validation. The 15-antigen signature achieved a higher accuracy in 984 out of the 1000 random antigen sets ( $p = 0.016$ ) ROC curves for ten random antigen sets are shown by gray dashed lines. The dashed blue line represents an AUC = 0.5 (i.e. random guess of the individual's status).

**Conservation of the 15-antigen Signature**—For the purpose of developing vaccines or monitoring tools toward highly variable parasite, the use of conserved antigens across different species and strains is desired. We therefore determined the degree of sequence conservation of our signature among all

TABLE II  
Performance of the decision rule applied to the 15-antigen signature for different  $z$  values in the Ghana cohort. For each  $z$  value the mean predictive accuracy, sensitivity and specificity were calculated.

$z$	Accuracy	Sensitivity	Specificity
1	0.5	1	0
2	0.5	1	0
3	0.5	1	0
4	0.5	1	0
5	0.53	0.99	0
6	0.59	0.98	0.14
7	0.65	0.97	0.29
8	0.68	0.95	0.4
9	0.72	0.86	0.51
10	0.80	0.80	0.84
11	0.87	0.76	1
12	0.69	0.45	1
13	0.58	0.29	1
14	0.52	0.06	1
15	0.51	0.03	1

available *P. falciparum* strains ( $n = 15$ ) as well as the presence and degree of homology for orthologous genes in other human *Plasmodium* species (*vivax*, *knowlesi*, *ovale*, *malariae*) ( $n = 11$  strains). Our 15-antigen signature is ~96% conserved in all *P. falciparum* strains (average percentage of sequence identity across all strains). Nine antigens showed more than 90% of similarity with all (100%) the *P. falciparum* isolated analyzed (supplemental Fig. S7A, supplemental Table S3) and 10 antigens shown at least 50% of similarity against the majority of other human infected parasite species (>50%) (*P. vivax*, *P.*



**FIG. 6. Performance of the decision rule applied to the 15-antigen signature.** Performance of our decision rule applied to the 15-antigen signature in the discovery cohort of Ghanaian children (A) and the validation cohort of individuals from Mali (B). Signal intensities were dichotomized using antigen-specific thresholds and binary values of antigens are shown in blue (0/low) and cream (1/high). An individual was then predicted as resistant if at least 11 of the 15 antigens scored high. In the lower heat maps, black and cream represent resistant and susceptible individuals (observed and predicted by the decision rule).

*malariae*, *P. knowlesi*, *P. ovale*) (supplemental Fig. S7B, supplemental Table S3). Only one protein (PF3D7\_0406100) had a high sequence identity to the human proteome (>50% of identity with more than 50% of coverage) whereas the other 11 proteins showed null or very low degree of homology indicating a low propensity of induction of autoimmune or cross-reactivity (supplemental Table S3).

#### DISCUSSION

Advances in “omics” technologies and analytical platforms for complex data have enabled large-scale analyses of natural and vaccine-induced immune responses. Such efforts have revealed novel insights into immunity and correlates of protection. Prominent examples are the immune response to yellow fever vaccination and influenza vaccination (13, 14), and tuberculosis vaccination and infection (15). However, discerning protective immune responses to complex pathogens such as *Plasmodium* spp has proved challenging. Major issues arise from the complexity of the parasite which expressed ~5300 proteins, the complexity of the host immune response, the high dimensionality of data generated by high-throughput technologies, the often small sample sizes of malaria cohort studies, and the limitations of standard statistical tests in accounting for interactions between different immune responses (30, 32, 33). In this study, we established the utility of a predictive modeling framework, which combines feature selection and machine learning to identify a predictive immune signature of clinical immunity to malaria. We carried out a prospective study of antibody responses against a large panel of *P. falciparum* antigens in a cohort of young children in the process of acquiring clinical immunity to malaria. Chil-

dren were monitored longitudinally for one year to identify malaria-resistant and susceptible individuals. Using our predictive modeling framework, we showed that an individual’s immune status can be accurately predicted by measuring IgG responses against a small set of 15 defined parasite antigens. To determine the reproducibility of this potentially valuable clinical tool for assessing individual malaria risk, we assessed the signature in an independent cohort of Malian individuals between the ages of 2–10 years and 18–25 years. The signature was replicated in this independent cohort, thus supporting the generalizability of our signature to estimate an individual’s immune status in different epidemiological settings. Moreover, we showed that all of the antigens included in the 15-antigen signature were relatively stable over the course of the longitudinal study, suggesting that their predictive value would remain stable despite fluctuations in malaria transmission.

This study also highlights the complexity of individual immune profiles. Our results show that an individual’s immune profile may reflect either a combination of high antibody reactivity to a low number of antigens, or conversely, low antibody reactivity to a high number of antigens. Therefore, we speculate that high antibody responses against dominant antigens may suppress responses to other antigens. This observation has important implications for vaccine design because it suggests that malaria vaccines based on a single immunodominant antigen or a small number of highly reactive antigens may not be effective in delivering a sufficiently high level of protection. Rather, next-generation malaria vaccines will need to target a combination of many antigens to induce synergistic effects.

The proteins in our 15-antigen signature, identified based on their association with clinical immunity, represent very promising candidates for next-generation malaria vaccines. None of these antigens has yet been studied as a malaria vaccine candidate. Moreover, this specific combination of 15 antigens was a much better predictor of immunity than single antigen, indicating that a vaccine comprising all the 15 antigens identified in our study would most likely induce a synergistic immune response against the parasite. Consistent with the concept that an effective malaria vaccine will likely need to target multiple stages of the *Plasmodium* parasite's lifecycle, all stages of the parasite life cycle are represented by our 15 signature proteins (predominantly pre-erythrocytic stage, 7/15 antigens; and merozoite, 5/15 antigens). These antigens are also conserved among different *P. falciparum* strains and other human *Plasmodium* species making them very suitable for the purpose of developing vaccines or monitoring tools toward highly variable parasite.

One strength of our study is that subjects were regularly contacted every two months, ensuring that most, if not all, clinical malaria episodes were detected and recorded. However, the identification of susceptible children in this age group is problematic as the presence of asymptomatic malaria parasite carriers is common, the clinical signs of malaria are nonspecific, and parasitemia accompanied by a fever may not be sufficient to indicate an episode of clinical malaria (34). We, therefore, investigated the relationship between sensitivity and specificity by varying the parasite threshold in the clinical case definitions of malaria and assessing the impact on the performance of the 15-antigen signature. We found that the 15-antigen signature has an excellent sensitivity and specificity in predicting the immune status of an individual from a threshold of at least 2500 parasites/ $\mu$ l. For parasite thresholds below this cut-off, the sensitivity and specificity decreased as the parasite burden decreased. Because fevers in the presence of low parasitemia may be because of non-malarial causes, it is not surprising that our 15-antigen immune signature would have lower sensitivity and specificity as the parasitemia threshold in the case definition is lowered.

Protein microarrays have proven to be powerful tools for detecting serum antibodies against a vast repertoire of parasite proteins thus offering the opportunity to correlate protective immunity with seroreactivity profiles. Classical univariate approaches to data analysis suggest that a single parasite antigen is probably not enough to elicit (predict) protective immunity, as demonstrated in different epidemiological settings (2–7). These approaches have the disadvantage of considering each variable independently from the others so that the interactions and potential synergies of multi-faceted immune responses are not considered (7, 33). This is an important drawback for the identification of correlates of protection in malaria and the responses to a specific antigen can be associated with protection in some studies or with increasing risk in others (35).

Recent studies have evaluated responses to multiple antigens simultaneously and have suggested that specific antigen combinations may be associated with immunity in field settings (4, 5) and laboratory studies (36). Other studies have shown that a combination of antigens targeting the same stage of the parasite lifecycle does not necessarily produce additive or synergistic protection (37) whereas a more effective combination of antigens may target multiple discrete stages of, for example, invasion (7). These studies represent an important proof-of-concept for the identification of synergistic combinations of antigens, but as they have only included a low number (<100) of antigens, most possible antigen combinations in the *P. falciparum* genome remained unexplored. How to advance beyond those earlier studies has not been obvious. In this study we employed a predictive modeling framework that allowed us to explore antibody responses to 1080 *P. falciparum* proteins. In the computational sciences, machine-learning methods were introduced because of their power to infer associations between correlated features that could not otherwise be made using conventional statistical methodologies (which usually require variable independence) (38–40). Support Vector Machines (SVMs) are now widely applied to transcriptomic, proteomic, and metabolomic datasets to develop predictive models for effective, sensitive and specific decision making. Although SVMs can achieve good predictive accuracy, the excess of features in the training datasets of omics experiments increases both the risk of overfitting and the prediction variability. The selection of a restricted number of variables related to the intended outcome before sample classification eliminates non-informative and correlating variables, thereby increasing the generalization ability of the classifier.

We therefore incorporated a features selection step before building a SVM-based classifier to reduce dimensionality and to make the model less complex and easier to interpret, avoid overfitting and consequently providing a more robust model. Features were selected by sparse partial least squares discriminant analysis (sPLS-DA), which has proven to be particularly appropriate for data with small sample sizes and a large number of correlated variables (25). Although our results show that the signature identified with our approach performs well in an area of intense malaria transmission where clinical disease is generally associated with children below 5 years of age, additional validation in area of lower transmission where many cases occur also in older children and adults is needed.

In summary, we have demonstrated the utility of a predictive modeling framework to decode the host antibody responses to the human malaria parasite *P. falciparum*, in order to identify a specific antigen signature of NAI. In this comprehensive analysis, we have demonstrated that a signature of only a few antigens can be used to predict the immune status at an individual level. Specifically, we have shown that the simultaneous detection of antibody responses to a specific set of 15 antigens is sufficient to discriminate clinically resist-

ant from susceptible children with a very high degree of accuracy and we further validated this in an independent and geographically and demographically distinct cohort of individuals naturally exposed to malaria. We also developed a simple decision rule based on the 15- antigen signature that could be potentially used for seroepidemiological surveillance.

Further evaluation of this signature in multicohort studies would support the utility of this 15-antigen signature as a tool for identifying individuals at high risk of clinical malaria and for monitoring the acquisition and maintenance of NAI. This study provides proof-of-concept for the power of applying advanced analytical approaches such as feature selections and machine learning algorithms to identify a predictive signature of clinical immunity. It paves the way for the development of a robust point-of-care test to identify individuals at high risk of disease and monitor the impact of vaccinations and other malaria interventions.

**Acknowledgments**—The participation of the parents and children of the Kassena Nankana District and the Navrongo Health Research Centre (NHRC) in the conduct of the original study is very much appreciated. We are also grateful to the individuals in Mali who participated in this study, and their families.

\* This work was supported by National Health and Medical Research Council of Australia. D.L.D. was supported by a NHMRC Principal Research Fellowship. The original samples were acquired as part of the studies undertaken under the Malaria Clinical Epidemiological Studies with support from NIAID under contract to NMIMR (Contract No: HSSN266200400016C; PI, K.A.K.). The work in Mali and P.D.C. are supported by the Division of Intramural Research, National Institute of Allergy and Infectious Diseases, National Institutes of Health. The authors declare that they have no conflicts of interest with the contents of this article.

☐ This article contains [supplemental material](#).

✉ To whom correspondence should be addressed: Centre for Molecular Therapeutics, Australian Institute of Tropical Health and Medicine, James Cook University, Cairns, QLD, Australia. E-mail: denise.doolan@jcu.edu.au.

||| These authors contributed equally to this work.

Author contributions: C.P., A.T., D.D., B.G., K.A.K., and W.O.R. performed research; C.P., L.K., A.T., K.A.K., W.O.R., T.L.R., P.C., P.F., and D.L.D. analyzed data; C.P., L.K., A.T., D.D., B.G., K.A.K., W.O.R., T.L.R., P.C., P.F., and D.L.D. wrote the paper; K.A.K., W.O.R., P.L.F., and D.L.D. designed research.

### REFERENCES

- Cohen, S., McGregor, I. A., and Carrington, S. (1961) Gamma-globulin and acquired immunity to human malaria. *Nature* **192**, 733–737
- Osier, F. H., Fegan, G., Polley, S. D., Murungi, L., Verra, F., Tetteh, K. K., Lowe, B., Mwangi, T., Bull, P. C., Thomas, A. W., Cavanagh, D. R., McBride, J. S., Lanar, D. E., Mackinnon, M. J., Conway, D. J., and Marsh, K. (2008) Breadth and magnitude of antibody responses to multiple *Plasmodium falciparum* merozoite antigens are associated with protection from clinical malaria. *Infect. Immun.* **76**, 2240–2248
- Trieu, A., Kayala, M. A., Burk, C., Molina, D. M., Freilich, D. A., Richie, T. L., Baldi, P., Felgner, P. L., and Doolan, D. L. (2011) Sterile protective immunity to malaria is associated with a panel of novel *P. falciparum* antigens. *Mol. Cell. Proteomics* **10**, M111.007948
- Osier, F. H., Mackinnon, M. J., Crosnier, C., Fegan, G., Kamuyu, G., Wanaguru, M., Ogada, E., McDade, B., Rayner, J. C., Wright, G. J., and Marsh, K. (2014) New antigens for a multicomponent blood-stage malaria vaccine. *Sci. Transl. Med.* **6**, 247ra102
- Daou, M., Kouriba, B., Ouedraogo, N., Diarra, I., Arama, C., Keita, Y., Sissoko, S., Ouologuem, B., Arama, S., Bousema, T., Doumbo, O. K., Sauerwein, R. W., and Scholzen, A. (2015) Protection of Malian children from clinical malaria is associated with recognition of multiple antigens. *Malar. J.* **14**, 56
- Proietti, C., Zakrzewski, M., Watkins, T. S., Berger, B., Hasan, S., Ratnatunga, C. N., Brion, M. J., Crompton, P. D., Miles, J. J., Doolan, D. L., and Krause, L. (2016) Mining, visualizing and comparing multidimensional biomolecular data using the Genomics Data Miner (GMine) Web-Server. *Sci. Rep.* **6**, 38178
- Bustamante, L. Y., Powell, G. T., Lin, Y. C., Macklin, M. D., Cross, N., Kemp, A., Cawkill, P., Sanderson, T., Crosnier, C., Muller-Sienerth, N., Doumbo, O. K., Traore, B., Crompton, P. D., Cicuta, P., Tran, T. M., Wright, G. J., and Rayner, J. C. (2017) Synergistic malaria vaccine combinations identified by systematic antigen screening. *Proc. Natl. Acad. Sci. U.S.A.* **114**, 12045–12050
- World Health Organization. (2017) Malaria Vaccine Rainbow Tables.
- Doolan, D. L., Mu, Y., Unal, B., Sundaresh, S., Hirst, S., Valdez, C., Randall, A., Molina, D., Liang, X., Freilich, D. A., Oloo, J. A., Blair, P. L., Aguiar, J. C., Baldi, P., Davies, D. H., and Felgner, P. L. (2008) Profiling humoral immune responses to *P. falciparum* infection with protein microarrays. *Proteomics* **8**, 4680–4694
- Crompton, P. D., Kayala, M. A., Traore, B., Kayentao, K., Ongoiba, A., Weiss, G. E., Molina, D. M., Burk, C. R., Waisberg, M., Jasinskas, A., Tan, X., Doumbo, S., Doumtabe, D., Kone, Y., Narum, D. L., Liang, X., Doumbo, O. K., Miller, L. H., Doolan, D. L., Baldi, P., Felgner, P. L., and Pierce, S. K. (2010) A prospective analysis of the Ab response to *Plasmodium falciparum* before and after a malaria season by protein microarray. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 6958–6963
- Barry, A. E., Trieu, A., Fowkes, F. J., Pablo, J., Kalantari-Dehaghi, M., Jasinskas, A., Tan, X., Kayala, M. A., Tavul, L., Siba, P. M., Day, K. P., Baldi, P., Felgner, P. L., and Doolan, D. L. (2011) The stability and complexity of antibody responses to the major surface antigen of *Plasmodium falciparum* are associated with age in a malaria endemic area. *Mol. Cell. Proteomics* **10**, M111.008326
- Proietti, C., and Doolan, D. L. (2014) The case for a rational genome-based vaccine against malaria. *Front. Microbiol.* **5**, 741
- Gaucher, D., Therrien, R., Kettaf, N., Angermann, B. R., Boucher, G., Filali-Mouhim, A., Moser, J. M., Mehta, R. S., Drake, D. R., 3rd, Castro, E., Akondy, R., Rinfret, A., Yassine-Diab, B., Said, E. A., Chouikh, Y., Cameron, M. J., Clum, R., Kelvin, D., Somogyi, R., Greller, L. D., Balderas, R. S., Wilkinson, P., Pantaleo, G., Tartaglia, J., Haddad, E. K., and Sekaly, R. P. (2008) Yellow fever vaccine induces integrated multilineage and polyfunctional immune responses. *J. Exp. Med.* **205**, 3119–3131
- Trombetta, C. M., and Montomoli, E. (2016) Influenza immunology evaluation and correlates of protection: a focus on vaccines. *Expert Rev. Vaccines* **15**, 967–976
- Haks, M. C., Bottazzi, B., Cecchinato, V., De Gregorio, C., Del Giudice, G., Kaufmann, S. H. E., Lanzavecchia, A., Lewis, D. J. M., Maertzdorf, J., Mantovani, A., Sallusto, F., Sironi, M., Uguccioni, M., and Ottenhoff, T. H. M. (2017) Molecular signatures of immunity and immunogenicity in infection and vaccination. *Front. Immunol.* **8**, 1563
- Helb, D. A., Tetteh, K. K., Felgner, P. L., Skinner, J., Hubbard, A., Arinaitwe, E., Mayanja-Kizza, H., Ssewanyana, I., Kanya, M. R., Beeson, J. G., Tappero, J., Smith, D. L., Crompton, P. D., Rosenthal, P. J., Dorsey, G., Drakeley, C. J., and Greenhouse, B. (2015) Novel serologic biomarkers provide accurate estimates of recent *Plasmodium falciparum* exposure for individuals and communities. *Proc. Natl. Acad. Sci. U.S.A.* **112**, E4438–E4447
- Appawu, M., Owusu-Agyei, S., Dadzie, S., Asoala, V., Anto, F., Koram, K., Rogers, W., Nkrumah, F., Hoffman, S. L., and Fryauff, D. J. (2004) Malaria transmission dynamics at a site in northern Ghana proposed for testing malaria vaccines. *Trop. Med. Int. Health* **9**, 164–170
- Koram, K. A., Owusu-Agyei, S., Fryauff, D. J., Anto, F., Atuguba, F., Hodgson, A., Hoffman, S. L., and Nkrumah, F. K. (2003) Seasonal profiles of malaria infection, anaemia, and bednet use among age groups and communities in northern Ghana. *Trop. Med. Int. Health* **8**, 793–802
- Baird, J. K., Owusu Agyei, S., Utz, G. C., Koram, K., Barcus, M. J., Jones, T. R., Fryauff, D. J., Binka, F. N., Hoffman, S. L., and Nkrumah, F. N. (2002) Seasonal malaria attack rates in infants and young children in northern Ghana. *Am. J. Trop. Med. Hyg.* **66**, 280–286

20. Oduro, A. R., Koram, K. A., Rogers, W., Atuguba, F., Ansah, P., Anyorigiya, T., Ansah, A., Anto, F., Mensah, N., Hodgson, A., and Nkrumah, F. (2007) Severe falciparum malaria in young children of the Kassena-Nankana district of northern Ghana. *Malar. J.* **6**, 96
21. Dodoo, D., Atuguba, F., Bosomprah, S., Ansah, N. A., Ansah, P., Lamptey, H., Egyir, B., Oduro, A. R., Gyan, B., Hodgson, A., and Koram, K. A. (2011) Antibody levels to multiple malaria vaccine candidate antigens in relation to clinical malaria episodes in children in the Kasena-Nankana district of Northern Ghana. *Malar. J.* **10**, 108
22. Dodoo, D., Hollingdale, M. R., Anum, D., Koram, K. A., Gyan, B., Akanmori, B. D., Ocran, J., Adu-Amankwah, S., Geneshan, H., Abot, E., Legano, J., Banania, G., Sayo, R., Brambilla, D., Kumar, S., Doolan, D. L., Rogers, W. O., Epstein, J., Richie, T. L., and Sedegah, M. (2011) Measuring naturally acquired immune responses to candidate malaria vaccine antigens in Ghanaian adults. *Malar. J.* **10**, 168
23. Osafo-Addo, A. D., Koram, K. A., Oduro, A. R., Wilson, M., Hodgson, A., and Rogers, W. O. (2008) HLA-DRB1\*04 allele is associated with severe malaria in northern Ghana. *Am. J. Trop. Med. Hyg.* **78**, 251–255
24. Bejon, P., Warimwe, G., Mackintosh, C. L., Mackinnon, M. J., Kinyanjui, S. M., Musyoki, J. N., Bull, P. C., and Marsh, K. (2009) Analysis of immunity to febrile malaria in children that distinguishes immunity from lack of exposure. *Infect. Immun.* **77**, 1917–1923
25. Le Cao, K. A., Boitard, S., and Besse, P. (2011) Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems. *BMC Bioinformatics* **12**, 253
26. Rohart, F., Gautier, B., Singh, A., and Le Cao, K. A. (2017) mixOmics: An R package for 'omics feature selection and multiple data integration. *PLoS Comput. Biol.* **13**, e1005752
27. Bates, D., Mächler, M., Bolker, B. M., and Walker, S. C. (2015) Fitting Linear Mixed-Effects Models Using lme4. *J. Statistical Software* **67**, 1–48
28. Pan, J., Song, G., Chen, D., Li, Y., Liu, S., Hu, S., Rosa, C., Eichinger, D., Pino, I., Zhu, H., Qian, J., and Huang, Y. (2017) Identification of Serological Biomarkers for Early Diagnosis of Lung Cancer Using a Protein Array-Based Approach. *Mol. Cell. Proteomics* **16**, 2069–2078
29. Dodoo, D., Aikins, A., Kusi, K. A., Lamptey, H., Remarque, E., Milligan, P., Bosomprah, S., Chilengi, R., Osei, Y. D., Akanmori, B. D., and Theisen, M. (2008) Cohort study of the association of antibody levels to AMA1, MSP119, MSP3 and GLURP with protection from clinical malaria in Ghanaian children. *Malar. J.* **7**, 142
30. Valletta, J. J., and Recker, M. (2017) Identification of immune signatures predictive of clinical protection from malaria. *PLoS Comput. Biol.* **13**, e1005812
31. Kangoye, D. T., Mensah, V. A., Murungi, L. M., Nkumama, I., Nebie, I., Marsh, K., Cisse, B., Bejon, P., Osier, F. H., Sirima, S. B., and Group, M. I. I. S. (2016) Dynamics and role of antibodies to Plasmodium falciparum merozoite antigens in children living in two settings with differing malaria transmission intensity. *Vaccine* **34**, 160–166
32. Marengo, E., and Robotti, E. (2014) Biomarkers for pancreatic cancer: recent achievements in proteomics and genomics through classical and multivariate statistical methods. *World J. Gastroenterol.* **20**, 13325–13342
33. Robotti, E. M., and Marengo, M. E. (2014) Biomarkers discovery through multivariate statistical methods: a review of recently developed methods and applications in proteomics. *J. Proteomics Bioinform.* **S3**
34. Mabunda, S., Aponte, J. J., Tiago, A., and Alonso, P. (2009) A country-wide malaria survey in Mozambique. II. Malaria attributable proportion of fever and establishment of malaria case definition in children across different epidemiological settings. *Malar. J.* **8**, 74
35. Fowkes, F. J., Richards, J. S., Simpson, J. A., and Beeson, J. G. (2010) The relationship between anti-merozoite antibodies and incidence of Plasmodium falciparum malaria: A systematic review and meta-analysis. *PLoS Med.* **7**, e1000218
36. Schussek, S., Trieu, A., Apte, S. H., Sidney, J., Sette, A., and Doolan, D. L. (2017) Novel Plasmodium antigens identified via genome-based antibody screen induce protection associated with polyfunctional T cell responses. *Sci. Rep.* **7**, 15053
37. Greenhouse, B., Ho, B., Hubbard, A., Njama-Meya, D., Narum, D. L., Lanar, D. E., Dutta, S., Rosenthal, P. J., Dorsey, G., and John, C. C. (2011) Antibodies to Plasmodium falciparum antigens predict a higher risk of malaria but protection from symptoms once parasitemic. *J. Infect. Dis.* **204**, 19–26
38. Sentelle, C. G., Anagnostopoulos, G. C., and Georgiopoulos, M. (2016) A Simple Method for Solving the SVM Regularization Path for Semidefinite Kernels. *IEEE Trans. Neural. Netw. Learn. Syst.* **27**, 709–722
39. Kim, H., Lee, S., Lee, D., Choi, S., Ju, J., and Myung, H. (2015) Real-time human pose estimation and gesture recognition from depth images using superpixels and SVM classifier. *Sensors* **15**, 12410–12427
40. Chen, H. L., Yang, B., Wang, G., Wang, S. J., Liu, J., and Liu, D. Y. (2012) Support vector machine based diagnostic system for breast cancer using swarm intelligence. *J. Med. Syst.* **36**, 2505–2519