There is no vaccine to protect from cryptosporidiosis, a leading cause of diarrhea in infants in low and middle income countries. Here we comprehensively identified parasite antigens associated with protection from reinfection. A Cryptosporidium protein microarray was constructed by in vitro transcription and translation of 1761 C. parvum, C. hominis or C. meleagridis antigens, including proteins with a signal peptide and/or a transmembrane domain. Plasma IgG and/or IgA from Bangladeshi children longitudinally followed for cryptosporidiosis from birth to three years of age, identified 233 seroreactive proteins. Seven of these were associated with protection from reinfection. These included Cp23 and Cp17, Gp900 and four additional antigens (CpSMP1, CpMuc8, CpCorA and CpCCDC1). Infection in the first year of life however often resulted in no detectable antigen-specific antibody response, and antibody responses, when detected, were (i) specific to the infecting parasite genotype, and (ii) decayed in the months post-infection. In conclusion humoral immune responses against specific parasite antigens were associated with acquired immunity. While antibody decay over time and parasite genotype-specificity may limit natural immunity, this work serves as a foundation for antigen selection for vaccine design.
Specific Cryptosporidium antigens associate with reinfection immunity and protection from cryptosporidiosis

Carol A. Gilchrist*, Joseph J. Campo*, Jozelyn V. Pablo, Jennie Z. Ma, Andy Teng, Amit Oberai, Adam D. Shandling, Masud Alam, Mamun Kabir, ASG Faruque, Rashidul Haque, William A. Petri Jr*

Department of Medicine, Public Health Science, Microbiology, Immunology and Cancer Biology, Pathology, University of Virginia, Charlottesville, Virginia, United States of America, Antigen Discovery Inc, Irvine, California, United States of America, and International Centre for Diarrhoeal Diseases Research, Bangladesh

*Co-first authors

#Corresponding Author: William A Petri  wap3g@virginia.edu

345 Crispell Ave, Carter Harrison Building (MR6) Rm1709A, Charlottesville, VA 22908

Conflict-of-interest statement: WP is a consultant for TechLab, Inc., a maker of diagnostic tests for cryptosporidiosis. The other authors have declared that no conflict of interest exists.

Brief Summary

Identification of Cryptosporidium proteins recognized by the humoral immune system of children and associated with immunity to reinfection provides a rational approach to vaccine development.

Abstract
There is no vaccine to protect from cryptosporidiosis, a leading cause of diarrhea in infants in low and middle income countries. Here we comprehensively identified parasite antigens associated with protection from reinfection. A Cryptosporidium protein microarray was constructed by in vitro transcription and translation of 1761 C. parvum, C. hominis or C. meleagridis antigens, including proteins with a signal peptide and/or a transmembrane domain. Plasma IgG and/or IgA from Bangladeshi children longitudinally followed for cryptosporidiosis from birth to three years of age, identified 233 seroreactive proteins. Seven of these were associated with protection from reinfection. These included Cp23 and Cp17, Gp900 and four additional antigens (CpSMP1, CpMuc8, CpCorA and CpCCDC1). Infection in the first year of life however often resulted in no detectable antigen-specific antibody response, and antibody responses, when detected, were (i) specific to the infecting parasite genotype, and (ii) decayed in the months post-infection. In conclusion humoral immune responses against specific parasite antigens were associated with acquired immunity. While antibody decay over time and parasite genotype-specificity may limit natural immunity, this work serves as a foundation for antigen selection for vaccine design.
Introduction

Cryptosporidium parasites are a common cause of diarrhea in infants and children in low and middle income countries, globally are the most common cause of outbreaks of waterborne disease, and in immunocompromised individuals can cause protracted infections (1). Evidence that a vaccine is an achievable goal includes that the incidence of infection declines with age (2–4) and that antibodies against the surface of the infectious oocyst (5) and invasive sporozoite parasite life stages are associated with partial immunity to reinfection (5,9). While the associated protective immune response involves both cellular and humoral immunity, identification of antibody responses against Cryptosporidium antigens associated with protection from reinfection may provide a foundation for the development of an effective anti-Cryptosporidium vaccine.

We had previously conducted a natural history study of cryptosporidiosis in infants longitudinally followed for infection from birth to three years of age (National Clinical Trial Identifier: NCT02764918) (6–9). The infants resided in a low income community of Dhaka, Bangladesh. Fecal DNA was extracted from monthly surveillance and diarrheal stool samples and symptomatic and asymptomatic cases of cryptosporidiosis were identified using a pan-Cryptosporidium qPCR assay. By the end of year one, 27.5% (n=109) of infants had been infected at least once. The immunity arising from this infection was however incomplete with most of these children reinfected by three years of age (Supplemental Figure 1) (10). These repeat infections however had a lower burden of parasites and were more likely to be sub-clinical in nature.
Here we present the results from probing a *Cryptosporidium* protein microarray with antisera from one year old infants. The association of antibody responses against specific proteins with immunity to reinfection in years 2 and 3 of life was then tested.
Results

Identification of Immunoreactive Proteins

The *Cryptosporidium* spp. protein microarray created for this study comprised a total of 1761 antigens representing 1250 unique genes from *C. parvum* (n=980), *C. hominis* (n=263) and *C. meleagridis* (n=7). *C. parvum* sequences were used as the backbone of the protein microarray due to the ready availability of *C. parvum* DNA and its superior assembly and annotation (11, 12) (Supplemental Figure 2). The selected *C. parvum* proteins included those which had been previously identified as potential vaccine candidates (Supplemental Table 1) (13–33). Open reading frames (ORFs) >3000 base pairs were cloned as overlapping segments to optimize in vitro translation. The array also included 15 genetically variant regions of the *gp60* gene common in this Bangladeshi population (9). Proteins with conserved sequences in the different *Cryptosporidium* species that account for the majority of cryptosporidiosis in humans (*C. hominis, C. parvum* and *C. meleagridis*), and that were annotated as having a signal peptide, and thus potentially a membrane protein and accessible to human antibodies, were prioritized for inclusion (Supplemental Figure 3).

The array was incubated with a 1:100 dilution of plasma collected at one year of age from 500 children in the cohort, and developed with anti-human IgG (DyLight650, Bethyl Laboratories) and Cy3 AffiniPure F(ab')₂ or anti-human IgA. The distribution of normalized fluorescence signal intensity (SI) values of each antigen was analyzed using a mixture modeling technique Supplemental Figure 4A) to identify the antigen specific background component of the spot signal and hence the appropriate intensity (cut-off value) that was defined as seropositive (Supplemental Figure 4B).
Antigens were classified as seroreactive in this population if ≥ 10% of the children had either IgA (Figure 1A) or IgG (Figure 1B) antibodies against the antigen. Using this criterion, 36 antigens were seroreactive to both IgG and IgA (Supplemental Figure 5), 57 antigens by IgA alone and 140 antigens by IgG alone (Figure 2). Antigens recognized varied greatly between children, with each antigen recognized by only a subset of the responding infants (Figure 2).

We found that orthologues encoded by *C. parvum* and *C. hominis* generated similar signals; for example the *C. parvum* Cp23 (cgd4_3620) and the *C. hominis* Cp23 orthologue (Chro.40414) signals were correlated (Pearson r value 0.844 \(p=2.11\times10^{-119}\)). In total there were 124 IgG-reactive proteins and 70 IgA-reactive *C. parvum* proteins. Antigens recognized were from multiple developmental stages of the *Cryptosporidium* parasite (34-38, 40) (Supplemental Figure 2).

**Humoral immune response diminished with time from infection**

The anti-*Cryptosporidium* IgA and IgG antibody profiles were analyzed using t-distributed stochastic neighbor embedding (t-SNE) as an unsupervised data reduction method for visualization of the trends in the antibody profile, based on the number of days since the first *Cryptosporidium* positive diagnostic qPCR and whether the child had a documented prior *Cryptosporidium* infection (Figure 3A). The antibody profile of children with more recent infections mapped to a distinct region within the t-SNE plot (Region 1, “R1”), while children with prior infection did not cluster separately from children without a prior infection (Figure 3A). By analyzing the most responsive antigens (top 100), the important factors which influenced the antibody profile of children in R1 of the t-SNE plot were the strength of the immune response (Figure 3B) and its breadth (the number of parasite antigens recognized by IgA and IgG) (Figure 3C)(39). Diminishing antibody responses over time were confirmed by a linear regression
analysis where most notably, the breadth of both the IgA and IgG anti-
Cryptosporidium immune responses significantly decreased over time (Figure 3D).

Impact of prior infection on antibody response

While the effect of time since Cryptosporidium infection on antibody levels was significant, it did not completely explain the failure to generate an anti-Cryptosporidium immune response in all cases. To focus on the impact of prior exposure to the Cryptosporidium parasite, the data were analyzed using partial least squares discriminant analysis (PLS-DA) (Figure 3E and F). This analysis demonstrated that a subset of children with prior infection differed from the population of children with no prior infection but that a substantial proportion of children with prior infection had antibody profiles similar to children without a documented prior infection. We concluded that some infections in this cohort were missed despite the active surveillance system in place.

To explore whether malnutrition or inflammation impacted the humoral immune response, we examined whether the immune profile correlated with growth failure (a measure of chronic malnutrition measured by child height-for-age Z scores (HAZ)) or biomarkers of systemic and local inflammation (sCD14, IL-1Beta, CRP) or immunoregulatory cytokines (IL-4), but no correlations were observed (Supplemental Figure 6).

Protection from reinfection was not associated with the breadth of the antibody response (i.e., number of antigens recognized by a given child)

Children for whom a previous infection had been identified had a greater number of parasite antigens recognized by IgA and IgG (greater “breadth”) (Figure 4A). However, no association was found between the breadth of the antibody response and resistance to reinfection using
either a data set restricted to the infants with a qPCR verified Cryptosporidium infection prior to year one (Figure 4B) or using the data from the entire study cohort (Figure 4C).

Validation of the array by examining the data obtained from antigens previously associated with a protective immune response

* C. hominis* Cp23 (Chro.40414) and Cp17, a conserved peptide encoded by the variable *C. hominis* gp60 gene (Chro.601380: variant laA25R3), are both potential vaccine candidates and previously shown to be associated with a delay in reinfection in our study population (7, 8). We investigated whether we could also detect an association with protection from reinfection between IgA antibodies recognizing *C. hominis* Cp23 (Chro.40414) and the *C. hominis* gp60 (Chro.601380) antigens on the Cryptosporidium array. As the IgA anti-Cp23 (Chro.40414) signal was low in our array data we were only able to analyze the anti-IgG Cp23 (Chro.40414) data. As expected, a protective association was observed between both anti-IgA and IgG *C. hominis* Gp60 (Chro.601380) (Figure 5 A-D) and anti-IgG *C. hominis* Cp23 (Chro.40414) (Figure 5 E and F). (Since Gp60 and Cp23 were a priori antigen candidates, p-values were not adjusted for the false discovery rate).

Impact of the polymorphisms in the gp60 gene on immune reactivity

The protein encoded by the *gp60* gene is processed by the parasite into Gp40 and Gp15 proteins (Figure 6A). The region of the *gp60* gene that encodes the Gp40 protein has three variable domains: a SNP-based allelic family “type”; a variable number of trinucleotide repeats “subtypes”; and a repeat sequence “R” (34). In the Bangladeshi infant cohort 15 different variants of *gp60* were identified (2 in *C. parvum* and 13 in *C. hominis*) (Figure 6A) all of which were included in the protein array (9). The *gp60* genotype of the infecting *Cryptosporidium* parasite was known in a subset of cases, and the data from the plasma collected from these children was examined to
see if an allele-specific immune signal could be observed. With one exception, the infecting *C. hominis* genotype matched the Gp40 antigen variant recognized by the child’s plasma (Figures 6A and 6B, Supplemental Figure 7). The number of trinucleotide repeats in the *gp60* gene however did not impact antibody recognition: children infected with the IaA18R3 subtype bound equivalently to the Ia antigens on the array which had different numbers of trinucleotide repeats (IaA27R3, IaA26R3, IaA25R3, IaA22R3, IaA19R3 and IaA18R3 (Figure 6A and B). We concluded that humoral immunity to the variable Gp40 antigen was genotype-specific.

Antigens associated with protection from reinfection

We tested if a delay in the time to reinfection was associated with the development of anti-*Cryptosporidium* antibodies against specific antigens. This analysis was done for the children on follow-up to ages 2 and 3 years who had qPCR-verified cryptosporidiosis during the first year of life (Table 1; Supplemental Figure 8). The analysis was also performed including all the children in the cohort (Supplemental Table 2; Supplemental Figure 9). To minimize false discoveries, as well as false exclusions, a feature selection antigen filtering step was employed using random forest ("RF") models on survival data. For the RF models, the children were stratified into seropositive vs seronegative for each of the 233 IgA- and/or IgG-reactive antigens, and variables that were important to the models over 100 iterations were identified (Figure 7A-B).

Among antigens with an average variable importance metric (VIMP) greater than one standard deviation above the mean of all antigen VIMP scores and with positive VIMP scores (i.e. important to the model) in at least 80% of iterations, 7 antigens in addition to Gp60 and Cp23 had hazard ratios less than 1 (protective) in all four modeling groups (Figure 7C). Additional RF comparisons are shown in Supplemental Figure 8) and are included in Table 1 along with the adjustment for
the number of antigens tested. These were selected for evaluation in Cox proportional hazards models (Figure 7D-E) with adjustment for the false discovery rate. In addition to the Gp60 and Cp23 antigens, a significant association with protection from cryptosporidiosis was observed for antibodies against the Gp900 mucin (cgd7_4020), the potential mucin CpMuc8 (cgd8_700), the putative metal ion transporter CpCorA (cgd2_1520), a small membrane protein (Chro.30111) and the coiled coil domain protein CpCCDC (cgd8_830) (additional RF comparisons are shown in Supplemental Figure 8). Parenthetically, cgd8_830 seropositivity was associated with significantly lower incidence of infection, particularly during the first year of follow-up post-sampling, but was found to be more abundant in children that ultimately were infected at the end of follow-up. Likewise, Chro.30111 antibody responses showed evidence of protection during follow-up and at the end of the first year of follow-up post-sampling, but not at the end of 2 years of follow-up.

A PLS-DA regression model was then used to evaluate the relative contribution of the selected antigens in defining the latent components (“loading weights”) that maximize discrimination of children by infection status (Figure 8A). The end point metric was complete protection from reinfection associated with antibody levels. The analysis was performed at age two (Figure 8B) and age three (Figure 8C). The contribution of each antibody to the PLS-DA profile at age two (Figure 8D) and three years (Figure 8E) is also shown.

**Discussion**

The most important discovery from this work is the identification of seven cryptosporidial antigens to which a humoral immune response is associated with protection from reinfection. These included previously identified vaccine candidates, Cp23 and Cp17 proteins (7, 8, 10, 35, 36) the Gp900 mucin (15, 25, 30) as well as CpMuc8 (a potential mucin); CpSMP1 (small membrane protein); a coiled-coil protein CpCCDC, and a potential metal transporter CpCorA.
The developmental stages of *Cryptosporidium* (37) include extracellular and intracellular forms that differ in protein repertoire (38, 39). We considered it a possibility that only antigens from proteins expressed at specific stages in the parasite’s lifecycle were protective. Antibodies that target the early extracellular life stages in apicomplexans may be more effective at preventing infection (40). However in our case, the protective antigens were present in more than one life-stage as assessed by mRNA transcripts (38, 39, 41). One drawback of using transcriptomics data is that post-transcriptional regulation may influence protein expression. We therefore examined the available proteomic data and found peptides from four of the seven candidates in the sporozoites/oocyst proteome (39, 42). In summary, the protective antigens are not all highly expressed in sporozoites. The available information suggests that protective antigens may not be restricted to the cryptosporidial sporozoite (39, 42).

While plasma IgG anti-Cp23 and Cp17 had not been previously observed to be protective in our cohort (7, 8), reanalysis of the IgG data in our study suggests that our use of the mixed models to determine the boundary of the positive response in the protein array data improved the specificity of the immunoassay results and accounted for the apparent difference in the study conclusions (Supplemental Figure 4) (44).

The association observed between antibody levels and time since detection of *Cryptosporidium* infection suggests that anti-*Cryptosporidium* antibodies may be short-lived in this young age group. It is possible that multiple repeat infections are needed to generate a more durable antibody response (45, 46). In addition a substantial proportion of the children with prior infection had largely undetectable antibody levels, similar to the larger population of immunologically naïve children. Failure to generate robust humoral immunity in infants has also been observed with infection with the apicomplexan parasite *Plasmodium*, with immunity to symptomatic disease developing only after repeated exposures to the pathogen (47). Also in *Plasmodium* a short-lived
non-sterile immunity is common in infants, and in this case may involve defects in antibody affinity
maturation in the host germinal centers (48–50). Whether this may also play a role in depressing
the antibody response to the Cryptosporidium parasite, and if this can be remedied by an
appropriate vaccination strategy remains to be discovered.

The immune mechanisms involved in the control of the Cryptosporidium parasite likely involve
both the innate and adaptive arms of the immune system (1). The adaptive immune response
consists of both cellular and humoral immunity. Pre-existing anti-Cryptosporidium IgG was shown
to be associated with immunity in this work and earlier in experimentally infected adults (51). The
importance of cellular immunity and in particular, interferon gamma is also evident (52–54). It
remains to be determined the contribution of these seven antigens to cellular immune responses.

In addition to the antibodies associated with protection, we identified some anti-Cryptosporidium
antigen-specific antibody responses that were associated with an increased risk of reinfection
(Figure 7B and Supplemental Figure 9). These antibodies occurred in a distinct subset of children
and their appearance was not correlated with that of the antibodies associated with the protective
immune response (Supplemental Figure 10). This observation is in line with our earlier finding
that not all antigen-antibody responses are associated with protection (Figure 4). We investigated
but found no association between the appearance of the antibodies targeting these nonprotective
antigens and biomarkers of inflammation or of IL-4 (which is involved in modulating the humoral
immune response) (Supplemental Figure 11). Further studies are needed to identify the
mechanism that promote a protective immune response.

A limitation of this study to identify antibody responses associated with protective immunity was
that not every infection was detected, despite active surveillance that included twice weekly home
visits. This coupled with the heterogeneity observed in the antibody profile among children with
prior exposure (i.e. lack of antibody responses in a subset of children) likely reduced the statistical
power of our analysis. Another limitation was the inability to probe the protein microarray with fecal IgA due to high background. A final limitation was that the microarray did not contain the entire proteome of the *C. parvum*, *C. hominis* and *C. meleagris* parasites. To offset this limitation we prioritized the inclusion of the potential vaccine candidates mentioned in earlier literature (n=22), as well as conserved transmembrane and secreted proteins (Supplemental Figure 3 and Supplemental Table 1) (13–33). Of the *C. parvum* genes with no introns, 336/761 ORF annotated as containing a transmembrane domain were included in the protein microarray. The remaining 425 ORFs were eliminated as being either absent in the other *Cryptosporidium* species of *C. meleagris* and *C. hominis*, or were represented by other orthologues on the array; or because they were part of the endomembrane system, components of the mitochondria, ribosome, or nucleus. We avoided potential metabolic proteins (annotated as DNA and RNA binding, involved in one of the biosynthetic pathways of the parasite or involved in the ubiquitin pathway). We also eliminated genes with a nonsynonymous/synonomous substitution rate >1 and that had at least 10 non-synonomous SNPs. Sixteen of the *C. parvum* ORF candidates failed at either the cloning step or during IVTT quality control. Seventy-one of the remaining ORFs that otherwise met our criteria as conserved membrane proteins could not be included on the array due to space limitations.

Strengths of this study included the intensive surveillance undertaken for cryptosporidiosis from birth to three years of life and the use of a proteome array that enabled the identification of previously unknown protective antigens (55). In addition to the antigens discussed these included some antigens that although interesting were not statistically significant (Supplemental Figure 11). This comprehensive identification of *Cryptosporidium* antigens to which an antibody response is associated with protection represents an important step forward in vaccine design. At the same time, the fact that many children do not develop a detectable antibody response to infection, and
that the antibody response when generated are short-lived and parasite genotype-specific may
offer insight into why recurrent infections are common.
Materials and Methods

Open Reading Frame Selection for the Cryptosporidium Protein Microarray

Although many Cryptosporidium species can infect humans, the most common human pathogens are the zoonotic C. parvum and C. meleagridis and the anthropotic C. hominis (6). The selected C. parvum proteins included those which had been previously identified as potential vaccine candidates (Supplemental Table 1) (13–33). The genome sequences of C. parvum and C. hominis are very conserved (95–97%) (56) and C. meleagridis is not very divergent (8.5%; C. parvum versus C. meleagridis) (57) (11, 12). The array included 522 conserved antigens derived from proteins (n=376) that contained signal peptides (C. parvum: antigens 382; C. hominis: 155 antigens [134 derived from the C. hominis orthologues of “high value” C. parvum proteins already on the array]); and C. meleagridis: 6 antigens [also represented by other orthologues on the array] (CryptoDB Database: Release 59 (58)). Due to discovery of export motifs such as PEXEL in other apicomplexan parasites, we did not restrict our antigen selections to only those with a signal peptide (59, 60). Additional antigens from proteins (n=327) without a signal peptide, but which nevertheless had a transmembrane domain, were included on the array (total antigens: 457; C. parvum antigens: 365; C. hominis antigens: 90 [66 were derived from the C. hominis orthologues of C. parvum proteins already on the array]; C. meleagridis 2 [both also represented by other orthologues on the array]).

Cryptosporidium Protein Microarray Fabrication

The protein microarray used in this study was produced by Antigen Discovery, Inc (ADI) (61, 62). Briefly, the open reading frames selected as described above were subcloned into a T7 expression vector pXI and expressed using an in vitro the Escherichia coli transcription and cell free translation (IVTT) system (Rapid Translation System, Biotechrabbit, Berlin, Germany). After
expression the proteins were printed onto nitrocellulose-coated AVID slides (Grace Bio-Labs Inc) using an Omni Grid Accent robotic microarray printer (Digilabs, Inc., Marlborough, MA, USA). As a positive control the purified recombinant Cryptosporidium Cp17 and Cp23 peptides were also spotted onto the array (8, 63).

Child cohort
A total of 500 children were enrolled within one week of birth from the Mirpur community of Dhaka, Bangladesh beginning in June 2014 through March 2016 (6)(ClinicalTrials.gov NCT02764918). (5) Infants were monitored for cryptosporidiosis through testing of diarrheal and monthly surveillance stool samples, through bi-weekly home visits by trained field investigators. Height and weight were measured quarterly to assess child growth (7). Children who had a HAZ score < -1 were defined as ‘at risk for stunting’ and HAZ < -2 as stunted (64). Select clinical metadata for this study is available on the NCBI’s dbGaP under accession number phs001665.v2.p1. The data for this study is collected as a sub-study of dbGaP phs001475.v2.p1.

Sampling and Specimen Testing
Fecal DNA was extracted from the diarrheal and monthly surveillance stools and a previously described multiplex qPCR assay which utilizes pan-Cryptosporidium primers and probes targeting the 18 S rDNA gene was used to identify infected infants (6). In select isolates the parasite was genotyped using standard protocols (9). Assays to measure select biomarkers of gut and systemic inflammation (CRP, sCD14, IL-1Beta, IL-4) were performed as previously described (6). A blood sample was drawn at 18 weeks, one year and then every 6 months (dbGAP Accession phs001665.v1.p1). The plasma samples from children which were collected at one year of age were diluted 1:100 and used to probe the Cryptosporidium proteome microarray using standard protocols (56). Samples were incubated on the arrays overnight at
4°C on a rocker, subsequently washed and incubated with polyclonal goat anti-human IgG-Fc fragment DyLight650 (Bethyl Laboratories, Cat#A80-104D5) or Cy3 AffiniPure F(ab’)2 fragment of the polyclonal goat anti-human serum IgA, α chain specific (Jackson ImmunoResearch Laboratories, Cat#109-166-011) for one hour at room temperature on a rocker, then washed, dried and stored in the dark until scanning. The exposed arrays were scanned, and the spot and background signal intensities (SI) were exported into R for statistical analysis (65). Spot SIs were adjusted for local background by subtraction, and values were floored to 1. Next, the data were normalized by dividing the Cryptosporidium protein spot values by the median of IVTT control spots (IVTT expression reactions with no Cryptosporidium ORFs), and values were log transformed using the base-2 logarithm. Thus, normalized data represented the log2 signal-to-noise ratio, where a value of 0 represents specific antibody SI equal to the background, 1.0 represented twice the background, 2.0 represented 4-fold over background, etc. For the purified Cp23 protein and Cp17 peptide printed on the microarrays, the data were unaffected by the IVTT background, and thus the normalization procedure was to floor the data to 1 and then transform values using the base-2 logarithm, resulting in normalized data that represent a doubling in fluorescence SI per unit increase. Thus, the scales of purified proteins and IVTT proteins differed due to the former being log-scale SI levels and the latter being log-scale signal-to-noise ratios. The results from the protein microarray are included in the supplemental data.

**Statistical Analysis**

Cryptosporidium protein responses were classified as seropositive or negative by taking the distribution of each spot individually between all samples to model negative and positive subpopulations using mixture models executed with the “normalmixEM” function in the mixtools package (66), and a seropositivity cutoff was established for each antigen as the mean and three standard deviations of the negative SI distribution. When mixture models failed to converge, a simple seropositivity cutoff of 1.0, or 2-fold over background, was applied. IgG and IgA “reactivity”
for each antigen was defined as a proportion of seropositive responses, or seroprevalence, of at least 10% among the study children. Overlap in IgG and IgA responses that recognized individual antigens was assessed using the VennDiagram package in R (68). Antibody breadth scores were calculated as the sum of seropositive responses per individual. Normalized SI and antibody breadth scores were visualized using the ComplexHeatmap package (67). Associations between normalized SI and protein features such as life cycle stage and signal peptides were analyzed using multivariable negative binomial regression for seropositivity classifications and ordinary least squares regression for normalized SI. Univariate groupwise analysis was performed using T-tests. Correlations were assessed using Pearson’s correlation coefficient. Unsupervised data reduction and trends was analyzed using t-distributed stochastic neighbor embedding (tSNE), using a perplexity parameter of 40, 1000 iterations and a theta of 0.0 after testing varying parameters for shape of the data (68, 69). To test differences in profile-wide normalized SI and to control for multiple measurements per subject, linear mixed effects regression (LMER) was performed using the lme4 package (44), allowing for random intercepts at the subject and antigen level—p-values for LMER models were obtained by ANOVA between full models and null variable models. Supervised data reduction and multi-antigen analysis was performed using partial least squares discriminant analysis (PLS-DA) using the mixOmics package (70). PLS-DA results were visualized using plots of PLS-DA scores of first and second latent variables for trends discriminating groups and by plotting the loading weights (i.e. PLS-DA regression coefficients) to highlight the importance of the variables used to define the latent components that maximize the covariance between antibody data and Cryptosporidium infection endpoints (infection prior to sampling and infection 1 and 2 years after sampling). Analysis of antibody breadth was performed using negative binomial regression.

Analysis of seropositivity to an antigen and association with risk of infection or reinfection (only among children with infections prior to 1 year of age) during the one-year and two-year follow-up...
periods following sample collection was performed using Cox proportional hazards models adjusted by the number of episodes a child had, HAZ score at birth, mother’s age, BMI and education level, household income and expenses, principal source of household drinking water, the water treatment method routinely used by the household, and proximity to one of the Dhaka water drainage channels. The two a priori vaccine candidate antigens Gp60 (containing the Cp17 peptide) and Cp23 were analyzed without correction for the false discovery rate (71).

To control for false discovery due to the multiple comparisons made, antibody responses were modeled using random forest (“RF”) with survival data using the “randomForestSRC” and “ggRandomForests” packages to perform feature selection (72). All reactive antigens were included in each of four RF models: 1 year follow-up post-sampling among all children or among only children with previous infections, and 2 year follow-up among the two groups of children. Each model performed 1,000 decision trees while allowing deterministic splitting (nsplit = 0) and computation of variable importance (VIMP) using permutation. Each model was repeated 100 times, and VIMP scores were computed for each repeat and then averaged for each antigen target. VIMP scores that were greater than 0 in at least 80% of repeats (80/100 RF models) were considered “potentially important variables” for the RF models. Further, a VIMP score cutoff was calculated as one standard deviation above the mean of all VIMP scores returned from the 100 RF model iterations. The survival analysis identified which antibodies were associated with a hazard ratio below 1 (protective response) vs hazard ratios greater than 1 (increased susceptibility). The criteria determined for selection of a minimal number of important variables was: (1) positive VIMP scores in at least 80% of repeats and average VIMP scores above the cutoff, and (2) seropositivity hazard ratios below 1 in all models. The selected features were used in Cox models, reporting both raw p-values and p-values adjusted for the false discovery rate.

Detailed information about the R workflow used in this paper is publicly available at

https://github.com/wikipetria/Cryptosporidium. Analysis of Gp40/15 (aka Gp60) sequence
variation and antibody associations was performed using the Prism 9 computer program (GraphPad).

Study approval

The study was approved by the Ethical and Research Review Committees of the International Centre for Diarrhoeal Disease Research, Bangladesh (PR-13092) and the Institutional Review Board of the University of Virginia (IRB# 20388). The ClinicalTrials.gov identifier is NCT02764918. Informed written consent was obtained from the parents or guardians for the participation of their child in the study.

Acknowledgements

This work was supported by the National Institute of Allergy and Infectious Diseases (NIAID) grants awarded to CAG and WAP (R01 AI-043596), CAG (R21 AI-109118); JJC (R43AI145607) and WAP (R21-AI154862) as well as the Bill and Melinda Gates Foundation Grant (OPP1100514) to ASGF. The funders had no role in study design, data collection and analysis or decision to submit for publication. We also wish to thank our colleagues Dr B Mann and Dr C. Marie Braun for careful reading of this manuscript and thoughtful feedback. We also wish to thank the field workers, nurses, laboratory staff of the Parasitology Laboratory of icddr,b who worked for this project, and parents and children at the icddr,b study sites who participated in this study and without whom we could not have completed this research. Work at icddr,b is supported by the core donors (Government of the People’s Republic of Bangladesh, GAC, Sida, UKAid).

Author contributions

Drafting of the manuscript was performed by CAG and JJC (co-first authors). The method used to assign initial authorship order was by mutual agreement. All authors edited and approved the final manuscript. CAG, WAP and RH conceived of the analysis plan and JJC performed the
bioinformatic analyses. JZM assisted in the statistical analysis. CAG designed and JJC, JVP, AT AO and ADS built the array and analyzed samples. WAP, RH, and ASGF founded the birth cohort and directed the study. Field work and data collection at the International Centre for Diarrhoeal Disease Research, Bangladesh (icddr,b) were performed by MA and MK, with supervision from ASGF and RH.
Bibliography


56. Obiero JM et al. Antibody Biomarkers Associated with Sterile Protection Induced by Controlled Human Malaria Infection under Chloroquine Prophylaxis. *mSphere* 2019;4(1).


69. van der Maaten L, Courville A, Fergus R, Manning C. Accelerating t-SNE using Tree-Based Algorithms [Internet]. *J Mach Learn Res* 2014;15(1)


765
Figure 1. Humoral immunity to Cryptosporidium antigens was isotype specific. Immune responses are shown for A) IgA and B) IgG antibodies. The Y-axis is signal intensity after normalization and the X-axis shows Cryptosporidium antigens ranked by median signal intensity. Bars represent the interquartile range of each antibody response and are shown as red if antibodies were present in \( \geq 10\% \) of infants (seroprevalent). C) The Venn diagram shows seroprevalent Cryptosporidium antigens with IgA- and IgG-specific and overlapping immune responses.
Figure 2. *Cryptosporidium* antigens recognized by IgA and IgG antibodies. The proteomic microarray was used to measure the parasite specific antibody response in the infants enrolled into our study cohort at one year in age. Previously infected children (columns) and the *Cryptosporidium* antigens (rows) that stimulated a strong IgG and/or IgA antibody response (present in > 10% of the children; n=232 antigens) are shown. The spot signals were normalized by first determining the specific background component by use of mixture models and setting this value to 0. Bar at the top of each heat map indicates the total number of *Cryptosporidium* antigens each child responds to (Antibody Breadth). The side bars indicate: a) the seroprevalence of each antigen and b) presence of a membrane targeting signal peptide (SP).
Figure 3. Antibody responses waned with time after a Cryptosporidium infection. A) The t-SNE plot identified a subset of children with a similar antibody profile. Each point corresponds to the immune profile of a child. Gray squares indicate children where no previous Cryptosporidium infections had been identified by qPCR in clinical or surveillance stool samples ("qPCR-"), and orange circles represent children that had previous infections detected by qPCR ("qPCR+"), with the intensity of the overlaid color indicating the days since the last Cryptosporidium qPCR+ stool sample was identified. A group of infants had similar antibody profiles and a high density of recent infections (R1). B) The split violin plot of antibody signals against the 100 most reactive antigens (Y-axis) for each isotype (X-axis) shows the responses of children within the R1 region of the t-SNE plot compared to the remainder of the samples in R2. The median and quartile values are shown as horizontal lines in each split violin. C) The split violin plot shows the same comparison as (B) using the antibody breadth (count of seropositive responses) among the 100 most reactive antigens. P-values above each split violin were calculated using linear mixed effects regression (LMER) and Wilcoxon’s rank sum tests for (B) and (C), respectively. D) Antibody breadth among the 100 most reactive antigens for each isotype is shown on the Y-axis after log$_{10}$ transformation with the interval (days) between the last Cryptosporidium qPCR positive diagnostic assay and the time of antibody measurement shown on the X-axis. Linear regression P-values and R$^2$ values are shown for IgG and IgA, as well as a line and confidence intervals (colored bands) fit to each. E and F) PLS-DA is shown for IgA and IgG responses respectively. Each point corresponds to the immune profile of a child. The purple circles indicate the antibody response obtained from plasma that was collected from children where none of the stool samples (diarrheal or surveillance) collected during the first year of life, prior to the plasma sampling time point, were ever qPCR positive for Cryptosporidium parasites ("Yr0-1 qPCR-"). Green triangles indicate that the child had a verified Cryptosporidium sub-clinical or symptomatic infection ("Yr-0-1 qPCR+"). The percentage of the variation in the child’s antibody profile accounted for by each axis is indicated.
Figure 4. The breadth of the anti-Cryptosporidium immune response was not correlated with protection from infection. A) Split violin plot of antibody breadth in plasma among the 100 most reactive antigens (Y-axis) for each isotype (X-axis) is shown for the comparison between children that had no stool samples (diarrheal or surveillance) qPCR positive for Cryptosporidium parasites (purple) and children who had a verified Cryptosporidium infection (green). B) Data is shown from one year old infants who had prior qPCR-confirmed Cryptosporidium infections (“Yr0-1 qPCR+”) that were subsequently uninfected (blue) or reinfected (orange) during the next two years. C) Data is shown from one year old infants that included both the immunologically naïve infants with no prior Cryptosporidium infections detected by qPCR in stool samples (diarrheal or surveillance) as well as those with qPCR+ stool samples during the first year of life. Medians and quartiles are indicated by horizontal lines in each split violin. Significant P-values from Wilcoxon’s rank sum tests are shown above violins. N.S.: not significant.
Figure 5. Children with antibodies that targeted the *C. hominis* peptides encoded by the *gp60* gene and Cp23 protein were associated with protection from reinfection. In the protein array data, IgA and IgG antibodies against the protein encoded by the *C. hominis gp60* gene (Chro.60183) and IgG against Cp23 (Chro.40414) were associated with a delay in *Cryptosporidium* reinfection among children with a qPCR verified *Cryptosporidium* infection during the first year of life (A, C, E) or among all children in the study (B, D, F). The X-axis shows days after the end of year one (when the assayed plasma samples were collected). The Y-axis shows the proportion of children who remained uninfected. Red lines represent children seronegative for the antigen, and blue lines represent seropositive children. The Kaplan-Meier curves show the probability of survival free of *Cryptosporidium* species, and the tables below the graphs indicate the number of children in the seropositive or seronegative categories at select time points. A-B) IgG against Gp60 (Chro.60183). C-D) IgA against Gp60 (Chro.60183). E-F) IgG against Cp23 (Chro.40414). Hazard ratios (HR), confidence intervals and p-values were calculated using multivariable Cox proportional hazards models.
**Figure 6. gp60 Genotype immune response.** A) Cartoon illustrating the proteins encoded by the *gp60* gene. Heat map showing the intensity and breadth of the B) IgA and C) IgG antibody responses to the polymorphic region of the Gp40 protein. The different alleles of the peptide encoded by the *gp60* allele (columns) and the signal obtained when using the plasma with antibodies raised in response to infection of parasite with different *gp60* genotypes (rows). Lines at the top of each heat map indicate the protein type and on the side the genotype of the infecting parasite. Parasite genotypes: rows 1-8 IaA18R3; 9; IaA19R3; 10-16 IaA25R3; 17-20 IbA9G3R2; 21-29 IdA15G1; 30 IfA13G1; 31-34 C. parvum IIdA15G1R1. Protein alleles columns A Ia27R3, B IaA26R3, C Ia25R3, D IaA22R3, E IaA19R3 F IaA18R3, G IbA9G3R2, H IdA14G1, I IdA15G1, J IeA11G3T3, K IfA13G1, L IfA16G1, M IIcA5G3a N IIdA13G1. Side panels show the intensity scale for the amount of antibody binding to alleles expressed by IVTT and spotted on the array. Antibody binding to the purified recombinant relatively conserved Cp17 peptide was included on the array as a positive control. Its signal intensity was higher than that of the IVTT values.
Figure 7. Random Forest analysis for selection of important antigens and analysis of risk during the first year after sampling. A) The scatter plot represents antigens and clinical variables ranked by variable importance (“VIMP”) scores in random forest (“RF”) using 1,000 trees constructed per model. Models were fit to survival data during one year of follow-up after sampling on seropositive and seronegative children that all previously had qPCR-confirmed Cryptosporidium infections. Models using the entire cohort of children and two year follow-up periods are shown in Supplemental Figure 8. Each model was repeated 100 times, and the VIMP score was averaged across all runs (Y-axis). For each antigen, the percentage of runs where VIMP was greater than 0 (i.e., important to the model) was calculated (X-axis). The red horizontal dashed lines represent the mean of all VIMP scores plus one standard deviation. The vertical dashed red lines represent antigens with at least 80% positive VIMP scores. The upper right quadrant shows the antigens selected as important variables in the model. B) The horizontal bar plot represents VIMP scores for each antigen with at least 80% positive VIMP scores. The vertical red dashed line represents the cutoff for selection of important variables (equivalent to the horizontal lines in A). Hazard ratios calculated in the survival analysis were shown as protective (HR < 1, teal) or not (HR > 1, magenta). C) Only protective antigens with at least 80% positive VIMP scores and VIMP scores above the importance cutoff were selected for individual antigen analysis. (D-G) The Kaplan Meier plots represent the two most significant previously unknown antigens associated with protection in children with prior qPCR+ stool samples or all children, respectively, after feature selection using random forest.
C Novel antigens identified from Random Forest analysis

<table>
<thead>
<tr>
<th>Gene</th>
<th>Description</th>
<th>Isotype</th>
</tr>
</thead>
<tbody>
<tr>
<td>cg04_4270</td>
<td>Uncharacterized protein</td>
<td>IgG</td>
</tr>
<tr>
<td>cg01_1520</td>
<td>CorA (previously uncharacterized protein)</td>
<td>IgG</td>
</tr>
<tr>
<td>cg03_3070</td>
<td>High mobility group box domain containing protein</td>
<td>IgG</td>
</tr>
<tr>
<td>cg07_4020</td>
<td>Gp900, Cryptosporidial mucin</td>
<td>IgG</td>
</tr>
<tr>
<td>cg08_700</td>
<td>Cpmuc8, secreted mucin (previously uncharacterized secreted protein)</td>
<td>IgG</td>
</tr>
<tr>
<td>cg08_830</td>
<td>CCDC, coiled-coil domain-containing protein (previously uncharacterized protein)</td>
<td>IgG</td>
</tr>
<tr>
<td>Chro.30111</td>
<td>Cpmuc8, small membrane protein (previously hypothetical protein)</td>
<td>IgG</td>
</tr>
</tbody>
</table>

Four separate survival RF models were fit: one year of follow-up post-sampling in children with year 0-1 qPCR+ stool samples; one year of follow-up post-sampling in all children; two years of follow-up post-sampling in children with year 0-1 qPCR+ stool samples; two years of follow-up post-sampling in all children. Criteria for selection of antigens were: (1) VIMP > 0 in at least 80% of RF model repetitions, (2) mean VIMP score greater than one standard deviation above mean of all VIMP scores, (3) hazard ratio in survival analysis < 1.0 and not > 1.0 in any of the four RF models.
Figure 8. *Cryptosporidium* antigens associated with the development of a protective immune response. A) The survival curve illustrates the two subgroups of children and two follow-up periods after plasma was collected (end of year one) that were analyzed for protection. The blue line follows only children who had a qPCR-verified *Cryptosporidium* infection (sub-clinical or symptomatic) during the first year of life, prior to sampling plasma. The red line follows all children in the array study and includes the immunologically naïve children that remained uninfected at one year of age as well as those known to be previously infected. Dotted and dashed lines indicate the time points (year 2 and year 3) selected for analysis of the differences between uninfected and infected groups looking specifically at the protective candidate antigens identified in Table 1. B and C) PLS-DA on the antibody profiles of the candidate antigens shown in Table 1 associated with either reinfection or protection. Each point represents the immune profile from one year old children with prior qPCR-verified *Cryptosporidium* infections who were subsequently uninfected (blue circles) or reinfected (orange triangles) during the one year follow-up period (B) or two year follow-up period (C) after plasma samples were collected. D and E) Predictor loadings derived from the PLS-DA analysis in (B) and (C) are shown, respectively. Antibody targets are shown on the Y-axis, and the X-axis shows the absolute value of the loading weights (or PLS-DA regression coefficients); the absolute value was used to focus attention on the importance of each antigen in maximizing the covariance between antibodies and *Cryptosporidium* infection outcomes. Orange bars indicate antibodies more abundant in the children who subsequently had a new *Cryptosporidium* infection, and blue bars indicate the antibodies more abundant in the uninfected children.
Table 1. Antibody responses associated with protection resulting from a *Cryptosporidium* infection

<table>
<thead>
<tr>
<th>Gene*</th>
<th>Description</th>
<th>Sero-prevalence</th>
<th>HR</th>
<th>P-value†</th>
</tr>
</thead>
<tbody>
<tr>
<td>IgG</td>
<td>Chro.601.38</td>
<td>Gp60, sporozoite antigen gp40/15</td>
<td>52/104 (50%)</td>
<td>0.41 (0.23-0.74)</td>
</tr>
<tr>
<td>IgG</td>
<td>Chro.601.38</td>
<td>Gp60, sporozoite antigen gp40/15</td>
<td>37/104 (35.6%)</td>
<td>0.47 (0.25-0.89)</td>
</tr>
<tr>
<td>IgG</td>
<td>Chro.404.14</td>
<td>Cp23, immunodominant antigen</td>
<td>55/104 (52.9%)</td>
<td>0.55 (0.31-0.99)</td>
</tr>
<tr>
<td>IgG</td>
<td>Cgd8_70_0</td>
<td>CpMuc8, secreted mucin</td>
<td>12/104 (11.5%)</td>
<td>0.26 (0.09-0.77)</td>
</tr>
<tr>
<td>IgG</td>
<td>Chro.301.11</td>
<td>CpSMP, small membrane protein (Signal peptide &amp; transmembrane domain)</td>
<td>33/104 (31.7%)</td>
<td>0.45 (0.23-0.87)</td>
</tr>
<tr>
<td>IgG</td>
<td>Cgd7_40_20</td>
<td>Gp900, Cryptosporidial mucin</td>
<td>57/104 (54.8%)</td>
<td>0.53 (0.3-0.95)</td>
</tr>
</tbody>
</table>

*Antibodies at one year of life associated with reduced risk of infection in year 2 (children with qPCR positive stool samples during the first year of life, n=104 children)*

<table>
<thead>
<tr>
<th>Gene*</th>
<th>Description</th>
<th>Sero-prevalence</th>
<th>HR</th>
<th>P-value†</th>
</tr>
</thead>
<tbody>
<tr>
<td>IgG</td>
<td>Chro.601.38</td>
<td>Gp60, sporozoite antigen gp40/15</td>
<td>88/435 (20.2%)</td>
<td>0.56 (0.37-0.85)</td>
</tr>
<tr>
<td>IgA</td>
<td>Chro.601.38</td>
<td>Gp60, sporozoite antigen gp40/15</td>
<td>50/435 (11.5%)</td>
<td>0.56 (0.33-0.94)</td>
</tr>
<tr>
<td>IgG</td>
<td>Chro.404.14</td>
<td>Cp23, immunodominant antigen</td>
<td>89/435 (20.5%)</td>
<td>0.58 (0.39-0.88)</td>
</tr>
<tr>
<td>IgG</td>
<td>Cgd2_15_20</td>
<td>CorA, possible magnesium transporter</td>
<td>76/435 (17.5%)</td>
<td>0.55 (0.36-0.83)</td>
</tr>
<tr>
<td>IgG</td>
<td>Cgd8_70_0</td>
<td>CpMuc8, secreted mucin</td>
<td>47/435 (10.8%)</td>
<td>0.53 (0.31-0.9)</td>
</tr>
<tr>
<td>IgG</td>
<td>Cgd8_83_0</td>
<td>CCDC, coiled-coil domain-containing protein</td>
<td>67/435 (15.4%)</td>
<td>0.58 (0.37-0.92)</td>
</tr>
</tbody>
</table>

*Antibodies at one year of life associated with reduced risk of infection in year 2 (all children in the cohort, n=435 children)*

*Antibodies at one year of life associated with reduced risk of infection in years 2 and 3 (children with qPCR positive stool samples during the first year of life, n=104)*
<table>
<thead>
<tr>
<th>Antibodies at one year of life associated with reduced risk of infection in years 2 and 3 (all children in the cohort, n=435)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>IgG Chro.601 38</strong></td>
</tr>
<tr>
<td><strong>IgA Chro.601 38</strong></td>
</tr>
<tr>
<td><strong>IgG Chro.404 14</strong></td>
</tr>
<tr>
<td><strong>IgG Chro.301 11</strong></td>
</tr>
<tr>
<td><strong>IgG Cgd7_40 20</strong></td>
</tr>
</tbody>
</table>

* Gene ID from CryptoDB; † P-values computed from multivariable Cox proportional hazards models, and P$_{FDR}$ represents P-values adjusted for the false discovery rate among the subset of protective antibodies identified from Random Forest analysis (hazard ratios < 1). Antibodies against the a priori vaccine candidate antigens Gp60 and Cp23 were not adjusted for the FDR. HR: hazard ratio.