Genome-wide association studies of diarrhea frequency and duration in the first year of life in Bangladeshi infants reveal links to enteric nervous system

Rebecca M. Munday[1], Rashidul Haque[2], Genevieve L. Wojcik[3], Poonum Korpe[3], Uma Nayak[4], Beth D. Kirkpatrick[5], William A. Petri Jr.[6], Priya Duggal[3#]

[1] Department of Genetic Medicine, Johns Hopkins School of Medicine, Baltimore, Maryland, USA

[2] International Centre for Diarrhoeal Disease Research, Bangladesh, Dhaka, Bangladesh

[3] Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, USA

[4] Center for Public Health Genomics and Department of Public Health Sciences, University of Virginia School of Medicine, Charlottesville, Virginia, USA

[5] University of Vermont College of Medicine and Vaccine Testing Center, Burlington, Vermont, USA

[6] Department of Medicine, Infectious Diseases and International Health, University of Virginia School of Medicine, Charlottesville, Virginia, USA

[#]Corresponding author: fax 410-955-0863, telephone 410-955-1213, email pduggal@jhu.edu

Main Text: 3492 words
Abstract: 200 words

Abstract

Background

Diarrhea is the second leading causing of death in children under 5 worldwide. The children most impacted by diarrhea are those living in low-to-middle-income countries (LMICs), averaging 3 diarrheal episodes in the first year alone. While many risk factors for diarrhea are known, we hypothesized that there may be host genetic variants which impact the frequency or duration.

Methods

Using three well-characterized birth cohorts from Dhaka, Bangladesh, we compared infants with no diarrhea in the first year of life to those with an abundance, measured by either frequency or duration. Cohorts were analyzed separately assuming an additive model and adjusting for sex and the first principal component. Results were meta-analyzed using a fixed-effects inverse-variance weighted model.

Results

In comparing frequency, we identified two genome-wide significant loci associated with no diarrhea: rs2827548 on chr 21 (OR=0.31, 95%CI=0.21-0.47, P=4.01x10^{-8}) and rs1915541 on chr 8 (OR=0.35, 95%CI=0.23-0.53, P=4.74x10^{-7}). In comparing duration, we identified two loci associated with no diarrhea: rs2827548 on chr 21 (OR=0.31, 95%CI=0.21-0.46, P=1.59x10^{-8}) and rs5026214 on chr 17 (OR=0.35, 95%CI=0.24-0.51, P=1.09x10^{-7}).

Conclusions

These loci are in or near genes involved in enteric nervous system development and intestinal inflammation. Any of these loci are potential drug targets for diarrhea therapeutics.

Key words: diarrhea, host genetics, association, GWAS, enterics, malnutrition

Background

Each year, 525,000 children under the age of 5 years old die from diarrhea, making it the second most common cause of death in that age group [1]. Efforts to reduce diarrhea have included rotavirus immunization, increasing access to clean water, improving toilet facilities, advocating exclusive breastfeeding, and health education programs [1]. While childhood mortality and overall rates of diarrhea have improved [2], there are still many children experiencing multiple bouts of diarrhea in their first year of life, particularly in low-to-middle-income countries (LMICs) [1]. Outside of public health measures and education, the reported heterogeneity of diarrhea even among children with similar exposures suggests that immunity and host genetics may play an important role [3–5].

Previous studies of the host genetic contribution to diarrheal disease generally fall into three categories: bowel disorders, inherited Mendelian disorders, and pathogen-specific analyses. Genetic studies of bowel disorders include irritable bowel syndrome [6], Celiac disease [7], and inflammatory bowel disease (Crohn's disease [8] and ulcerative colitis [9]), and hundreds of susceptibility genes have been identified. Studies of inherited Mendelian disorders, such as congenital sodium diarrhea [10] and Trichohepatoenteric syndrome [11], have also identified associated genes. Early pathogen-specific analyses focused on candidate genes, such as *FUT2, FUT3,* and *ABO* [12,13], but more recently have expanded to evaluating genes across the human genome. These include the identification of host genes associated with *Entamoeba histolytica* [14], *Cryptosporidium* [15], *Shigella* [16], and *Campylobacter* [17] diarrheal infections. A large genome-wide association study of European and American children enrolled in several cohorts identified the *FUT2* locus associated with infant diarrhea [18] in European ancestry children not vaccinated for rotavirus.

In this study we interrogated the human genome for associations with diarrhea among poor infants in the first year of life from Dhaka, Bangladesh. Focusing this study in an urban slum with limited sanitation and overcrowding we aimed to identify those children on the tail

ends of the spectrum: fully protected or highly susceptible to multiple diarrheal episodes in the first year of life. The identification of genes associated with these extreme outcomes could provide ideal targets for drug development or therapeutics for all children.

Methods

Institutional Review Boards at Johns Hopkins Bloomberg School of Public Health, University of Virginia, and International Centre for Diarrhoeal Disease Research, Bangladesh (icddr,b) approved study protocols. Parent/guardian of each participant provided written informed consent.

Study participants

Infants were ascertained from three independent birth cohorts, all from Mirpur, Dhaka, Bangladesh: the Dhaka Birth Cohort (DBC) [19], the “Performance of Rotavirus and Oral Polio Vaccines in Developing Countries” (PROVIDE) Study [20], and the Cryptosporidiosis Birth Cohort (CBC) [21]. In each cohort, enrollment took place within one week after birth and individuals were followed for at least the first year of life. Field research assistants visited families multiple times per week, collecting information about feeding habits and symptoms of illness, including but not limited to, diarrhea, vomiting, coughing, and fever. Anthropometric data were recorded at time of enrollment and subsequently every three months.

Clinical definitions

For each participant, the first year of life was defined as an interval beginning on the day of birth and ending on the first birthday. For those born on February 29, 2012, the first birthday was defined as March 1, 2013. Any diarrheal episodes occurring within the interval were then counted and used to define the phenotype. Diarrhea was defined as 3 or more abnormally loose stools in a 24-hour period. An episode of diarrhea was defined as a period of days during which

a caregiver reported that an infant had diarrhea, including no more than 2 consecutive non-diarrheal days. In PROVIDE and CBC, some infants underwent lactulose-mannitol testing during the first year. As lactulose is an osmotic laxative, incidents of diarrhea that occurred on the same day as or the day following a lactulose-mannitol test were excluded from the total count. Total duration of diarrhea was defined as the sum of all days in all episodes in the first year, including non-diarrheal days within a given episode. For example, if an infant had two days of diarrhea followed by two non-diarrheal days followed by one more day of diarrhea, this would be a single episode of five days duration.

We considered two case definitions based on extremes in either number of diarrheal episodes, or days with diarrhea, as determined by examining the distributions. Infants were defined as controls if they had zero episodes of diarrhea and/or had zero days of diarrhea in the first year of life. The first case definition were infants with six or more diarrheal episodes in the first year of life. The second case definition were infants reporting 25 or more days of diarrhea in the first year of life.

To compare exposure profiles, we analyzed all available surveillance (non-diarrheal) stool samples in the birth cohorts. Both DBC and CBC had monthly stool collection which were tested via RT-PCR for three pathogens: *Entamoeba histolytica*, *Cryptosporidium*, and *Giardia*. In DBC, 22/26 children without diarrhea had at least one surveillance sample tested and 314/354 of the remaining cohort had at least one surveillance sample tested. In CBC, all 365 children had at least one surveillance sample tested. PROVIDE did not conduct monthly stool collection but clinical samples obtained at 6 weeks and 10 weeks were tested for a variety of Enteropathogens via TaqMan array card. Using this method, 20/32 children without diarrhea had samples tested and 242/443 of the remaining cohort had samples tested.

Socioeconomic definitions

Household income was reported as total monthly income in Bangladeshi taka (BDT). For comparison, during the time of data collection, the international poverty line for a family of five was approximately 21000 BDT per month ($1.90 USD per day per person [22] * 5 people * 30 days per month * 75 BDT per USD (average exchange rate over study periods)).

Sanitation was defined as either “Improved” toilets including those utilizing a sewer or septic tank, as well as water-sealed or slab latrines, and ventilated-improved pit latrines [23] or “Unimproved” toilets including other pit latrines, open latrines, hanging latrines, and those with no facility. Drinking water was considered “Protected” if from municipality-supplied piped water, pumped water, and tube wells [23] and “Unprotected” if from an open well or surface water such as from a river, pond, or canal.

Genotyping, quality control, and imputation

Methods for genotyping, quality control, and imputation have been previously described [14] and are outlined in detail in Figure S1. All steps were done for each cohort separately. Briefly, infants from DBC were genotyped using 3 different Illumina arrays (Human1M-duoV3, HumanOmni1-Quad v1.0, and HumanOmni2.5-4v1), infants from PROVIDE were genotyped using the Expanded Multi-Ethnic Genotyping Array (MEGA-EX) from Illumina, and infants from CBC were genotyped using Illumina’s Infinium Multi-Ethnic Global Array (MEGA). Standard QC measures included checking clustering, SNP missingness, individual missingness, and heterozygosity rate. PCA outliers, heterozygosity outliers, and one member from each pair of relatives were removed. Variants were filtered for minor allele frequency (MAF) >0.005 and Hardy-Weinberg equilibrium *P* value $>5x10^{-5}$. After this process, phasing and imputation were conducted using SHAPEIT [24,25] and IMPUTE2 [26–30], with 1000 Genomes Project phase 3 as reference [31]. After imputation, each cohort was rechecked for any additional cryptic relatedness using PLINK [32].

Inter-cohort relatedness was assessed using KING [33], and one member from each pair of relatives was removed (n=24). CBC was restricted to the Mirpur site, resulting in removal of 221 individuals. Infants were removed if they left the study before their first birthday, to ensure proper phenotyping. Final sample sizes for each cohort were 380 in DBC, 475 in PROVIDE, and 365 in CBC.

Association analysis

We used SNPTEST [26,28,34] to run logistic regression for each of these case-control analyses, assuming an additive model of inheritance, and including sex and the first principal component as covariates. In DBC, genotyping batch was also included as a covariate to account for any potential batch effects. Each cohort was analyzed separately, and results were filtered for MAF ≥0.05 and information score (INFO) ≥0.7. Filtered summary statistics were used in META [35] for inverse-variance weighted, fixed-effects meta-analyses. Meta-analysis results were restricted to variants that were observed in all 3 cohorts and passed heterogeneity testing (P_{het} >0.05). We conducted a sensitivity analysis by adding variables for socioeconomic status to the model and analyzing the top loci again. We added income as a continuous variable and access to improved toilet as a discrete binary variable, with all other aspects of the model as specified above. We assessed potential interaction between the top SNPs and sex, income, and sanitation using nested models in R.

Annotation and functional assessment

Results from the meta-analysis were privately uploaded to LocusZoom [36] and FUMA [37]. Complete methods for each of the FUMA tests can be found on the website (https://fuma.ctglab.nl/). Briefly, annotation included MAGMA competitive gene set analysis [38], functional consequences of SNPs on genes as assigned by ANNOVAR [39], overlap with regulatory elements, regulatory potential from RegulomeDB [40], expression quantitative trait

loci (eQTLs) from the Genotype-Tissue Expression (GTEx) project [41], and chromatin interactions. Where applicable, expression data were obtained from The Human Protein Atlas (proteinatlas.org) [42]. Global allele frequencies were visualized using the Geography of Genetic Variants Browser (popgen.uchicago.edu/ggv/) [43].

Results

Clinical and sociodemographic characteristics of each cohort

Overall, the male to female ratio was about 50% for each cohort with an average household size of 5-6 (Table S1). Maternal age at time of enrollment was similar in each cohort. Median monthly household income steadily increased over time reflecting overall changes in the Dhaka community, from 6,000 BDT in DBC (2008-2010) to 10,000 BDT in PROVIDE (2011-2012) and 14,000 BDT in CBC (2014-2016). All three cohorts were well below the international poverty line (21,000 BDT for family of 5). In all cohorts the infants had an average of approximately four episodes of diarrhea and total duration of 16-18 days over the course of the first year. At birth, the length-for-age Z-scores (LAZ), representing long-term nutritional status, were similar in all cohorts. Weight-for-length Z-scores (WLZ) at birth, representing acute nutritional status, improved over time but still fell more than one standard deviation below the global median. At one year of age, the average LAZ in each cohort was lower than at birth, indicating chronic undernutrition over the course of the first year. The proportion of households using improved toilets, including sewer, septic tank, water- or slab-sealed pits, and ventilated-improved pits, sharply increased from 38% in DBC to 95% in PROVIDE and 94% in CBC. Very few households in any cohort had unprotected drinking water (<1%).

The number of episodes of diarrhea and total days of diarrhea are detailed in Figure S2. We defined controls as infants with no reported diarrhea in the first year (0 episodes/0 days) and cases as those with 6+ episodes and/or 25+ days. The two groups with extreme diarrhea

had considerable but not complete overlap of individuals. In DBC and PROVIDE, the children with no diarrhea were more likely to live in households with access to improved toilets (Table 1). In all 3 cohorts we noted an inverse relationship between median age at first diarrheal episode and both total number of episodes and total duration in days. In DBC and PROVIDE, children with the most diarrhea (6+ episodes and/or 25+ days) had their first episode at a median age <60 days, and the same infants in CBC were <70 days. This is in stark contrast to the children with only 1-2 episodes and/or 1-14 days, who were closer to 150 days old at the time of their first diarrheal episode (Fig S3).

We evaluated whether the children with no diarrhea were exposed to pathogens in the environment at a level comparable to those with diarrhea (Fig 1). In the available samples, the children with no diarrhea had similar rates of asymptomatic infection compared to their respective cohorts. This indicates that these children were exposed to the same pathogens as their peers but did not develop diarrhea. PROVIDE non-diarrheal samples were collected at 6 weeks and 10 weeks of age (Fig S4).

Genome-wide association analyses

We evaluated approximately 7.36 million variants in 383 infants. Comparing infants with 6 or more diarrheal episodes to those with no diarrhea, we identified two protective genome-wide significant (P <5x10^{-7}) regions on chromosomes 21 and 8 (Fig 2A). The top SNP was rs2827548 located on chromosome 21 in an intron of the non-coding RNA *AP000959* (OR=0.31, 95%CI=0.21-0.47, P=4.01x10^{-8}). The protective effect size was consistent across cohorts (P_{het}=0.24; Table 2). Interestingly, in the meta-analysis for no diarrhea versus ≥ 25 days of diarrhea, we observed the same association with SNP rs2827548 (OR=0.31, 95%CI=0.21-0.46, P=1.59x10^{-8}) (Fig 2B, Table 2).

The other protective locus for 0 vs 6+ episodes of diarrhea was rs1915541 on chromosome 8 (EAF=0.25), in the first intron of the sterile alpha motif domain containing 12

(*SAMD12*) gene (OR=0.35, 95%CI=0.23-0.53, $P=4.74\times10^{-7}$). In the analysis of 0 vs 25+ days of diarrhea, another protective locus was identified on chromosome 17 for rs5026214, near *WSCD1* (EAF=0.55, OR=0.35, 95%CI=0.24-0.51, $P=1.09\times10^{-7}$). To account for any underlying socioeconomic changes in the cohorts over time, we performed a sensitivity analysis on the top loci, and we included monthly household income and access to improved toilet facility. All of the top SNPs remained significant, and the effect sizes were consistent (Table S2).

The allele frequencies for the effect allele were similar in both diarrhea case categories across the three studies (Table S3). Interestingly, the children with no diarrhea for rs2827548 had an effect allele frequency of 0.42, higher than the 1000 Genomes Bengali in Bangladesh population (0.33) and higher than any other 1000 Genomes population in Asia or Europe (Fig 3). Consistent with the protective effect, the effect allele frequencies were higher in the no diarrhea group as compared to the case categories. We evaluated the genetic principal components by case definition, and there was no substructure detected to suggest that the children with no diarrhea are genetically distinct from the other children at a genome-wide level (Fig S5).

Assessment of interaction

We were interested in whether there were any interactions between our top loci and sociodemographic factors, including sex, income, and sanitation. When we stratified the infants in DBC by access to improved toilets, we observed that the top SNP on chr 21, rs2827548, was only protective against 6 or more episodes of diarrhea for those with access to improved toilets and those with unimproved toilets showed no difference in case status by genotype (Fig S6A). We were unable to include a similar stratification for the other cohorts due to most families having improved toilets. Including an interaction term in the generalized linear model revealed a significant interaction between access to improved toilet and the top SNP on chr 21, rs2827548, for the outcome of 0 vs 6+ episodes of diarrhea in DBC (Table S4). The same interaction was

significant in DBC for the analysis of 0 vs 25+ days of diarrhea (Table S5). In PROVIDE, we noted a significant interaction between sex and the top SNP on chr 21, rs2827548, for both outcomes (Tables S6, S7). In PROVIDE, we also observed a significant interaction between income and the top SNP on chr 17, rs5026214, for the analysis of 0 vs 25+ days of diarrhea. No significant interactions were observed in CBC (Tables S8, S9).

Annotation and functional assessment

One way to assess potential function of non-coding variation is to look for chromatin interactions, or areas of the genome that are in close contact and become ligated together during the experiment. FUMA specifically annotates interactions between a locus of interest (the significant region from the GWAS) and the promoter region of a gene, which indicates potential regulatory function. For the chromosome 21 locus, FUMA identified significant (FDR $\leq 1x10^{-6}$) chromatin interactions which mapped to 40 different genes (Fig S7). The chromosome 8 locus had significant (FDR $\leq 1x10^{-6}$) chromatin interactions mapped to 26 different genes (Fig S8). These included *SAMD12*, *COLEC10*, *NOV*, and *MAL2*. Additionally, the top SNP in this locus, rs1915541, was identified as an eQTL for *SAMD12* in 10 different tissues, including esophagus mucosa and whole blood. The protective allele (T) is associated with lower expression of *SAMD12* in esophagus mucosa (normalized effect size (NES)= -0.40, $P=7.1x10^{-15}$) and higher expression of *SAMD12* in whole blood (NES= 0.38, $P=1.1x10^{-14}$). Finally, the chromosome 17 locus had significant (FDR $\leq 1x10^{-6}$) chromatin interactions mapped to 25 different genes (Fig S9). These included *DHX33*, *NLRP1*, *DERL2*, *WSCD1*, and *OR1G1*.

Discussion

In this study of Bangladeshi infants, we identified three loci which were protective against extreme frequency or duration of diarrhea in the first year of life. These infants lived in an urban and poor city slum with limited sanitation, thus they were exposed to a variety of pathogens

284 throughout the first year of life. We identified a small group in each cohort who despite these
285 exposures (surveillance sampling) did not have a diarrheal episode in the first year of life.
286 Furthermore, we noted an interaction between our top SNP, rs2827548, and access to
287 improved toilets, which demonstrated that protective genetic variants may not be sufficient to
288 guard against extreme illness when there is insufficient sanitation. The comprehensive
289 characterization of these children with household visits afforded us the opportunity to identify
290 these protective loci that may serve as therapeutic targets for diarrheal disease across
291 populations.

292 The first locus on chromosome 21 was shared between both frequency and duration of
293 diarrhea. The intronic variant in the lncRNA *AP000959* had several significant chromatin
294 interactions, including with the neural cell adhesion molecule 2 gene (*NCAM2)*. *NCAM2* is
295 involved in development of the enteric nervous system (ENS), and shows higher expression in
296 wild type mouse gut vs the aganglionic model used for Hirschsprung's disease [44–46]. The
297 enteric nervous system has many functions, including responding to sensory stimuli from the
298 wall of the bowel, controlling intestinal motility, regulating intestinal secretion, and controlling
299 intestinal blood flow. Furthermore, *NCAM2* is downregulated in ulcerative colitis [47], although it
300 is unclear whether this is a cause or consequence of disease.

301 Another significant locus, *SAMD12*, was identified for frequency of diarrhea and reached
302 suggestive associations for duration of diarrhea. *SAMD12* has high protein abundance in the gut
303 [42] and the intronic variant we identified is an eQTL for *SAMD12* in esophagus mucosa and
304 whole blood. Additionally, *SAMD12* is differentially expressed in a FOXO3 knockout model of
305 inflammatory colon cancer [48] and intronic repeat expansions in this gene are associated with
306 benign adult familial myoclonic epilepsy (MIM 601068).

307 An additional locus on chromosome 17 was significant in the analysis of days and
308 suggestive in the analysis of episodes. The lead SNP, rs5026214, is near the *WSCD1* gene,
309 which is downregulated in pediatric inflammatory bowel disease (IBD), both Crohn's and

ulcerative colitis [49]. According to the Human Protein Atlas, this protein is most abundant in the gastrointestinal tract [42]. Another SNP, rs146313367, upstream of *WSCD1* (but not in allelic association with our lead SNP) was suggestive in a GWAS of IBS in the UK Biobank [50]. This locus also had chromatin interactions with several genes that could be related to gut inflammation, including *DHX33*, *OR1G1*, and *NLRP1*.

It is remarkable that these novel genetic associations are distinct from our previous studies [14–17] of genetic susceptibility to specific pathogens. This suggests that these loci are not pathogen-specific but rather reflect pathology of protection from diarrhea generally. What may be distinct is that we have identified associations with genetic loci and not having diarrhea despite exposure to causative factors. This is also reflected in the similar GWAS results between extremes of the frequency of incident diarrheal episodes and duration due to the same children classified as not having diarrhea and thus not having days of diarrhea. However, there were loci that differed between the two analyses, indicating that protection from incident diarrhea may be different than protection for total duration of diarrhea. If we can identify infants at risk for repeat diarrheal episodes or prolonged bouts of diarrhea, we may be able to target treatment or therapeutics to limit both the acute and chronic effects of infant diarrhea.

There were limitations to this study. First, to get a clear distinction between those with diarrhea and those who did not have any diarrhea in this high exposure environment, we opted to analyze children at the extreme ends of the distributions. However, this limited our sample size, and power. Second, children get diarrhea for many different reasons, including pathogens, stress, dietary changes, and bowel disorders. We could not evaluate the cause of the diarrhea, but by using the extremes of the distribution we highlight those kids most prone to either frequent or long episodes.

In our genome-wide analyses of extreme frequency and prolonged duration of diarrhea in the first year of life in Bangladeshi infants, we identified protective genes with plausible functional links to the enteric nervous system, intestinal inflammation, and response to

pathogens. Understanding the genetic architecture of these variants may provide targets for treatment and prevention of diarrhea.

Funding

This work was supported by the Maryland Genetics Epidemiology and Medicine Training Program, funded by Burroughs-Wellcome Fund. Additional sources of funding were grants from the National Institute of Allergy and Infectious Disease [AI108790, AI043596], as well as research grants to W.A.P., Jr. from the Bill and Melinda Gates Foundation and the Henske family. Funders had no role in study design, analysis, or publication.

Acknowledgments

We thank the families of Mirpur that participated in these cohort studies, as well as the laboratory staff and field research assistants. icddr,b is grateful to the governments of Bangladesh, the UK, Sweden, and Canada for core unrestricted support.

References

1. Diarrhoeal disease [Internet]. [cited 2022 Jun 17]. Available from: https://www.who.int/news-room/fact-sheets/detail/diarrhoeal-disease

2. World Health Organization, United Nations Children's Fund (UNICEF). Ending preventable child deaths from pneumonia and diarrhoea by 2025 : the integrated global action plan for pneumonia and diarrhoea (GAPPD) [Internet]. Geneva: World Health Organization; 2013 [cited 2022 Jul 20]. Available from: https://apps.who.int/iris/handle/10665/79200

3. Platts-Mills JA, Babji S, Bodhidatta L, et al. Pathogen-specific burdens of community diarrhoea in developing countries: a multisite birth cohort study (MAL-ED). Lancet Glob Health. **2015**; 3(9):e564-575.

4. Anders KL, Thompson CN, Thuy NTV, et al. The epidemiology and aetiology of diarrhoeal disease in infancy in southern Vietnam: a birth cohort study. Int J Infect Dis. **2015**; 35:3–10.

5. Sarkar R, Gladstone BP, Warier JP, et al. Rotavirus and other diarrheal disease in a birth cohort from Southern Indian community. Indian Pediatr. **2016**; 53(7):583–588.

6. Eijsbouts C, Zheng T, Kennedy NA, et al. Genome-wide analysis of 53,400 people with irritable bowel syndrome highlights shared genetic pathways with mood and anxiety disorders. Nat Genet. Nature Publishing Group; **2021**; 53(11):1543–1552.

7. Serena G, Lima R, Fasano A. Genetic and Environmental Contributors for Celiac Disease. Curr Allergy Asthma Rep. **2019**; 19(9):40.

8. Garza-Hernandez D, Sepulveda-Villegas M, Garcia-Pelaez J, et al. A systematic review and functional bioinformatics analysis of genes associated with Crohn's disease identify more than 120 related genes. BMC Genomics. **2022**; 23(1):302.

9. Fachal L, O.B.O. the International IBD Genetics Consortium, on behalf of the International IBD Genetics Consortium. OP11 Expanded genome-wide association study of Inflammatory Bowel Disease identifies 174 novel loci and directly implicates new genes in disease susceptibility. Journal of Crohn's and Colitis. **2022**; 16(Supplement_1):i011–i013.

10. Müller T, Wijmenga C, Phillips AD, et al. Congenital sodium diarrhea is an autosomal recessive disorder of sodium/proton exchange but unrelated to known candidate genes. Gastroenterology. **2000**; 119(6):1506–1513.

11. Hartley JL, Zachos NC, Dawood B, et al. Mutations in TTC37 Cause Trichohepatoenteric Syndrome (Phenotypic Diarrhea of Infancy). Gastroenterology. **2010**; 138(7):2388-2398.e2.

12. Marionneau S, Ruvoën N, Le Moullac-Vaidye B, et al. Norwalk virus binds to histo-blood group antigens present on gastroduodenal epithelial cells of secretor individuals. Gastroenterology. **2002**; 122(7):1967–1977.

13. Tan M, Jiang X. Norovirus and its histo-blood group antigen receptors: an answer to a historical puzzle. Trends in Microbiology. **2005**; 13(6):285–293.

14. Wojcik GL, Marie C, Abhyankar MM, et al. Genome-wide association study reveals genetic link between diarrhea-associated Entamoeba histolytica infection and inflammatory bowel disease. mBio. **2018**; 9(5).

15. Wojcik GL, Korpe P, Marie C, et al. Genome-wide association study of cryptosporidiosis in infants implicates PRKCA. mBio. **2020**; 11(1).

16. Duchen D, Haque R, Chen L, et al. Host genome wide association study of infant susceptibility to Shigella-associated diarrhea. Infection and Immunity. **2021**; 89(6):e00012-21.

17. Munday RM, Haque R, Jan N-J, et al. Genome-Wide Association Study of Campylobacter-Positive Diarrhea Identifies Genes Involved in Toxin Processing and Inflammatory Response. mBio. **2022**; 13(3):e0055622.

18. Bustamante M, Standl M, Bassat Q, et al. A genome-wide association meta-analysis of diarrhoeal disease in young children identifies FUT2 locus and provides plausible biological pathways. Hum Mol Genet. **2016**; 25(18):4127–4142.

19. Mondal D, Minak J, Alam M, et al. Contribution of Enteric Infection, Altered Intestinal Barrier Function, and Maternal Malnutrition to Infant Malnutrition in Bangladesh. Clinical Infectious Diseases. **2012**; 54(2):185–192.

20. Kirkpatrick BD, Colgate ER, Mychaleckyj JC, et al. The "Performance of Rotavirus and Oral Polio Vaccines in Developing Countries" (PROVIDE) study: description of methods of an interventional study designed to explore complex biologic problems. The American Journal of Tropical Medicine and Hygiene. **2015**; 92(4):744–751.

21. Steiner KL, Ahmed S, Gilchrist CA, et al. Species of Cryptosporidia causing subclinical infection associated with growth faltering in rural and urban Bangladesh: a birth cohort study. Clin Infect Dis. **2018**; 67(9):1347–1355.

22. Indicator Metadata Registry Details [Internet]. [cited 2022 Jul 7]. Available from: https://www.who.int/data/gho/indicator-metadata-registry/imr-details/4744

23. Improved sanitation facilities and drinking-water sources [Internet]. [cited 2022 Jul 7]. Available from: https://www.who.int/data/nutrition/nlis/info/improved-sanitation-facilities-and-drinking-water-sources

24. Delaneau O, Zagury J-F, Marchini J. Improved whole-chromosome phasing for disease and population genetic studies. Nat Methods. **2013**; 10(1):5–6.

25. Delaneau O, Marchini J, McVean G. Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel. Nat Commun. **2014**; 5(1):1–9.

26. Marchini J, Howie B, Myers S, McVean G, Donnelly P. A new multipoint method for genome-wide association studies by imputation of genotypes. Nat Genet. **2007**; 39(7):906–913.

27. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLOS Genetics. **2009**; 5(6):e1000529.

28. Marchini J, Howie B. Genotype imputation for genome-wide association studies. Nat Rev Genet. **2010**; 11(7):499–511.

29. Howie B, Marchini J, Stephens M. Genotype imputation with thousands of genomes. G3: Genes, Genomes, Genetics. **2011**; 1(6):457–470.

30. Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat Genet. **2012**; 44(8):955–959.

31. Auton A, Abecasis GR, Altshuler DM, et al. A global reference for human genetic variation. Nature. **2015**; 526(7571):68–74.

32. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience. **2015**; 4(1):7.

33. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen W-M. Robust relationship inference in genome-wide association studies. Bioinformatics. **2010**; 26(22):2867–2873.

34. Burton PR, Clayton DG, Cardon LR. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. **2007**; 447(7145):661–678.

35. Liu JZ, Tozzi F, Waterworth DM, et al. Meta-analysis and imputation refines the association of 15q25 with smoking quantity. Nat Genet. **2010**; 42(5):436–440.

36. Pruim RJ, Welch RP, Sanna S, et al. LocusZoom: regional visualization of genome-wide association scan results. Bioinformatics. **2010**; 26(18):2336–2337.

37. Watanabe K, Taskesen E, Bochoven A van, Posthuma D. Functional mapping and annotation of genetic associations with FUMA. Nat Commun. Nature Publishing Group; **2017**; 8(1):1826.

449 38. Leeuw CA de, Mooij JM, Heskes T, Posthuma D. MAGMA: Generalized Gene-Set
450 Analysis of GWAS Data. PLOS Computational Biology. Public Library of Science; **2015**;
451 11(4):e1004219.

452 39. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from
453 high-throughput sequencing data. Nucleic Acids Research. **2010**; 38(16):e164.

454 40. Boyle AP, Hong EL, Hariharan M, et al. Annotation of functional variation in personal
455 genomes using RegulomeDB. Genome Res. **2012**; 22(9):1790–1797.

456 41. GTEx Consortium. Genetic effects on gene expression across human tissues. Nature. **2017**;
457 550(7675):204–213.

458 42. Uhlén M, Fagerberg L, Hallström BM, et al. Proteomics. Tissue-based map of the human
459 proteome. Science. **2015**; 347(6220):1260419.

460 43. Marcus JH, Novembre J. Visualizing the geography of genetic variants. Bioinformatics.
461 **2017**; 33(4):594–595.

462 44. Vohra BPS, Tsuji K, Nagashimada M, et al. Differential gene expression and functional
463 analysis implicates novel mechanisms in enteric nervous system precursor migration and
464 neuritogenesis. Dev Biol. **2006**; 298(1):259–271.

465 45. Schriemer D, Sribudiani Y, IJpma A, et al. Regulators of gene expression in Enteric Neural
466 Crest Cells are putative Hirschsprung disease genes. Developmental Biology. **2016**;
467 416(1):255–265.

46. Avetisyan M. Development of Enteric Neurons and Muscularis Macrophages. Arts & Sciences Electronic Theses and Dissertations. **2019;** 1781.

47. Schniers A, Anderssen E, Fenton CG, et al. The Proteome of Ulcerative Colitis in Colon Biopsies from Adults - Optimized Sample Preparation and Comparison with Healthy Controls. PROTEOMICS – Clinical Applications. **2017**; 11(11–12):1700053.

48. Penrose HM, Cable C, Heller S, et al. Loss of Forkhead Box O3 Facilitates Inflammatory Colon Cancer: Transcriptome Profiling of the Immune Landscape and Novel Targets. Cell Mol Gastroenterol Hepatol. **2018**; 7(2):391–408.

49. Fang K, Grisham MB, Kevil CG. Application of Comparative Transcriptional Genomics to Identify Molecular Targets for Pediatric IBD. Front Immunol. **2015**; 6:165.

50. Bonfiglio F, Zheng T, Garcia-Etxebarria K, et al. Female-specific Association Between Variants on Chromosome 9 and Self-reported Diagnosis of Irritable Bowel Syndrome. Gastroenterology. **2018**; 155(1):168–179.

482 **Table 1: Demographics by phenotype and cohort.**

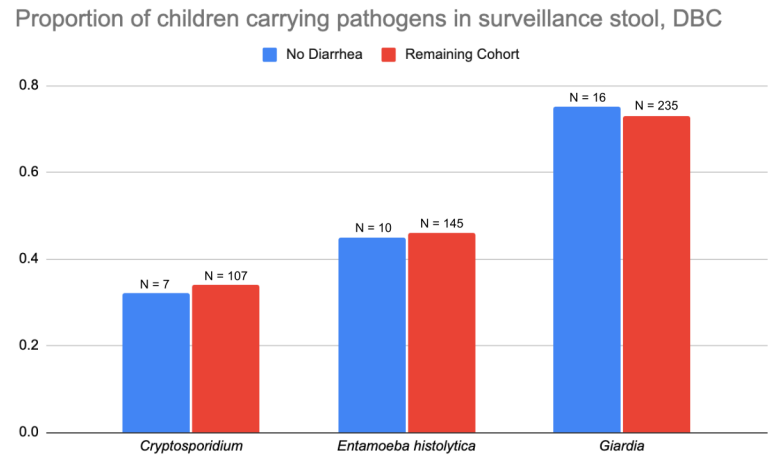
	DBC			PROVIDE			CBC		
	No Diarrhea N = 26	6+ diarrheal episodes N = 96	25+ days of diarrhea N = 70	No Diarrhea N = 32	6+ diarrheal episodes N = 117	25+ days of diarrhea N = 130	No Diarrhea N = 32	6+ diarrheal episodes N = 80	25+ days of diarrhea N = 93
Female, %	58	47	47	66	39	41	59	54	48
Maternal age at enrollment, mean (sd)	25.6 (5.93)	25.5 (5.01)	25.4 (5.30)	25.2 (4.74)	24.5 (4.37)	24.4 (4.14)	24.6 (5.06)	25.0 (4.66)	24.8 (4.81)
Household size, mean (sd)	5.77 (2.76)	5.43 (1.93)	5.41 (2.34)	5.28 (2.29)	5.44 (2.35)	5.35 (2.29)	5.41 (2.62)	5.54 (2.39)	5.24 (2.20)
Household income, median (IQR)	7.50 (6.05-8.87)	5.50 (4.29-8.00)	5.35 (4.22-7.15)	10.0 (7.75-15.0)	9.00 (6.00-14.0)	9.00 (6.50-15.0)	15.0 (9.00-21.2)	12.0 (10.0-18.0)	14.0 (10.0-18.0)
Improved toilet, %	65	28	24	100	94	92	88	99	98
Birth WLZ, mean	-1.51 (1.01)	-1.28 (1.27)	-1.35 (1.33)	-1.20 (0.78)	-1.39 (0.99)	-1.37 (0.96)	-0.91 (1.41)	-1.26 (1.17)	-1.39 (1.19)
Birth LAZ, mean	-0.52 (0.94)	-1.05 (1.03)	-1.10 (1.05)	-1.09 (0.78)	-0.85 (0.95)	-0.85 (0.94)	-1.05 (0.96)	-0.92 (1.04)	-0.92 (1.00)
1 year WLZ, mean	-0.94 (0.79)	-1.04 (1.15)	-1.14 (1.11)	-0.82 (0.92)	-0.62 (1.11)	-0.59 (1.03)	-0.54 (0.99)	-0.32 (1.14)	-0.38 (1.18)
1 year LAZ, mean	-1.38 (0.98)	-1.81 (0.98)	-1.89 (0.93)	-1.62 (1.24)	-1.58 (1.13)	-1.58 (1.11)	-1.48 (0.84)	-1.38 (0.98)	-1.23 (0.93)
Age at first diarrheal episode, median (IQR)	N/A	50.0 (23.0-78.0)	39.0 (21.2-76.7)	N/A	58.0 (31.0-93.0)	58.0 (29.2-96.7)	N/A	62.5 (33.0-86.5)	66.0 (39.0-91.0)

Household income is reported in thousands of Bangladeshi taka (BDT). Improved toilet includes sewer or septic tank, water-sealed or slab latrines, and ventilated-improved pit latrines. WLZ, weight-for-length Z score. LAZ, length-for-age Z score. Shading indicates phenotype. The white columns are those with no diarrhea, the light gray columns are those with 6 or more diarrheal episodes in the first year, and the darker gray columns are those with 25 or more days of diarrhea in the first year. The two diarrhea groups are not mutually exclusive.

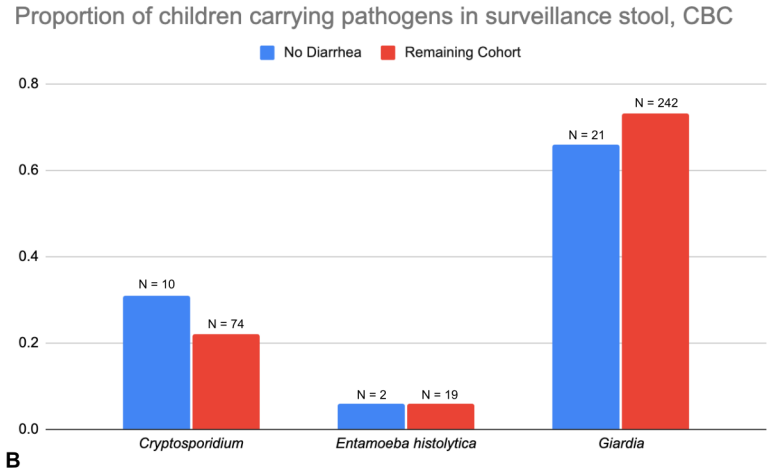
Table 2: Odds ratios for top SNP at each locus show similarity across cohorts.

			DBC			PROVIDE			CBC			META		
Diarrheal episodes, 0 vs 6+														
rsID	**Chr:Pos**	**A0:A1**	**EAF**	**OR (95 % CI)**	***P***	**EAF**	**OR (95% CI)**	***P***	**EAF**	**OR (95% CI)**	***P***	**EAF**	**OR (95% CI)**	***P***	P_{het}
rs2827548	21:23878342	G:C	0.29	0.52 (0.24-1.15)	0.11	0.26	0.21 (0.10-0.43)	1.48E-05	0.24	0.31 (0.15-0.60)	6.57E-04	0.26	0.31 (0.21-0.47)	4.02E-08	0.24
rs1915541	8:119605603	C:T	0.24	0.64 (0.27-1.53)	0.31	0.23	0.31 (0.15-0.62)	1.02E-03	0.30	0.28 (0.15-0.52)	6.42E-05	0.25	0.35 (0.23-0.53)	4.74E-07	0.30
rs6841521	4:30536556	C:T	0.27	0.27 (0.12-0.63)	2.27E-03	0.33	0.35 (0.20-0.62)	3.04E-04	0.31	0.53 (0.29-0.94)	0.03	0.31	0.39 (0.27-0.57)	5.26E-07	0.40
Days of diarrhea, 0 vs 25+														
rsID	**Chr:Pos**	**A0:A1**	**EAF**	**OR (95 % CI)**	***P***	**EAF**	**OR (95% CI)**	***P***	**EAF**	**OR (95% CI)**	***P***	**EAF**	**OR (95% CI)**	***P***	P_{het}
rs2827548	21:23878342	G:C	0.27	0.32 (0.14-0.75)	7.99E-03	0.26	0.25 (0.13-0.49)	5.23E-05	0.24	0.37 (0.19-0.70)	2.42E-03	0.26	0.31 (0.21-0.46)	1.59E-08	0.72
rs5026214	17:6030892	T:C	0.57	0.41 (0.19-0.84)	0.02	0.54	0.25 (0.13-0.48)	3.31E-05	0.53	0.43 (0.23-0.81)	9.47E-03	0.55	0.35 (0.24-0.51)	1.09E-07	0.46
rs11635981	15:58569265	G:C	0.11	0.32 (0.11-0.91)	0.03	0.09	0.54 (0.19-1.58)	0.26	0.13	0.15 (0.07-0.32)	1.81E-06	0.11	0.25 (0.15-0.44)	7.26E-07	0.13
rs1873995	15:57879990	A:G	0.44	0.35 (0.17-0.69)	2.61E-03	0.36	0.58 (0.30-1.11)	0.10	0.40	0.31 (0.17-0.57)	1.28E-04	0.40	0.39 (0.27-0.57)	7.80E-07	0.36

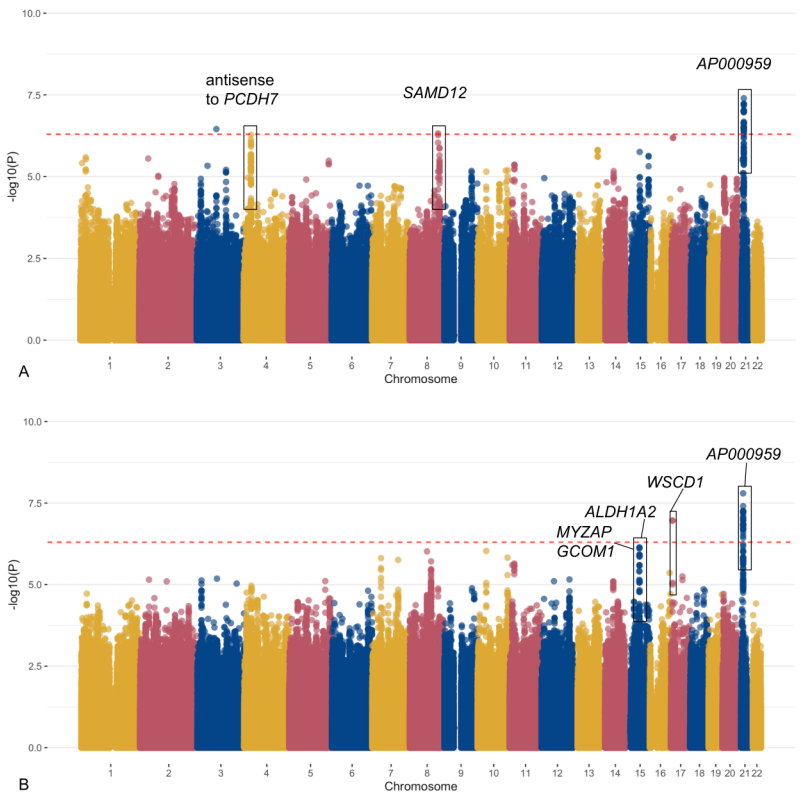
rsID, reference SNP identifier; Chr:Pos, chromosome:position on hg19; A0:A1, non-effect allele:effect allele; EAF, effect allele frequency; OR, odds ratio; 95% CI, 95% confidence interval; *P*, *P* value from frequentist association test in SNPTEST for each cohort, and *P* value from meta-analysis in META; P_{het}, *P* value of heterogeneity from Cochran's Q in META. SNPs are arranged by analysis and then *P* value in the meta-analysis. Upper portion shows the top SNP for each locus outlined in the Manhattan for the analysis of diarrheal frequency, 0 vs 6+ episodes. Lower portion shows the top SNP for each locus outlined in the Manhattan for the analysis of duration of diarrhea, 0 vs 25+ days.


Proportion of children carrying pathogens in surveillance stool, DBC
No Diarrhea
Remaining Cohort
0.8
0.6
0.4
0.2
0.0
N = 7
N = 107
N = 10
N = 145
N = 16
N = 235
Cryptosporidium
Entamoeba histolytica
Giardia

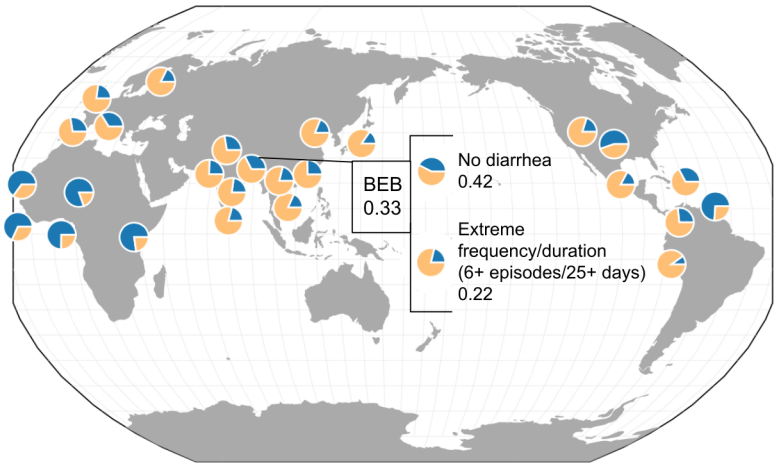
A


Proportion of children carrying pathogens in surveillance stool, CBC
No Diarrhea
Remaining Cohort
0.8
0.6
0.4
0.2
0.0
N = 10
N = 74
N = 2
N = 19
N = 21
N = 242
Cryptosporidium
Entamoeba histolytica
Giardia

B


10.0
7.5
5.0
2.5
0.0
-log10(P)
antisense
to PCDH7
SAMD12
AP000959
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
Chromosome
A
10.0
7.5
5.0
2.5
0.0
-log10(P)
AP000959
WSCD1
ALDH1A2
MYZAP
GCOM1
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
Chromosome
B

chr21:23878342 C/G



Frequency Scale = Proportion out of 1

The pie below represents a minor allele frequency of 0.25

Fig 1. Proportion of children carrying pathogens in surveillance stool. (A) Surveillance stool samples from DBC. Blue bars represent proportion of kids without diarrhea in the first year who tested positive for a given pathogen at least once in surveillance (total N = 22). Red bars represent proportion of remaining cohort who tested positive for a given pathogen at least once in surveillance (total N = 314). (B) Surveillance stool samples from CBC. Blue bars represent proportion of kids without diarrhea in the first year who carried a given pathogen in surveillance stool (total N = 32). Red bars represent proportion of remaining cohort who carried pathogen in surveillance stool (total N = 333).

Fig 2. Manhattan plots of genome-wide association meta-analyses. (A) This comparison is having no diarrheal episodes versus 6 or more diarrheal episodes in the first year of life. Three peaks are identified, one suggestive and two significant. (B) This comparison is no days of diarrhea versus 25 or more days of diarrhea. Four peaks are identified, chromosomes 17 and 21 are genome-wide significant. Chromosome 15 encompasses two suggestive peaks. Each point is a single variant. The x-axis is chromosomal position, and the y-axis is the $-\log_{10}(P)$. The red dashed line indicates the genome-wide significance threshold, $P = 5 \times 10^{-7}$.

Fig 3. Frequency of chromosome 21 effect allele compared to global populations. Each pie chart represents the frequencies of the C (blue) and G (yellow) alleles in the 1000 Genomes Project population closest to its location on the map. Some circles are slightly shifted to avoid overlap. BEB is the Bengali in Bangladesh population, and to the right of that are the allele frequencies in our meta-analyses.

1. We declare no conflicts of interest.
2. Funding for this work was provided to P.D. and R.M. by the Maryland Genetics Epidemiology and Medicine Training Program, supported by the Burroughs-Wellcome Fund. Additional sources of funding were grants from the National Institute of Allergy and Infectious Disease (AI108790, AI043596), as well as research grants to W.A.P., Jr. from the Bill and Melinda Gates Foundation and the Henske family. Funders had no role in study design, analysis, or publication.
3. Part of this work will be presented at the ASTMH annual meeting, October 2022, in Seattle, WA (Abstract 1214).
4. Corresponding author: Priya Duggal 615 N. Wolfe Street, Baltimore MD 21205, fax 410-955-0863, telephone 410-955-1213, email pduggal@jhu.edu