HMG Advance Access published August 23, 2016

A genome-wide association meta-analysis of diarrhoeal disease in young children identifies *FUT2* locus and provides plausible biological pathways

Mariona Bustamante* [1, 2, 3, 4], Marie Standl [5], Quique Bassat [6, 7], Natalia Vilor-Tejedor [1, 3, 4], Carolina Medina-Gomez [8, 9, 10], Carolina Bonilla [11, 12], Tarunveer S Ahluwalia [13], Jonas Bacelis [14], Jonathan P. Bradfield [15], Carla M.T. Tiesler [5, 16], Fernando Rivadeneira [8, 9, 10], Susan Ring [11,12], Nadja H. Vissing [13], Nadia R. Fink [13], Astanand Jugessur [17], Frank D. Mentch [15], Ferran Ballester [18, 4], Jennifer Kriebel [19, 20], Jessica C. Kiefte-de Jong[10, 21, 22], Helene M. Wolsk [13], Sabrina Llop [18,4], Elisabeth Thiering [5, 16], Systke A. Beth [8, 21], Nicholas J. Timpson [11, 12], Josefine Andersen [13], Holger Schulz [5], Vincent W.V. Jaddoe [8, 21], David M. Evans [11, 12, 23], Johannes Waage [13], Hakon Hakonarson [15, 24, 25], Struan F.A. Grant [15, 24, 25, 26], Bo Jacobsson [14, 17], Klaus Bønnelykke [13], Hans Bisgaard [13], George Davey Smith [11, 12], Henriette A. Moll [21], Joachim Heinrich [5, 27], Xavier Estivill [2, 3, 28, 4, 29], Jordi Sunyer [1, 3, 28, 4]

1 ISGlobal, Center for Research in Environmental Epidemiology (CREAL), Barcelona, Spain.

2 Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Barcelona, Spain.

3 Pompeu Fabra University (UPF), Barcelona, Spain.

4 CIBER Epidemiología y Salud Pública (CIBERESP), Barcelona, Spain.

5 Institute of Epidemiology I, Helmholtz Zentrum München - German Research Center for Environmental Health, Neuherberg, Germany.

6 ISGlobal, Barcelona Ctr. Int. Health Res. (CRESIB), Hospital Clínic, Universitat de Barcelona, Barcelona, Spain.

7 Centro de Investigação em Saúde de Manhiça (CISM), Maputo, Mozambique.

8 The Generation R Study Group, Erasmus MC, Rotterdam, The Netherlands.

9 Department of Internal Medicine, Erasmus MC, Rotterdam, The Netherlands.

10 Department of Epidemiology, Erasmus MC, Rotterdam, The Netherlands.

11 MRC/University of Bristol Integrative Epidemiology Unit, University of Bristol, Bristol, UK.

12 School of Social and Community Medicine, University of Bristol, Bristol, UK.

13 COPSAC, Copenhagen Prospective Studies on Asthma in Childhood, Herlev and Gentofte Hospital, University of Copenhagen, Copenhagen, Denmark.

14 Department of Obstetrics and Gynecology, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden.

15 Center for Applied Genomics, The Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, USA.

16 Ludwig-Maximilians-University of Munich, Dr. von Hauner Children's Hospital, Division of Metabolic Diseases and Nutritional Medicine, Munich, Germany.

17 Department of Genetics and Bioinformatics, Area of Health Data and Digitalisation, Institute of Public Health, Oslo, Norway.

18 Epidemiology and Environmental Health Joint Research Unit, FISABIO–Universitat Jaume I–Universitat de València, València, Spain.

19 Research Unit of Molecular Epidemiology, Helmholtz Zentrum München - German Research Center for Environmental Health, Neuherberg, Germany.

20 Institute of Epidemiology II, Helmholtz Zentrum München - German Research Center for Environmental Health, Neuherberg, Germany.

21 Department of Pediatrics, Erasmus MC, Rotterdam, The Netherlands.

22 Leiden University College, The Hague, The Netherlands.

23 University of Queensland Diamantina Institute, Translational Research Institute, Brisbane, Queensland, Australia.

24 Division of Human Genetics, The Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, USA.

25 Department of Pediatrics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA.

26 Division of Endocrinology, The Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, USA.27 Institute and Outpatient Clinic for Occupational, Social and Environmental Medicine, Inner City Clinic, University Hospital of Munich (LMU), Munich, Germany.

28 IMIM (Hospital del Mar Medical Research Institute), Barcelona, Spain

29 Experimental Genetics, Sidra Medical and Research Centre, Doha, Qatar.

The authors wish it to be known that, in their opinion, the two first and last two authors should be regarded as joint First and Last Authors.

*** Corresponding author**: Dr Aiguader 88, 08003, Barcelona, Spain, Tel +34 93 214 73 00, Fax +34 93 214 73 02, mbustamante@creal.cat

Contributions

Study design: M.B., M.S., N.V.T., J.P.B., F.D.M., F.B., S.L., H.S., V.W.V.J., H.H., S.F.A.G., K.B., H.B., H.A.M., G.D.S., H.B., A.J., B.J., J.H., X.E., J.S.

Phenotyping: A.J., J.P.B., N.H.V., N.R.F., B.J., F.D.M., F.B., J.C.K-dJ., H.M.W., S.L., E.T., S.B., J.A., H.H., S.F.A.G., J.B., K.B., H.B., J.H., J.S.

Genotyping: M.B., C.M-G., C.M.T.T., F.R., S.R., B.J., J.K., N.J.T., D.M.E., J.W., H.H., S.F.A.G., J.B., K.B., H.B., X.E.

Analysis: M.B., M.S., N.V-T., C.M-G., C.B., T.S.A., J.P.B., F.R., F.D.M., J.C.K., J.B.

Writing manuscript: M.B., M.S., Q.B., X.E., J.S.

Revising and reviewing paper: M.B., M.S., Q.B., N.V-T., C.M-G., C.B., T.S.A., A.J., J.P.B., C.M.T.T., F.R., S.R., N.H.V., N.R.F., B.J., F.D.M., F.B., J.K., J.C.K-dJ., H.M.W., S.L., E.T., S.B., N.J.T., J.A., H.S., V.V.W.J., D.M.E., J.W., H.H., S.F.A.G., G.D.S, J.B., K.B., H.B., H.A.M., J.H., X.E., J.S.

Abstract

More than a million childhood diarrhoeal episodes occur worldwide each year, and in developed countries a considerable part of them are caused by viral infections. In this study we aimed to search for genetic variants associated with diarrhoeal disease in young children by meta-analyzing genome-wide association studies, and to elucidate plausible biological mechanisms.

The study was conducted in the context of the Early Genetics and Lifecourse Epidemiology (EAGLE) consortium. Data about diarrhoeal disease in two time windows (around one year of age and around two years of age) was obtained via parental questionnaires, doctor interviews or medical records. Standard quality control and statistical tests were applied to the 1000 Genomes imputed genotypic data.

The meta-analysis (N=5,758) followed by replication (N=3,784) identified a genome-wide significant association between rs8111874 and diarrhoea at age one year. Conditional analysis suggested that the causal variant could be rs601338 (W154X) in the *FUT2* gene. Children with the A allele, which results in a truncated FUT2 protein, had lower risk of diarrhoea. FUT2 participates in the production of histo-blood group antigens and has previously been implicated in the susceptibility to infections, including *Rotavirus* and *Norovirus*. Gene-set enrichment analysis suggested pathways related to the histo-blood group antigen production, and the regulation of ion transport and blood pressure. Among others, the gastrointestinal tract, and the immune and neuro-secretory systems were detected as relevant organs.

In summary, this genome-wide association meta-analysis suggests the implication of the *FUT2* gene in diarrhoeal disease in young children from the general population.

Introduction

Diarrhoea, defined as three or more loose stools within the previous 24 hours, is probably one of the most common symptoms in children, with an estimated 1,370 million annual episodes in children younger than five years in 2010 (1). Two percent of these episodes progress to severe disease, and 700,000 episodes lead to death, mainly in low-income countries (1). In developed countries, diarrhoeal disease is a common reason for attendance at a general practitioner, especially in children under five years of age (2).

Several pathogens can account for infections associated with diarrhoeal disease, including viruses, bacteria and parasites. The GEMS study (Global Enteric Multicenter Study), conducted as a case-control study in seven African and Asian sites, identified *Rotavirus*, *Cryptosporidium*, enterotoxigenic *Escherichia coli*, and *Shigella* as most responsible attributable pathogens for cases of moderate-to-severe diarrhea (3). In developed countries, improvements in public health infrastructure (water and sewage management), has caused a shift in the main causes of acute pediatric diarrhoea, moving from bacterial and parasite etiologies to viruses. *Rotavirus* is the primary cause of diarrhoeal disease globally, and is responsible for almost half of the gastroenteritis cases requiring hospitalization in Western countries, followed by *Norovirus*, *Adenovirus* and *Salmonella* (4). The peak age for infection with *Rotavirus* is between three months and two years, coinciding with reduced protection by transplacental transfer of maternal antibodies (5) and the end of the lactation period (6). At the age of five, almost all children have been infected with *Rotavirus*, and progressively develop natural immunity against this virus (7). Enteric pathogens damage small bowel enterocytes and cause impaired intestinal absorption, low grade fever and watery diarrhoea as a result of the deregulation of ion transport and stimulation of the enteric nervous system (8, 9). In children, factors other than enteric pathogens can cause diarrhoea, including among others extra-intestinal infections, intolerances or food allergies (10), nutrient deficiencies, antimicrobials, or hereditary diseases, such as cystic fibrosis, but these represent a small proportion of all cases of diarrhoea.

The exposure to infectious agents plays a major role in the acquisition of the pathogen and development of diarrhoeal disease. However, not all individuals are equally susceptible to

infection, and if infected, they may differ in their immunological response. Host genetic factors can explain part of the differences in susceptibility and response to infection. In a study on families with adopted children, it was shown that premature death due to infection of the biological parent, but not of the adopting parent, increased the relative risk of death due to infection in the adopted child by 5.8 fold, a higher risk than observed for vascular disease or cancer (11). Twin studies on infectious diseases have shown higher correlation between monozygotic than dizygotic twins (12-14). Heritability for early childhood diarrhoea was estimated to be 54% in a pedigree-based design in Brazil (15).

Candidate gene studies for infectious gastroenteritis have identified genetic variants in genes involved in the innate and acquired immune responses and in genes that participate in the production of histo-blood group antigens (HBGAs), which serve as receptors for numerous pathogens. In particular, the non-secretor phenotype, associated with null or inactivating mutations in *FUT2* gene resulting in a lack of certain antigens in secretions and epithelial mucosa, confers strain-specific protection against *Norovirus* (16-18) and *Rotavirus* (19). A review of the associations between enteric pathogens and genetic variants can be found elsewhere (20, 21).

The aims of the present study were: 1) to identify genetic variants that confer susceptibility to diarrhoeal disease in young children from the general population of developed countries through a genome-wide association meta-analysis, and 2) to elucidate potential biological mechanisms involved in diarrhoeal disease using pathway analysis approaches.

Results

Sample

Four different traits, any diarrhoea and doctor's confirmed diagnosis of diarrhoea around 1 year of age (D1Y and DD1Y) and around 2 years of age (D2Y and DD2Y), were explored. Samples analyzed for each outcome in the discovery and in the replication phase are shown in Table 1. Information on diarrhoea was collected through questionnaires, doctor interviews or medical records (see Supplementary Material – Annex A).

In the discovery phase, 46.8% of the children had had at least one diarrhoeal episode around the age of one year (D1Y), while this proportion decreased to 21.8% for the cases diagnosed by a doctor (DD1Y). The proportions for children around the age of two years were 50.9% for diarrhoeal disease (D2Y) and 20.6% for doctor's diagnosis (DD2Y). Similar frequencies of diarrhoea were observed in the replication samples. The number of diarrhoea cases in Generation R and CHOP was lower than in other cohorts. For some of the diarrhoeal definitions, ALSPAC and GINIplus showed higher proportion of cases among males than among females.

Discovery phase

The Q-Q plots and the Manhattan plots for each outcome are shown in Figure 1 (D1Y) and in the Supplementary Material Figures S1 (DD1Y), S2 (D2Y) and S3 (DD2Y). Genomic inflation factor, lambda (λ), ranged from 0.9952 to 1.0031.

For D1Y, the meta-analysis of 5,758 samples revealed a genome-wide significant variant, rs8111874, at 19q13.33 [odds ratio (OR) (95% confidence interval (CI))=1.32 (1.21-1.44); p-value=1.06E-09] (Table 2, Figure 2). This lead SNP was also nominally associated with DD1Y [OR (95%CI)=1.26 (1.14-1.39); p-value=1.05E-05] and with DD2Y [OR (95%CI)=1.32 (1.17-1.48); p-value=4.6E-07], and with D2Y [OR (95%CI)=1.08 (1.00-1.18); p-value=6.56E-02]. In D1Y, the G allele increased the risk of diarrhoeal disease in all the studies (p-value for heterogeneity=2.99E-02, Table 2).

Another locus at 4q21.23, was associated with DD1Y at genome-wide significance [N=6,403, OR (95%CI)=1.31 (1.19-1.44); p-value=2.92E-08] (Table S1). The lead variant, rs1481779, was

located in an intron of *ARHGAP24* gene. This variant was also nominally associated with D1Y [OR (95%CI)=1.09 (1.01-1.19); p-value=3.7E-02], but not with D2Y or DD2Y.

Variants with a p-value<1E-05 are shown in Table 2 (D1Y) and in Supplementary Tables S1 (DD1Y), S2 (D2Y) and S3 (DD2Y). With the exception of the signal at 19q13.33, no major overlap was observed among the suggestive variants for each outcome (data not shown). A description of the potential function of the genetic variants and genes in each loci detected at a p value < 1E-05 can be found in Supplementary Tables S4-S7.

Replication phase

Seventy-two loci (p-value<1E-05 in any of the four diarrhoea definitions) were followed for replication in an independent dataset (Table 1). After multiple-testing adjustment, the SNP rs8111874 at chromosome 19q13.33 was associated with D1Y [replication: OR (95%CI)=1.25 (1.13-1.39); p-value=1.69E-05] (Table 2), and it was nominally associated for the other diarrhoea outcomes (Figure 2). Other variants at the same locus were also associated with DD1Y and DD2Y (Supplementary Tables S1 and S3). None of the variants in other loci replicated, and neither of them reached genome-wide significance in the combined analyses. Results of the replication phase can be found in Table 2 (D1Y) and in Supplementary Tables S1 (DD1Y), S2 (D2Y) and S3 (DD2Y).

Chromosome 19 locus: FUT2 gene

The regional association plot of rs8111874 at 19q13.33 shows a linkage disequilibrium block that overlaps several genes, including *FUT2* (Figure 3A). *FUT2* participates in the production of histo-blood group antigens (HBGAs) and contains a stop mutation (rs601338, W154X) known to confer protection against certain infections. We conditioned the top SNP at 19q13.33 (rs8111874) on the stop mutation (rs601338) and *viceversa*: the odds ratios from these analyses were attenuated (Table 3). After conditioning to rs601338, no secondary signals were observed in the region (window size 1 Mb) (Figure 3B). The forestplots for rs601338 (W154X) are shown in Supplementary Figure S4.

Enrichment analysis

Two prediction programs, the Meta-Analysis Gene-set Enrichment of variaNT Associations (MAGENTA) and the Data-driven Expression Prioritized Integration for Complex Traits (DEPICT) were used to investigate gene-sets enriched among the variants with the lowest p-values. Using MAGENTA, the "KEGG_GLYCOSPHINGOLIPID_BIOSYNTHESIS_GLOBO_SERIES" gene-set was identified in DD2Y at a 5% False Discovery Rate (FDR) (Supplementary Table S8). Six out of 14 genes of this pathway had SNPs in the 95th percentile of lowest p-values. This gene-set was also weakly associated with diarrhoea at age one year (D1Y). Other gene-sets with nominal evidence are listed in Supplementary Table S8, including the "KEGG_GLYCOSPHINGOLIPID_BIOSYNTHESIS_GANGLIO_SERIES" and the blood pressure regulatory gene-set "KEGG_RENIN_ANGIOTENSIN_SYSTEM".

The top 10 gene-sets detected with DEPICT for each outcome are shown in Supplementary Table S9. The gene-sets "ENSG00000147955 - SIGMAR1 PPI subnetwork" (involved in ion channels regulation and modulation of neurotransmitter release), "MP:0001675 - abnormal ectoderm development", "GO:0007492 - endoderm development" and "ENSG00000140612 - SEC11A PPI subnetwork" (component of the microsomal signal peptidase complex) remained after multiple-testing for DD1Y. Genetic variants showing suggestive p-values were linked to genes with enriched expression in the gastrointestinal tract (D1Y), the immune system (DD1Y) and in the neuro-secretory system (D2Y), among others (Supplementary Table S10). Exclusion of genetic variants in 1 Mb around the *FUT2* locus gave similar gene-sets and tissue/organs, but p-values were attenuated (data not shown).

Overlap with known variants and genes for related diseases

We compared the results from this study with variants reported in the literature as associated with inflammatory bowel disease (IBD) risk (22) and with viral infection and response to vaccination (23-36). Nine out of the 162 variants identified in IBD were nominally associated with D1Y (Supplementary Table S11). Two of these variants passed the multiple-testing correction: one of them located in the *FUT2* gene (p-value=4.67E-09, opposite effect direction), and the other one in *CARD11* locus (p-value=1.52E-05, same effect direction). In addition, one

out of 139 variants associated with infections was also associated with D2Y after correction for multiple-testing (Supplementary Table S12). Specifically, the A allele of rs17793829, located in *TTC7B* gene, was associated with higher anti *Cytomegalovirus* IgG titer (31) and higher risk for D2Y (p value=2.12E-04). Finally, we evaluated 86 genes retrieved from OMIM (Online Mendelian Inheritance in Man) with the entry "diarrhea". None of them was associated with childhood diarrhoeal disease after multipletesting correction (Supplementary Table S13).

Discussion

This study suggested the implication of the *FUT2* locus in the diarrhoeal risk and provided evidences supporting the role of the histo-blood group antigen (HBGA) production and the regulation of ion transport pathways. The gastrointestinal tract, and the immune and neuro-secretory systems were detected as relevant organs.

The genome-wide association meta-analysis followed by replication identified an association between rs8111874 mapping to the 19q13.33 locus and diarrhoea around one year of age (D1Y). Although with different strength, the association was also observed in all the different diarrhoeal outcomes investigated. The rs8111874 variant is located in an intronic region of the *NTN5* (*Netrin 5*) and *SEC1P* (*Secretory Blood Group 1, Pseudogene*) genes and close to *FUT2*, previously associated with susceptibility to infection. *FUT2* encodes the *Fucosyltransferase 2* enzyme that participates in the production of histo-blood group antigens (HBGAs) by catalyzing the addition of a fucose residue in α1,2 linkage to the galactose of O- or N-glycoproteins and globo-, ganglio- or lacto-series of glycolipids (37). The FUT2 enzymatic activity is polymorphic, exhibiting the non-secretor phenotype (lack of certain antigens in the gut and epithelial mucosa) when inactivating mutations are present in the *FUT2* gene. The most common *FUT2* inactivating variant in Caucasians (Europeans and Iranians) and in Africans, is the stop mutation W154X (rs601338), while in Asians it is the missense variant A385T (rs1047781) (38). In order to investigate whether the signal observed at 19q13.33 locus could be caused by the stop mutation W154X (rs601338), the lead SNP in the region (rs8111874) was conditioned to the inactivating mutation and *viceversa*. The magnitudes of the effects of the conditional analyses were substantially attenuated, suggesting the presence of one single association signal. Children with the A allele at rs601338 (W154X), which results in a truncated FUT2 protein and the non-secretor phenotype, had lower risk of diarrhoeal disease during the first years of life. In agreement with our results from a population-based design, the non-secretor phenotype has been associated with protection against *Rotavirus* (19), *Norovirus* (16-18) and *Helicobacter pylori* (39, 40) in small settings of very well characterized hospitalized subjects. It is known that *Norovirus* (41) and *Rotavirus* (42-44) bind to the antigen associated

with the FUT2 secretor phenotype to enter the cells. In contrast, the FUT2 non-secretor phenotype has been associated with higher risk of urinary tract infections (45), acute pyelonephritis (46), oral or vaginal *Candida* infections (12, 47, 48), *Haemophilus influenza* (49), *Neisseria meningitides* and *Streptococcus pneumoniae* infections (50). In addition, non-secretor individuals are at a higher risk of developing autoimmune diseases such as inflammatory bowel disease (22, 51), psoriasis (52, 53), and Behcet's disease (54) and they also have higher vitamin B12 plasma levels (55, 56). Recently, the non-secretor phenotype has been associated with gut microbiota at both the compositional and functional level (57). Non-secretors have lower species richness than the secretors (58), and the secretion status is a modifying factor for gut microbiota composition in Chron's disease (59). Fucose from fucosylated proteins synthesized by FUT2 in response to the activation of the innate immunity can be used by microbes as an energy source, and this has been shown to reduce bacterial infection and downregulate the expression of virulence genes (60, 61). Fucosylation appears to be a protective mechanism to maintain host-microbial interactions during pathogen-induced stress, but on the other hand it facilitates viral entrance into the cells.

Although the non-secretor phenotype related mutation in *FUT2* (W154X, rs601338) is the most likely causal variant of childhood diarrhoeal disease at 19q13.33 locus, the functional role of other variants in the region cannot be completely ruled out based on statistical analysis. Another variant, located in an intron of the *ARHGAP24* gene (rs1481779), reached genome-wide significance in DD1Y, but it could not be replicated. *ARHGAP24* codes a Rho GTPase-activating protein involved in cell polarity, cell morphology and cytoskeletal organization; and SNPs in it have been reported to be associated with blood pressure regulation (62).

In order to gain insight into potential molecular mechanisms underlying diarrhoeal disease in young children, we performed enrichment studies. The analysis using two different programs identified several gene-sets at 5% FDR. The "KEGG_GLYCOSPHINGOLIPID_BIOSYNTHESIS-GLOBO_SERIES" gene-set contains genes related to the production of histo-blood group antigens (HBGAs) from globo-series glycolipids and was at least nominally associated with several definitions of diarrhoea (D1Y and

DD2Y). Apart from *FUT2*, already discussed above, *A4GALT* (*Alpha 1,4-Galactosyltransferase*) in the gene-set also showed some evidence of association with D2Y. A similar gene-set, "KEGG_GLYCOSPHINGOLIPID_BIOSYNTHESIS-GANGLIO_SERIES", was also identified.

A second gene-set consisting of a SIGMAR1 protein interaction network ("ENSG00000147955 - SIGMAR1 PPI subnetwork") was detected for DD1Y. *SIGMAR1* is an endoplasmic reticulum-resident two-transmembrane chaperone that regulates voltage-gated ion channels including calcium, sodium, and potassium channels (63). We also identified other pathways related to hydro-electrolytic balance such as the "KEGG_RENIN_ANGIOTENSIN_SYSTEM", the "GO:0005227 – calcium activated cation channel activity", the "GO:0015299 - solute:hydrogen antiporter activity", and the "ENSG00000187446 - ENSG00000187446 PPI subnetwork", related to cell pH regulation by controlling plasma membrane-type Na(+)/H(+) exchange activity (64). Deregulation of transport of ions is a central mechanism in the pathophysiology of enteric pathogens (8, 9). Host genetic variants might compensate or accelerate watery stools after enteric infection.

Other gene-sets related to ectoderm and endoderm development were observed for DD1Y: "MP:0001675 - abnormal ectoderm development" and "GO:0007492 - endoderm development". The endoderm is the inner most germ layer that develops into the gastrointestinal tract, the lungs and associated tissues. Finally, the "ENSG00000140612 – SEC11A PPI subnetwork" was also identified. SEC11A is a component of the microsomal signal peptidase complex which removes signal peptides from nascent proteins as they are translocated into the lumen of the endoplasmic reticulum.

As expected, the gastrointestinal tract and the immune system were detected as relevant organs for diarrhoeal disease. The gastrointestinal tract was identified with the general definition of diarrhoea around one year of age (D1Y); while doctor diagnosis of diarrhoeal disease (DD1Y), probably more related to severe cases, highlighted the immune system. In addition, the neuro-secretory system was found to be associated with diarrhoea around two years of age (D2Y). The central nervous system communicates with the intestine through the hypothalamic-pituitary-

adrenal axis. The autonomic enteric nervous system regulates gastrointestinal motility, secretion, barrier function, and inflammatory response at the mucosa (65). Indeed, gene-sets related to neurotransmission were also identified: "Panther - Nicotinic_acetylcholine_receptor_signaling_pathway", "GO:0090278 - negative regulation of peptide hormone secretion" or a protein network regulating neurotransmitter release "ENSG00000147955 - SIGMAR1 PPI subnetwork". Several organs, such as the urogenital and musculoskeletal systems, not a priori of relevance for diarrhoeal disease, were identified for DD2Y.

Finally, given the known opposite effect of the *FUT2* locus in infectious diseases compared to autoimmune diseases, we decided to compare our GWAS results with the results reported in the largest GWAS of IBD, for which an infectious component has been suspected (22). Besides the *FUT2* locus, where we confirmed the opposite direction of the association, the same allele in a SNP at *CARD11* locus, was associated with higher risk of IBD and higher risk of D1Y. *CARD11* (*Caspase Recruitment Domain Family, Member 11*) is involved in the co-stimulatory signal essential for T-cell receptor (TCR)-mediated T-cell activation. We also compared our results with several GWAS of viral infections and response to vaccination (23-36). The *TTC7B* (*Tetratricopeptide Repeat Domain 7B*) locus, associated with higher anti *Cytomegalovirus* IgG titer (31), a marker of either new infection or frequent viral reactivation, was associated with higher risk for D2Y. Other studies have reported the participation of *TTC7B* in *Chikungunya* virus infection (66), and a paralog of *TTC7B*, *TTC7A*, has been implicated in a rare hereditary disease characterized by intestinal obstructions and profound immune defects (67).

The main limitations of the study are the low specificity of the phenotype definition, and the lack of underlying etiological information for the defined episodes. Information on diarrhoeal disease was retrieved using standard epidemiological tools for population-based birth cohort studies: questionnaires (ALSPAC, Generation R, GINIplus and LISAplus, INMA_VAL and INMA_SAB and MoBa), medical interviews (COSPAC) and medical records (CHOP). Apart from medical record definitions, the most specific assessment of infectious gastroenteritis was in Generation R and in MoBa studies, where the questionnaires included the following

statement: “infections of the stomach/intestine or gastric flu”. In fact, Generation R showed a low number of diarrhoeal disease cases and a high effect of the variant at *FUT2* locus, suggesting less misclassification problems. We acknowledge that misclassification and heterogeneity (i.e. infection by different underlying etiologies, bacterial species or viral strains or seasonality) may have decreased the statistical power in the discovery phase and might explain the lack of replication of suggestive hits besides the *FUT2* locus. Furthermore, cohorts with different study designs participated in the analyses: mainly population-based birth cohorts, but also a random collection of pediatric children from a hospital centre (CHOP, replication), and a population-based birth cohort of infants born to mothers with a history of asthma (COPSAC2000, discovery). The potential relationship between asthma and infection diseases might have affected the associations. To our knowledge there are no population-based studies of this sample size with molecular diagnosis of infectious gastroenteritis. Although we analyzed all available samples from the EArly Genetics and Life course Epidemiology (EAGLE) consortium following a flexible inclusion criteria the sample size is still intermediate for genome-wide scale studies. Therefore, increasing the specificity of the outcome and the sample size in population-based designs might reveal novel loci for childhood diarrhoeal disease and confirm the role of host genetics in infectious diseases during the first years of life. Finally, although diarrhoeal disease in young children is mainly caused by *Rotavirus*, which is a wide-spread virus producing seasonal break-outs in all the countries included in this study, we cannot completely exclude the possibility that the identified variants reflect different levels of exposure to the virus, rather than a higher susceptibility to infection.

In summary, the genome-wide association meta-analysis of diarrhoeal disease in children suggested the implication of the *FUT2* locus at the population level, and has pointed to W154X (rs601338) as the most likely causal variant. The histo-blood group antigen (HBGA) production and the regulation of ion transport were plausible underlying biological mechanisms accounting for part of the host genetic variability of diarrhoeal disease.

Material and Methods

Sample and diarrhoeal disease definition

This study was performed within the framework of the Early Genetics and Life course Epidemiology (EAGLE) Consortium (http://www.wikigenes.org/e/art/e/348.html) and it was divided in two phases: discovery and replication. The following population-based birth cohorts or studies settled up in developed countries participated in the discovery and/or replication phases (Table 1): the Avon Longitudinal Study of Parents And Children (ALSPAC), the Children's Hospital of Philadelphia (CHOP) study, the Copenhagen prospective studies on asthma in childhood (COPSAC2000 and COPSAC2010), the Generation R study, the Influence of Life-style related factors on the development of the Immune System and Allergies in East and West Germany plus the influence of traffic emissions and genetics (LISAplus) study, the Study on the influence of Nutrition Intervention plus Air pollution and Genetic on Allergy development (GINIplus), the Infancia y Medio Ambiente (INMA) project, and the national Norwegian Mother and Child Cohort Study (MoBa). Cohorts were allocated in the discovery or in the replication set with the aim of making both sets comparable. Diarrhoeal disease was defined in two different time windows: around age 1 year (from 0 months to 18 months) and around age 2 years (from 12 to 30 months). At each time point diarrhoea (D1Y and D2Y) and doctor diagnosis of diarrhoea (DD1Y and DD2Y) were studied. Data were collected from parental questionnaires, doctor interviews or medical records. A detailed description of diarrhoeal disease definitions in each cohort at each time point is described in the Supplementary Material – Annex A. A comparison of the year of initiation of the cohort vs. the year of introduction of *Rotavirus* vaccination in each country can be seen in Supplementary Material – Annex B. The vaccine was introduced in the USA at the time when children from CHOP study were enrolled, and thus we excluded vaccinated children, identified through medical records, from the analysis.

Each cohort obtained the ethical approval from the respective Ethical Committees and a written consent including permission to perform GWAS analyses was obtained from participating parents.

Genotyping, quality control and imputation

Genotypes within each cohort were collected using high-density SNP arrays on Illumina (ALSPAC, CHOP, COPSAC2000, COPSAC2010, Generation R, INMA_SAB, INMA_VAL, MoBa) and Affymetrix (GINIplus and LISAplus) platforms. Each cohort imputed up to ~30 M variants using MACH (68) or IMPUTE2 (69) considering the 1000 Genomes Project CEU release March 2012 (http://mathgen.stats.ox.ac.uk/impute/ALL_1000G_phase1integrated_v3_impute.tgz) as the reference population panel. More details on the process followed by each cohort are described in Supplementary Material – Annex A.

Analysis, meta-analysis and replication

The study included only at term Caucasian singletons and children with congenital anomalies were excluded. The association between diarrhoeal disease and the variant dose was assessed in each study using logistic regression analyses assuming an additive genetic model. Sex and principal components accounting for genetic sub-stratification were added as covariates. Chromosome X was analysed under the same statistical model but without sex adjustment. More details on the programs used by each study to perform the analysis can be found in Supplementary Material – Annex A.

Only variants with a Minor Allele Frequency (MAF) ≥0.01 and with a quality of imputation ≥ 0.4 (INFO) or ≥ 0.3 (R2) were considered. Due to the limited sample size of some cohorts, an additional filtering based on expected minor allele counts (EMAC) was performed. This parameter is related to both the sample size and the quality of imputation (2*N*MAF*quality of imputation). Variants that did not reach an EMAC≥50 were excluded. After quality control, from 5.4 to 8.7 million variants were kept for the analysis in each cohort. The genomic inflation factor lambda (λ) was calculated for each study. A summary of the quality control procedure is shown in Supplementary Table S14. Marker names and alleles were harmonized among studies.

A fixed effect meta-analysis weighted by inverse variance was conducted using GWAMA (70). The genomic control approach was applied to the meta-analysis results. Only variants with data for at least 5,000 samples were considered. Genome-wide level of significance was defined at p-

value≤5E-08 and suggestive associations at p-value ≤1E-05. Quantile-quantile (Q-Q) plots, calculation of lambda (λ) and Manhattan plots were performed in R software environment version 3.2.3 (71). Regional association plots were performed with Locus Zoom (72).

In total seventy-two variants were followed for replication in an independent dataset. They were selected among the four outcomes based on the statistical significance (p-value<1E-05). Association p-values from the replication phase were corrected for multiple testing using Bonferroni correction (for each trait independently). Exclusion of CHOP cohort, as it comprises potentially vaccinated children, from the replication phase did not revealed any new replicated genetic variant, and the association of replicated variants was maintained (data not shown).

Conditional analysis

We conditioned the analysis of the leading SNP identified at 19q13.33 (rs8111874) to a stop mutation situated 37.7 kb apart (rs601338, W154X) using the GCTA program (73). As reference we used the INMA 1000 Genomes imputation (restricted to variants with MAF>0.01 and imputation quality (INFO) >0.8).

A similar analysis was performed to search for secondary signals in +/- 500 kb surrounding the stop mutation (rs601338) or the top SNP at 19q13.33 (rs8111874). In this case, the significance threshold was calculated by Bonferroni correction, where the number of independent tests was the effective number of variants in this region estimated using Nyholt's procedure and the 1000G reference data for Europeans (74).

Annotation and enrichment analysis

Genetic variants annotation (nearest gene, eQTLs, protein binding, and regulatory features) was done with the HaploReg v4.1 program (75). In addition, a second search for eQTLs and for expression levels in tissues was performed with Genotype-Tissue Expression (GTEx) data (http://www.gtexportal.org/). GeneCard (http://www.genecards.org) and the USC genome browser (http://genome.ucsc.edu/) were used to search for gene functions and GWAS signals, respectively.

Two different tools were used to explore gene-set enrichment analysis. We performed an analysis with MAGENTA software that uses genome-wide summary statistics (76). Briefly,

first, the program links variants to genes considering flanking regions and then computes the gene-set enrichment comparing the variants with the lowest p-values (95th percentile) versus the rest. Gene-set databases evaluated in MAGENTA were Panther, KEGG (Kyoto Encyclopaedia of Genes and Genomes) and Ingenuity. In addition, we used DEPICT to identify enriched genes-sets as well as tissues/cell types where genes from associated loci are highly expressed (77). In this case, only variants associated with diarrhoea with a p-value≤ 1E-05 and a sample size >5,000 were considered. Both programs estimate adjusted p-values using the FDR method. In the case of MAGENTA, FDR was applied within each database.

Overlap with known variants and genes for related diseases

We investigated the association between diarrhoea and known variants for inflammatory bowel disease (IBD) (22). We also looked at variants associated with viral infection susceptibility, disease progression and response to vaccination against different viruses (23-26). Variants were selected from GWAS retrieved from the GWAS catalog (http://www.ebi.ac.uk/gwas/, date: April 2015). All variants reported in European populations, regardless of their statistical significance, and their replication status, were evaluated. Finally, we evaluated 86 genes retrieved from OMIM (Online Mendelian Inheritance in Man; date: July 2016) with the entry "diarrhea". To test the association of these "candidate" genes, we performed a gene-based analysis using VEGAS2 (Versatile Gene based Association Study, http://vegas2.qimrberghofer.edu.au/) (78), considering linkage disequilibrium patterns described in European populations and a flanking region of +/- 50 kb around the gene. The corrected statistical significance level was calculated using Bonferroni correction accounting for the number of variants/genes within each analysis (inflammatory bowel disease, viral infection, or hereditary diarrhoeal diseases) as independent tests.

Genome-wide summarized results of the discovery phase can be found at INMA's web page (Infancia and Medio Ambiente project, http://www.proyectoinma.org/) and at the EAGLE consortium web page (http://www.wikigenes.org/e/art/e/348.html)."

Acknowledgements

ALSPAC

The UK Medical Research Council and the Wellcome Trust [102215/2/13/2] and the University of Bristol provide core support for ALSPAC. GDS's and NJT's works are supported by the UK Medical Research Council Integrative Epidemiology Unit at the University of Bristol [MC_UU_12013_1 and MC_UU_12013/3, respectively].

We are extremely grateful to all the families who took part in this study, the midwives for their help in recruiting them, and the whole ALSPAC team, which includes interviewers, computer and laboratory technicians, clerical workers, research scientists, volunteers, managers, receptionists and nurses. GWAS data was generated by Sample Logistics and Genotyping Facilities at the Wellcome Trust Sanger Institute and LabCorp (Laboratory Corportation of America) using support from 23andMe. This publication is the work of the authors and Carolina Bonilla will serve as guarantor for the contents of this paper.

CHOP

This research was financially supported by the Daniel B. Burke Chair for Diabetes Research, an Institute Development Award from the Children's Hospital of Philadelphia, a Research Development Award from the Cotswold Foundation and the National Institutes of Health [R01 HD056465].

We thank the network of primary care clinicians, their patients and families for their contribution to this project and clinical research facilitated through the Pediatric Research Consortium (PeRC) at The Children's Hospital of Philadelphia. Rosetta Chiavacci, Elvira Dabaghyan, Hope Thomas, Kisha Harden, Andrew Hill, Kenya Fain, Crystal Johnson-Honesty, Cynthia Drummond, Shanell Harrison and Sarah Wildrick, Cecilia Kim, Edward Frackelton, George Otieno, Kelly Thomas, Cuiping Hou, Kelly Thomas and Maria L. Garris provided expert assistance with genotyping or data collection and management. We would also like to thank Smari Kristinsson, Larus Arni Hermannsson and Asbjörn Krisbjörnssonof Raförninnehf for their extensive software design and contribution.

COPSAC

COPSAC is funded by private and public research funds all listed on www.copsac.com. The Lundbeck Foundation [R16-A1694]; The Danish Ministry of Health [903516]; Danish Council for Strategic Research [0603-00280B]; and The Capital Region Research Foundation have provided core support for COPSAC. Genotyping was supported by The Danish Council for Independent Research [10-082884 and 271-08-0815]. No pharmaceutical company was involved in the study. The funding agencies did not have any role in design and conduct of the study; collection, management, and interpretation of the data; or preparation, review, or approval of the manuscript.

We express our deepest gratitude to the children and families of the $COPSAC_{2000}$ and $COPSAC_{2010}$ study cohorts for all their support and commitment. We acknowledge and appreciate the unique efforts of the COPSAC research team.

Generation R

The general design of Generation R Study is made possible by financial support from the Erasmus Medical Center, Rotterdam, the Erasmus University Rotterdam, the Netherlands Organization for Health Research and Development (ZonMw), the Netherlands Organisation for Scientific Research (NWO), the Ministry of Health, Welfare and Sport. VWVJ received grants from the Netherlands Organization for Health Research and Development [ZonMw 907.00303, ZonMw 916.10159, VIDI 016.136.361], and from the European Research Council [ERC-2014-CoG-648916].

The Generation R Study is conducted by the Erasmus Medical Center in close collaboration with the Faculty of Social Sciences of the Erasmus University, the Municipal Health Service Rotterdam area, the Rotterdam Homecare Foundation and the Stichting Trombosedienst & Artsenlaboratorium Rijnmond (STAR). We gratefully acknowledge the contribution of general practitioners, hospitals, midwives and pharmacies in Rotterdam. The generation and management of GWAS genotype data for the Generation R Study was done at the Genetic Laboratory of the Department of Internal Medicine, Erasmus MC, The Netherlands. We thank Pascal Arp, Mila Jhamai, Marijn Verkerk, Lizbeth Herrera and Marjolein Peters for their help in creating, managing and QC of the GWAS database. Also Karol Estrada and Carolina Medina-

Gomez for their support in creation and analysis of imputed data. We would like to thank Karol Estrada, Fernando Rivadeneira, and Anis Abuseiris (Erasmus MC Rotterdam, The Netherlands), for their help in creating GRIMP, and we thank Big GRID for access to their grid computing resources.

GINIplus

The GINIplus study was mainly supported for the first 3 years of the Federal Ministry for Education, Science, Research and Technology (interventional arm) and Helmholtz Zentrum Munich (former GSF) (observational arm). The 4 year, 6 year, 10 year and 15 year follow-up examinations of the GINIplus study were covered from the respective budgets of the 5 study centres (Helmholtz Zentrum Munich (former GSF), Research Institute at Marien-Hospital Wesel, LMU Munich, TU Munich and from 6 years onwards also from IUF - Leibniz Research-Institute for Environmental Medicine at the University of Düsseldorf) and a grant from the Federal Ministry for Environment [IUF Düsseldorf, FKZ 20462296]. Further, the 15 year follow-up examination was supported by the European Commission [FP7-HEALTH-2010-261357], as well by the companies Mead Johnson and Nestlé.

The authors thank all the families for their participation in the GINIplus study. Furthermore, we thank all members of the GINIplus Study Group for their excellent work. The GINIplus Study group consists of the following: Institute of Epidemiology I, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg (Heinrich J, Brüske I, Schulz H, Flexeder C, Zeller C, Standl M, Schnappinger M, Sußmann M, Thiering E, Tiesler C); Department of Pediatrics, Marien-Hospital, Wesel (Berdel D, von Berg A); Ludwig-Maximilians-University of Munich, Dr von Hauner Children's Hospital (Koletzko S); Child and Adolescent Medicine, University Hospital rechts der Isar of the Technical University Munich (Bauer CP, Hoffmann U); IUF- Environmental Health Research Institute, Düsseldorf (Schikowski T, Link E, Klümper C).

LISAplus

The LISAplus study was mainly supported by grants from the Federal Ministry for Education, Science, Research and Technology and in addition from Helmholtz Zentrum Munich (former

GSF), Helmholtz Centre for Environmental Research - UFZ, Leipzig, Research Institute at Marien-Hospital Wesel, Pediatric Practice, Bad Honnef for the first 2 years. The 4 year, 6 year, 10 year and 15 year follow-up examinations of the LISAplus study were covered from the respective budgets of the involved partners (Helmholtz Zentrum Munich (former GSF), Helmholtz Centre for Environmental Research - UFZ, Leipzig, Research Institute at Marien-Hospital Wesel, Pediatric Practice, Bad Honnef, IUF – Leibniz-Research Institute for Environmental Medicine at the University of Düsseldorf) and in addition by a grant from the Federal Ministry for Environment [IUF Düsseldorf, FKZ 20462296]. Further, the 15 year follow-up examination was supported by the European Commission [FP7-HEALTH-2010-261357].

The authors thank all the families for their participation in the LISAplus study. Furthermore, we thank all members of the LISAplus Study Group for their excellent work. The LISAplus Study group consists of the following: Helmholtz Zentrum München, German Research Center for Environmental Health, Institute of Epidemiology I, Munich (Heinrich J, Schnappinger M, Brüske I, Sußmann M, Lohr W, Schulz H, Zeller C, Standl M); Department of Pediatrics, Municipal Hospital "St. Georg", Leipzig (Borte M, Gnodtke E); Marien Hospital Wesel, Department of Pediatrics, Wesel (von Berg A, Berdel D, Stiers G, Maas B); Pediatric Practice, Bad Honnef (Schaaf B); Helmholtz Centre of Environmental Research – UFZ, Department of Environmental Immunology/Core Facility Studies, Leipzig (Lehmann I, Bauer M, Röder S, Schilde M, Nowak M, Herberth G , Müller J, Hain A); Technical University Munich, Department of Pediatrics, Munich (Hoffmann U, Paschke M, Marra S); Clinical Research Group Molecular Dermatology, Department of Dermatology and Allergy, Technische Universität München (TUM), Munich (Ollert M).

INMA

This project was funded by grants from Instituto de Salud Carlos III: FIS-FEDER [CB06/02/0041, G03/176, PI041436, PI081151, PI041705, PI061756, PI091958, and PS09/00432, 03/1615, 04/1509, 04/1112, 04/1931, 05/1079, 05/1052, 06/1213, 07/0314, 09/02647, 11/01007, 11/02591, 11/02038, 13/1944, 13/2032] and Miguel Servet-FEDER

[CP11/0178, MS15/0025, and CP11/00269]; Spanish Ministry of Science and Innovation [SAF2008-00357]; European Commission [FP7-HEALTH-2010-261357, FP7-ENV-2011-282957]; Fundació La Marató de TV3; Generalitat de Catalunya [CIRIT 1999SGR 00241] and Conselleria de Sanitat Generalitat Valenciana. The work at XE laboratory was funded by "Retos de la Sociedad 2013: Europa Redes y Gestores" Programme from the Spanish Ministry of Economy and Competitiveness [SAF2013-49108-R], the Generalitat de Catalunya [AGAUR 2014 SGR-1138], the European Commission [FP7/2007-2013-262055]. XE lab acknowledges support of the MINECO, 'Centro de Excelencia Severo Ochoa 2013-2017' [SEV-2012-0208].

The authors would like to thank all the participants for their generous collaboration. The authors are grateful to Silvia Fochs, Anna Sànchez, Maribel López, NuriaPey, Muriel Ferrer, Amparo Quiles, Sandra Pérez, Gemma León, Elena Romero, Maria Andreu, Nati Galiana, Maria Dolores Climent, Amparo Cases and Cristina Capo for their assistance in contacting the families and administering the questionnaires. A full roster of the INMA project investigators can be found at http://www.proyectoinma.org/presentacion-inma/listado-investigadores/en_listado-investigadores.html. Some of the DNA extractions and genotyping were performed at the Spanish National Genotyping Centre (CEGEN-Barcelona).

MoBa (Mother and Child Cohort of NIPH)

This work was supported by grants from the Norwegian Research Council [FUGE 183220/S10, FRIMEDKLI-05 ES236011], Swedish Medical Society [SLS 2008-21198], Jane and Dan Olsson Foundations and Swedish government grants to researchers in the public health service [ALFGBG-2863, ALFGBG-11522, ALFGBG-426411], Swedish Medical Research Council [2015-02559] and the European Commission [HEALTH-F4-2007-201413]. The Norwegian Mother and Child Cohort Study was also supported by the Norwegian Ministry of Health and the Ministry of Education and Research, NIH/NIEHS [N01-ES-75558], NIH/NINDS [UO1 NS 047537-01 and UO1 NS 047537-06A1], and the Norwegian Research Council/FUGE [151918/S10; FRI-MEDBIO 249779].

We are grateful to all the participating families in Norway who take part in this ongoing cohort study. Researchers interested in using data or biological material from MoBa must obtain approval from the Scientific Management Committee and from the Regional Committee for Medical and Health Research Ethics. Researchers will be required to follow the terms of an Assistance Agreement containing a number of clauses designed to ensure protection of privacy and compliance with relevant laws. For further information, contact the principal investigator of MoBa, Per Magnus (per.magnus@fhi.no).

Conflict of Interest Statement

All the researchers from the study declare no conflict of interests.

References

1 Walker, C.L., Rudan, I., Liu, L., Nair, H., Theodoratou, E., Bhutta, Z.A., O'Brien, K.L., Campbell, H. and Black, R.E. (2013) Global burden of childhood pneumonia and diarrhoea. *Lancet*, **381**, 1405-1416.

2 Elliott, E.J. (2007) Acute gastroenteritis in children. *Bmj*, **334**, 35-40.

3 Kotloff, K.L., Nataro, J.P., Blackwelder, W.C., Nasrin, D., Farag, T.H., Panchalingam, S., Wu, Y., Sow, S.O., Sur, D., Breiman, R.F. *et al.* (2013) Burden and aetiology of diarrhoeal disease in infants and young children in developing countries (the Global Enteric Multicenter Study, GEMS): a prospective, case-control study. *Lancet*, **382**, 209-222.

4 Wiegering, V., Kaiser, J., Tappe, D., Weissbrich, B., Morbach, H. and Girschick, H.J. (2011) Gastroenteritis in childhood: a retrospective study of 650 hospitalized pediatric patients. *Int. J. Infect. Dis.*, **15**, e401-407.

5 Ray, P.G., Kelkar, S.D., Walimbe, A.M., Biniwale, V. and Mehendale, S. (2007) Rotavirus immunoglobulin levels among Indian mothers of two socio-economic groups and occurrence of rotavirus infections among their infants up to six months. *J. Med. Virol.*, **79**, 341-349.

6 Prameela, K.K. and Vijaya, L.R. (2012) The importance of breastfeeding in rotaviral diarrhoeas. *Malays. J. Nutr.*, **18**, 103-111.

7 Velazquez, F.R., Matson, D.O., Guerrero, M.L., Shults, J., Calva, J.J., Morrow, A.L., Glass, R.I., Pickering, L.K. and Ruiz-Palacios, G.M. (2000) Serum antibody as a marker of protection against natural rotavirus infection and disease. *J. Infect. Dis.*, **182**, 1602-1609.

8 Ramig, R.F. (2004) Pathogenesis of intestinal and systemic rotavirus infection. *J. Virol.*, **78**, 10213-10220.

9 Greenberg, H.B. and Estes, M.K. (2009) Rotaviruses: from pathogenesis to vaccination. *Gastroenterology*, **136**, 1939-1951.

10 Thapar, N. and Sanderson, I.R. (2004) Diarrhoea in children: an interface between developing and developed countries. *Lancet*, **363**, 641-653.

11 Sorensen, T.I., Nielsen, G.G., Andersen, P.K. and Teasdale, T.W. (1988) Genetic and environmental influences on premature death in adult adoptees. *N. Engl. J. Med.*, **318**, 727-732.

12 Ben-Aryeh, H., Blumfield, E., Szargel, R., Laufer, D. and Berdicevsky, I. (1995) Oral Candida carriage and blood group antigen secretor status. *Mycoses*, **38**, 355-358.

13 Burgner, D., Jamieson, S.E. and Blackwell, J.M. (2006) Genetic susceptibility to infectious diseases: big is beautiful, but will bigger be even better? *Lancet Infect. Dis.*, **6**, 653-663.

14 Hill, A.V. (2006) Aspects of genetic susceptibility to human infectious diseases. *Annu. Rev. Genet.*, **40**, 469-486.

15 Pinkerton, R.C., Oria, R.B., Kent, J.W., Jr., Kohli, A., Abreu, C., Bushen, O., Lima, A.A., Blangero, J., Williams-Blangero, S. and Guerrant, R.L. (2011) Evidence for genetic susceptibility to developing early childhood diarrhea among shantytown children living in northeastern Brazil. *Am. J. Trop. Med. Hyg.*, **85**, 893-896.

16 Thorven, M., Grahn, A., Hedlund, K.O., Johansson, H., Wahlfrid, C., Larson, G. and Svensson, L. (2005) A homozygous nonsense mutation (428G-->A) in the human secretor (FUT2) gene provides resistance to symptomatic norovirus (GGII) infections. *J. Virol.*, **79**, 15351-15355.

17 Carlsson, B., Kindberg, E., Buesa, J., Rydell, G.E., Lidon, M.F., Montava, R., Abu Mallouh, R., Grahn, A., Rodriguez-Diaz, J., Bellido, J. *et al.* (2009) The G428A nonsense mutation in FUT2 provides strong but not absolute protection against symptomatic GII.4 Norovirus infection. *PLoS One*, **4**, e5593.

18 Kindberg, E., Akerlind, B., Johnsen, C., Knudsen, J.D., Heltberg, O., Larson, G., Bottiger, B. and Svensson, L. (2007) Host genetic resistance to symptomatic norovirus (GGII.4) infections in Denmark. *J. Clin. Microbiol.*, **45**, 2720-2722.

19 Imbert-Marcille, B.M., Barbe, L., Dupe, M., Le Moullac-Vaidye, B., Besse, B., Peltier, C., Ruvoen-Clouet, N. and Le Pendu, J. (2014) A FUT2 gene common polymorphism determines resistance to rotavirus A of the P[8] genotype. *J. Infect. Dis.*, **209**, 1227-1230.

20 Petri, W.A., Jr., Miller, M., Binder, H.J., Levine, M.M., Dillingham, R. and Guerrant, R.L. (2008) Enteric infections, diarrhea, and their impact on function and development. *J. Clin. Invest.*, **118**, 1277-1290.

21 Flores, J. and Okhuysen, P.C. (2009) Genetics of susceptibility to infection with enteric pathogens. *Curr. Opin. Infect. Dis.*, **22**, 471-476.

22 Jostins, L., Ripke, S., Weersma, R.K., Duerr, R.H., McGovern, D.P., Hui, K.Y., Lee, J.C., Schumm, L.P., Sharma, Y., Anderson, C.A. *et al.* (2012) Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature*, **491**, 119-124.

23 Limou, S., Le Clerc, S., Coulonges, C., Carpentier, W., Dina, C., Delaneau, O., Labib, T., Taing, L., Sladek, R., Deveau, C. *et al.* (2009) Genomewide association study of an AIDS-nonprogression cohort emphasizes the role played by HLA genes (ANRS Genomewide Association Study 02). *J. Infect. Dis.*, **199**, 419-426.

24 Le Clerc, S., Limou, S., Coulonges, C., Carpentier, W., Dina, C., Taing, L., Delaneau, O., Labib, T., Sladek, R., Deveau, C. *et al.* (2009) Genomewide association study of a rapid progression cohort identifies new susceptibility alleles for AIDS (ANRS Genomewide Association Study 03). *J. Infect. Dis.*, **200**, 1194-1201.

25 Fellay, J., Ge, D., Shianna, K.V., Colombo, S., Ledergerber, B., Cirulli, E.T., Urban, T.J., Zhang, K., Gumbs, C.E., Smith, J.P. *et al.* (2009) Common genetic variation and the control of HIV-1 in humans. *PLoS Genet.*, **5**, e1000791.

26 Herbeck, J.T., Gottlieb, G.S., Winkler, C.A., Nelson, G.W., An, P., Maust, B.S., Wong, K.G., Troyer, J.L., Goedert, J.J., Kessing, B.D. *et al.* (2010) Multistage genomewide association study identifies a locus at 1q41 associated with rate of HIV-1 disease progression to clinical AIDS. *J. Infect. Dis.*, **201**, 618-626.

27 Pereyra, F., Jia, X., McLaren, P.J., Telenti, A., de Bakker, P.I., Walker, B.D., Ripke, S., Brumme, C.J., Pulit, S.L., Carrington, M. *et al.* (2010) The major genetic determinants of HIV-1 control affect HLA class I peptide presentation. *Science*, **330**, 1551-1557.

28 Bol, S.M., Moerland, P.D., Limou, S., van Remmerden, Y., Coulonges, C., van Manen, D., Herbeck, J.T., Fellay, J., Sieberer, M., Sietzema, J.G. *et al.* (2011) Genome-wide association study identifies single nucleotide polymorphism in DYRK1A associated with replication of HIV-1 in monocyte-derived macrophages. *PLoS One*, **6**, e17190.

29 Troyer, J.L., Nelson, G.W., Lautenberger, J.A., Chinn, L., McIntosh, C., Johnson, R.C., Sezgin, E., Kessing, B., Malasky, M., Hendrickson, S.L. *et al.* (2011) Genome-wide association study implicates PARD3B-based AIDS restriction. *J. Infect. Dis.*, **203**, 1491-1502.

30 Chen, D., McKay, J.D., Clifford, G., Gaborieau, V., Chabrier, A., Waterboer, T., Zaridze, D., Lissowska, J., Rudnai, P., Fabianova, E. *et al.* (2011) Genome-wide association study of HPV seropositivity. *Hum. Mol. Genet.*, **20**, 4714-4723.

31 Kuparinen, T., Seppala, I., Jylhava, J., Marttila, S., Aittoniemi, J., Kettunen, J., Viikari, J., Kahonen, M., Raitakari, O., Lehtimaki, T. *et al.* (2012) Genome-wide association study does not reveal major genetic determinants for anti-cytomegalovirus antibody response. *Genes Immun.*, **13**, 184-190.

32 Liu, L., Li, J., Yao, J., Yu, J., Zhang, J., Ning, Q., Wen, Z., Yang, D., He, Y., Kong, X. *et al.* (2011) A genome-wide association study with DNA pooling identifies the variant rs11866328 in the GRIN2A gene that affects disease progression of chronic HBV infection. *Viral Immunol.*, **24**, 397-402.

33 Kennedy, R.B., Ovsyannikova, I.G., Pankratz, V.S., Haralambieva, I.H., Vierkant, R.A. and Poland, G.A. (2012) Genome-wide analysis of polymorphisms associated with cytokine responses in smallpox vaccine recipients. *Hum. Genet.*, **131**, 1403-1421.

34 Duggal, P., Thio, C.L., Wojcik, G.L., Goedert, J.J., Mangia, A., Latanich, R., Kim, A.Y., Lauer, G.M., Chung, R.T., Peters, M.G. *et al.* (2013) Genome-wide association study of

spontaneous resolution of hepatitis C virus infection: data from multiple cohorts. *Ann. Intern. Med.*, **158**, 235-245.

35 McLaren, P.J., Coulonges, C., Ripke, S., van den Berg, L., Buchbinder, S., Carrington, M., Cossarizza, A., Dalmau, J., Deeks, S.G., Delaneau, O. *et al.* (2013) Association study of common genetic variants and HIV-1 acquisition in 6,300 infected cases and 7,200 controls. *PLoS Pathog.*, **9**, e1003515.

36 Kennedy, R.B., Ovsyannikova, I.G., Haralambieva, I.H., Lambert, N.D., Pankratz, V.S. and Poland, G.A. (2014) Genome-wide SNP associations with rubella-specific cytokine responses in measles-mumps-rubella vaccine recipients. *Immunogenetics*, **66**, 493-499.

37 Marionneau, S., Cailleau-Thomas, A., Rocher, J., Le Moullac-Vaidye, B., Ruvoen, N., Clement, M. and Le Pendu, J. (2001) ABH and Lewis histo-blood group antigens, a model for the meaning of oligosaccharide diversity in the face of a changing world. *Biochimie*, **83**, 565-573.

38 Ferrer-Admetlla, A., Sikora, M., Laayouni, H., Esteve, A., Roubinet, F., Blancher, A., Calafell, F., Bertranpetit, J. and Casals, F. (2009) A natural history of FUT2 polymorphism in humans. *Mol. Biol. Evol.*, **26**, 1993-2003.

39 Ikehara, Y., Nishihara, S., Yasutomi, H., Kitamura, T., Matsuo, K., Shimizu, N., Inada, K., Kodera, Y., Yamamura, Y., Narimatsu, H. *et al.* (2001) Polymorphisms of two fucosyltransferase genes (Lewis and Secretor genes) involving type I Lewis antigens are associated with the presence of anti-Helicobacter pylori IgG antibody. *Cancer Epidemiol. Biomarkers Prev.*, **10**, 971-977.

40 Azevedo, M., Eriksson, S., Mendes, N., Serpa, J., Figueiredo, C., Resende, L.P., Ruvoen-Clouet, N., Haas, R., Boren, T., Le Pendu, J. *et al.* (2008) Infection by Helicobacter pylori expressing the BabA adhesin is influenced by the secretor phenotype. *J. Pathol.*, **215**, 308-316.

41 Ruvoen-Clouet, N., Belliot, G. and Le Pendu, J. (2013) Noroviruses and histo-blood groups: the impact of common host genetic polymorphisms on virus transmission and evolution. *Rev. Med. Virol.*, **23**, 355-366.

42 Hu, L., Crawford, S.E., Czako, R., Cortes-Penfield, N.W., Smith, D.F., Le Pendu, J., Estes, M.K. and Prasad, B.V. (2012) Cell attachment protein VP8* of a human rotavirus specifically interacts with A-type histo-blood group antigen. *Nature*, **485**, 256-259.

43 Huang, P., Xia, M., Tan, M., Zhong, W., Wei, C., Wang, L., Morrow, A. and Jiang, X. (2012) Spike protein VP8* of human rotavirus recognizes histo-blood group antigens in a type-specific manner. *J. Virol.*, **86**, 4833-4843.

44 Ramani, S., Cortes-Penfield, N.W., Hu, L., Crawford, S.E., Czako, R., Smith, D.F., Kang, G., Ramig, R.F., Le Pendu, J., Prasad, B.V. *et al.* (2013) The VP8* domain of neonatal rotavirus strain G10P[11] binds to type II precursor glycans. *J. Virol.*, **87**, 7255-7264.

45 Kinane, D.F., Blackwell, C.C., Brettle, R.P., Weir, D.M., Winstanley, F.P. and Elton, R.A. (1982) ABO blood group, secretor state, and susceptibility to recurrent urinary tract infection in women. *Br. Med. J. (Clin. Res. Ed.)*, **285**, 7-9.

46 Ishitoya, S., Yamamoto, S., Mitsumori, K., Ogawa, O. and Terai, A. (2002) Non-secretor status is associated with female acute uncomplicated pyelonephritis. *BJU Int*, **89**, 851-854.

47 Thom, S.M., Blackwell, C.C., MacCallum, C.J., Weir, D.M., Brettle, R.P., Kinane, D.F. and Wray, D. (1989) Non-secretion of blood group antigens and susceptibility to infection by Candida species. *FEMS Microbiol. Immunol.*, **1**, 401-405.

48 Chaim, W., Foxman, B. and Sobel, J.D. (1997) Association of recurrent vaginal candidiasis and secretory ABO and Lewis phenotype. *J. Infect. Dis.*, **176**, 828-830.

49 Blackwell, C.C., Jonsdottir, K., Hanson, M.F. and Weir, D.M. (1986) Non-secretion of ABO blood group antigens predisposing to infection by Haemophilus influenzae. *Lancet*, **2**, 687.

50 Blackwell, C.C., Jonsdottir, K., Hanson, M., Todd, W.T., Chaudhuri, A.K., Mathew, B., Brettle, R.P. and Weir, D.M. (1986) Non-secretion of ABO antigens predisposing to infection by Neisseria meningitidis and Streptococcus pneumoniae. *Lancet*, **2**, 284-285.

51 McGovern, D.P., Jones, M.R., Taylor, K.D., Marciante, K., Yan, X., Dubinsky, M., Ippoliti, A., Vasiliauskas, E., Berel, D., Derkowski, C. *et al.* (2010) Fucosyltransferase 2 (FUT2) non-secretor status is associated with Crohn's disease. *Hum. Mol. Genet.*, **19**, 3468-3476.

52 Ellinghaus, D., Ellinghaus, E., Nair, R.P., Stuart, P.E., Esko, T., Metspalu, A., Debrus, S., Raelson, J.V., Tejasvi, T., Belouchi, M. *et al.* (2012) Combined analysis of genome-wide association studies for Crohn disease and psoriasis identifies seven shared susceptibility loci. *Am. J. Hum. Genet.*, **90**, 636-647.

53 Tang, H., Jin, X., Li, Y., Jiang, H., Tang, X., Yang, X., Cheng, H., Qiu, Y., Chen, G., Mei, J. *et al.* (2014) A large-scale screen for coding variants predisposing to psoriasis. *Nat. Genet.*, **46**, 45-50.

54 Xavier, J.M., Shahram, F., Sousa, I., Davatchi, F., Matos, M., Abdollahi, B.S., Sobral, J., Nadji, A., Oliveira, M., Ghaderibarim, F. *et al.* (2013) FUT2: filling the gap between genes and environment in Behcet's disease? *Ann. Rheum. Dis.*, **74**, 618-624.

55 Grarup, N., Sulem, P., Sandholt, C.H., Thorleifsson, G., Ahluwalia, T.S., Steinthorsdottir, V., Bjarnason, H., Gudbjartsson, D.F., Magnusson, O.T., Sparso, T. *et al.* (2013) Genetic architecture of vitamin B12 and folate levels uncovered applying deeply sequenced large datasets. *PLoS Genet.*, **9**, e1003530.

56 Hazra, A., Kraft, P., Lazarus, R., Chen, C., Chanock, S.J., Jacques, P., Selhub, J. and Hunter, D.J. (2009) Genome-wide significant predictors of metabolites in the one-carbon metabolism pathway. *Hum. Mol. Genet.*, **18**, 4677-4687.

57 Tong, M., McHardy, I., Ruegger, P., Goudarzi, M., Kashyap, P.C., Haritunians, T., Li, X., Graeber, T.G., Schwager, E., Huttenhower, C. *et al.* (2014) Reprograming of gut microbiome energy metabolism by the FUT2 Crohn's disease risk polymorphism. *Isme J.*, **8**, 2193-2206.

58 Wacklin, P., Tuimala, J., Nikkila, J., Sebastian, T., Makivuokko, H., Alakulppi, N., Laine, P., Rajilic-Stojanovic, M., Paulin, L., de Vos, W.M. *et al.* (2014) Faecal microbiota composition in adults is associated with the FUT2 gene determining the secretor status. *PLoS One*, **9**, e94863.

59 Rausch, P., Rehman, A., Kunzel, S., Hasler, R., Ott, S.J., Schreiber, S., Rosenstiel, P., Franke, A. and Baines, J.F. (2011) Colonic mucosa-associated microbiota is influenced by an interaction of Crohn disease and FUT2 (Secretor) genotype. *Proc. Natl. Acad. Sci. U. S. A.*, **108**, 19030-19035.

60 Goto, Y., Obata, T., Kunisawa, J., Sato, S., Ivanov, II, Lamichhane, A., Takeyama, N., Kamioka, M., Sakamoto, M., Matsuki, T. *et al.* (2014) Innate lymphoid cells regulate intestinal epithelial cell glycosylation. *Science*, **345**, 1254009.

61 Pickard, J.M., Maurice, C.F., Kinnebrew, M.A., Abt, M.C., Schenten, D., Golovkina, T.V., Bogatyrev, S.R., Ismagilov, R.F., Pamer, E.G., Turnbaugh, P.J. *et al.* (2014) Rapid fucosylation of intestinal epithelium sustains host-commensal symbiosis in sickness. *Nature*, **514**, 638-641.

62 Kato, N., Loh, M., Takeuchi, F., Verweij, N., Wang, X., Zhang, W., Kelly, T.N., Saleheen, D., Lehne, B., Mateo Leach, I. *et al.* (2015) Trans-ancestry genome-wide association study identifies 12 genetic loci influencing blood pressure and implicates a role for DNA methylation. *Nat. Genet.*, **47**, 1282-1293.

63 Kinoshita, M., Matsuoka, Y., Suzuki, T., Mirrielees, J. and Yang, J. (2012) Sigma-1 receptor alters the kinetics of Kv1.3 voltage gated potassium channels but not the sensitivity to receptor ligands. *Brain Res.*, **1452**, 1-9.

64 Pang, T., Hisamitsu, T., Mori, H., Shigekawa, M. and Wakabayashi, S. (2004) Role of calcineurin B homologous protein in pH regulation by the Na+/H+ exchanger 1: tightly bound Ca2+ ions as important structural elements. *Biochemistry*, **43**, 3628-3636.

65 de Jonge, W.J. (2013) The Gut's Little Brain in Control of Intestinal Immunity. *ISRN Gastroenterol.*, **2013**, 630159.

66 Bourai, M., Lucas-Hourani, M., Gad, H.H., Drosten, C., Jacob, Y., Tafforeau, L., Cassonnet, P., Jones, L.M., Judith, D., Couderc, T. *et al.* (2012) Mapping of Chikungunya virus interactions with host proteins identified nsP2 as a highly connected viral component. *J. Virol.*, **86**, 3121-3134.

67 Chen, R., Giliani, S., Lanzi, G., Mias, G.I., Lonardi, S., Dobbs, K., Manis, J., Im, H., Gallagher, J.E., Phanstiel, D.H. *et al.* (2013) Whole-exome sequencing identifies tetratricopeptide repeat domain 7A (TTC7A) mutations for combined immunodeficiency with intestinal atresias. *J. Allergy Clin. Immunol.*, **132**, 656-664 e617.

68 Li, Y., Willer, C.J., Ding, J., Scheet, P. and Abecasis, G.R. (2010) MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet. Epidemiol.*, **34**, 816-834.

69 Howie, B.N., Donnelly, P. and Marchini, J. (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.*, **5**, e1000529.

70 Magi, R. and Morris, A.P. (2010) GWAMA: software for genome-wide association meta-analysis. *BMC Bioinformatics*, **11**, 288.

71 R-Core-Team. (2015) R: A language and environment for statistical computing. *R Foundation for Statistical Computing,Vienna, Austria.URL https://www.R-project.org/.R-Core-Team*, in press.

72 Pruim, R.J., Welch, R.P., Sanna, S., Teslovich, T.M., Chines, P.S., Gliedt, T.P., Boehnke, M., Abecasis, G.R. and Willer, C.J. (2010) LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics*, **26**, 2336-2337.

73 Yang, J., Lee, S.H., Goddard, M.E. and Visscher, P.M. (2011) GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.*, **88**, 76-82.

74 Nyholt, D.R. (2004) A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. *Am. J. Hum. Genet.*, **74**, 765-769.

75 Ward, L.D. and Kellis, M. (2012) HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res.*, **40**, D930-934.

76 Segre, A.V., Groop, L., Mootha, V.K., Daly, M.J. and Altshuler, D. (2010) Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemic traits. *PLoS Genet.*, **6**.

77 Pers, T.H., Karjalainen, J.M., Chan, Y., Westra, H.J., Wood, A.R., Yang, J., Lui, J.C., Vedantam, S., Gustafsson, S., Esko, T. *et al.* (2015) Biological interpretation of genome-wide association studies using predicted gene functions. *Nat. Commun.*, **6**, 5890.

78 Mishra, A. and Macgregor, S. (2015) VEGAS2: Software for More Flexible Gene-Based Testing. *Twin Res. Hum. Genet.*, **18**, 86-91.

Legends to Figures

Figure 1.A) Quantile-quantile (Q-Q) plots showing the probability values from GWAS meta-analysis of diarrhoeal disease at age 1y (D1Y). The line indicates the distribution under the null hypothesis. Lambda value is shown. **B)** Manhattan plot of the GWAS meta-analysis of diarrhoeal disease at age 1y (D1Y). The x-axis represents the autosomal chromosomes and the y-axis represents–$\log_{10}(p)$.The dotted line indicates genome-wide significance (p=5.00E-08), and the dashed line indicates suggestive genome-wide significance (p=1.00E-05).

Figure 2. Forest plots for rs8111874 at 19q13.33 for the four diarrhoeal outcomes: **A)** Diarrhoea at age one year (D1Y); **B)** Doctor diagnosis of diarrhoea at age one year (DD1Y); **C)** Diarrhoea at age two years (D2Y); **D)** Doctor diagnosis of diarrhoea at age two years (DD2Y). In the vertical panel, the studies participating in the discovery or replication phase are presented. In the horizontal lines, the sizes of the boxes represent precision and the lines the confidence intervals. The diamond shapes represent the pooled effect estimates, for both the fixed- and random-effect models. The horizontal axis shows the scale of the effect estimates. The effect allele is G, and the other allele is A.

Figure 3. A) Regional association of 19q13.33 locus (top SNP: rs8111874, chr19:49168942) with diarrhoeal disease at age one year (D1Y) (N= 5,758). The top variant is indicated with a diamond in purple and the flanking variants in circles, colored according to their linkage disequilibrium (LD). Variant rs601338 is shown in a black circle. The plot was constructed using the 1000 Genomes CEU population (Northern and Western European ancestry). **B)** Regional association of 19q13.33 locus (top SNP: rs8111874, chr19:49168942) with diarrhoeal disease at age one year (D1Y) conditioned to rs601338 (N= 5,758). The associations of the genetic variants in the region were attenuated.

Tables

Table 1. Samples included in the study at age 1 year and at age 2 years by diarrhoeal disease definition and study phase.

	Age 1 year									
	Diarrhoea (D1Y)					Doctor diagnosis of diarrhoea (DD1Y)				
	Cohort	Assessment (period in months)[a]	N total	N cases	% cases	Cohort	Assessment (period in months)[a]	N total	N cases	% cases
Discovery	ALSPAC (Disc)[d]	Questionnaire (6-18)	3363	2001	59.5	ALSPAC (Disc)[d]	Questionnaire (6-18)	3363	893	26.6
	Generation R	Questionnaire (6-12)[c]	2033	469	23.1	Generation R	Questionnaire (6-12)[c]	2033	179	8.8
	INMA_SAB	Questionnaire (6-14)	362	223	61.6	COPSAC2000	Doctor interview (6-12)	345	81	23.5
						LISAplus	Questionnaire (6-12)	662	186	28.1
	TOTAL		5758	2693	46.8	TOTAL		6403	1339	21.8
Replication	ALSPAC (Repl)	Questionnaire (6-18)	3361	2047	60.9	ALSPAC (Repl)	Questionnaire (6-18)	3361	871	25.9
	MoBa	Questionnaire (6-18)[c]	407	255	62.6	COPSAC2010	Doctor interview (6-12)	547	244	44.6
						INMA_VAL	Questionnaire (0-12)	334	149	44.6
						CHOP[b]	Medical records (6-18)	3223	147	4.6
	TOTAL		3768	2302	61.1	TOTAL		7465	1411	18.9
	Age 2 years									
	Diarrhoea (D2Y)					Doctor diagnosis of diarrhoea (DD2Y)				
	Cohort	Assessment (period in months)[a]	N total	N cases	% cases	Cohort	Assessment (period in months)[a]	N total	N cases	% cases
Discovery	ALSPAC (Disc)[d]	Questionnaire (18-30)	3189	1746	54.8	ALSPAC (Disc)	Questionnaire (18-30)	3189	514	16.1

	Generation R	Questionnaire (18-24)[c]	2058	943	45.8	Generation R	Questionnaire (18-24)[c]	2058	190	9.2
	INMA_SAB	Questionnaire (12-24)	361	166	46	COPSAC2000	Doctor interview (18-24)	319	86	27
						LISAplus	Questionnaire (18-24)	667	200	30
	TOTAL		5608	2855	50.9	TOTAL		6233	990	20.6
	ALSPAC (Repl)[d]	Questionnaire (18-30)	3187	1759	55.2	ALSPAC (Repl)	Questionnaire (18-30)	3187	485	15.2
	INMA_VAL	Questionnaire (12-14)	329	213	64.7	COPSAC2010	Doctor interview (18-24)	518	271	52.3
Replication						GINIplus[d]	Questionnaire (12-24)	794	328	41.3
						CHOP[b]	Medical records (18-30)	3223	190	5.9
	TOTAL		3516	1972	56.1	TOTAL		7722	1274	16.5

[a]Period refered in the questionnaire, medical record or doctor interview

[b]Only children not vaccinated against *Rotavirus* were included

[c]It refers specifically to gastric flu or gastroenteritis

[d]Statistically significant differences in the proportions of diarrhoeal disease among males and females (p-value < 0.05) [ALSPAC- D1Y-Discovery: 61.2% males vs 57.7% females; ALSPAC--DD1Y-Discovery: 29.2% vs. 23.8%; ALSPAC-D2Y-Discovery: 56.6% vs. 52.8%; ASLPAC -D2Y-Replication: 58.5% vs. 51.7%; and GINIplus- DD2Y- Replication: 46.7% vs. 36.0%].

Cohort designs: ALSPAC, COPSAC2010, Generation R, INMA_VAL, INMA_GIP, LISAplus, GINIplus, and MOBA are unselected population-based birth cohorts. COPSAC2000 is a prospective clinical study of a birth cohort of infants born to mothers with a history of asthma. CHOP is a random collection of paediatric patients from a hospital centre.

Table 2. Results from the fixed effect meta-analysis for diarrhoeal disease at age one year (D1Y) by discovery and replication phase. Variants with a p-value<1.E-05 in the discovery phase are shown.

				Discovery[a]											Replication[b]					Combined				
Marker	rs_number	Chr	Pos	EA	NEA	EAF	N	OR	lCI	uCI	p-value	p-value het	Effects	N variants/ locus	Gene	N	OR	lCI	UCI	p-value	N	OR	lCI	UCI	p-value
19:49168942:SNP	rs8111874	19	49168942	G	A	0.57	5758	1.32	1.21	1.44	1.06E-09	2.99E-02	+++	71	*NTN5* and *SEC1P* (intronic)	3768	1.25	1.13	1.39	1.69E-05	9526	1.29	1.21	1.38	8.05E-14
15:50562847:SNP	rs62020330	15	50562847	A	G	0.97	5396	2.16	1.61	2.88	2.02E-07	5.05E-02	++?	2	4.7kb 5' of *HDC*	3195	0.99	0.69	1.43	9.72E-01	8591	1.60	1.27	2.00	4.97E-05
7:2930941:SNP	rs1713926	7	2930941	C	T	0.69	5758	1.37	1.21	1.55	8.06E-07	5.68E-01	+++	1	15kb 3' of *CARD11*	1956	1.07	0.92	1.25	3.58E-01	7714	1.24	1.13	1.37	1.04E-05
7:63621349:SNP	rs139755348	7	63621349	C	T	0.73	5758	0.79	0.71	0.87	1.14E-06	9.29E-01	---	89	46kb 5' of *ZNF735*	3768	1.06	0.95	1.19	2.96E-01	9526	0.89	0.83	0.96	2.59E-03
7:1599067:SNP	rs112411182	7	1599067	T	C	0.95	5396	1.7	1.36	2.13	4.10E-06	8.75E-01	++?	1	3kb 5' of *TMEM184A*	3768	0.78	0.60	1.02	7.26E-02	9164	1.23	1.04	1.46	1.76E-02
12:130325960:SNP	rs34180477	12	130325960	G	A	0.93	5396	0.58	0.46	0.73	4.30E-06	4.93E-01	--?	1	*TMEM132D* (intronic)	3009	1.21	0.96	1.53	1.04E-01	8405	0.84	0.71	0.98	3.24E-02
14:38210286:SNP	rs74731421	14	38210286	G	A	0.97	5396	1.81	1.4	2.33	4.87E-06	8.07E-01	++?	1	146kb 5' of *FOXA1*	3768	0.91	0.69	1.22	5.42E-01	9164	1.34	1.11	1.63	2.32E-03
12:106438211:INDEL[c]		12	106438211	R	D	0.52	5758	1.26	1.14	1.39	4.95E-06	8.07E-01	+++	1	19kb of 3' *NUAK1*	407	1.20	0.85	1.68	3.01E-01	6165	1.26	1.14	1.38	2.73E-06
7:96366342:SNP	rs12704876	7	96366342	T	C	0.51	5758	0.83	0.77	0.9	5.04E-06	8.45E-01	---	1	27kb 5' of *SHFM1*	3768	1.10	1.00	1.20	4.94E-02	9526	0.93	0.88	0.99	2.74E-02
5:1695532:SNP	rs79411306	5	1695532	C	T	0.95	5396	0.55	0.42	0.71	6.17E-06	4.86E-01	--?	1	13kb 3' of *MIR4277*	3136	0.89	0.71	1.13	3.55E-01	8532	0.72	0.60	0.85	1.79E-04
5:40474267:SNP	rs116560909	5	40474267	C	T	0.97	5396	1.9	1.44	2.52	6.90E-06	5.52E-01	++?	2	206kb 5' of *PTGER4*	3768	1.08	0.78	1.49	6.63E-01	9164	1.50	1.21	1.85	2.03E-04
17:52598239:INDEL[c]		17	52598239	R	D	0.94	5396	1.57	1.29	1.92	6.90E-06	6.76E-01	++?	1	380kb 5' of *TOM1L1*	407	1.29	0.59	2.80	5.20E-01	5803	1.56	1.28	1.88	6.00E-06
2:116776614:SNP	rs12615869	2	116776614	A	G	0.94	5396	1.54	1.28	1.87	7.02E-06	5.62E-01	++?	2	174kb 3' of *DPP10*	3768	1.11	0.90	1.37	3.38E-01	9164	1.33	1.16	1.53	6.57E-05
4:184501675:SNP	rs7662749	4	184501675	C	T	0.72	5758	0.81	0.74	0.89	7.24E-06	8.53E-01	---	1	59kb 3' of *RWDD4*	3768	1.02	0.92	1.14	6.55E-01	9526	0.90	0.84	0.96	1.69E-03

22:49732481:SNP	rs5770255	22	49732481	G	A	0.76	5758	1.23	1.12	1.35	8.77E-06	7.32E-01	+++	2	281kb 3' of *C22orf34*	3768	0.94	0.84	1.05	2.83E-01	9526	1.10	1.03	1.19	5.95E-03

Only the most significant variant per locus is shown. Variants in the same locus are defined in a 1 Mb window. N variants/locus indicates the number of SNPs in the locus with a p-value < 1E-05

Only genetic variants with a minimal of 5,000 samples are shown

EA: effect allele; NEA: non effect allele; EAF: effect allele frequency; N: sample size; OR: odds ratio; lCI: 95% lower confidence interval; uCI: 95% upper confidence interval; p-value het: p-value for the heterogeneity test

[a]Cohorts in the discovery phase are included in alphabetical order: ALSPAC (Disc), Generation R and INMA_SAB

[b]Cohorts included in the replication phase are: ALSPAC (Repl) and MoBa

[c]ALSPAC (Repl) has no data for these markers

Bonferroni correction for 15 variants considered in the replication phase: p-value = 3.33E-03

Table 3. Conditional analysis at 19q13.33 locus (rs8111874 and rs601338).

							Crude results for rs8111874				Results for rs8111874 conditioned to rs601338				
Marker	rs_number	EA	NEA	EAF	IMP	Outcome	OR	lCI	uCI	p-value	N	OR	lCI	uCI	p-value
19:49168942:SNP	rs8111874	G	A	0.57	0.82-0.93	D1Y	1.32	1.21	1.44	1.06E-09	5214.7	1.08	1.01	1.14	1.40E-02
						DD1Y	1.26	1.13	1.39	1.17E-05	5576.7	1.05	0.98	1.12	2.00E-01
						D2Y	1.08	0.99	1.18	6.56E-02	5385.9	1.00	0.95	1.06	9.12E-01
						DD2Y	1.32	1.17	1.48	2.74E-06	5672.7	1.08	1.00	1.17	4.77E-02
							Crude results for rs601338				Results for rs601338 conditioned to rs8111874				
Marker	rs_number	EA	NEA	EAF	IMP	Outcome	OR	lCI	uCI	p-value		OR	lCI	uCI	p-value
19:49206674:SNP	rs601338	G (W)	A (X)	0.52	1	D1Y	1.28	1.18	1.39	2.74E-09	6224.8	1.06	1.00	1.12	3.93E-02
						DD1Y	1.25	1.14	1.37	1.99E-06	6872.8	1.07	1.01	1.14	2.48E-02
						D2Y	1.10	1.02	1.18	1.84E-02	6471.1	1.04	0.99	1.09	1.40E-01
						DD2Y	1.27	1.14	1.4	5.90E-06	6976.7	1.06	0.99	1.13	1.16E-01

EA: effect allele; NEA: non effect allele; EAF: effect allele frequency; IMP: imputation quality (from - to); D1Y: any diarrhoea at age one year; DD1Y: doctor's confirmed diagnosis of diarrhoea at age one year; D2Y any diarrhoea at age two years; DD2y doctor's confirmed diagnosis of diarrhoea at age two years

Distance between SNPs is 37.7 kb

Abbreviations

CI: confidence interval

D1Y: diarrhoea at age one year

DD1Y: doctor diagnosis of diarrhoea at age one year

D2y: diarrhoea at age two years

DD2Y: doctor diagnosis of diarrhoea at age two years

EA: effect allele

EAF: effect allele frequency

EMAC: expected minor allele counts

eQTL: expression quantitative trait

HWE: Hardy-Weinberg Equilibrium

MAF: minor allele frequency

NEA: non effect allele

OR: odds ratio

SNP: single nucleotide polymorphism

Figure 1

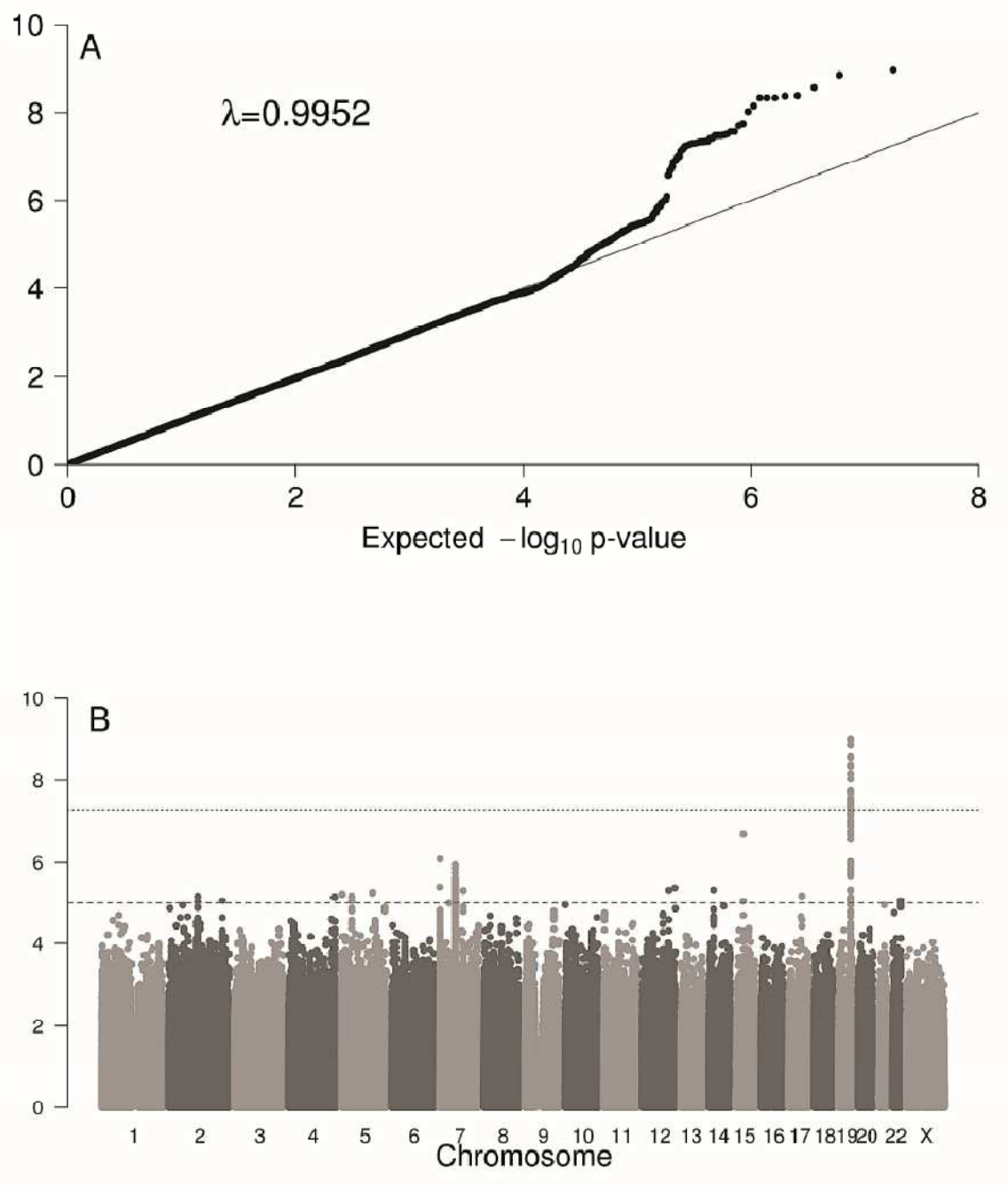

Figure 2

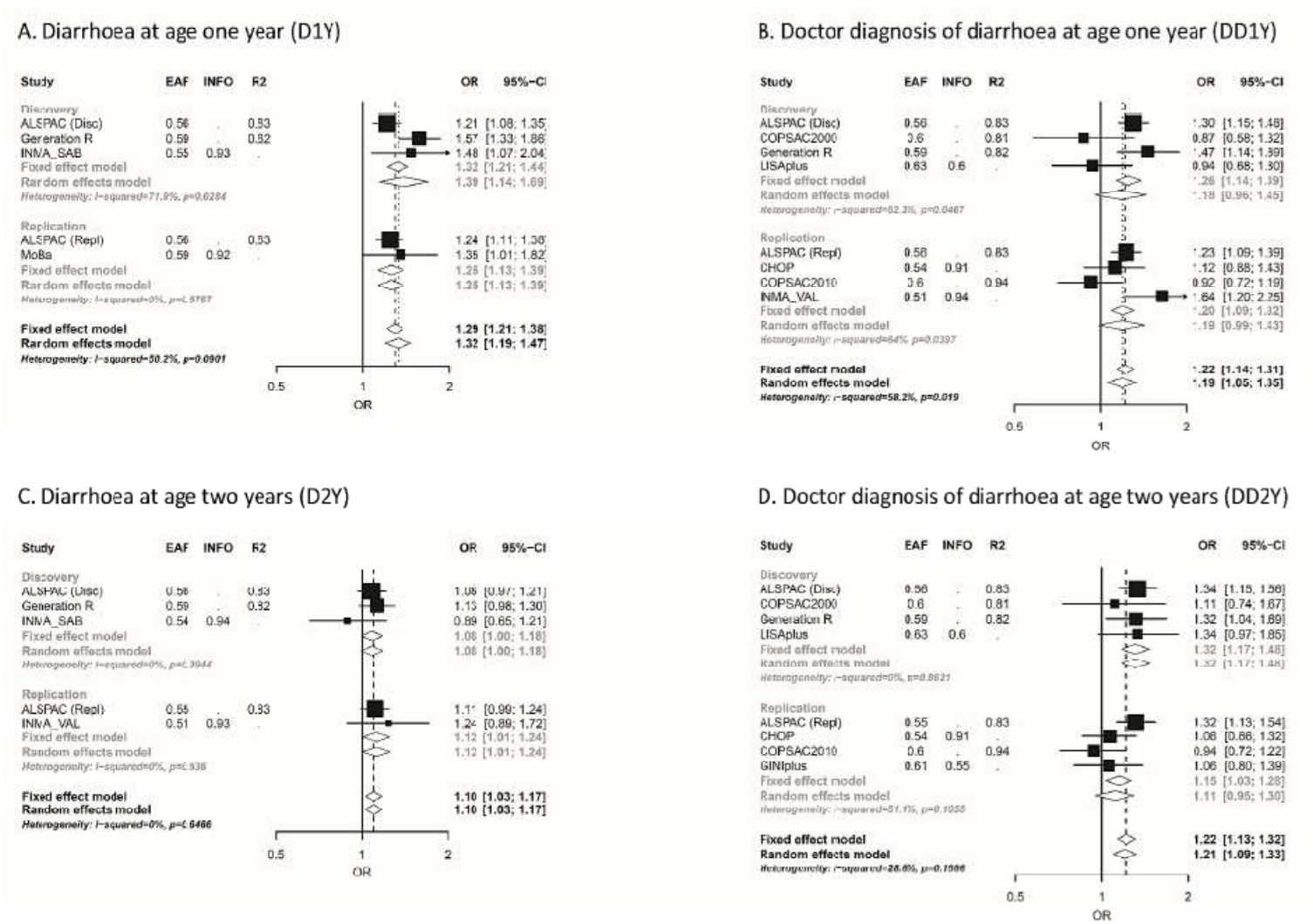

Figure 3